Genetic Determinants of Blood Cell Traits Influence Susceptibility to Childhood Acute Lymphoblastic Leukemia

Linda Kachuri[1], Soyoung Jeon[2], Andrew T. DeWan[3,4], Catherine Metayer[5], Xiaomei Ma[4], John S. Witte[1,6,7], Charleston W. K. Chiang[2,8], Joseph L. Wiemels[2,9], Adam J. de Smith[2,9]

Affiliations:

1. Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA
2. Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA
3. Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health, New Haven, CT
4. Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT
5. Division of Epidemiology & Biostatistics, School of Public Health, University of California Berkeley, Berkeley, CA
6. Institute for Human Genetics, University of California, San Francisco, San Francisco, USA
7. Department of Urology, University of California, San Francisco, San Francisco, USA
8. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA
9. Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA

Corresponding Author:

Adam J. de Smith

Center for Genetic Epidemiology
Keck School of Medicine, University of Southern California
1450 Biggy Street, Health Sciences Campus
Los Angeles, CA
desmith@usc.edu

ABSTRACT

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer. Despite overlap between genetic risk loci for ALL and hematologic traits, the etiological relevance of dysregulated blood cell homeostasis remains unclear. We investigated this question in a genome-wide association study (GWAS) of ALL (2666 cases, 60,272 controls) and multi-trait GWAS of 9 blood cell indices in the UK Biobank. We identified 3000 blood cell trait-associated ($P<5.0\times10^{-8}$) variants, explaining 4.0% to 23.9% of trait variation, and including 115 loci associated with blood cell ratios (LMR: lymphocyte/monocyte, NLR: neutrophil/lymphocyte, PLR: platelet/lymphocyte). ALL susceptibility was genetically correlated with lymphocyte counts (r_g=0.088, p=4.0×10^{-4}) and PLR (r_g= -0.072, p=0.0017). In Mendelian randomization analyses, genetically predicted increase in lymphocyte counts was associated with increased ALL risk (Odds ratio (OR)=1.16, p=0.031) and strengthened after accounting for other cell types (OR=1.48, p=8.8×10^{-4}). We observed positive associations with increasing LMR (OR=1.22, p=0.0017) and inverse effects for NLR (OR=0.67, p=3.1×10^{-4}) and PLR (OR=0.80, p=0.002). Our study shows that a genetically induced shift towards higher lymphocyte counts, overall and in relation to monocytes, neutrophils, and platelets, confers an increased susceptibility to childhood ALL.

INTRODUCTION

The hematopoietic system is remarkably orchestrated and responsible for some of the most important physiological functions, such as the production of adaptive and innate immunity, nutrient transport, clearance of toxins, and wound healing. Genetic factors contribute significantly to inter-individual variation in blood cell phenotypes, with heritability estimates for most blood cell traits ranging from 50-90% in twin studies to 30-40% in population-based studies of array-based heritability[1, 2, 3, 4]. Genome-wide association studies (GWAS) conducted in large population-based studies have revealed the highly polygenic nature of blood cell traits, with over 5,000 independently associated genetic loci identified to date[4, 5, 6]. These studies have provided insights into the genetic regulation of hematopoiesis and how dysregulation in blood cell development can lead to disease[7]. Genetic variants associated with blood cell variation have been implicated in the risk of immune-related conditions such as asthma, rheumatoid arthritis, and type 1 diabetes, and in rare blood disorders[4, 5, 6]. Positive genetic correlation was found between counts of varying blood cell types and the risk of myeloproliferative neoplasms, a group of diseases primarily of older age and characterized by the overproduction of mature myeloid cells[8]. However, the contribution of heritable variation in blood cell traits to the risk of other hematologic cancers has not been examined.

Acute lymphoblastic leukemia (ALL) is a malignancy of white blood cells, developing from immature B-cells or T-cells, and is the most common cancer diagnosed in children under 15 years of age[9]. Despite significant advances in treatment in recent decades and the corresponding improvements in survival rates[10], ALL remains one of the leading causes of pediatric cancer mortality in the United States[11]. In addition, childhood ALL patients may endure severe toxicities during treatment, and survivors face long-term treatment-related morbidities and mortality[12, 13]. Thus, understanding the etiology of ALL remains important to identify avenues for disease prevention as well as potential novel treatment targets.

In most cases, the development of ALL is thought to follow a two-hit model of leukemogenesis, with *in utero* formation of a preleukemic clone and subsequent postnatal acquisition of secondary somatic mutations that drive progression to overt leukemia[14]. Epidemiologic studies have identified several genetic and non-genetic risk factors for ALL (reviewed in Williams et al.[15] and Greaves[14]), but the biological mechanisms through which they promote leukemogenesis are largely unknown. GWAS of childhood ALL have revealed at least 12 common genetic risk loci to date, including at genes involved in hematopoiesis and early lymphoid development[16], such as *ARID5B*, *IKZF1*, *CEBPE*, *GATA3*, *BMI1*, *IKZF3*, and *ERG* [17, 18, 19, 20, 21, 22, 23, 24]. Intriguingly, several childhood ALL risk regions have also been associated with variation in blood cell traits[4, 6, 22, 23, 25] and a recent phenome-wide association study (PheWAS) of childhood ALL identified platelet count as the most enriched trait among known ALL risk loci[26]. A comprehensive study of the role of blood cell trait variation in the etiology of childhood ALL is, therefore, warranted.

In this study, we utilize genome-wide data available from the UK Biobank (UKB) resource[27] to perform a GWAS of blood cell traits and apply the discovered loci to a GWAS of childhood ALL in 2,666 cases and 60,272 controls of European ancestry. We assess the shared genetic architecture between blood cell phenotypes and childhood ALL and conduct Mendelian randomization (MR) and mediation analyses to disentangle putative causal effects of variation in blood cell homeostasis on ALL susceptibility.

RESULTS

Genetic determinants of blood cell traits

Genome-wide analyses revealed a substantial genetic contribution to blood cell trait variation. Heritability (h_g) estimated from GWAS summary statistics on the full analytic cohort (median n=335,030) ranged from 3.1% for basophils to 21.8% for platelets (Figure 1, Supplementary Table 1). There was significant genetic correlation between all blood cell populations, which supports our rationale for using MTAG to leverage this shared genetic basis (Figure 2, Supplementary Table 2). Among non-composite traits, the largest correlations were observed between pairs of white blood cells: monocytes and neutrophils (r_g=0.45,

P=1.8×10^{-83}), basophils and neutrophils (r_g=0.44, P=4.0×10^{-33}), and lymphocytes and monocytes (r_g=0.41, P=1.3×10^{-68}). Platelet counts were also significantly correlated with neutrophils (r_g=0.24, P=1.7×10^{-26}), lymphocytes (r_g=0.22, P=3.8×10^{-35}), and monocytes (r_g=0.21, P=6.8×10^{-26}). Cell type ratios LMR and NLR were primarily correlated with their component traits and with each other. PLR was significantly correlated with all phenotypes, including monocytes (r_g=-0.20, P=2.5×10^{-20}), neutrophils (r_g=-0.16, P=2.7×10^{-8}), basophils (r_g=-0.21, P=1.1×10^{-10}), and eosinophils (r_g=-0.11, P=6.3×10^{-7}).

After applying our instrument selection criteria (discovery P_{MTAG}<5×10^{-8}, replication P_{MTAG}<0.05; LD r^2<0.05 within 10 Mb) we identified 3000 variants that were independent within, but not across hematological phenotypes (Supplementary Data 1). Of these, 2500 were associated with a single phenotype, 378 were associated with two, and 122 were instruments for 3 or more blood cell traits. The number of available instruments ranged from 157 for basophils to 692 for platelets (Supplementary Table 3). The proportion of trait variation accounted for by each set of instruments was estimated in the replication sample (100,284 to 100,764 individuals) and ranged between 5.1% for basophils to 24.4% for platelets (Supplementary Table 3). All traits had strong instruments with mean F statistics exceeding 60. Previous GWAS have not examined cell type ratios, while we identified 770 instruments specifically for ratio phenotypes: LMR, NLR, PLR. To assess whether these signals are captured by existing associations with cell counts or proportions, we performed clumping (LD r^2<0.05 within 10 Mb) with loci reported in Vuckovic et al.[6], a meta-analysis of UK Biobank and Blood Cell Consortium cohorts. This yielded 225 independent, ratio-specific variants in 115 cytoband loci, including six missense mutations (Figure 3, Supplementary Data 2).

In-silico functional annotations identified overlap with multiple regulatory elements among all genetic instruments (Supplementary Data 1). A total of 324 variants were predicted to be in the top 10% of deleterious substitutions genome wide (CADD scores >10)[28] and 138 had significant (p<0.05) evidence of overlap with open chromatin (FAIRE, DNAse, Pol-II, CTCF, and Myc) based on ENCODE data from up to 14 cell types. Over 80% of all instruments (n=2405) were expression quantitative trait loci (eQTL) in whole blood (FDR<0.05) based on results from eQTLGen[29] (Supplementary Table 4). Fewer immune cell eQTLs were identified, although these reference datasets were much smaller. The highest proportion of eQTLs were observed in monocytes (27.0%), T-cells (23.6%), and neutrophils (21.1%), followed by B-cells (11.4%) (Supplementary Table 4). The proportion of immune cell eQTLs was broadly similar across categories of instruments, ranging from 26% for neutrophils to 16% for platelets (Supplementary Figure 2). For every instrument class, T-cell eQTLs were the most common. Lymphocytes were the most prevalent instrument class among cell-specific immune eQTLs.

Among instruments specific to one blood cell trait with eQTL effects in >2 cell types (Supplementary Figure 3A), the largest number of target genes was observed for platelet-specific and monocyte-specific instruments. Instruments for >4 blood cell traits with eQTL effects in >3 cell types (Supplementary Figure 3B) had a predominance of immune function genes in the HLA region and a previously identified ALL risk gene, *BAK1*. Among instruments for >5 blood cell traits with eQTL effects in a single cell/tissue type (Supplementary Figure 4) notable findings included multiple ALL risk genes (*IKZF3*, *CDKN2A, CDKN2B*, *IRF1*), and *FLT3*, a receptor tyrosine kinase that serves as a key regulator of hematopoiesis and is frequently mutated in ALL and acute myeloid leukemia (AML).

Impact of blood cell variation on ALL risk

Associations between genetic determinants of blood cell trait variation and ALL susceptibility were investigated using a GWAS meta-analysis comprised of 2666 cases and 60,272 controls. Heritability of ALL was 18.1% (h_g=0.181, SE=0.013), converted to the liability scale using SEER estimates of ALL lifetime risk in Non-Hispanic whites (0.15%) (Figure 1). At the genome-wide level, we observed positive correlations with ALL risk for increasing lymphocyte counts (r_g=0.088, P=4.0×10^{-4}), LMR (r_g=0.065, P=0.012), and neutrophils (r_g=0.051, P=0.027). Increasing PLR, corresponding to higher levels of platelets compared to lymphocytes, was inversely correlated (r_g= -0.072, P=1.7×10^{-3}) with ALL risk (Figure 4).

Next, we conducted Mendelian Randomization (MR) analyses using genetic instruments developed in the UK Biobank to assess the putative causal relevance of blood cell trait variation in childhood ALL etiology (Figure 5, Supplementary Table 5). Primary analyses were based on the maximum likelihood (ML) and inverse variance weighted multiplicative random effects estimators (IVW-mre). To ensure that the observed results were robust against potential violations of MR assumptions, particularly horizontal pleiotropy, these estimators were complemented with shrinkage-based MR RAPS (Robust Adjusted Profile Score)[30, 31] and MR pleiotropy residual sum and outlier (PRESSO) approaches[32].

We did not detect evidence of directional horizontal pleiotropy based on the MR Egger intercept test for any blood cell traits (Supplementary Table 6). However, there was indication of balanced horizontal pleiotropy for all phenotypes based on Cochran's Q (P_Q<0.05) and the PRESSO Global test (P_{Global}<0.05). Among white blood cells, a 1-SD increase in lymphocyte counts was associated with a modest increase in ALL risk (OR_{ML}=1.16, 95% CI: 1.01-1.33, p=0.035; $OR_{IVW\text{-}mre}$=1.15, 0.99-1.34, p=0.061). This effect was slightly attenuated in pleiotropy-corrected analyses (OR_{PRESSO}=1.14, 0.98-1.32, p=0.087; OR_{RAPS}=1.16, 1.01-1.34, p=0.033), but the effect size distortion was not significant (P_{Dist}=0.88). There was no significant association between counts of other white blood cell types (monocytes, neutrophils, basophils, eosinophils) or platelets and ALL risk (Figure 4, Supplementary Table 5).

Considering ratios, indicating a genetic predisposition to a shift in the counts of one cell type relative to another, revealed several associations. An increase in LMR was associated with an approximately 22% increase in ALL risk (per 1-unit increase: OR_{ML}=1.23, 1.07-1.41 p=4.5×10^{-3}; $OR_{IVW\text{-}mre}$=1.22, 1.00-1.50, p=0.052). Accounting for the influence of potentially pleiotropic outliers slightly attenuated this effect (OR_{RAPS}=1.14, 0.99-1.32; OR_{PRESSO}=1.18, 1.01-1.38). An inverse association with ALL risk was observed for increasing NLR (OR_{ML}=0.67, 0.54-0.83, p=3.1×10^{-4}; $OR_{IVW\text{-}mre}$=0.67, 0.49-0.92, p=0.012), denoting a shift to higher levels of neutrophils compared to lymphocytes. Increased PLR was also associated with a lower risk of ALL (OR_{ML}=0.80, 0.70-0.92, p=2.0×10^{-3}; $OR_{IVW\text{-}mre}$=0.80, 0.67-0.96, p=0.012). Associations with ALL for both phenotypes remained stable in sensitivity analyses correcting for pleiotropy (NLR: OR_{PRESSO}=0.77, 0.66-0.98, p=0.036; PLR: OR_{PRESSO}=0.82, 0.70-0.96, p=0.014) and outliers (NLR: OR_{RAPS}=0.73, 0.58-0.91, p=5.6×10^{-3}; PLR: OR_{RAPS}=0.85, 0.73-0.98, p=0.025).

In addition to analytically correcting for pleiotropy, we also conducted analyses using a filtered set of genetic instruments excluding variants that showed evidence of heterogeneity based on Cochran's Q in the observed causal effect estimates (Supplementary Table 7). These sensitivity analyses confirmed our previous findings showing that an increase in lymphocyte counts ($OR_{IVW\text{-}mre}$=1.18, p=7.4×10^{-3}) and LMR ($OR_{IVW\text{-}mre}$=1.19, p=0.016) conferred a modest increase in ALL risk, while increased NLR ($OR_{IVW\text{-}mre}$=0.67, p=7.4×10^{-3}) and PLR ($OR_{IVW\text{-}mre}$=0.82, p=5.8×10^{-3}) were associated with lower risk.

Assessment of additional diagnostic tests indicated that our analysis was robust against main threats to validity (Supplementary Table 6). Associations were not affected by violations of the no measurement error assumption (I^2_{GX}>0.98). MR Steiger directionality test was used to orient the causal effects and confirmed that instruments for blood cell traits were affecting ALL susceptibility, not the reverse. The direction of this effect was robust across all traits, including lymphocytes (P=5.0×10^{-135}), LMR (P=5.4×10^{-98}), NLR (P=2.1×10^{-10}), and PLR (P=1.2×10^{-109}). Our analyses were powered to at least 80% to detect a minimum OR of 1.17 (equivalent to 0.85) for LMR and PLR, OR of 1.20 for lymphocytes, and OR of 1.28 (equivalent to 0.78) for NLR (Supplementary Figure 5).

Next, we conducted multivariable MR analyses to assess the degree to which specific blood cell subtypes exert independent direct effects on ALL (Supplementary Table 8). Among lymphocytes, monocytes, neutrophils, and platelets, only lymphocytes were independently associated with ALL (OR_{MVMR}=1.18, 1.06-1.31, p=3.3×10^{-3}) based on all variants and when restricting to instruments associated with each exposure ($OR_{MVMR\text{-}mod}$=1.43, 1.16-1.76, p=8.8×10^{-4}). This was confirmed using MV LASSO, which only retained lymphocytes. Among cell type ratios PLR was associated with ALL when considering all variants for all

traits (OR_{MVMR}=0.90, 0.82-0.99, p=0.033), but not in the instrument-specific analysis. PLR was the only trait selected by MV LASSO. Lastly, we explored the degree to which causal effects observed for ratio phenotypes were mediated by any of their component traits. We did not observe any statistically significant indirect effects, which suggests that the impact on ALL susceptibility observed for LMR, NLR, and PLR could not be attributed to effects on the counts of lymphocytes, monocytes, neutrophils, or platelets (Supplementary Table 9).

Exploring mechanisms of ALL susceptibility

Lastly, MR-Clust[33] was applied to blood cell traits associated with ALL to identify subgroups of variants with homogenous causal effects and novel ALL risk variants (Supplementary Figure 6; Supplementary Table 10). Clustering instruments for lymphocytes identified 10 variants indicating a large effect of increasing lymphocyte counts on ALL (OR=7.63). LMR instruments contained a cluster of 9 variants (OR=3.64). The largest cluster was identified for PLR, with 18 variants (OR of 0.27 per 1-unit increase in the ratio). A cluster comprised of 10 variants implied an extremely large inverse effect of NLR on ALL (OR=0.039). The substantive clusters were largely distinct, with one variant shared by all four traits (rs28447467: P_{ALL}=0.026). Across all clusters, two variants were statistically significantly associated with ALL after correcting for the number of independent variants across all phenotypes tested ($P_{ALL}<5\times10^{-5}$): rs6430608-C (OR=1.28, 1.15-1.41, $P_{ALL}=2.5\times10^{-6}$) near *CXCR4* on 2q22.1 and rs76428106-C (OR=1.79, 1.36-2.35, $P_{ALL}=3.2\times10^{-5}$) in *FLT3* on 13q12.2. The former is an intergenic variant specific to NLR with cis-effects on whole blood gene expression of *MCM6* ($P_{eQTL}=4.0\times10^{-28}$) and *DARS1* ($P_{eQTL}=3.7\times10^{-37}$), based on data from eQTLGen[29]. On the other hand, rs76428106, an intronic variant in *FLT3* and an eQTL for *FLT3* in whole blood ($P_{eQTL}=1.0\times10^{-11}$)[29] was included in substantive clusters for lymphocytes and PLR, but assigned to the "junk" cluster for LMR. Annotation of variants in substantive clusters revealed a predominance of previously reported associations with blood cell trait variation, as well as autoimmune and allergic conditions, such as type 1 diabetes, Crohn's, asthma, and IgA deficiency (Supplementary Table 10).

Notable instruments assigned to "junk" clusters for LMR, PLR, and NLR included established ALL risk variants rs4948492 and rs4245597 (*ARID5B*, 10q21.2), rs2239630 (*CEBPE*, 14q11.2), rs78697948 (*IKZF1*, 7p12.2), and rs74756667 (8q24.2). These variants were also classified as outliers based on Cochran's Q, suggesting that their effects on ALL susceptibility are predominantly mediated through pathways other than regulation of blood cell profiles. We formally tested this hypothesis using mediation analysis[34] by decomposing the total SNP effect on ALL into direct and indirect (mediated) effects. For variants that were instruments for more than one blood cell phenotype, mediation was only explored for phenotypes significantly associated with ALL. Mediator-outcome effects were obtained from MR results excluding outliers (see Supplementary Table 7). For rs4245597 (*ARID5B*), only a small proportion of its effect on ALL risk was mediated by blood cell traits, 1.65% (0.79 – 2.51) was attributed to NLR and 0.84% (0.23 – 1.44) to PLR (Supplementary Table 11). Mediated effects attributed to LMR were not statistically significant for rs4245597 (*ARID5B*; 0.75%), rs2239630 (2.16%; *CEBPE*), and rs78697948 (1.72%; *IKZF1*).

Modest, but statistically significant, mediated effects were observed for putative novel ALL risk variants rs6430608 (NLR: 4.72%, 2.26 – 7.20) and rs76428106 (PLR: 2.43%, 0.68 – 4.19; lymphocytes: 2.51%, 0.67 – 4.35). The LMR-mediated effect of rs76428106 was larger (11.39%), but in the opposite direction from the effect of LMR on ALL, indicating pleiotropic effects consistent with the assignment of rs76428106 to the "junk" cluster for LMR. Of the six traits linked to this variant, its effect on monocytes (rs76428106-C: β=0.484, $P=1.3\times10^{-310}$) was by far the strongest. In MR analyses monocyte counts were not implicated in ALL susceptibility, suggesting that rs76428106 may be influencing ALL via other pathways or broad effects on hematopoiesis.

DISCUSSION

Hematopoiesis is a tightly regulated hierarchical process designed to maintain optimal physiological ranges. Abnormalities in blood counts may be indicative of systemic inflammation or the presence of infections, and serve as indicators for a wide range of potentially adverse health conditions including inborn defects in hematopoiesis. While the responsive and sensitive nature of blood cell counts makes them useful clinical biomarkers, this poses a challenge for etiological studies. Elevated white blood cell counts are an established diagnostic feature of childhood ALL reflecting the overproduction of immature lymphocytes, or lymphoblasts, in the bone marrow. However, blood counts at a single time point, particularly in cancer cases, may not be representative of the individual's stable, pre-diagnostic blood count profile, making it difficult to disentangle disease correlates from risk factors.

In this study, we leveraged the highly heritable nature of blood cell variation to evaluate its role in ALL pathogenesis without the limitations inherent in observational blood count measures. Our overarching finding is the convergence of genetic mechanisms resulting in increased lymphocyte counts and increased ALL susceptibility. Using genetic correlation and Mendelian randomization (MR) approaches, we observed a significant positive relationship between a genetically predicted increase in lymphocyte counts and ratio of lymphocytes to monocytes (LMR) and risk of ALL. Conversely, genetic predisposition to an increased ratio of platelets to lymphocytes (PLR) and neutrophils to lymphocytes (NLR) was inversely associated with ALL risk. These effects were largely robust to analytic corrections for horizontal pleiotropy and in some cases the removal of instruments contributing to heterogeneity strengthened the observed associations. Taken together, these results provide new insight into ALL etiology and point to a specific shift in blood cell homeostasis which confers an increased susceptibility.

However, the ways in which a genetic predisposition to over-production of lymphocytes may confer ALL risk are likely multifactorial. Stable and consistent causal effect estimates for lymphocytes, PLR, and NLR do not imply a single biological mechanism, even if they are estimated using valid instruments that primarily regulate the target blood cell trait. Acknowledging this, we propose two distinct, though not necessarily mutually exclusive, biological mechanisms related to the two-hit model of childhood ALL development that warrant further investigation. First, the initiating genetic lesions in childhood ALL, such as *ETV6-RUNX1* gene fusions, arise prenatally in most cases and require additional somatic mutations to progress to overt leukemia[14, 35]. The presence of common alleles across the spectrum of variants that subtly tune lymphocyte production may lead to an elevated risk of ALL by increasing the reservoir of preleukemic cell clones which, in turn, may increase the chances of acquiring "second-hit" oncogenic events and progression to ALL.

Second, the "delayed infection" hypothesis posits that children who lack early microbial exposures may have an unmodulated immune network that results in dysregulated immune responses to infectious stimuli later in childhood, and an increased risk of ALL[14]. This is supported by epidemiological evidence, with proxies for early-life infectious exposure, including daycare attendance and higher birth order, associated with a reduced risk of ALL[36, 37], and by experimental models that demonstrated higher ALL incidence in mice with delayed exposure to pathogens[38, 39]. Further, children who develop ALL have been found to have different cytokine profiles at birth[40, 41]. Genetic variants that influence the blood cell phenotypes associated with ALL risk in our study may confer their effects via modulation of neonatal immune development and of immune responses to infections in childhood that may trigger ALL development. A shift towards increased lymphocytes to neutrophils is suggestive of increased adaptive immunity and lymphocyte activation in response to infections. This is consistent with our findings for NLR, a marker of increased inflammation, which was associated with reduced ALL risk. Similarly, reduced immune-inflammatory responses and increased activation of lymphocytes would be denoted by a higher ratio of lymphocytes to monocytes and lymphocytes to platelets[42], both of which were associated with increased ALL risk in our study.

Previous studies have noted an overlap between ALL risk loci and genomic regions associated with blood cell phenotypes[21, 22, 23, 24], however, ours is the first study to systematically analyze the contribution and

causal effects of genetic variation across blood cell traits in ALL etiology. In a recent PheWAS, ALL risk variants were found to be enriched for regulation of platelet levels, but the overall association between platelet counts and ALL was null in Mendelian randomization and genetic score analyses using 223 platelet-associated variants[26]. This is consistent with our findings using over 600 genetic instruments for platelets, which indicate that variation in platelet counts alone does not influence ALL susceptibility, whereas PLR, which captures dysregulation of platelets in relation to lymphocytes, has a significant impact.

Indeed, we identified the cell type ratios LMR, NLR, and PLR as independent risk factors for ALL, with distinct genetic mechanisms not captured by their component traits. In multivariable MR analyses that concurrently modeled the effects of lymphocyte, monocyte, neutrophil, and platelet counts on ALL, lymphocytes remained as the only independent risk factor and this association with ALL strengthened compared to univariate analyses. However, there was no evidence that the total MR effects for LMR, NLR, and PLR were mediated either by lymphocytes or by the other cell populations. This implies that while dysregulation of lymphocyte homeostasis seems to be a key factor, it should be considered in the broader context of other blood cell subtypes.

In addition to identifying novel susceptibility pathways, our study also provides insights into the underlying mechanisms of several established ALL risk variants in *ARID5B*, *CEBPE*, *IKZF1*, and at chromosome 8q24. Despite significant associations with LMR and, in the case of *ARID5B*, with NLR and PLR, these variants were flagged as pleiotropic outliers in MR analyses, which was subsequently confirmed in mediation analyses. This supports that the overall effects of these loci on ALL risk are largely mediated by pathways other than regulation of blood cell trait variation, although we cannot rule out potential effects of these variants on early stages of hematopoiesis that may influence ALL development. Our MR clustering analysis also identified two putative novel ALL risk variants among genetic instruments for various blood cell traits: rs6430608 on 2q22.1 and rs76428106 in *FLT3* on 13q12.2.

While additional studies are needed to confirm their association with ALL, the locus at *FLT3* is of particular interest as this same variant was recently associated with an increased risk of autoimmune thyroid disease and AML[43]. The AML/ALL risk-increasing allele, rs76428106-C, has a frequency of approximately 1% in the general population (1.3% in UKB, 1.4% in ALL GWAS) and is reported as a splicing QTL in GTEx ($P_{sQTL}=1.3\times10^{-8}$). Indeed, rs76428106-C was shown to generate a cryptic splice site resulting in truncation of the FLT3 protein, but an increase in FLT3 ligand levels[43]. Gain-of-function somatic mutations in *FLT3* are relatively frequent in childhood ALL[44] and, while rs76428106 has greater effects on the production of myeloid cells than lymphocytes in our analyses, its putative effects on ALL risk may largely occur via activation of the RAS/MAPK pathway. This activation is likely to be restricted to key developmental decisions of hematopoietic cells given the delimited expression of FLT3 to hematopoietic stem and progenitor cells[45]. Less is known about the 2q22.1 variant, rs6430608, which is an eQTL for *MCM6* in blood and *CXCR4* in multiple tissues. *MCM6* is upregulated in multiple cancers and is believed to regulate DNA replication and activate MAPK/ERK signaling[46]. *CXCR4* is a chemokine receptor that facilitates HIV cell entry and regulates immune cell migration, including retention of B-cell precursors in the bone-marrow, and is being pursed as a therapeutic target in ALL and AML[47, 48].

Several limitations of this study should be acknowledged. First, genetic instruments were developed for blood cell phenotypes measured in adult participants in the UK Biobank due to a paucity of adequately powered GWAS of blood cell traits in newborns or children. Environmental exposures throughout the life course influence blood cell dynamics, which has implications for the accuracy of genetic association estimates at different time points across the lifespan. While studies of blood cell development in pediatric populations should be pursued, the true underlying genetic architecture is not affected by age. This is also supported by studies of other complex traits, such as BMI, which showed that genetic risk scores developed in adults accurately predict weight gain in early childhood[49]. Therefore, we would expect any error in our genetic instruments developed in adults to bias MR results towards the null.

Second, our analysis was limited to broad classes of cell types, such as lymphocytes, and in future studies it will be important to distinguish between subpopulations of B-cell and T-cell lymphocytes. B-cell precursor ALL, the most common subtype, likely has distinct etiologic mechanisms from T-cell ALL[15]. Of relevance to our findings, the epidemiological evidence for the two-hit model of leukemogenesis is more compelling for B-cell ALL than for T-cell ALL[14] and GWAS have revealed that hematopoietic transcription factor genes confer stronger effects on B-cell ALL risk[24]. We were also unable to characterize the effect of blood cell traits on B-cell ALL versus T-cell ALL or on specific molecular subtypes, or to explore the potential for germline-somatic interactions with specific ALL mutational signatures.

Finally, MTAG assumes that the variance-covariance matrix of effect sizes is homogeneous across all variants, which may be violated for SNPs that are null for one trait but non-null for other traits[50]. Replication is the best way to assess the credibility of observed associations, therefore our two-stage discovery and replication approach should minimize false positives. Furthermore, in a two-sample setting false positive instruments would bias Mendelian randomization estimates towards the null, not induce a spurious signal.

Despite some limitations, this study has important strengths that support the robustness of our findings. Our instrument development approach was optimized for Mendelian randomization studies of cancer etiology. The large sample size of the UK Biobank cohort allowed us to apply appropriate exclusions while retaining a sufficient number of participants for a two-stage discovery and replication analysis. Furthermore, applying the MTAG framework increased statistical power for identifying genetic determinants of specific blood cell traits, while taking into account the correlation between these phenotypes. This resulted in a set of strong genetic instruments explaining between 5% and 24% of variation in the target blood cell trait. These variants were enriched for multiple regulatory features, with over 80% having significant effects on gene expression in whole blood and up to 27% of instruments classified as eQTLs in immune cell subtypes, albeit with a limited degree of cell-type specificity in the eQTL effects across instrument classes. In addition, we characterized the genetic determinants of blood cell ratios, specifically LMR, NLR, and PLR, which have received considerably less attention in genetic association studies. A GWAS of PLR and NLR was conducted in 5901 healthy Dutch individuals, which identified one significant locus for PLR[51]. Examining these ratio phenotypes revealed novel ALL susceptibility pathways and helped contextualize the observed results for lymphocytes and platelets. Finally, the causal interpretation of our results depends on the validity of fundamental MR assumptions and, to this end, we employed a range of MR estimation methods with different underlying assumptions and conducted multiple diagnostic tests to interrogate the robustness of our results with respect to confounding, horizontal pleiotropy, and weak instrument bias.

In conclusion, we demonstrate for the first time that a genetic propensity for overproduction of lymphocytes, particularly in relation to other blood cell types, is associated with an increased risk of childhood ALL in Europeans. It will be important to elucidate the underlying biological mechanisms of our findings, and to assess their transferability to non-European populations.

METHODS

Population for Blood Trait Instrument Development

The UK Biobank (UKB) is a population-based prospective cohort of over 500,000 individuals aged 40-69 years at enrollment in 2006-2010 who completed extensive questionnaires on health-related factors, underwent physical assessments, and provided blood samples[27]. The UKB has been developed as a research resource and all participants provided informed consent at recruitment under the broad consent model. Four Beckman Coulter LH750 instruments were used to analyze blood samples collected in 4ml EDTA vacutainers. The LH750 instrument is a quantitative, automated hematology analyzer and leukocyte differential counter for in vitro diagnostic use in clinical laboratories. Samples were analyzed at the UK Biobank central laboratory within 24 hours of blood draw.

Quality control steps for this dataset have been previously described[52]. Briefly, genetic association analyses were restricted to individuals of predominantly European ancestry identified based on self-report and refined by excluding samples with any of the first two genetic ancestry principal components (PCs) outside of 5 standard deviations (SD) of the population mean. We removed samples with discordant self-reported and genetic sex, as well as one sample from each pair of first-degree relatives identified using KING [53]. Using a subset of genotyped autosomal variants with minor allele frequency (MAF)≥0.01 and call rate ≥97%, we filtered samples with call rates <97% or heterozygosity >5 standard deviations (SD) from the mean, leaving 413,810 individuals available for analysis.

Additional exclusions were applied to optimize our dataset for developing genetic instruments for studies of cancer etiology by removing subjects with medical conditions that would alter blood cell proportions by pathophysiological conditions (n=13,597), such as pre-malignant myelodysplastic syndromes, autoimmune diseases, and immunodeficiencies, including HIV (Supplementary Figure 1). Blood counts (10^9 cells/L) outside of the LH750 reportable range and extreme outliers (>99th percentile) were excluded. Remaining values were converted to normalized z-scores with mean=0 and SD=1. In addition to overall blood cell counts we also examined relative concentrations: lymphocyte to monocyte ratio (LMR), neutrophil to lymphocyte ratio (NLR), and platelet to lymphocyte ratio (PLR).

Genome-Wide Association Study of Blood Cell Traits

UKB participants were genotyped on the UK Biobank Affymetrix Axiom array (89%) or the UK BiLEVE array (11%) with imputation performed using the Haplotype Reference Consortium (HRC) and the merged UK10K and 1000 Genomes (1000G) phase 3 reference panels[27]. We excluded variants that were out of Hardy-Weinberg equilibrium in cancer-free individuals ($P_{HWE}<1\times10^{-5}$) or had low imputation quality (INFO<0.30). Analyses were restricted to 10,369,434 variants with MAF≥0.005.

Genome-wide association analyses were conducted using linear regression in PLINK 2.0 (October 2017 version). Blood cell traits were analyzed using a two-stage GWAS, with a randomly sampled 70% of the cohort used for discovery and the remaining 30% reserved for replication, followed by Multi-Trait Analysis of GWAS (MTAG)[50]. Models for each trait were adjusted for age, age^2, sex, genotyping array, the first 15 PC's, cigarette pack-years, blood count device ID, and assay date. The resulting summary statistics were analyzed using MTAG, which has been shown to increase power to detect associations for correlated phenotypes [50]. A key feature of this framework is distinguishing between different sources of correlation, such as genetic correlation and correlations due to sample overlap or biases in GWAS effect sizes due to population stratification or cryptic relatedness [50]. MTAG integrates this information via a generalization of inverse-variance-weighted meta-analysis and outputs trait-specific effect estimates for each SNP [50]. Genetic instruments were selected from MTAG results and defined as independent variants (linkage disequilibrium (LD) $r^2<0.05$ in a clumping window of 10,000 kb) with $P<5\times10^{-8}$ in the discovery stage and $P<0.05$ and consistent direction of effect in the replication stage.

The functional relevance of the genetic instruments for blood cell traits was assessed using in-silico functional annotations: Combined Annotation Dependent Depletion (CADD) scores[28] and Regulome DB[54]. We also explored associations with gene expression in whole blood in eQTLGen[29], a meta-analysis of 31,684 subjects, and immune-cell specific effects in: DICE (Database of Immune Cell Expression)[55], a dataset of 91 healthy blood donors; BLUEPRINT[56] (n=197 healthy blood donors); and CEDAR (Correlated Expression and Disease Association Research)[57] (n=322 healthy individuals from a cancer screening cohort). Gene expression datasets were accessed from the FUMA platform[58]: https://fuma.ctglab.nl.

Childhood Acute Lymphoblastic Leukemia Datasets

Genetic associations with childhood ALL were obtained from a meta-analysis of 2666 cases and 60,272 controls from two separate genome-wide scans (details in Supplementary Note). The first GWAS consisted of a pooled dataset of 1162 cases and 1229 controls from the California Cancer Records Linkage Project (CCRLP)[21] with 56,112 additional controls from the Kaiser Permanente Genetic Epidemiology Research on Ageing (GERA) cohort. Details of the CCRLP study and combined CCRLP/GERA GWAS have been previously described[21], with the present analysis including additional GERA controls and imputation using the HRC reference panel (version r1.1 2016). In brief, newborn dried bloodspots were obtained from the California Biobank Program for CCRLP childhood ALL cases and cancer-free controls, identified via linkage between the statewide birth records maintained by the California Department of Public Health (1982-2009) and cancer diagnosis data from the California Cancer Registry (for the years 1988-2011). All CCRLP and GERA participants were genotyped on the Affymetrix Axiom World array. The second ALL GWAS included 1504 ALL cases from the Children's Oncology Group (COG) and 2931 cancer-free controls from the Welcome Trust Case-Control Consortium (WTCCC), genotyped on either the Affymetrix Human SNP Array 6.0 (WTCCC, COG trials AALL0232 and P9904/9905)[25] or the Affymetrix GeneChip Human Mapping 500K Array (COG P9906 and St. Jude Total Therapy XIIIB/XV)[59]. GWAS meta-analysis was restricted to individuals of predominantly European ancestry.

Standard quality control steps were implemented, removing variants with $P_{HWE}<1\times10^{-5}$ in controls and imputation INFO<0.30. Additional filters were applied to minimize potential for bias due to the inclusion of external controls (Supplementary Figure 1). Variants associated with control group (CCRLP vs. GERA) at $P<1\times10^{-5}$ were removed (n=443). We also excluded variants if their MAF differed by >50% or ≥0.10 from the average MAF across CCRLP, GERA, and WTCCC controls (MAF≥0.05: n=3029; MAF<0.05: n=198,632). Lastly, allele frequencies in CCRLP/GERA and WTCCC controls were compared to the gnomAD non-Finnish European reference dataset and variants with absolute MAF differences ≥0.10 were filtered out (n=21,863).

Heritability and Genetic Correlation

LD Score regression [60] was used to estimate heritability (h_g) for each blood cell phenotype and for ALL, as well as the genetic correlation (r_g) between each blood cell phenotype and ALL. We used a reference panel of LD scores generated from all variants that passed QC with MAF>0.0001 using a random sample of 10,000 European ancestry UKB participants. UKB LD scores were used to estimate h_g for each blood cell trait phenotype and r_g with ALL.

Mendelian Randomization

Mendelian randomization (MR) analyses were carried out to investigate the potential causal relationship between blood cell trait variation and ALL. Genetic instruments excluded multi-allelic and non-inferable palindromic variants with intermediate allele frequencies (MAF>0.42). To minimize potential for bias due to differences in allele frequencies between exposure (UKB) and outcome (ALL) populations, analyses were restricted to variants with MAF≥0.01 and MAF difference<0.10. For instruments that were unavailable in the ALL dataset (n=294), LD proxies (r^2>0.95) were obtained. MR analyses estimated odds ratios (OR) and corresponding 95% confidence intervals (CI) for a genetically predicted 1-SD increase in the normalized z-

score for lymphocytes, monocytes, neutrophils, basophils, and eosinophils. For LMR, NLR, and PLR effects were estimated per 1-unit increase in the ratio.

Multiple MR estimators were used to strengthen inference by evaluating consistency in the observed effects. Maximum likelihood (ML) provides an unbiased estimate in the absence of horizontal pleiotropy, while inverse variance weighted multiplicative random-effects (IVW-mre) accounts for non-directional pleiotropy[61, 62]. Weighted median (WM)[63] provides unbiased estimates when up to 50% of the weights are from invalid instruments. Shrinkage-based MR RAPS (Robust Adjusted Profile Score)[30, 31] models additive pleiotropic effects and incorporates a robust loss function to limit the influence of invalid instruments. MR pleiotropy residual sum and outlier (PRESSO)[32] regresses variant effects on the outcome on their exposure effects and compares the observed distance of all instruments to this regression line with the expected distance under the null hypothesis of no horizontal pleiotropy[32].

To assess potential violations of MR assumptions we examined the following diagnostic tests: i) deviation of the MR Egger intercept from 0 ($p<0.05$) indicative of directional horizontal pleiotropy; ii) $I^2_{GX}<0.90$ indicative of regression dilution bias due to exposure measurement error which may inflate the MR Egger pleiotropy test[64]; iii) Cochran's Q-statistic $P_Q<0.05$ or MR PRESSO $P_{Global}<0.05$ indicative of heterogeneity due to balanced horizontal pleiotropy. We also report MR PRESSO distortion p-values, which test for significant differences between the original and pleiotropy-corrected effect estimates.

Next, we conducted multivariable MR (MVMR) analyses to estimate direct effects of specific blood cell traits on ALL, after accounting for related phenotypes. MVMR requires SNP effects for all instruments across all exposures so that they can be regressed against the outcome together, weighting for the inverse variance of the outcome (MVMR-IVW). We also applied a modified analysis where the instruments are selected for each exposure based on $P<5\times10^{-8}$ and then all exposures for those SNPs are regressed together (MVMR-IVW_{mod}). Feature selection was also performed using MV LASSO.

For ratios we conducted summary-based mediation analysis to decompose the observed total MR effects into direct and indirect effects mediated by each of the component traits[65]. For instance, for LMR we quantified indirect effects on ALL risk that were mediated through regulation of lymphocyte and monocyte counts, respectively, as well as direct LMR effects on ALL.

Lastly, we applied MR-Clust[33], a heterogeneity-based clustering method for detecting distinct values of the causal effect that are evidenced by multiple genetic variants. MR-Clust assigns variants to *K* substantive clusters where all variants indicate the same causal effect, a null cluster, and a "junk" cluster, which includes non-null variants which do not fit into any of the substantive clusters. This approach may reveal different causal or pleiotropic pathways and identify potentially novel ALL risk variants due to a reduced burden of multiple testing compared with GWAS. Variants were assigned to a cluster if their conditional probability of cluster membership was greater than 0.50. Clusters were formed with a minimum of 4 variants.

All statistical analyses were conducted using R (version 4.0.2). Mendelian randomization analyses were conducted using the TwoSampleMR R package (version 0.5.5).

DATA AVAILABILITY

This research was conducted with approved access to UK Biobank data under application number 14105 (PI: Witte) and in accordance with the UK Biobank Ethics and Governance Framework. UK Biobank data are publicly available by request from https://www.ukbiobank.ac.uk. Ethics approval for establishing the UK Biobank resource was obtained from the North West Centre for Research Ethics Committee (11/NW/0382).

This study included the analysis of data derived from biospecimens from the California Biobank Program (CCRLP study). Any uploading of genomic data and/or sharing of these biospecimens or individual data derived from these biospecimens has been determined to violate the statutory scheme of the California Health and Safety Code Sections 124980(j), 124991(b), (g), (h), and 103850 (a) and (d), which protect the confidential nature of biospecimens and individual data derived from biospecimens. This study was approved by Institutional Review Boards at the California Health and Human Services Agency, University of Southern California, Yale University, and the University of California San Francisco. The de-identified newborn dried blood spots for the CCRLP (California Biobank Program SIS request # 26) were obtained with a waiver of consent from the Committee for the Protection of Human Subjects of the State of California.

This study makes use of data from the Kaiser Permanente (KP) Research Program on Genes, Environment and Health (RPGEH) Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort, available from dbGaP (Study Accession: phs000788.v1.p2).

This study also makes use of data generated by the Wellcome Trust Case-Control Consortium available by request from the European Genotype Archive: http://www.ebi.ac.uk/ega (Study Accession: EGAD00000000021). Genotype data for COG ALL cases are available for download from dbGaP (Study Accession: phs000638.v1.p1).

ACKNOWLEDGEMENTS

This work was supported by research grants from the National Institutes of Health (NIH) National Cancer Institute (NCI): R03CA245998 (AJD and LK), K99CA246076 (LK), R01CA155461 (JLW, XM) and R01CA175737 (JLW, XM). The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries, under cooperative agreement 5NU58DP003862-04/DP003862; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors. A subset of the CCRLP data used in this study were obtained from the California Biobank Program at the California Department of Public Health (CDPH), SIS request number 26, in accordance with Section 6555(b),17CCR. The CDPH is not responsible for the results or conclusions drawn by the authors of this publication.

References

1. Evans DM, Frazer IH, Martin NG. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res* **2**, 250-257 (1999).

2. Garner C*, et al.* Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* **95**, 342-346 (2000).

3. Pilia G*, et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).

4. Astle WJ*, et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e1419 (2016).

5. Group CCHW. Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. *Nat Genet* **48**, 867-876 (2016).

6. Vuckovic D*, et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231 e1211 (2020).

7. Liggett LA, Sankaran VG. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* **182**, 1384-1400 (2020).

8. Bao EL*, et al.* Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* **586**, 769-775 (2020).

9. Linet MS, Ries LA, Smith MA, Tarone RE, Devesa SS. Cancer surveillance series: recent trends in childhood cancer incidence and mortality in the United States. *J Natl Cancer Inst* **91**, 1051-1058 (1999).

10. Hunger SP*, et al.* Improved survival for children and adolescents with acute lymphoblastic leukemia between 1990 and 2005: a report from the children's oncology group. *J Clin Oncol* **30**, 1663-1669 (2012).

11. Curtin SC, Minino AM, Anderson RN. Declines in Cancer Death Rates Among Children and Adolescents in the United States, 1999-2014. *NCHS Data Brief*, 1-8 (2016).

12. Turcotte LM*, et al.* Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. *JAMA* **317**, 814-824 (2017).

13. Mulrooney DA*, et al.* The changing burden of long-term health outcomes in survivors of childhood acute lymphoblastic leukaemia: a retrospective analysis of the St Jude Lifetime Cohort Study. *Lancet Haematol* **6**, e306-e316 (2019).

14. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. *Nat Rev Cancer* **18**, 471-484 (2018).

15. Williams LA, Yang JJ, Hirsch BA, Marcotte EL, Spector LG. Is There Etiologic Heterogeneity between Subtypes of Childhood Acute Lymphoblastic Leukemia? A Review of Variation in Risk by Subtype. *Cancer Epidemiol Biomarkers Prev* **28**, 846-856 (2019).

16. Gocho Y, Yang JJ. Genetic defects in hematopoietic transcription factors and predisposition to acute lymphoblastic leukemia. *Blood* **134**, 793-797 (2019).

17. Papaemmanuil E*, et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1006-1010 (2009).

18. Trevino LR*, et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1001-1005 (2009).

19. Migliorini G*, et al.* Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood* **122**, 3298-3307 (2013).

20. Xu H*, et al.* Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. *J Natl Cancer Inst* **105**, 733-742 (2013).

21. Wiemels JL*, et al.* GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. *Nat Commun* **9**, 286 (2018).

22. de Smith AJ*, et al.* BMI1 enhancer polymorphism underlies chromosome 10p12.31 association with childhood acute lymphoblastic leukemia. *Int J Cancer* **143**, 2647-2658 (2018).

23. de Smith AJ*, et al.* Heritable variation at the chromosome 21 gene ERG is associated with acute lymphoblastic leukemia risk in children with and without Down syndrome. *Leukemia* **33**, 2746-2751 (2019).

24. Qian M*, et al.* Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. *Blood* **133**, 724-729 (2019).

25. Vijayakrishnan J*, et al.* Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk. *Nat Commun* **10**, 5348 (2019).

26. Semmes EC, Vijayakrishnan J, Zhang C, Hurst JH, Houlston RS, Walsh KM. Leveraging Genome and Phenome-Wide Association Studies to Investigate Genetic Risk of Acute Lymphoblastic Leukemia. *Cancer Epidemiol Biomarkers Prev* **29**, 1606-1614 (2020).

27. Bycroft C*, et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

28. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).

29. Võsa U*, et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*, 447367 (2018).

30. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann Statist* **48**, 1742-1769 (2020).

31. Zhao Q, Chen Y, Wang J, Small DS. Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *Int J Epidemiol* **48**, 1478-1492 (2019).

32. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50**, 693-698 (2018).

33. Foley CN, Mason AM, Kirk PDW, Burgess S. MR-Clust: Clustering of genetic variants in Mendelian randomization with similar causal estimates. *Bioinformatics*, (2020).

34. Kachuri L*, et al.* Mendelian Randomization and mediation analysis of leukocyte telomere length and risk of lung and head and neck cancers. *Int J Epidemiol* **48**, 751-766 (2019).

35. Wiemels JL*, et al.* Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet* **354**, 1499-1503 (1999).

36. Rudant J*, et al.* Childhood acute lymphoblastic leukemia and indicators of early immune stimulation: a Childhood Leukemia International Consortium study. *Am J Epidemiol* **181**, 549-562 (2015).

37. Urayama KY*, et al.* Early life exposure to infections and risk of childhood acute lymphoblastic leukemia. *Int J Cancer* **128**, 1632-1643 (2011).

38. Cobaleda C, Vicente-Duenas C, Sanchez-Garcia I. Infectious triggers and novel therapeutic opportunities in childhood B cell leukaemia. *Nat Rev Immunol*, (2021).

39. Martin-Lorenzo A*, et al.* Infection Exposure is a Causal Factor in B-cell Precursor Acute Lymphoblastic Leukemia as a Result of Pax5-Inherited Susceptibility. *Cancer Discov* **5**, 1328-1343 (2015).

40. Chang JS, Zhou M, Buffler PA, Chokkalingam AP, Metayer C, Wiemels JL. Profound deficit of IL10 at birth in children who develop childhood acute lymphoblastic leukemia. *Cancer Epidemiol Biomarkers Prev* **20**, 1736-1740 (2011).

41. Soegaard SH, Rostgaard K, Skogstrand K, Wiemels JL, Schmiegelow K, Hjalgrim H. Neonatal Inflammatory Markers Are Associated with Childhood B-cell Precursor Acute Lymphoblastic Leukemia. *Cancer Res* **78**, 5458-5463 (2018).

42. Gasparyan AY, Ayvazyan L, Mukanova U, Yessirkepov M, Kitas GD. The Platelet-to-Lymphocyte Ratio as an Inflammatory Marker in Rheumatic Diseases. *Ann Lab Med* **39**, 345-357 (2019).

43. Saevarsdottir S*, et al.* FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature* **584**, 619-623 (2020).

44. Annesley CE, Brown P. The Biology and Targeting of FLT3 in Pediatric Leukemia. *Front Oncol* **4**, 263 (2014).

45. Kazi JU, Ronnstrand L. FMS-like Tyrosine Kinase 3/FLT3: From Basic Science to Clinical Implications. *Physiol Rev* **99**, 1433-1466 (2019).

46. Liu M*, et al.* MCM6 promotes metastasis of hepatocellular carcinoma via MEK/ERK pathway and serves as a novel serum biomarker for early recurrence. *J Exp Clin Cancer Res* **37**, 10 (2018).

47. Katsura M*, et al.* Correlation between CXCR4/CXCR7/CXCL12 chemokine axis expression and prognosis in lymph-node-positive lung cancer patients. *Cancer Sci* **109**, 154-165 (2018).

48. Cancilla D, Rettig MP, DiPersio JF. Targeting CXCR4 in AML and ALL. *Front Oncol* **10**, 1672 (2020).

49. Khera AV*, et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596 e589 (2019).

50. Turley P*, et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229-237 (2018).

51. Lin BD*, et al.* 2SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio. *J Hum Genet* **62**, 979-988 (2017).

52. Kachuri L*, et al.* Immune-mediated genetic pathways resulting in pulmonary function impairment increase lung cancer susceptibility. *Nat Commun* **11**, 27 (2020).

53. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

54. Dong S, Boyle AP. Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum Mutat* **40**, 1292-1298 (2019).

55. Schmiedel BJ*, et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e1716 (2018).

56. Chen L*, et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e1324 (2016).

57. Momozawa Y*, et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun* **9**, 2427 (2018).

58. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

59. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

60. Bulik-Sullivan BK*, et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).

61. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658-665 (2013).

62. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* **36**, 1783-1802 (2017).

63. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304-314 (2016).

64. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol* **45**, 1961-1974 (2016).

65. Burgess S, Thompson DJ, Rees JMB, Day FR, Perry JR, Ong KK. Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. *Genetics* **207**, 481-487 (2017).

Figure 1: Array-based heritability (h_g) estimates for acute lymphoblastic leukemia (ALL) and blood cell subtypes: lymphocytes, monocytes, neutrophils, basophils, eosinophils, basophils, platelets, lymphocyte to monocyte ratio (LMR), neutrophil to lymphocyte ratio (NLR), and platelet to lymphocyte ratio (PLR) estimated using LD score regression.

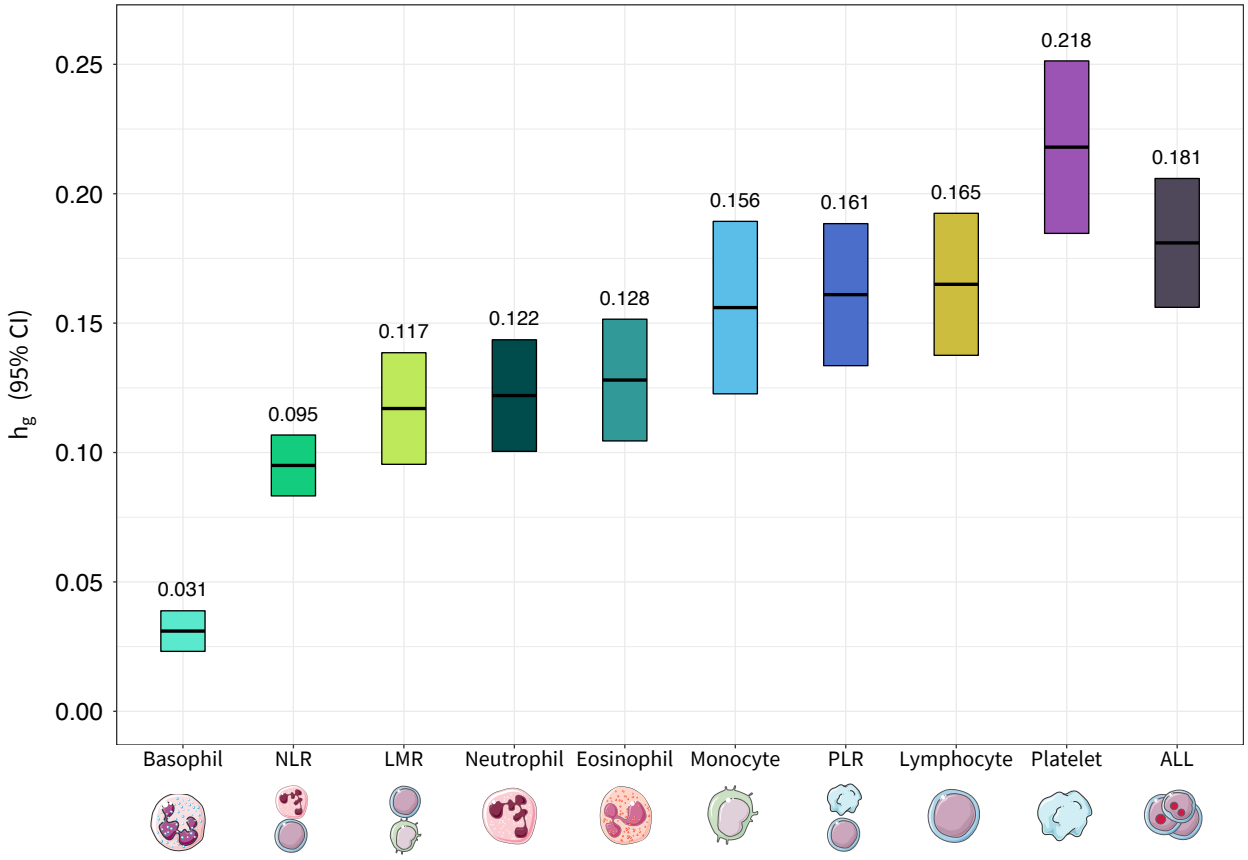


Figure 2: Genetic correlation (r_g) estimates between blood cell subtypes: lymphocytes, monocytes, neutrophils, basophils, eosinophils, basophils, platelets, lymphocyte to monocyte ratio (LMR), neutrophil to lymphocyte ratio (NLR), and platelet to lymphocyte ratio (PLR) estimated using LD score regression. Associations with $P<1.4\times10^{-3}$ were considered statistically significant after Bonferroni correction for 36 pairs tested, with corresponding r_g estimates labeled in black font.

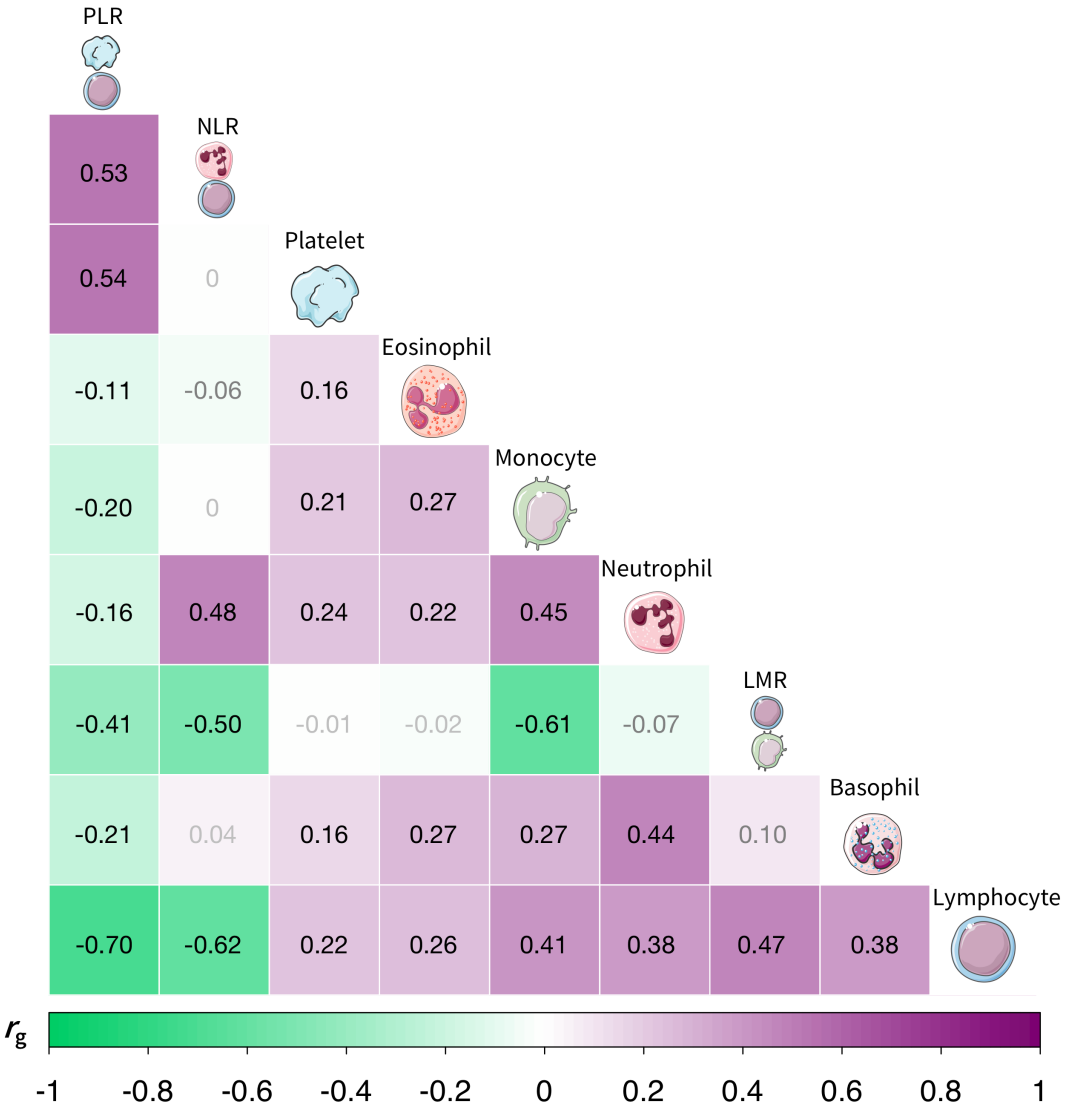


Figure 3: Genome-wide significant (P<5×10^{-8}) associations for cell type ratios and their component count traits. Points with black borders denote variants that were selected only as instruments for the given ratio trait and are not in linkage disequilibrium (r^2<0.05 within 10Mb) with previously reported loci for its component cell type counts or proportions. Labeled genes contain variants with specific functional features (CADD score >10, Regulome DB rank 1-3a, missense mutations) and/or P<1×10^{-20}.

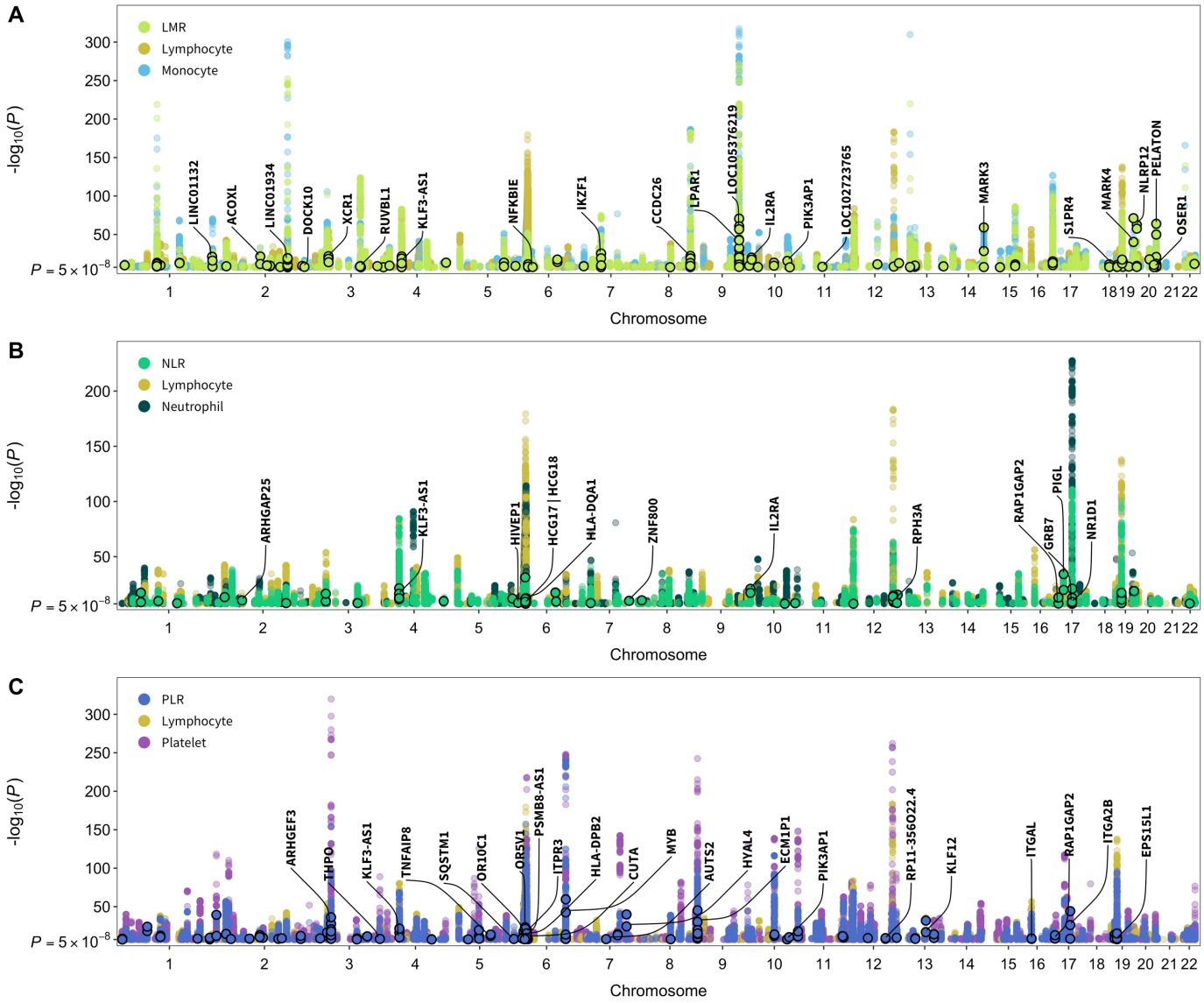
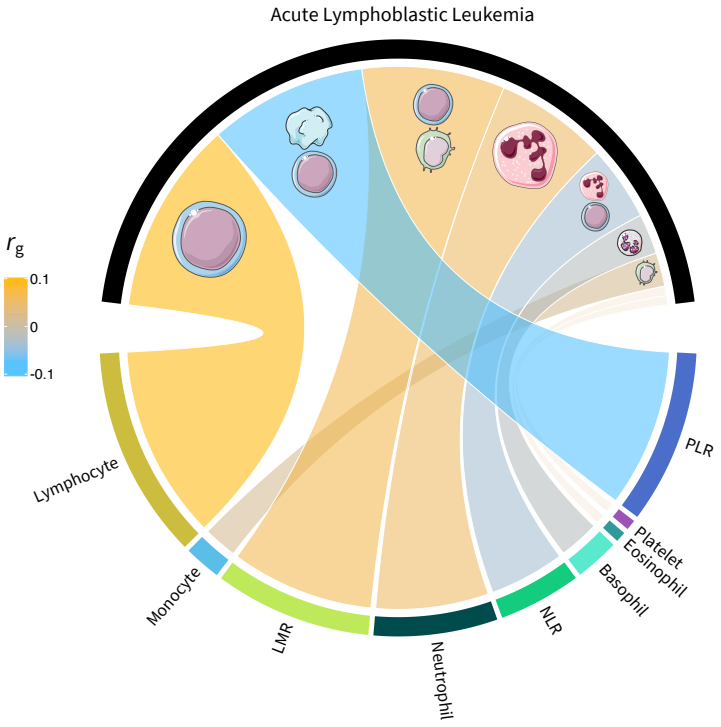
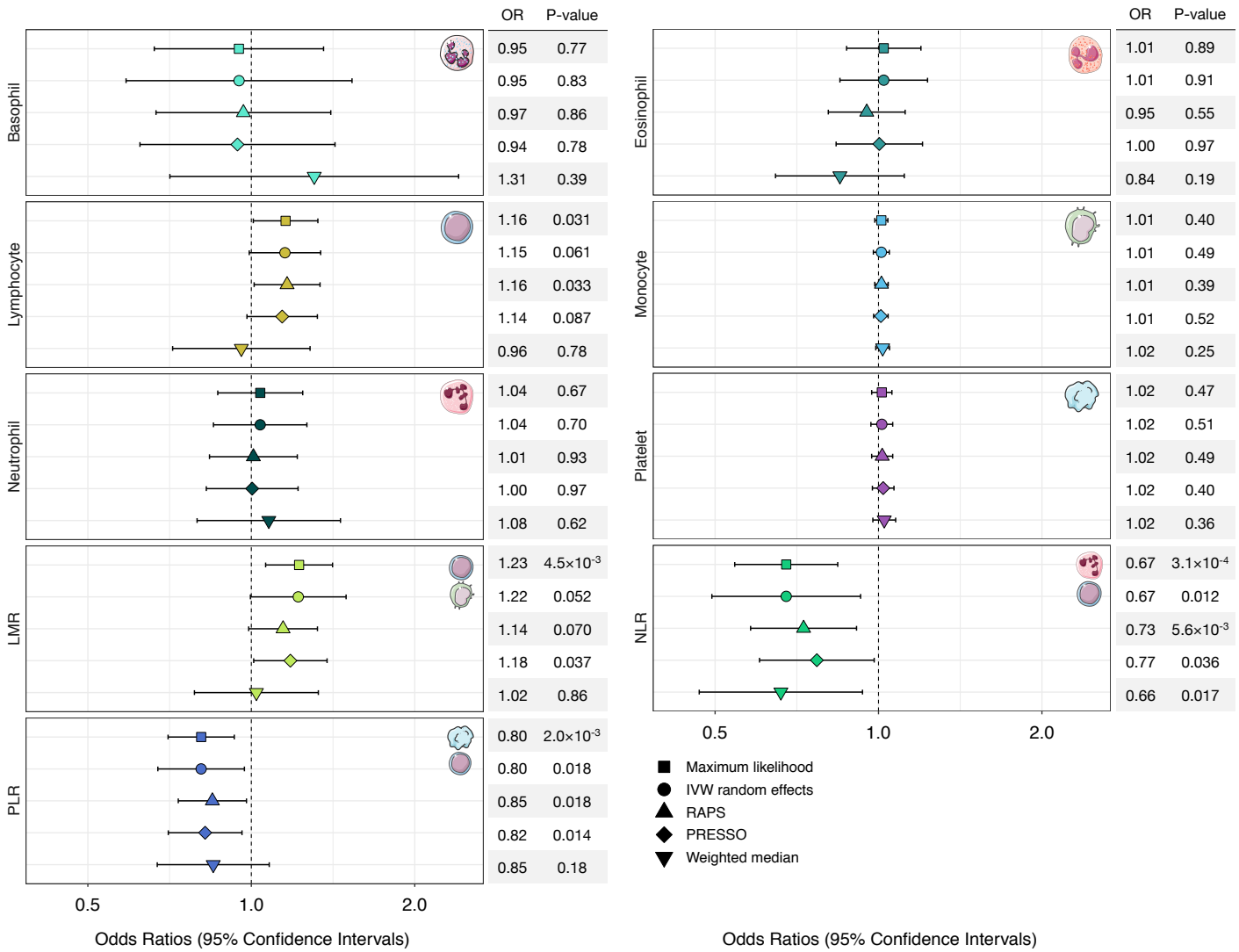


Figure 4: Genetic correlation (r_g) between blood cell subtypes and acute lymphoblastic leukemia (ALL) based on genome-wide summary statistics. The colors in the circos plot correspond to the direction of genetic correlation, with warm shades depicting positive correlations between increasing blood cell counts or ratios and ALL risk, cool tones corresponding to inverse associations, and faded gray shades corresponding to null correlations. The width of each band in the circos plot is proportional to the magnitude of the absolute value of the r_g estimate.

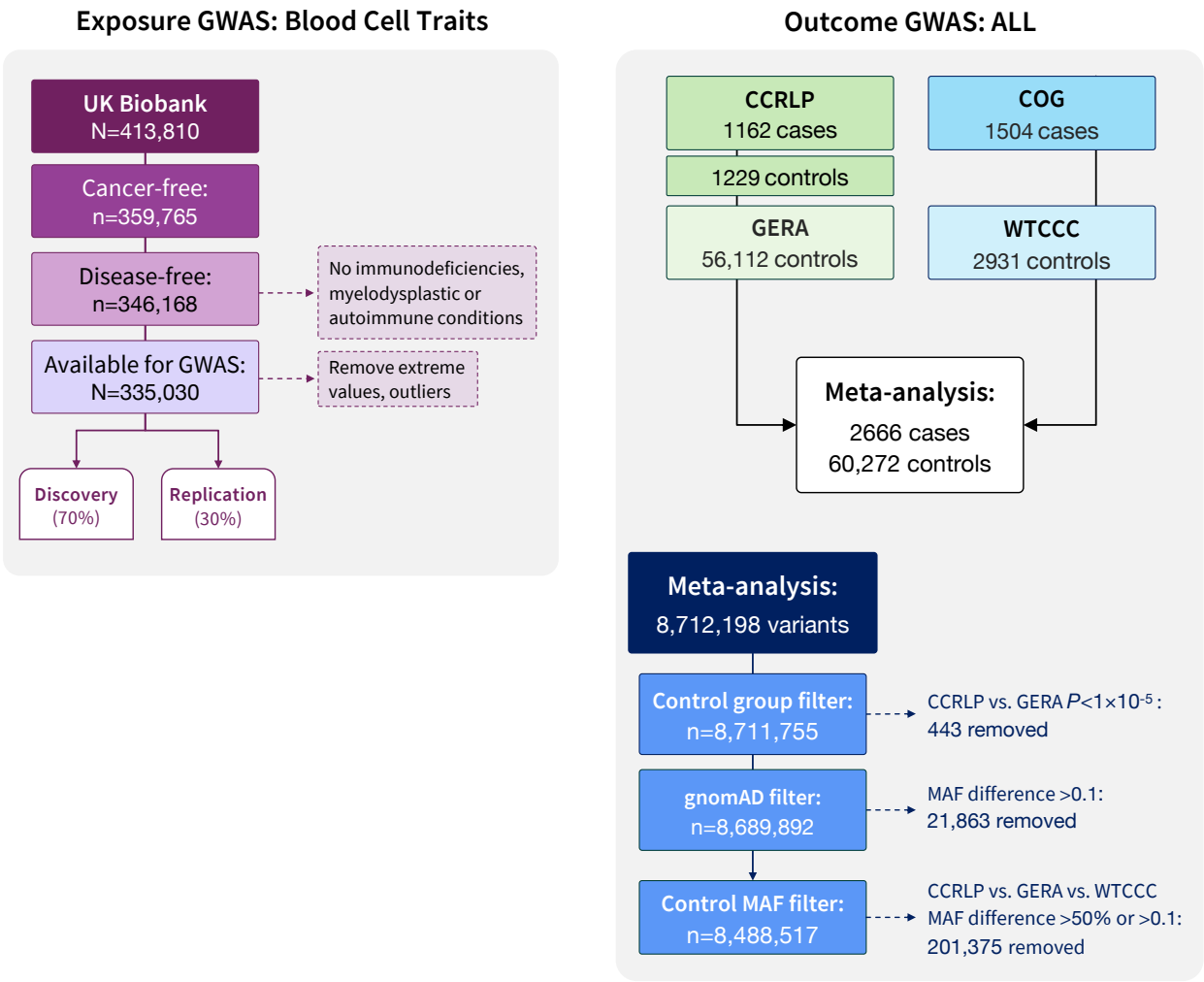


Phenotype	r_g	(SE)	P
Lymphocyte	0.088	(0.025)	4.0×10^{-4}
PLR	-0.072	(0.023)	1.7×10^{-3}
LMR	0.065	(0.026)	0.012
Neutrophil	0.051	(0.023)	0.027
NLR	-0.034	(0.026)	0.19
Basophil	-0.018	(0.031)	0.56
Monocyte	0.015	(0.028)	0.58
Eosinophil	0.004	(0.028)	0.87
Platelet	0.004	(0.025)	0.87

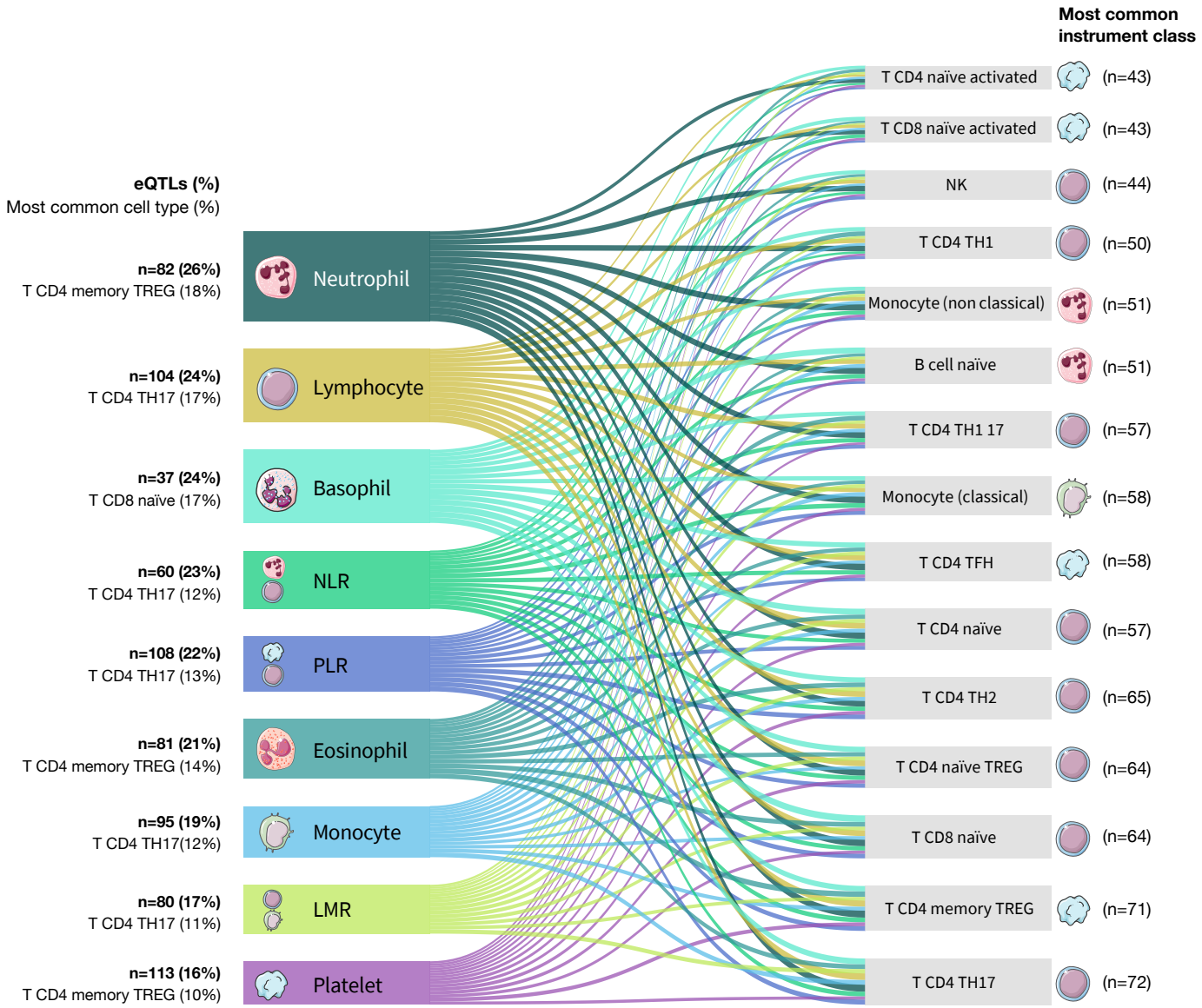
Figure 5: Odds ratios (OR) and 95% confidence intervals (CI) for the effect of increasing blood cell counts or cell type ratios on the risk of acute lymphoblastic leukemia (ALL), estimated using Mendelian Randomization (MR). Association results based on five different MR estimators are shown.



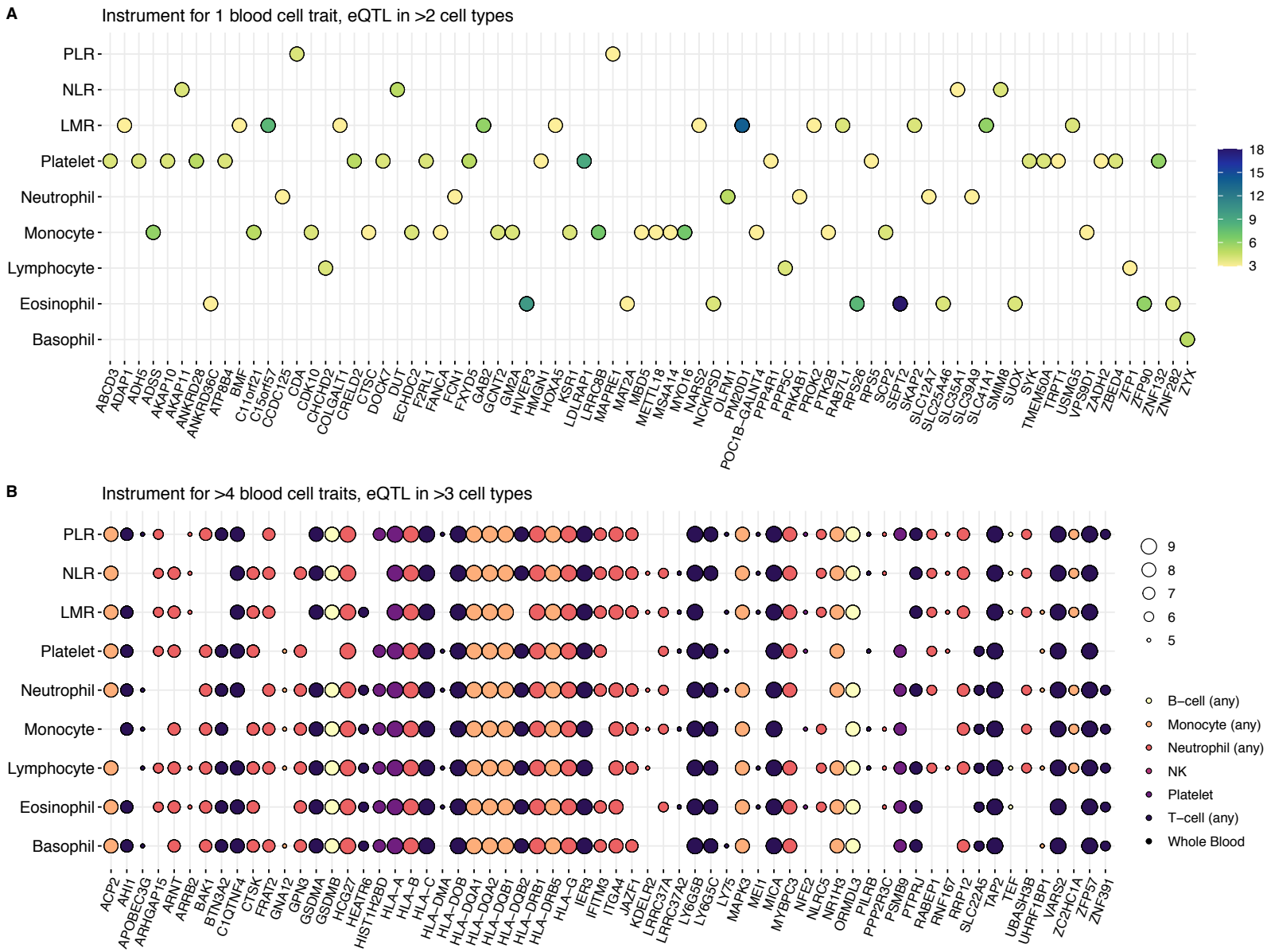
Supplementary Figure 1: Flowchart outlining the study populations and additional quality control steps. Participants were restricted to individuals of predominantly European ancestry.



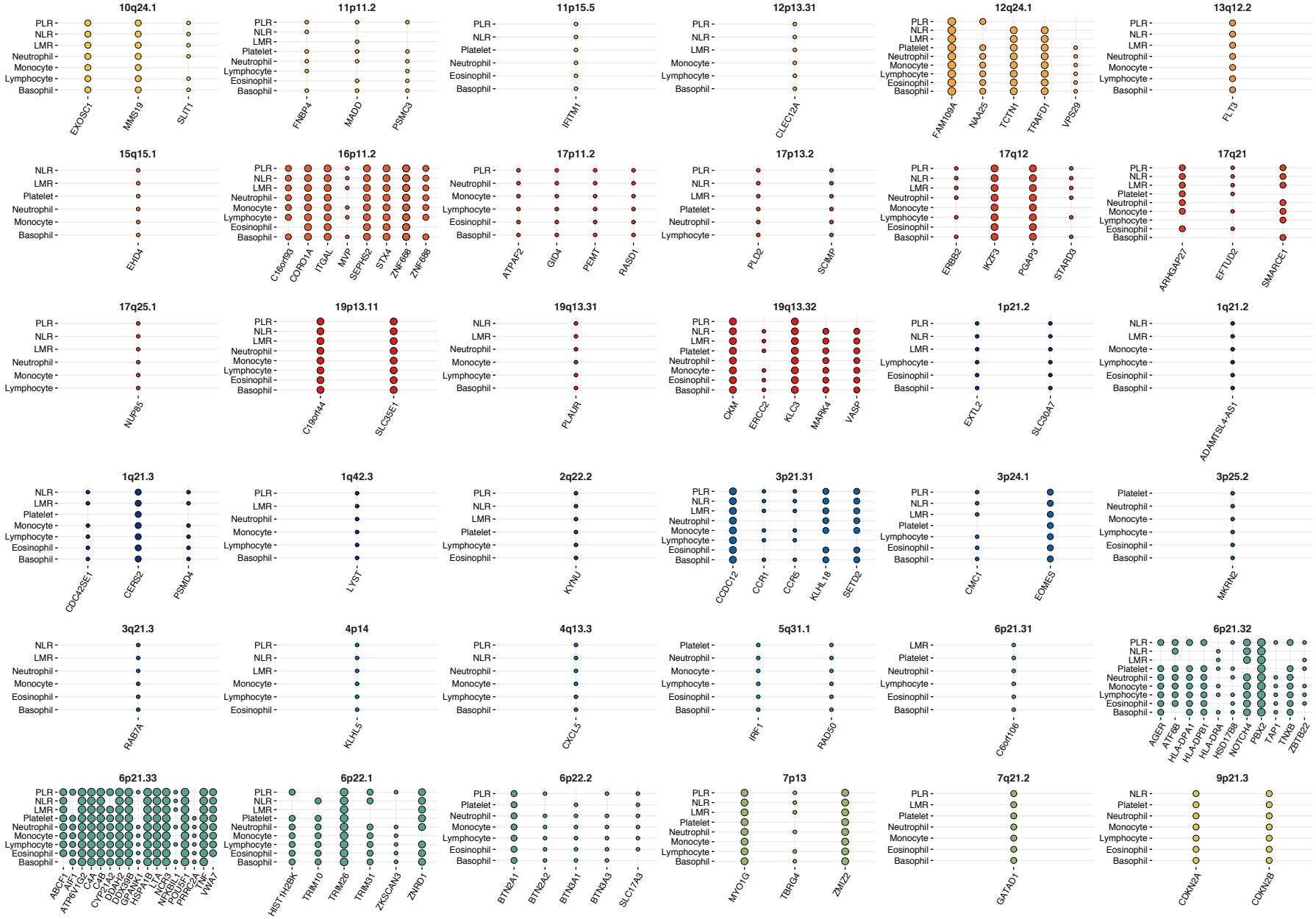
Supplementary Figure 2: Sankey style plot illustrating the pattern of cell-type specific effects on gene expression obtained from DICE (Database of Immune Cell Expression) for blood cell trait genetic instruments. Genetic instruments were considered expression quantitative trait loci (eQTL) if their associations with gene expression were significant at FDR<0.05. The color of each band corresponds to the blood cell type and the width of each band is scaled to reflect the proportion of all eQTL signals accounted for by a specific cell type.



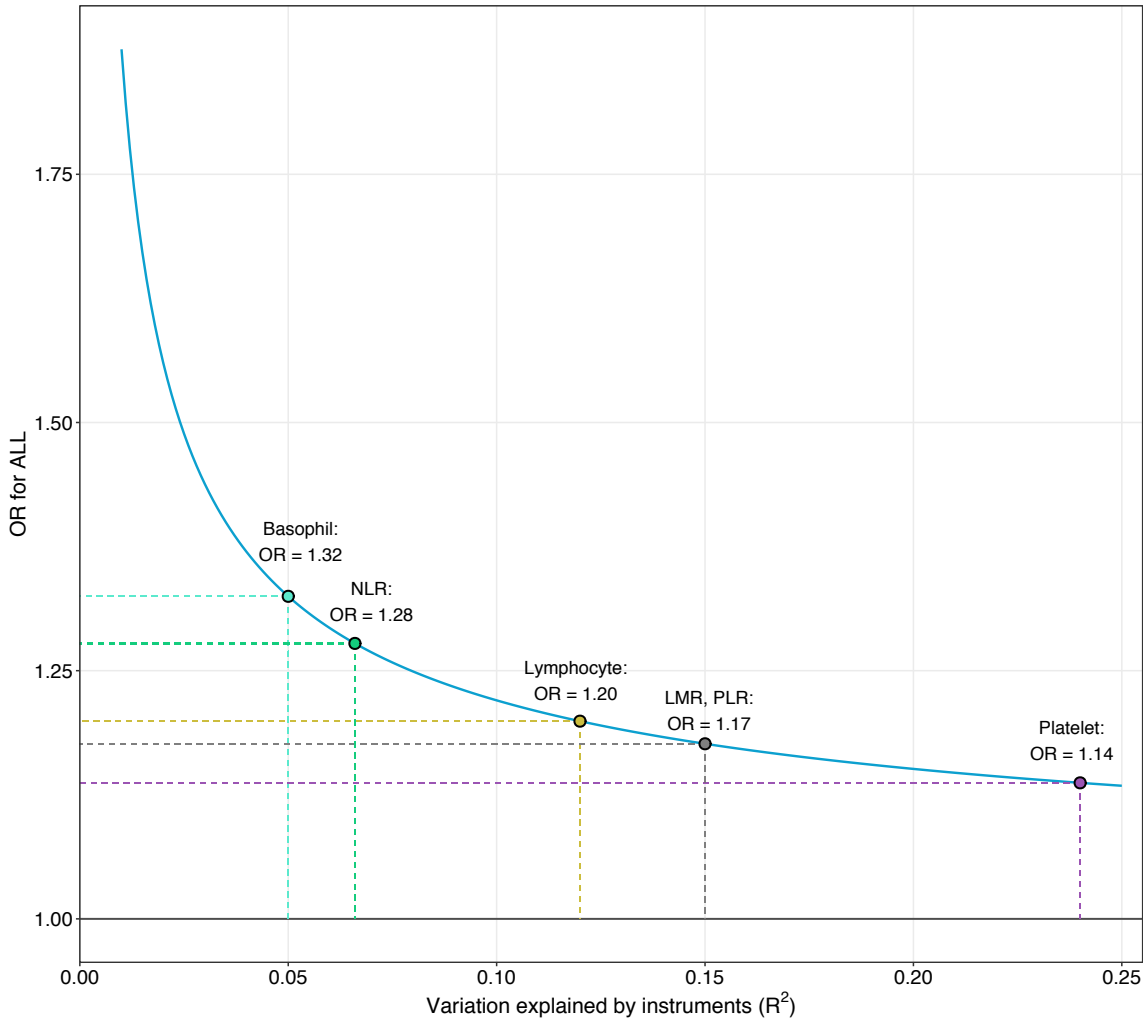
Supplementary Figure 3: Genes with expression trait loci (eQTL) in multiple cell types found among instruments for blood cell traits



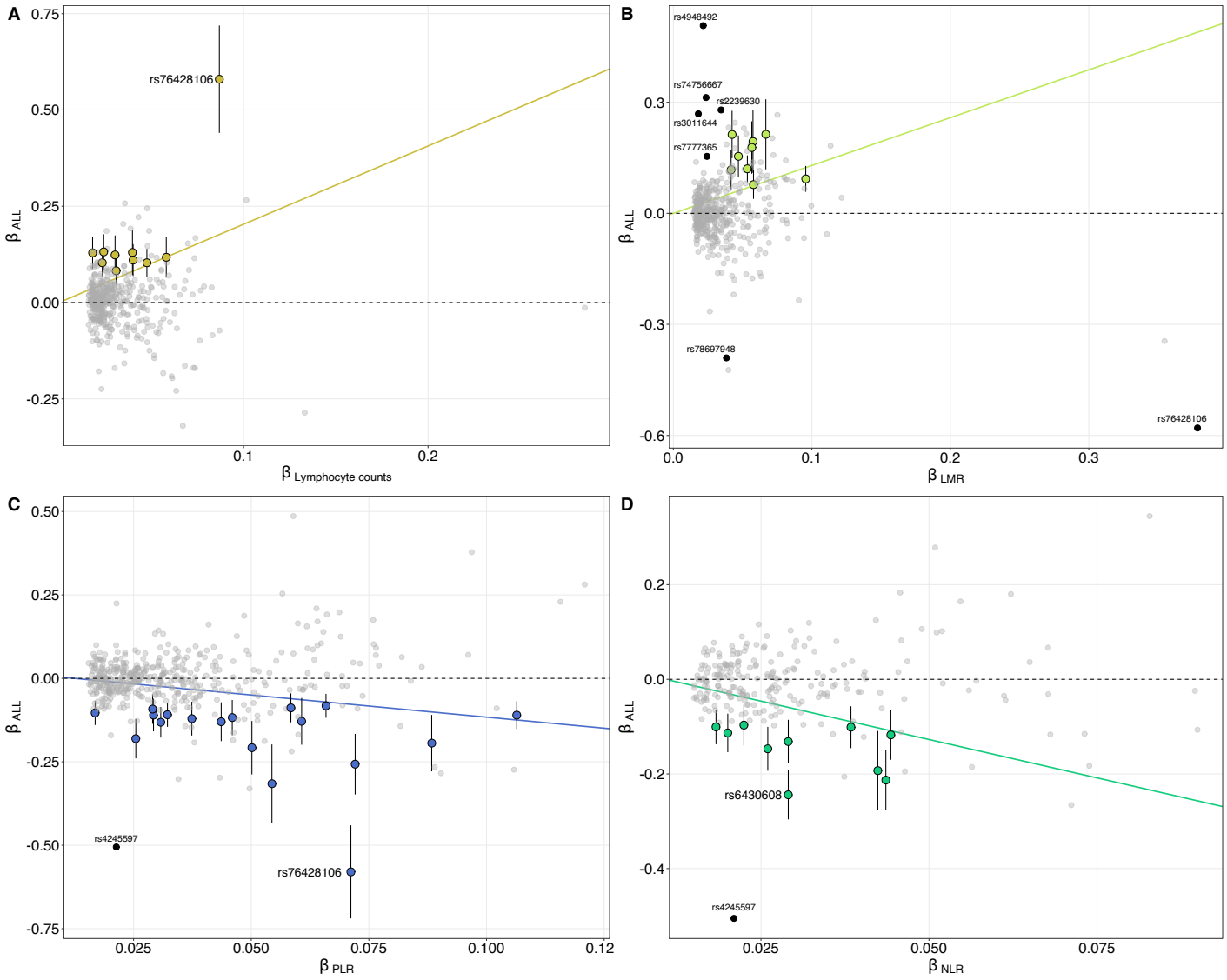
Supplementary Figure 4: Genes with expression trait loci (eQTL) in one cell type found among genetic instruments for >5 blood cell traits



Supplementary Figure 5: Minimum detectable odds ratio at 80% power for each blood cell trait based on the observed instrument strength and available sample size for the outcome



Supplementary Figure 6: Visualization of MR Clust results identifying subgroups of variants with homogenous causal effects and novel ALL risk variants. The slope of the line corresponds to the mean causal effect within each cluster. Substantive clusters included variants with an assignment probability of greater than 50%. Outliers are denoted by solid black circles. Only variants significantly associated with ALL at the Bonferroni-corrected threshold (P_{ALL}<5×10^{-5}) are labeled.



Supplementary Table 1: Array-based heritability (hg) estimates for acute lymphoblastic leukemia (ALL) and blood cell subtypes: lymphocytes, monocytes, neutrophils, basophils, eosinophils, basophils, platelets, lymphocyte to monocyte ratio (LMR), neutrophil to lymphocyte ratio (NLR), and platelet to lymphocyte ratio (PLR) estimated with LD score regression using UKB LD scores from European ancestry participants as the reference panel (7,166,343 variants).

Phenotype		h_g	(SE)
White blood cells		0.141	(0.009)
Lymphocytes		0.165	(0.014)
Monocytes		0.156	(0.017)
Neutrophils		0.122	(0.011)
Eosinophils		0.128	(0.012)
Basophils		0.031	(0.004)
Platelets		0.218	(0.017)
NLR		0.095	(0.006)
LMR		0.117	(0.011)
PLR		0.161	(0.014)
ALL	Lifetime risk: 0.15%	0.181	(0.013)
	Lifetime risk: 0.10%	0.169	(0.012)

Note: for ALL heritability was transformed to the liability scale using SEER estimates of lifetime for non-Hispanic whites (0.15%), with a sensitivity analysis using a lower lifetime risk estimate.

Supplementary Table 2: Genetic correlation (r_g) matrix for blood cell subtypes: lymphocytes, monocytes, neutrophils, basophils, eosinophils, basophils, platelets, lymphocyte to monocyte ratio (LMR), neutrophil to lymphocyte ratio (NLR), and platelet to lymphocyte ratio (PLR) estimated using LD score regression. Correlations with $p<1.4\times10^{-3}$ were considered statistically significant after Bonferroni correction for 36 pairs tested.

	Lymphocytes	Monocytes	Neutrophils	Eosinophils	Basophils	Platelets	LMR	NLR	PLR
Lymphocytes	1.0	0.405 $p = 1.26\times10^{-8}$	0.384 $p = 1.40\times10^{-51}$	0.255 $p = 5.47\times10^{-21}$	0.378 $p = 1.02\times10^{-26}$	0.224 $p = 3.83\times10^{-35}$	0.468 $p = 2.77\times10^{-48}$	-0.616 $p = 1.34\times10^{-145}$	-0.702 $p = 2.77\times10^{-393}$
Monocytes	0.405 $p = 1.26\times10^{-8}$	1.0	0.449 $p = 1.77\times10^{-83}$	0.265 $p = 1.84\times10^{-26}$	0.267 $p = 8.28\times10^{-17}$	0.213 $p = 6.81\times10^{-26}$	-0.612 $p = 1.88\times10^{-74}$	-0.004 p = 0.89	-0.196 $p = 2.45\times10^{-20}$
Neutrophils	0.384 $p = 1.40\times10^{-51}$	0.449 $p = 1.77\times10^{-83}$	1.0	0.219 $p = 1.52\times10^{-16}$	0.442 $p = 4.00\times10^{-33}$	0.236 $p = 1.65\times10^{-26}$	-0.069 p = 0.017	0.480 $p = 5.89\times10^{-64}$	-0.158 $p = 2.74\times10^{-8}$
Eosinophils	0.255 $p = 5.47\times10^{-21}$	0.265 $p = 1.84\times10^{-26}$	0.219 $p = 1.52\times10^{-16}$	1.0	0.268 $p = 2.43\times10^{-17}$	0.155 $p = 2.24\times10^{-9}$	-0.022 p = 0.32	-0.061 p = 0.017	-0.112 $p = 6.29\times10^{-7}$
Basophils	0.378 $p = 1.02\times10^{-26}$	0.267 $p = 8.28\times10^{-17}$	0.442 $p = 4.00\times10^{-33}$	0.268 $p = 2.43\times10^{-17}$	1.0	0.158 $p = 2.31\times10^{-8}$	0.104 $p = 1.9\times10^{-3}$	0.037 p=0.40	-0.208 $p = 1.12\times10^{-10}$
Platelets	0.224 $p = 3.83\times10^{-35}$	0.213 $p = 6.81\times10^{-26}$	0.236 $p = 1.65\times10^{-26}$	0.155 $p = 2.24\times10^{-9}$	0.158 $p = 2.31\times10^{-8}$	1.0	-0.005 p = 0.74	-0.004 p = 0.82	0.535 $p = 3.08\times10^{-123}$
LMR	0.468 $p = 2.77\times10^{-48}$	-0.612 $p = 1.88\times10^{-74}$	-0.069 p = 0.017	-0.022 p = 0.32	0.104 $P = 1.9\times10^{-3}$	-0.005 p = 0.74	1.0	-0.496 $p = 1.37\times10^{-63}$	-0.408 $p = 9.35\times10^{-39}$
NLR	-0.616 $p = 1.34\times10^{-145}$	-0.004 p = 0.89	0.480 $p = 5.89\times10^{-64}$	-0.061 p = 0.017	0.037 p = 0.40	-0.004 p = 0.82	-0.496 $p = 1.37\times10^{-63}$	1.0	0.532 $p = 4.68\times10^{-80}$
PLR	-0.702 $p = 2.77\times10^{-393}$	-0.196 $p = 2.45\times10^{-20}$	-0.158 $p = 2.74\times10^{-8}$	-0.112 $p = 6.29\times10^{-7}$	-0.208 $p = 1.12\times10^{-10}$	0.535 $p = 3.08\times10^{-123}$	-0.408 $p = 9.35\times10^{-39}$	0.532 $p = 4.68\times10^{-80}$	1.0

Supplementary Table 3: Overview of the properties of genetic instruments for blood cell traits. Proportion of trait variation explained was estimated in the independent UK Biobank replication sample of over 100,000 individuals. Instruments applied includes available variants and proxies (LD r^2>0.95) in the outcome GWAS meta-analysis of acute lymphoblastic leukemia.

Trait	Instruments Available		Instruments Applied		
	N	Variation (%)	N	Variation (%)	Mean F
Lymphocytes	429	11.9	406	10.7	71.1
Monocytes	505	16.9	477	16.3	85.6
Neutrophils	313	9.4	301	8.7	69.4
Eosinophils	387	11.1	363	10.4	71.8
Basophils	157	5.1	144	4.5	64.6
Platelets	692	24.4	661	23.4	85.8
NLR	266	6.6	246	6.0	63.8
LMR	464	15.0	432	14.4	83.8
PLR	489	15.1	462	14.0	76.5

Supplementary Table 4: Summary of expression quantitative trait loci (eQTL) identified among genetic instruments for blood cell traits.

Cell Group	eQTL		eQTL (P<5×10^{-8})			Cell Type	eQTL		eQTL (P<5×10^{-8})			Study and Sample Size	
	N_{SNP}	(%)	N_{SNP}	%	Genes		N_{SNP}	%	N_{SNP}	%	Genes		
Whole Blood	2405	(80.2)	2253	(75.1)	2981	Whole Blood	2405	(80.2)	2253	(75.1)	2981	Vosa bioRxiv: 447367	31684
Monocytes (any)	810	(27.0)	514	(17.1)	431	Monocyte	559	(18.6)	370	(12.3)	328	Chen PMID: 27863251	197
						Monocyte (CD14 classical)	376	(12.5)	231	(7.7)	148	Momozawa PMID: 29930244	322
						Monocyte (CD14 classical)	228	(7.6)	83	(2.8)	68	Schmiedel PMID: 30449622	91
						Monocyte (non-classical)	206	(6.9)	64	(2.1)	58	Schmiedel PMID: 30449622	91
Neutrophils (any)	633	(21.1)	419	(14)	341	Neutrophil	458	(15.3)	289	(9.6)	274	Chen PMID: 27863251	197
						Neutrophil (CD15)	391	(13)	253	(8.4)	168	Momozawa PMID: 29930244	322
T-cells (any)	708	(23.6)	362	(12.1)	327	T-cell	430	(14.3)	277	(9.2)	233	Chen PMID: 27863251	197
						T CD8 (naïve)	221	(7.4)	87	(2.9)	84	Schmiedel PMID: 30449622	91
						T CD4 TH17	242	(8.1)	78	(2.6)	75	Schmiedel PMID: 30449622	91
						T_CD4 (memory TREG)	230	(7.7)	79	(2.6)	73	Schmiedel PMID: 30449622	91
						T_CD4 (naïve TREG)	228	(7.6)	77	(2.6)	79	Schmiedel PMID: 30449622	91
						T_CD4_TH2	222	(7.4)	74	(2.5)	69	Schmiedel PMID: 30449622	91
						T_CD4 (naïve)	223	(7.4)	69	(2.3)	71	Schmiedel PMID: 30449622	91
						T_CD4_TFH	216	(7.2)	66	(2.2)	65	Schmiedel PMID: 30449622	91
						T_CD4_TH1_17	207	(6.9)	64	(2.1)	61	Schmiedel PMID: 30449622	91
						T_CD4_TH1	182	(6.1)	51	(1.7)	45	Schmiedel PMID: 30449622	91
						T_CD8 (naïve activated)	148	(4.9)	48	(1.6)	39	Schmiedel PMID: 30449622	91
						T_CD4 (naïve activated)	138	(4.6)	45	(1.5)	42	Schmiedel PMID: 30449622	91
B-cells (any)	341	(11.4)	166	(5.5)	119	B-cell (CD19)	233	(7.8)	125	(4.2)	69	Momozawa PMID: 29930244	322
						B-cell (naïve)	185	(6.2)	73	(2.4)	62	Schmiedel PMID: 30449622	91
Platelets	119	(4.0)	72	(2.4)	51	Platelets	119	(4.0)	72	(2.4)	51	Momozawa PMID: 29930244	322
NK cells	154	(5.1)	48	(1.6)	44	NK cells	154	(5.1)	48	(1.6)	44	Schmiedel PMID: 30449622	91

Supplementary Table 5: Odds ratios (OR) and 95% confidence intervals (CI) for the effect of increasing blood cell counts or cell type ratios on the risk of acute lymphoblastic leukemia (ALL), estimated using Mendelian Randomization (MR).

Trait	MR Estimator	N_{SNP}	Association Results		
			OR	(95% CI)	P
Basophils	IVW (multiplicative random effects)	144	0.95	(0.59 – 1.53)	0.83
	Maximum likelihood		0.95	(0.66 – 1.36)	0.77
	Weighted median		1.31	(0.71 – 2.39)	0.39
	PRESSO (corrected)		0.94	(0.62 – 1.43)	0.78
	RAPS shrinkage		0.97	(0.67 – 1.41)	0.86
Eosinophils	IVW (multiplicative random effects)	363	1.01	(0.84 – 1.22)	0.91
	Maximum likelihood		1.01	(0.86 – 1.18)	0.89
	Weighted median		0.84	(0.64 – 1.09)	0.19
	PRESSO (corrected)		1.00	(0.84 – 1.21)	0.97
	RAPS shrinkage		0.95	(0.81 – 1.12)	0.55
Lymphocytes	IVW (multiplicative random effects)	406	1.15	(0.99 – 1.34)	0.061
	Maximum likelihood		1.16	(1.01 – 1.33)	0.035
	Weighted median		0.96	(0.71 – 1.29)	0.78
	PRESSO (corrected)		1.14	(0.98 – 1.32)	0.087
	RAPS shrinkage		1.16	(1.01 – 1.34)	0.033
Monocytes	IVW (multiplicative random effects)	477	1.01	(0.98 – 1.05)	0.49
	Maximum likelihood		1.01	(0.98 – 1.04)	0.40
	Weighted median		1.02	(0.99 – 1.05)	0.25
	PRESSO (corrected)		1.01	(0.98 – 1.04)	0.52
	RAPS shrinkage		1.01	(0.98 – 1.04)	0.39
Neutrophils	IVW (multiplicative random effects)	301	1.04	(0.85 – 1.27)	0.70
	Maximum likelihood		1.04	(0.87 – 1.25)	0.67
	Weighted median		1.08	(0.80 – 1.46)	0.62
	PRESSO (corrected)		1.00	(0.83 – 1.22)	0.97
	RAPS shrinkage		1.01	(0.84 – 1.22)	0.93
Platelets	IVW (multiplicative random effects)	661	1.02	(0.97 – 1.06)	0.51
	Maximum likelihood		1.02	(0.97 – 1.06)	0.47
	Weighted median		1.02	(0.97 – 1.08)	0.36
	PRESSO (corrected)		1.02	(0.97 – 1.07)	0.40
	RAPS shrinkage		1.02	(0.97 – 1.06)	0.49
LMR	IVW (multiplicative random effects)	432	1.22	(1.00 – 1.50)	0.052
	Maximum likelihood		1.23	(1.07 – 1.41)	4.5×10^{-3}
	Weighted median		1.02	(0.80 – 1.31)	0.86
	PRESSO (corrected)		1.18	(1.01 – 1.38)	0.037
	RAPS shrinkage		1.14	(0.99 – 1.32)	0.070

	IVW (multiplicative random effects)		0.67	(0.49 – 0.92)	0.012
	Maximum likelihood		0.67	(0.54 – 0.83)	3.1×10^{-4}
NLR	Weighted median	246	0.66	(0.47 – 0.93)	0.017
	PRESSO (corrected)		0.77	(0.60 – 0.98)	0.036
	RAPS shrinkage		0.73	(0.58 – 0.91)	5.6×10^{-3}
	IVW (multiplicative random effects)		0.80	(0.67 – 0.96)	0.018
	Maximum likelihood		0.80	(0.70 – 0.92)	2.0×10^{-3}
PLR	Weighted median	462	0.85	(0.67 – 1.08)	0.18
	PRESSO (corrected)		0.82	(0.70 – 0.96)	0.014
	RAPS shrinkage		0.85	(0.73 – 0.98)	0.025

Supplementary Table 6: Summary of diagnostic tests carried out for Mendelian Randomization analyses

Trait	Heterogeneity[1]			MR Egger[2]		MR PRESSO		I^2_{GX}[5]	$P_{Steiger}$[6]
	Q	DF	$P_{Q\text{-value}}$	β_0	P_{Egger}	P_{Global}[3]	P_{Dist}[4]		
Basophils	262.3	143	4.7×10^{-9}	-0.0099	0.40	$<6.7\times10^{-5}$	0.98	0.985	1.8×10^{-14}
Eosinophils	509.9	360	3.1×10^{-7}	0.0028	0.69	$<6.7\times10^{-5}$	0.04	0.986	1.6×10^{-75}
Lymphocytes	512.1	398	9.4×10^{-5}	0.0082	0.09	$<6.7\times10^{-5}$	0.88	0.986	5.0×10^{-135}
Monocytes	708.8	472	9.0×10^{-12}	-0.0040	0.15	2.7×10^{-4}	0.69	0.988	1.8×10^{-269}
Neutrophils	368.3	299	3.8×10^{-3}	-0.0098	0.16	3.7×10^{-3}	0.03	0.986	3.8×10^{-120}
Platelets	764.5	654	1.8×10^{-3}	-0.0002	0.93	3.1×10^{-3}	0.75	0.988	$<1\times10^{-300}$
LMR	899.0	428	1.2×10^{-35}	0.0097	0.15	$<6.7\times10^{-5}$	0.63	0.988	5.4×10^{-98}
NLR	524.1	242	6.8×10^{-23}	-0.0103	0.39	$<6.7\times10^{-5}$	0.15	0.984	2.1×10^{-10}
PLR	798.9	456	4.1×10^{-21}	-0.0071	0.30	$<6.7\times10^{-5}$	0.83	0.987	1.2×10^{-109}

[1] Cochran's Q based on modified second order weights, p-values<0.05 indicate statistically significant heterogeneity

[2] Presence of statistically significant directional pleiotropy is indicated by non-zero intercept values with p<0.05

[3] Estimated based on 15000 replicates; p-values<0.05 indicate statistically significant pleiotropy

[4] Estimated based on 15000 replicates; p-values<0.05 indicate statistically significant distortion in causal effect estimates

[5] I^2_{GX} <0.90 indicates weak instrument bias due to violations of the no exposure measurement error assumption

[6] Steiger directionality test for orienting the causal effect as exposure → outcome

Supplementary Table 7: Odds ratios (OR) and 95% for acute lymphoblastic leukemia (ALL), estimated using Mendelian Randomization (MR) for selected traits after filtering out instruments that significantly contributed to heterogeneity.

Trait	Method	N / N_{SNP}	Association Results			Diagnostics	
			OR	(95% CI)	P	$P_{Q\text{-value}}$	P_{Egger}
Lymphocytes	IVW (mre)	356 / 406	1.18	(1.05 – 1.33)	7.4×10^{-3}	1.00	0.17
	Maximum likelihood		1.19	(1.02 – 1.37)	0.023		
LMR	IVW (mre)	403 / 432	1.19	(1.03 – 1.37)	0.016	0.93	0.22
	Maximum likelihood		1.19	(1.02 – 1.38)	0.022		
NLR	IVW (mre)	218 / 246	0.67	(0.55 – 0.82)	1.4×10^{-4}	0.98	0.81
	Maximum likelihood		0.67	(0.53 – 0.84)	6.5×10^{-4}		
PLR	IVW (mre)	429 / 462	0.82	(0.71 – 0.94)	5.8×10^{-3}	0.72	0.72
	Maximum likelihood		0.82	(0.71 – 0.95)	7.3×10^{-3}		

Supplementary Table 8: Odds ratios (OR) and 95% for acute lymphoblastic leukemia (ALL), estimated using multivariable Mendelian Randomization (MVMR) for traits that were individually associated with ALL or were used to derive relevant ratio phenotypes.

Model	Trait	MVMR Method	Association Results		
			OR	(95% CI)	P
Cell types	Lymphocytes	IVW	1.18	(1.06 – 1.31)	3.31×10^{-3}
		IVW (instrument specific)	1.43	(1.16 – 1.76)	8.83×10^{-4}
		LASSO	1.05		
	Monocytes	IVW	1.02	(0.92 – 1.13)	0.70
		IVW (instrument specific)	0.94	(0.83 – 1.08)	0.40
		LASSO	-		
	Neutrophils	IVW	0.93	(0.77 – 1.11)	0.41
		IVW (instrument specific)	0.88	(0.68 – 1.14)	0.33
		LASSO	-		
	Platelets	IVW	0.92	(0.85 – 0.99)	0.040
		IVW (instrument specific)	0.95	(0.86 – 1.05)	0.30
		LASSO	-		
Cell type ratios	LMR	IVW	1.00	(0.91 – 1.09)	0.97
		IVW (instrument specific)	1.10	(0.92 – 1.31)	0.28
		LASSO	-		
	NLR	IVW	0.95	(0.81 – 1.12)	0.54
		IVW (instrument specific)	0.83	(0.60 – 1.14)	0.25
		LASSO	-		
	PLR	IVW	0.90	(0.82 – 0.99)	0.034
		IVW (instrument specific)	0.91	(0.77 – 1.09)	0.30
		LASSO	0.82		

Supplementary Table 9: Mediation analyses that decompose the total effect of blood cell ratios on ALL risk into effects mediated by their component cell types. Total effects are based on Mendelian randomization results after filtering out instruments that significantly contributed to heterogeneity, as seen in Supplementary Table 6.

Trait	Effects	log(OR)	(95% CI)[1]	Mediated (%)	
LMR	Total	0.173	(0.024, 0.322)		
	Indirect (mediated): Lymphocytes	-0.025	(-0.098, 0.053)	14.4	(0, 16.5)
	Indirect (mediated): Monocytes	0.066	(-0.128, 0.264)	38.1	(0, 82.1)
NLR	Total	-0.399	(-0.627, -0.171)		
	Indirect (mediated): Lymphocytes	-0.044	(-0.333, 0.245)	10.9	(0, 53.1)
	Indirect (mediated): Neutrophils	0.008	(-0.046, 0.063)	2.1	(0, 7.3)
PLR	Total	-0.242	(-0.449, -0.034)		
	Indirect (mediated): Lymphocytes	-0.053	(-0.255, 0.138)	21.9	(0, 56.7)
	Indirect (mediated): Platelets	0.043	(-0.100, 0.202)	17.8	(0, 22.3)

[1] Bootstrapped 95% confidence intervals for direct and indirect (mediated) effects were obtained using 1000 replicates.

Supplementary Table 10: Annotation of blood cell trait instruments assigned to substantive clusters using the MR Clust algorithm with an assignment probability of greater than 50%. For each blood cell trait, the substantive cluster mean corresponds to the causal effect on ALL, log odds ratio (OR) per 1 unit increase, indicated by the variants comprising that cluster. Published associations with other phenotypes were obtained by querying the PhenoScanner database.

SNP	P_{ALL}	log(OR_{ALL})				PMID	Study	Traits with P <5.0×10^{-8}
		Lymphocyte	LMR	NLR	PLR			
rs76428106	3.2E-05	2.032	-	-	-1.327	27863252	Astle W	Granulocyte count | Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells | Myeloid white cell count | Sum neutrophil eosinophil counts | White blood cell count
rs1555137	2.0E-03	2.032	-	-	-	27863252	Astle W	Granulocyte count | Myeloid white cell count | Sum basophil neutrophil counts | White blood cell count
rs274555	3.7E-03	2.032	-	-	-	27863252	Astle W	Eosinophil count | Eosinophil percentage of granulocytes | Eosinophil percentage of white cells | Granulocyte count | High light scatter percentage of red cells | High light scatter reticulocyte count | Immature fraction of reticulocytes | Mean platelet volume | Monocyte count | Myeloid white cell count | Neutrophil count | Neutrophil percentage of granulocytes | Platelet count | Sum basophil neutrophil counts | Sum eosinophil basophil counts | Sum neutrophil eosinophil counts | White blood cell count
						26192919	IBDGC	Crohns disease
rs1633043	4.2E-03	2.032	-	-3.240	-1.327	27863252	Astle W	Hemoglobin concentration | High light scatter percentage of red cells | High light scatter reticulocyte count | Immature fraction of reticulocytes | Lymphocyte count | Lymphocyte percentage of white cells | Mean corpuscular hemoglobin | Mean corpuscular hemoglobin concentration | Mean corpuscular volume | Neutrophil percentage of white cells | Red cell distribution width | Reticulocyte count | Reticulocyte fraction of red cells
rs3132571	4.3E-03	2.032	-	-	-	27863252	Astle W	Granulocyte count | Lymphocyte count | Mean platelet volume | Monocyte count | Myeloid white cell count | Neutrophil count | Platelet count | Reticulocyte count | Reticulocyte fraction of red cells | Sum basophil neutrophil counts | Sum eosinophil basophil counts | Sum neutrophil eosinophil counts | White blood cell count
						27723758	Bronson P	IgA deficiency
						21323541	Stanescu HC	Idiopathic membranous nephropathy
						17632545	Hakonarson H	Type 1 diabetes
						27992413	Ji S	Primary sclerosing cholangitis

rs7776054	7.7E-03	2.032	-	-	-1.327	27863252	Astle W	Eosinophil count | Granulocyte count | Hematocrit | Hemoglobin concentration | High light scatter percentage of red cells | Lymphocyte count | Mean corpuscular hemoglobin | Mean corpuscular hemoglobin concentration | Mean corpuscular volume | Monocyte count | Myeloid white cell count | Neutrophil count | Platelet count | Plateletcrit | Red blood cell count | Red cell distribution width | Reticulocyte fraction of red cells | Sum basophil neutrophil counts | Sum eosinophil basophil counts | Sum neutrophil eosinophil counts | White blood cell count
						19862010	Ganesh SK	Erythrocyte count | Hematocrit Hct | Mean corpuscular hemoglobin MCH | Mean corpuscular hemoglobin concentration | Mean corpuscular volume | Mean corpuscular hemoglobin | Erythrocyte indices
rs11757367	0.015	2.032	-	-	-	27863252	Astle W	Eosinophil count | Granulocyte count | Mean corpuscular hemoglobin | Mean corpuscular volume | Mean platelet volume | Monocyte count | Myeloid white cell count | Neutrophil count | Platelet count | Plateletcrit | Red blood cell count | Sum basophil neutrophil counts | Sum eosinophil basophil counts | Sum neutrophil eosinophil counts | White blood cell count
						27723758	Bronson P	IgA deficiency
						20453842	Stahl E	Rheumatoid arthritis
						27992413	Ji S	Primary sclerosing cholangitis
						24390342	Okada Y	Rheumatoid arthritis
rs2872516	0.024	2.032	-	-	-	27863252	Astle W	Basophil count | Eosinophil percentage of granulocytes | Eosinophil percentage of white cells | Granulocyte count | Granulocyte percentage of myeloid white cells | Lymphocyte count | Lymphocyte percentage of white cells | Monocyte percentage of white cells | Myeloid white cell count | Neutrophil count | Neutrophil percentage of granulocytes | Neutrophil percentage of white cells | Sum basophil neutrophil counts | Sum neutrophil eosinophil counts | White blood cell count
						29083406	Ferreira M	Allergic disease
						26192919	IBDGC	Inflammatory bowel disease
rs12722496	0.025	2.032	-	-	-1.327	27863252	Astle W	Lymphocyte count | Lymphocyte percentage of white cells
						29083406	Ferreira M	Allergic disease
rs28447467	0.026	2.032	1.292	-3.240	-1.327	27863252	Astle W	Lymphocyte count | Lymphocyte percentage of white cells | Neutrophil percentage of white cells | White blood cell count
rs61839660	8.6E-04	-	1.292	-3.240	-	27863252	Astle W	Lymphocyte count | Lymphocyte percentage of white cells
						29083406	Ferreira M	Allergic disease | Allergic disease asthma hay fever or eczema
						22293688	Huang J	Type 1 diabetes | Diabetes mellitus type 1
rs4385425	9.7E-04	-	1.292	-	-	27863252	Astle W	Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells | Neutrophil percentage of white cells

rs62447197	6.5E-03	-	1.292	-	-	27863252	Astle W	Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells
rs10980797	7.6E-03	-	1.292	-	-	27863252	Astle W	Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells | Myeloid white cell count | Neutrophil percentage of white cells | White blood cell count
rs57822871	0.012	-	1.292	-	-	27863252	Astle W	Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells
rs62447173	0.022	-	1.292	-	-	27863252	Astle W	Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells
rs67483792	0.0242	-	1.292	-	-	27863252	Astle W	Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells
rs369173	0.04424	-	1.292	-	-	27863252	Astle W	Granulocyte count | Granulocyte percentage of myeloid white cells | Monocyte count | Monocyte percentage of white cells | Neutrophil count | Neutrophil percentage of white cells | Sum basophil neutrophil counts | Sum neutrophil eosinophil counts
rs6430608	2.5E-06	-	-	-3.240	-	-	-	-
rs57153090	1.5E-03	-	-	-3.240	-	27863252	Astle W	Lymphocyte count | Lymphocyte percentage of white cells
rs2415288	5.3E-03	-	-	-3.240	-	27863252	Astle W	Lymphocyte count
rs35045014	6.0E-03	-	-	-3.240	-	27863252	Astle W	Eosinophil count | Eosinophil percentage of granulocytes | Eosinophil percentage of white cells | Granulocyte percentage of myeloid white cells | Monocyte percentage of white cells | Neutrophil percentage of granulocytes | Sum eosinophil basophil counts
rs192665233	0.021	-	-	-3.240	-	27863252	Astle W	Granulocyte count | Lymphocyte percentage of white cells | Neutrophil count | Neutrophil percentage of white cells | Sum basophil neutrophil counts | Sum neutrophil eosinophil counts
rs3132713	0.022	-	-	-3.240	-	27863252	Astle W	Hemoglobin concentration | High light scatter percentage of red cells | High light scatter reticulocyte count | Immature fraction of reticulocytes | Lymphocyte count | Lymphocyte percentage of white cells | Mean corpuscular hemoglobin | Mean corpuscular hemoglobin concentration | Mean corpuscular volume | Neutrophil percentage of white cells | Red cell distribution width | Reticulocyte count | Reticulocyte fraction of red cells
rs9533128	0.024	-	-	-3.240	-	-	-	-
rs1984021	2.2E-03	-	-	-	-1.327	27863252	Astle W	Platelet count | Plateletcrit

rs4795397	2.7E-03	-	-	-	-1.327	27863252	Astle W	Basophil count | Eosinophil percentage of granulocytes | Eosinophil percentage of white cells | Granulocyte count | Granulocyte percentage of myeloid white cells | Lymphocyte count | Lymphocyte percentage of white cells | Monocyte percentage of white cells | Myeloid white cell count | Neutrophil count | Neutrophil percentage of granulocytes | Neutrophil percentage of white cells | Sum basophil neutrophil counts | Sum neutrophil eosinophil counts | White blood cell count
						29083406	Ferreira M	Allergic disease
						24097068	GLGC	High density lipoprotein
						22961000	Liu	Primary biliary cirrhosis
						26192919	IBDGC	Crohns disease | Inflammatory bowel disease | Ulcerative colitis
						24390342	Okada Y	Rheumatoid arthritis
						29273806	TAGC	Asthma
rs7654909	4.4E-03	-	-	-	-1.327	-	-	-
rs72974176	0.005	-	-	-	-1.327	27863252	Astle W	Hematocrit | Mean corpuscular hemoglobin | Mean corpuscular hemoglobin concentration | Mean corpuscular volume | Platelet count | Plateletcrit | Red blood cell count | Red cell distribution width
rs113107801	7.2E-03	-	-	-	-1.327	27863252	Astle W	Lymphocyte count
rs77619009	9.5E-03	-	-	-	-1.327	26192919	IBDGC	Inflammatory bowel disease | Ulcerative colitis
		-	-	-		24390342	Okada Y	Rheumatoid arthritis
rs2428514	0.017	-	-	-	-1.327	27863252	Astle W	White blood cell count
						27992413	Ji S	Primary sclerosing cholangitis
						24390342	Okada Y	Rheumatoid arthritis
rs114694170	0.022	-	-	-	-1.327	27863252	Astle W	Mean corpuscular hemoglobin | Mean corpuscular volume | Mean platelet volume | Platelet count | Platelet distribution width | Plateletcrit | Red blood cell count
rs7098181	0.022	-	-	-	-1.327	27863252	Astle W	Granulocyte count | Granulocyte percentage of myeloid white cells | Mean platelet volume | Monocyte percentage of white cells | Myeloid white cell count | Neutrophil count | Neutrophil percentage of white cells | Platelet count | Platelet distribution width | Sum basophil neutrophil counts | Sum neutrophil eosinophil counts | White blood cell count
rs7098181	0.022				-1.327	27225129	SSGAC	Years of educational attainment in males | Years of educational attainment
rs10066265	0.025	-	-	-	-1.327	27863252	Astle W	Mean platelet volume | Platelet count | Red cell distribution width
rs145125135	0.037	-	-	-	-1.327	27863252	Astle W	Mean corpuscular hemoglobin | Mean corpuscular volume | Platelet count | Plateletcrit | Red blood cell count | Red cell distribution width

rs2523466	0.042	-	-	-	-1.327	27863252	Astle W	Basophil count | Granulocyte count | Hematocrit | Hemoglobin concentration | High light scatter percentage of red cells | High light scatter reticulocyte count | Lymphocyte count | Monocyte count | Myeloid white cell count | Neutrophil count | Reticulocyte count | Reticulocyte fraction of red cells | Sum basophil neutrophil counts | Sum neutrophil eosinophil counts | White blood cell count
rs2523466	0.042	-	-	-	-1.327	27723758 27992413 24390342	Bronson P Ji S Okada Y	IgA deficiency Primary sclerosing cholangitis Rheumatoid arthritis
rs56043070	0.066	-	-	-	-1.327	27863252	Astle W	High light scatter percentage of red cells | High light scatter reticulocyte count | Mean platelet volume | Platelet count | Platelet distribution width | Plateletcrit | Reticulocyte count | Reticulocyte fraction of red cells

Supplementary Table 11: Mediation analyses that decompose the total effect of selected ALL risk variants into direct and indirect effects, mediated via regulation of blood cell profiles. Mediator-outcome effects were obtained from Mendelian randomization analyses excluding outliers (as seen in Supplementary Table 7).

		OR_{ALL}	P_{ALL}	β_{ALL}	(95% CI)[1]	Mediated (%)	
Mediator: PLR		0.82	5.8×10^{-3}	-0.200			
rs4245597 (*ARID5B*, 10q21.2)	Total	0.60	5.3×10^{-46}	-0.505	(-0.575, -0.436)		
	Direct			-0.500	(-0.571, -0.431)		
	Indirect (mediated)			-0.004	(-0.007, -0.001)	0.84	(0.23, 1.44)
rs76428106 (*FLT3*, 13q12.2)	Total	0.56	3.2×10^{-5}	-0.580	(-0.853, -0.307)		
	Direct			-0.562	(-0.841, -0.291)		
	Indirect (mediated)			-0.014	(-0.024, -0.004)	2.43	(0.68, 4.19)
Mediator: NLR		0.67	1.4×10^{-4}	-0.400			
rs4245597 (*ARID5B*, 10q21.2)	Total	0.60	5.3×10^{-46}	-0.505	(-0.575, -0.436)		
	Direct			-0.496	(-0.567, -0.426)		
	Indirect (mediated)			-0.008	(-0.013, -0.004)	1.65	(0.79, 2.51)
rs6430608 (2q22.1)	Total	0.78	2.5×10^{-6}	-0.244	(-0.346, -0.142)		
	Direct			-0.231	(-0.334, -0.129)		
	Indirect (mediated)			-0.012	(-0.018, -0.006)	4.72	(2.26, 7.20)
Mediator: LMR		1.19	0.016	0.173			
rs4948492 (*ARID5B*, 10q21.2)	Total	1.66	5.1×10^{-46}	0.507	(0.437, 0.577)		
	Direct			0.503	(0.433, 0.574)		
	Indirect (mediated)			0.004	(-0.001, 0.008)	0.75	(0, 1.63)
rs74756667 (8q24.2)	Total	1.37	3.3×10^{-10}	0.313	(0.215, 0.410)		
	Direct			0.310	(0.210, 0.407)		
	Indirect (mediated)			0.004	(0.001, 0.007)	1.31	(0.24, 2.38)
rs2239630 (*CEBPE*, 14q11.2)	Total	1.32	1.2×10^{-13}	0.278	(0.206, 0.353)		
	Direct			0.275	(0.199, 0.348)		
	Indirect (mediated)			0.006	(-0.001, 0.013)	2.16	(0, 4.71)
rs78697948* (*IKZF1*, 7p12.2)	Total	0.68	1.2×10^{-10}	-0.391	(-0.509, -0.272)		
	Direct			-0.395	(-0.516, -0.276)		
	Indirect (mediated)			0.007	(-0.001, 0.015)	1.72	(0, 3.75)
rs76428106* (*FLT3*, 13q12.2)	Total	0.56	3.2×10^{-5}	-0.580	(-0.853, -0.307)		
	Direct			-0.637	(-0.918, -0.368)		
	Indirect (mediated)			0.066	(0.012, 0.120)	11.39	(2.07, 20.7)
Mediator: Lymphocytes		1.18	7.4×10^{-3}	0.166	(0.045, 0.288)		
rs76428106 (*FLT3*, 13q12.2)	Total	1.79	3.2×10^{-5}	0.580	(0.307, 0.853)		
	Direct			0.569	(0.290, 0.840)		
	Indirect (mediated)			0.015	(0.004, 0.025)	2.51	(0.67, 4.35)

[1] Bootstrapped 95% confidence intervals for direct and indirect (mediated) effects were obtained using 1000 replicates.

* Variants had statistically significant effects, and of larger magnitude, on monocyte counts than LMR or lymphocytes