medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Genome-wide association study reveals the unique genetic structure of active blood donors

Jonna Clancy*[1], Jarkko Toivonen[1], Jouni Lauronen[1], Satu Koskela[1], Jukka Partanen[1], FinnGen[2], Mikko Arvas[1], Jarmo Ritari[1]

[1] Finnish Red Cross Blood Service, Vantaa, Finland

[2] List of FinnGen members is available as a supplementary file

* jonna.clancy@veripalvelu.fi

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

The healthy donor effect (HDE) refers to the observed lower mortality rate among blood donors compared to the general population. While membership biases such as HDE arise due to healthier individuals being more likely to participate, the extent to which it is influenced by genetic differences remains largely unclear. To elucidate the genetic basis of HDE, we conducted a genome-wide association study (GWAS) involving 53,688 active blood donors with extensive donation histories and 228,060 controls from biobank cohorts within the FinnGen project. Our results identified 2,973 genome-wide significant loci associated with repeated blood donation and impacting several health-related endpoints. Importantly, significant variants related not only to blood groups but also influenced the predisposition to somatic and mental diseases, suggesting that HDE is not solely tied to blood donation eligibility criteria. In conclusion, in this work we show that HDE is partially explained by genetic factors affecting various disease categories.

Keywords: blood donor, biobank, healthy donor effect, GWAS, blood group, HLA

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Blood donors play an important role as an excellent healthy control cohort in genetics and biomedical research due to their commitment and altruistic nature[1]. Their voluntary and frequent blood donations offer unique opportunities for longitudinal sampling. In addition, blood donors exhibit positive attitudes toward the use of their samples and data in scientific research, if not directly required in patient care[2,3].

However, blood donors are a highly selected population due to the health status criteria they must meet to be eligible to donate blood. Along with a general survey of the donor's current state of health, basic restrictions on blood donation suitability include serious illnesses such as severe cardiovascular conditions, cancer and epilepsy[4]. Some conditions that restrict blood donation suitability have strong well-known genetic background, such as the association of *HLA* genes in type 1 diabetes (T1D)[5] or in certain autoimmune diseases[6–8]. As a minimum hemoglobin level is required for donation eligibility, genetic variants affecting iron and hemoglobin levels may also cause donor selection. Due to these factors, blood donor selection results in a known phenomenon, *healthy donor effect*, HDE, which can lead to detrimental effects in a research setting if not taken into account[9].

Comprehensive selection of different blood groups, and especially a sufficient amount of emergency blood type O RhD neg suitable for most recipients, are basic requirements in blood banks. Due to the high need for emergency blood O RhD neg and strong immunogenicity of certain blood group antigens, such as ABO, Rh and Kell[10–13], donor recruitment is partially based on blood group genotype and results in enrichment of certain blood group antigens in a blood donor pool. Well known associations between ABO blood group antigens and cardiovascular conditions exist, more precisely between red cell antigen O and reduced thromboembolic diseases[14,15]. Risk for bleeding disorder, Von Willebrand disease, is known to be higher among individuals with red cell O antigen[16]. Pregnancy-induced hypertension has been previously described to occur more in RhD positive

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

individuals[15]. Moreover, in certain medical conditions, such as in hemochromatosis, blood donation may be beneficial, and donors with these conditions may be enriched in the blood donor population[17].

Enrichment of genetic variants with known disease associations or biological functions may result in unpredicted effects in biomedical research. To reveal the genetic qualities of blood donors and to understand the possible genetic selection leading to the HDE, we conducted a genome-wide association study (GWAS) between the Finnish Red Cross Blood Service Biobank, and FinnGen (Figure 1). In addition, our aim was to discover genetic variants potentially contributing to blood transfusion efficacy.

To our knowledge, no previous genome studies have focused on the possible selection bias caused by blood group and health questionnaire-based requirement of blood donors. We found altogether 2,973 genome-wide significant associations. Genetic correlation analysis revealed strong negative correlation between blood donorship and diseases such as mental disorders, pain disorders, cardiovascular diseases, addiction disorders, autoimmune disorders, and Alzheimer's disease. *HLA* association analysis confirmed the enrichment of hemochromatosis related *HLA* alleles and the lower occurrence of autoimmune related *HLA* alleles in the blood donor population and indicated a risk SNP tagging both hemochromatosis and autoimmune related *HLA* alleles. The genome-wide significant variants associated with 112 plasma protein level alterations. The results of this study demonstrate the genetic basis of the HDE, which is not limited to immediate blood donorship inclusion criteria.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

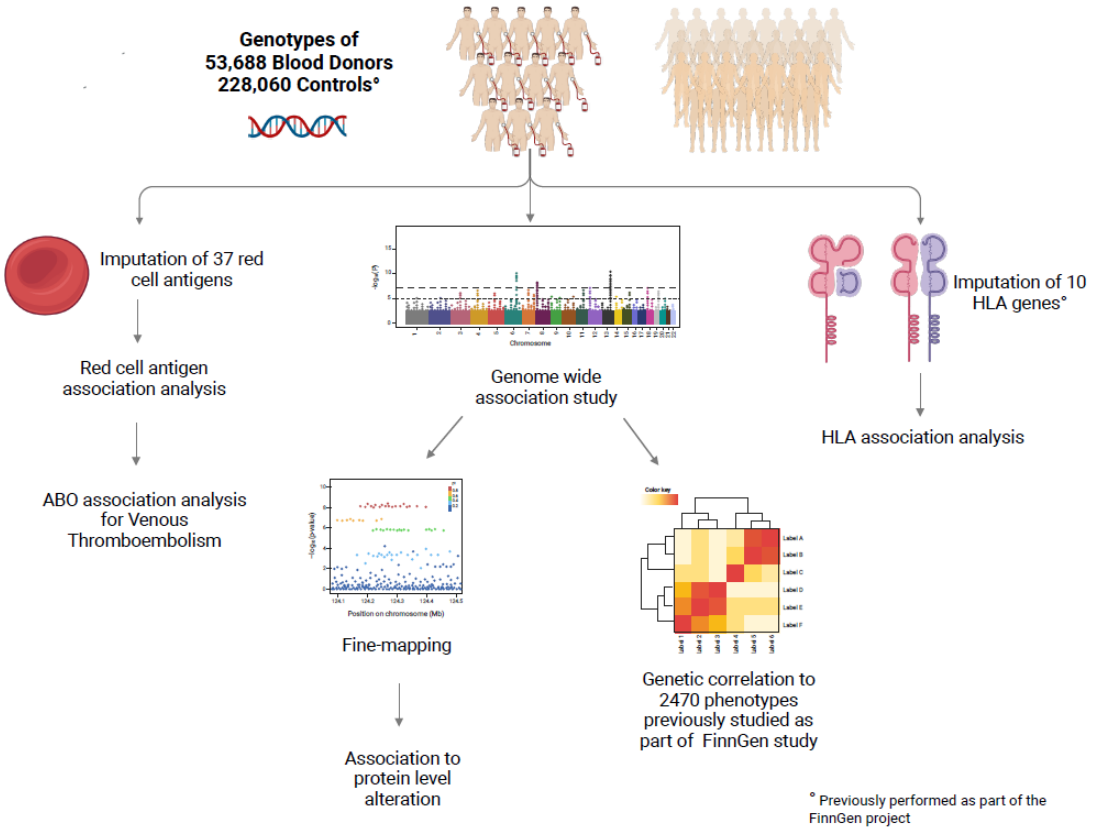

Figure 1 General workflow of the study. GWAS was performed on 53,688 cases (blood donors) and 228,060 controls. Genome wide significant GWAS results were fine-mapped and further used for several association analyses. Genetic correlation of 2470 phenotypes to blood donorship were calculated based on GWAS results. In addition, we imputed red cell antigens and tested their association to blood donorship. Previously imputed HLA alleles in FinnGen project were used to perform association analysis to blood donorship. Figure is created with BioRender.com

Results

Demographic information and differences in diagnoses between the two cohorts are described in Table 1. In both sexes the blood donors were younger, taller and in males also heavier. More female participants were seen in the blood donor cohort. A clear and significant difference between the two cohorts was seen in both sexes in all the diagnose groups apart from ´XVI Certain conditions originating in the perinatal period´ where more incidents were found among blood donors in both

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

genders. FinnGen cohort consisted of donors in hospital biobanks but also donors in population cohorts and in occupational health (Supplementary Table 1).

	Females		Males	
	Blood donors	**FinnGen**	**Blood donors**	**FinnGen**
Median age at sampling, years (IQR)	44.47 (18.2-70.8)	62.7 (36.2-89.2)	50.26 (25.5-75.0)	70.46 (51.3-89.6)
Female sex, n (%)	32,460 (60.46)	120,279 (52.74)	21,228 (39.54)	107,781 (47.26)
Median weight, kg (IQR)	72 (54-90) [a]	72 (51-93) [a]	86 (68-104)	84 (64-104)
Median height, cm (IQR)	167 (159-175)	164 (156-172)	180 (171-189)	177 (168-186)
Median BMI (IQR)	25.83 (19.3-32.4)	26.64 (19.1-34.2)	26.58 (21.6-31.6)	26.88 (21.2-32.6)
FinnGen endpoint category [b]				
I Certain infectious and parasitic diseases (AB1)	7,466 (23)	47,901 (39.82)	5,565 (26.22)	48,487 (44.99)
II Neoplasms from cancer register (ICD-O-3)	968 (2.98)	34,299 (28.52)	791 (3.73)	38,891 (36.08)
II Neoplasms from hospital discharges (CD2)	6,960 (21.44)	67,026 (55.73)	2,804 (13.21)	52,708 (48.9)
III Blood and immune system (D3)	964 (2.97)	17,017 (14.15)	534 (2.52)	15,606 (14.48)
IV Endocrine, nutritional and metabolic (E4)	10,269 (31.64)	83,780 (69.65)	3,029 (14.27)	58,530 (54.3)
IX Circulatory system (I9)	6,276 (19.33)	69,429 (57.72)	4,587 (21.61)	78,876 (73.18)
V Mental and behavioural disorders (F5)	7,871 (24.25)	42,676 (35.48)	3,648 (17.18)	33,080 (30.69)
VI Nervous system (G6)	6,431 (19.81)	54,710 (45.49)	4,501 (21.2)	51,142 (47.45)
VII Eye and adnexa (H7)	8,088 (24.92)	58,567 (48.69)	4,660 (21.95)	47,174 (43.77)
VIII Ear and mastoid process (H8)	5,316 (16.38)	29,249 (24.32)	3,893 (18.34)	26,561 (24.64)
X Respiratory system (J10)	14,408 (44.39)	69,022 (57.38)	11,532 (54.32)	69,209 (64.21)
XI Digestive system (K11)	27,542 (84.85)	106,938 (88.91)	17,037 (80.26)	90,482 (83.95)
XII Skin and subcutaneous tissue (L12)	7,710 (23.75)	43,402 (36.08)	4,096 (19.3)	32,486 (30.14)
XIII Musculoskeletal and connective tissue (M13)	13,715 (42.25)	83,931 (69.78)	9,118 (42.95)	68,570 (63.62)
XIV Genitourinary system (N14)	17,334 (53.4)	91,665 (76.21)	4,463 (21.02)	49,729 (46.14)
XIX Injury, poisoning, external (ST19)	15,936 (49.09)	73,725 (61.29)	13,051 (61.48)	70,375 (65.29)
XV Pregnancy, childbirth and the puerperium (O15)	20,560 (63.34)	86,799 (72.16)	0 (0)	0 (0)
XVI Perinatal period (P16)	535 (1.65)	1,000 (0.83)	263 (1.24)	514 (0.48)
XVII Congenital malformations (Q17)	2,018 (6.22)	10,032 (8.34)	1,231 (5.8)	7,258 (6.73)
XVIII Abnormal clinical and laboratory findings (R18)	16,800 (51.76)	93,458 (77.7)	9,125 (42.99)	80,419 (74.61)
XX External causes of mortality (VWXY20)	13,025 (40.13)	64,874 (53.94)	10,741 (50.6)	62,034 (57.56)
XXI Contact with health services (Z21)	24,581 (75.73)	10,4068 (86.52)	10,544 (49.67)	72,915 (67.65)
XXII Codes for special purposes (U22)	1,417 (4.37)	5,906 (4.91)	595 (2.8)	4,310 (4)

[a] Not significant
[b] Shortened for clarity

Table 1 Demographic information of the cohorts. Age, sex, height, weight, BMI and FinnGen endpoint categories in blood donors and FinnGen cohort are shown. Apart from weight in females, difference between the cohorts in each phenotype was statistically significant.

Genome-wide association analysis

Genome-wide association analysis revealed 2,973 significant ($p < 5x10^{-8}$) genetic loci associated with blood donorship. After fine-mapping, 5 coding (4 missense and 1 inframe deletion) and 36 non-coding variants were detected (Figure 2A). In addition to these 41 variants, 5 non-coding lead

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

variants in the extended MHC region (Figure 2B and C) were included in the downstream analyses (Table 2).

The strongest genome-wide associations were seen in variants related to blood group antigens: rs55794721, p = 8.88×10^{-89}, in chromosome 1, rs687621, p = 5.70×10^{-81}, in chromosome 9 in *ABO* and rs8176058, p = 1.16×10^{-31}, in chromosome 7 in *KEL* (Figure 1A, Table 2).

Despite of the adjustment of the GWAS for BMI, we observed significant GWAS variants known to influence height, weight, and BMI. rs66781921, an inframe deletion variant in chr2, previously shown to have a negative association to height, weight and body mass index[18], was rarer in blood donors (beta = -1.97, p = 1.89×10^{-11}). Missense variant rs34811474, previously shown to be negatively associated with body mass index[18], was seen to be more frequent in blood donors (beta = 0.064, p = 1.21×10^{-12}). Variants rs10947114 (p = 2.88×10^{-09}, beta = 0.055), rs3130906 (p = 5.17×10^{-11}, beta = 0.055) and rs4713637 (p = 1.39×10^{-08}, beta = -0.047), previously described to be associated with spondylopathies and celiac disease[18], were negatively associated with blood donorship. A rare intron variant, rs528492111, in chromosome 3 previously shown to have a weak association to benign neoplasm of conjunctiva[18], was more frequent in blood donors (beta = 1.87, p = 4.09×10^{-10}).

All the fine-mapped and MHC lead variants in the are listed in Table 2.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

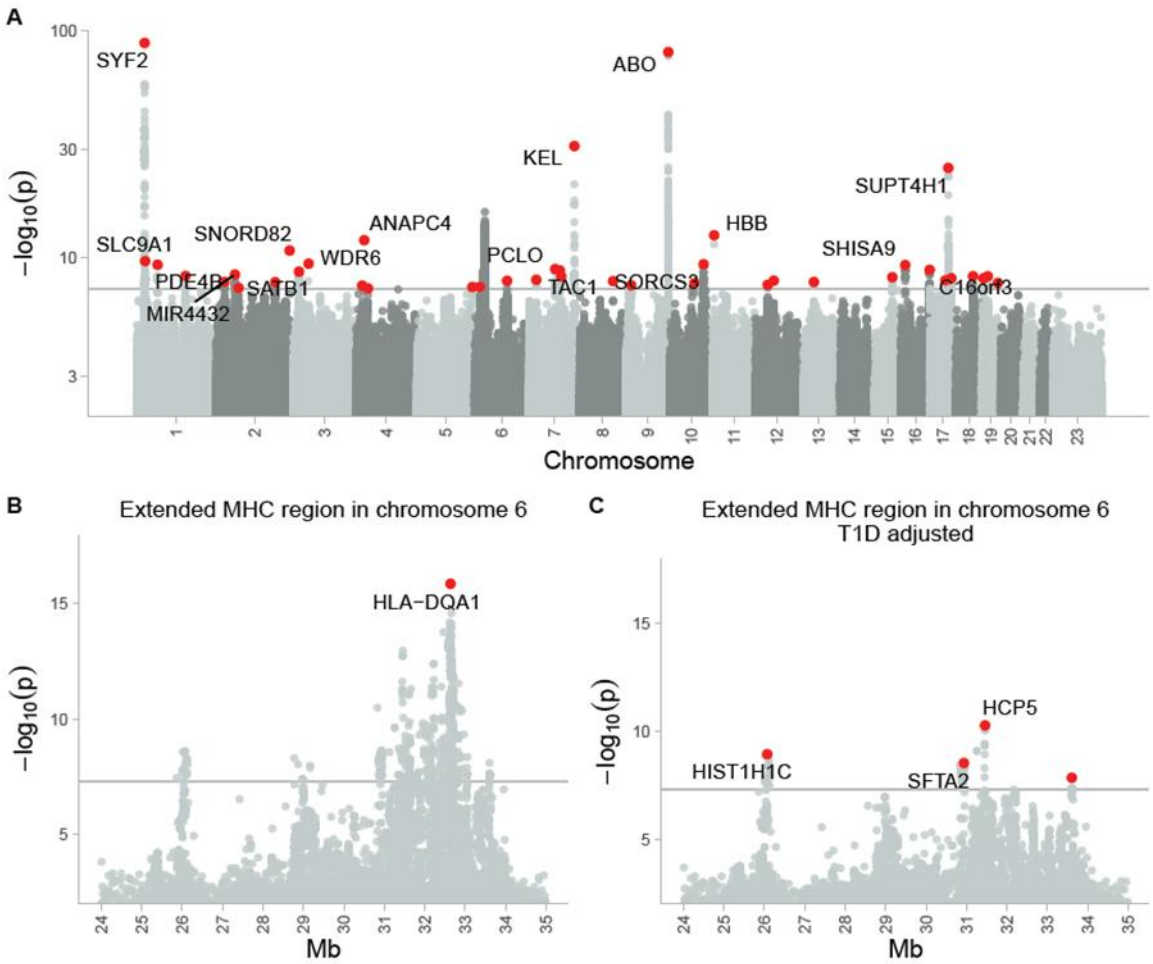

Figure 2 A) Manhattan plot of the genome-wide association study of blood donorship. B) Extended MHC region of the chromosome 6 of the GWAS. C) Chromosome 6 association study adjusted for type 1 diabetes, T1D. Red dot indicates fine-mapped variants in panel A and in panels B and C the highest variant in each peak reaching the genome wide significance, $<5 \times 10^{-8}$. The name of the nearest gene is shown for these variants if the p-value is $\leq 5 \times 10^{-9}$. The horizontal line has been set to 5×10^{-8}.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

	Genome wide association study					pQTL
rs ID	p-value	Beta	Effect	Target gene	Nearest	Target protein
rs66781921	1,89E-11	-1,973	inframe deletion	NCL	SNORD82	-
rs1135071	2,93E-13	-0,573	missense	HBB	HBB	AHSP
rs199598395	1,55E-25	-0,383	missense	RNF43	SUPT4H1	FTL, HAMP, FGF23
rs8176058	1,16E-31	-0,327	missense	KEL	KEL	KEL
rs191302298	3,89E-08	-0,292	intron	LINC01411	MSX2	-
rs10096658	1,35E-08	-0,277	intron	ZFPM2	LRP12	-
rs62113181	8,06E-09	-0,135	intron	KDM4B	KDM4B	-
rs75738358	4,89E-08	-0,080	intergenic	NA	MIR4801	-
rs12539059	5,68E-09	-0,071	downstream gene	COL26A1	LINC01007	KLK11
rs9272324	1,36E-16	-0,064	intron		HLA-DQA1	HERC5, METTL2B, MICB, HLA-DRA, C4B*
rs36035346	5,48E-09	-0,062	intron	CRTC1	CRTC1	-
rs60678519	1,06E-08	-0,055	downstream gene	AC079780 ,1	RPS2P32	GPNMB
rs10947114	2,88E-09	-0,055	intron		SFTA2	DDR1
rs3130906	5,17E-11	-0,055	intron		HCP5	MICA, CDSN, MICB
rs80319585	1,92E-08	-0,055	intron	KCNMA1	KCNMA1	-
rs7795945	1,28E-09	-0,049	intron	PCLO	PCLO	-
rs34483988	1,65E-08	-0,049	intron	FAM171B	FAM171B	-
rs6775319	2,17E-09	-0,048	intron	SATB1-AS1	SATB1	-
rs7907026	4,63E-10	-0,048	intergenic	NA	SORCS3	-
rs3806234	5,16E-09	-0,047	upstream gene	SNAPIN	SNAPIN	-
rs4713637	1,39E-08	-0,047	intergenic		LINC00336	-
rs62083414	5,35E-09	-0,046	intron	DCC	DCC	-
rs207350	3,03E-08	-0,045	intergenic	NA	LCORL	-
rs594398	1,18E-08	-0,044	intron	SUMO2P17	ATP5G1	-
rs10880819	2,64E-08	0,043	downstream gene	RF00019	ALG10B	-
rs35440643	4,54E-08	0,044	intron	ZNF638	ZNF638	-
rs10166897	1,57E-08	0,044	intron	MAPRE3	MAPRE3	CGREF1
rs113236240	6,53E-09	0,045	intron	PPCDC	PPCDC	MPI, PPCDC
rs55763604	2,84E-08	0,049	intron	BNC2	BNC2	-
rs55732343	2,31E-10	0,049	regulatory region	NA	SLC9A1	-
rs9968910	1,12E-09	0,050	intergenic	NA	HIST1H1C	-
rs5821444	7,98E-09	0,051	3 prime UTR	C17orf58	C17orf58	-
rs6498415	5,47E-10	0,053	intron	U91319 ,1	SHISA9	-
rs12444334	1,51E-09	0,053	intron	GAS8	C16orf3	CDH15
rs11675100	3,88E-09	0,054	intron	AC007100 ,1	MIR4432	-
rs9367942	3,81E-08	0,055	intron	STMND1	STMND1	-
rs61937595	1,25E-08	0,062	intron	R3HDM2	R3HDM2	INHBC
rs34811474	1,21E-12	0,064	missense	ANAPC4	ANAPC4	-
rs1500956	5,14E-10	0,075	intron	PDE4B	PDE4B	-
rs77448558	1,28E-08	0,078	intergenic	NA	GRIK2	-
rs10268629	1,71E-09	0,078	intergenic	NA	TAC1	-
rs75350584	1,58E-08	0,094	intron	DLEU1	ST13P4	-
rs687621	5,70E-81	0,147	intron	ABO	ABO	63 different
rs55794721	8,88E-89	0,157	upstream gene	SYF2	SYF2	ICAM4
rs118101548	1,90E-08	0,299	intron	POLD1	POLD1	-
rs528492111	4,09E-10	1,875	intron	WDR6	WDR6	-

Table 2. Fine-mapped and MHC lead variants. Significance, effect size, the most serious effect and nearest gene for each variant is shown. If the variant was associated with altering protein expression levels, the target proteins are shown.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Genetic correlation

Altogether 593 phenotypes were negatively correlated with blood donorship (FDR < 0.05, Supplementary file 2). Height was the only phenotype showing weak positive correlation with blood donorship, r_g = 0.07 (Figure 3A) despite the GWAS having been adjusted for BMI. Phenotypes associated with some of the individual variants detected by the GWAS with a relatively strong effect size did not show strong correlation with blood donorship. In turn, some phenotypes associated with individual variants with small effect size in GWAS were relatively strongly correlated to blood donorship (Figure 3B).

Blood donorship was negatively correlated with r_g < -0.5 to 30 phenotypes (Supplementary file 2). The strongest negative correlation (r_g < -0.6) in blood donors was seen with Anemias (r_g = -0.66, FDR = 2.81 x 10^{-112}), ´Other reaction to severe stress and adjustment disorders´ (r_g = -0.64, FDR = 3.36 x 10^{-40}) and ´Other disorders of teeth and supporting structures´ (r_g = -0.62, FDR = 2.74x10^{-03}). Furthermore, negative correlation was observed in phenotypes such as autoimmune diseases (r_g = -0.33, FDR = 1.87 x 10^{-52}), cardiovascular diseases (excluding rheumatic etc) (r_g = -0.42, FDR = 3.79 x 10^{-64}) and bleeding disorders (r_g = -0.39, largest FDR = 7.10 x 10^{-11}) (Figure 3C). In addition, negative correlation to blood donorship was observed in phenotypes such as addiction disorders (r_g = -0.33, largest FDR = 7.09 x 10^{-04}), pain disorders (r_g = -0.45, largest FDR = 1.41 x 10^{-10}) and Alzheimer's disease (r_g = -0.24, FDR = 0.03) which are not directly restricted by blood donation eligibility criteria, but which undeniably effect on donation suitability or donor's knowledge on their current state of health. Another example of phenotypes not directly restricted by blood donation eligibility but with relatively strong negative correlation towards blood donorship are infectious phenotypes, such as acute bronchitis (r_g = -0.58, FDR = 1.76 x 10^{-16}) and 'other or unspecified bacterial infection' (r_g = -0.50, FDR = 1.83 x 10^{-09}) (Figure 3A, Supplementary file 2).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

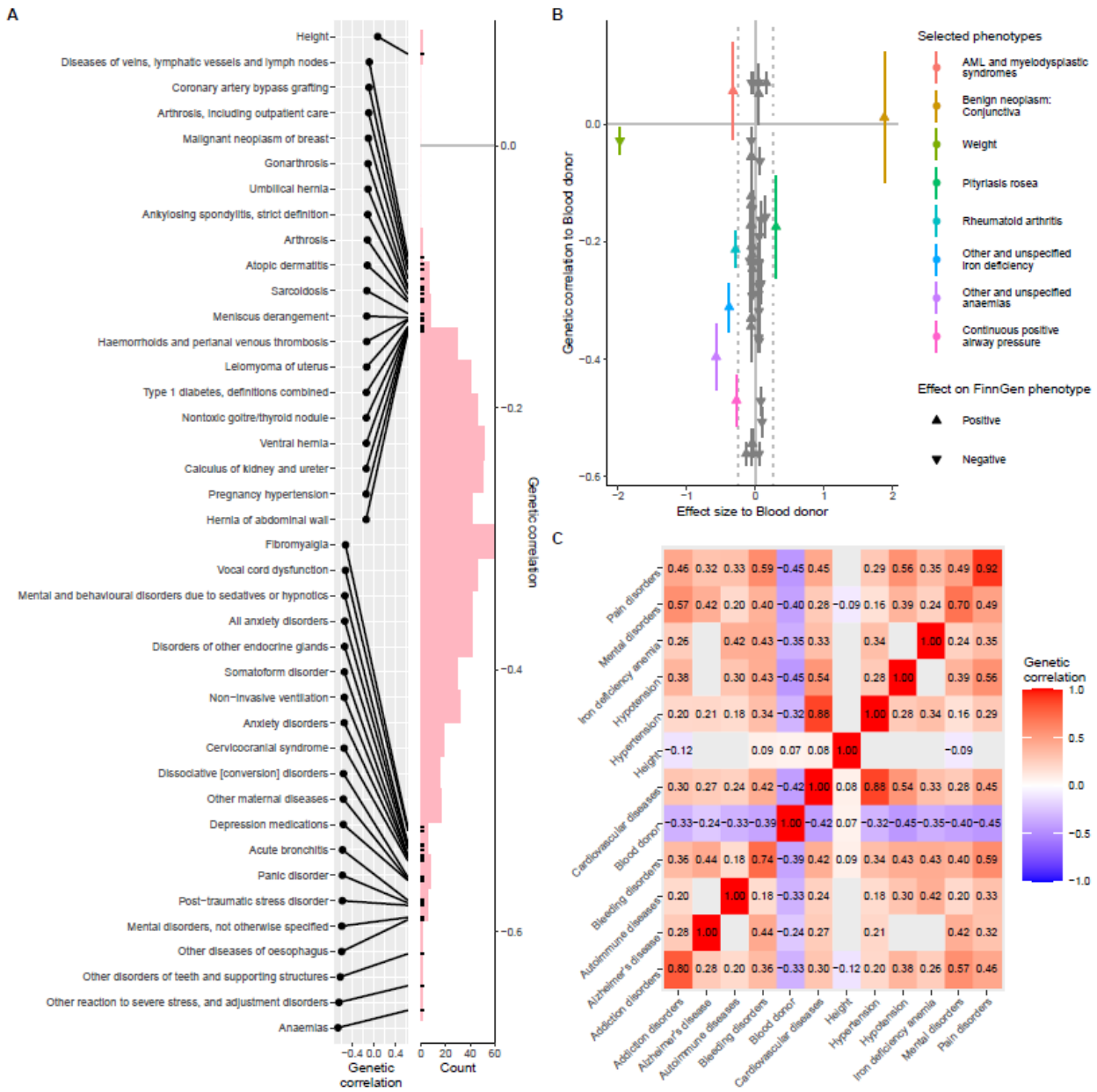

Figure 3. Genetic correlation of blood donorship to 594 phenotypes. A) Genetic correlation of 40 phenotypes representing the strongest negative and strongest positive correlation to blood donorship. B) For each of the 46 variants and their most significantly associated FinnGen phenotypes the effect size of the variant and the genetic correlation of the phenotype to blood donorship is shown. The phenotype is labeled when the absolute value of the effect size is higher than 0.25. The vertical bars indicate the standard error of the genetic correlation estimate. C) Genetic correlation matrix of selected phenotypes. Pain disorders, mental disorders, bleeding disorders and addiction disorders have been grouped from several phenotypes belonging to same phenotype category and average genetic correlation is shown. This results in some of the diagonal elements being less than one.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Red cell antigen related associations

To understand whether blood donor selection leads to enrichment of blood groups in the blood donors, we evaluated the association of imputed red cell antigens with blood donorship using a logistic regression model. The model revealed a strong negative association for blood group antigens Kell, RhD, and AB and a positive association for blood group antigen O (Figure 4A). Results of the GWAS covariates and red cell antigens are available in Supplementary file 2.

We further analyzed the blood group related significant GWAS variants in chromosomes 1, 7 and 9 with linear regression models. Lead variant in chromosome 1, rs55794721, was associated with the Rh blood group system (previously shown to be associated with height[18]). Rh antigens E, C and D were negatively associated with rs55794721, whereas e and c antigens of the most common Rh negative haplotype in the Caucasoid population, dce[11], were positively associated with rs55794721 (Figure 4B). 91% of individuals that were positive for Rh c and e alleles and had no Rh D, C or E alleles, were homozygotes (dosage value > 1.5) for rs55794721. rs55794721 is in LD[19] (R^2 = 0.39) with variant rs586178 tagging certain RhCE blood group alleles[20]. Lead variant in chromosome 7, rs8176058, a known tag SNP for Kell blood group antigen[20] was positively associated with K antigen (Figure 4C). Lead variant in chromosome 9, rs687621[19], is in strong LD, R^2 = 1 with blood group O tagging SNP rs687289[21]. We could also see a strong positive association of rs687621 with blood group O, a positive association with blood group antigen B and a negative association with blood group antigen AB (Figure 4D).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

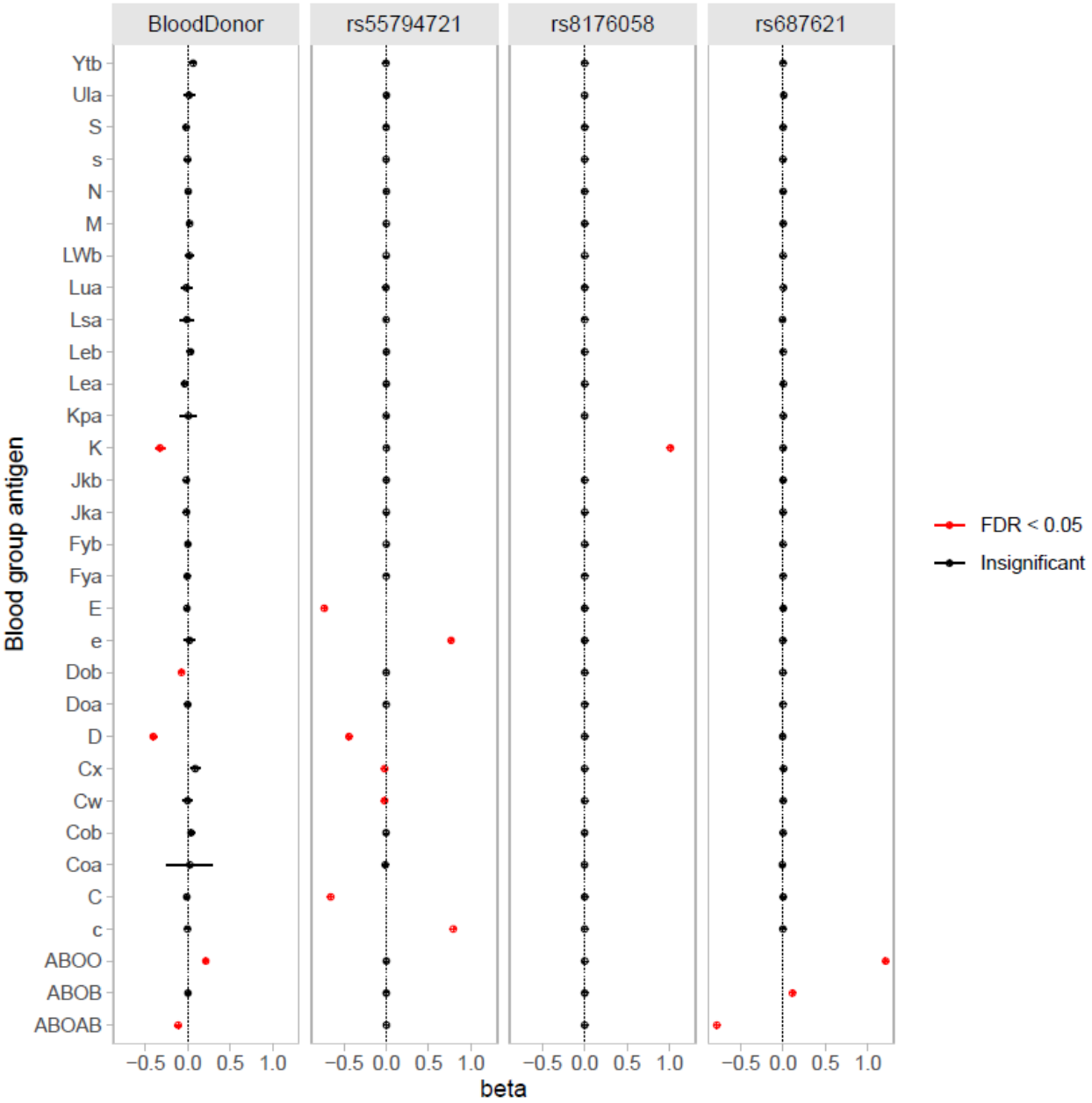

Figure 4 Association of red blood cell antigens with blood donorship, rs55794721, rs8176058 and rs687621. Red cell antigens K, Dob, D and AB were negatively, and red cell antigen O positively associated to blood donorship. Red cell antigens c and e belonging to the most common Rh-negative haplotype Rh dce were positively associated to rs55794721, whereas red cell antigens D, C and E were negatively associated with rs55794721. Red cell antigens Cw and Cx were negatively associated with rs55794721. Red cell antigen K was positively associated with rs8176058. Red cell antigens O and B were positively associated and red cell antigen AB negatively associated with rs687621 in ABO gene.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Hemoglobin related associations

As hemoglobin value within a specified interval is a required for blood donation eligibility, we investigated whether some of the genome-wide significant variants were associated with hemoglobin. Hemoglobin-related associations were seen in chromosomes 6, 9, 11 and 17 in variants rs9968910, rs3130906, rs687621, rs1135071 and rs199598395, respectively (Supplementary Figure 1). We observed a negative association between rs199598395 in RNF43 gene in chromosome 17 and blood donorship. rs199598395 has been previously shown to be Finnish enriched and associated with iron deficiency anemia[22] and to have an effect on blood donation eligibility[23]. Negative association in blood donorship was also seen with rs1135071 in chromosome 11, in a known Finnish enriched Hb variant, Hb Tacoma[24], in HBB gene. Linear regression analysis of blood service operational finger-prick Hb measurements showed a decreasing effect of rs199598395, rs1135071 and rs3130906 to hemoglobin level -2.99 g/L ($p = 6.43 \times 10^{-30}$, 95% CI [-3.49, -2.5]), -6.46 g/L ($p = 6.66 \times 10^{-28}$, 95% CI [-7.57, -5.35]) and -0.22 g/L ($p = 0.002$, 95% CI [-0.32, -0.11]) respectively, and moderate increasing effect of rs9968910 and rs687621, 0.52 g/L ($p = 4.03 \times 10^{-22}$, 95% CI [0.42-0.62]) and to 0.25 g/L ($p = 1.72 \times 10^{-05}$, 95% CI [0.15-0.35]), respectively (Supplementary file 2). In summary, all the alleles of these five variants that had an increasing effect on hemoglobin, were more frequent in blood donors (Table 2).

We detected a higher allele frequency, 0.69, of rs9968910, previously shown to be associated with diseases of iron metabolism[18], among patients with clinical hereditary hemochromatosis (ICD10 code E83.1) in FinnGen (N=477) compared to population level in Finland, 0.33[25]. rs9968910 is in weak LD with known HFE mutations causing hemochromatosis: rs1799945 (H63D, $R^2 = 0.18$) and rs1800562 (C282Y, $R^2 = 0.1$)[19].

Association of the *HLA* alleles

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Because of the known strong role of the HLA region in immunology and the predisposing risk of certain HLA alleles in autoimmunity[5–8,26], we performed an association study with imputed HLA alleles for blood donor endpoint. We could see a strong association ($p = 1.36 \times 10^{-16}$) with rs9272324 in chromosome 6 in HLA-DQA1 gene in the genome-wide association study of blood donorship (Figure 1B). However, after adjusting with T1D (ICD10 E10 and ICD9 250), no significant association in *HLA-DQA1* gene was seen (Figure 1C). *HLA* association analysis for the blood donor phenotype (Figure 5, upper panel) revealed negative association with HLA-alleles of the well-known T1D risk genotypes *HLA-DQ2* and *DQ8* which are in linkage with HLA haplotypes *DRB1*03:01-DRB3*01:01-DQA1*05:01-HLADQB1*02:01* and *DRB1*04:01- DRB4*01:03- DQA1*03:01- DQB1*03:02,* respectively. The *HLA* alleles in linkage with HLA-DQ8 were negatively associated with blood donorship: *DRB1*04:01* (beta = -0.12, $p = 2.2 \times 10^{-11}$), *DRB4*01:03* (beta = -0.06, $p = 2.75 \times 10^{-8}$), *DQA1*03:01* (beta = - 0.09, $p = 1.16 \times 10^{-8}$) and *DQB1*03:02* (beta = -0.09, $p = 9.99 \times 10^{-9}$). HLA-alleles previously shown to be associated with HFE C282Y allele in Finland[17] *B*07:02* (beta = 0.07, p = 5.45e-10^{5}), *DRB1*15:01* (beta = 0.06, $p = 5.45 \times 10^{-5}$), *DQA1*01:02* (beta = 0.05, p = 5.45 x -10^{-5}) and *DQB1*06:02* (beta = 0.06, p = 9.24 x -10^{-5}) were positively associated with blood donorship. *HLA-DRB5*01:01* (beta = 0.06, $p = 5.45 \times 10^{-5}$) and *HLA-DPB1*05:01* (beta = 0.1, p = 0.04) were positively associated with blood donorship. Furthermore, *HLA-B*27:05* (beta = -0.07, $p = 1.96 \times 10^{-3}$), a well-known ankylosing spondylitis risk allele[26], was negatively associated with blood donorship. Adjustment with chr6 lead variant, rs9272324 (Figure 5, middle panel), decreased the effect sizes and increased p-values of the above-mentioned *HLA* alleles in linkage with *HLA-DQ8* and *HLA-B*27:05* and increased the p-values of the above mentioned HFE C282Y related *HLA*-alleles above the significance level (Supplementary file 2). Furthermore, when adjusted with the lead *HLA*-allele, *DRB1*04:01* (Figure 5, lower panel), *HLA-B*27:05* (beta = -0.08, $p = 2.84 \times 10^{-04}$) and *HLA-C*02:02* (beta = -0.06, $p = 2.95 \times 10^{-02}$) in LD ($r^2 = 0.3$) with *HLA-B*27:05*, HLA-alleles in linkage with the T1D risk genotype *HLA-DQ2*; *DRB1*03:01* (beta = -0.05, $p = 4.31 \times 10^{-03}$), *DQA1*05:01* (beta = -0.06, p = 2.45×10^{-03}), *DQB1*02:01* (beta = -0.06, $p = 4.31 \times 10^{-03}$) and *HLA-B*08:01* (beta = -0.05, p = 4.2 x 10^{-}

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[02]), a known *HLA-B* allele in in linkage with *HLA-DQ2* in Finland, were negatively associated with blood donorship (Figure 4, lower panel). In *HLA-DRB1*04:01* adjusted analysis, HFE C282Y related *HLA* alleles *B*07:02*, *DRB1*15:01*, *DQA1*01:02* and *DQB1*06:02* were positively associated with blood donorship, but with lower p-values than in *HLA*-association analysis without any adjustment (Supplementary file 2).

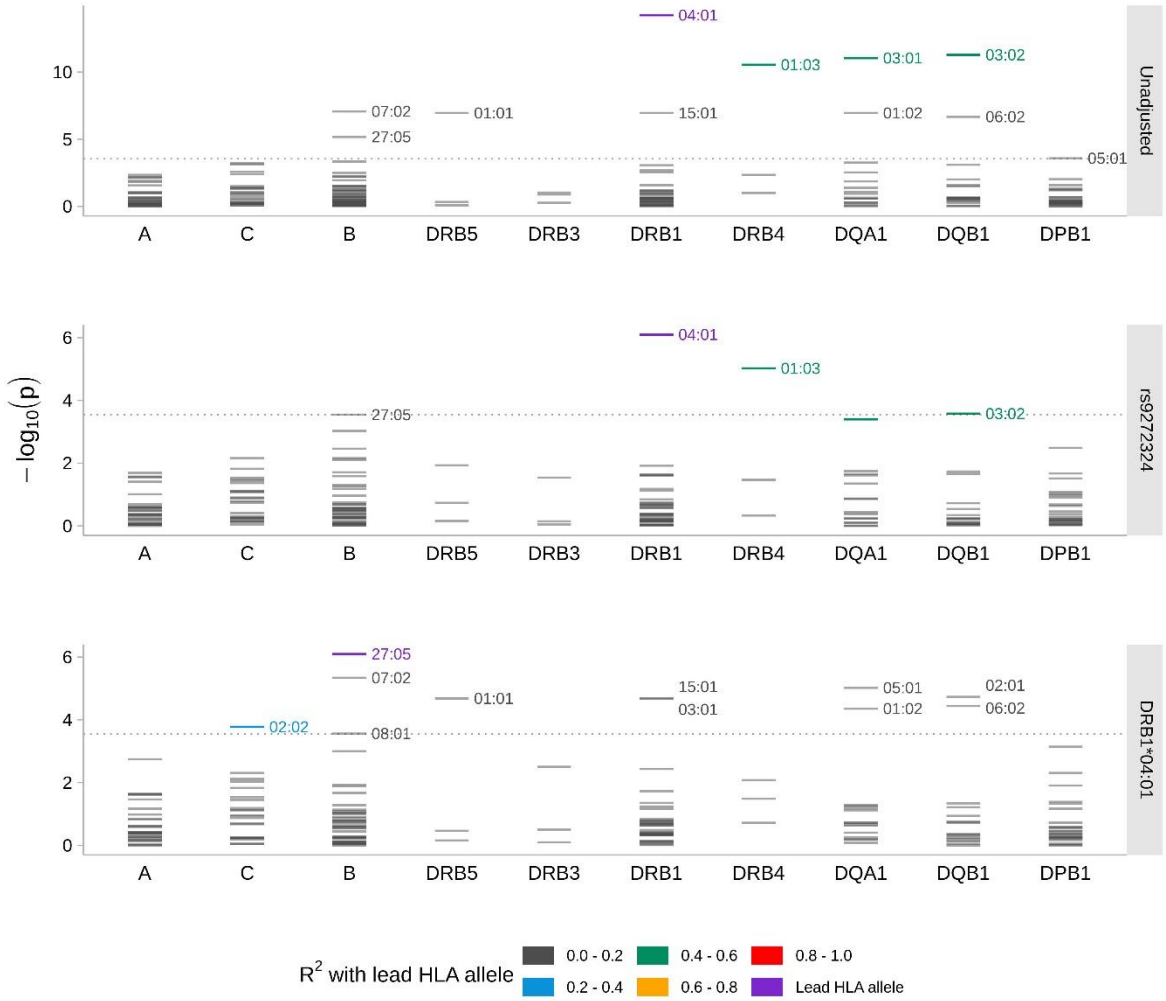

Figure 5. Conditional *HLA* association analysis. Upper panel: *HLA* association analysis for blood donorship. Middle panel: *HLA* association analysis for blood donorship adjusted with rs9272324. Lower panel: *HLA* association analysis for blood donorship adjusted with *HLA-DRB1**04:01. Covariates used in all the analysis: age, sex, BMI, PC1-10, birth region and FinnGen genotyping array version. Dotted horizontal line corresponds to the highest p-value that still fits the FDR limit of 0.05.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Variants altering protein expression levels

To investigate the function of the lead GWAS variants more closely, we employed protein quantitative trait loci (pQTL) and functional enrichment analyses. Of the 41 fine-mapped and five MHC region variants (Table 2), 14 were found to be significantly associated (FDR <0.05) with 81 unique plasma protein expression levels. Most of the associations, comprising 63 unique proteins, were found with rs687621 located in the *ABO* gene. These included genes that encode for proteins fundamental to the histo-blood group ABO formation, such as *ABO* antigens, and *FUT1*, the *H*-antigen. The latter is converted by glycosyltransferase expressed by individuals with *A*, *B* or *AB* into antigen *A* or *B* while individuals with the *O* blood group antigen lack this activity[27]. Decreasing effects on expression were seen for e.g. *FUT1*, *VWF*, *F8* and *CLEC4M*, and increasing effects for *ALPI* and *SELE*. Importantly, rs55794721 exerted a decreasing effect on *ICAM4*, a gene encoding for LW glycoprotein present in the RhD protein complex (Figure 6). In addition, two variants were associated with protein expression levels of genes putatively associated with carcinogenesis (Figure 6). rs12539059 regulates protein levels of *KLK11*, a kallikrein suspected to play a role in carcinogenesis in human tissues[28]. Another putatively cancer-related[27] protein association was seen for rs60678519 and *GPNMB*. Overexpression of GPNMB has been shown to lead into immunosuppression within tumor immune microenvironment and further into poorer prognosis in variety of cancers[29].

The STRING[30] functional enrichment analysis tool identified several functional processes (PPI enrichment p-value < 1×10^{-16}) among the 81 proteins. Gene Ontology (GO) revealed the enrichment of cell adhesion molecules in biological processes and molecular functions. These molecules were further refined to ICAM, C-type lectin and immunoglobulins in protein domains. Other functional

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

categories were blood groups, hemoglobin and iron metabolism, immunology, and cell-cell adhesion (Supplementary file 3).

We detected three associations with protein expression level alteration that weren't involved with any of the enriched biological or cellular processes (Supplementary file 3). rs113236240 showed an increasing effect (beta = 0.36) on *PPCDC*, encoding for a protein needed in coentzyme A biosynthesis[27]. rs687621 showed a decreasing effect (beta = -0.4) on *GOLM2*, a proto-oncogene whose overexpression associates with proto-oncogene *HER2/neu*, found in human breast and ovarian cancers[27]. rs1135071 showed a strong increasing effect (beta 2.57) on *AHSP* (Figure 6), which encodes for chaperone protein binding free alpha-globin and therefore having a stabilizing effect in hemoglobin assembly[27].

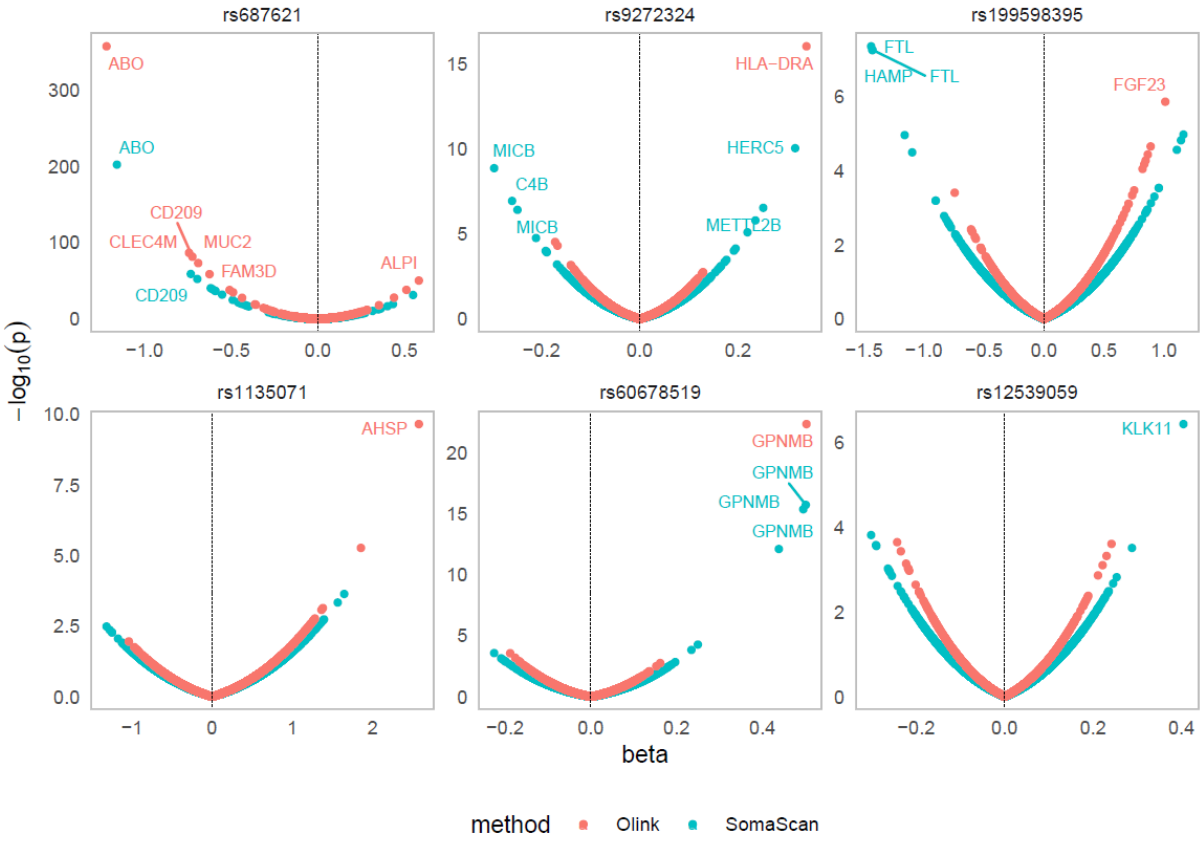

Figure 6 Association of the variants to protein expression level alteration. Six variants with association to protein level alteration in pQTL analysis are shown in the figure. The lead variant in ABO gene rs687621, was found to be associated to 63 unique protein level alterations. Lead variant in MHC region in unadjusted GWAS, rs9272324, showed association mainly to immunity related protein level alteration. Variants with decreasing effect on hemoglobin level (Supplementary file 2)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

rs199598395 and rs1135071 showed association to iron metabolism genes and hemoglobin related gene, respectively. rs6067859 and rs12539059 showed increasing effect on expression of putative cancer-related genes. All the variants effecting on protein levels can be found in Table 2 and in more detail in Supplementary file 3.

Putative bias caused by red cell antigen enrichment in research setting

Protective association of red cell antigen O to venous thromboembolism (VTE) has been shown previously by multiple studies[14–16]. To investigate the putative bias an ABO enrichment could cause to VTE association study, we conducted an association analysis with VTE as an endpoint in chromosome 9 (Supplementary Figure 2A) by adjusting for imputed ABO antigens (Supplementary Figure 2). Our results showed that after the adjustment, the p-value of the lead variant failed to reach the genome-wide significance level 5×10^{-8} (rs582118 unadjusted $p = 3.28 \times 10^{-115}$, ABO adjusted $p = 1.16 \times 10^{-6}$) but the signal was not completely removed as some genome-wide significant variants still remained (Supplementary Figure 2A-B). The adjustment not only removed associations of some SNPs but also made visible other variants not detected in the basic analysis (Supplementary Figure 2C).

We could see a strong association of rs55794721 (beta = 0.16, $p = 8.88 \times 10^{-89}$) to blood donorship and furthermore we showed the association of this variant to RhD negative haplotype, Rh dce (Figure 4). rs55794721 is previously shown to be associated with height[18]. The blood donors are taller than the control population (Table 1) and due to the strong immunogenity of RhD blood group system[11], the Rh dce tagging rs55794721 is enriched in blood donor population (Table 2). To understand the possible bias the RhD negative enrichment could cause, we conducted an association study for height and adjusted the primary analysis with cohort information. Our results show that RhD dce enrichment can cause bias in research setting. Prior the adjustment, rs55794721 was statistically significant ($p = 1.74 \times 10^{-13}$) but after the adjustment failed to reach reduced the genome-wide significance level 5×10^{-8} (after adjustment $p = 6.86 \times 10^{-7}$). Furthermore, the effect

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

size of the variant varied depending on the adjustment (rs55794721 unadjusted beta = 0.14, adjusted beta = 0.09) (Supplementary Figure 3).

Discussion

In the present study we conducted a genome-wide comparison between blood donor biobank donors with relatively long and active blood donation history[17] and FinnGen. The FinnGen control population included biobank participants from hospital biobanks as well as from population cohorts and patients at occupational health clinics. As the latter two cohorts comprise largely healthy individuals, and the disease-specific legacy cohorts were excluded from the analysis, the study controls are likely representative of the general Finnish population.

Negative genetic correlation towards blood donorship was observed with several phenotypes, which could be a consequence of HDE. Strong negative correlation to anaemias in blood donors could be a result of blood donation eligibility but just as well certain autoimmune diseases, such as Crohn's disease, which may cause anaemia through secondary blood loss.

The observed protection of blood donors from certain diseases can be directly explained by donor eligibility criteria that selects healthy individuals. For example, blood donors lack viral hepatitis due to the exclusion of infected individuals. Surprisingly, our results reveal that blood donors not only exhibit a protective genetic profile against viral hepatitis, but also a broader negative genetic correlation extending beyond viral hepatitis to infectious diseases in general, even though these are not explicitly targeted by donor eligibility criteria. This kind of negative genetic correlation also manifests in various mental illnesses and Alzheimer's disease. While pleiotropic effects may explain the phenomenon, our analyses indicate that genetic correlation with the phenotype targeted by the eligibility criteria is not consistently strong. Further detailed investigations are needed to fully

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

understand pleiotropy in blood donor selection and to explore the potential benefits of an active blood donor population for genetic and biomedical research.

Expectedly, genetic variants in the MHC region played strong role in several genetic associations as well as in genetic correlation. HLA alleles with known autoimmune associations occurred with lower frequency in the blood donor population than in the control population, whereas hemochromatosis related HLA alleles were more frequent among the blood donors. We have previously shown the enrichment of *HFE C282Y* homozygotes and certain *HLA* haplotype within these donors in the Blood Service Biobank[17]. These findings were repeated in the present study with new variants, rs9968910, more frequent among patients with clinical hemochromatosis diagnosis (ICD10: E83.1) and rs9272324 tagging several autoimmune related *HLA* haplotypes, including hemochromatosis related *HLA* haplotype. However, we could not see the ancestral hemochromatosis marker *HLA-A*03:01*[31,32] in the HLA-association analysis. *HLA-DPB1*05:01* hasn't been previously shown to be associated with the *HLA-A*03:01-C*07:02-B*07:02-DRB1*15:01-DQA1*01:02-DQB1*06:02* haplotype but instead has been shown to be enriched in with *HLA-A*02:01-C*07:02-B*07:02-DRB1*15:01-DQA1*01:02-DQB1*06:02*[33]. However, our unpublished data shows that *HLA-A*03:01-C*07:02-B*07:02-DRB1*15:01-DQA1*01:02-DQB1*06:02-DPB1*05:01* haplotype occurs in the Blood Service Biobank with a frequency of 0.00011. Taken the enrichment of *C282Y* homozygotes in the Blood Service Biobank and the positive association of these *HLA* alleles with blood donorship in this study into consideration, it could be possible that this haplotype is associated with clinical hemochromatosis. However, more studies are needed to confirm the *HLA* haplotype among donors with clinical hemochromatosis.

In addition to tagging certain autoimmune related haplotypes, rs9272324 was shown to enhance the protein levels of *HLA-DRA* encoding the nonpolymorphic *DR* alpha-chain[34] essential for the stabile expression of DR molecules on cell surface. Taken this and the known autoimmune associations of *HLA-DRB1*15:01* into consideration, it could be possible that rs9272324 plays a role in the known

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

immunological abnormalities often met in patients with clinical hemochromatosis. Furthermore, the significantly higher frequency of rs9968910 and rs9272324 among individuals with clinical hemochromatosis diagnosis, could shed a light in the long discussed low penetrance of the well described *HFE* variants *C282Y* and *H63D* in clinical hemochromatosis.

The strong immunogenicity of certain blood group antigens[10–13] as well as the constant need of the emergency blood, O RhD neg, causes enrichment of these red cell antigens into the blood donor population as demonstrated in the present study. Multiple studies have shown the relationship between ABO blood group and cardiovascular conditions and more precisely the decreased risk of blood group O to thromboembolic end points[15,35]. In research setting where blood donors are used as a healthy control population, confounding factors as the blood group O enrichment should be taken into consideration as shown in this study by VTE association analysis (Supplementary Figure 2). Furthermore, the blood group antigen related variants rs55794721 allele A and rs687621 allele A decreased the expression level of genes known to have a role in the function or structure of the cognate blood group molecule, such as RhD accessory protein ICAM4, a Landsteiner-Wiener glycoprotein[11], and FUT1, fucosyltransferase needed in the synthesis of red cell antigens A and B[27,36]. rs687621 alone altered expression levels of 63 genes, whose products are involved in various of biological pathways (Supplementary file 3). Due to the blood group O enrichment in the blood donor pool, these included e.g., decreasing effect on proteins with a role in hemostasis and thrombosis such as VWF and F8. An increasing effect of protein levels of rs687621 allele A was seen with ALPI and SELE, the former having a role in the defense system of gut mucosal and is thought to prevent bacterial translocation and the latter been speculated to have role in atherosclerosis. Both of these immunity and cardiovascular related biological functions are well-known roles of ABO blood group antigens[14,15,36,37]. Furthermore, another red cell antigen related finding in this work was the variant rs55794721, which was shown to tag the most frequent RhD negative haplotype in the European ancestry, Rh dce.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

In addition to hemochromatosis, the results of the present study highlight genetic factors with possible effect on recipient matching or donation activity. We show a lower expression level of AHSP protein in blood donors due to the lower occurrence of rs1135071 allele A, which was also shown to decrease the hemoglobin level in this work. Consequently, blood donors had higher hemoglobin levels. AHSP plays a role in hemoglobin assembly by stabilizing the αHb subunits and preventing the self-aggregation of the excess αHb subunits prior the formation of HbA1 subunit[38]. To our understanding, no earlier studies indicate the increasing effect of *Hb Tacoma* variant rs1135071 on AHSP levels. Taken the speculated role of AHSP as a genetic modifier in β thalassemia[38,39] and the described normal clinical phenotype of *Hb Tacoma* heterozygotes in account[40], rs1135071 could be of further interest in β thalassemia related research. Furthermore, regarding the donor health, it may be useful to analyze whether blood donation is a harmless event for *Hb Tacoma* heterozygotes or whether the blood products are suitable for recipients with e.g. sickle cell disease[41]. The known association of rs199598395 with iron deficiency[18], its effect on blood donation eligibility[23] and the decreasing effect on hemoglobin level demonstrated in this study support the need for evaluating its role in blood donation eligibility or interval in more detail.

In biomedical research, active blood donors offer valuable opportunities due to their positive attitude toward research, availability for longitudinal sampling and unique genetic makeup. However, when using blood donors exclusively as a control group, critical evaluation and interpretation of results are essential. For instance, our VTE and height association analysis highlights the need to consider possible blood group enrichment and its impact on disease risk to adjust for cohort information to minimize biases. Moreover, genome-wide association studies in FinnGen have been conducted without adjustment for the cohort information. Hence, the possibility of HDE or blood group enrichment caused biases in the FinnGen results and therefore also in the phenotype associations of this study, cannot be completely excluded for all the variants as was shown with our height association study. In addition, Blood Service Biobank is one of the cohorts used for FinnGen association studies[18,22], which results are utilized on this study. Although blood

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

donors represent minority in FinnGen project, the putative accumulating impact on the effect sizes and p-values of this study, cannot be completely excluded.

In this study we present, to our knowledge for the first time, the genetic underpinnings of HDE and blood group-based selection in active blood donors. Our findings highlight the unique genetic makeup of this donor population and its impact on health and blood donation suitability. The results demonstrate the role of genetics in maintaining good health status and active blood donation career that is also connected to a better mental health at the genetic level. Furthermore, research using blood donors as a control group should account for this distinct genetic structure, even after donation ceases. Additionally, the enrichment of specific blood group antigens among blood donors warrants consideration in research. Further investigations are necessary to fully elucidate the implications of these discoveries, but they underscore the genetic structure behind HDE and the potential of active blood donor cohorts for advancing biomedical genetic research.

Material and Methods

Ethics statement

Study subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017. Full list of permits numbers are listed in Supplementary information.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

All the sample donors had given a written biobank consent. The use of the data in the present study is in accordance with the biobank consent and meets the requirements of the Finnish Biobank Act 688/2012. A formal biobank decision 004-2022 for the research project has been granted by the Finnish Red Cross Blood Service Biobank. The use of the FinnGen data has been accepted with FinnGen proposal F_2022_086.

Data acquisition

Genome data used in the present study originated from samples that were genotyped in the FinnGen project with the FinnGen ThermoFisher Axiom custom array v1 or v2 and genome imputation version R12. Genotyping, quality control, and genome imputation protocols are described in detail in FinnGen Gitbook[42]. In short, genotype calling was performed with GenCall or zCall algorithm for Illumina and AxiomGT1 algorithm for Affymetrix arrays. Prior the imputation, genotyped samples were pre-phased with Eagle 2.3.5 with the default parameters, except the number of conditioning haplotypes was set to 20,000. Genotype imputation was performed using the population-specific imputation reference panel SISu v4.2 with Beagle 4.1 (version 27Jan18.7e1).

In addition to genome data, the following electronic health registry data and supplementing data were utilized: ICD (ICD9 and 10) code for type 1 diabetes, occurrence of venous thromboembolism, age, BMI, sex, region of birth and genetic principal components (PC) 1-10. Blood donation data contained blood donation history and hemoglobin value measured at the point of each blood donation event and was used when applicable. Donors with three or less donation events were filtered from the hemoglobin analyses. Blood donation data was available from 44,675 donors.

Due to the known selection of healthy individuals as blood donors, legacy samples[42], highly enriched in certain common diseases, were filtered out from the study cohort to minimize possible bias in downstream analyses. Individuals under 18 years of age were removed as 18 is the age limit of blood

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

482 donation. In addition, individuals born in regions in Finland that were ceded to Soviet Union were
483 filtered out due to age structure of the region and the maximum blood donation age of 70 years. As
484 BMI might influence donor eligibility, individuals lacking BMI information were filtered out. After
485 filtering, total number of FinnGen genome data used in this study was 53,688 cases (blood donors)
486 and 228,060 controls (other FinnGen participants).

487

488 Methods

489 Statistical analyses

490 If not stated otherwise, all statistical analyses were performed in R[43] version 4.3 or 4.4, with
491 RStudio[44]. The genome-wide significance p-value threshold $p < 5 \times 10^{-8}$ was used in GWAS. In all
492 other analyses p-values were adjusted with the Benjamini–Yekutieli procedure[45], and significance
493 threshold of 0.05 was used. All the phenotypes described here, have been previously studied and
494 described in FinnGen in DF12 dataset. At the present, only DF10 data set is publicly available[18].

495

496 Association analysis and fine-mapping

497 Genome-wide association study, GWAS, with an additive genetic model, was performed in FinnGen
498 pipeline with Regenie[46,47] v2.2.4 with blood donor status as a binary endpoint. Sex, age, BMI, PC1-10,
499 birth region, and FinnGen genotyping array version were used as covariates in the GWAS. To
500 discover associations independent of type 1 diabetes (T1D) in chromosome 6, the FinnGen T1D
501 endpoint [22] was used as an additional covariate in the chromosome 6 association analysis.

502 Fine-mapping for the genome wide significant signals ($p < 5 \times 10^{-8}$) of the association study was
503 performed in FinnGen pipeline[47] with SuSiE[48] excluding MHC region (GRCh38, chr6:25-34 Mb)[22,42].
504 For each fine-mapped locus we accepted the credible set as an independent hit if the strongest
505 primary p-value was $p < 5 \times 10^{-8}$. In case of multiple credible sets within a single locus, to accept the

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

secondary hit, we required genome-wide significance and $\log_{10}$(Bayes factor) > 2. As the MHC region is excluded from the fine-mapping pipeline due to its strong LD, the lead variants from the T1D adjusted association analysis, rs9968910, rs10947114, rs3130906 and rs4713637, were selected for the downstream analysis. Moreover, the lead variant in the MHC region in GWAS, rs9272324, was included in the further analyses due to its strong association in autoimmunity[18] and the hypothesized lower association with autoimmunity among blood donors.

Association analysis for all the covariates used in GWAS and the imputed blood groups to the blood donor end point was performed with logistic regression in R. We further analyzed the association of all the covariates used in GWAS and imputed blood group antigens to the lead variants in dosage form in chromosomes 1 (rs55794721), 7 (rs8176058) and 9 (rs687621) by using linear regression.

Association testing of each classical *HLA* allele *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQA1*, *-DQB1*, *DPB1* and *HLA-DRB3-5* for the blood donor endpoint was performed using Regenie v3.0.1. First, a genotype data set for fitting Regenie null models was formed with Plink2 v2.00a4LM using FinnGen DF12 HapMap3 variants as a basis and randomly thinning the variants to 500 000 with further filtering for minor allele frequency at least 0.1, Hardy-Weinberg equilibrium greater than p-value 1e-15 and minimum allele count at least 100. Second, the Regenie null models were fitted on the filtered genotype data and used in the association step to take consanguinity into account. To dissect the HLA signal further, two additional association tests were performed: i) conditioning for GWAS chromosome 6 lead variant rs9272324 allele dosage and ii) conditioning for the HLA allele dosage with the strongest association, *HLA-DRB1*04:01*, in the primary HLA association analysis.

To understand the putative bias caused by red cell antigen, ABO, enrichment in the blood donor pool, we conducted an association study in chromosome 9 in FinnGen pipeline with Regenie[46,47] v2.2.4 with FinnGen venous thromboembolism[22] as a binary endpoint. Sex, age, PC1-10, FinnGen genotyping array version, and FinnGen batch information were used as covariates. In the primary analysis there were 19,957 cases and 344,672 controls. The primary analysis was conditioned with

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

imputed ABO red cell antigens. After filtering out 20 individuals based on low posterior probability of the imputed ABO antigens, 19,956 cases and 344,653 controls remained in the conditional analysis.

To further understand putative bias in FinnGen association results, we conducted an association study in chromosome 1 in FinnGen pipeline with Regenie[46,47] v2.2.4 with FinnGen height[22] as a non-binary endpoint. Sex, age, PC1-10, FinnGen genotyping array version, and FinnGen batch information were used as covariates. The study had 364,629 cases. The primary analysis was conditioned with cohort information, to minimize the possible bias of RhD dce enriched haplotype in blood donors.

Genetic correlation

All pair-wise genetic correlations between blood donorship, hemoglobin and all the 2470 phenotypes previously studied as part of FinnGen study were computed using LD Score Regression v1.0.1 using Finnish LD panel in FinnGen pipeline[47]. The computation was based on the GWAS summary statistics previously obtained using Regenie for each phenotype. We used the Benjamini-Yekutieli method to control FDR of genetic correlation at level 0.05. Genetic correlation between groups of phenotypes was approximated as the average of all the pair-wise genetic correlations between the groups.

HLA and blood group imputation

The alleles of the classical HLA genes, *HLA-A*, *-C*, *-B*, *-DRB1*, *DRB3-5*, *-DQA1*, *-DQB1*, and *-DPB1*, had been previously imputed at four-field resolution (i.e., defining protein sequence variation) in FinnGen using HIBAG[49] algorithm with population-specific refence panel as reported earlier[50]. The dosage value of each HLA-allele was used in downstream analyses. Altogether 37 Red Blood cell, *RBC*, antigens in 14 different blood group systems were imputed using random forest models as described by Hyvärinen et al[51]. After imputation, *P1*, *A1* and *A2* antigens were excluded from

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

downstream analyses due to uncertain frequency at the population level. In addition, blood group antigens *HrS* and *HrB*, occurring mainly in African population, are in high correlation with blood group antigen *e* and have low clinical significance (consultation with expert at Finnish Red Cross Blood Service), hence they were excluded from further analyses. After imputation, the samples were filtered based on the posterior probability, PP, value; samples with PP < 0.5 were considered as antigen negative and samples with PP ≥ 0.5 were considered as antigen positive. Altogether 13 individuals were excluded from downstream blood group related analyses due to all PP values in ABO being < 0.5.

pQTL

Protein quantitative trait loci, pQTL, analysis had been previously performed in FinnGen[47,52] with multiplex antibody-based immunoassay and multiplex aptamer-based immunoassay, Olink and SomaScan[53], respectively, to detect associations between genetic variants and measured protein levels. We extracted the proteomics summary statistics data for the 41 fine-mapped and the five MHC lead variants in the present study from both datasets, Olink and SomaScan, and filtered the results for statistical significance at FDR < 0.05 level.

To evaluate the enrichment of the associated proteins in functional processes, we used STRING interaction network and functional enrichment analysis tool[30].

Data Availability

FinnGen summary statistics are available at www.finngen.fi. Data supporting the current study are available from the authors upon reasonable request and with permission of FinnGen.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The GWAS results for the blood donor phenotype are available at GWAS catalogue under accession number XXXXX.

Code Availability

Code is available in GitHub XXXXXX

References

1. Mitchell, R. Blood banks, biobanks, and the ethics of donation. *Transfusion* **50**, 1866–1869 (2010).
2. Raivola, V., Snell, K., Helén, I. & Partanen, J. Attitudes of blood donors to their sample and data donation for biobanking. *Eur. J. Hum. Genet.* **27**, 1659–1667 (2019).
3. Raivola, V., Snell, K., Helen, I. & Partanen, J. Attitudes of blood donors to their sample and data donation for biobanking. *Eur. J. Hum. Genet.* **27**, 1659–1667 (2019).
4. Finnish Red Cross Blood Service - donation eligibility. https://www.veripalvelu.fi/en/faq/?search=&category=eligibility.
5. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).
6. De Silvestri, A. *et al.* The Involvement of HLA Class II Alleles in Multiple Sclerosis: A Systematic Review with Meta-analysis. *Dis. Markers* **2019**, 1409069 (2019).
7. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
8. Ritari, J., Koskela, S., Hyvärinen, K., FinnGen & Partanen, J. HLA-disease association and pleiotropy landscape in over 235,000 Finns. *Hum. Immunol.* **83**, 391–398 (2022).
9. Atsma, F., Veldhuizen, I., Kort, W. De & Vegt, F. De. Healthy donor effect: its magnitude in health research among blood donors. *Transfusion* **51**, 1820–1828 (2011).
10. Marsh, W. L. & Redman, C. M. The Kell blood group system: a review. *Transfusion* **30**, 158–167 (1990).
11. Avent, N. D. & Reid, M. E. The Rh blood group system: a review. *Blood* **95**, 375–387 (2000).
12. Levine, P. & Stetson, R. E. Landmark article July 8, 1939. An unusual case of intra-group agglutination. By Philip Levine and Rufus E Stetson. *JAMA* **251**, 1316–1317 (1984).
13. GIBLETT, E. R. A critique of the theoretical hazard of inter vs. intra-racial transfusion. *Transfusion* **1**, 233–238 (1961).
14. Stowell, S. R. & Stowell, C. P. Biologic roles of the ABH and Lewis histo-blood group antigens part II: thrombosis, cardiovascular disease and metabolism. *Vox Sang.* **114**, 535–552 (2019).
15. Dahlén, T., Clements, M., Zhao, J., Olsson, M. L. & Edgren, G. An agnostic study of associations

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

between ABO and RhD blood group and phenome-wide disease risk. *Elife* **10**, (2021).

16. Ward, S. E., O'Sullivan, J. M. & O'Donnell, J. S. The relationship between ABO blood group, von Willebrand factor, and primary hemostasis. *Blood* **136**, 2864–2874 (2020).

17. Clancy, J. *et al.* Blood donor biobank as a resource in personalised biomedical genetic research. *Eur. J. Hum. Genet.* (2024) doi:10.1038/s41431-023-01528-0.

18. FinnGen. FinnGen Result Browser. https://r10.finngen.fi/ (2024).

19. National Institutes of Health LDlink. https://ldlink.nih.gov/?tab=home.

20. ISBT Blood Group Allele Tables. https://www.isbtweb.org/isbt-working-parties/rcibgt/blood-group-allele-tables.html.

21. Li, S. & Schooling, C. M. A phenome-wide association study of ABO blood groups. *BMC Med.* **18**, 334 (2020).

22. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

23. Toivonen, J., Allara, E., Castrén, J., di Angelantonio, E. & Arvas, M. The value of genetic data from 665,460 individuals in managing iron deficiency anaemia and suitability to donate blood. *Vox Sang.* **119**, 34–42 (2024).

24. Kangastupa, P. *et al.* The prevalence of hemoglobin Tacoma in Finland detected by HbA1c capillary electrophoresis. *Scand. J. Clin. Lab. Invest.* **83**, 51–57 (2023).

25. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

26. Bowness, P. HLA-B27. *Annu. Rev. Immunol.* **33**, 29–48 (2015).

27. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

28. Borgoño, C. A., Michael, I. P. & Diamandis, E. P. Human tissue kallikreins: physiologic roles and applications in cancer. *Mol. Cancer Res.* **2**, 257–280 (2004).

29. Lazaratos, A.-M., Annis, M. G. & Siegel, P. M. GPNMB: a potent inducer of immunosuppression in cancer. *Oncogene* **41**, 4573–4590 (2022).

30. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

31. Simon, M. *et al.* A study of 609 HLA haplotypes marking for the hemochromatosis gene: (1) mapping of the gene near the HLA-A locus and characters required to define a heterozygous population and (2) hypothesis concerning the underlying cause of hemochromatosis-HLA associa. *Am. J. Hum. Genet.* **41**, 89–105 (1987).

32. Simon, M., Bourel, M., Fauchet, R. & Genetet, B. Association of HLA-A3 and HLA-B14 antigens with idiopathic haemochromatosis. *Gut* **17**, 332 LP – 334 (1976).

33. Linjama, T. *et al.* Extended HLA Haplotypes and Their Impact on DPB1 Matching of Unrelated Hematologic Stem Cell Transplant Donors. *Biol. Blood Marrow Transplant.* **25**, 1956–1964 (2019).

34. Brown, J. H. *et al.* Three-dimensional structure of the human class II histocompatibility

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

antigen HLA-DR1. *Nature* **364**, 33–39 (1993).

35. Groot, H. E. *et al.* Genetically Determined ABO Blood Group and its Associations With Health and Disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 830–838 (2020).

36. Stowell, C. P. & Stowell, S. R. Biologic roles of the ABH and Lewis histo-blood group antigens Part I: infection and immunity. *Vox Sang.* **114**, 426–442 (2019).

37. Franchini, M., Favaloro, E. J., Targher, G. & Lippi, G. ABO blood group, hypercoagulability, and cardiovascular and cancer risk. *Crit. Rev. Clin. Lab. Sci.* **49**, 137–149 (2012).

38. Che Yaacob, N. S., Islam, M. A., Alsaleh, H., Ibrahim, I. K. & Hassan, R. Alpha-hemoglobin-stabilizing protein (AHSP): a modulatory factor in β-thalassemia. *Int. J. Hematol.* **111**, 352–359 (2020).

39. dos Santos, C. O. & Costa, F. F. AHSP and beta-thalassemia: a possible genetic modifier. *Hematology* **10**, 157–161 (2005).

40. Giardine, B. *et al.* Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. *Nucleic Acids Res.* **42**, D1063-9 (2014).

41. Rutherford-Parker, N. J., Colby, J. M., Gehrie, E. A. & Booth, G. S. Unrecognized Hemoglobin Variants in the Donor Blood Supply Are Detectable in the Transfused Population. *Am. J. Clin. Pathol.* **154**, 494–498 (2020).

42. FinnGen. FinnGen Documentation of data release. https://finngen.gitbook.io/documentation/ (2024).

43. R Core Team. https://www.r-project.org/.

44. R studio. https://www.rstudio.com/.

45. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).

46. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

47. FinnGen gitbook. https://github.com/FINNGEN/.

48. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

49. Zheng, X. *et al.* HIBAG - HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).

50. Ritari, J. *et al.* Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genomics Bioinforma.* **2**, (2020).

51. Hyvärinen, K. *et al.* A machine-learning method for biobank-scale genetic prediction of blood group antigens. *PLoS Comput. Biol.* **20**, e1011977 (2024).

52. FinnGen. FinnGen Documentation. https://finngen.gitbook.io/finngen-handbook (2024).

53. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Acknowledgements:

We want to acknowledge the participants and investigators of FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis Pharma AG, and Boehringer Ingelheim International GmbH. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank (https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB.

In addition, we want to acknowledge Dr Katri Haimila for her valuable advice with red cell antigen immunogenetics, Dr Kati Hyvärinen for her expertise and kind help with red cell antigen imputation and the personnel of Blood Service Biobank for their kind help and support for the study.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.18.24309089; this version posted June 19, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Author Contributions

Original study concept by JC and JR. Genetic and statistical analysis of the study: JC, JT, MA and JR. Visualization: JC, JT and JR. Medical expertise: JL. FinnGen: data curation and resources. Procedures in the Blood Service Biobank: SK. Original manuscript: JC. JC, JT, JL, SK, JP, FinnGen, MA, JR read, commented, and approved the final manuscript.

Competing Interests

The authors declare no competing interests.