

1 **Genome-wide analyses of variance in blood cell phenotypes provide new**
2 **insights into complex trait biology and prediction**
3

4 Ruidong Xiang^{1-5,*}, Yang Liu^{1,2,6-9}, Chief Ben-Eghan^{2,6-9}, Scott Ritchie^{1,2,6-9}, Samuel A.
5 Lambert^{1,2,6-9}, Yu Xu^{2,6-9}, Fumihiko Takeuchi^{1,10} and Michael Inouye^{1,2,6-9,*}
6

7 1. Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne,
8 Victoria, Australia

9 2. Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care,
10 University of Cambridge, Cambridge, UK

11 3. Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia

12 4. Baker Department of Cardiovascular Research, Translation and Implementation, La Trobe University,
13 Melbourne, VIC, 3086, Australia

14 5. Baker Department of Cardiometabolic Health, The University of Melbourne, VIC, 3010, Australia

15 6. British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary
16 Care, University of Cambridge, Cambridge, UK

17 7. Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, UK

18 8. Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge,
19 Cambridge, UK

20 9. British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

21 10. Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global
22 Health and Medicine, Tokyo, Japan

23 * Correspondence: R.X. (ruidong.xiang@agriculture.vic.gov.au) and M.I. (mi336@cam.ac.uk)
24
25

26 **Abstract**

27 Blood cell phenotypes are routinely tested in healthcare to inform clinical decisions.
28 Genetic variants influencing mean blood cell phenotypes have been used to understand
29 disease aetiology and improve prediction; however, additional information may be captured
30 by genetic effects on observed variance. Here, we mapped variance quantitative trait loci
31 (vQTL), i.e. genetic loci associated with trait variance, for 29 blood cell phenotypes from the
32 UK Biobank (N~408,111). We discovered 176 independent blood cell vQTLs, of which 147
33 were not found by additive QTL mapping. vQTLs displayed on average 1.8-fold stronger
34 negative selection than additive QTL, highlighting that selection acts to reduce extreme blood
35 cell phenotypes. Variance polygenic scores (vPGSs) were constructed to stratify individuals
36 in the INTERVAL cohort (N~40,466), where genetically less variable individuals (low
37 vPGS) had increased conventional PGS accuracy (by ~19%) than genetically more variable
38 individuals. Genetic prediction of blood cell traits improved by ~10% on average combining
39 PGS with vPGS. Using Mendelian randomisation and vPGS association analyses, we found
40 that alcohol consumption significantly increased blood cell trait variances highlighting the
41 utility of blood cell vQTLs and vPGSs to provide novel insight into phenotype aetiology as
42 well as improve prediction.

43

44

45 **Introduction**

46 The complete blood count is amongst the most routinely ordered clinical laboratory
47 tests performed globally¹. Blood cells play crucial roles in a variety of biological processes,
48 such as oxygen transport, iron homeostasis, and pathogen clearance²⁻⁴, and serve as key
49 biological conduits for interactions between an individual and their environment. The genetic
50 architecture of blood cell traits has been recently elucidated by genome-wide association
51 studies (GWAS)^{5,6} and, consistent with their well-known role in disease and clinical testing,
52 blood cell traits are both highly heritable and have been genetically linked to many diseases,
53 including cardiovascular diseases⁷, mental disorders⁸ and autoimmune diseases⁹.

54 Despite the success of GWAS, our understanding of the genetic architecture of complex
55 traits has been limited by a focus on mean trait values and how these change with respect to
56 genotype. The genetics of trait variance, how individual measurements deviate from the mean
57 trait value across genotypes, is far less studied. It has long been known that trait variance, e.g.
58 for gene expression^{10,11} and metabolic rate¹², plays a role in an organism's fitness and
59 phenotypic penetrance. Theories support the existence of selection on trait variance to improve
60 fitness^{13,14}. However, there are limited observations of selection on clinically significant traits.
61 Variance quantitative trait loci (vQTLs) have been identified for human body composition
62 traits, such as BMI^{15,16}, and for cardiometabolic biomarkers¹⁷. vQTLs have also been linked to
63 gene-by-environment interactions (GxE) or gene-by-gene interactions (GxG)¹⁵⁻¹⁸. vQTL
64 studies of blood cell traits are currently lacking, despite their central role in biological processes
65 and ubiquity in clinical testing.

66 Polygenic scores (PGS) are being intensively studied in various ways to determine their
67 utility in clinical practice¹⁹⁻²¹. PGS for blood cell traits, in particular, are both highly predictive
68 and show sex- and age-specific interactions^{6,7}. How to treat trait variance and vQTLs with
69 respect to phenotype prediction is relatively unexplored. A variance PGS (vPGS) to predict the

70 trait variance may be estimated from the effect sizes obtained from a genome-wide vQTL
71 analysis. In theory, a PGS is different from a vPGS, where the former may be used to stratify
72 individuals based on the inherited trait level while the latter stratifies individuals based on the
73 inherited deviation of individuals from the population mean. It is known that the accuracy of a
74 PGS varies across individuals as a function of the genetic distance from the reference
75 population²². As a vPGS may represent the outcome of GxE¹⁶ or GxG due to the nature of
76 vQTLs¹⁵, examining a PGS alongside vPGS may reveal individual variability in PGS accuracy
77 that can be accommodated.

78 Here, we conduct genome-wide vQTL analysis for 29 blood cell traits in the UK
79 Biobank^{6,7} and the INTERVAL cohort²³. We compared the discovered vQTL with
80 conventional QTL and analysed vPGS with conventional PGS in the prediction of blood cell
81 traits. We found novel vQTL, not identified by previous conventional GWAS and displayed
82 strong selection to reduce blood cell trait variances. Finally, we demonstrate the use of vPGS
83 in stratifying individuals, resulting in differing PGS performance, and then show that PGS
84 performance within vPGS strata is associated with lifestyle factors.

85

86 **Results**

87

88 *Genome-wide discovery and annotation of vQTLs in the UK Biobank*

89 We performed GWAS of variance in 29 blood cell traits from the UKB^{17,18} (Average
90 sample size = 402,142, **Supplementary Table 1**). The processing of phenotypes and genotypes
91 followed previously established protocols with stringent quality control and normalisation
92 procedures⁵⁻⁷. Levene's test²⁴, as implemented in OSCA²⁵, was used to map vQTLs for each of
93 the 29 blood cell traits and the inflation factors and lambda GC were assessed using LD Score
94 regression (LDSC)²⁶. Across the 29 traits, the average lambda GC and LDSC intercepts were

95 1.03 and 1.007, respectively (**Supplementary Table 2**), indicating negligible inflation. At a
96 study-wide significance level of $p < 4.6 \times 10^{-9}$ and with clumping $r^2 < 0.01$, we identified 176
97 independent vQTLs (**Figure 1a, Supplementary Table 3, Methods**).

98 Basophil cell count (baso) and basophil percentage of white cells (baso_p) yielded the
99 largest number of independent vQTLs ($N = 27$ and 23 , respectively), whereas high light scatter
100 reticulocyte count (hlc) did not have any study-wide significant vQTLs (**Supplementary Table**
101 **4**). Most vQTL were associated with the variance of only one or two traits and many of these
102 traits are correlated (**Supplementary Figure 1 and Supplementary Table 3**). By counting the
103 number of blood cell traits associated, the most pleiotropic lead vQTL was located in gene
104 *HBM* (hemoglobin subunit mu) and was associated with the variance of four traits (red blood
105 cell count, mean corpuscular volume, mean corpuscular hemoglobin and mean corpuscular
106 hemoglobin concentration, **Supplementary Table 3**). The second pleiotropic lead vQTL
107 related to long intergenic non-coding RNA *LINC02768* was associated with 3 traits [monocyte
108 percentage of white cells (mono_p), baso and baso_p, **Figure 1b**]. To account for the
109 phenotypic correlations, the pleiotropy of trait variance was further assessed using HOPS²⁷,
110 which found that 495 SNPs (out of 71,216 input SNPs) showed significant pleiotropy
111 (**Supplementary Table 5**). In this analysis, the most significant pleiotropic locus was
112 *LINC02768* (**Supplementary Table 5**).

113 vQTLs were largely distinct from additive QTLs. Of 176 lead vQTLs, 147 were not
114 detected as additive QTLs by Vuckovic et al⁶, the largest GWAS to date of blood cell traits.
115 vQTLs had an average r^2 of 0.33 (SD=0.12) with the lead additive QTLs from Vuckovic et al⁶
116 (**Supplementary Figure 2**). We repeated the OSCA²⁵ analysis fitting the trait level as a
117 covariate, i.e. effects of vQTL conditioned on the trait level. The correlation of the effects of
118 these vQTLs between the original and conditional analysis was 0.99 (**Supplementary Figure**
119 **3**), consistent with vQTL effects being independent of those for mean trait level.

120 Across 29 traits, the magnitude of the genetic correlation between trait variance and
121 trait level, as estimated by LDSC²⁶, was on average 0.328 (SD=0.24) (**Figure 1c**) and the
122 genetic correlation between trait variance and value was not significant for 21 out of 29 traits
123 after adjusting for multi-testing. Notably, red cell distribution width (rdw) and neutrophil
124 percentage of white cells (neut_p) had significant negative genetic correlations after adjustment
125 for multiple testing, indicating genetic control of trait variance so it is reduced at high levels of
126 rdw or neut_p. Rdw is itself a measure of variation; however, high rdw is an indicator of iron
127 or other nutrient deficiencies, thus our results suggest a potential simultaneous genetic
128 stabilisation when rdw is genetically high. Similarly, high neut_p is an indicator of microbial
129 or inflammatory stress, thus a negative genetic correlation suggests a stabilisation at genetically
130 high neut_p levels.

131 With many known trait-associated alleles under negative selection²⁸, we also assessed
132 the extent to which QTLs for trait variability are under selection. We used Bayes(S)²⁸, to
133 compare the selection coefficient (S) between vQTLs and additive QTLs across 29 blood cell
134 traits (**Figure 1d**). We found that, on average S is 1.8 times stronger on trait variance (-0.82,
135 SD=0.07) than trait level (-0.45, SD=0.05) (**Figure 1d**). These results show a much stronger
136 negative selection on blood cell trait variance than on trait level. While it can be difficult to
137 differentiate between negative and stabilising selection, our results are consistent with
138 evolution acting on blood cell traits to remove extreme phenotypes from the population.

139 We used FUMA²⁵ to annotate the lead vQTLs for each trait (**Supplementary Data 1**
140 **and Supplementary Table 6**) and perform a trait enrichment analysis with GWAS Catalog²³.
141 We found multiple significant overlaps between vQTL and additive QTL related to alcohol
142 consumption. Significant vQTLs (rs191673261 in LD with lead vQTL rs572454376) for
143 platelet crit (pct) were located proximal to *ALDH2*, a well-known gene contributing to alcohol
144 consumption²⁹ (**Figure 2a**). We subsequently performed Summary-data-based Mendelian

145 Randomisation (GSMR)³⁰ between GWAS of alcohol consumption (as exposure, obtained
146 from Cole et al 2020³¹) and variances of blood cell traits (as outcome). We also used MR-
147 PRESSOR³² and MR-weighted median³³ to validate our results with different assumptions. We
148 did not find statistically significant causal links between alcohol consumption and pct.
149 However, at multi-testing adjusted $p < 0.05$ level, increased alcohol consumption was
150 genetically predicted to increase variance in mean corpuscular volume (mcv) and mean sphered
151 corpuscular volume (mscv) (**Figure 2b-d**). At nominal significance ($p < 0.05$ for each of the
152 three MR methods), increased alcohol consumption was genetically predicted to increase
153 variance in red blood cell count (rbc) and neutrophil percentage of white cells (neut_p) (**Figure**
154 **2b**). The positive effects of alcohol consumption on neutrophil count (neut) were significant in
155 GSMR (nominal $p = 0.014$) and MR-PRESSO (nominal $p=0.008$), but insignificant (nominal
156 $p=0.1$) in MR-weighted median. Overall, our results support alcohol consumption as affecting
157 particular blood cell trait variances.

158 FUMA-enabled ANNOVAR²⁴ was used to study the enrichment of vQTLs in
159 different functional annotation classes. We found that vQTLs for mean sphered corpuscular
160 volume (mscv), reticulocyte count (ret) and reticulocyte fraction of red cells (ret_p) were
161 significantly enriched in exonic variants related to protein-coding functions (**Supplementary**
162 **Figure 4a**). However, vQTLs for many other traits were enriched in regulatory regions. For
163 example, vQTLs for mean corpuscular hemoglobin concentration, red blood cell count and
164 hemoglobin concentration (hgb) were enriched for upstream gene regulatory sites. vQTLs for
165 eosinophil count (eo), mean corpuscular hemoglobin and mean corpuscular volume were
166 enriched for downstream regulatory sites of genes. vQTLs for platelet distribution width
167 (pdw) and basophil percentage of white cells (baso_p) were enriched for UTR-3' sites
168 (**Supplementary Figure 4a**). We used pathway enrichment analyses within FUMA to further
169 investigate whether vQTLs were enriched for gene regulation, finding that vQTLs for mean

170 corpuscular hemoglobin were enriched for many epigenetic regulatory mechanisms including
171 DNA methylation and histone modifications (**Supplementary Figure 4b**).

172

173 *Polygenic scores of blood cell trait variance*

174 Polygenic scores are conventionally constructed for differences in trait level. Using
175 the vQTL results from the UK Biobank, we constructed polygenic scores for blood cell trait
176 variance (vPGS) using PRSICE³⁴ and the INTERVAL study as an external validation cohort
177 (**Supplementary Table 1, Methods**). For conventional PGS we utilised those from Xu et al⁷.
178 Across traits, there was nearly zero Pearson correlation between vPGS and PGS (mean
179 0.00028, range [-0.018, 0.023]; **Supplementary Figure 5**), consistent with PGS for trait
180 variance being independent from those for mean trait levels.

181 A potential use of vPGS is to stratify a population by trait variance, thus identifying
182 subgroups where predictive models may have increased performance. For each trait, we
183 stratified individuals into the top and bottom 5% of vPGS. As vPGS were trained to estimate
184 SNP effects on trait variance, individuals with lower or higher vPGS were expected to
185 display less or more variation around the trait mean, respectively. We then compared the
186 correlation of PGSs for each trait between these more (high-vPGS) or less variable (low-
187 vPGS) groups. Across the 27 blood cell traits, we found the less variable group (bottom 5%
188 of vPGS) had a significantly higher PGS-trait correlation than the more variable group (top
189 5% vPGS) (**Figure 3**). Across all traits, the mean relative difference in PGS-trait correlation
190 (Pearson) between the less and more variable groups was +6.5% [-7%, 18%] (**Figure 3**), with
191 a mean difference of +6.6% [-9%, 19%] for spearman correlation (**Supplementary Figure**
192 **6**).

193 Next, we analysed the effects of interaction between PGS and vPGS for each trait. We
194 found that 6 out of 27 blood cell traits displayed statistically significant ($p < 0.05$) effects of

195 interaction between PGS and vPGS (**Figure 4a**), suggesting that associations between PGS
196 and blood cell trait level can depend on vPGS (**Figure 4b-c**).

197 Next, for all INTERVAL individuals, we examined whether adding vPGS to PGS
198 increased the prediction of blood cell trait level. For each blood cell trait, we estimated the
199 difference in the variance explained (R^2) between PGS models with or without vPGS (**Figure**
200 **5, Methods**). Across all 27 traits, the mean R^2 increase was +1.8% (range [0%, 5%]) and 9
201 traits showed a statistically significant³⁵ increase in R^2 (**Figure 5, Methods**). We further
202 tested whether multi-trait vPGSs also increase prediction power³⁶, and found that adding
203 multi-trait vPGSs to PGS increased R^2 by a mean of +3.5% (range [0%, 10%]) and the
204 increase was statistically significant in 16 traits (**Figure 5**).

205

206 *Lifestyle effects on blood cell trait variance*

207 To investigate why some individuals have highly variable blood cell trait levels we
208 assessed two major lifestyle factors, namely alcohol consumption and smoking behaviour.
209 We first identified distinct groups of individuals with high or low trait variance in
210 INTERVAL. For the high variability trait group, we identified individuals who were in the
211 top 5% of vPGS for at least 4 blood cell traits and, for the low variability trait group, with
212 individuals in the bottom 5% of vPGS for at least 4 traits (**Methods, Figure 6**). Our analysis
213 found that those in the high variability trait group were more likely to be current or previous
214 consumers of alcohol (**Figure 6a**). Further, we applied this analysis to mcv, neut_p and rbc,
215 finding significant causal effects of alcohol consumption in GSMR analyses (**Figure 2a,**
216 mscv not available in INTERVAL). Consistent with the results from GSMR, individuals with
217 high variability in mcv, neut_p and rbc were more likely to be alcohol consumers (**Figure**
218 **6b**). These results support the hypothesis that alcohol consumption increases variation in
219 blood cell traits.

220

221 **Discussion**

222 The analysis of vQTL and vPGS may yield new insights into locus and GxE
223 discovery as well as the use of human genetics for patient stratification. Our study explored
224 vQTL analysis in 29 blood cell traits in the UK Biobank, where the majority (84%) of vQTLs
225 did not overlap with and were largely independent of genetic variants identified in
226 conventional GWAS of trait mean. We investigated the functional annotation, pathway-level
227 associations and selection of vQTLs. The potential utility of using vQTLs to construct vPGS
228 and using the latter to stratify the population into groups of trait variance was demonstrated.
229 Finally, our analysis also showed trait variance to be related to non-genetic factors, finding
230 that alcohol consumption had a putatively causal effect on increasing blood cell trait
231 variances.

232 Both blood cell trait variance and level display significant negative selection.
233 Stabilising selection of human traits has been reported¹⁴. However, to our knowledge,
234 negative selection on blood cell trait variance, particularly its strength relative to that on trait
235 level, has not yet been identified. Strong negative selection of blood cell trait variances
236 suggests that extreme blood cell morphologies, which may be indicative of diseases, have not
237 been favoured.

238 Many vQTLs tagged loci implicated in GxG, GxE and under epigenetic regulation,
239 consistent with previous studies of vQTLs^{18,37}. We found blood cell vQTLs tagged genes
240 related to diet. Previous GWAS of diet identified loci related to blood lipids³⁸ and glycated
241 hemoglobin³⁹ but not to blood cell traits analysed here; however, others have reported that
242 alcohol intake increases mean corpuscular volume independent of the genetic contribution to
243 the level of mean corpuscular volume⁴⁰. In our study, alcohol consumption-related loci

244 significantly overlapped with vQTLs for platelet count, the function of which can be
245 significantly affected by alcohol drinking⁴¹.

246 Stratification by vPGS was shown to identify groups with significantly different PGS
247 prediction accuracy, indicating that some groups are intrinsically harder to predict by PGSs
248 than others. Such information may be important for the implementation of PGSs in
249 healthcare. Interestingly, our analysis also found multiple significant interactions between
250 PGS and vPGS, suggesting that the non-additive and GxE components related to PGSs could
251 impact prediction accuracy. These findings are consistent with previous observations^{42,43}.

252 Our results also showed that alcohol consumption and, to some extent increased BMI,
253 were significant contributors to increased genetic variability in blood cell traits. Previously
254 reports have found that blood cell traits can be significantly influenced by alcohol intake⁴⁴
255 and BMI⁴⁵. However, to our knowledge, this is the first study to report lifestyle risk factors
256 contributing to genetically predicted variation in blood cell traits.

257 In conclusion, our study provides an in-depth analysis of human genetic effects on the
258 variance of blood cell traits, including the discovery of loci and strong negative selection,
259 improved genomic prediction and stratification, and identification of GxE. vPGSs may
260 provide a generalisable approach to incorporate individual differences to improve trait and
261 disease risk prediction. This study demonstrates that there is substantive human biology and
262 potential clinical utility in studying trait variances alongside conventional studies of trait
263 means.

264

265 **Methods and Materials**

266

267 **Study Cohorts and Methods**

268 **UK Biobank.** The UK Biobank^{46,47} (<https://www.ukbiobank.ac.uk/>) is a cohort
269 including 500,000 individuals living in the UK who were recruited between 2006 and 2010,
270 aged between 40 and 69 years at recruitment. Ethics approval was obtained from the North
271 West Multi-Center Research Ethics Committee. The current analysis was approved under UK
272 Biobank Project 30418. The participants with the measurements of the 29 blood cell traits
273 and who were identified as European ancestry based on their genetic component analysis
274 were included in our study. The detailed sample sizes used for vQTL detection were shown in
275 **Supplementary Table 1.**

276 **INTERVAL Study.** INTERVAL²³ (<https://www.intervalstudy.org.uk/>) is a
277 randomised trial of 50,000 healthy blood donors, aged 18 years or older at recruitment. The
278 participants with measurements of the 27 considered blood cell traits were included in our
279 study. The detailed sample sizes were shown in **Supplementary Table 1.** All participants
280 have given informed consent and this study was approved by the National Research Ethics
281 Service (11/EE/0538).

282 **Data quality control.** For trait levels of 29 blood cell traits in the UK Biobank and
283 matching 27 traits in the INTERVAL, we adopted previously established protocols for
284 quality controls⁵⁻⁷ to adjust technical and other confounders and the first 10 genetic principal
285 components. For trait levels, adjusted technical variables include the time between
286 venepuncture and full blood cell analysis, seasonal effects, center of sample collection, the
287 time-dependent drift of equipment, and systematic differences in equipment; other adjusted
288 variables included sex, age, diet, smoking and alcohol consumption. Quality control and
289 imputation of the genotype data have been described previously^{5,47}, which filtered the
290 samples to the European ancestry only.

291 **vQTL analysis.** Genome-wide analysis of vQTL used Levene's test. As detailed in
292 ^{11,15}, the test statistic of Levene's test is:

293
$$\frac{(n - k) \sum_{i=1}^k n_i (z_{i.} - z_{..})^2}{(k - 1) \sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - z_{i.})^2}$$

294 where n is the total sample size, k is the number of groups ($k = 3$ in vQTL analysis), n_i is the
 295 sample size of the i th group (one of three genotypes), z_{ij} is the absolute difference between
 296 the phenotype value in sample j from genotype i and the median value in genotype i , $z_{i.}$ is the
 297 average z value in genotype i , and $z_{..}$ is the average z value across all samples. OSCA-
 298 implemented Levene's test also provides beta and se estimates based on p-value and minor
 299 allele frequency¹⁵ and the beta estimates were used to construct vPGSs described later.

300 We estimated the study-wise significance for vQTL as $4.6 \times 10^{-9} = 5 \times 10^{-8} / 10.2$
 301 where 10.2 is the effective number of traits analysed in the study. The effective number of
 302 traits is estimated using $\frac{(\sum_{k=1}^p \lambda_k)^2}{\sum_{k=1}^p \lambda_k^2}$, where $\lambda_1 \dots \lambda_p$ is principal component variances or the
 303 ordered eigenvalues^{15,17}. To identify lead vQTL with relative independence, we used plink-
 304 clumping⁴⁸ using a p-value threshold of 4.6×10^{-9} , $r^2 < 0.01$ and window size of 5000kb (the
 305 same parameter used by¹⁵). The LD analysis between vQTL and lead QTL reported by
 306 Vuckovic et al⁶ used plink 1.9 with the function of --ld. The novel vQTL was defined as
 307 those lead vQTL after clumping with GWAS p-value $> 4.6 \times 10^{-9}$ in Vuckovic et al and with
 308 LD- $r^2 < 0.8$ with lead QTL reported by Vuckovic et al. For vQTL mapping results of each
 309 trait, we used LDSC²⁶ to estimate lambda-GC and intercept to check inflation. We also used
 310 FUMA⁴⁹ to annotate significant vQTL for each trait with default settings. Results from
 311 FUMA functions of SNP2GENE and GENE2FUNC were presented in the results.

312 To explore the potential causal relationships between alcohol consumption and blood
 313 cell trait variances, we used GSMR³⁰ to discover the causal relationships and used MR-
 314 PRESSO³² and weighted-mean implemented in MendelianRandomisation³³ as validation. The
 315 GWAS summary data for alcohol consumption was obtained from Cole et al³¹. Default
 316 settings for nominated software were used and SNPs with p-value $< 5e-8$ and $r^2 < 0.05$ were

317 used in the analysis. Significant results were defined as the multi-testing adjusted p-value
318 from GSMR < 0.05 and the nominal significance was defined as the Mendelian
319 Randomisation had raw p-value < 0.05 in all 3 methods.

320 **Analysis of vPGS and PGS.** PGS trained using the elastic net from Xu 2022 et al⁷
321 was used. For training vPGS, we followed the protocol described by Miao et al¹⁶ reported
322 successful implementation of vPGS for BMI using PRSICE³⁴, we used the same procedure
323 described by Miao et al to construct vPGS in the INTERVAL using PRSICE, i.e., `-clump-p1`
324 `1 -clump-p2 1 -clump-r2 0.1` and `-clump-kb 1000`. When vPGS was computed for each trait,
325 they were used to rank INTERVAL individuals where the top and bottom 5% of individuals
326 were stratified. As vPGS was trained based on SNP effects on phenotypic variance, i.e., the
327 extent to which the individual measurement deviates from the mean, vPGS was expected to
328 genetically predict such variation of individuals for the corresponding trait. Therefore,
329 individuals ranked in the top 5% of vPGS were called the genetically more variable group
330 and individuals ranked in the bottom 5% of vPGS were called the genetically less variable
331 group. Then, for each trait, within the more variable and less variable groups, we estimated
332 the PGS accuracy, i.e., the correlation between PGS and the corresponding trait. We then
333 compared the PGS accuracy between the more variable and less variable groups for each trait
334 and the relative increase was calculated as $\frac{r_{less\ variable} - r_{more\ variable}}{r_{more\ variable}}$ where $r_{less\ variable}$ is the
335 PGS accuracy in the less variable group defined by vPGS and $r_{more\ variable}$ is the PGS
336 accuracy in the more variable group defined by vPGS.

337 The effects of interaction between PGS and vPGS on the corresponding trait in
338 INTERVAL were tested on corrected blood cell traits (described above). As the traits were
339 already corrected for covariates, only the main effects and interaction of PGS and vPGS were
340 fitted for each blood cell trait in the `lm()` function in R: $y = PGS + vPGS + PGS * vPGS$,
341 where y was each of the blood cell trait. The effects of interaction on specific traits (e.g.,

342 eo_p and neut) were visualised using the function of plot_model in the R package sjPlot
343 (version 2.8.15).

344 To evaluate if adding vPGS improves PGS model predictability, we tested two sets of
345 vPGS, where one set is the original single-trait vPGSs for 27 traits computed by PRSICE, and
346 the other set is estimated using the multi-trait BLUP (SMTpred³⁶) combining information
347 from single-trait vPGSs. Following the instructions from
348 <https://github.com/uqrmaie1/smtpred>, we used the LDSC²⁶ wrapper (ldsc_wrapper.py) with
349 default options in SMRpred to estimate the genetic parameters for each trait which are
350 required inputs by the multi-trait BLUP. Then, the script smtpred.py was used by default
351 options with the estimated genetic parameters to combine single-trait vPGSs to construct
352 multi-trait vPGSs. Then, we used r2redux³⁵ to quantify the difference in variance explained
353 (R^2) between PGS models with and without vPGS. As described by Momin et al³⁵, r2redux
354 can powerfully detect R^2 differences between models for the out-of-sample genomic
355 prediction which is suitable to our case where the PGS and vPGS models were trained in the
356 UK Biobank and predicted into INTERVAL. We followed the instructions provided by
357 (<https://github.com/mommy003/r2redux>) to compare the R^2 of models with vPGS and
358 without PGS using the nested method and obtained p-values testing the significance of the
359 increase in R^2 when adding vPGS. The relative increase in R^2 was expressed as the absolute
360 difference in R^2 divided by the heritability estimated using LDSC²⁶.

361 To characterise the individuals that were identified as genetically variable across
362 traits, we first counted the number of times (out of 27 blood cell traits) an individual was
363 ranked in the top 5% by PGS for each trait. We also counted the number of times an
364 individual was ranked in the bottom 5% by PGS for each trait. We then identified 2,465
365 individuals who always ranked in the top 5% vPGS, and 2,362 individuals who always
366 ranked in the bottom 5% vPGS across multiple blood traits. Individuals in the top group were

367 ranked in the top 5% vPGS for 4 to 17 traits with a mean of 5 and individuals in the bottom
368 group were ranked in the bottom 5% vPGS for 4 to 23 traits with a mean of 9. Then, the top
369 group was labelled as 1 and the bottom group was labelled as 0 and this 0/1 vector was
370 analysed as a binary outcome for a logistic regression analysis against lifestyle factors: $y =$
371 $age + sex + BMI + smoking_status + drinking_status$, where the average age is 46.1
372 (SD=14.3) and the average BMI is 26.2 (SD=4.6); for sex, there are 2,419 women; for
373 smoking status, there are 2,728 people never smoked, 378 current smokers, 1,634 previous
374 smokers and 87 with no answers; for alcohol drinking status, there are 118 who never drunk,
375 4,178 current drinkers, 323 previous drinkers and 208 with no answers. The logistic
376 regression used the function `glm()` in R and for sex the male was set to the reference level, for
377 smoking the level of never smoked was set to the reference and for drinking the level of
378 never drunk was set to the reference. The same analysis was also applied to individual blood
379 cell traits of mean corpuscular volume (mcv), neutrophil percentage of white cells (neut_p)
380 and red blood cell count (rbc) which were significant in Mendelian Randomisation analyses.

381

382 **Data availability**

383 Full summary statistics of vQTL mapping are available via the GWAS Catalog
384 (<https://www.ebi.ac.uk/gwas/>) under the accession number NNNNNNNNNN (*to be generated*
385 *upon acceptance of peer-reviewed manuscript*). All data described are available through the
386 UK Biobank subject to approval from the UK Biobank access committee. See
387 <https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access> for further details.

388 INTERVAL study data from this paper are available to bona fide researchers from
389 helpdesk@intervalstudy.org.uk and information, including the data access policy, is available
390 at <http://www.donorhealth-btru.nihr.ac.uk/project/bioresource>.

391

392 **Code availability**

393 The manuscript does not produce original code. vQTL mapping used OSCA:
394 <https://yanglab.westlake.edu.cn/software/osca/#Overview>; genetic correlation analysis used
395 LDSC: <https://github.com/bulik/ldsc>; pleiotropy analysis: <https://github.com/rondolab/HOPS>.
396 Mendelian randomisation used GSMR: <https://yanglab.westlake.edu.cn/software/gsmr/>, MR-
397 PRESSO: <https://github.com/rondolab/MR-PRESSO> and MendelianRandomisation:
398 <https://cran.r-project.org/web/packages/MendelianRandomization/index.html>; Analysis of
399 selection used GCTB-BayesS:
400 <https://cnsgenomics.com/software/gctb/#SummaryBayesianAlphabet>; vPGS analysis used
401 PRSICE: <https://choishingwan.github.io/PRSice/> and plink2: <https://www.cog->
402 [genomics.org/plink/2.0/](https://www.cog-genomics.org/plink/2.0/); multi-trait GBLUP used SMTpred:
403 <https://github.com/uqrmaie1/smtpred>; significance tests of R^2 increase used r2redux:
404 <https://github.com/mommy003/r2redux>; logistic regression analysis used glm():
405 <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm>.

406

407 **Competing interests**

408 M.I. is a trustee of the Public Health Genomics (PHG) Foundation, a member of the
409 Scientific Advisory Board of Open Targets, and has research collaborations with
410 AstraZeneca, Nightingale Health and Pfizer which are unrelated to this study.

411

412 **Funding**

413 The authors are grateful to UK Biobank for access to data to undertake this study (Projects
414 #30418).

415

416 This work was supported by core funding from the: British Heart Foundation
417 (RG/18/13/33946) and NIHR Cambridge Biomedical Research Centre (NIHR203312),
418 Cambridge BHF Centre of Research Excellence (RE/18/1/34212) and BHF Chair Award
419 (CH/12/2/29428) and by Health Data Research UK, which is funded by the UK Medical
420 Research Council, Engineering and Physical Sciences Research Council, Economic and
421 Social Research Council, Department of Health and Social Care (England), Chief Scientist
422 Office of the Scottish Government Health and Social Care Directorates, Health and Social
423 Care Research and Development Division (Welsh Government), Public Health Agency
424 (Northern Ireland), British Heart Foundation and Wellcome.

425

426 X.J. was funded by the British Heart Foundation (CH/12/2/29428) and Wellcome Trust
427 (227566/Z/23/Z). S.C.R. was funded by a British Heart Foundation (BHF) Programme Grant
428 (RG/18/13/33946). S.C.R. was also funded by the National Institute for Health and Care
429 Research (NIHR) Cambridge BRC (BRC-1215-20014; NIHR203312) [*]. Y.X. and M.I.
430 were supported by the UK Economic and Social Research Council (ES/T013192/1). M.I. and
431 S.L. were supported by NIH U24 (5U24HG012542-02). M.I. is supported by the Munz Chair
432 of Cardiovascular Prediction and Prevention and the NIHR Cambridge Biomedical Research
433 Centre (NIHR203312) [*]. *The views expressed are those of the authors and not necessarily
434 those of the NIHR or the Department of Health and Social Care.

435

436 **Author contributions**

437 R.X. and M.I. conceived of the study. R.X. performed the analyses with assistance from Y.X.
438 and M.I.. R.X., M.I., F.T., Y.L., C.B.E., X.J., S.R., S.A.L., and Y.X. drafted and revised the
439 manuscript. All authors read and approved the final version of the manuscript.

440

441 **Acknowledgements**

442 We thank Dr. Emmanuela Bonglack and Dr. Xilin Jiang, for their insightful comments. The
443 authors are grateful to Prof. Michael Goddard for discussions on the selection of vQTL.

444

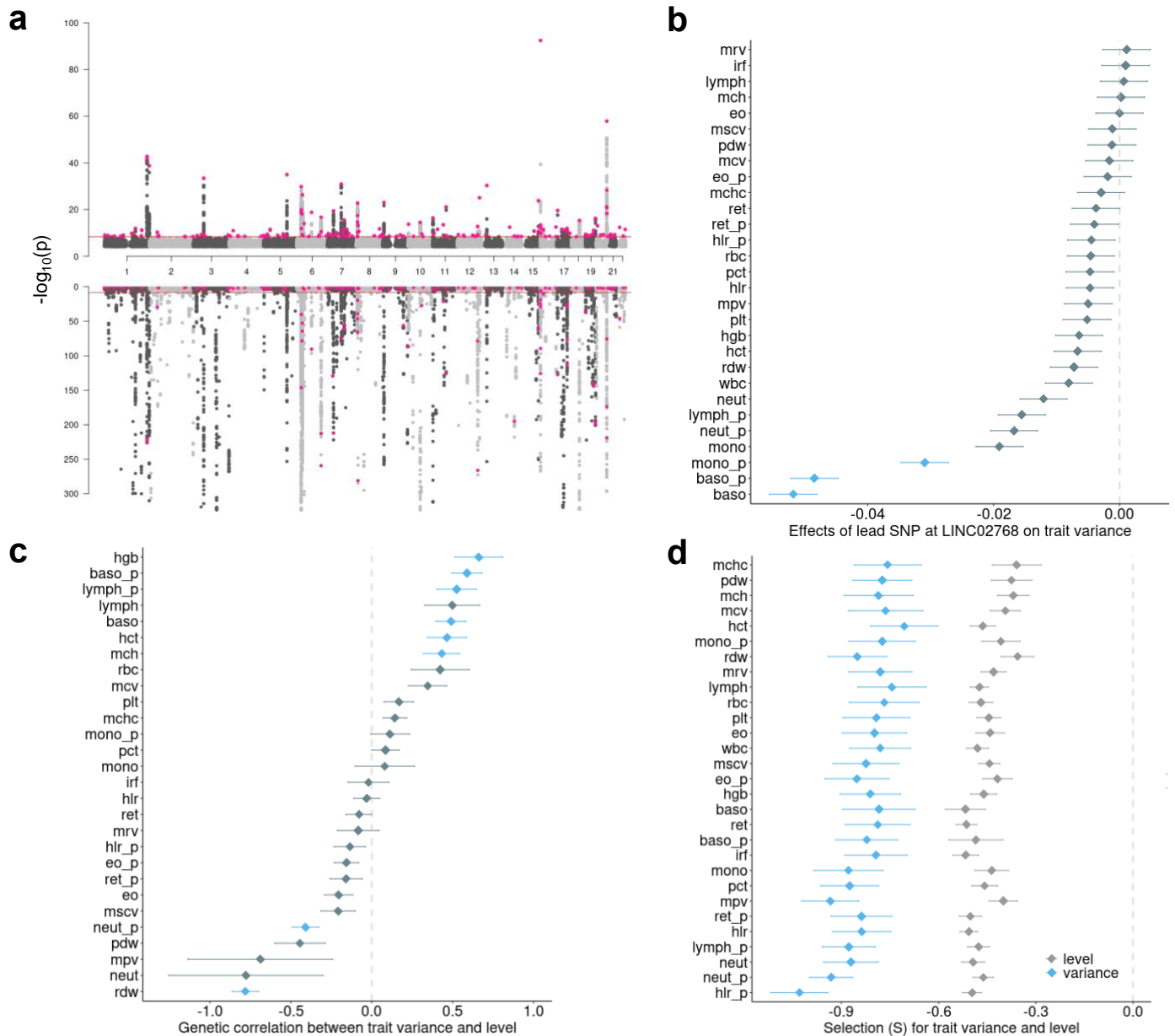
445

446 **References**

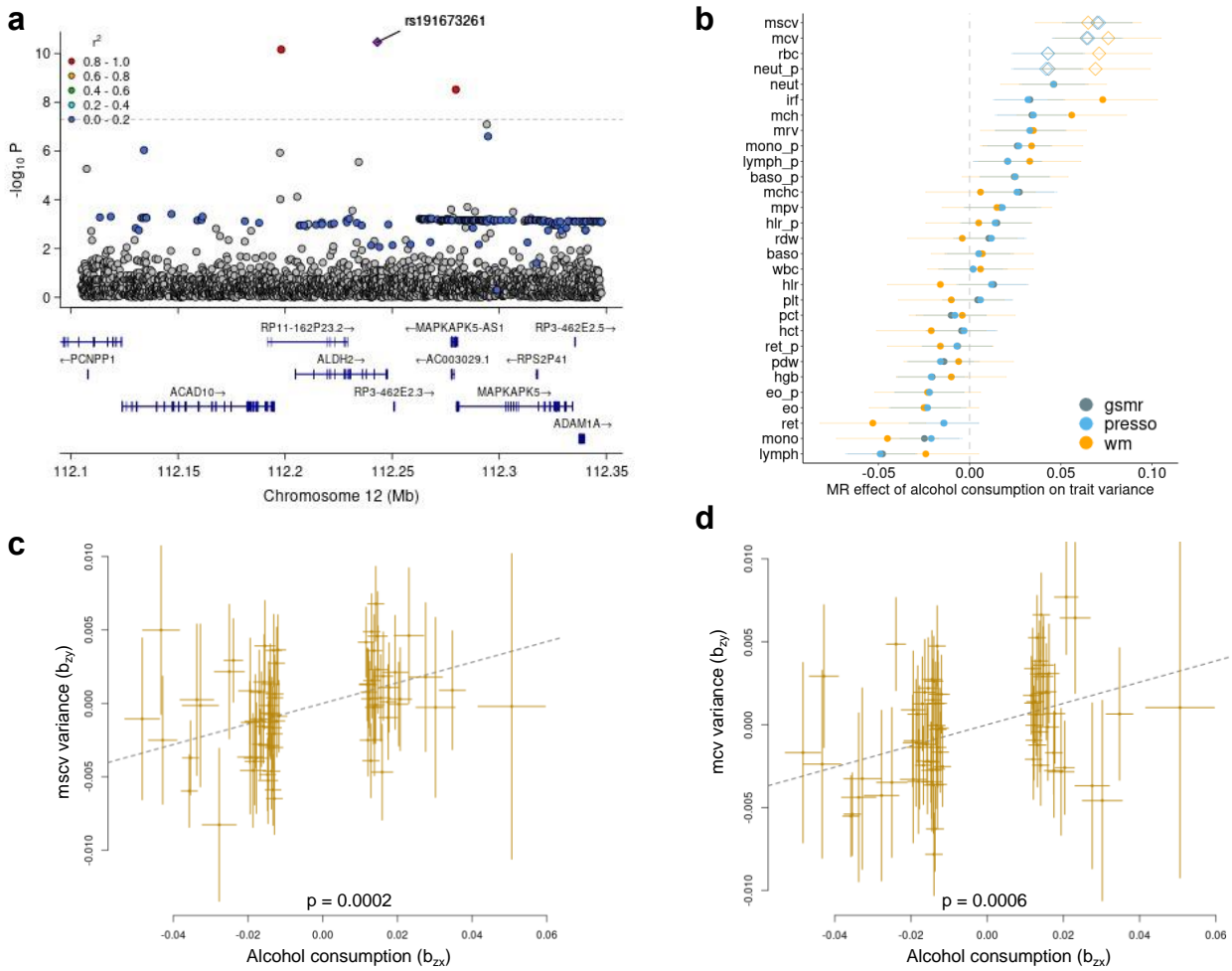
- 447 1 Horton, S. *et al.* The top 25 laboratory tests by volume and revenue in five different
448 countries. *Am J Clin Pathol* **151**, 446-451 (2019).
- 449 2 Jensen, F. B. The dual roles of red blood cells in tissue oxygen delivery: oxygen carriers
450 and regulators of local blood flow. *J Exp Biol* **212**, 3387-3393 (2009).
- 451 3 Jenne, C., Urrutia, R. & Kubes, P. Platelets: bridging hemostasis, inflammation, and
452 immunity. *Int J Lab Hematol* **35**, 254-261 (2013).
- 453 4 Nagata, S. Apoptosis and clearance of apoptotic cells. *Annu Rev Immunol* **36**, 489-517
454 (2018).
- 455 5 Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to
456 common complex disease. *Cell* **167**, 1415-1429. e1419 (2016).
- 457 6 Vuckovic, D. *et al.* The polygenic and monogenic basis of blood traits and diseases. *Cell*
458 **182**, 1214-1231. e1211 (2020).
- 459 7 Xu, Y. *et al.* Machine learning optimized polygenic scores for blood cell traits identify
460 sex-specific trajectories and genetic correlations with disease. *Cell Genomics* **2** (2022).
- 461 8 Yang, Y. *et al.* The shared genetic landscape of blood cell traits and risk of neurological
462 and psychiatric disorders. *Cell Genomics* **3** (2023).
- 463 9 Akbari, P. *et al.* A genome-wide association study of blood cell morphology identifies
464 cellular proteins implicated in disease aetiology. *Nat Commun* **14**, 5023 (2023).
- 465 10 Duveau, F. *et al.* Fitness effects of altering gene expression noise in *Saccharomyces*
466 *cerevisiae*. *eLife* **7**, e37272 (2018). <https://doi.org:10.7554/eLife.37272>
- 467 11 Sarkar, A. K. *et al.* Discovery and characterization of variance QTLs in human induced
468 pluripotent stem cells. *PLoS Genet* **15**, e1008045 (2019).
469 <https://doi.org:10.1371/journal.pgen.1008045>
- 470 12 Pettersen, A. K., Marshall, D. J. & White, C. R. Understanding variation in metabolic
471 rate. *J Exp Biol* **221**, jeb166876 (2018).
- 472 13 Kimura, M. A stochastic model concerning the maintenance of genetic variability in
473 quantitative characters. *PNAS* **54**, 731-736 (1965).
- 474 14 Sanjak, J. S., Sidorenko, J., Robinson, M. R., Thornton, K. R. & Visscher, P. M. Evidence
475 of directional and stabilizing selection in contemporary humans. *PNAS* **115**, 151-156
476 (2018).
- 477 15 Wang, H. *et al.* Genotype-by-environment interactions inferred from genetic effects on
478 phenotypic variability in the UK Biobank. *Science advances* **5**, eaaw3538 (2019).
- 479 16 Miao, J. *et al.* A quantile integral linear model to quantify genetic effects on phenotypic
480 variability. *PNAS* **119**, e2212959119 (2022).
- 481 17 Westerman, K. E. *et al.* Variance-quantitative trait loci enable systematic discovery of
482 gene-environment interactions for cardiometabolic serum biomarkers. *Nat Commun* **13**,
483 3993 (2022).

- 484 18 Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per
485 genotype as a tool to identify quantitative trait interaction effects: a report from the
486 Women's Genome Health Study. *PLoS Genet* **6**, e1000981 (2010).
- 487 19 Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk
488 scores. *Hum Mol Genet* **28**, R133-R142 (2019).
- 489 20 Alliance, P. R. S. T. F. o. t. I. C. D. Responsible use of polygenic risk scores in the clinic:
490 potential benefits, risks and gaps. *Nat Med* **27**, 1876-1884 (2021).
- 491 21 Xiang, R. *et al.* Recent advances in polygenic scores: translation, equitability, methods
492 and FAIR tools. *Genome Med* **16**, 33 (2024).
- 493 22 Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum.
494 *Nature* **618**, 774-781 (2023). <https://doi.org/10.1038/s41586-023-06079-4>
- 495 23 Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood
496 donations can be safely and acceptably decreased to optimise blood supply: study
497 protocol for a randomised controlled trial. *Trials* **15**, 1-11 (2014).
- 498 24 Levene, H. Robust tests for equality of variances. *Contributions to probability and*
499 *statistics*, 278-292 (1960).
- 500 25 Zhang, F. *et al.* OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol*
501 **20**, 1-13 (2019).
- 502 26 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
503 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).
- 504 27 Jordan, D. M., Verbanck, M. & Do, R. HOPS: a quantitative score reveals pervasive
505 horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of
506 human traits and diseases. *Genome Biol* **20**, 1-18 (2019).
- 507 28 Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human
508 complex traits. *Nat Genet* **50**, 746-753 (2018).
- 509 29 Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563
510 individuals yields insights into biology and relationships with other traits. *Nat Neurosci*
511 **23**, 809-818 (2020).
- 512 30 Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from
513 GWAS summary data. *Nat Commun* **9**, 1-12 (2018).
- 514 31 Cole, J. B., Florez, J. C. & Hirschhorn, J. N. Comprehensive genomic analysis of dietary
515 habits in UK Biobank identifies hundreds of genetic associations. *Nat Commun* **11**, 1467
516 (2020).
- 517 32 Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal
518 pleiotropy in causal relationships inferred from Mendelian randomization between
519 complex traits and diseases. *Nat Genet* **50**, 693-698 (2018).
- 520 33 Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing
521 Mendelian randomization analyses using summarized data. *Int J Epidemiol* **46**, 1734-1739
522 (2017).
- 523 34 Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale
524 data. *Gigascience* **8**, giz082 (2019).
- 525 35 Momin, M. M., Lee, S., Wray, N. R. & Lee, S. H. Significance tests for R² of out-of-
526 sample prediction using polygenic scores. *Am J Hum Genet* **110**, 349-358 (2023).
- 527 36 Maier, R. M. *et al.* Improving genetic prediction by leveraging genetic correlations among
528 human diseases and traits. *Nat Commun* **9**, 989 (2018).
- 529 37 Rönnegård, L. & Valdar, W. Detecting major genetic loci controlling phenotypic
530 variability in experimental crosses. *Genetics* **188**, 435-447 (2011).
- 531 38 Pirastu, N. *et al.* Using genetic variation to disentangle the complex relationship between
532 food intake and health outcomes. *PLoS Genet* **18**, e1010162 (2022).

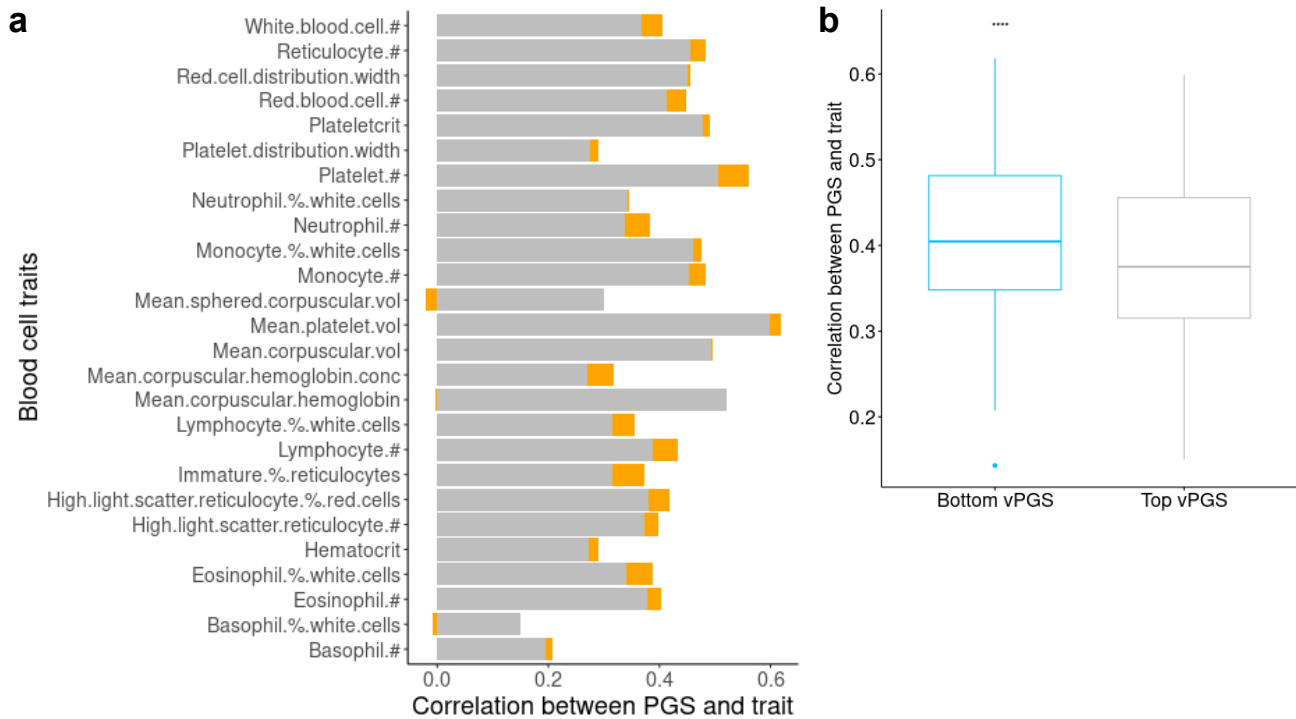
- 533 39 Westerman, K. E. *et al.* Genome-wide gene–diet interaction analysis in the UK Biobank
534 identifies novel effects on hemoglobin A1c. *Hum Mol Genet* **30**, 1773-1783 (2021).
- 535 40 Thompson, A., King, K., Morris, A. P. & Pirmohamed, M. Assessing the impact of
536 alcohol consumption on the genetic contribution to mean corpuscular volume. *Hum Mol*
537 *Genet* **30**, 2040-2051 (2021).
- 538 41 Pashek, R. E. *et al.* Alcohol intake including wine drinking is associated with decreased
539 platelet reactivity in a large population sample. *Int J Epidemiol*, dyad099 (2023).
- 540 42 Selzam, S. *et al.* Comparing within-and between-family polygenic score prediction. *Am J*
541 *Hum Genet* **105**, 351-363 (2019).
- 542 43 Abdellaoui, A., Dolan, C. V., Verweij, K. J. & Nivard, M. G. Gene–environment
543 correlations across geographic regions affect genome-wide association studies. *Nat Genet*
544 **54**, 1345-1354 (2022).
- 545 44 Ballard, H. S. The hematological complications of alcoholism. *Alcohol Health Res World*
546 **21**, 42 (1997).
- 547 45 Thom, C. S., Wilken, M. B., Chou, S. T. & Voight, B. F. Body mass index and adipose
548 distribution have opposing genetic impacts on human blood traits. *Elife* **11**, e75317
549 (2022).
- 550 46 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a
551 wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
- 552 47 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
553 *Nature* **562**, 203-209 (2018).
- 554 48 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
555 datasets. *Gigascience* **4**, 7 (2015). <https://doi.org/10.1186/s13742-015-0047-8>
- 556 49 Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and
557 annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
- 558
- 559



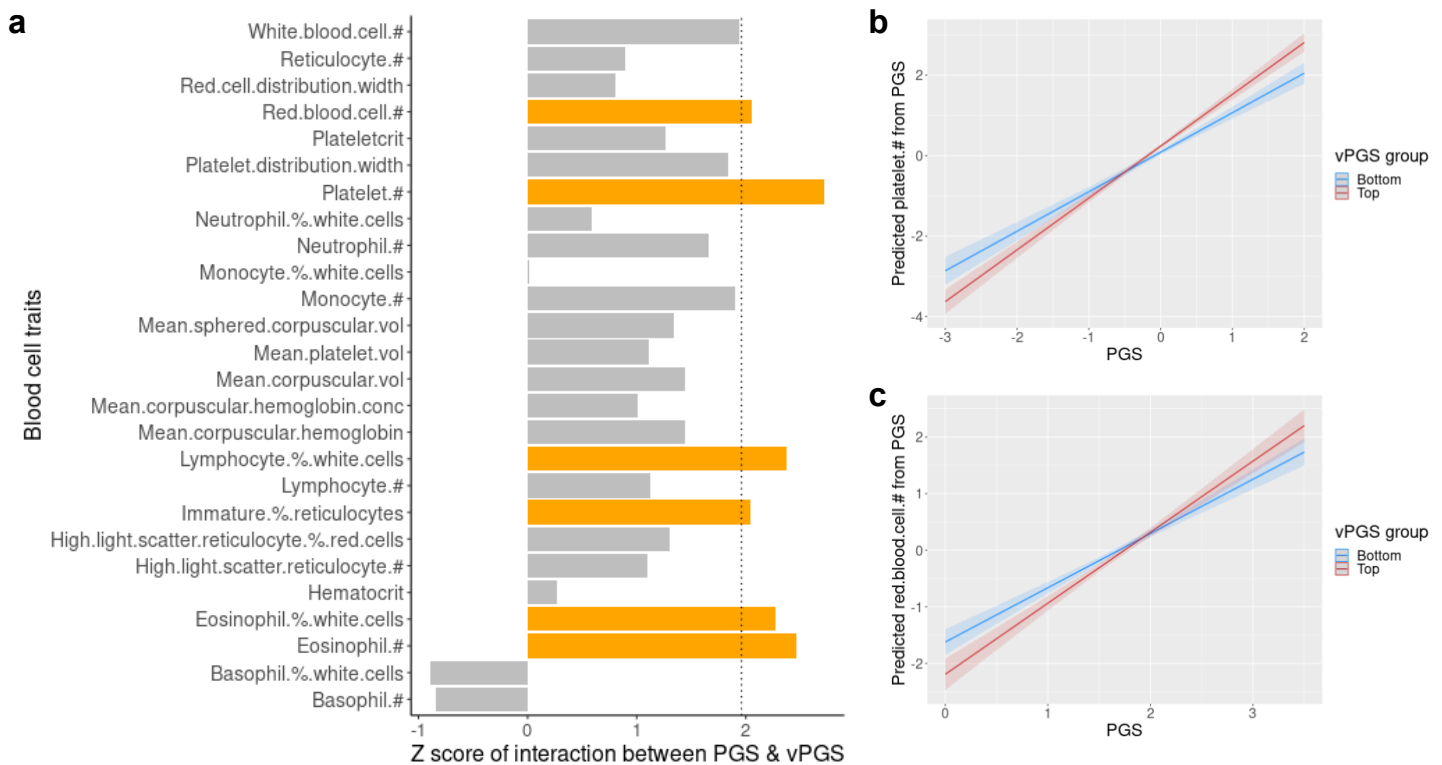
560 **Figure 1.** vQTLs for 29 blood cell traits and their comparison with additive QTLs. **a:** Miami
561 plot showing the best (smallest p-value) vQTLs across 29 blood cell traits (top plot) and the
562 corresponding best additive QTLs (bottom plot). Red dots are genome-wide significant
563 independent vQTLs. **b:** Example of pleiotropic effects of the C allele of rs10803164 for the
564 long non-coding RNA *LINC02768* on blood cell trait variance. Blue indicates the effect on
565 trait variance had $p < 4.6 \times 10^{-9}$ (study-wide GWAS significance). **c:** Genetic correlation
566 between blood cell trait variance and trait level. Blue indicates the correlation had multi-
567 testing adjusted $p < 0.05$. **d:** Selection coefficient estimated by BayesS²⁸ for trait variance
568 and level.
569



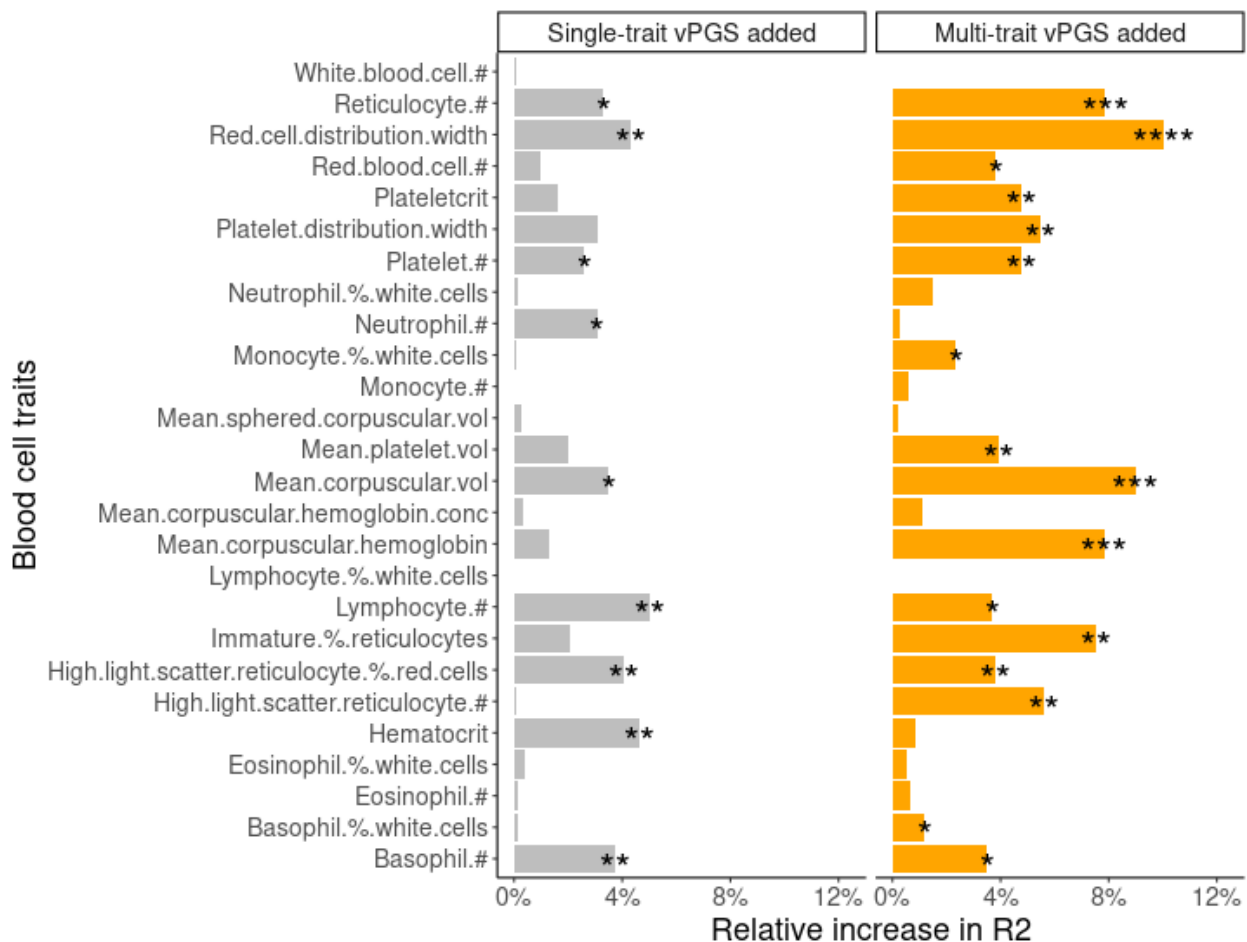
571 **Figure 2. a:** LocusZoom plot of variance QTL mapping for platelet crit (pct) variance at
 572 ALDH2 gene; **b:** Mendelian randomization (MR) of alcohol consumption on variance of blood
 573 cell traits using GSMR³⁰, MR-PRESSO (presso)³² and weighted-median (wm)³³. Diamonds:
 574 significant in 3 methods; grey dots: $p \geq 0.05$; the error bars indicate standard errors; **c:** Effects
 575 of MR of alcohol consumption on variance of corpuscular hemoglobin concentration (mscv);
 576 **d:** Effects of MR of alcohol consumption on variance of corpuscular volume variance (mcv).
 577 Dashed fitted lines indicate the coefficient of Mendelian Randomisation ($b_{xy}=0.07$, $se_{xy}=0.019$
 578 for mscv and $b_{xy}=0.064$, $se_{xy}=0.0188$ for mcv).
 579



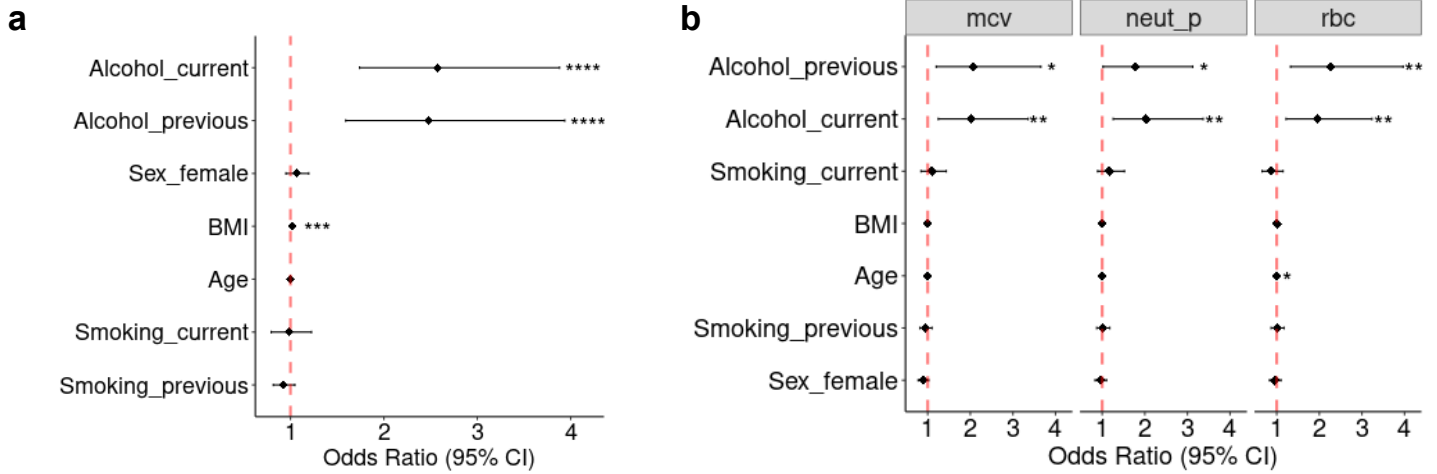
580 **Figure 3.** The variation in the accuracy of PGSs for 27 blood cell traits (Pearson correlation)
 581 between the top and bottom vPGS groups. **a:** Accuracy of PGS in the top vPGS group (more
 582 variable group, grey colour) and the difference (orange) of PGS between the top vPGS group
 583 and the bottom vPGS group (less variable group). #: count; % percentage; vol: volume; conc:
 584 concentration. **b:** Difference of accuracy of PGS between the bottom and top vPGS groups
 585 across 27 blood cell traits. ****: $p < 0.0001$.
 586



588 **Figure 4.** Effects of interaction between PGS and vPGS on blood cell traits. **a:** Effects of
 589 interaction across 27 traits in INTERVAL. The vertical dashed line indicates the z-score
 590 value = 1.96 which equals p-value = 0.05 and bars with z-score value > 1.96 (p < 0.05) are in
 591 orange color. #: count; % percentage; vol: volume; conc: concentration. **b-c:** Examples of
 592 visualised effects of interaction for eosinophil percentage of white cells (eo_p) and neutrophil
 593 count (neut).
 594



595 **Figure 5.** The difference in the variance explained (R^2) between PGS models with or without
596 vPGS. Each bar represents the relative increase in R^2 for the blood cell trait when the PGS
597 model added vPGS. In the left panel, the single-trait vPGS was added to PGS. In the right
598 panel, multi-trait vPGS was added to PGS. #: count; % percentage; vol: volume; conc:
599 concentration. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$ and ****: $p < 0.0001$. P-values were
600 estimated by comparing models with and without vPGS using *r2redux*³⁵.
601



602 **Figure 6.** Association between BMI, age, alcohol drinking and smoking and individuals to be
 603 genetically variable across blood cell traits in INTERVAL. **a:** an overall estimate across 27
 604 blood cell traits. **b:** Estimates for mean corpuscular volume (mcv), neutrophil percentage of
 605 white cells (neut_p) and red blood cell count (rbc) which were significant Mendelian
 606 Randomisation analyses. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$ and ****: $p < 0.0001$.
 607