Human Genomics

RESEARCH

Open Access

A genome-wide association study of neutrophil count in individuals associated to an African continental ancestry group facilitates studies of malaria pathogenesis

Andrei-Emil Constantinescu[1,2,4], David A. Hughes[1,2,3], Caroline J. Bull[1,2,4,5], Kathryn Fleming[6], Ruth E. Mitchell[1,2], Jie Zheng[7,8,9], Siddhartha Kar[1,2,10], Nicholas J. Timpson[1,2], Borko Amulic[6*†] and Emma E. Vincent[1,2,4*†]

Abstract

Background 'Benign ethnic neutropenia' (BEN) is a heritable condition characterized by lower neutrophil counts, predominantly observed in individuals of African ancestry, and the genetic basis of BEN remains a subject of extensive research. In this study, we aimed to dissect the genetic architecture underlying neutrophil count variation through a linear-mixed model genome-wide association study (GWAS) in a population of African ancestry ($N = 5976$). Malaria caused by *P. falciparum* imposes a tremendous public health burden on people living in sub-Saharan Africa. Individuals living in malaria endemic regions often have a reduced circulating neutrophil count due to BEN, raising the possibility that reduced neutrophil counts modulate severity of malaria in susceptible populations. As a follow-up, we tested this hypothesis by conducting a Mendelian randomization (MR) analysis of neutrophil counts on severe malaria (MalariaGEN, $N = 17{,}056$).

Results We carried out a GWAS of neutrophil count in individuals associated to an African continental ancestry group within UK Biobank, identifying 73 loci ($r^2 = 0.1$) and 10 index SNPs (GCTA-COJO loci) associated with neutrophil count, including previously unknown rare loci regulating neutrophil count in a non-European population. BOLT-LMM was reliable when conducted in a non-European population, and additional covariates added to the model did not largely alter the results of the top loci or index SNPs. The two-sample bi-directional MR analysis between neutrophil count and severe malaria showed the greatest evidence for an effect between neutrophil count and severe anaemia, although the confidence intervals crossed the null.

Conclusion Our GWAS of neutrophil count revealed unique loci present in individuals of African ancestry. We note that a small sample-size reduced our power to identify variants with low allele frequencies and/or low effect sizes in our GWAS. Our work highlights the need for conducting large-scale biobank studies in Africa and for further exploring the link between neutrophils and severe malaria.

Keywords Malaria, Neutrophil count, Mendelian randomization, GWAS, African ancestry

†Borko Amulic and Emma E. Vincent are senior authors.

*Correspondence:
Borko Amulic
borko.amulic@bristol.ac.uk
Emma E. Vincent
emma.vincent@bristol.ac.uk
Full list of author information is available at the end of the article

Introduction

Malaria is a mosquito-transmitted disease that annually affects approximately 215 million people [1, 2]. The disease is caused by protozoan parasites of the *Plasmodium* genus: *Plasmodium falciparum* (*P. falciparum*) causes life-threatening disease in sub-Saharan Africa and accounts for almost all malaria deaths, while *P. vivax* leads to a milder disease that is nonetheless associated with a significant public health burden in diverse geographical regions [2].

P. falciparum malaria causes approximately 400,000–600,000 deaths each year, primarily in African children under the age of five [1]. The majority of *P. falciparum* malaria cases consist of uncomplicated febrile illness, however a portion of nonimmune infected individuals succumb to severe malaria, which can manifest as cerebral malaria, severe anemia, acute respiratory distress or kidney injury [3]. *Plasmodium* resides and proliferates in red blood cells (RBCs) and pathology is triggered by cytoadherence of infected RBCs (iRBCs) to microcapillary endothelia in different organs, which can lead to vascular occlusion and endothelial permeability [3]. Inflammation plays a key role in both facilitating iRBC sequestration [4] and in tissue damage [3, 5, 6]. In cerebral malaria, the deadliest form of the disease, iRBCs sequester in the neurovasculature, provoking blood brain barrier permeabilization, vascular leak and brain swelling [3].

Malaria has been the biggest cause of childhood deaths over the past 5000 years [7]. As such, it has exerted the strongest known selective pressure on the human genome and has resulted in the selection of various polymorphisms that confer *Plasmodium* tolerance or resistance. Among the most prominent examples are haemoglobin S (Hbs; sickle cell trait) [8] and alpha-thalassemia variants [9], both of which are common in malaria endemic regions despite causing disease in the homozygous state [7]. The HbS polymorphism in the heterozygous state confers the greatest protection (effect size > 80%; [7, 10]). The heritability of severe malaria is estimated to be around 30% [11, 12] but the cumulative effect of the aforementioned variants is thought to only be 2% [7, 11], suggesting that polygenic interactions may account for a large part of the missing heritability of this complex disease.

Individuals living in malaria-endemic regions, as well as those descended from them, often have reduced numbers of neutrophils in circulation as compared to those living in non-endemic regions. This heritable phenomenon is called 'benign ethnic neutropenia' (BEN) and is distinct from life-threatening severe neutropenia. BEN is prominent in South Mediterranean, Middle Eastern, sub-Saharan African and West Indies populations [13]. BEN is estimated to occur in 25–50% of Africans [13–15] and 10.7% of Arabs [16] but in less than 1% of people of European ancestry living in the Americas [17]. Neutrophils are essential for immune defense against bacteria and fungi [18], however BEN does not lead to significantly greater susceptibility to infection in the United States [13]. Nevertheless, it remains curious that selection for lower neutrophil counts occurred in sub-Saharan Africa, a region associated with a high infectious disease burden. This observation is partly explained by the finding that in populations of African and Yemenite Jewish ancestry, BEN is strongly associated with a polymorphism in the atypical chemokine receptor 1 (ACKR1/DARC), which encodes the Fy/Duffy antigen, a surface receptor utilized by *P. vivax* to invade RBCs [19]. This variant abolishes expression of ACKR1 on RBCs and is thought to contribute to low prevalence of *P. vivax* in sub-Saharan Africa, where the polymorphism is found at levels close to fixation [7]. ACKR1, in addition to serving as one of the invasion receptors for *P. vivax*, controls circulating levels of chemokines [20], which also regulate blood neutrophil numbers [20]. It is unclear to what extent other polymorphisms contribute to BEN in individuals living in malaria endemic regions [21].

Neutrophils have recently been shown to have a detrimental role in malaria, promoting pathogenesis by enhancing sequestration of iRBCs [4] and contributing to inflammatory tissue damage [6, 22, 23]. Altered neutrophil responses have also been linked to severe malarial anemia in paediatric patients [24]. On the other hand, neutrophils have also been suggested to participate in parasite clearance [25] and in shaping the *Plasmodium* antigenic repertoire [26]. These studies raise the possibility that neutropenia in malaria endemic regions may modulate severity of *P. falciparum* malaria, in addition to conferring resistance to *P. vivax*. However, observational studies, such as the ones referenced above, are prone to confounding and reverse causation [27–29]. It is therefore essential to employ additional methods, such as those in population genetics, to study the link between neutrophil count and *P. falciparum* severe malaria, with the overarching aim to improve the health outcomes of the people residing in endemic regions.

Mendelian randomization (MR) is a method in genetic epidemiology which uses genetic variants as proxies with the aim of providing evidence for causal inference between an exposure and an outcome [27]. As the majority of alleles are assigned randomly at birth, an MR analysis is analogous to that of a randomized control trial (RCT), the most reliable method for evaluating the effectiveness of an intervention [30]. Large-scale studies, such as UK Biobank (UKBB) [31], have increased the potential of MR studies due to the

increase in power to detect associations in genome-wide association studies (GWASs) that comes with such a large sample size.

Recent efforts in genetics have resulted in the generation of hundreds of GWAS using UKBB's non-European participants for many traits in a hypothesis-free manner (https://pan.ukbb.broadinstitute.org/). However, the same covariates were used for each trait, and the impact of confounding due to population structure was not studied, this represents a potential limitation for constructing reliable instruments for a MR analysis [32]. A recent study by Chen et al. used individuals of non-European ancestry in UKBB to perform trans-ancestry GWAS of blood cell traits (BCTs) [33]. However, the African continental ancestry groups (CAGs) of UKBB display strong population structure [34]. It therefore remains unclear whether a GWAS of a complex trait, such as neutrophil count, would result in associations that are linked to a biological mechanism, or whether the associations would be a product of confounding due to residual population structure. In order to answer these questions, a more thorough investigation of the sampled dataset is warranted. This becomes even more important when aiming to conduct causal inference analyses in genetic epidemiology, such as two-sample Mendelian randomization [35, 36].

To test the hypothesis that reduced neutrophil counts modulates severity of malaria in susceptible populations, we first performed a GWAS of neutrophil count in individuals associated to the UKBB African continental ancestry group (CAG), described in our previous study [34]. Here, we conducted a series of sensitivity analyses to describe the GWAS results and selection of genetic instruments to proxy for neutrophil count in a MR analysis. We then conducted bi-directional MR to estimate the casual relationship between neutrophil count and SM using data from the MalariaGEN consortium [37].

Materials and methods

Study design

6,653 people representing the UKBB African CAG were identified as part of our previous study [34]. After PCA outlier filtering [34], we also excluded those without neutrophil count data and blood-related disorders [38], resulting in a final sample of 5,976. The primary GWAS of neutrophil count used in all other analyses was generated with BOLT-LMM. Several analyses were undertaken afterwards to test the validity of the primary GWAS estimates. Following this, an MR analysis was performed between neutrophil count and severe malaria caused by *P. falciparum* using data from MalariaGEN (Fig. 1).

UK Biobank genetic data

UK Biobank's "non-white" British data was studied previously, where 6653 people corresponded to the African CAG, of which 6504 remained (5989 unrelated; 515 related) after filtering for principal component analysis (PCA) outliers [34]. These were further assigned into seven clusters based on a K-means clustering algorithm (K1 = 527; K2 = 1,177; K3 = 1176; K4 = 1001; K5 = 1206; K6 = 862; K7 = 184) [34]. This dataset (N = 6504) included both directly genotyped (N = 784,256) and imputed (N = 29,363,284) SNPs filtered with a minor allele count of > 20. We filtered out SNPs with an INFO score threshold of 0.3, as it gives the best balance between data quality and quantity. Another filtering process was a Hardy–Weinberg equilibrium (HWE) test (P < 1e−10), used to identify SNPs of poor genotyping quality [39]. Finally, related individuals from the dataset were removed, resulting in 5,509 unrelated people in the filtered African CAG dataset. SNPs with a minor allele count of less than 17 (corresponding to the new sample-size from 20) were removed. 23,530,028 SNPs remained after filtering by INFO score, HWE test and minor allele count.

UK Biobank phenotypic data

Haematological samples were analysed using four Beckman Coulter LH750 instruments [40]. Total white blood cell (WBC) count and neutrophil percentage (%) were measured through the Coulter method, with neutrophil count derived as "neutrophil %/100 × total WBC" and expressed as 10^9 cells/Litre [40]. Afterwards, the sample collection date was split into year, month, day, and minutes (passed since the start of the day of the appointment visit), while the neutrophil count measurement variable was log-transformed into a variable named "nc_log", which was used as the default neutrophil count variable throughout the study. Other variables that were used in the main analyses were: age, genetic sex, blood sample device ID, UKBB assessment centre and principal components (PCs) 1 to 100. Filtering was performed based on the selection criteria described by Astle et al. [38] and Chen et al. [33]. Briefly, individuals with disorders/diseases that could affect blood counts (e.g. HIV, leukaemia, congenital anaemias, cirrhosis) were removed, bringing the final sample size to 5,976. This dataset is referred to as "AFR_CAG".

BOLT-LMM GWAS

BOLT-LMM was used as the software to run the primary (main) GWAS. Linkage disequilibrium (LD) scores were generated from the directly genotyped dataset that is required by BOLT-LMM to calibrate the test statistics. After preparing the phenotypic data to match the desired

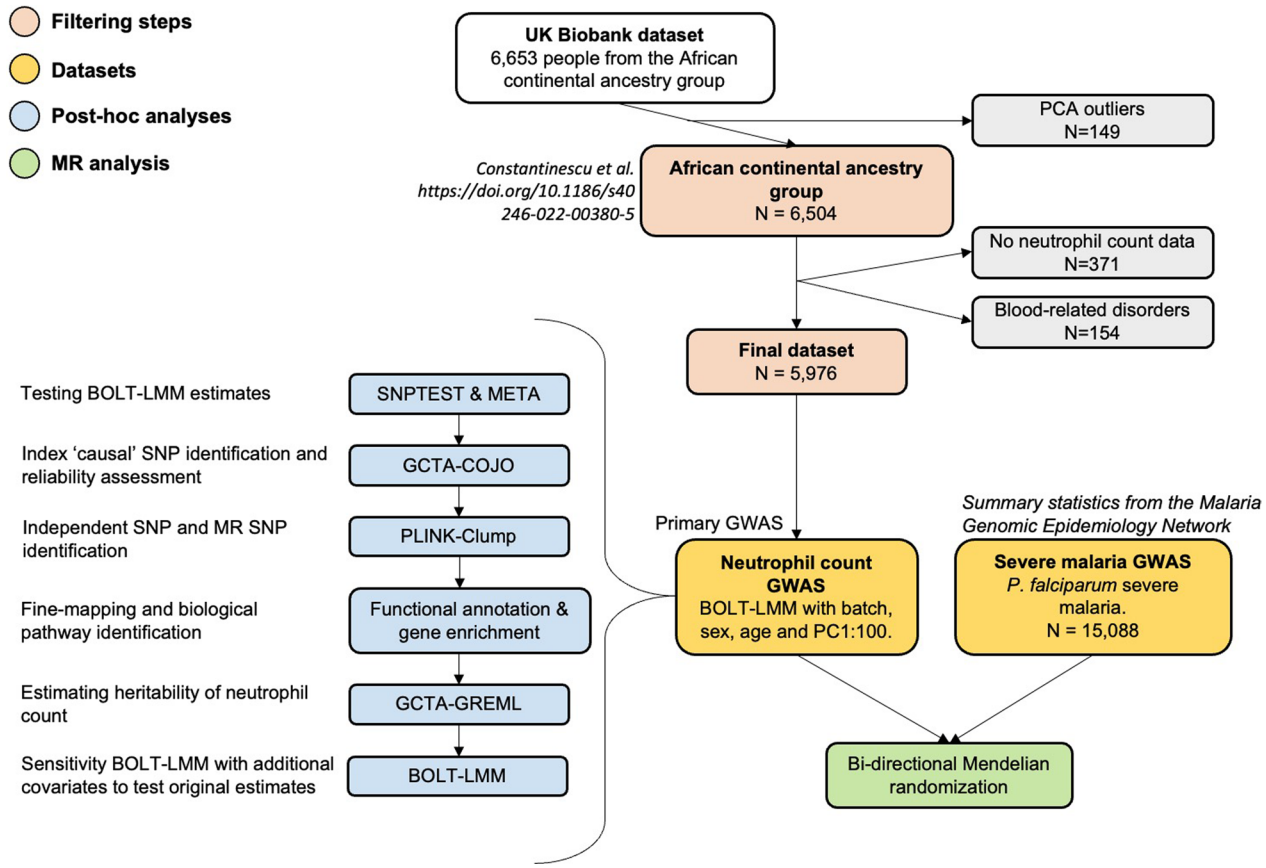

Fig. 1 Study design of the project

input, BOLT-LMM was run on AFR_CAG adjusting for age, genetic sex, UKBB assessment centre, blood sampling device, sampling year, sampling month, sampling day, minutes passed in sampling day and the first 100 principal components (PCs). Two linear model GWAS in SNPTEST were also completed on each K-means cluster and then meta-analysed: one without accounting for the Duffy SNP rs2814778 called "META-WOD", and one where the Duffy SNP was included as a covariate, called "META-WD". Another BOLT-LMM sensitivity run was done with additional covariates to further study the validity of the main GWAS findings (Additional file 1: Methods).

Conditional and joint association analysis

We used GCTA-COJO [41, 42] to identify independent signals from the BOLT-LMM GWAS, as well as to detect any possible secondary signals arising from a stepwise selection model. SNPs which are close together are usually in LD i.e. their alleles are not random, but correlated [39]. Before running GCTA-COJO, related individuals were filtered out of the dataset. PLINK was then used on this resulting output to perform a greedy filtering of related individuals. Following this step, GCTA-COJO was run on the AFR_CAG filtered dataset to identify conditionally independent SNPs. These were referred to as "index" SNPs in the text.

PLINK clumping

After GCTA-COJO, we used PLINK to perform clumping with three different thresholds. The first two represent the thresholds for defining LD independent SNPs for running analyses on the online variant annotation platform Functional Mapping and Annotation (FUMA) [43], while the latter being the clumping conditions used for conducting a Mendelian randomization analysis [44, 45].

1. –clump-p1 = 5e–8, –clump-r2 = 0.6, –clump-kb = 250
2. –clump-p1 = 5e–8, –clump-r2 = 0.1, –clump-kb = 250

3. –clump-p1 = 5e−8, –clump-r2 = 0.001, –clump-kb = 10,000

Heritability analysis

An analysis was conducted with GCTA-GREML to estimate the proportion of variance in neutrophil count explained by all genetic variants present in the filtered AFR_CAG dataset [46], with and without adjusting for the Duffy SNP rs2814778.

P. falciparum severe malaria genetic data

GWAS summary statistics for *P. falciparum* severe malaria were downloaded from a case–control study that spanned nine African and two Asian countries [37]. In brief, controls samples were gathered from cord blood, and in some cases, from the general population. Cases were diagnosed according to WHO definitions of severe malaria [47] and were categorised according to CM, severe malarial anemia (SMA) and other severe malaria (OTHER) symptoms (Additional file 3: Table S1). The majority of the RSIDs in the MalariaGEN dataset used older identifiers, and some of them had the "kgp" prefix that comes with the Illumina-HumanOmni2.5 M array. Ideally, in a two-sample MR setting, the two samples would have a perfect match in the available genetic variants. It is desirable to at least maximise the number of matching variants to test. Therefore, RSID information for the MalariaGEN variants was updated in R by using the filtered AFR_CAG dataset as a reference panel.

Meta-analysis of severe malaria African populations

Summary statistics for severe malaria and its sub-phenotypes were generated from a meta-analysis which included individuals from two non-African countries—Vietnam and Papua New Guinea. The inclusion of SNP effect sizes from GWAS conducted in heterogenous population might bias MR estimates [48]. Therefore, per-population summary statistics were downloaded (https://www.malariagen.net/sppl25/) for each African country in the study and a meta-analysis was conducted on them using METAL [49–51].

Mendelian randomization analysis

The "TwoSampleMR" R package [52, 53] was used to perform the MR analyses. The two datasets were harmonised i.e. orientated on the same strand and if SNPs were not found in the outcome dataset, we searched for SNP proxies. We then conducted a bi-directional MR analysis, where the effect of neutrophil count on overall severe malaria, along with the three sub-phenotypes was estimated and vice-versa. The main analysis was conducted using an IVW model [54]. Additionally, we ran a sensitivity MR analysis to outline the effect estimates of each SNP on the desired outcome, with IVW and MR-Egger [55, 56] estimates where the number of instruments was larger than two and three, respectively.

Results

Analysis of study sample

5,976 out of 6,504 individuals in AFR_CAG remained after filtering for missing data and traits affecting blood cells. The mean value for neutrophil count was 2.9×10^9 cells/litre, as expected this was lower than a European sample (4.21×10^9 cells/Litre) [33, 38]. The GWAS sample had a larger proportion of females (57%), was of a higher mean age (39 vs. 58.1 years) [57] and slightly higher body mass index (BMI) (27.6 vs. 29.8 kg/m^2) [58] than the general UK population (Additional file 3: Table S2).

We used the natural log-transformation, nc_log, in the GWAS. There was some variation in nc_log between each K-means cluster (Kpop) (Additional file 2: Fig. S1B), although this was low, with the median hovering around 1 (Additional file 2: Fig. S1A).

Next, we conducted a power calculation supposing a linear, additive, GWA model [59–61]. The power to detect an association was > 80% when the proportion of variance explained by SNPs was higher than 0.75% (Additional file 2: Fig. S2).

Genome-wide association study

We used BOLT-LMM for the main GWAS, which employs a linear-mixed model algorithm for conducting association testing [62]. It is unknown how well linear mixed model using PCs and kinship matrixes performs in highly stratified population samples with complex demographic histories and unique allele frequencies and linkage disequilibrium [63]. To ensure that results derived by a linear mixed model as implemented by BOLT-LMM were reliable, we also aimed to conduct additional GWAS using a standard linear model on less stratified sub-samples of our sample population—as identified using an unsupervised machine learning methodology Additional file 1: Methods).

This AFR_CAG filtered sample was taken forward for further analyses. 704 genetic variants passed the GWAS significance threshold of P < 5e-8 in the primary GWAS. Most of these signals were in chromosome 1, in the proximity of the ACKR1-associated rs2814778, which had the lowest P-value across the genome (2.7E-87) (Additional file 2: Fig. S2A). The META-WOD GWAS had 373 variants passing the threshold, while

the META-WD (with Duffy adjustment) GWAS had 31 significant SNPs, evidencing that most of the identified top signals in META-WOD were likely in LD with rs2814778. The QQ-plot of the BOLT-LMM GWAS did not display an early deviation from the expected P-value, indicating low likelihood of systemic bias in association statistics [64] (Additional file 2: Fig. S2B).

Next, we aimed to identify which SNPs might causally associate with neutrophil count. To do this, we used a conservative GCTA-COJO approach [42], which yielded 10 index SNPs (Fig. 2A, Table 1). Genomic location context of each index SNP is available in Additional file 2: Figs. S3–S5.

A sensitivity BOLT-LMM GWAS was conducted with six additional covariates on 5,310 individuals: UN region of birth, K-means cluster, smoking status, alcohol drinker status, menstrual status and BMI (Additional file 3: Table S4, Additional file 2: Fig. S11). The association statistics of this sensitivity run and the main BOLT-LMM GWAS run were compared, showing very similar results (Additional file 3: Table S5). This provides evidence that the effect of these additional variables on the main GWAS were modest, and that the PCs and kinship matrix derived by BOLT-LMM appears to have accounted for any population stratification. As a follow-up, we aimed to assess if "missing" or "prefer not to answer" data in these additional covariates associated with differences in neutrophil count,. Even after adjusting for these additional variables, there was no evidence of a difference in neutrophil count (Additional file 3: Table S6).

The effect sizes of the primary GWAS index SNPs were compared with those from the SNPTEST/META GWAS. The direction was consistent and effect sizes were similar between the three GWAS, with those generated from the BOLT-LMM run (primary GWAS) being slightly larger, most likely due to the improved sample size (minor allele count) and power of the linear-mixed model (Fig. 3). As expected, the META-WD effect size for the rs2814778 SNP was zero.

We next investigated the association statistics of the index SNPs in each Kpop. This was done to detect discrepancies in directionality and effect sizes, which could indicate residual population structure or a SNP association with a specific Kpop. Overall, there was agreement in direction, and some variation in effect sizes was detected across Kpops (Additional file 2: Fig. S6).

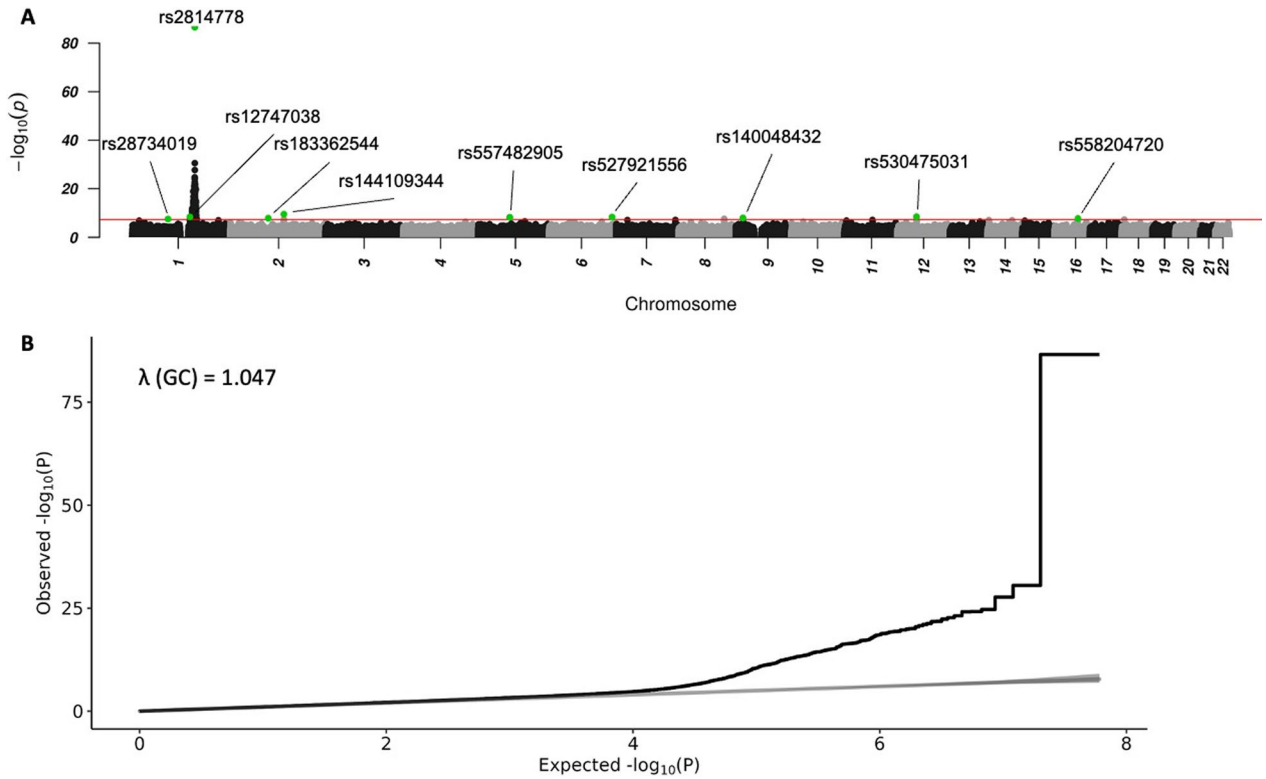

Fig. 2 Manhattan plot of neutrophil count GWAS. The x-axis is the base-pair position inside each chromosome, while the y-axis is the -log of the association P-value. A GWAS significance line is drawn to correspond to P = 5e−8 on the −log(P) axis (**A**). Index SNPs from the GCTA-COJO run are highlighted in green. QQ-Plot of observed vs. expected P-values for each SNP, along with the genomic inflation factor on the top-left (**B**)

Table 1 GCTA-COJO index SNPs. BETA, SE and P BOLT are the regression statistics of the BOLT-LMM neutrophil count GWAS. BETA, SE and P for META-WOD and META-WD are the regression statistics of the meta-analysed GWAS done on each Kpop, without and with adjustment for the Duffy SNP, respectively, META-N and META-N-Studies indicate the number of meta-analysed individuals in each Kpop, and the number of Kpops included in the meta-analysis

SNP	BETA-BOLT	SE-BOLT	P-BOLT	BETA-META-WOD	BETA-META-WD	META-N	META-N-Studies	CHR	BP (GRCh37)	EA	NEA	EAF
rs2814778	0.43	0.02	2.66E−87	0.28	0	5,793	6	1	159,174,683	T	C	0.036
rs144109344	−0.12	0.02	3.12E−10	−0.06	−0.06	5,976	7	2	136,787,730	C	T	0.964
rs530475031	0.73	0.12	3.16E−09	0.47	0.46	4,952	5	12	48,810,860	G	T	0.998
rs12747038	−0.22	0.04	3.89E−09	−0.13	−0.08	5,976	7	1	146,651,428	T	G	0.990
rs527921556	0.4	0.07	4.48E−09	0.33	0.31	5,793	6	6	160,605,701	T	C	0.996
rs557482905	0.55	0.1	5.79E−09	0.42	0.38	3,778	4	5	80,629,499	C	T	0.998
rs140048432	−0.33	0.06	1.11E−08	−0.25	−0.25	5,976	7	9	17,700,893	T	C	0.996
rs183362544	0.61	0.11	1.27E−08	0.28	0.28	2,717	4	2	97,045,902	C	T	0.998
rs558204720	0.52	0.09	1.67E−08	0.37	0.33	1,486	2	16	59,472,815	T	C	0.998
rs28734019	−0.65	0.12	2.89E−08	−0.53	−0.5	4,124	4	1	90,800,573	C	T	0.998

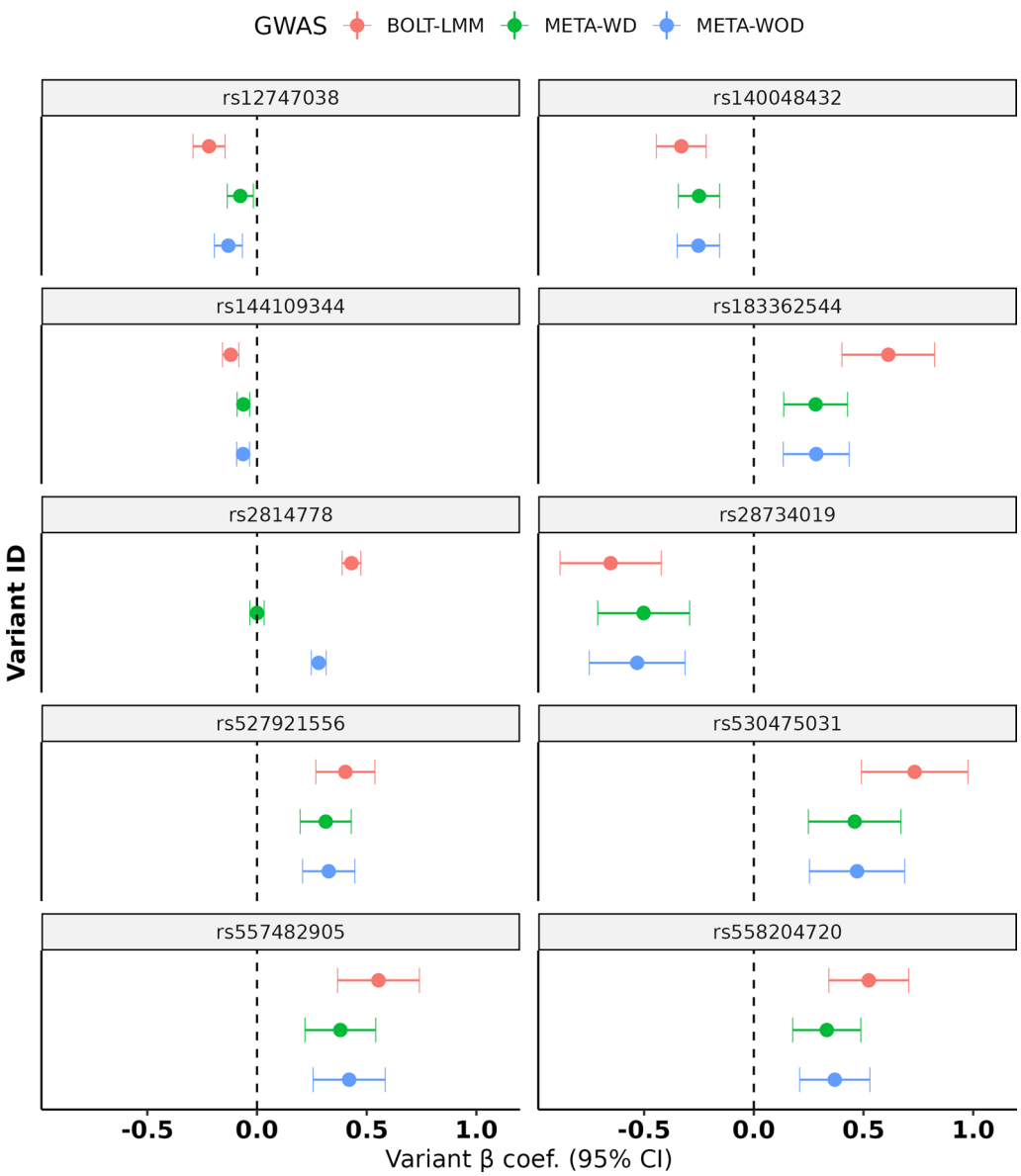

Fig. 3 Effect estimates of the index SNPs. The beta coefficient for each index SNP is displayed along with 95% CIs. These are displayed for the BOLT-LMM, META-WOD and META-WD GWAS

The GCTA-COJO analysis was also run on the two SNPTEST/META GWASs. The META-WOD analysis identified rs2814778, rs138163369 and rs570518709 as index SNPs. Similarly, the META-WD analysis identified rs138163369 and rs570518709. These two latter SNPs were not identified as index SNPs in the BOLT-LMM analysis, but their P-values were similar (rs138163369 – 4.90E-08, 2.28E-08, 1.22E-08; rs570518709 – 8.10E-08, 1.07E-09, 3.03E-09) (Additional file 3: Table S6). As another sensitivity analysis to test the reliability of the BOLT-LMM results, the effect sizes of all GCTA-COJO SNPs were compared in a pair-wise manner across the three GWAS. A regression line was fit through the scatter plots, showing a large degree of correlation between the BOLT-LMM effect sizes and the SNPTEST/META runs (META-WOD $R^2 = 0.91$, META-WD $R^2 = 0.93$) (Additional file 2: Fig. S7).

Two PLINK clumping analyses were performed on the filtered AFR_CAG summary statistics using the same clumping parameters on the well-known FUMA platform [43]. Here, 193 SNPs were identified as loci at the relaxed threshold of $r^2 = 0.6$ and 73 independent loci at the stringent threshold of $r^2 = 0.1$. Finally, 12 top loci were

identified at $r^2 = 0.001$ and a 10 Mb window, which are the very conservative MR clumping parameters [44, 45]. Furthermore, a FUMA analysis was run on the filtered AFR_CAG dataset for the top loci ($r^2 = 0.1$). This was done to visualise which genomic locations are affecting neutrophil count and if they are more likely to have a particular genetic function compared to the whole genome i.e. functional variants [65]. Seventeen genomic risk loci were identified (Additional file 2: Fig. S8A). The ANNOVAR analysis [66] showed evidence for changes in genetic function enrichment relative to all SNPs in the reference panel. In brief, seven genomic regions were enriched, all indicating an enrichment in genic rather than intergenic spaces (Additional file 2: Fig. S8B).

Next, we investigated the independent SNPs in the GWAS Catalog [67], as we aimed to see if they have been previously associated with WBC count or immunity. Here, SNPs predominantly showed associations with white blood cell count variation, further improving the reliability of the GWAS (Additional file 3: Table S7). We compared the AFR_CAG GWAS with a neutrophil count GWAS meta-analysis of Africans from UKBB and additional studies from Chen et al. [33], and found that 81.71% of the GWAS significant SNPs from Chen et al. were replicated (using the same covariates) in the AFR_CAG dataset ($P < 0.05$) (Additional file 3: Table S8). The Manhattan plots also visually showed a good degree of overlap (Additional file 2: Fig. S9), in contrast with a GWAS of neutrophil count in Europeans Additional file 2: Fig. S10) [38]. Finally, SNPs that were top loci at $r^2 = 0.1$ were investigated in the Astle et al. [38] and Chen et al. [33] summary statistics, as well as in the GWAS Catalog [67]. Nineteen genetic variants were not present in these three datasets, 7 of which were index SNPs (Table 2). All novel SNPs were rare if aligned to a European genomic reference panel.

Heritability analysis

Without adjusting for rs2814778, the genetic variance was estimated at 0.101 (10.1%) (SE = 0.018), and the phenotypic variance at 0.133 (13.3%) (SE = 0.003) with an analysis *P*-value of 2.29e−09. When adjusting for the ACKR1/Duffy SNP, the genetic variance was estimated at 0.050 (5%) (SE = 0.017), twice as low as in the previous analysis, and the phenotypic variance was estimated at 0.123 (12.3%) (SE = 0.002), with the analysis *P*-value of 1.36E−03 (Additional file 3: Table S9).

Mendelian randomization

Finally, a bi-directional MR was performed between neutrophil count and severe malaria. For the latter, we used summary statistics from the MalariaGEN study [37]. Only 3 SNPs were available to proxy for neutrophil count after data harmonization with the malaria dataset. For severe malaria as an exposure, 7 SNPs were available for overall severe malaria, 2 for CM and 3 for OTHER.

Table 2 Top loci not found in other studies. Only independent SNPs clumped at $r = 0.1$ are shown

SNP	CHR	BP (GRCh37)	EAF	r0.001 lead?	cojo_index	Novel?	Nearest gene	Type
rs28734019	1	90,800,573	0.998	Yes	Yes	Yes	RNU6-695P	Intergenic
rs61823703	1	159,542,164	0.987	No	No	Yes	OR10AE1P	Intergenic
rs539456851	1	158,731,459	0.982	No	No	Yes	OR6N1	Intergenic
rs371178711	1	158,186,653	0.969	No	No	Yes	RP11-404O13.5	Intergenic
rs146677619	1	158,995,984	0.991	No	No	Yes	IFI16	Intronic
rs11576058	1	161,111,446	0.979	No	No	Yes	UFC1	Intergenic
1:158777618_CT_C	1	158,777,618	0.050	No	No	Yes	OR10AA1P	Downstream
rs183362544	2	97,045,902	0.998	Yes	Yes	Yes	NCAPH	Intergenic
rs11422063	1	159,799,599	0.022	No	No	Yes	SLAMF8	Intronic
rs112483667	1	151,651,180	0.974	No	No	Yes	SNX27	Intronic
rs12406899	1	157,540,651	0.911	No	No	Yes	FCRL4	Intergenic
rs1103805	1	158,924,741	0.929	No	No	Yes	PYHIN1	Intronic
rs557482905	5	80,629,499	0.998	Yes	Yes	Yes	ACOT12	Intronic
rs527921556	6	160,605,701	0.996	Yes	Yes	Yes	SLC22A2	Intronic
rs10096834	8	116,281,087	0.573	Yes	No	Yes	TRPS1	Intergenic
rs140048432	9	17,700,893	0.996	Yes	Yes	Yes	SH3GL2	Intronic
rs530475031	12	48,810,860	0.998	Yes	Yes	Yes	C12orf54	Intronic
rs558204720	16	59,472,815	0.998	Yes	Yes	Yes	LOC105371298	Intronic
rs138163369	18	6,492,075	0.998	Yes	No	Yes	CTD-2124B20.2	Intergenic

EAF Effect allele frequency

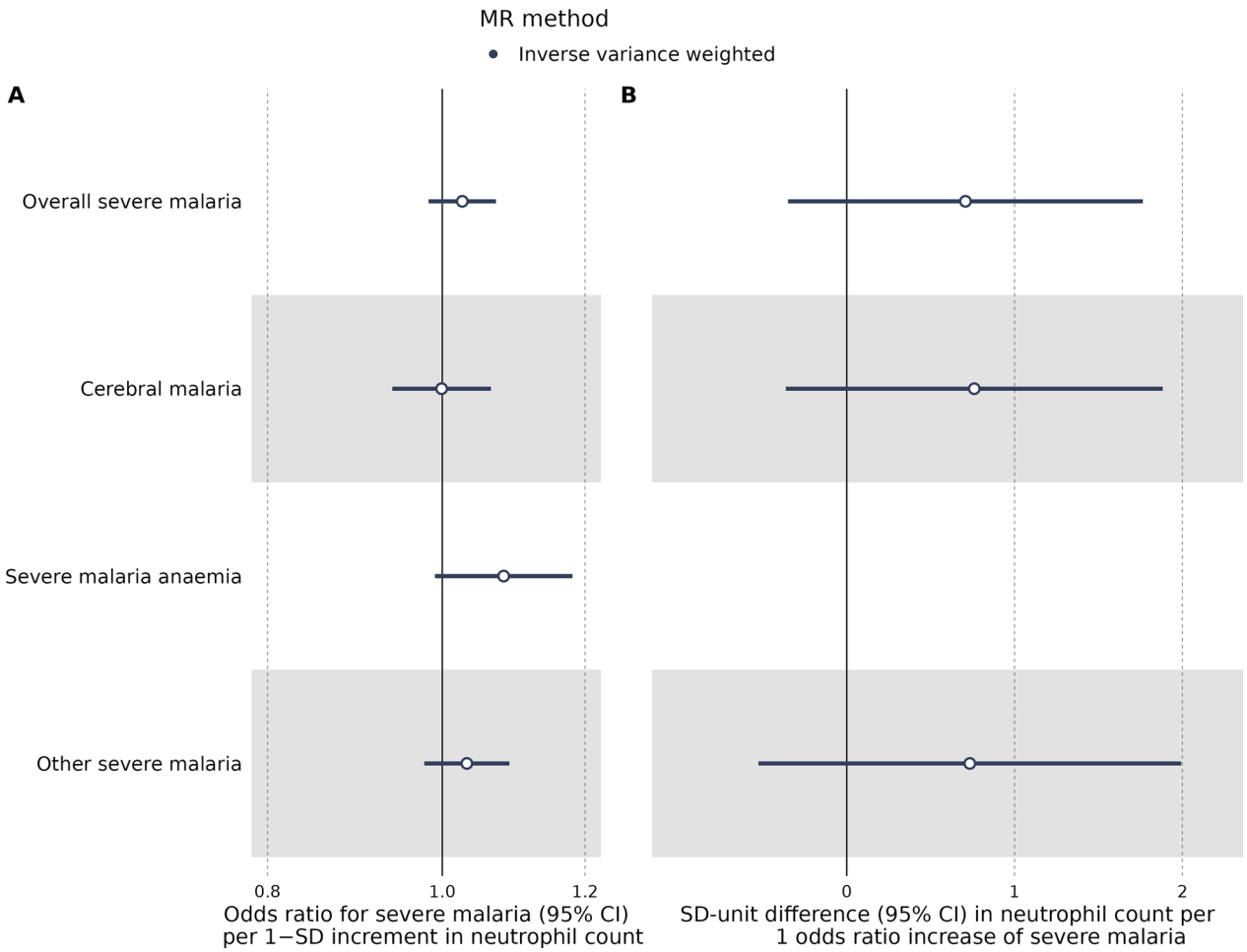

Fig. 4 Bi-directional Mendelian randomization. Forest plot of the IVW MR analysis with neutrophil count as an exposure (**A**) and severe malaria as an exposure (**B**). Overall severe malaria and its sub-phenotypes are listed on the y-axis, with the effect estimates on the x-axis. In the first instance, the MR results are interpreted as an OR increase severe malaria per 1-SD increase in neutrophil count, while in the latter as a 1-SD unit difference in neutrophil count per 1-OR

The MR analysis did not suggest an effect of increasing neutrophil count on CM risk (IVW OR: 1.00, 95% CI: 0.94–1.06; $P=0.98$. There was limited evidence of an effect of neutrophil count on overall severe malaria (IVW OR: 1.03, 95% CI: 0.98–1.07; P = 0.24), OTHER (IVW OR: 1.03, 95% CI: 0.98–1.09; $P=0.26$) and SMA (IVW OR: 1.08, 95% CI: 0.99–1.18; $P=0.08$), although the effect estimates were trending towards an increased risk of severity, particularly for SMA (Fig. 4A, Additional file 3: Table S10). When running the MR analysis in the other direction, there was little evidence of an effect of overall severe malaria (IVW OR: 2.03, 95% CI: 0.70 to 5.84; P = 0.19), CM (IVW OR: 2.14, 95% CI: 0.70–6.57; P = 0.18) and OTHER (IVW OR: 2.08, 95% CI: 0.59–7.34; $P=0.25$) on neutrophil count. However, there was a directional agreement in effect estimates towards an increase in neutrophil count (Fig. 4B, Additional file 3: Table S11). No SNPs passed the GWAS significance threshold for SMA, meaning this analysis could not be conducted.

A single-SNP MR analysis was performed to study the effect of each genetic variant on the outcome. For neutrophil count as the exposure, SNPs rs2325919 (proxy for rs2814778), rs7460611 (proxy for rs10096834), and rs144109344 were used. There was little evidence of an effect by any single SNP, although the general direction was towards an increased risk of severe malaria (Additional file 3: Table S11, Additional file 2: Fig. S12). The estimated conditional F-statistic for SNPs rs2325919, rs7460611 and rs144109344 were 182, 16 and 36 respectively. For severe malaria as an exposure, SNPs rs113892119, rs116423146, rs1419114, rs553707144, rs557568961, rs57032711, rs8176751 were used to proxy for overall severe malaria, rs113892119 and rs543034558 for CM, and rs113892119, rs116423146, rs557568961 for OTHER (Additional file 3: Supplementary Table S12,

Additional file 2: Supplementary Fig. S13). The estimated conditional F-statistic for SNPs rs113892119, rs116423146, rs1419114, rs553707144, rs557568961, rs57032711 and rs8176751 were 96, 32, 30, 38, 119, 32 and 44 respectively.

Discussion

Here, we conducted a GWAS of neutrophil count in individuals from the AFR CAG in UKBB. Seventy-three independent loci were identified, of which nineteen were novel and rare (when contrasted to a European reference panel). Ten index SNPs were found using the conservative GCTA-COJO approach, and another two through MR clumping. Moreover, BOLT-LMM was found to be reliable in conducting GWAS on UKBB participants of African ancestry. As a follow-up application example, we ran a MR analysis between neutrophil count and *P. falciparum* severe malaria.

An aim of our study was to assess whether BOLT-LMM could provide reliable results when performing a GWAS in people of non-European ancestry, such as those in the UKBB AFR CAG. In their meta-analysis of BCT in non-European datasets, Chen et al. used a linear model in PLINK to run their GWAS, restricting BOLT-LMM only to the European dataset [33]. Compared to our META-WD and META-WOD GWAS, the BOLT-LMM approach was more similar to that of Chen et al. conducted with a larger sample-size (N = 15,171). These findings indicate that a linear mixed model framework using a kinship matrix might reliably account for extensive population structure in a complex data set such as that seen in the African CAG used here. If this observation holds true this would be advantageous in identifying more causal ancestry-specific SNPs in future studies, as the power of BOLT-LMM scales with increasing GWAS sample-size [62].

Next, we found a marked difference between the genetic architecture of neutrophil count in people of African vs. European ancestry [38]. Interestingly, tissue expression for BCTs has been found to vary between ancestries as well [68], further showing the importance of conducting GWAS in diverse populations to improve the understanding of BCT biology. We investigated some of the GCTA-COJO index SNPs in relation to a biological mechanism that could explain how allele variation might affect neutrophil count levels in people of African ancestry. Not all index SNPs had evidence in the literature or online databases in terms of their potential biological function(s) and we have included only those SNPs for which information was available.

One such SNP is rs12747038, an index SNP located on chromosome 1 (1q21.1), was also identified by Chen et al. and Hu et al. to be associated with neutrophil count and they found a similar effect size to us (AFR_CAG: BETA = −0.22, *P*-value = 3.90e−09; Chen et al.: BETA = −0.31, *P*-value = 3e−20; Hu et al.: BETA = −0.21, *P*-value = 8e−36) [33, 69]. Interestingly, rs12747038 has a role as a splicing QTL (sQTL) i.e. affecting alternative splicing to make different protein isoforms [70], which can be more relevant mechanistically to a phenotype compared to expression data [71]. The strongest association as an sQTL was with *NBPF12* gene (NES = 0.49, *P*-value = 2.9e−9) in the thyroid. McCartney et al. had found that rs11239931, an sQTL for *NBPF12*, was also associated with a decrease in granulocyte count (BETA = −0.23, *P*-value = 4e−12) in people of African ancestry (N = 6152) [72]. *NBPF12* is part of the neuroblastoma breakpoint family, which has been associated with an array of traits, such as autism, psoriasis and various cancers [73].

The rs2814778 (chromosome 1q23.2) index SNP has been the most replicated genetic variant in people of African ancestry known to affect neutrophil count [33, 74–79], with the CC genotype (most common in Africans) associated with decreased neutrophil count [20]. The exact location of rs2814778 is inside a promoter upstream of the *ACKR1*/*DARC* (Atypical Chemokine Receptor 1/Duffy Antigen Receptor for Chemokines) gene [13]. The CC genotype inhibits the binding of the GATA transcription factor and therefore *ACKR1* expression in erythrocytes, preventing the production of a glycosylated transmembrane receptor [20]. This receptor is heavily involved in chemokine signalling, such as CXCL8 and CCL5 [13].

rs144109344 is an index variant on chromosome 2 (2q21.3), and its association was similar to that in the studies of Chen et al. and Soremekun et al. (N = 17,802 Africans): AFR_CAG BETA = −0.12, *P*-value = 3.10e−10; Chen BETA = −0.27, *P*-value = 3.39e−14; Soremekun BETA = −0.21, *P*-value = 2e−13) [33, 77]. Similarly, other SNPs mapping to the *DARS*/*CXCR4* (Aspartyl-TRNA Synthetase 1/C-X-C Motif Chemokine Receptor 4) genes have been associated with neutrophil and monocyte count [33, 38, 80–83]. CXCR4 is a chemokine receptor which binds to CXCL12 [84], and is known to regulate the release of neutrophils from the bone marrow during both homeostasis and infections [85]. Interestingly, CXCR4 has been implicated in *P. falciparum* pathogenesis. Macrophage migration inhibitory factor (MIF) can interact with CXCR4 to recruit neutrophils [86], and *P. falciparum* is known to also produce MIF (PfMIF) [87]. A previous laboratory study using both murine (*P berghei*) and human (*P falciparum*) models found impairment of the parasite liver-cycle in both genetically deficient and drug-targeted CXCR4 [88].

We note that the process of mapping SNPs to a biological function is a difficult process. This particularly applies to rare SNPs, such as those identified in our study, due to multiple factors (not limited to): rare SNPs being harder to detect in the first place [89], lack of information in databases on SNPs found in non-Euroepan population and especially Africa [90, 91], no straight forward way to find function (e.g. splicing vs non-coding) [92], context-dependent and interaction-dependent SNP effects along with small effects on multiple traits (pleiotropy) [93]. Therefore, the brief discussion above only serves as an inquiry into a possible explanation for the primary GWAS results.

Finally, in the MR analysis, there was limited evidence for an effect of increased circulating neutrophil on the risk of SM. The strongest effect was observed for the SMA sub-phenotype, however, this did not reach statistical significance. Interestingly, a recent report demonstrated an association between circulating neutrophil transcriptional activity and levels of anaemia in children with malaria [24], highlighting the need for further pathophysiological studies. We also observed little evidence for an effect of SM on neutrophil count. Previously, Band et al. performed a MR analysis between neutrophil count and *P. falciparum* SM [37], however, they used SNPs for neutrophil count generated from a GWAS in Europeans from UKBB [38], where they found no evidence of an effect on SM (AFR_CAG BETA = 0.03, P-value = 0.24; Band BETA = 0.00, P-value = 0.87) [37].

Our study has certain limitations. Firstly, the novel genetic variants identified here may be a result of Winner's curse [94]—SNPs can pass the "significance" threshold (commonly set at 5e-8 [95, 96]) in GWAS by chance in the first discovery study, which is then not replicated in subsequent studies [97, 98].

Secondly, only a limited number of instruments were available to proxy for neutrophil count in the MR analysis. Seven index SNPs had a very high effect allele count, which might have been fixed in the MalariaGEN study population and so could not be used in the MR analysis. The rs2814778 SNP (associated with the *ACKR1* gene) most likely had a very small allele frequency and might have been eliminated, although we were able to use another SNP in LD with it as a proxy. While LD proxies are useful, they can also come with the caveat of not precisely instrumenting the trait [36].

Finally, the most impactful limitation in this study is the small sample-size and hence statistical power. As mentioned previously, we have chosen to use BOLT-LMM here to best address the issues of a small sample-size and the presence of population structure. Current studies performed on people living in sub-Saharan Africa have been small [33, 75–77] compared to those currently being carried out in Europe, East Asia and the US [31, 83, 99]. Having a large-scale study akin to UKBB in sub-Saharan African would allow for finding common SNPs with smaller effect sizes that could be used reliably for polygenic risk score generation or MR analyses for complex traits such as neutrophil count.

In conclusion, our GWAS of neutrophil count in people from the UKBB African CAG identified several SNPs associated with neutrophil count. Additionally, our analyses would support a conclusion that linear mixed model frameworks can properly account for possible confounding due to population stratification in complex highly stratified sample populations. Finally, while the MR results were largely inconclusive, this only demonstrates the importance of conducting large-scale biobank studies in Africa.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-024-00585-w.

Additional file 1. Supplementary Methods on GWAS with references.

Additional file 2. Supplementary Figures S1–S13.

Additional file 3. Supplementary Tables ST1–ST12.

Acknowledgements

We are grateful to the UK Biobank study and its participants. This research has been conducted using the UK Biobank resource under Application 15825. We thank the Malaria GEN Network for their study and their participants.

Author contributions

AC, BA, DA, EEV, CJB and REM conceived the study. AC conducted the analysis. All authors contributed to the interpretation of the findings. AC, KF, EEV, CJB, DH and BA wrote the manuscript. All authors critically revised the paper for intellectual content and approved the final version of the manuscript.

Funding

AC acknowledges funding from grant MR/N0137941/1 for the GW4 BIOMED MRC DTP, awarded to the Universities of Bath, Bristol, Cardiff and Exeter from the Medical Research Council (MRC)/UKRI. NJT and REM acknowledge funding from the MRC (MC_UU_00011/1). NJT is the PI of the Avon Longitudinal Study of Parents and Children (MRC & Wellcome Trust 217065/Z/19/Z) and is supported by the University of Bristol NIHR Biomedical Research Centre (BRC-1215-2001). NJT and DAH acknowledge funding from the Wellcome Trust (202802/Z/16/Z). EEV, CJB, and NJT also acknowledge funding by the CRUK Integrative Cancer Epidemiology Programme (C18281/A29019). EEV and CJB are supported by Diabetes UK (17/0005587) and the World Cancer Research Fund (WCRF UK), as part of the World Cancer Research und International grant program (IIG_2019_2009). S.K. is supported by a United Kingdom Research and Innovation Future Leaders Fellowship (MR/T043202/1). JZ is supported by Shanghai Thousand Talents Program and the National Health Commission of the PR China. BA acknowledges funding from the Medical Research Council (MR/R02149x/1). The funders of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Availability of data and materials

Genetic data from UK Biobank were made available as part of project code 15,825. Analytical code is available on GitHub at https://github.com/andrewcon/AFR-GWAS-neutrophil.

Declarations

Ethics approval and consent to participate

UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382; 16/NW/0274) and was conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent before enrolment in the study.

Consent for publication

All authors consented to the publication of this work.

Competing interests

The authors declare no competing interests.

Author details

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [2]Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK. [3]Louisiana State University, Louisiana, USA. [4]School of Translational Health Sciences, University of Bristol, Bristol, UK. [5]Health Data Research UK, London, UK. [6]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [7]Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, People's Republic of China. [8]Shanghai National Clinical Research Center for Metabolic Diseases, Key Laboratory for Endocrine and Metabolic Diseases, National Health Commission, Shanghai, People's Republic of China. [9]Shanghai National Center for Translational Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, People's Republic of China. [10]Early Cancer Insitute, University of Cambridge, Cambridge, UK.

Received: 19 October 2023 Accepted: 12 February 2024
Published online: 15 March 2024

References

1. WHO Africa. World malaria report 2019. 2019.
2. Price RN, Commons RJ, Battle KE, Thriemer K, Mendis K. *Plasmodium vivax* in the era of the shrinking *P. falciparum* map. Trends Parasitol. 2020;36:560–70. https://doi.org/10.1016/j.pt.2020.03.009.
3. Moxon CA, Gibbins MP, McGuinness D, Milner DA, Marti M. New insights into malaria pathogenesis. Annu Rev Pathol. 2020;15:315–43. https://doi.org/10.1146/annurev-pathmechdis-012419-032640.
4. Knackstedt SL, Georgiadou A, Apel F, Abu-Abed U, Moxon CA, Cunnington AJ, et al. Neutrophil extracellular traps drive inflammatory pathogenesis in malaria. Sci Immunol. 2019;4:336. https://doi.org/10.1126/SCIIMMUNOL.AAW0336.
5. Sierro F, Grau GER. The ins and outs of cerebral malaria pathogenesis: immunopathology, extracellular vesicles, immunometabolism, and trained immunity. Front Immunol. 2019;10:830. https://doi.org/10.3389/fimmu.2019.00830.
6. Cela D, Knackstedt SL, Groves S, Rice CM, Kwon JTW, Mordmüller B, et al. PAD4 controls chemoattractant production and neutrophil trafficking in malaria. J Leukoc Biol. 2021. https://doi.org/10.1002/JLB.4AB1120-780R.
7. Kariuki SN, Williams TN. Human genetics and malaria resistance, vol. 139. Cham: Springer; 2020. https://doi.org/10.1007/S00439-020-02142-6.
8. Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. Br Med J. 1954;1:290–4. https://doi.org/10.1136/bmj.1.4857.290.
9. Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. The American Journal of Human Genetics. 2005;77:171–92. https://doi.org/10.1086/432519.
10. Ndila CM, Uyoga S, Macharia AW, Nyutu G, Peshu N, Ojal J, et al. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. Lancet Haematol. 2018;5:e333–45. https://doi.org/10.1016/S2352-3026(18)30107-8.
11. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN. Heritability of Malaria in Africa. PLoS Med. 2005;2:e340. https://doi.org/10.1371/journal.pmed.0020340.
12. Sakuntabhai A, Ndiaye R, Casadémont I, Peerapittayamonkol C, Rogier C, Tortevoye P, et al. Genetic determination and linkage mapping of *Plasmodium falciparum* malaria related traits in senegal. PLoS ONE. 2008;3:e2000. https://doi.org/10.1371/journal.pone.0002000.
13. Atallah-Yunes SA, Ready A, Newburger PE. Benign ethnic neutropenia. Blood Rev. 2019;37:100586. https://doi.org/10.1016/j.blre.2019.06.003.
14. Shoenfeld Y, Alkan ML, Asaly A, Carmeli Y, Katz M. Benign familial leukopenia and neutropenia in different ethnic groups. Eur J Haematol. 1988;41:273–7. https://doi.org/10.1111/j.1600-0609.1988.tb01192.x.
15. Rippey JJ. Leucopenia in West Indians and Africans. The Lancet. 1967;290:44. https://doi.org/10.1016/S0140-6736(67)90086-4.
16. Denic S, Showqi S, Klein C, Takala M, Nagelkerke N, Agarwal MM. Prevalence, phenotype and inheritance of benign neutropenia in Arabs. BMC Blood Disord. 2009;9:3. https://doi.org/10.1186/1471-2326-9-3.
17. Hsieh MM, Everhart JE, Byrd-Holt DD, Tisdale JF, Rodgers GP. Prevalence of neutropenia in the U.S. population: age, sex, smoking status, and ethnic differences. Ann Intern Med. 2007;146:486. https://doi.org/10.7326/0003-4819-146-7-200704030-00004.
18. Amulic B, Cazalet C, Hayes GL, Metzler KD, Zychlinsky A. Neutrophil function: from mechanisms to disease. Annu Rev Immunol. 2012;30:459–89. https://doi.org/10.1146/ANNUREV-IMMUNOL-020711-074942.
19. Reich D, Nalls MA, Kao WHL, Akylbekova EL, Tandon A, Patterson N, et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS Genet. 2009;5:360. https://doi.org/10.1371/journal.pgen.1000360.
20. Rappoport N, Simon AJ, Amariglio N, Rechavi G. The Duffy antigen receptor for chemokines, ACKR 1,–'Jeanne DARC'of benign neutropenia. Br J Haematol. 2019;184(4):497–507. https://doi.org/10.1111/bjh.15730.
21. Palmblad J, Höglund P. Ethnic benign neutropenia: a phenomenon finds an explanation. Pediatr Blood Cancer. 2018;65:e27361. https://doi.org/10.1002/pbc.27361.
22. Amulic B, Moxon CA, Cunnington AJ. A more granular view of neutrophils in malaria. Trends Parasitol. 2020;36(6):501–3.
23. Aitken EH, Alemu A, Rogerson SJ. Neutrophils and malaria. Front Immunol. 2018;9:3005. https://doi.org/10.3389/fimmu.2018.03005.
24. Anyona S, Cheng Q, Guo Y, Seidenberg P, Schneider K, Lambert C, et al. Entire expressed peripheral blood transcriptome in pediatric severe malarial anemia. Res Square 2023. https://doi.org/10.21203/RS.3.RS-3150748/V1.
25. Garcia-Senosiain A, Kana IH, Singh S, Das MK, Dziegiel MH, Hertegonne S, et al. Neutrophils dominate in opsonic phagocytosis of *P. falciparum* blood-stage merozoites and protect against febrile malaria. Commun Biol. 2021. https://doi.org/10.1038/S42003-021-02511-5.
26. Zelter T, Strahilevitz J, Simantov K, Yajuk O, Jensen AR, Dzikowski R, et al. Neutrophils impose strong selective pressure against PfEMP1 variants implicated in cerebral malaria. BioRxiv 2021:2021.05.09.443317. https://doi.org/10.1101/2021.05.09.443317.
27. Smith GD, Ebrahim S. "Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1–22. https://doi.org/10.1093/ije/dyg070.
28. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol. 2004;33:30–42. https://doi.org/10.1093/ije/dyh132.
29. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23:R89-98. https://doi.org/10.1093/hmg/ddu328.
30. Zheng J, Baird D, Borges M-C, Bowden J, Hemani G, Haycock P, et al. Recent developments in mendelian randomization studies. Curr Epidemiol Rep. 2017;4:330–45. https://doi.org/10.1007/s40471-017-0128-6.
31. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9. https://doi.org/10.1038/s41586-018-0579-z.
32. Skrivankova VW, Richmond RC, Woolf BAR, Davies NM, Swanson SA, Vanderweele TJ, et al. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. BMJ 2021;375. https://doi.org/10.1136/BMJ.N2233.
33. Chen MH, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667

individuals from 5 global populations. Cell. 2020;182:1198-1213.e14. https://doi.org/10.1016/j.cell.2020.06.045.
34. Constantinescu A-E, Mitchell RE, Zheng J, Bull CJ, Timpson NJ, Amulic B, et al. A framework for research into continental ancestry groups of the UK Biobank. Hum Genomics. 2022;16:1–14. https://doi.org/10.1186/S40246-022-00380-5.
35. Davey Smith G, Ebrahim S, Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1–22. https://doi.org/10.1093/ije/dyg070.
36. Hartwig FP, Davies NM, Hemani G, Smith GD. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. Int J Epidemiol. 2016;45:1717–26.
37. Network MGE. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nat Commun. 2019;10:5732. https://doi.org/10.1038/s41467-019-13480-z.
38. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167:1415-1429.e19. https://doi.org/10.1016/j.cell.2016.10.042.
39. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27:1–10. https://doi.org/10.1002/mpr.1608.
40. Sheard S, Nicholls R, Froggatt J. UK Biobank Haematology Data Companion Document n.d.
41. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.
42. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of AnTC, Consortium DialaGRAM (DIAGRAM), et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44:369-75. https://doi.org/10.1038/ng.2213.
43. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1–11. https://doi.org/10.1038/s41467-017-01261-5.
44. Choi KW, Stein MB, Nishimi KM, Ge T, Coleman JRI, Chen CY, et al. An exposure-wide and mendelian randomization approach to identifying modifiable factors for the prevention of depression. Am J Psychiatry. 2020;177:944–54. https://doi.org/10.1176/APPI.AJP.2020.19111158.
45. Noyce AJ, Bandres-Ciga S, Kim J, Heilbron K, Kia D, Hemani G, et al. The Parkinson's disease Mendelian randomization research portal. Mov Disord. 2019;34:1864. https://doi.org/10.1002/MDS.27873.
46. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of heritability for human height. Nat Genet. 2010;42:565. https://doi.org/10.1038/NG.608.
47. WHO Africa. Severe Malaria 2014https://doi.org/10.1111/tmi.12313
48. Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? Hum Genet. 2020;139:23–41. https://doi.org/10.1007/s00439-019-02014-8.
49. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190–1. https://doi.org/10.1093/BIOINFORMATICS/BTQ340.
50. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, et al. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet. 2008;40:198–203. https://doi.org/10.1038/NG.74.
51. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008;40:161–9. https://doi.org/10.1038/NG.76.
52. Hemani G, Tilling K, Davey SG. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. 2017;13:e1007081. https://doi.org/10.1371/JOURNAL.PGEN.1007081.
53. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408.
54. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. Stat Med 2016;35:1880–906. https://doi.org/10.1002/sim.6835.
55. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44:512–25. https://doi.org/10.1093/ije/dyv080.
56. Bowden J, Hemani G, Davey SG. Invited Commentary: Detecting individual and global horizontal pleiotropy in mendelian randomization—a job for the humble heterogeneity statistic? Am J Epidemiol. 2018;187:2681–5. https://doi.org/10.1093/AJE/KWY185.
57. Age groups - GOV.UK Ethnicity facts and figures n.d. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest (accessed August 17, 2022).
58. Health Survey for England: Weight n.d. http://healthsurvey.hscic.gov.uk/data-visualisation/data-visualisation/explore-the-trends/weight.aspx (accessed August 17, 2022).
59. Klein RJ. Power analysis for genome-wide association studies. BMC Genet. 2007;8:58. https://doi.org/10.1186/1471-2156-8-58.
60. Pierce BL, Burgess S. Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol. 2013;178:1177. https://doi.org/10.1093/AJE/KWT084.
61. Visscher PM, Hemani G, Vinkhuyzen AAE, Chen GB, Lee SH, Wray NR, et al. Statistical power to detect genetic (Co)variance of complex traits using SNP data in unrelated samples. PLoS Genet. 2014;10:e1004269. https://doi.org/10.1371/JOURNAL.PGEN.1004269.
62. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47:284–90. https://doi.org/10.1038/ng.3190.
63. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, et al. Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores. Nat Genet. 2022;54:450. https://doi.org/10.1038/S41588-022-01036-9.
64. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348-54.https://doi.org/10.1038/ng.548
65. Lichou F, Trynka G. Functional studies of GWAS variants are gaining momentum. Nature Commun. 2020;11:1–4. https://doi.org/10.1038/s41467-020-20188-y.
66. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164. https://doi.org/10.1093/NAR/GKQ603.
67. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D005-12. https://doi.org/10.1093/NAR/GKY1120.
68. Wen J, Xie M, Rowland B, Rosen JD, Sun Q, Chen J, et al. Transcriptome-wide association study of blood cell traits in african ancestry and hispanic/latino populations. Genes (Basel). 2021;12:1049. https://doi.org/10.3390/genes12071049.
69. Hu Y, Bien SA, Nishimura KK, Haessler J, Hodonsky CJ, Baldassari AR, et al. Multi-ethnic genome-wide association analyses of white blood cell and platelet traits in the Population Architecture using Genomics and Epidemiology (PAGE) study. BMC Genomics. 2021;22:1–11. https://doi.org/10.1186/S12864-021-07745-5.
70. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463:457. https://doi.org/10.1038/NATURE08909.
71. Garrido-Martín D, Borsari B, Calvo M, Reverter F, Guigó R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nature Commun. 2021;12:1–16. https://doi.org/10.1038/s41467-020-20578-2.
72. McCartney DL, Min JL, Richmond RC, Lu AT, Sobczyk MK, Davies G, et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. Genome Biol. 2021;22:25. https://doi.org/10.1186/S13059-021-02398-9.
73. Zhou F, Xing Y, Xu X, Yang Y, Zhang J, Ma Z, et al. NBPF is a potential DNA-binding transcription factor that is directly regulated by NF-κB. Int J Biochem Cell Biol. 2013;45:2479–90. https://doi.org/10.1016/J.BIOCEL.2013.07.022.

74. Moore CB, Verma A, Pendergrass S, Verma SS, Johnson DH, Daar ES, et al. Phenome-wide association study relating pretreatment laboratory parameters with human genetic variants in AIDS clinical trials group protocols. Open Forum Infect Dis. 2015;2:113. https://doi.org/10.1093/OFID/OFU113.
75. Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. Cell. 2019;179:984. https://doi.org/10.1016/J.CELL.2019.10.004.
76. Reiner AP, Lettre G, Nalls MA, Ganesh SK, Mathias R, Austin MA, et al. Genome-Wide Association Study of White Blood Cell Count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT). PLoS Genet. 2011;7:e1002108. https://doi.org/10.1371/journal.pgen.1002108.
77. Soremekun O, Soremekun C, Machipisa T, Soliman M, Nashiru O, Chikowore T, et al. Genome-wide association and Mendelian randomization analysis reveal the causal relationship between white blood cell subtypes and asthma in Africans. Front Genet. 2021;12:749415. https://doi.org/10.3389/FGENE.2021.749415/FULL.
78. Jain D, Hodonsky CJ, Schick UM, Morrison JV, Minnerath S, Brown L, et al. Genome-wide association of white blood cell counts in Hispanic/Latino Americans: the Hispanic Community Health Study/Study of Latinos. Hum Mol Genet. 2017;26:24. https://doi.org/10.1093/HMG/DDX024.
79. Legge SE, Pardiñas AF, Helthuis M, Jansen JA, Jollie K, Knapper S, et al. A genome-wide association study in individuals of African ancestry reveals the importance of the Duffy-null genotype in the assessment of clozapine-related neutropenia. Mol Psychiatry. 2019;24:328–37. https://doi.org/10.1038/s41380-018-0335-7.
80. Kachuri L, Jeon S, DeWan AT, Metayer C, Ma X, Witte JS, et al. Genetic determinants of blood-cell traits influence susceptibility to childhood acute lymphoblastic leukemia. Am J Hum Genet. 2021;108:1823–35. https://doi.org/10.1016/J.AJHG.2021.08.004.
81. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The polygenic and monogenic basis of blood traits and diseases. Cell. 2020;182:1214-1231.e11. https://doi.org/10.1016/J.CELL.2020.08.008/ATTACHMENT/347CE04A-7337-4664-BB5B-5ED6234B8F9E/MMC11.DOCX.
82. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genetics. 2021;53:1415–24. https://doi.org/10.1038/s41588-021-00931-x.
83. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genetics. 2018;50:390–400. https://doi.org/10.1038/s41588-018-0047-6.
84. De Filippo K, Rankin SM. CXCR4, the master regulator of neutrophil trafficking in homeostasis and disease. Eur J Clin Invest. 2018;48:12949. https://doi.org/10.1111/ECI.12949.
85. Eash KJ, Means JM, White DW, Link DC. CXCR4 is a key regulator of neutrophil release from the bone marrow under basal and stress granulopoiesis conditions. Blood. 2009;113:4711. https://doi.org/10.1182/BLOOD-2008-09-177287.
86. Weber C, Kraemer S, Drechsler M, Lue H, Koenen RR, Kapurniotu A, et al. Structural determinants of MIF functions in CXCR2-mediated inflammatory and atherogenic leukocyte recruitment. Proc Natl Acad Sci U S A. 2008;105:16278. https://doi.org/10.1073/PNAS.0804017105.
87. Ghosh S, Jiang N, Farr L, Ngobeni R, Moonah S. Parasite-produced MIF cytokine: role in immune evasion, invasion, and pathogenesis. Front Immunol. 2019;10:1995. https://doi.org/10.3389/FIMMU.2019.01995/BIBTEX.
88. Bando H, Pradipta A, Iwanaga S, Okamoto T, Okuzaki D, Tanaka S, et al. CXCR4 regulates *Plasmodium* development in mouse and human hepatocytes. J Exp Med. 2019;216:1733.
89. Young KL, Fisher V, Deng X, Brody JA, Graff M, Lim E, et al. Whole-exome sequence analysis of anthropometric traits illustrates challenges in identifying effects of rare genetic variants. Hum Genet Genomics Adv. 2023;4:100163. https://doi.org/10.1016/J.XHGG.2022.100163.
90. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177:26–31. https://doi.org/10.1016/J.CELL.2019.02.048.
91. Cooke Bailey JN, Bush WS, Crawford DC. Editorial: the importance of diversity in precision medicine research. Front Genet. 2020;11:875. https://doi.org/10.3389/fgene.2020.00875.
92. Mousas A, Ntritsos G, Chen MH, Song C, Huffman JE, Tzoulaki I, et al. Rare coding variants pinpoint genes that control human hematological traits. PLoS Genet. 2017;13:e1006925. https://doi.org/10.1371/JOURNAL.PGEN.1006925.
93. Fadason T, Farrow S, Gokuladhas S, Golovina E, Nyaga D, O'Sullivan JM, et al. Assigning function to SNPs: considerations when interpreting genetic variation. Semin Cell Dev Biol. 2022;121:135–42. https://doi.org/10.1016/j.semcdb.2021.08.008.
94. Thaler RH. Anomalies: the winner's curse. J Econ Perspect. 1988;2:191–202. https://doi.org/10.1257/JEP.2.1.191.
95. Panagiotou OA, Ioannidis JPA, Hirschhorn JN, Abecasis GR, Frayling TM, McCarthy MI, et al. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol. 2012;41:273–86. https://doi.org/10.1093/IJE/DYR178.
96. Chen Z, Boehnke M, Wen X, Mukherjee B. Revisiting the genome-wide significance threshold for common variant GWAS. G3 Genes|Genomes|Genetics 2021;11:jkaa056. https://doi.org/10.1093/G3JOURNAL/JKAA056.
97. Kraft P. Curses: Winner's and otherwise—In genetic epidemiology. Epidemiology. 2008;19:649–51. https://doi.org/10.1097/EDE.0B013E318181B865.
98. Ioannidis JPA. Why most discovered true associations are inflated. Epidemiology. 2008;19:640–8. https://doi.org/10.1097/EDE.0B013E31818131E7.
99. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70:214–23. https://doi.org/10.1016/J.JCLINEPI.2015.09.016.

Publisher's Note