Study Descriptions

Age, Gene/Environment Susceptibility (AGES)-Reykjavik Study
The AGES-Reykjavik Study is a single center prospective cohort study based on the Reykjavik Study. The Reykjavik Study was initiated in 1967 by the Icelandic Heart Association to study cardiovascular disease and risk factors. The cohort included men and women born between 1907 and 1935 who lived in Reykjavik at the 1967 baseline examination. Re-examination of surviving members of the cohort was initiated in 2002 as part of the AGES-Reykjavik Study.[1]

Atherosclerosis Risk in Communities (ARIC) Study
The ARIC study has been described in detail previously.[2] Men and women aged 45-64 years at baseline were recruited from four communities: Forsyth County, North Carolina; Jackson, Mississippi; Minneapolis, Minnesota; and Washington County, Maryland. A total of 15,792 individuals, predominantly White and African American, participated in the baseline examination in 1987-1989, with three additional triennial follow-up examinations and a fifth exam in 2011-2013.

*The Mount Sinai Institute for Personalized Medicine Bio*Me *Biobank (Bio*Me*)*
The Bio*Me* Biobank Program is an ongoing, prospective, hospital- and outpatient- based population research program operated by The Charles Bronfman Institute for Personalized Medicine (IPM) at Mount Sinai. Bio*Me* has enrolled over 33,000 participants from over 30 clinical care sites between September 2007 and April 2015. Bio*Me* is an Electronic Medical Record (EMR)-linked biobank that integrates research data and clinical care information for consented patients at The Mount Sinai Medical Center, which serves diverse local communities of upper Manhattan with broad health disparities. IPM Bio*Me* populations include 25% of African American ancestry (AA), 36% of Hispanic Latino ancestry (HL), 30% of white European ancestry (EA), and 9% of other ancestry. Information on anthropometrics, demographics, blood cell traits was derived from participants EMR.

Cardiovascular Health Study (CHS)
The CHS has been described in detail previously.[3] The CHS is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers. The original predominantly Caucasian cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists, and an additional 687 African-Americans were enrolled subsequently for a total sample of 5,888. DNA was extracted from blood samples drawn on all participants at their baseline examination in 1989-90.

The NHLBI Family Heart Study (FamHS)
The FamHS (https://dsgweb.wustl.edu/fhscc/) began in 1992 with the ascertainment of 1,200 families with approximately 6,000 subjects, half randomly sampled and half selected because of an excess of CHD or risk factor abnormalities as compared with age- and sex-specific population rates.[4] The participants attended a clinic visit between the years 1994-1996 and a broad range

of phenotypes were assessed in the general domains of coronary heart disease (CHD), atherosclerosis, hematologic markers, lifestyle, medical history and medication use, among other cardiovascular risk factors. Approximately 8 years later, 2,756 European American (EA) subjects belonging to the 510 largest and most informative pedigrees were invited for a second clinical exam (2002-04). The most important CHD risk factors were measured again and medical history and medication use were updated.

Framingham Heart Study (FHS)
The FHS is a three generational prospective cohort that has been described in detail previously.[5,6] Individuals were initially recruited in 1948 in Framingham, USA to evaluate cardiovascular disease risk factors. The second generation cohort (5,124 offspring of the original cohort) was recruited between 1971 and 1975. The third generation cohort (4,095 grandchildren of the original cohort) was collected between 2002 and 2005.

Health2006
Health2006 is a population-based epidemiological study of general health, diabetes and cardiovascular disease of individuals aged 18-74 years.[7] In addition to fasting biochemistry these individuals have had step test to objectively quantify physical fitness. Health2006 was conducted at the Research Centre for Prevention and Health in Glostrup, Denmark.

Health2008
The Health2008 is an extension of the Health2006 study, where a random sample of the general population aged 30 to 60 years from the same regional areas in Copenhagen County was drawn from the Civil Registration System. Essentially the same examinations as in Health2006 were made (anthropometrics, basal biochemistry, health and lifestyle questionnaire), however, extended with a 2 hour OGTT, a bicycle test to estimate cardio-respiratory fitness as maximum oxygen uptake (VO2 max), Acti-Heart® for objective measurement of physical activity, and a test of intelligence.

Health ABC (HABC) study
Health ABC is a longitudinal, prospective cohort of well-functioning older men and women recruited from Memphis, TN and Pittsburgh, PA using Medicare records.[8] Recruitment began in 1997-1998 when participants were between 70 and 79 years of age and entry into study was dependent on participants' ability to walk one-quarter mile and climb 10 steps without difficulty. Thirty-three percent of the men are African-Americans as are 46% of the women.

Jackson Heart Study (JHS)
The JHS is a large single-site, prospective, epidemiologic investigation of cardiovascular disease among African American adults in the three counties (Hinds, Madison, and Rankin) that comprise the Jackson, Mississippi metropolitan area.[9] The Jackson Heart Study involves a collaboration among three institutional partners, the Jackson community, and the National Institutes of Health to discover best practices for eliminating health disparities.

The Lothian Birth Cohorts of 1921 and 1936 (LBC2921/LBC1936)
The Lothian Birth Cohorts are follow-up studies of the Scottish Mental Surveys of 1932 and 1947.[10-12] The surveys had, respectively, tested the intelligence of almost every child born in 1921 or 1936 and attending school in Scotland in the month of June in those years. Therefore, tracing, recruiting and re-testing people who had taken part in the Surveys offered a rare opportunity to examine the distribution and causes of cognitive ageing across most of the human life course.

Multi-Ethnic Study of Atherosclerosis (MESA)
The Multi-Ethnic Study of Atherosclerosis[13] is a National Heart, Lung and Blood Institute-sponsored, population-based investigation of subclinical cardiovascular disease and its progression. A total of 6,814 individuals, aged 45 to 84 years, were recruited from six US communities (Baltimore City and County, MD; Chicago, IL; Forsyth County, NC; Los Angeles County, CA; New York, NY; and St. Paul, MN) between July 2000 and August 2002. Participants were excluded if they had physician-diagnosed cardiovascular disease prior to enrollment, including angina, myocardial infarction, heart failure, stroke or TIA, resuscitated cardiac arrest or a cardiovascular intervention (e.g., CABG, angioplasty, valve replacement, or pacemaker/defibrillator placement). Pre-specified recruitment plans identified four racial/ethnic groups (White European-American, African-American, Hispanic-American, and Chinese-American) for enrollment, with targeted oversampling of minority groups to enhance statistical power. MESA consists of 38% European descent, 28% African-American, 23% Hispanic and 11% Asian, primarily of Chinese descent participants. Only European, African, and Hispanic American participants were included in this study.

Rotterdam Study (RS)
The Rotterdam Study is an ongoing prospective population-based cohort study, focused on chronic disabling conditions of the elderly. The study comprises an outbred ethnically homogenous population of Dutch Caucasian origin. The rationale of the study has been described in detail elsewhere.[14-16] In summary, 7,983 men and women aged 55 years or older, living in Ommoord, a suburb of Rotterdam, the Netherlands, were included in the baseline exam.

Utrecht Health Project (UHP)
The Utrecht Health Project (in Dutch: Leidsche Rijn GezondheidsProject, LRGP) is a population based cohort study of the inhabitants of Leidsche Rijn, a district in the west of the city Utrecht.[17] Residents of Leidsche Rijn who register at one of the academic healthcare centers receive a written invitation to participate in the project from their general practitioner.

Women's Health Initiative (WHI)
WHI is one of the largest (n=161,808) studies of women's health ever undertaken in the United States.[18] There are two major components of WHI: (1) a Clinical Trial (CT) that enrolled and randomized 68,132 women ages 50 – 79 into at least one of three placebo-control clinical trials (hormone therapy, dietary modification, and calcium/vitamin D); and (2) an Observational Study (OS) that enrolled 93,676 women of the same age range into a parallel prospective cohort

study. A diverse population including 26,045 (17%) women from minority groups were recruited from 1993-1998 at 40 clinical centers across the U.S. The design has been published.[19,20] For the CT and OS participants enrolled in WHI and who had consented to genetic research, DNA was extracted by the Specimen Processing Laboratory at the Fred Hutchinson Cancer Research Center (FHCRC) using specimens that were collected at the time of enrollment in to the study (between 1993 and 1998).

The Cardiovascular Risk in Young Finns Study (YFS)

The Young Finns Study is a population-based follow up-study started in 1980.[21] The main aim of the YFS is to determine the contribution made by childhood lifestyle, biological and psychological measures to the risk of cardiovascular diseases in adulthood. In 1980, over 3,500 children and adolescents all around Finland participated in the baseline study. The follow-up studies have been conducted mainly with 3-year intervals. The latest 30-year follow-up study was conducted in 2010-11 (ages 33-49 years) with 2,063 participants. The study was approved by the local ethics committees (University Hospitals of Helsinki, Turku, Tampere, Kuopio and Oulu) and was conducted following the guidelines of the Declaration of Helsinki.

The Peking University – University of Michigan Study of Atherosclerosis (PUUMA) Beijing Shijingshan Cohort

The Peking University – University of Michigan Study of Atherosclerosis (PUUMA) is based upon the enrollment of individuals at two hospitals in the Peking University Health Science system: PKU First Hospital and PKU Third Hospital. There were several sources of samples, including the cardiac catheterization laboratories of the hospitals and a community-based enrollment in Beijing Shijingshan district. The analyses reported in this study were based upon the community based enrollment of 5,274 unrelated individuals (confirmed by exome wide genotypes) which represents a population sample of Han Chinese from Beijing. Genotyping was performed using the Illumina Exome Plus chip with additional custom content based upon sequencing of additional individuals of Asian ancestry. QC was performed as described elsewhere.[22] Association analyses were implemented in PLINK, and genomic control was applied to the final association statistics. For blood count phenotypes, complete blood counts were performed on peripheral blood obtained through standard venipuncture techniques using an automated hematology analyzers (BC-3000 manufactured by Shenzhen Mindray Bio-Medical Electronics Co., Ltd., China).

Expression quantitative trait loci (eQTL) analysis methods

We identified proxy SNPs in high linkage disequilibrium (r2>0.8) with associated index SNPs in 3 HapMap builds and 1000 Genomes with SNAP[23] for European-ancestry populations. The SNPs with Hispanic-ancestry and African-ancestry specific signals were queries for proxies in corresponding HapMap populations. SNP rsIDs were searched for primary SNPs and LD proxies against a collected database of expression SNP (eSNP) results. The collected eSNP results met criteria for statistical thresholds for association with gene transcript levels as described in the original papers.

Blood cell related eQTL studies included fresh lymphocytes[24] , fresh leukocytes[25], leukocyte samples in individuals with Celiac disease[26], whole blood samples[15,27-36], lymphoblastoid cell lines (LCL) derived from asthmatic children[37,38], HapMap LCL from 3 populations[39], a separate study on HapMap CEU LCL[40], additional LCL population samples[41-45], CD[40]+ B cells[46], primary PHA-stimulated T cells[42,44] CD4+ T cells[47], peripheral blood monocytes[46,48,49] CD11+ dendritic cells before and after Mycobacterium tuberculosis infection[50] . Micro-RNA QTLs[51] and DNase-I QTLs[52] were also queried for LCL.

Non-blood cell tissue eQTLs searched included omental and subcutaneous adipose[27,30,43,53] stomach[53], endometrial carcinomas[54], ER+ and ER- breast cancer tumor cells[55], brain cortex[48,56,57] pre-frontal cortex[58-60], parietal lobe[61] frontal cortex[59,62] temporal cortex[57,59,62], hippocampus[59], thalamus[59], pons[62], cerebellum[57,59,61,62], 3 additional large studies of brain regions including prefrontal cortex, visual cortex and cerebellum, respectively[63], liver[53,64-66], osteoblasts[67], intestine[68], skeletal muscle[69], breast tissue (normal and cancer)[70], lung[27,71,72], skin[27,43,73], primary fibroblasts[42,44], sputum[74], and heart tissue from left ventricles[27] and left and right atria[75]. Micro-RNA QTLs were also queried for gluteal and abdominal adipose[76].

Additional eQTL data was integrated from online sources including ScanDB, the Broad Institute GTex browser, and the Prichard Lab (eqtl.uchicago.edu). Cerebellum, parietal lobe and liver eQTL data was downloaded from ScanDB and cis-eQTLs were limited to those with $P<1.x10^{-6}$ and trans-eQTLs with $P<5.0x10^{-8}$. The top 1000 eQTL results were downloaded from the GTex Browser at the Broad Institute for 9 tissues on 11/26/2013: thyroid, leg skin (sun exposed), tibial nerve, tibial artery, skeletal muscle, lung, heart (left ventricle), whole blood, and subcutaneous adipose5. All GTex results had associations with $P<8.4x10^{-7}$.

Genes and variants previously known to be associated with hematologic traits

Many of the associations observed in this meta-analysis were common coding single nucleotide polymorphisms (SNPs) well-established to be associated with hematologic traits (e.g., *TMPRSS6* Val736Ala, *HFE* Cys282Tyr and His63Asp, *SH2B3* Arg262Trp, and the *G6PD* A-variant) or previously reported non-coding SNPs from GWAS for erythrocyte or WBC traits that were included on the exome chip (e.g., *DARC*, *CSF3-PSMD3*, *IL1F10*, *IL1RL1*, *IRF8, UB2L3, ATR, MYB-HBS1L, CCND3, CITED2, PRKCE, KIT, BCL9, TFR2, RCL1, PRKAG2*).

Sensitivity analysis for ultra-rare variants with large effect

We lowered the MAC threshold to MAC≥10 (MAF>0.0001) to determine if there were any ultra-rare variants with large effect sizes. We identified three additional novel associations with MCH (exm1244478/rs150764393 in CCDC135, exm373009/rs553434720 in ATP13A5, and exm1645679/rs149819079 in NHSL2 which was also associated with MCV) that were not identified by MAC≥40. While all of these variants showed consistent effects across multiple cohorts and races (**Supplementary Figure 8**), none of them replicated in WHI. However, we lacked power to replicate findings with such low minor allele frequency.

Dissecting Gene Based Tests

Several genes were only significant because critical variants that are common in one population but absent in another averaged out to below the MAF<0.05 threshold for inclusion in the trans-ethnic analysis. For instance, the *HFE* gene is only significant because one of the canonical hemochromatosis variants (Cys282Tyr) is MAF>0.05 in EAs, it is MAF<0.05 in the combined EA+AA+HA analyses and is therefore able to drive the finding.

Similarly, G6PD is associated with several traits at the aggregate gene level, but mainly because the missense variant with a strong effect that is common in AAs (rs1050828; AA MAF=0.11; AA $p=4.4x10^{-19}$) is brought below the 5% MAF threshold when combined with EAs (EA MAF<0.0001; EA p=0.75) and HAs (HA MAF=0.02; HA p=0.07). Of note, there is also a rare missense variant in G6PD nominally associated with Hb in EAs (exm1666749/rs5030868; EA MAF=0.0007; EA $p=4.7x10^{-5}$).

Other genes were driven by a single rare variant, such as the recently reported[77] association between *EPO* and lower Hb/Hct which is driven by a rare (MAF=0.001) missense variant rs62483572 (Ala70Asn). There was no LD (r^2=0.0) between rs62483572 and previously known GWAS SNPs in the *TFR2-EPO* region associated with hematocrit, MCV and RBC (index SNPs rs7385804, rs7786877 and rs2075671, respectively) (**Supplementary Table 10**). Moreover, conditional analysis demonstrated that the magnitude of the p-values for each marker remained the same even when the other variant was included as a covariate in the analysis.

Conditional analyses

Long range LD in the HFE region: A missense mutation in *MOG* (rs3130253) and a variant in the intergenic region 1 Mb downstream of *HFE* (rs13194491) were conditionally dependent on the canonical *HFE* Cys282Tyr and His63Asp hemochromatosis variants. In contrast, *ZSCAN31* appears to represent an independent association 2 Mb downstream of *HFE*. In the region around *HFE* on chromosome 6, multiple variants are associated with Hb, Hct, MCV, MCH, and MCHC in EAs only. The index SNP with the most extreme p-value in this meta-analysis was the canonical His63Asp variant (exm521702), which is independent from the other hemochromatosis variant Cys282Tyr, which was the strongest finding identified through GWAS (rs1800562). Linkage disequilibrium between the two variants is low (r^2=0.01) and conditioning on one strengthens the association of the other (**Supplementary Table 10**).

Independence of ANK1 variant from the GWAS variant: In the *ANK1* region, we identified a low frequency variant (Ala1462Val /rs34664882; EA MAF=0.029; AA MAF=0.015; HA MAF=0.013) associated with MCHC that is independent of the original GWAS variant (rs4737009; 1000G CEU MAF=0.27; ARIC EA MAF=0.24). The two variants are not in linkage disequilibrium (r^2=0.007), and conditioning on rs4737009 did not attenuate the signal in the ARIC EA sample (**Supplementary Table 10**).

Pleiotropy at the *HBS1L/MYB* locus: Common variation at the *HBS1L/MYB* locus on chromosome 6 has previously been robustly associated with erythrocyte traits, and here we have identified association with total WBC. The novel WBC associated SNP, rs7776054 (MAF=0.20-0.26), is in perfect LD with the previously known erythrocyte trait-associated variant in this region (rs7775698), demonstrating pleiotropy of this region for multiple blood cell lineages (**Supplementary Table 10**).

Independence of *ITFG3* variant from GWAS variants near *HBA1* and *HBA2*: The GWAS variants associated with MCV in the region containing *HBA1-HBA2* and *ITFG3* were not on the exome chip and required the merging of genetic datasets. Since the associations are strongest in AAs, the ARIC AA exome chip was merged with the corresponding Affymetrix 6.0 data. The index variant from the exome chip is a low frequency missense variant in *ITFG3* (Asp534Asn; EA MAF=0.0001; AA MAF=0.017; HA MAF=0.0076) and in the in ARIC AAs alone was strongly associated with MCV ($p=9x10^{-16}$) and MCHC ($p=4x10^{-5}$), but not with Hb (p=0.27). None of the SNPs on the Affymetrix 6.0 array that were index SNPs from previous GWAS meta-analyses (rs11248850, rs1211375, rs13335629) were associated with these traits in the ARIC AA sample alone (all p>0.60), indicating that these variants are not tagging the rare variant in *ITFG3*. Asp534Asn in *ITFG3* was one of several lower frequency missense variants nominally associated with Hb levels in a prior study;[78] however, this is the first time the variant was significantly associated with an erythrocyte-related trait.

Additional rare variation near *HBA1* and *HBA2* significantly associated with RBC traits: In addition to the Asp534Asn variant in *ITFG3*, several other missense and loss of function variants were significantly associated with RBC traits. All of them were found predominantly in AAs and HAs (EA MAF<0.0001 for all five variants). A variant in *MRPL28* (Gly47Glu/rs80158709, AA MAF=0.028, HA MAF=0.013) was associated with MCV ($p=3.4x10^{-}$

[12]), MCHC ($p=1.3x10^{-8}$) and MCH ($p=1.2x10^{-7}$). A variant in *NARFL* (Val38Met/rs8045850, AA MAF=0.07, HA MAF=0.024) was associated MCV ($p=9.9x10^{-8}$). A stop-gain in *RGS11* (Arg112Stop/rs149201684, AA MAF=0.006, HA MAF=0.002) was associated with MCV ($p=4.3x10^{-9}$). A variant in *TMEM8A* (Pro445Leu/rs144557907, AA MAF=0.003, HA MAF=0.001) was associated with MCHC ($p=1.8x10^{-8}$). And a variant in *TPSD1* (Asp225His/rs145927179, AA MAF=0.009, HA MAF=0.003) was associated with MCH ($p=2.7x10^{-9}$). Linkage disequilibrium between these variants, as well as between these variants and the *ITFG3* variant, was low ($r^2<0.02$), with the exception of *MRPL28*-Gly47Glu and *ITFG3*-Asp534Asn (D'=0.59; $r^2=0.21$). Conditioning on *ITFG3*-Asp534Asn in the ARIC AA sample abrogated the association between MCV and *MRPL28*-Gly47Glu (from p=0.0004 to p=0.76), indicating that *MRPL28*-Gly47Glu likely does not have an independent effect.

***CEP89* with Hb/Hct and kidney function (eGFR)**: Conditional analyses were performed within the ARIC EA individuals. rs4805834 in CEP89 was significantly associated with Hb ($p=2.3x10^{-6}$), with Hct ($p=7.7x10^{-6}$), with eGFR-Cr (p=0.002), and with eGFR-Cys (p=0.01). Hb and Hct were highly correlated (Pearson correlation=0.96), eGFR-Cr and eGFR-Cys were moderately correlated (Pearson correlation=0.57), and Hb was only slightly negatively correlated with either eGFR measurement (Pearson correlation=-0.10). When including eGFR in the model as a covariate, the SNP was still associated with Hb and Hct at a similar level ($p=2.8x10^{-6}$). Similarly, when Hb or Hct were included as covariates, the SNP also remained associated with eGFR-Cr (p=0.002) and with eGFR-Cys (p=0.04).

***SHROOM3* with Hb/Hct and kidney function (eGFR)**: Conditional analyses were also attempted for SHROOM3 using ARIC EA individuals. However, rs13146355 was only associated with eGFR-Cr ($p=1.5x10^{-5}$) and eGFR-Cys ($p=8.1x10^{-7}$) and not with Hb (p=0.40) or Hct (p=0.23) in this sample. Similar to *CEP89*, conditioning on Hb did not affect the associations with eGFR-Cr ($p=1.6x10^{-5}$) and eGFR-Cys ($p=1.3x10^{-6}$). And conditioning on eGFR did not affect the associations with Hb (p=0.44) or Hct (p=0.26).

***APOE* with RDW and LDL:** While the ε2 variant that we found to be associated with RDW is on the array, the ε4 allele is not. There is another nearby SNP on the array, rs4420638, that tags the ε4 allele reasonably well (D'=0.86, $r^2=0.60$) and has been found to be associated many traits, including coronary artery disease.[79] Unfortunately, this variant did not pass quality control in most studies and was not analyzed. Manual reclustering in the ARIC samples led to high confidence genotypes that maintained Hardy Weinberg equilibrium and were found to be nominally associated with RDW in the expected, opposite direction from the ε2 variant (p=0.02). In the ARIC study, the ε2 variant accounted for 2.8% of the variance in LDL (after accounting for age, race, and HT meds) and 1.4% of RDW levels, while the ε4 variant accounted for 0.4% of variance of LDL and 0.2% of variance of RDW. When both variants were included in the same model, the ε4 variant remained significant ($p=9x10^{-9}$) while the ε4 tag did not (p=0.13). The RDW and LDL traits were not correlated in the 3,085 EAs that had both traits assayed (Pearson correlation=-0.0002), so the RDW associations remained robust even when LDL was included as a covariate in the model.

Pleiotropy in the associated loci

In addition to pleiotropy between our novel findings and the known associations with kidney function (*CEP89* and *SHROOM3*) and with dementia and dyslipidemia (*APOE*), we also identified variants with pleiotropy across multiple blood cell lineages. The *HBS1L/MYB* locus, previously associated with erythrocyte traits,[80] was significantly associated with total WBC. Similarly, a 1.1Mb haplotype on chromosome 12q24 in EA participants containing *SH2B3*, previously associated with erythrocyte traits and eosinophil count, was associated with lymphocyte count. At each trait-locus meeting our defined exome-wide significance criteria (**Supplementary Table 6**), we systematically searched for evidence of modest association ($p<1x10^{-4}$) with other traits (**Supplementary Table 13**). Subthreshold associations were noted across blood cell lineages near *HFE* (MCV $p=1.9x10^{-7}$; TotalWBC $p=6.2x10^{-5}$), *ABO* (Hb $p=2.8x10^{-12}$; Monocytes/TotalWBC $p=1.4x10^{-5}$), *SPTA1* (MCHC $p=8.5x10^{-11}$; TotalWBC $p=6.7x10^{-5}$), and *TRIM58* (Hct $p=4.5x10^{-12}$; Neutrophils/Monocytes $p=8.0x10^{-6}$).

Differences in continental ancestry

Participants in the IPM BioMe and MESA cohorts who self-identify as Hispanic/Latino reported to be of Puerto Rican (39%), Dominican (23%), Central/South American (17%), Mexican (5%) or other Hispanic (16%) ancestry. Principal component analysis confirmed that this group was highly diverse with respect to how much of their DNA originated from the African, European and [Native] American continents. These principal components were used as covariates in the analysis to account for genetic diversity among this group; nevertheless, this diversity, coupled with low sample number of Hispanic/Latino participants, hampers our ability to identify new associations.

The only markers that replicated in PUUMA were the missense variant in *S1PR4* and the *CEP89* variant associated with Hb. This is not entirely surprising given the difference in continental ancestry of the discovery cohorts. Those variants that change an amino acid, like the one in *S1PR4*, have a distinct mechanism of action and that effect is likely to be preserved among individuals from different continents. But an association will only be seen if the variant is present in those distinct populations, and that was not the case for the rare variants in IQCJ, SEC24D and IL17RA. Conversely, the same effect is often not preserved for variants that are intragenic or intronic, including those variants that were only on the exome chip because they were identified through prior GWAS studies, generally of European Americans. It is likely that many of these GWAS variants are not the functional themselves but rather in LD with the functional variant. If the true functional variant is at a lower frequency or absent in a different population, or if the LD patterns are different, then we would not expect to see the same association in such a population. This may explain why the variants that replicated in WHI (*SHROOM3*, *FADS2*, *MYB/HBS1L, and NLRP12)* did not replicate in PUUMA. Finally, we could not attempt replication of *APOE* and *IRF1* because the corresponding traits (RDW and Eosinophils) were not available in PUUMA.

References

1. Harris, T.B. *et al.* Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am. J. Epidemiol.* **165**, 1076-87 (2007).
2. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *American journal of epidemiology* **129**, 687-702 (1989).
3. Fried, L.P. *et al.* The Cardiovascular Health Study: design and rationale. *Annals of epidemiology* **1**, 263-76 (1991).
4. Higgins, M. *et al.* NHLBI Family Heart Study: objectives and design. *Am. J. Epidemiol.* **143**, 1219-28 (1996).
5. Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M. & Castelli, W.P. The Framingham Offspring Study. Design and preliminary data. *Preventive medicine* **4**, 518-25 (1975).
6. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. & Stokes, J., 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. *Annals of internal medicine* **55**, 33-50 (1961).
7. Thuesen, B.H. *et al.* Cohort Profile: the Health2006 cohort, research centre for prevention and health. *Int. J. Epidemiol.* **43**, 568-75 (2014).
8. Goodpaster, B.H. *et al.* Attenuation of skeletal muscle and strength in the elderly: The Health ABC Study. *J. Appl. Physiol. (1985)* **90**, 2157-65 (2001).
9. Sempos, C.T., Bild, D.E. & Manolio, T.A. Overview of the Jackson Heart Study: a study of cardiovascular diseases in African American men and women. *Am. J. Med. Sci.* **317**, 142-6 (1999).
10. Deary, I.J., Whiteman, M.C., Starr, J.M., Whalley, L.J. & Fox, H.C. The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *J. Pers. Soc. Psychol.* **86**, 130-47 (2004).
11. Deary, I.J., Gow, A.J., Pattie, A. & Starr, J.M. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **41**, 1576-84 (2012).
12. Deary, I.J. *et al.* The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr.* **7**, 28 (2007).
13. Bild, D.E. *et al.* Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology* **156**, 871-81 (2002).
14. Hofman, A. *et al.* The Rotterdam Study: 2010 objectives and design update. *European journal of epidemiology* **24**, 553-72 (2009).
15. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238-43 (2013).
16. Hofman, A., Grobbee, D.E., de Jong, P.T. & van den Ouweland, F.A. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *European journal of epidemiology* **7**, 403-22 (1991).
17. Grobbee, D.E. *et al.* The Utrecht Health Project: optimization of routine healthcare data for research. *Eur. J. Epidemiol.* **20**, 285-7 (2005).
18. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Controlled clinical trials* **19**, 61-109 (1998).
19. Hays, J. *et al.* The Women's Health Initiative recruitment methods and results. *Annals of epidemiology* **13**, S18-77 (2003).
20. Anderson, G.L. *et al.* Implementation of the Women's Health Initiative study design. *Annals of epidemiology* **13**, S5-17 (2003).
21. Raitakari, O.T. *et al.* Cohort profile: the cardiovascular risk in Young Finns Study. *Int. J. Epidemiol.* **37**, 1220-6 (2008).

22. Tang, C. *et al.* Exome-wide Association Analysis Reveals Novel Coding Sequence Variants Associated with Lipid Traits in Chinese. *Nature Communications* **in press**(2016).
23. Johnson, A.D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938-9 (2008).
24. Goring, H.H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208-16 (2007).
25. Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* **42**, 62-7 (2010).
26. Heap, G.A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2**, 1 (2009).
27. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580-5 (2013).
28. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14-24 (2014).
29. Benton, M.C. *et al.* Mapping eQTLs in the Norfolk Island genetic isolate identifies candidate genes for CVD risk traits. *Am. J. Hum. Genet.* **93**, 1087-99 (2013).
30. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
31. Fehrmann, R.S. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
32. Landmark-Hoyvik, H. *et al.* Genome-wide association study in breast cancer survivors reveals SNPs associated with gene expression of genes belonging to MHC class I and II. *Genomics* **102**, 278-87 (2013).
33. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* **21**, 48-54 (2013).
34. Sasayama, D. *et al.* Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: an eQTL study in the Japanese population. *PLoS ONE* **8**, e54967 (2013).
35. van Eijk, K.R. *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13**, 636 (2012).
36. Zhernakova, D.V. *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9**, e1003594 (2013).
37. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202-7 (2007).
38. Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* **23**, 716-26 (2013).
39. Stranger, B.E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
40. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40**, 225-31 (2008).
41. Cusanovich, D.A. *et al.* The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum. Mol. Genet.* **21**, 2111-23 (2012).
42. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-50 (2009).
43. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084-9 (2012).

44. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
45. Mangravite, L.M. *et al.* A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature* **502**, 377-80 (2013).
46. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502-10 (2012).
47. Murphy, A. *et al.* Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum. Mol. Genet.* **19**, 4745-57 (2010).
48. Heinzen, E.L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* **6**, e1 (2008).
49. Zeller, T. *et al.* Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**, e10693 (2010).
50. Barreiro, L.B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1204-9 (2012).
51. Huang, R.S. *et al.* Population differences in microRNA expression and biological implications. *RNA Biol.* **8**, 692-701 (2011).
52. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-4 (2012).
53. Greenawalt, D.M. *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* **21**, 1008-16 (2011).
54. Kompass, K.S. & Witte, J.S. Co-regulatory expression quantitative trait loci mapping: method and application to endometrial cancer. *BMC Med. Genomics* **4**, 6 (2011).
55. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
56. Webster, J.A. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* **84**, 445-58 (2009).
57. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet.* **8**, e1002707 (2012).
58. Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519-23 (2011).
59. Kim, S., Cho, H., Lee, D. & Webster, M.J. Association between SNPs and gene expression in multiple regions of the human brain. *Transl. Psychiatry* **2**, e113 (2012).
60. Liu, C. *et al.* Whole-genome association mapping of gene expression in the human prefrontal cortex. *Mol. Psychiatry* **15**, 779-84 (2010).
61. Gamazon, E.R. *et al.* Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol. Psychiatry* **18**, 340-6 (2013).
62. Gibbs, J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
63. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707-20 (2013).
64. Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* **7**, e1002078 (2011).
65. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
66. Schroder, A. *et al.* Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *Pharmacogenomics J.* **13**, 12-20 (2013).

67. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19**, 1942-52 (2009).
68. Kabakchiev, B. & Silverberg, M.S. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* **144**, 1488-96, 1496 e1-3 (2013).
69. Keildson, S. *et al.* Expression of phosphofructokinase in skeletal muscle is influenced by genetic variation and associated with insulin sensitivity. *Diabetes* **63**, 1154-65 (2014).
70. Quigley, D.A. *et al.* The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Mol. Oncol.* **8**, 273-84 (2014).
71. Gao, C. *et al.* HEFT: eQTL analysis of many thousands of expressed genes while simultaneously controlling for hidden factors. *Bioinformatics* **30**, 369-76 (2014).
72. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**, e1003029 (2012).
73. Ding, J. *et al.* Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.* **87**, 779-89 (2010).
74. Qiu, W. *et al.* Genetics of sputum gene expression in chronic obstructive pulmonary disease. *PLoS ONE* **6**, e24395 (2011).
75. Lin, H. *et al.* Gene expression and genetic variation in human atria. *Heart Rhythm* **11**, 266-71 (2014).
76. Rantalainen, M. *et al.* MicroRNA expression in abdominal and gluteal adipose tissue is associated with mRNA expression levels and partly genetically driven. *PLoS ONE* **6**, e27338 (2011).
77. Auer, P.L. *et al.* Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat. Genet.* **46**, 629-34 (2014).
78. Auer, P.L. *et al.* Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am. J. Hum. Genet.* **91**, 794-808 (2012).
79. Willer, C.J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics* **40**, 161-9 (2008).
80. Ganesh, S.K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41**, 1191-8 (2009).

Study-specific Acknowledgements

AGES: Age, Gene/Environment Susceptibility (AGES) Reykjavik Study is funded by NIH contract N01-AG-12100, the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association) and the Althingi (the Icelandic Parliament). The study is approved by the Icelandic National Bioethics Committee, VSN: 00-063.

ARIC: The Atherosclerosis Risk in Communities (ARIC) study is carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419).

Cardiovascular Health Study: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants HL080295, HL087652, HL103612, HL105756, HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through AG023629 from the National Institute on Aging (NIA). A full list of CHS investigators and institutions can be found at http://chs-nhlbi.org/. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

FamHS: The Family Heart Study was funded by grant R01DK089256 from the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center, and by grants R01-HL-087700, R01-HL-088215, and R01-HL-117078 from the National Heart, Lung, and Blood Institute.

FHS: The National Heart, Lung, and Blood Institute's Framingham Heart Study (FHS) is a joint project of the National Institutes of Health and Boston University School of Medicine and was supported by the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract No. N01-HC-25195) and its contract with Affymetrix, Inc. for genotyping services (contract No. N02-HL-6-4278). Analyses reflect the efforts and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGAII) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center.

Health2006: The Health2006 study was financially supported by grants from the Velux Foundation; the Danish Medical Research Council, Danish Agency for Science, Technology and Innovation; the Aase and Ejner Danielsens Foundation; ALK-Abello´ A/S (Horsholm, Denmark), Timber Merchant Vilhelm Bangs Foundation, MEKOS Laboratories (Denmark) and Research Centre for Prevention and Health, the Capital Region of Denmark.

Health2008: The Health2008 study was supported by the Timber Merchant Vilhelm Bang's Foundation, the Danish Heart Foundation (07-10-R61-A1754-B838-22392F), and the Health Insurance Foundation (2012B233).

The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk).

HABC: The Health ABC Study was supported by NIA contracts N01AG62101, N01AG62103, and N01AG62106 and, in part, by the NIA Intramural Research Program. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov).

JHS: The Jackson Heart Study is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities.

LBC1921/LBC1936: Genotyping for the Lothian Birth Cohort 1921 and Lothian Birth Cohort 1936was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK and the Royal Society of Edinburgh and was conducted at the Wellcome Trust Clinical Research Facility Genetics Core at Western General Hospital, Edinburgh, UK. Phenotype collection for LBC1921 was supported by the BBSRC, the Royal Society and the Chief Scientist Office of the Scottish Government. Phenotype collection for LBC1936 was supported by Age UK (The Disconnected Mind project). The work was undertaken by The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (MR/K026992/1). Funding from the BBSRC and MRC is gratefully acknowledged. We thank the cohort participants and team members who contributed to this study.

MESA: The Multi-Ethnic Study of Atherosclerosis (MESA) and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts N01 HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169 and RR-024156. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-6-4278.

PUUMA: The project is a collaboration between Peking University Health Science Center and The University of Michigan Medical School. Funding for the project was provided by the University of Michigan Medical School and the Peking University Health Sciences Center Joint Institute for Clinical and Translational Research.

RS: The GWA database of the Rotterdam Study was funded through the Netherlands Organisation of Scientific Research NWO (nr. 175.010.2005.011). The Rotterdam Study is supported by the Erasmus Medical Center and Erasmus University, Rotterdam; the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the

Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The Exome chip array data set was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO)-sponsored Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-810); the Netherlands Organization for Scientific Research (NWO; project number 184021007) and by the Rainbow Project (RP10; Netherlands Exome Chip Project) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl). We thank Ms. Mila Jhamai, Ms. Sarah Higgins, and Mr. Marijn Verkerk for their help in creating the exome chip database, and Carolina Medina-Gomez, BSc, Lennard Karsten, BSc, and Dr. Linda Broer for QC and variant calling.

UHP: The Utrecht Health Project (LRGP) received grants from the Ministry of Health, Welfare and Sports (VWS), the University of Utrecht, the Province of Utrecht, the Dutch Organisation of Healthcare Insurance Companies and the University Medical Centre of Utrecht. The exome chip data were generated in a research project that was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007). We thank inhabitants and GPs of the “Leidsche Rijn” district, Utrecht, for their cooperation in this project. Folkert W. Asselbergs is supported by a Dekker scholarship-Junior Staff Member 2014T001 – Netherlands Heart Foundation and UCL Hospitals NIHR Biomedical Research Centre.

The Women’s Health Initiative: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.” The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf. Exome-chip data and analysis were supported through the Women’s Health Initiative Sequencing Project (NHLBI RC2 HL-102924), the Genetics and Epidemiology of Colorectal Cancer Consortium (NCI CA137088), the Genomics and Randomized Trials Network (NHGRI U01-HG005152), and an NCI training grant (R25CA094880).

The Young Finns Study: YFS has been financially supported by the Academy of Finland: grants 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi); the Social Insurance Institution of Finland; Kuopio, Tampere and Turku University Hospital Medical Funds (grant X51001 for T.L.); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation of Cardiovascular Research (T.L.); Finnish Cultural Foundation; Tampere Tuberculosis Foundation (T.L.); Emil Aaltonen Foundation (T.L.); and Yrjö Jahnsson Foundation (T.L.). The expert technical assistance in the statistical analyses by Ville Aalto and Irina Lisinen is gratefully acknowledged.

This work was supported in part by the NIDDK Division of Intramural Research.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

S.K.G. was supported by R01 HL122684 and the Doris Duke Charitable Foundation.

Supplementary Figures

Supplementary Figure 1: LocusZoom plot for *S1PR4* gene region. –log *P* values are shown for exome chip single variant meta-analysis results for total WBC count. Estimated LD with Arg365Leu is shown according to the color code for the r^2 ranges depicted

Supplementary Figure 2: Peripheral blood circulating neutrophil counts in mice null for *S1pr4* as compared to wild-type mice, demonstrating outliers. Box plots of blood neutrophil counts are shown with outliers which were defined by |Q1-1.5(Q3-Q1), Q3+1.5(Q3-Q1)|, which were the points outside the fence or whisker in the boxplots (Lower fence: Q1-1.5*IQR=Q1-1.5(Q3-Q1); Upper fence: Q3+1.5*IQR=Q3+1.5(Q3-Q1)) in which Q1, Q2, Q3 are the quantile numbers at 25%, 50%(median), and 75%. The outliers are the points shown as free dots outside the whiskers of the boxplots in all data from the 48 mice analyzed Peripheral blood circulating neutrophil counts (Gr1+% of total cells per ul of peripheral whole blood and absolute neutrophil counts) are shown for all *S1pr4* $^{-/-}$ and *S1pr4* $^{+/+}$ mouse studied by genotype groups (A) and by genotype groups broken out by sex (B).

Supplementary Figure 3: Bone marrow and spleen neutrophil counts and blood monocyte counts in mice null for *S1pr4* as compared to wild-type littermate mice. Boxplots are shown for bone marrow mature neutrophil counts (A), bone marrow immature neutrophil counts (B), spleen neutrophil counts (C) and peripheral blood circulating monocyte counts (D) in the *S1pr4*$^{-/-}$ mice and *S1pr4*$^{+/+}$ mice.

Supplementary Figure 4: Blood monocyte numbers in *S1pr4*$^{-/-}$ mice. Cells from blood of *S1pr4*$^{+/+}$ and *S1pr4*$^{-/-}$ 8-week-old mice were stained with anti-Gr-1 and anti-CD11b antibodies and analyzed by flow cytometry. Monocytes were identified as Gr-1^{low} CD11b^{+}. Results are shown as absolute numbers μl of blood (A) and as the percentage of cells analyzed (B). The bars represent mean values (n=12 for each genotype). ns, not significant.

Supplementary Figure 5: Bone marrow myeloid progenitors in *S1pr4*$^{-/-}$ mice. (A, B) Cells from bone marrow of *S1pr4*$^{+/+}$ (open bars) and *S1pr4*$^{-/-}$ (red bars) 8-week-old mice were stained with anti-Gr-1 and anti-CD11b antibodies and analyzed by flow cytometry. Myeloid progenitors were identified as Gr-1^{low} CD11b^{+} (immature) and Gr-1^{high} CD11b^{+}(mature). Results are shown as absolute numbers per femur (A) and as the percentage of cells analyzed (B). No outliers were removed for these bar graphs. The bars represent mean values (n=8 for each genotype). ns, not significant. (C, D) Adhesion molecule expression on mature myeloid progenitors. Bone marrow cells were analyzed by flow cytometry for the expression of CD49d (C) and CXCR4 (D). Results are shown as mean fluorescence intensity (MFI) from *S1pr4*$^{+/+}$ and *S1pr4*$^{-/-}$ mice. The bars represent mean values (n=8 for each genotype). ns, not significant.

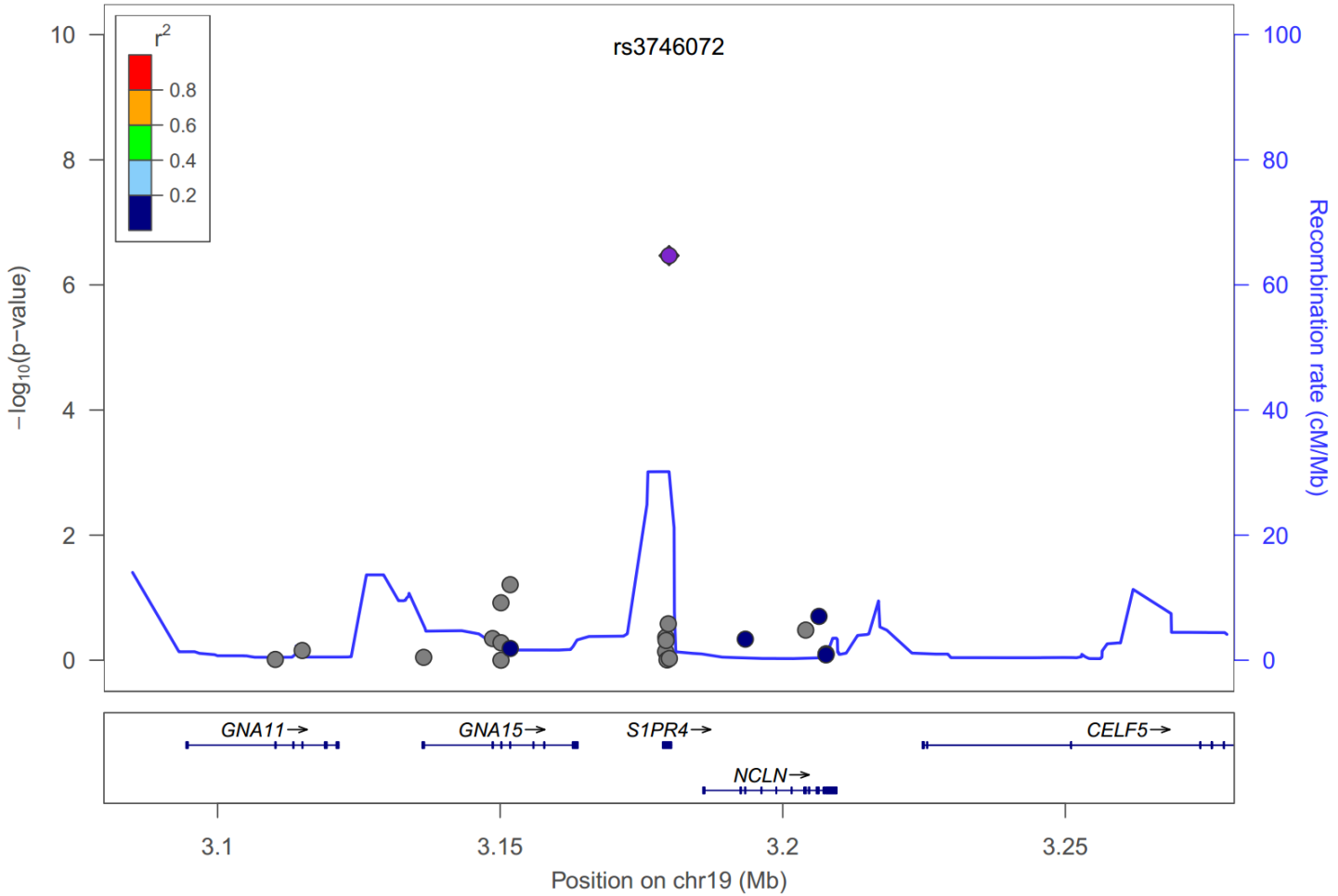
Supplementary Figure 6: Spleen neutrophils in *S1pr4*$^{-/-}$ mice. (A, B) Splenocytes from *S1pr4*$^{+/+}$ and *S1pr4*$^{-/-}$ 8-week-old mice were stained with anti-Gr-1 and anti-CD11b antibodies and analyzed by flow cytometry. Neutrophils were identified as Gr-1^{high} CD11b^{+}. Results are shown as absolute numbers per spleen (A) and as the percentage of splenocytes analyzed (B). The bars represent mean values (n=12 for each genotype). (C-E) Adhesion molecule expression on spleen neutrophils. Spleen neutrophils were analyzed by flow cytometry for the expression of

CD49d (C), CD62L (D) and CXCR4 (E). Results are shown mean fluorescence intensity (MFI) from *S1pr4*$^{+/+}$ and *S1pr4*$^{-/-}$ mice. The bars represent mean values (n=12 for each genotype). *S1pr4*$^{+/+}$ (open bars) and *S1pr4*$^{-/-}$ (red bars). ns, not significant.

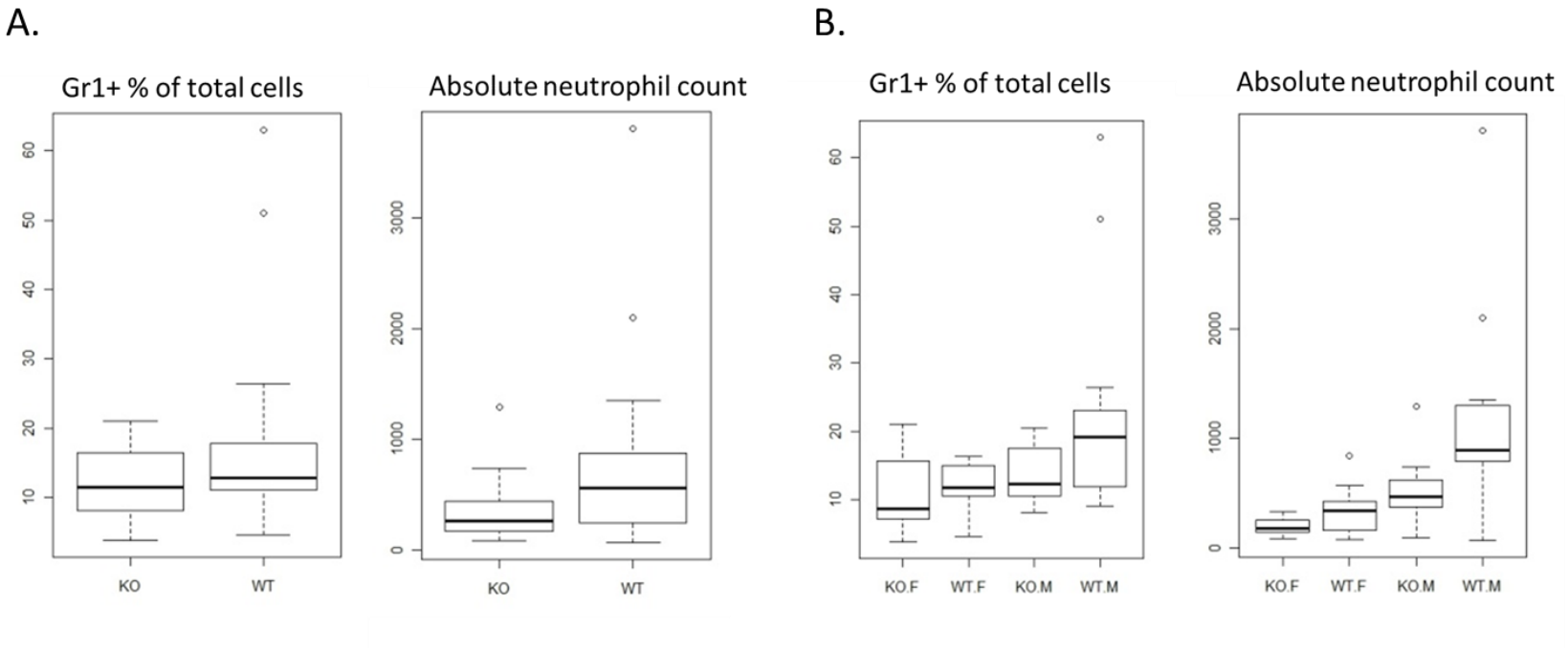
Supplementary Figure 7. Neutrophil immunohistochemistry of mouse liver and lung. Immunohistochemistry was performed to identify and quantify neutrophils in *S1pr4*$^{-/-}$ and *S1pr*$^{+/+-}$ mice liver (n=6 in each group) (A) and lung (n=6 in each group) (B).

Supplementary Figure 8: Forest plots for associations identified in the analysis of variants with ultra-low minor allele count (MAC).

Supplementary Figure 1: LocusZoom plot for *S1PR4* gene region (http:// locuszoom.sph.umich.edu)

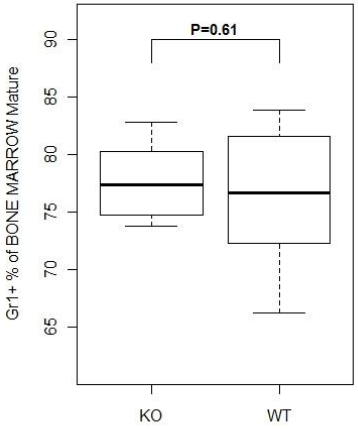


Supplementary Figure 2: Peripheral blood circulating neutrophil counts (Gr1+% of total cells per ul of peripheral whole blood and absolute neutrophil counts) are shown for all *S1pr4* $^{-/-}$ and *S1pr4* $^{+/+}$ mouse studied by genotype groups (A) and by genotype groups broken out by sex (B).

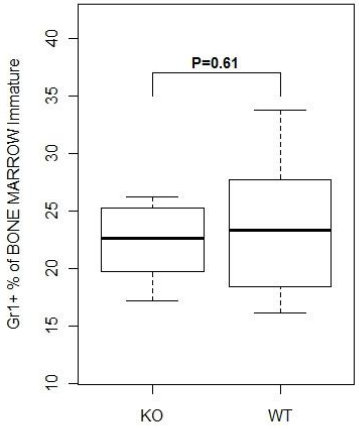


Supplementary Figure 3: Boxplots of bone marrow and spleen neutrophil counts, and peripheral blood monocyte counts in *S1pr4* $^{-/-}$ and *S1pr4* $^{+/+}$ littermates (A-D).

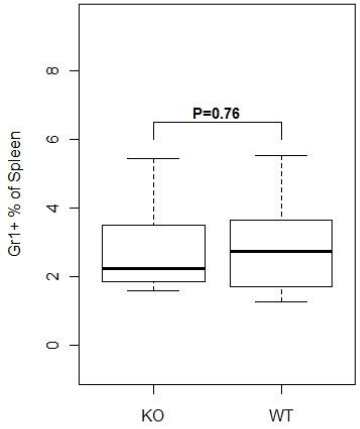
A.



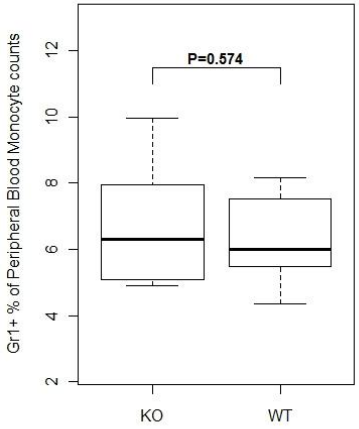
B.



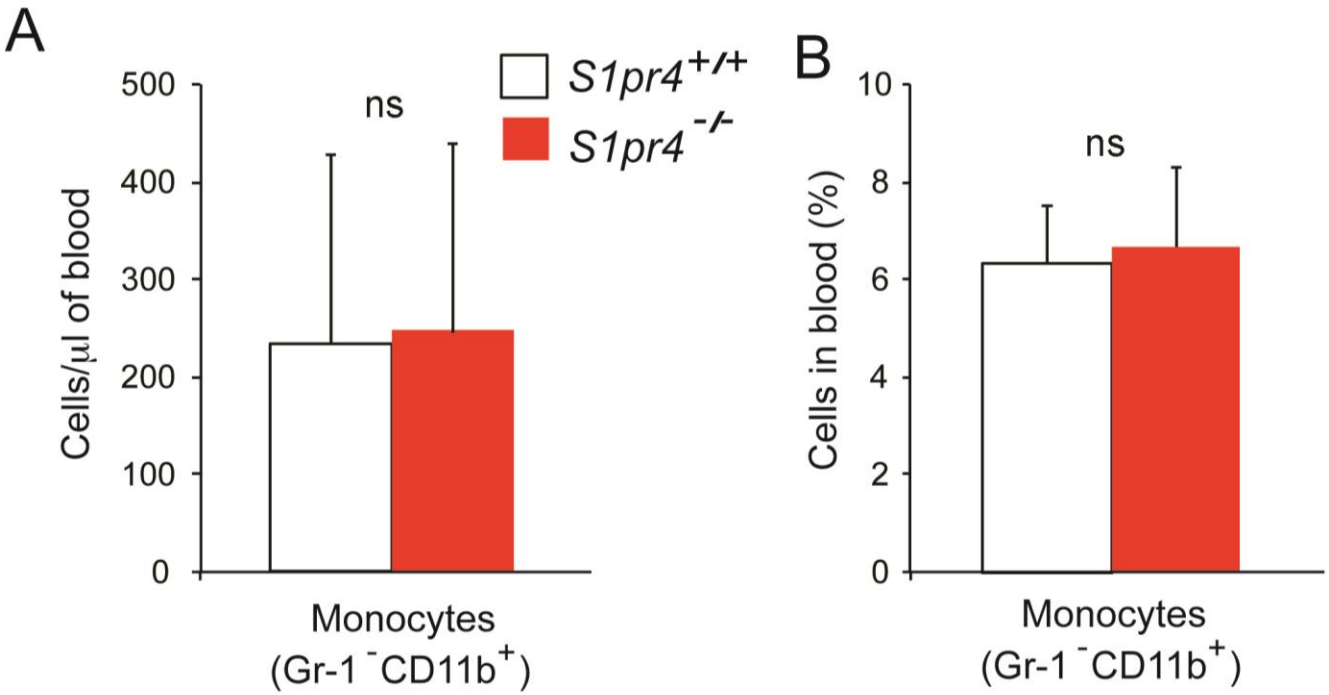
C.



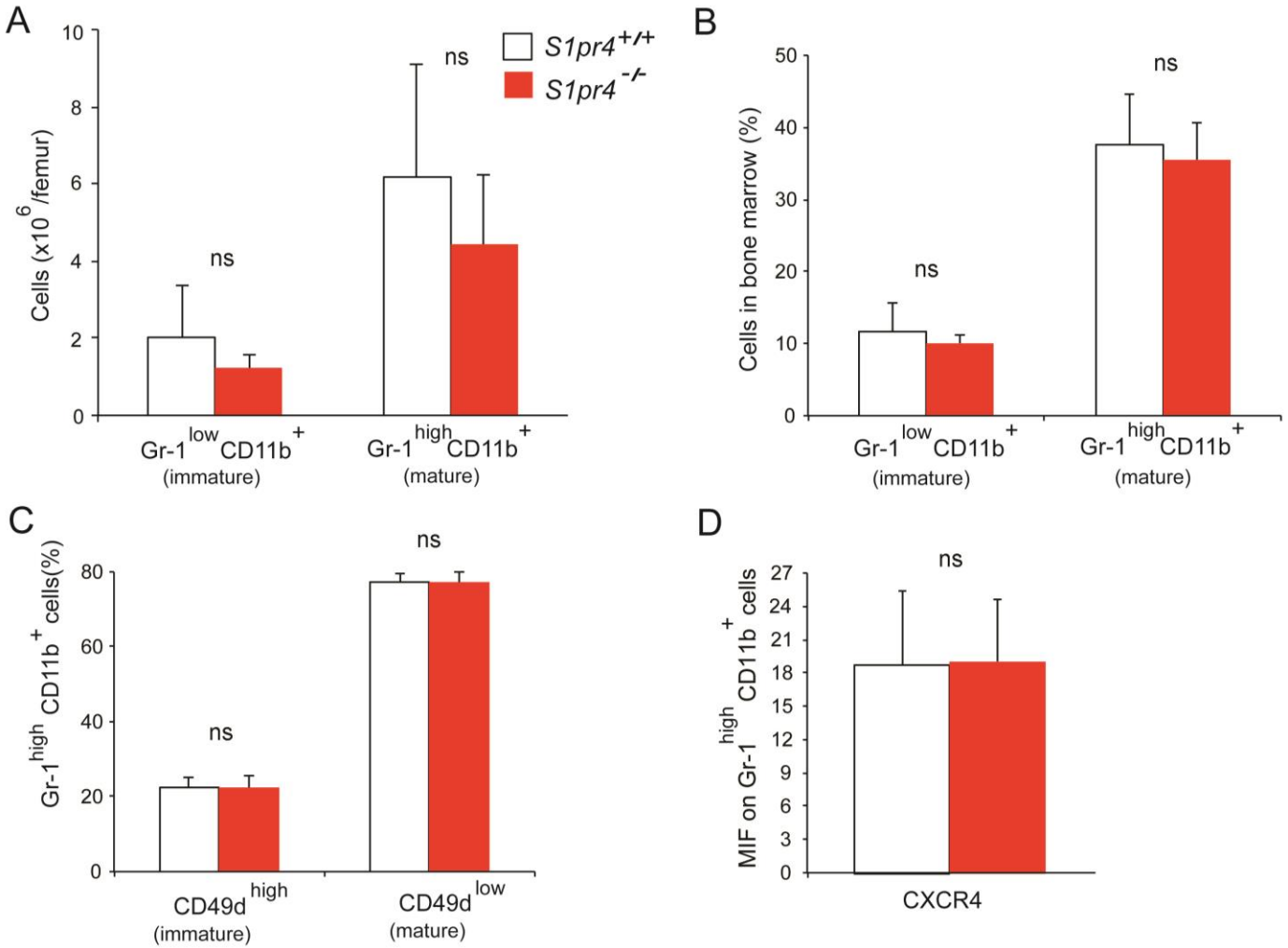
D.



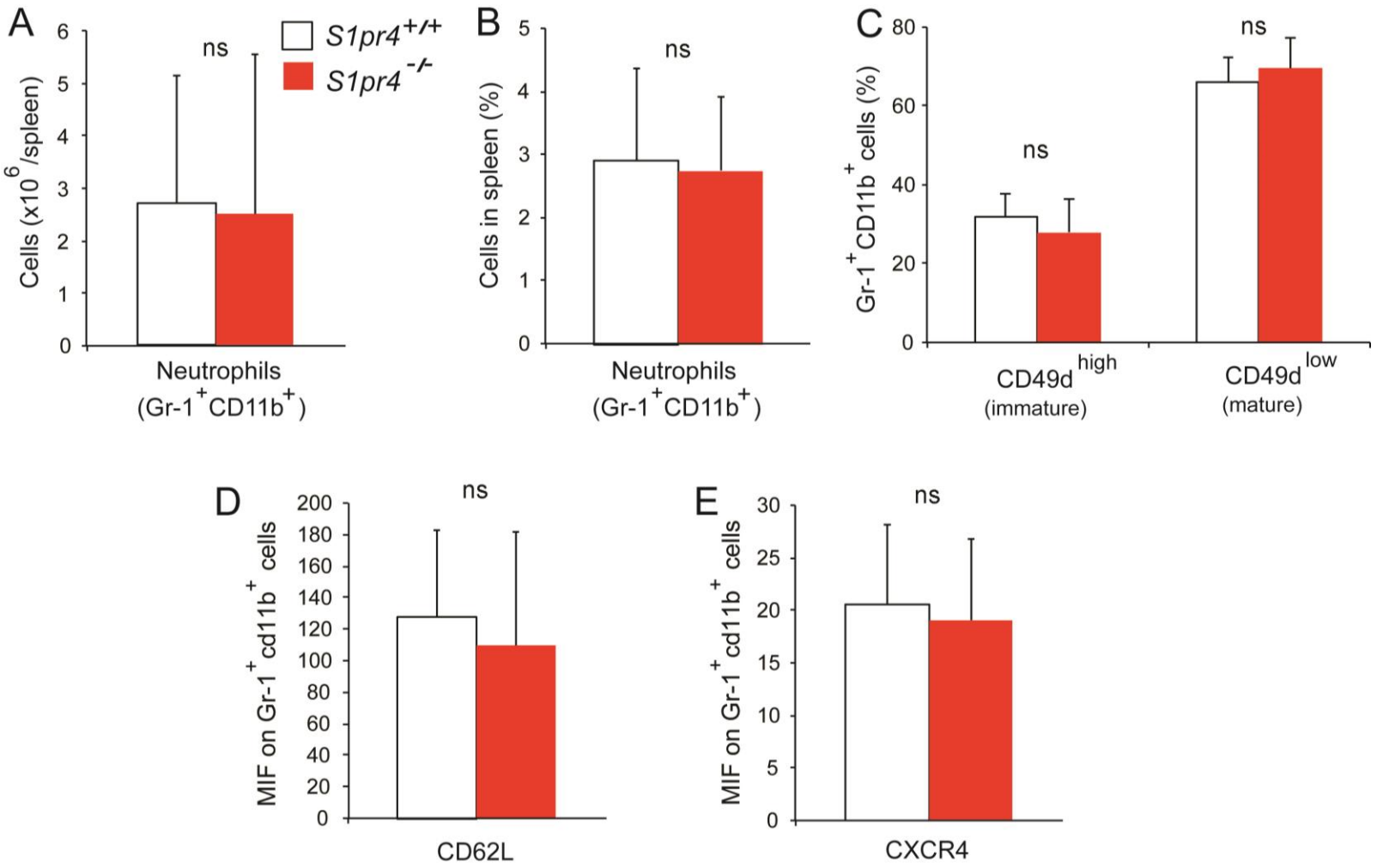
Supplementary Figure 4: Blood Monocyte mouse data



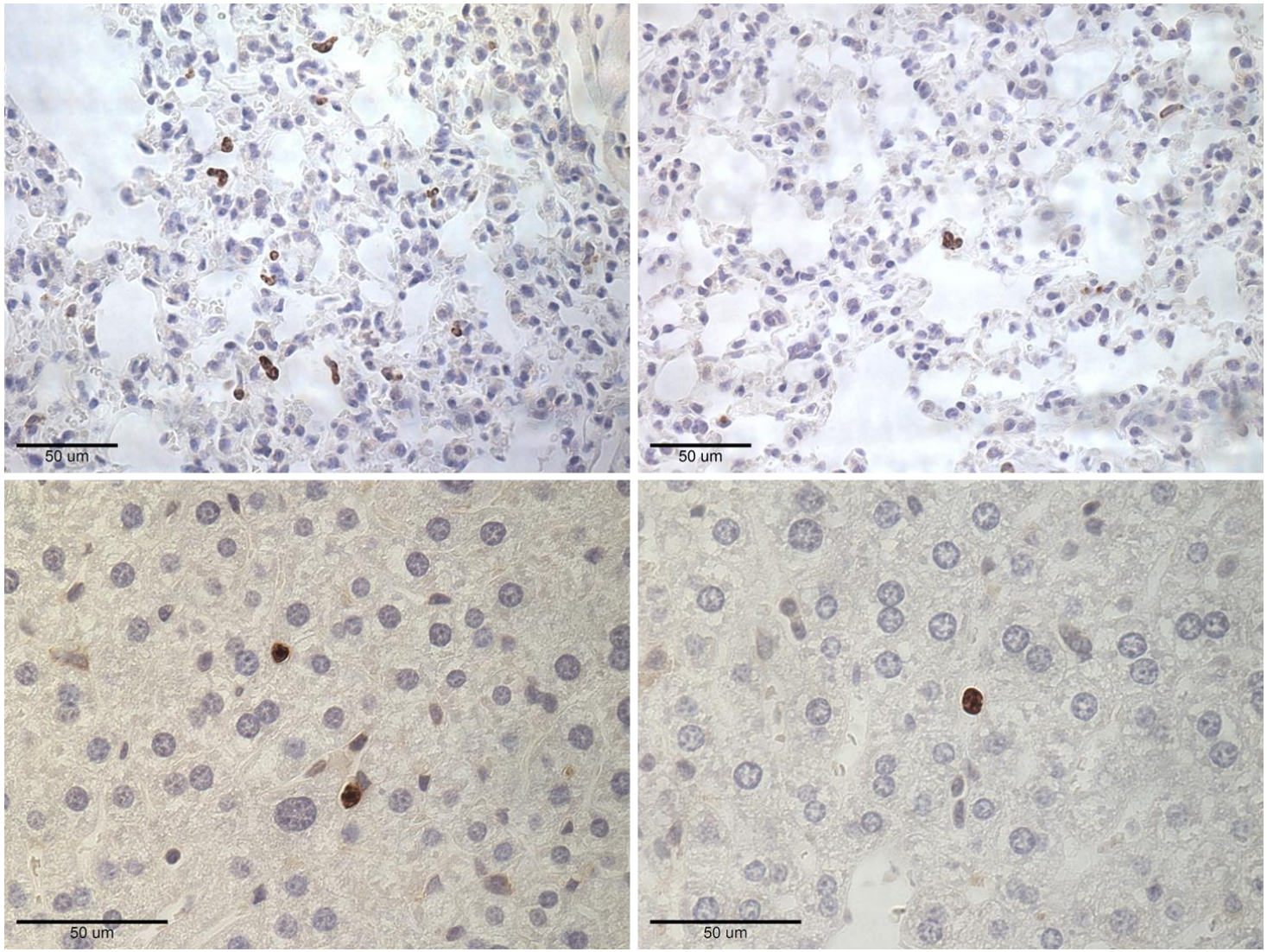
Supplementary Figure 5: Bone marrow mouse data



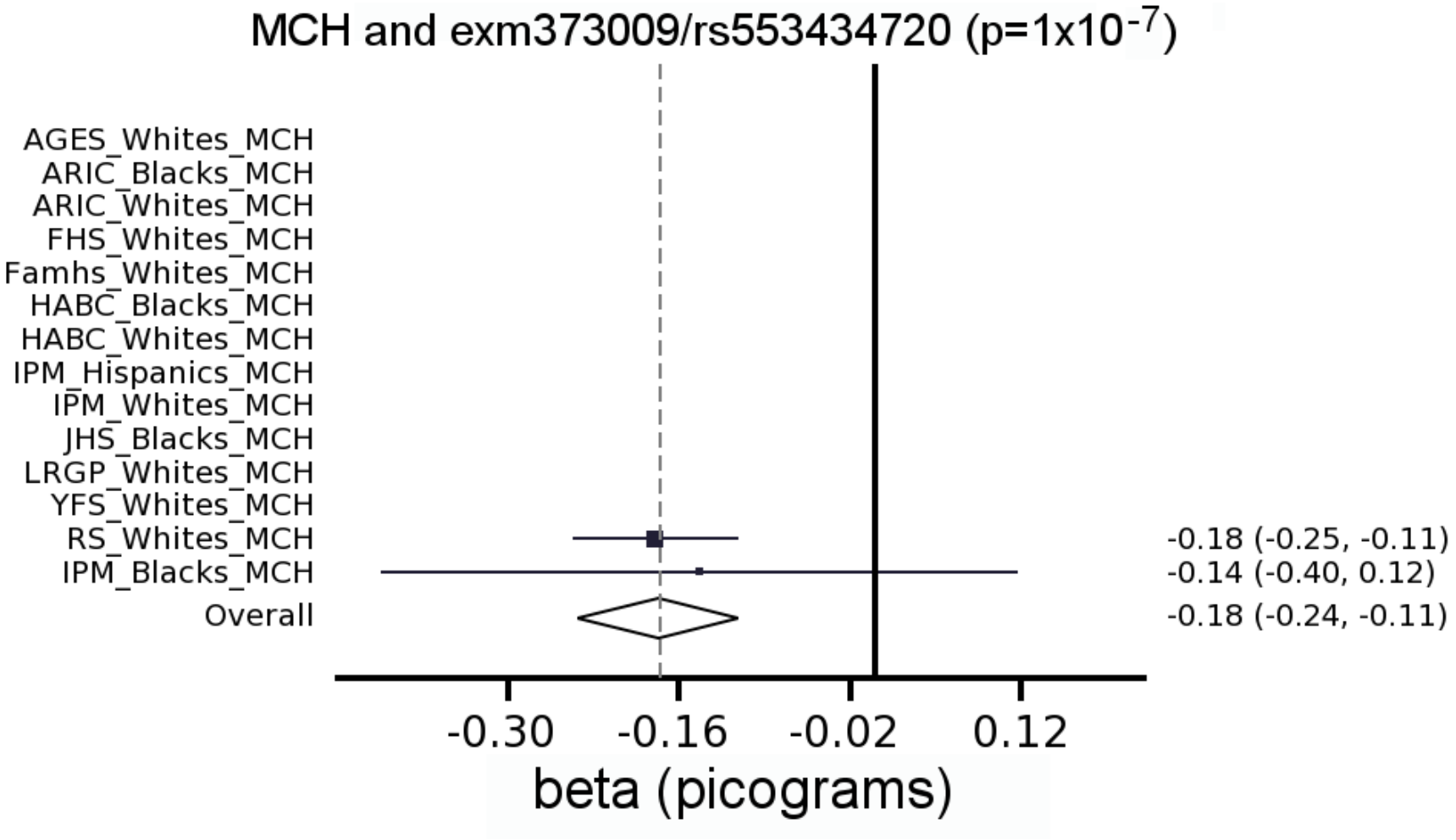
Supplementary Figure 6: Spleen mouse data

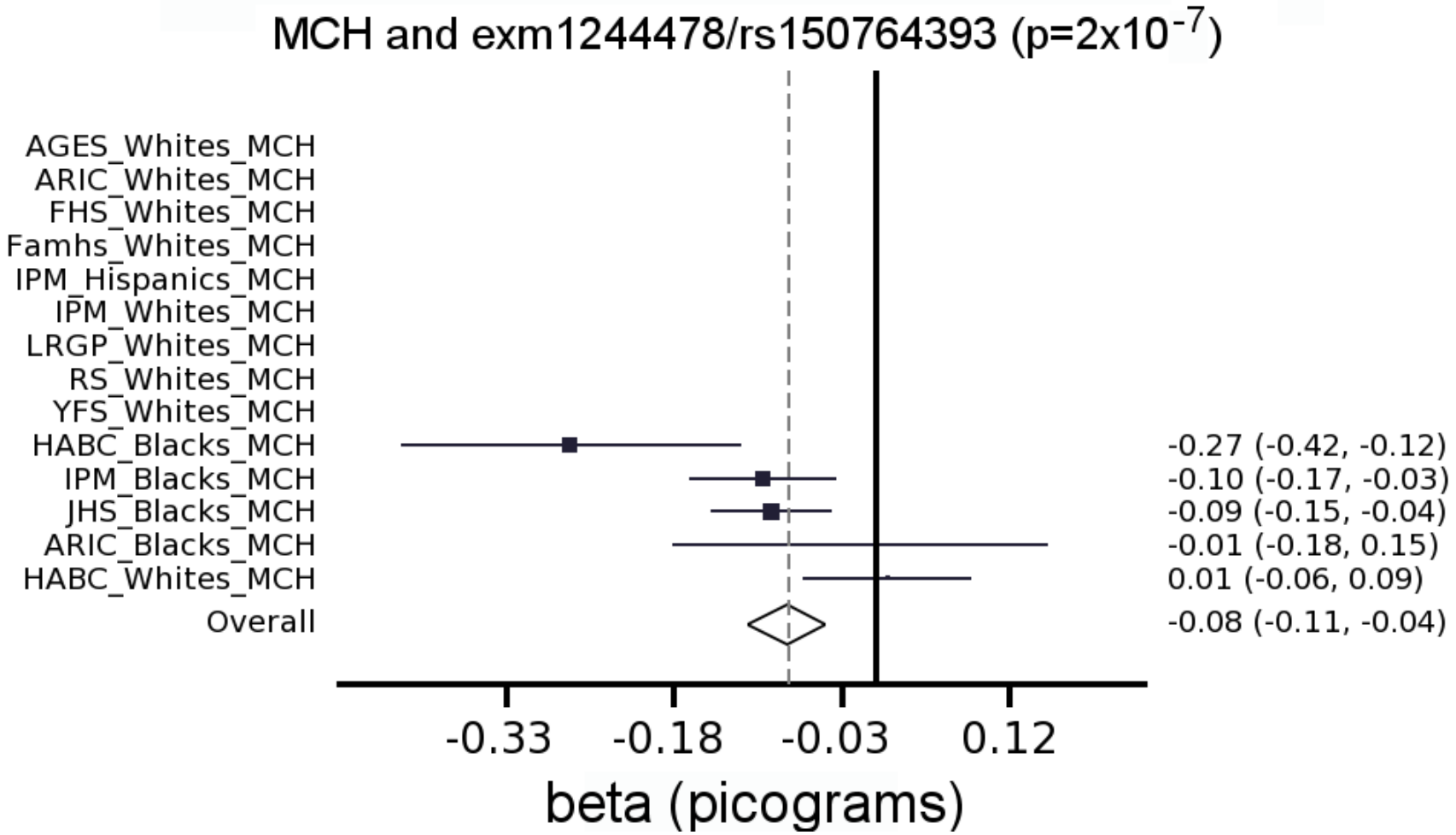


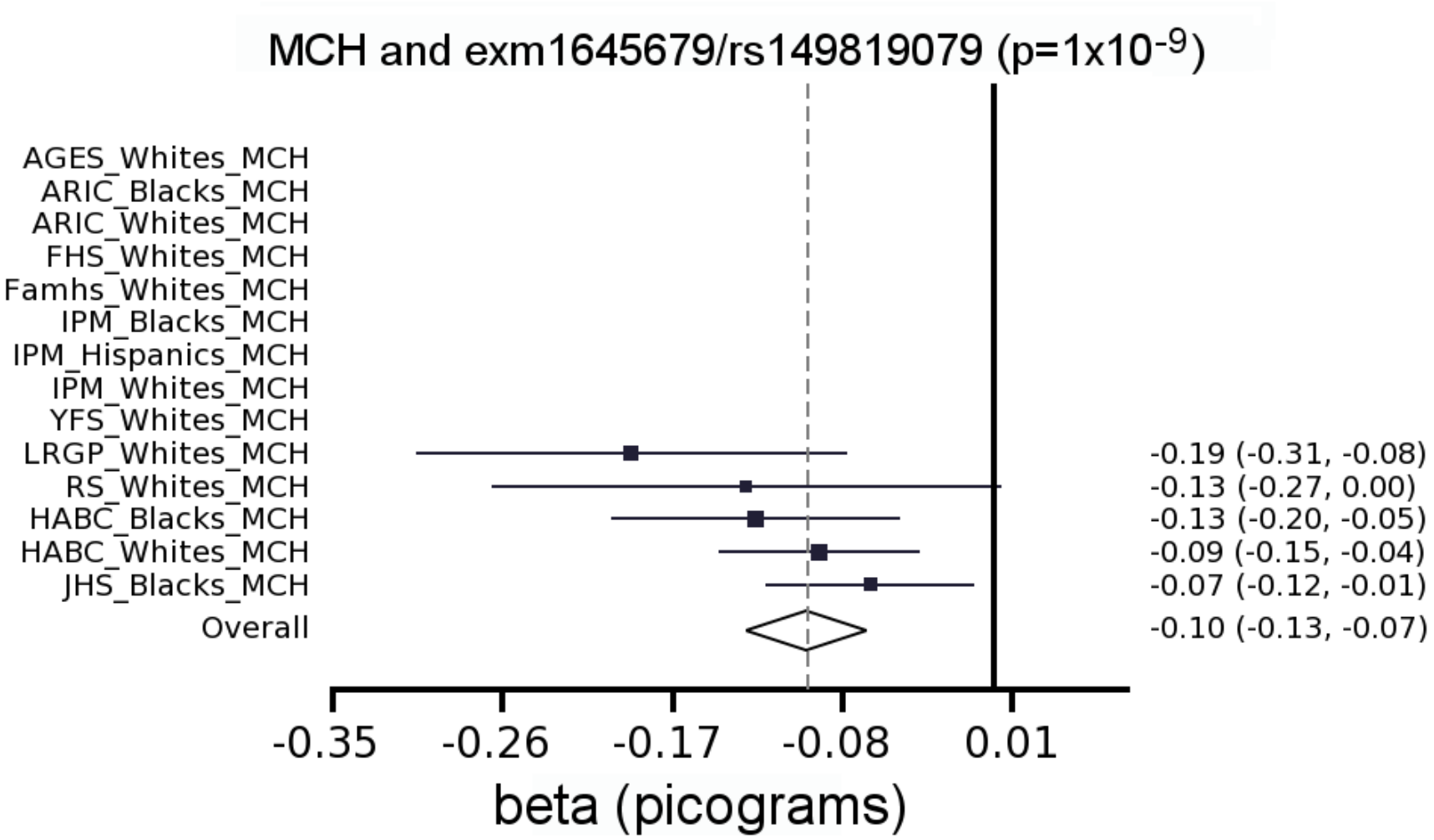
Supplementary Figure 7. Neutrophil immunohistochemistry of liver and lung from *S1pr4* $^{-/-}$ and *S1pr4* $^{+/+}$ mice. Lung images are shown for *S1pr4* $^{+/+}$ (40X) (left upper panel), and *S1pr4* $^{-/-}$ (40X) (right upper panel) and liver images are shown for *S1pr4* $^{+/+}$ (60X) (left lower panel), and *S1pr4* $^{-/-}$ (60X) (right lower panel).



Supplementary Figure 8: Forest Plots for the three ultra-rare variants associated with MCH: (a) exm373009/rs553434720 (b) exm1244478/rs150764393 (c) exm1645679/rs149819079.




MCH and exm1244478/rs150764393 (p=2x10-7)
AGES_Whites_MCH
ARIC_Whites_MCH
FHS_Whites_MCH
Famhs_Whites_MCH
IPM_Hispanics_MCH
IPM_Whites_MCH
LRGP_Whites_MCH
RS_Whites_MCH
YFS_Whites_MCH
HABC_Blacks_MCH
IPM_Blacks_MCH
JHS_Blacks_MCH
ARIC_Blacks_MCH
HABC_Whites_MCH
Overall
-0.27 (-0.42, -0.12)
-0.10 (-0.17, -0.03)
-0.09 (-0.15, -0.04)
-0.01 (-0.18, 0.15)
0.01 (-0.06, 0.09)
-0.08 (-0.11, -0.04)
-0.33
-0.18
-0.03
0.12
beta (picograms)


MCH and exm1645679/rs149819079 (p=1x10-9)
AGES_Whites_MCH
ARIC_Blacks_MCH
ARIC_Whites_MCH
FHS_Whites_MCH
Famhs_Whites_MCH
IPM_Blacks_MCH
IPM_Hispanics_MCH
IPM_Whites_MCH
YFS_Whites_MCH
LRGP_Whites_MCH
RS_Whites_MCH
HABC_Blacks_MCH
HABC_Whites_MCH
JHS_Blacks_MCH
Overall
-0.19 (-0.31, -0.08)
-0.13 (-0.27, 0.00)
-0.13 (-0.20, -0.05)
-0.09 (-0.15, -0.04)
-0.07 (-0.12, -0.01)
-0.10 (-0.13, -0.07)
-0.35
-0.26
-0.17
-0.08
0.01
beta (picograms)