Supplemental Methods

Study population

The study population was derived from 14,974 genotyped participants of reported Black race from Vanderbilt University Medical Center's (VUMC) electronic health record (EHR)-linked DNA biobank resource (BioVU).[1] Participants had 1) prior genotyping of the rs2814778 variant, and 2) a WBC or ANC measurement collected during a health maintenance exam. African ancestry was confirmed by genetic principal components analysis in conjunction with HAPMAP reference populations, and participants whose genetic race fell within the white European ancestry range were excluded. Subjects were also excluded if they had ICD-9 or ICD-10 diagnosis codes for hematological malignancies, lymphoma, radiation or chemotherapy, anemia, sepsis, chronic infections (e.g. HIV), systemic lupus erythematosus, or organ transplant.[2,3]

This study was approved by the VUMC Institutional Review Board.

Genetic data

The rs2814778 variant was genotyped on the Illumina Infinium MEGA^{EX} platform. Quality control analyses were implemented, as previously described.[4,5] Subjects with outlying heterozygosity estimates were excluded, as were one of each pair of related subjects with a pi-hat>0.3. The allele frequency for the rs2814778-C variant was 0.79 and the Hardy-Weinberg equilibrium p-value was 0.02. PLINK v 1.90β3.42 was used for these analyses.[6]

Phenotype data

Only white blood cell (WBC) counts and ANC measurements obtained on the same day as a health maintenance examination (i.e. a preventive care visit), defined by presence of ICD-9 ('V20', 'V20.1', 'V20.2', 'V70', 'V70.0', 'V70.9') or ICD-10 ('Z00.00', 'Z00.8', 'Z00.129') codes, were examined. Laboratory results with WBC counts < 35,000 cells/µl were included. The first and minimum values were identified for each participant, as some participants had multiple WBC or ANC measurements collected over time.

Diagnoses of "Neutropenia" and "Decreased WBC count" where based on phenome wide association (PheWAS) codes (v1.2, https://phewascatalog.org/phecodes), which are collections of related ICD-9 and ICD-10 diagnostic billing codes.[2,3] For each phenotype, cases have one or more instances of the code in their EHR.[7] Controls had no closely related codes.

Analysis

Because lower leukocyte counts are associated with the CC genotype, but not CT or TT, the rs2814778 genotypes were grouped as CC vs CT/TT (binary). Select demographic characteristics, WBC counts, and ANC are presented for the entire population and by genotype. 95% confidence intervals (CI) for the differences in WBC or ANC values by genotype were determined by boot-strapping.

To explore longitudinal trends, we performed a repeated measures analysis for participants who had more than one measurement of WBC and ANC. For this analysis, linear mixed models were fitted adjusting for age, sex, genotype and the interaction between age and genotype as fixed effects, and participants as a random effect. These fitted models were plotted stratified by genotypes.

The proportions of subjects whose minimum or first ANC value fell below thresholds of 1500, 1000 and 500 (cells/µL) are presented, stratified by age (<18 and >=18 years) and genotype.

These ANC categories are not mutually *exclusive.* Values for WBC counts and ANC corresponding to the median, and *2.5th and 97.5th* percentiles of the population distribution are presented by age group and genotype.

Associations with a diagnosis of low WBC and rs2814778-CC were tested using multivariable logistic regression, adjusting for sex, age and 5 genetic principal components, generating odds ratios for the CC versus CT/TT genotypes.[8]

R v4.0.0 was used for data analysis and visualization.

Data sharing statement
For original data, please contact jonathan.d.mosley@vumc.org.

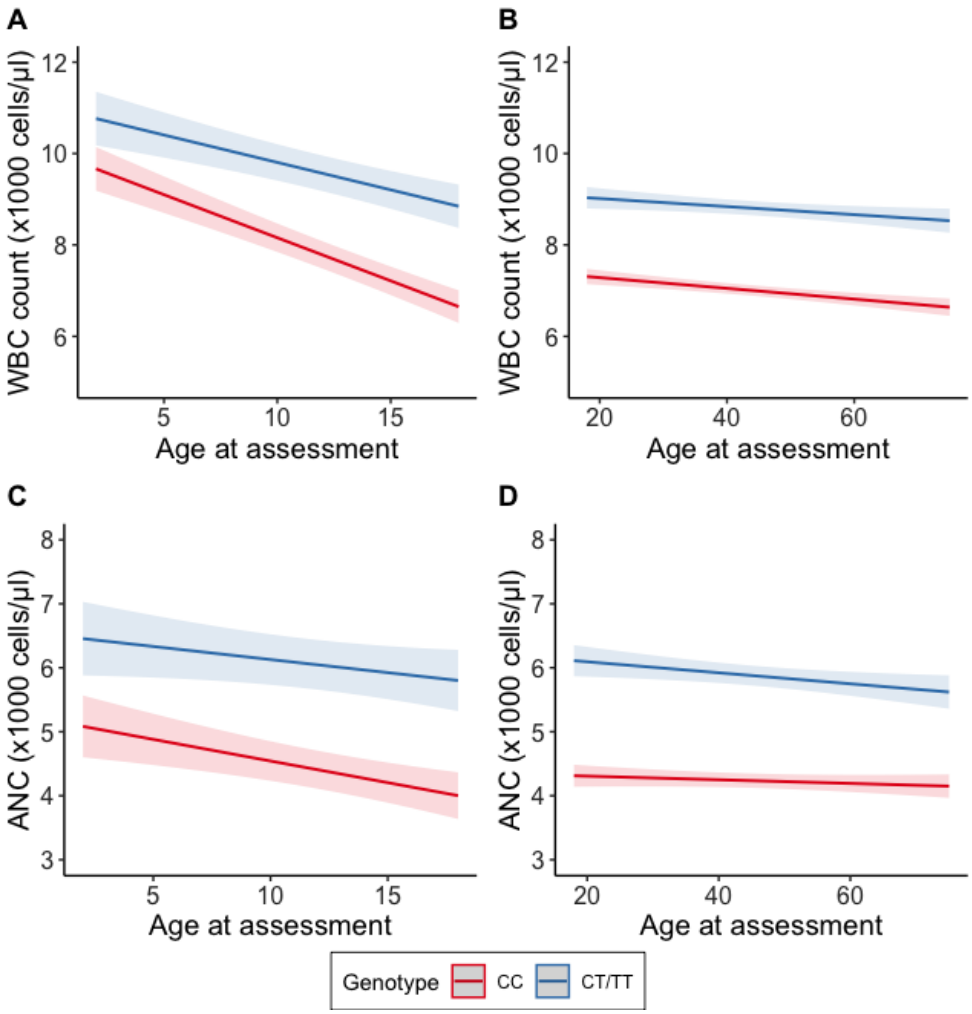
Supplemental table 1. Population characteristics.

Characteristic[1]	Total (n=3739)	CC genotype (n=2437)	CT/TT genotype (n=1302)
Median Age (years)	36 (17-53)	37 (18-54)	35 (15-53)
Age < 18 years (%)	968 (26%)	594 (24%)	374 (29%)
Females (%)	2384 (64%)	1598 (66%)	786 (60%)
Number of test results per participant[2]	7 (3-19)	8 (3-20)	7 (3-17)
Duration of follow-up (years)[3]	8.1 (3.1-14.3)	8.4 (3.2-14.6)	7.7 (2.7-13.9)
First[4] WBC ($x10^3$ cells/µl)	7 (5.5-9.2)	6.4 (5-8.3)	8.2 (6.6-10.5)
First[4] ANC (cells/µL)	3800 (2610-5680)	3270 (2260-4875)	4830 (3570-6760)

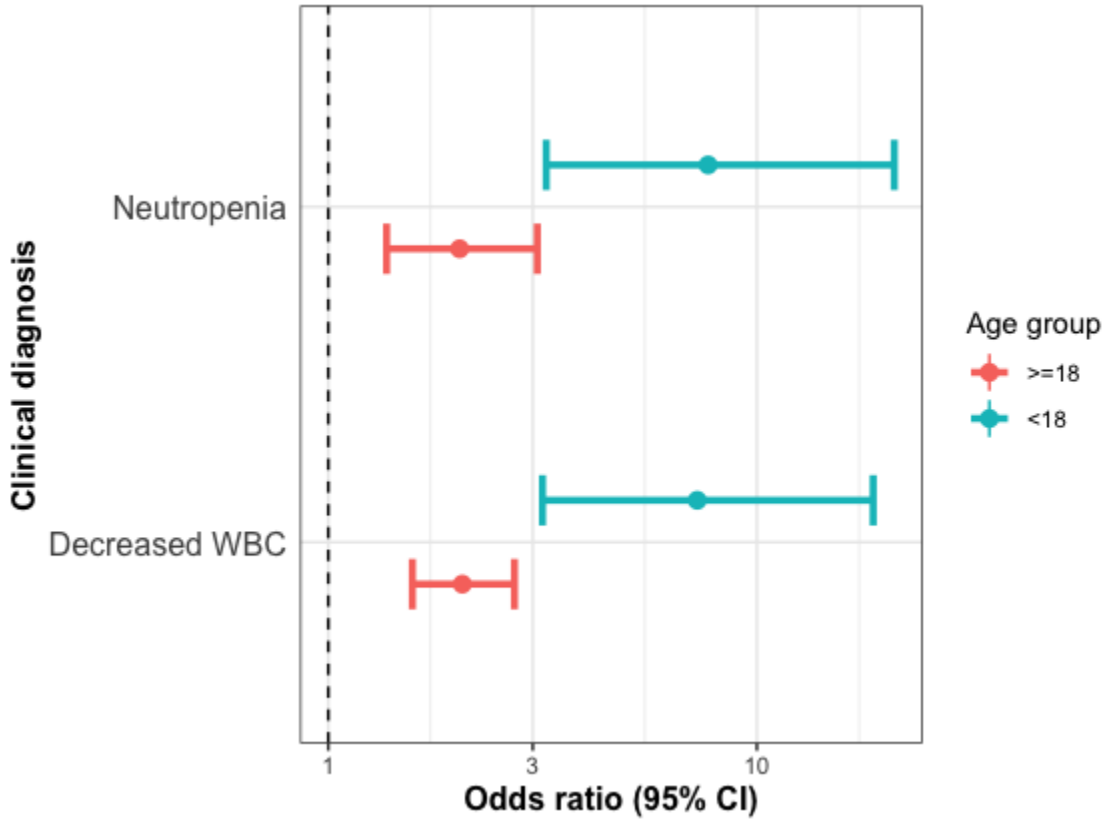
Footnotes:

1. Results are presented in median (IQR) or n (%)
2. The total number of WBC or ANC measurements collected during a health maintenance exam.
3. The duration of time between the first and last WBC or ANC measurement.
4. Summary measure based on the first observed value for a participant with multiple measurements.
5. Abbreviations: *IQR*: interquartile range, *WBC*: white blood cell

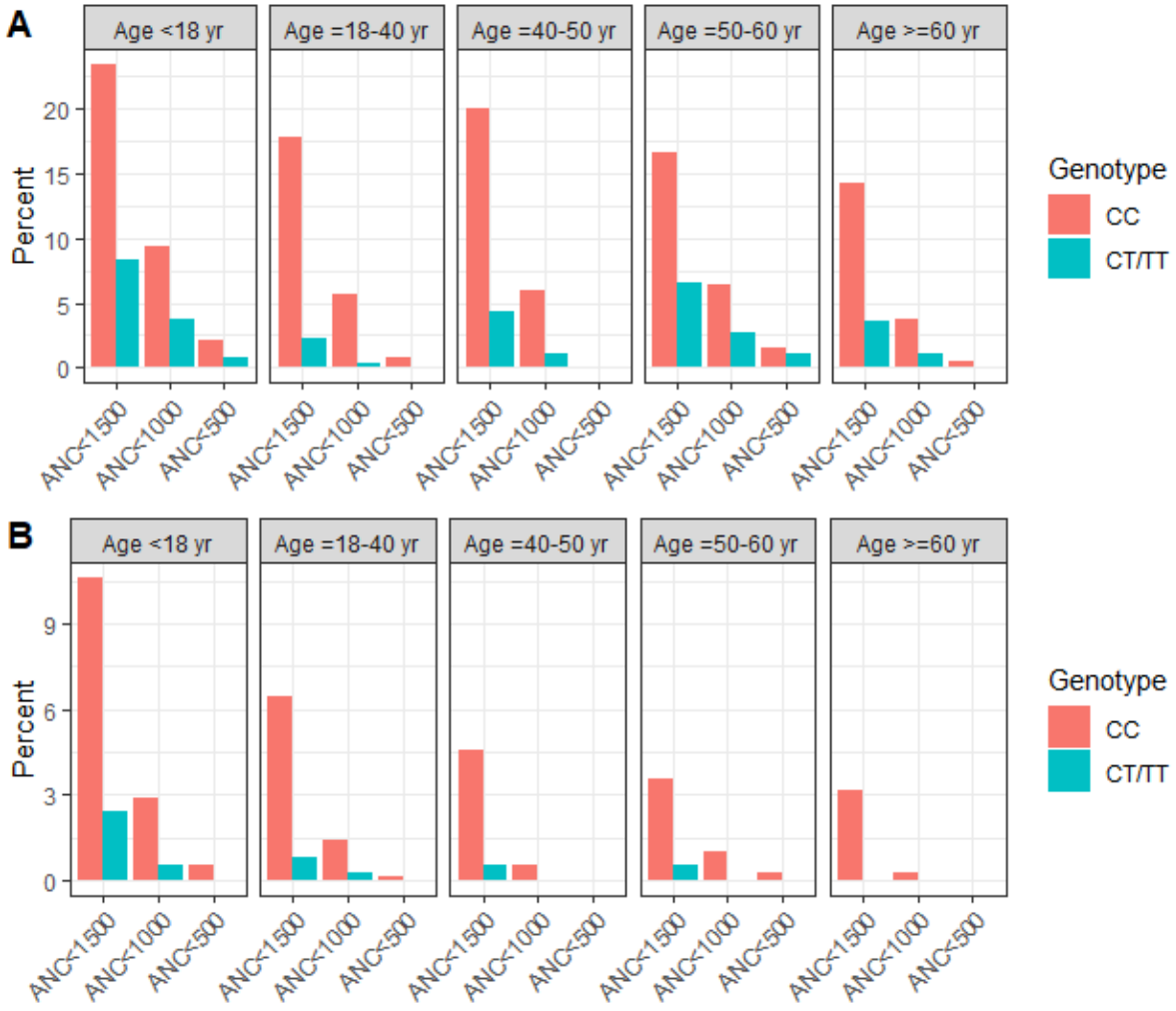
Supplemental figure 1. White blood cell and absolute neutrophil counts by age and genotype among participants with more than one measurement. To explore longitudinal trends, we performed a repeated measures analysis for participants who had more than one measurement of WBC and ANC. Linear mixed models were fitted, adjusting for age, sex, genotype and the interaction between age and genotype as fixed effects, and participants as a random effect. Panels (A & C) show WBC and ANC values, respectively, for participants under 18 and panels (B & D) are for participants >=18 years. The lines represent the fitted values, by genotype, across the age spectrum, and the shaded regions represent standard errors.



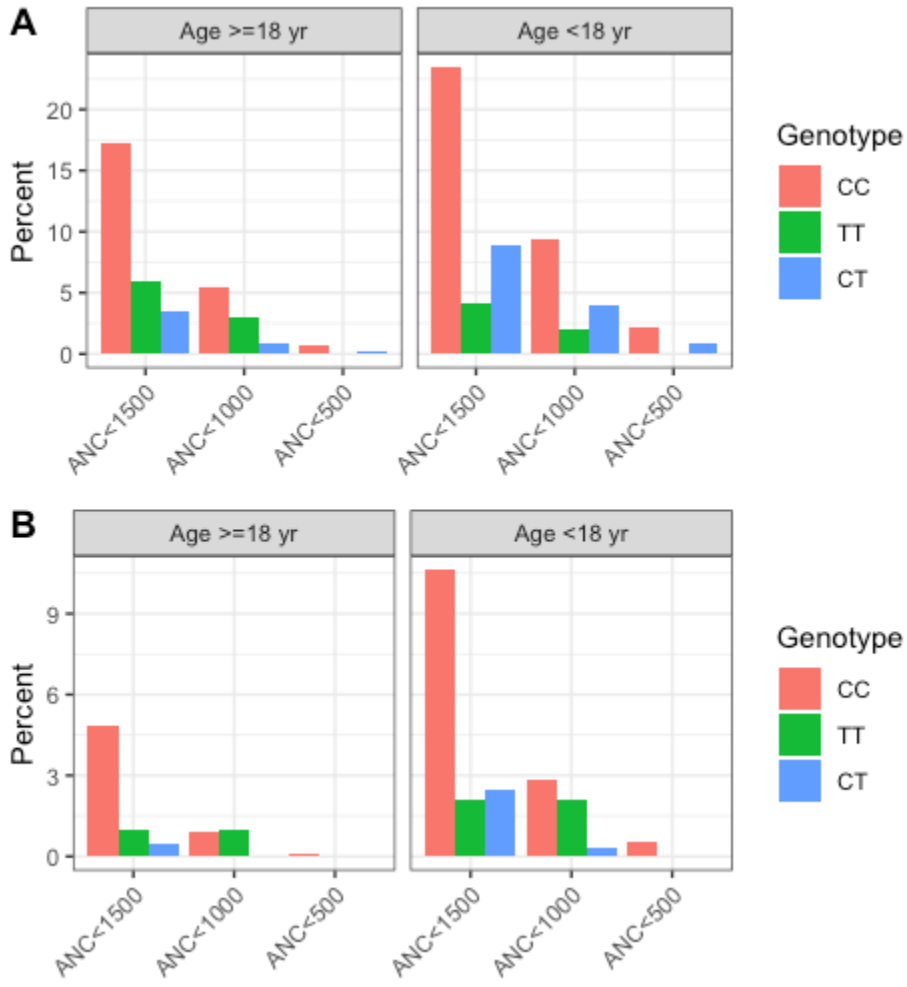
Supplemental figure 2. Association of the genotype (CC vs CT/TT) with a billing code related to low WBC counts. Odds-ratios are for the presence of billing codes related to "Neutropenia" or "Decreased white blood cell count" and are based on a multivariable logistic regression analysis adjusted for age, gender and principal components of ancestry. Odds-ratios reflect the risk associated with the CC vs CT/TT genotypes.



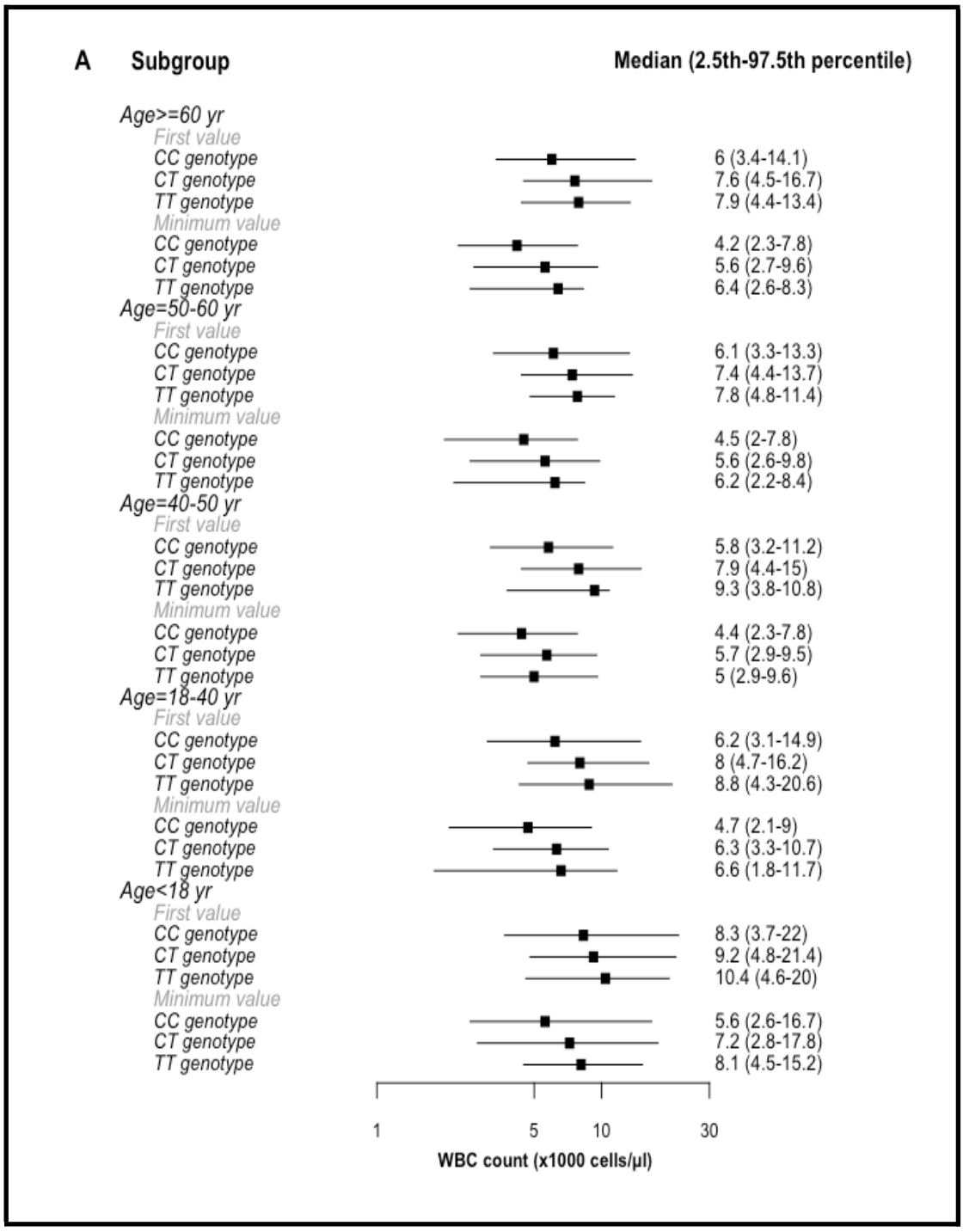
Supplementary Figure 3. Histogram showing the proportion of participants with ANC measurements that fell below select thresholds, stratified by age. (A) Histograms show the proportions of patients whose lowest ANC measurement fell below the indicated thresholds. (B) The proportions of patients whose first ANC measurement fell below the indicated thresholds. ANC thresholds represent cells/µL.



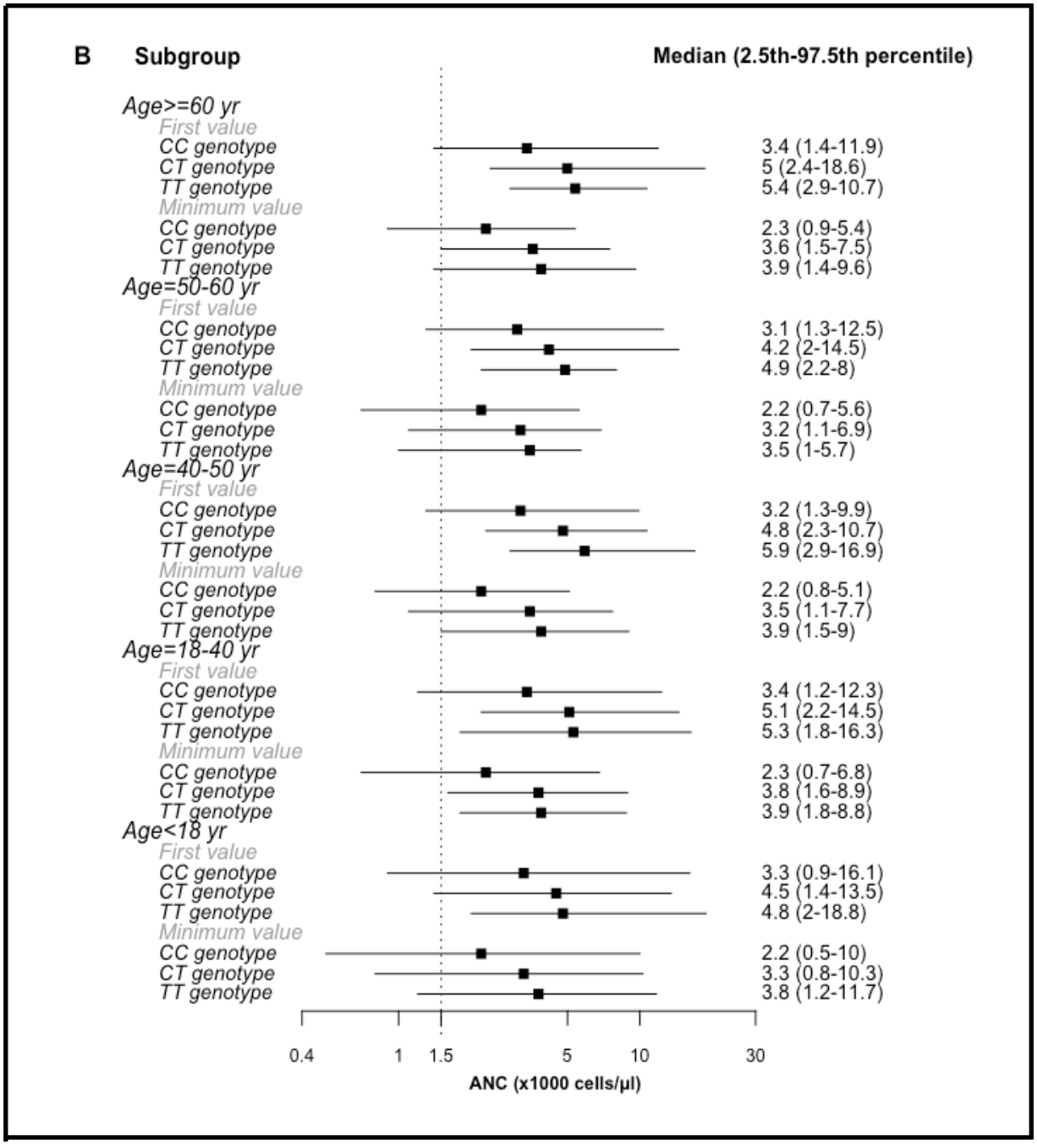
Supplemental Figure 4. Histogram showing the proportion of participants with ANC measurements that fell below select thresholds, stratified by age and genotype. (A) Histograms show the proportions of participants whose lowest ANC measurement fell below the indicated thresholds. (B) The proportions of participants whose first ANC measurement fell below the indicated thresholds. ANC thresholds represent cells/µL.



Supplementary Figure 5A. Observed ranges for WBC counts by age group and genotype. Ranges are based on either the first or minimum measurement for each participant. The median, 2.5th and 97.5th thresholds correspond the overall distribution of values within the indicated population subset.



Supplementary Figure 5B. Observed ranges for ANC by age group and genotype. Ranges are based on either the first or minimum measurement for each participant. The median, 2.5th and 97.5th thresholds correspond the overall distribution of values within the indicated population subset.



Bibliography

1. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
2. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* **26**, 1205–1210 (2010).
3. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
4. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).
5. Mosley, J. D. *et al.* Identifying genetically driven clinical phenotypes using linear mixed models. *Nat. Commun.* **7**, 11433 (2016).
6. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
7. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS One* **12**, e0175508 (2017).
8. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma. Oxf. Engl.* **30**, 2375–2376 (2014).