Published in final edited form as:

Genes Immun. 2019 September ; 20(7): 555–565. doi:10.1038/s41435-018-0051-y.

Unfolding of hidden white blood cell count phenotypes for gene discovery using latent class mixed modeling

Taryn O. Hall^{1,*}, Ian B. Stanaway¹, David S. Carrell², Robert J. Carroll³, Joshua C. Denny³, Hakon Hakonarson⁴, Eric B. Larson², Frank D. Mentch⁴, Peggy L. Peissig⁵, Sarah A. Pendergrass⁶, Elizabeth A. Rosenthal⁷, Gail P. Jarvik⁷, David R. Crosslin^{1,**}

¹Department of Biomedical Informatics Medical Education, School of Medicine, University of Washington, Seattle, WA, 98109, USA

²Kaiser Permanente Washington Health Research Institute (Formerly Group Health Cooperative-Seattle), Kaiser Permanente, Seattle, WA, 98109, USA

³Departments of Biomedical Informatics and Medicine, Vanderbilt University, Nashville, TN, 37235, USA

⁴Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

⁵Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, 54449, USA

⁶Geisinger Research, Rockville, MD, 20850, USA

⁷Division of Medical Genetics, School of Medicine, University of Washington, Seattle, WA, 98105, USA

Abstract

Resting-state white blood cell (WBC) count is a marker of inflammation and immune system health. There is evidence that WBC count is not fixed over time and there is heterogeneity in WBC trajectory that is associated with morbidity and mortality. Latent class mixed modeling (LCMM) is a method that can identify unobserved heterogeneity in longitudinal data and attempts to classify individuals into groups based on a linear model of repeated measurements. We applied LCMM to repeated white blood cell (WBC) count measures derived from electronic medical records of participants of the National Human Genetics Research Institute (NHRGI) electronic Medical Record and GENomics (eMERGE) network study, revealing two WBC count trajectory phenotypes. Advancing these phenotypes to GWAS, we found genetic associations between trajectory class membership and regions on chromosome 1p34.3 and chromosome 11q13.4. The chromosome 1 region contains *CSF3R*, which encodes the granulocyte colony stimulating factor receptor. This protein is a major factor in neutrophil stimulation and proliferation. The association on chromosome 11 contain genes *RNF169* and *XRRRA1*; both involved in the regulation of double strand break DNA repair.

*Correspondence: tohall@uw.edu. **Correspondence: davidcr@uw.edu.

Conflict of Interest

The authors declare no conflict of interest.

Introduction

White blood cell count (WBC) count is a marker of systemic inflammation and immune system health. WBC count varies acutely in response to infection and other environmental exposures. However, resting-state WBC count -- the WBC level when the immune system is neither challenged nor suppressed -- may be an indicator of chronic disease risk. Elevated resting WBC count has been associated with metabolic syndrome¹⁻⁴, cardiovascular disease^{5,6} and mortality⁷⁻¹¹. This may reflect excess inflammation as evidenced by WBC count, or leukocytes may contribute directly to disease¹².

While WBC count is impacted by modifiable factors such as smoking¹³⁻¹⁵ and body composition¹⁶⁻¹⁸, resting-state WBC count is also influenced by ancestry, and has been found to be partly under genetic control, with heritability estimated at around 40%¹⁹. Individuals with African ancestry, on average, have lower WBC count compared to individuals with European ancestry, attributed to lower neutrophil count^{20,21}. Among those with African ancestry, total WBC and neutrophil count has been associated with SNP rs2814778 in the *ACKR1/DARC* gene, via admixture mapping^{22,23}. This association was replicated in several genome-wide association studies, including our own, and a meta-analysis²⁴⁻²⁷.

There is evidence that resting-state WBC count is not fixed over time. Longitudinal analysis has shown a U-shaped pattern in WBC counts over the lifespan, dipping around age 60 and then increasing⁹. Similarly, cross-sectional data has shown higher WBC count in individuals older than 65 years old¹⁰. Heterogeneity in WBC count trajectory also exists and some trajectories are associated with morbidity and mortality⁸. Because WBC count is also influenced by adiposity, changes in resting-state WBC count may reflect age-related change in body composition. However, in a mouse model, different strains exhibited different WBC count trajectories, indicating these trajectories may be under genetic control²⁸.

Deep phenotyping aims to increase the granularity of a phenotype in hopes that a more precise phenotype will increase the power of a genome-wide association study (GWAS) and lead to larger effect size estimates²⁹. Extending a phenotype over time by harnessing the information contained in longitudinal data instead of simple aggregation is one strategy to deepen phenotype^{30,31}. Different trajectories of WBC count over the lifespan may be a fruitful deep phenotype to use in GWAS.

Trajectory heterogeneity may be difficult to discern in large, observational datasets using standard statistical methods. Trajectory analysis is a method that can identify unobserved heterogeneity in longitudinal data and attempts to classify individuals into groups based on a linear model of repeated measurements over time³². As such, this method is particularly suited to the type of data gathered in the electronic medical records (EMR) which contains information about multiple traits, gathered repeatedly over time. Trajectory analysis, applied to EMR data, has been used to characterize and identify risk factors for multimorbidity³³, depression³⁴, dementia-related cognitive decline³⁵, and adverse birth weight outcomes³⁶. Trajectory-based phenotypes have been shown to be heritable, used in candidate gene studies, linkage analysis and GWAS, and have been associated with genetic risk scores for a

number of complex traits (e.g. systolic blood pressure, BMI, schizophrenia, alcohol use and smoking behavior, ADHD and Autism)^{37–44}.

Here we applied a trajectory analysis, using latent class mixed modeling (LCMM)⁴⁵, to longitudinal WBC count data obtained from the EMR from the electronic MEDical Record and GENomics (eMERGE) Network study. We then conducted a GWAS and identified genetic variants associated with the trajectory classes derived in the deep phenotyping step.

Results

Resting-state WBC count data was identified for 14 018 participants. LCMM requires a minimum number of repeated measurements to appropriately model trajectory (here a minimum of three data points for a quadratic model). In our sample, 4 762 participants were excluded due to insufficient data. Excluded participants were younger than included participants (56.6 vs 64.1 year, respectively). There was also a higher proportion of participants of genetically determined African Ancestry (AA) among those excluded. A higher proportion of participants from the Vanderbilt University site and a lower proportion of participants from the Marshfield Clinic site were excluded.

LCMM selection

We evaluated model fit based on Bayesian Information Criteria (BIC), average posterior probability of class membership 70%, and minor class sample size 10%. A summary of the models fit are presented in Table 1. Details for all models tested are available in the Supplementary Materials. Based on these criteria, we determined that the two-class solution was the best fitting model tested.

Participants were assigned to a trajectory class based on the class for which they had a higher posterior probability of membership, given their data and the model fit. Figure 1A shows the mean predicted trajectory based on the LCMM for each class. Class 1 is modeled by the equation $-0.10137 * age_at_event + 0.00043 * age_at_event^2$. The equation for the Class 2 trajectory is $-4.00651 - 0.07115 * age_at_event + 0.00082 * age_at_event^2$. Figure 1B shows the mean observed trajectory and 95% confidence interval for classed participant data. Class 2 was the major trajectory identified, representing 65% of the participants, showed a stable resting-state WBC count trajectory and then increased after age 60. The Class 1 WBC count trajectory decreased steadily across the lifespan and accounted for 35% of sample participants. The trajectories cross at about age 70 and the 95% confidence intervals overlap from ages 68 to 72. The average posterior probability of Class 1 and Class 2 membership was moderately high at 73% and 75%, respectively, but the entropy (a measure of confidence, bounded by 0 and 1) of classification was low (0.27).

The median number of observations, age-at-event, years of follow-up, WBC counts, and BMI were similar in magnitude between classes (Table 2). Likewise, the distributions of males, those in genetically determined ancestry (GDA) groups, and study site for each class were comparable. The differences between Class 1 and Class 2 demographics were statistically significant, but this may reflect the large sample size rather than meaningful differences in cohort makeup.

GWAS

The results of the joint and GDA group stratified GWAS analyses comparing Class 1 to Class 2 trajectory phenotype are summarized in Figure 2. The Q-Q plots and lambda values of 0.9926, 1.0216 and 0.9888 indicate good control for population stratification in the Joint, AA and European Ancestry (EA) groups, respectively (Figure 2). The Manhattan plots show three regions of interest, with p-values less than 10^{-7} , for genetic associations with our trajectory phenotype classes: 1q23.2, 1p34.3, and 9q33.1 (Figure 2).

The strongest region associated with trajectory class membership we identified was on chromosome 1q23.3 in the joint and AA group analyses with the lead SNP rs2814778 (p-value = 9.83×10^{-9} , joint analysis; 9.56×10^{-9} , AA group analysis). In the AA group, the T allele of SNP rs2814778 was associated with a two-fold higher risk of having the Class 1 trajectory phenotype (OR: 2.23 95% CI: 1.23–4.07). The T allele is the minor allele among individuals of AA, with an allele frequency of 0.21 in our study. By contrast, the T allele is nearly fixed in the EA group, with an allele frequency of 0.996. SNP rs2814778 is located in the first exon of the *ACKR1/DARC* gene, and together with rs12075, polymorphisms at these loci determine the Duffy blood group. Among individuals of AA, homozygotes for the C allele of rs2814778 have the Duffy-Null phenotype, which exhibits a strong association with low neutrophil count²³.

We identified an additional significant region on chromosome 1p34.3 in the joint analysis (p-value = 3.54×10^{-8} , lead SNP rs12094900). This SNP is in moderate to high linkage disequilibrium with several other SNPs in a region that contains the genes *MRPS15*, *OSCP1*, and *CSF3R* (Figure 3). The *MRPS15* gene encodes a mitochondrial ribosomal protein. *OSCP1*, also known as *NOR1*, is a tumor-suppressor gene associated with nasopharyngeal cancer. *CSF3R* encodes the receptor for the granulocyte colony-stimulating factor (G-CSF) cytokine. This cytokine-receptor complex stimulates the creation of granulocytes and activates neutrophils.

In the AA group analysis, two SNPs on chromosome 9q33.1 we associated with trajectory class membership at the genome-wide threshold (3.40×10^{-9} , lead SNP rs55736771). This SNP is located the third intron of the *ASTN2* gene. The protein encoded by this gene is expressed in the brain and has been associated with age of Alzheimer's disease onset, schizophrenia, and neurodevelopmental disorders in males^{46–48}.

Given that rs2814778 polymorphism in the *ACKR1/DARC* gene has a strong effect on resting-state WBC count among individuals of AA and, in our analysis, Class 1 trajectory members had a significantly higher median WBC count compared to Class 2 members, we were concerned that median WBC count was confounding an association between rs2814778 and the trajectory phenotype. To assess potential confounding, we ran an additional set of GWAS analyses, this time adjusting for median WBC count. A comparison of our minimally- and fully-adjusted models is presented in Table 3.

Adjusting for median WBC removed the association signal at rs2814778 in both the joint and AA group analyses and attenuated the association on chromosome 9. The association between trajectory class membership and rs12094900 was attenuated slightly, after

adjustment for median WBC count ($p=3.62 \times 10^{-7}$). Interestingly, adjusting for median WBC count revealed a new region of interest for association with trajectory class membership on chromosome 11q13.4 in both the joint and EA group analyses (1.10×10^{-7} , joint analysis lead SNP rs143804085; 3.21×10^{-7} , EA group analysis lead SNP rs79852880). The regional association plot for the joint analysis shows a large number of SNPs in high LD with rs143804085, with an equivalent level of association, across a four megabase region (Figure 4). This region contains six genes; rs143804085 falls within an intron of *RNF169*, the five remaining genes are: *CHRD2*, *MIR4696*, *XRR1*, *SPCS2*, and *NEU3*. The ring finger protein specified by *RNF169* is involved with the regulation of DNA double strand break repair⁴⁹. Chordin-like 2 protein, encoded by *CHRD2*, associates with members of the TGF- superfamily and may play a role in myoblast and osteoblast differentiation and maturation⁵⁰. *MIR4696* encodes a microRNA, which are involved in post-translational gene expression regulation⁵¹. The *XRR1* gene product is believed to regulate cell response to X-radiation exposure; the lead SNP in the EA group analysis falls within the *XRR1* gene⁵². *SPCS2* encodes a subunit of the microsomal signal peptidase complex, which removes signaling peptides from newly-formed proteins as they move to the endoplasmic reticulum⁵³. Neuraminidase 3, the gene product of *NEU3*, catabolizes gangliosides in the brain, thereby regulating neuronal function⁵⁴.

Discussion

The LCMM we used to identify unobserved heterogeneity in these longitudinal data for deep phenotype discovery identified two distinct latent resting-state WBC count trajectories. The major class trajectory, Class 2, was slightly U-shaped (concave up), with the point of inflection at approximately 60 years of age. The predicted trajectory corresponded well to previous reports of WBC count change over the lifespan^{9,10}. The Class 1 steady-state WBC count trajectory decreased across the lifespan and may indicate individuals with important differences in inflammation and immune health. While the average posterior probabilities of each class indicate distinct trajectories we discovered by the LCMM, the entropy value of 0.27 suggests a degree of imprecision in the clustering of participants into trajectory classes⁵⁵. Here, the low entropy value likely reflects the crossing in the predicted trajectories. Individuals with most of their data in the region of the crossing have posterior probabilities of class membership close to 50% for both classes and entropy nearing zero, driving the entropy of the whole model down, despite the moderately high average posterior probabilities for trajectory class membership.

Though not free from misclassification, using trajectory class membership predicted by LCMM as our phenotype of interest, we identified three regions associated with longitudinal change in WBC level in our GWAS; two on chromosome 1 at 1p34.3 and 1q23.2, and one at chromosome 9q33.1. Our previous work with this cohort identified WBC count quantitative trait loci in the 1q23.2 region at the *ACKR1/DARC* locus among those of AA and the 17q21.1 region among those of European ancestry²⁶. In this study, we found the T allele of the rs2814778 polymorphism in the *ACKR1/DARC* was also associated with decreasing WBC count with age (Class 1 phenotype) in the AA group. Individuals that carry the T allele -- all of our participants of EA and 37% of our participants of AA -- express the Duffy antigen on their RBCs. The Duffy antigen is a chemokine receptor and has been found to

preferentially bind inflammatory chemokines⁵⁶. Given the proinflammatory nature of the Duffy antigen, we would expect to see a positive association between the T allele of rs2814778 and the Class 2 phenotype, but this is not the case. Rather, because there was a significant difference in median WBC count between trajectory phenotype classes, we believe the association to Duffy region is reflecting the strong QTL there. Indeed, when we adjusted our analysis for potential confounding by median WBC count, the association between trajectory class membership and rs2814778 disappeared.

Similarly, adjusting for median WBC count removed the suggested association at chromosome 9q33.1. The associated SNPs are in *ASTN2*. The gene appears to be primarily expressed in the brain, prostate, and testis, with a lower expression level in the adrenal gland. Variation in this gene has been associated with neurological disorders^{46–48}. Though *ASTN2* has also been associated with osteoarthritis, that mechanism was attributable to influences on femur shape⁵⁷. There is no obvious mechanism for an association with median WBC count or trajectory class.

The association with region on 1p34.3 and WBC trajectory was slightly attenuated after adjusting for potential confounding by median WBC count, but the signal remained. Of the genes in this region, *CSF3R* is the most biologically plausible candidate explaining this association. *CSF3R* encodes the receptor for the granulocyte colony-stimulating factor (G-CSF) cytokine. The gene which encodes this cytokine, *CSF3*, is in close proximity to the 17q21.1 locus, which we found to be associated with WBC count in a previous study of this cohort²⁶. G-CSF, working with its receptor expressed on the surface of hematopoietic progenitor cells and neutrophilic granulocytes, stimulates granulopoiesis and activates neutrophils. Deficiency in G-CSF is associated with severe neutropenia, and G-CSF therapy is the major treatment for neutropenia, regardless of cause^{58,59}.

Adjusting for median WBC count revealed an additional area of interest on chromosome 11q13.4, with several SNPs in high LD across a four megabase region, just above the genome-wide significance threshold. Of the six genes in this region, two (*RNF169* and *XRRAI*) are involved in DNA repair and cell cycle arrest and are the most biologically plausible in relation to WBC count trajectory, as bone marrow, which gives rise to the WBCs, is the most rapidly replacing tissue in the body. In response to DNA damage, RNF169 protein negatively regulates the ubiquitin-dependent signaling cascade for double strand break repair, turning off the DNA damage signal and promoting mitosis after cell recovery⁶⁰. *RNF169* is also overexpressed in peripheral blood mononuclear cells, but not granulocytes (www.genecards.org)⁶¹. Similarly, *XRRAI* appears to regulate the cell cycle in response to X-radiation. It is expressed in normal tissue, including WBCs, as well as cancer cells⁵². This LD region is associated with decreasing resting-state WBC count over time and polymorphisms within may be contributing to this phenotype by decreasing mitosis or increasing apoptosis rates in response to cell damage with age.

While we successfully revealed two latent WBC count trajectory phenotypes in our longitudinal EMR-derived data and found novel genetic associations with these trajectories, it is difficult to determine the biological or clinical relevance of these trajectory phenotypes

with the available demographic and diagnosis codes in our dataset, given their complex relationships.

We used very little *a priori* biological knowledge to inform model construction. Incorporating longitudinal measures of BMI and smoking status may improve the precision of discovered trajectory phenotypes and make for easier interpretation of clinical relevance. However, age matched BMI measurements were available for less than half of participants with longitudinal WBC count data and LCMM with multiple imputed datasets is computationally intensive. Smoking status of eMERGE participants was not available. Care should be taken when incorporating longitudinal EHR covariate data into analyses that require “complete” data as this requirement can bias the individuals available for the study⁶². Indeed, requiring a minimum number of longitudinal WBC count data points preferentially excluded younger eMERGE participants and participants with AA.

To further the goals of precision medicine, more precise phenotyping of complex disease is needed. The pattern of change over time can be controlled by loci not apparent when considering cross-sectional data. Though it has limitations, we have shown that trajectory analysis with LCMM is a useful tool to integrate longitudinal measures for deep phenotype construction. This method was also a fruitful way to partition phenotypic heterogeneity, with respect to gene discovery.

Materials and Methods

Participants

Resting-state WBC count data was identified from the EMR of 14 018 participants in the electronic Medical Records and Genomics (eMERGE) Network. Currently, the eMERGE Network is a consortium of twelve U.S. cohorts linked to EMR data for conducting large-scale, high-throughput genetic research⁶³. Our study used a subset of participating sites include the following: (1) Kaiser Permanente Washington (formerly Group Health Cooperative) and University of Washington partnership, Seattle, WA; (2) Marshfield Clinic, Marshfield, WI; (3) Mayo Clinic, Rochester, MN; (4) Northwestern University, Evanston, IL; and (5) Vanderbilt University, Nashville, TN⁶⁴. Participant informed consent was obtained by the recruiting eMERGE site. The study was approved by each site’s internal Institutional Review Board.

The WBC data extraction algorithm used has been previously published²⁶. Generally, we excluded participants and visits where the participant’s diagnosis or medication use may have perturbed WBC count outside of resting-state. To further exclude abnormal WBC counts, we excluded visits where the WBC count was outside two standard deviations of the median, within subject. Additionally, identifying a trajectory-based deep phenotype through LCMM requires that participants have one more WBC count measure than the order of magnitude of the model fitted. Because WBC count appears to be slightly U-shaped over the lifespan, we chose to fit a quadratic model and therefore participants with fewer than three WBC count visits were excluded. A summary of subject- and visit-level exclusion criteria are listed in Table 4. After exclusions, 9 742 participants were available for trajectory model analysis.

Deep phenotype

LCMM was fitted using the R package `lcmm`^{45,65}. This method uses a modified Marquardt iterative algorithm and maximum likelihood theory to estimate the LCMM. The Bayesian Information Criterion (BIC), posterior probability of class membership, and percentage of class membership were used to evaluate model fit. When comparing models, that with the lower BIC is the preferred model. An average posterior probability of class membership of 70% and 10% of the population assigned to the minor class are also indications that distinct trajectories are being modelled^{33,66}. We followed the model fitting procedure suggested by Andruff and colleagues, first fitting the mixed model specifying one latent class and then adding classes until the model fit criteria are satisfied⁶⁷. In all models, age at event (i.e. WBC count draw) was our time varying covariate, fit with a linear and quadratic fixed effect and a linear random effect. The `lcmm` package calculates the posterior probability of class membership for each participant in each class using Bayes Theorem as the probability of class membership given the participant's data and the model fit⁴⁵. Participants were then assigned membership in a class based on the class for which they had the highest posterior probability of class membership. Participant class membership was used as the phenotype in subsequent analyses. To further characterize the discrimination of the latent classes, we calculated the entropy of each model using the participant posterior probability of class membership, as suggested by van de Schoot and colleagues⁶⁸. Entropy, E_k , was calculated as: $E_k = 1 - \left(\sum_i \sum_k \hat{P}_{ik} \ln \hat{P}_{ik} \right) / (I \ln K)$, for i individuals in k classes. E_k is 0 when the posterior probabilities for all participants are equal, indicating no separation of classes, and 1 when the classes are discrete partitions⁵⁵.

Statistics

We described the characteristics of each class, comparing several covariates. Continuous values were compared using the Kruskal-Wallis rank sum test and categorical values using the chi-squared test. Median age-at-event, WBC count and BMI were first calculated within subject and then the median was calculated for each class, i.e. the grand median. All analyses were performed in R version 3.3.0⁶⁹. R code is available upon request.

Genotyping

As reported previously²⁶, most subjects were genotyped on the Illumina Human660W-Quad v1_A (660 W) genotyping platform. A subset of subjects who were self-reported (Northwestern University) or observer reported (Vanderbilt University) to have African ancestry was genotyped on the Illumina Human1M-Duo (1 M) genotyping platform. Genotyping calls for both platforms were made at CIDR and Broad using BeadStudio version 3.3.7 and Gentrain version 1.0. Both samples and SNPs were assessed for quality and subsequently filtered from the production data, if thresholds were not met⁷⁰. We produced a unified genotype variant dataset for the different genotyping platforms, imputing missing genotype calls using the Michigan Imputation Server with the HRC1.1 haplotype reference set^{71–73}. Cryptic relatedness was assessed for all sites, and pairs at half-sibling level (kinship coefficient $\theta = k_1/4 + k_2/2 = 1/8$) or higher were randomly broken (by dropping one) before assessing whole-genome association. Subjects identified for filtering at each particular site through the quality control/quality assurance process were subsequently

filtered for the entire merged dataset. The eMERGE imputed genotype dataset is available on dbGaP study accession phs001584.v1.p1.

Assigning genetically-determined ancestry (GDA)

Principal components analysis was performed using independent, autosomal SNPs with missing call rates < 5.0% and minor allele frequency > 5.0% across the merged data set of 17 150 unique subjects, as described previously²⁶. We used k-means clustering on the first two principal components, specifying three groups, to assign an individual's GDA as either European Ancestry (EA), African Ancestry (AA) or Asian Ancestry⁷⁴.

GWAS

GWAS analyses of the WBC count trajectory phenotypes discovered in the phenotyping phase were performed in PLINK⁷⁵. The minor class, Class 1, was set as the “at risk” phenotype. We performed analyses pooling subjects of all genetic ancestries (Joint analysis) as well as analyses stratified by GDA (EA group and AA group only, due to small Asian Ancestry sample size). All analyses were adjusted for sex, median BMI, and principal components 1 and 2, to account for possible confounding by ancestry. Joint and EA group analyses were also adjusted for study site, but this covariate was dropped from the AA group analysis due to collinearity. We assumed an additive genetic model, with SNP genotypes were coded as 0, 1 and 2 copies of the minor allele. We filtered out SNPs with a minor allele frequency of < 0.03. Manhattan and Q-Q plots we made using the GWAS Tools R package⁷⁶. Regional association plots were generated by LocusZoom⁷⁷.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The eMERGE Network was initiated and funded by NHGRI through the following grants:

Phase III: U01HG8657 (Kaiser Permanente Washington, formerly Group Health Cooperative/University of Washington, Seattle); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine)

Phase II: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center), U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers, and U01HG004438 (CIDR) serving as a Sequencing Center.

Phase I: U01-HG-004610 (Group Health Cooperative/University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

References

1. Shim WS, Kim HJ, Kang ES, Ahn CW, Lim SK, Lee HC et al. The association of total and differential white blood cell count with metabolic syndrome in type 2 diabetic patients. *Diabetes Res Clin Pract* 2006; 73: 284–291. [PubMed: 16563549]
2. Chao T-T, Hsieh C-H, Lin J-D, Wu C-Z, Hsu C-H, Pei D et al. Use of white blood cell counts to predict metabolic syndrome in the elderly: a 4 year longitudinal study. *Aging Male* 2014; 17: 230–237. [PubMed: 24456526]
3. Pei C, Chang J-B, Hsieh C-H, Lin J-D, Hsu C-H, Pei D et al. Using white blood cell counts to predict metabolic syndrome in the elderly: A combined cross-sectional and longitudinal study. *Eur J Intern Med* 2015; 26: 324–329. [PubMed: 25910617]
4. Babio N, Ibarrola-Jurado N, Bulló M, Martínez-González MÁ, Wärnberg J, Salaverría I et al. White blood cell counts as risk markers of developing metabolic syndrome and its components in the PREDIMED study. *PLoS One* 2013; 8: e58354. [PubMed: 23526980]
5. Huh JY, Ross GW, Chen R, Abbott RD, Bell C, Willcox B et al. Total and differential white blood cell counts in late life predict 8-year incident stroke: the Honolulu Heart Program. *J Am Geriatr Soc* 2015; 63: 439–446. [PubMed: 25739422]
6. Loimaala A, Rontu R, Vuori I, Mercuri M, Lehtimäki T, Nenonen A et al. Blood leukocyte count is a risk factor for intima-media thickening and subclinical carotid atherosclerosis in middle-aged men. *Atherosclerosis* 2006; 188: 363–369. [PubMed: 16378612]
7. Nilsson G, Hedberg P, Ohrvik J. White blood cell count in elderly is clinically useful in predicting long-term survival. *J Aging Res* 2014; 2014: 475093. [PubMed: 24624295]
8. Ruggiero C, Metter EJ, Cherubini A, Maggio M, Sen R, Najjar SS et al. White blood cell count and mortality in the Baltimore Longitudinal Study of Aging. *J Am Coll Cardiol* 2007; 49: 1841–1850. [PubMed: 17481443]
9. Chmielewski PP, Boryśłowski K, Chmielewicz K, Chmielewicz J, Strzelec B. The association between total leukocyte count and longevity: Evidence from longitudinal and cross-sectional data. *Ann Anat* 2016; 204: 1–10. [PubMed: 26434754]
10. Brown DW, Ford ES, Giles WH, Croft JB, Balluz LS, Mokdad AH. Associations between White Blood Cell Count and Risk for Cerebrovascular Disease Mortality: NHANES II Mortality Study, 1976–1992. *Ann Epidemiol* 2004; 14: 425–430. [PubMed: 15246331]
11. Ahmadi-Abhari S, Luben RN, Wareham NJ. Seventeen year risk of all-cause and cause-specific mortality associated with C-reactive protein, fibrinogen and leukocyte count in men and women: the EPIC-Norfolk *European journal of* 2013 <http://link.springer.com/article/10.1007/s10654-013-9819-6>.
12. Collier BS. Leukocytosis and ischemic vascular disease morbidity and mortality: is it time to intervene? *Arterioscler Thromb Vasc Biol* 2005; 25: 658–670. [PubMed: 15662026]
13. Smith MR, Kinmonth A-L, Luben RN, Bingham S, Day NE, Wareham NJ et al. Smoking status and differential white cell count in men and women in the EPIC-Norfolk population. *Atherosclerosis* 2003; 169: 331–337. [PubMed: 12921986]
14. Schwartz J, Weiss ST. Cigarette smoking and peripheral blood leukocyte differentials. *Ann Epidemiol* 1994; 4: 236–242. [PubMed: 8055125]
15. 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: 486–492. [PubMed: 17404350]
16. Dixon JB, O'Brien PE. Obesity and the white blood cell count: changes with sustained weight loss. *Obes Surg* 2006; 16: 251–257. [PubMed: 16545154]
17. Church TS, Finley CE, Earnest CP, Kampert JB, Gibbons LW, Blair SN. Relative associations of fitness and fatness to fibrinogen, white blood cell count, uric acid and metabolic syndrome. *Int J Obes Relat Metab Disord* 2002; 26: 805–813. [PubMed: 12037651]
18. Womack J, Tien PC, Feldman J, Shin JH, Fennie K, Anastos K et al. Obesity and immune cell counts in women. *Metabolism* 2007; 56: 998–1004. [PubMed: 17570264]
19. Pilia G, Chen W-M, Scuteri A, Orrú M, Albai G, Dei M et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* 2006; 2: e132. [PubMed: 16934002]

20. Haddy TB, Rana SR, Castro O. Benign ethnic neutropenia: what is a normal absolute neutrophil count? *J Lab Clin Med* 1999; 133: 15–22. [PubMed: 10385477]
21. Rana SR, Castro OL, Haddy TB. Leukocyte counts in 7,739 healthy black persons: effects of age and sex. *Ann Clin Lab Sci* 1985; 15: 51–54. [PubMed: 3970517]
22. Nalls MA, Wilson JG, Patterson NJ, Tandon A, Zmuda JM, Huntsman S et al. Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am J Hum Genet* 2008; 82: 81–87. [PubMed: 18179887]
23. 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: e1000360. [PubMed: 19180233]
24. 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. [PubMed: 21738479]
25. Li J, Glessner JT, Zhang H, Hou C, Wei Z, Bradfield JP et al. GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Hum Mol Genet* 2013; 22: 1457–1464. [PubMed: 23263863]
26. Crosslin DR, McDavid A, Weston N, Nelson SC, Zheng X, Hart E et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum Genet* 2012; 131: 639–652. [PubMed: 22037903]
27. Keller MF, Reiner AP, Okada Y, van Rooij FJA, Johnson AD, Chen M-H et al. Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum Mol Genet* 2014; 23: 6944–6960. [PubMed: 25096241]
28. Teliëps T, Köhler M, Treise I, Foertsch K, Adler T, Busch DH et al. Longitudinal Frequencies of Blood Leukocyte Subpopulations Differ between NOD and NOR Mice but Do Not Predict Diabetes in NOD Mice. *J Diabetes Res* 2016; 2016: 4208156. [PubMed: 26966692]
29. Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M. The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS One* 2013; 8: e76295. [PubMed: 24146854]
30. Tracy RP. ‘Deep phenotyping’: characterizing populations in the era of genomics and systems biology. *Current opinion in lipidology* 2008; 19: 151–157. [PubMed: 18388695]
31. Chiu Y-F, Justice AE, Melton PE. Longitudinal analytical approaches to genetic data. *BMC Genet* 2016; 17 Suppl 2: 4. [PubMed: 26866891]
32. Nagin DS. Group-based trajectory modeling: an overview. *Ann Nutr Metab* 2014; 65: 205–210. [PubMed: 25413659]
33. Strauss VY, Jones PW, Kadam UT, Jordan KP. Distinct trajectories of multimorbidity in primary care were identified using latent class growth analysis. *J Clin Epidemiol* 2014; 67: 1163–1171. [PubMed: 25063556]
34. Gunzler DD, Morris N, Perzynski A, Ontaneda D, Briggs F, Miller D et al. Heterogeneous depression trajectories in multiple sclerosis patients. *Mult Scler Relat Disord* 2016; 9: 163–169. [PubMed: 27645367]
35. Baker E, Iqbal E, Johnston C, Broadbent M, Shetty H, Stewart R et al. Trajectories of dementia-related cognitive decline in a large mental health records derived patient cohort. *PLoS One* 2017; 12: e0178562. [PubMed: 28591196]
36. Pugh SJ, Albert PS, Kim S, Grobman W, Hinkle SN, Newman RB et al. Patterns of gestational weight gain and birthweight outcomes in the Eunice Kennedy Shriver National Institute of Child Health and Human Development Fetal Growth Studies-Singletons: a prospective study. *Am J Obstet Gynecol* 2017. doi:10.1016/j.ajog.2017.05.013.
37. Justice AE, Howard AG, Chittoor G, Fernandez-Rhodes L, Graff M, Voruganti VS et al. Genome-wide association of trajectories of systolic blood pressure change. *BMC Proc* 2016; 10: 321–327. [PubMed: 27980656]
38. Dick DM, Cho SB, Latendresse SJ, Aliev F, Nurnberger JI Jr, Edenberg HJ et al. Genetic influences on alcohol use across stages of development: GABRA2 and longitudinal trajectories of

- drunkenness from adolescence to young adulthood. *Addict Biol* 2014; 19: 1055–1064. [PubMed: 23692184]
39. Lessov-Schlaggar CN, Kristjansson SD, Bucholz KK, Heath AC, Madden PAF. Genetic influences on developmental smoking trajectories. *Addiction* 2012; 107: 1696–1704. [PubMed: 22385035]
 40. Riglin L, Collishaw S, Thapar AK, Dalsgaard S, Langley K, Davey Smith G et al. Contribution of Genetic Risk Variants to Attention-Deficit Hyperactivity Disorder Trajectories in the General Population. *JAMA Psychiatry* 2016; : 1–8.
 41. Holliday EG, McLean DE, Nyholt DR, Mowry BJ. Susceptibility locus on chromosome 1q23–25 for a schizophrenia subtype resembling deficit schizophrenia identified by latent class analysis. *Arch Gen Psychiatry* 2009; 66: 1058–1067. [PubMed: 19805696]
 42. Chen WJ. Taiwan Schizophrenia Linkage Study: lessons learned from endophenotype-based genome-wide linkage scans and perspective. *Am J Med Genet B Neuropsychiatr Genet* 2013; 162B: 636–647. [PubMed: 24132895]
 43. Bureau A, Croteau J, Tayeb A, Mérette C, Labbe A. Latent class model with familial dependence to address heterogeneity in complex diseases: adapting the approach to family-based association studies. *Genet Epidemiol* 2011; 35: 182–189. [PubMed: 21308764]
 44. Wickrama KKAS, O'Neal CW, Lee TK. Early community context, genes, and youth body mass index trajectories: an investigation of gene-community interplay over early life course. *J Adolesc Health* 2013; 53: 328–334. [PubMed: 23711506]
 45. Proust-Lima C, Philipps V, Lique B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcomm. *Journal of Statistical Software, Articles* 2017; 78: 1–56.
 46. Lionel AC, Tammimies K, Vaags AK, Rosenfeld JA, Ahn JW, Merico D et al. Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Hum Mol Genet* 2014; 23: 2752–2768. [PubMed: 24381304]
 47. Wang K-S, Tonarelli S, Luo X, Wang L, Su B, Zuo L et al. Polymorphisms within ASTN2 gene are associated with age at onset of Alzheimer's disease. *J Neural Transm* 2015; 122: 701–708. [PubMed: 25410587]
 48. Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, Strengman E, Genetic Risk and Outcome in Psychosis (GROUP) Consortium, Sabatti C et al. Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet* 2008; 83: 504–510. [PubMed: 18940311]
 49. Poulsen M, Lukas C, Lukas J, Bekker-Jensen S, Mailand N. Human RNF169 is a negative regulator of the ubiquitin-dependent response to DNA double-strand breaks. *J Cell Biol* 2012; 197: 189–199. [PubMed: 22492721]
 50. Oren A, Toporik A, Biton S, Almogly N, Eshel D, Bernstein J et al. hCHL2, a novel chordin-related gene, displays differential expression and complex alternative splicing in human tissues and during myoblast and osteoblast maturation. *Gene* 2004; 331: 17–31. [PubMed: 15094188]
 51. Hammond SM. An overview of microRNAs. *Adv Drug Deliv Rev* 2015; 87: 3–14. [PubMed: 25979468]
 52. Mesak FM, Osada N, Hashimoto K, Liu QY, Ng CE. Molecular cloning, genomic characterization and over-expression of a novel gene, XRR1, identified from human colorectal cancer cell HCT116Clone2_XRR and macaque testis. *BMC Genomics* 2003; 4: 32. [PubMed: 12908878]
 53. Kalies KU, Hartmann E. Membrane topology of the 12- and the 25-kDa subunits of the mammalian signal peptidase complex. *J Biol Chem* 1996; 271: 3925–3929. [PubMed: 8632014]
 54. Pan X, De Aragão CDBP, Velasco-Martin JP, Priestman DA, Wu HY, Takahashi K et al. Neuraminidases 3 and 4 regulate neuronal function by catabolizing brain gangliosides. *FASEB J* 2017; 31: 3467–3483. [PubMed: 28442549]
 55. Jedidi K, Ramaswamy V, Desarbo WS. A maximum likelihood method for latent class regression involving a censored dependent variable. *Psychometrika* 1993; 58: 375–394.
 56. Gardner L, Patterson AM, Ashton BA, Stone MA, Middleton J. The human Duffy antigen binds selected inflammatory but not homeostatic chemokines. *Biochem Biophys Res Commun* 2004; 321: 306–312. [PubMed: 15358176]

57. Lindner C, Thiagarajah S, Wilkinson JM, Panoutsopoulou K, Day-Williams AG, arcOGEN Consortium et al. Investigation of association between hip osteoarthritis susceptibility loci and radiographic proximal femur shape. *Arthritis Rheumatol* 2015; 67: 2076–2084. [PubMed: 25939412]
58. Ohno R Granulocyte colony-stimulating factor, granulocyte-macrophage colony-stimulating factor and macrophage colony-stimulating factor in the treatment of acute myeloid leukemia and acute lymphoblastic leukemia. *Leuk Res* 1998; 22: 1143–1154. [PubMed: 9922078]
59. Zeidler C, Welte K. Kostmann syndrome and severe congenital neutropenia. *Semin Hematol* 2002; 39: 82–88. [PubMed: 11957189]
60. Chen J, Feng W, Jiang J, Deng Y, Huen MSY. Ring finger protein RNF169 antagonizes the ubiquitin-dependent signaling cascade at sites of DNA damage. *J Biol Chem* 2012; 287: 27715–27722. [PubMed: 22733822]
61. Fishilevich S, Zimmerman S, Kohn A, Iny Stein T, Olender T, Kolker E et al. Genic insights from integrated human proteomics in GeneCards. *Database* 2016; 2016. doi:10.1093/database/baw030.
62. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K et al. Biases introduced by filtering electronic health records for patients with ‘complete data’. *J Am Med Inform Assoc* 2017; 24: 1134–1141. [PubMed: 29016972]
63. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011; 4: 13. [PubMed: 21269473]
64. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008; 84: 362–369. [PubMed: 18500243]
65. CRAN - Package lcomm. <https://cran.r-project.org/web/packages/lcomm/index.html> (accessed 29 Jun2017).
66. Chassin L, Fora DB, King KM. Trajectories of alcohol and drug use and dependence from adolescence to adulthood: the effects of familial alcoholism and personality. *J Abnorm Psychol* 2004; 113: 483–498. [PubMed: 15535782]
67. Andruff H, Carraro N, Thompson A, Gaudreau P. Latent class growth modelling: A tutorial. *Tutor Quant Methods Psychol* 2009; 5: 11–24.
68. van de Schoot R, Sijbrandij M, Winter SD, Depaoli S, Vermunt JK. The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies. *Struct Equ Modeling* 2017; 24: 451–467.
69. R Core Team. R: A Language and Environment for Statistical Computing 2017<https://www.R-project.org/>.
70. Zuvich RL, Armstrong LL, Bielinski SJ, Bradford Y, Carlson CS, Crawford DC et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet Epidemiol* 2011; 35: 887–898. [PubMed: 22125226]
71. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; 48: 1279–1283. [PubMed: 27548312]
72. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016; 48: 1284–1287. [PubMed: 27571263]
73. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016; 48: 1443–1448. [PubMed: 27694958]
74. Stanaway IB, Hall TO, Rosenthal EA, Palmer M, Naranbhai V, Knevel R et al. The eMERGE Genotype Set of 83,717 Subjects Imputed to ~40 Million Variants Genome Wide and Association with the Herpes Zoster Medical Record Phenotype. *Genet Epidemiol* accepted 28-8-2018.
75. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81: 559–575. [PubMed: 17701901]
76. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 2012; 28: 3329–3331. [PubMed: 23052040]

77. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; 26: 2336–2337. [PubMed: 20634204]

Figure 1. Predicted mean class-specific trajectory (A) and observed mean class-specific trajectory and 95% confidence interval (B) of WBC count by age from the LCMM.

Figure 2.

Manhattan and Q-Q plots summarizing the results of the WBC count trajectory phenotype GWAS, joint and GDA stratified analyses. A) Manhattan plot for the joint analysis. B) Q-Q plot for the joint analysis. C) Manhattan plot for the AA analysis. D) Q-Q plot for the AA analysis. E) Manhattan plot for the EA analysis. F) Q-Q plot for the EA analysis.

Figure 3.
Regional association plot for chromosome 1p34.3, joint analysis.

Figure 4.
Regional association plot for chromosome 11p13.4, joint analysis.

Table 1.

LCMM fit statistics.

Model	Link	Classes	BIC	Entropy	Average Posterior Probability	Sample Size (%)
1a	linear	1	339342.22	--	--	9742 (100)
1b	beta	1	301289.38	--	--	9742 (100)
2	beta	2	299820.82	0.27	0.73	3349 (35) 6292 (65)
3	beta	3	299145.93	0.44	0.75	209 (2) 6725 (70) 2702 (28)
4	beta	4	298777.82	0.52	0.77	151 (1) 7639 (78) 1595 (16) 357 (4)

Table 2.

Descriptive characteristics of each trajectory class.

	Class 1 (N=3349)	Class 2 (N=6292)	p-value
MEDIAN [IQR]			
OBSERVATIONS	8 [5–12]	7 [4–10]	< 0.0001
AGE AT EVENT	62.5 [52.3–71.8]	64.6 [55.1–72.9]	< 0.0001
NUMBER YEARS FOLLOW-UP	11.7 [6.5–17.5]	10.9 [5.9–15.9]	< 0.0001
WBC COUNT	7.4 [6.2–8.6]	6.2 [5.4–7.1]	< 0.0001
BMI	28.9 [25.1–33.3]	27.9 [24.9–31.3]	< 0.0001
%			
MALE	41	45	< 0.0001
GENETICALLY DETERMINED ANCESTRY			
EUROPEAN	88	91	0.0002
AFRICAN	11	8	
ASIAN	1	1	
SITE			
GROUP HEALTH	21	22	<0.0001
MARSHFIELD	37	31	
MAYO	12	23	
NORTHWESTERN	11	7	
VANDERBILT	19	16	

Table 3.

Comparison of significant GWAS results before and after adjusting for median WBC count.

CHR	SNP	Minor allele	Closest gene		OR	P	OR*	P*
Joint								
1	rs2814778	C	ACKR1	exon	0.50	9.83E-09	0.68	0.01131
1	rs12094900	G	CSF3R	intergenic	0.83	3.54E-08	0.80	3.62E-07
3	rs75135222	C	DCPIA	intron	1.46	5.53E-07	1.46	9.59E-05
11	rs143804085	G	RNF169	intron	1.31	3.97E-05	1.55	1.10E-07
11	rs117711119	A	XRRA1	intron	1.30	6.52E-05	1.55	1.86E-07
11	rs117258637	G	SPCS2	intron	1.29	9.74E-05	1.54	2.29E-07
11	rs11606575	G	CEP164	intron	0.86	2.11E-04	0.76	8.30E-07
AA								
1	rs2814778	T	ACKR1	exon	2.15	9.59E-09	1.26	0.2174
1	rs4657616	G	POP3	intron	2.69	1.24E-07	1.59	0.06193
1	rs856068	C	IFI16	intron	1.97	2.94E-07	1.40	0.05618
1	rs2501339	G	VSIG8	intron	2.17	6.35E-07	1.76	0.006375
6	rs7761344	A	SASH1	intron	1.42	7.89E-04	2.12	5.60E-07
9	rs55736771	A	ASTN2	intron	1.89	3.40E-09	1.79	6.80E-05
EA								
11	rs79852880	G	XRRA1	intron	1.30	7.65E-05	1.54	3.21E-07
11	rs143804085	G	RNF169	intron	1.29	1.65E-04	1.53	5.42E-07
11	rs117258637	G	SPCS2	intron	1.27	2.86E-04	1.51	7.24E-07

Table 4.

Summary of subject- and visit-level exclusions for EMR WBC count data.

Subject-level exclusion criteria	
Any indication at any time of HIV	
Dialysis at any time	
<3 WBC count records	
Visit-level exclusion criteria	
Inpatient or emergency visit	
Splenectomy record prior to lab	
Prior diagnosis of myelodysplastic syndrome	
Medications with minor impacts on WBC (aspirin at high doses)	
Strongly immune affecting medications (oral or IV steroids chemotherapeutic agents such as methotrexate)	
Indication of concurrent “active chemotherapy” regimen 6 months prior to 3 months after index visit	
Prior indication of Alzheimer’s disease	
Blood dyscrasia (leukemia, myeloma, bone marrow failure, aplastic anemia, etc.)	
“Active infection” in prior or subsequent 30 days	
Other acute and chronic infections	
Within subject outlier WBC count	