

# Whole genome sequencing based analysis of inflammation biomarkers in the Trans-Omics for Precision Medicine (TOPMed) consortium

Min-Zhi Jiang <sup>1,†</sup>, Sheila M. Gaynor <sup>2,3,†</sup>, Xihao Li <sup>1,4</sup>, Eric Van Buren<sup>2</sup>, Adrienne Stilp <sup>5</sup>, Erin Buth<sup>5</sup>, Fei Fei Wang<sup>5</sup>, Regina Manansala <sup>6</sup>, Stephanie M. Gogarten <sup>5</sup>, Zilin Li<sup>7</sup>, Linda M. Polfus <sup>8</sup>, Shabnam Salimi<sup>9</sup>, Joshua C. Bis <sup>10</sup>, Nathan Pankratz <sup>11</sup>, Lisa R. Yanek <sup>12</sup>, Peter Durda<sup>13</sup>, Russell P. Tracy<sup>13</sup>, Stephen S. Rich<sup>14</sup>, Jerome I. Rotter <sup>15</sup>, Braxton D. Mitchell <sup>16</sup>, Joshua P. Lewis<sup>16</sup>, Bruce M. Psaty <sup>10,17</sup>, Katherine A. Pratte<sup>18</sup>, Edwin K. Silverman<sup>19</sup>, Robert C. Kaplan<sup>20</sup>, Christy Avery<sup>21</sup>, Kari E. North <sup>21</sup>, Rasika A. Mathias<sup>22</sup>, Nauder Faraday <sup>23</sup>, Honghuang Lin <sup>24</sup>, Biqi Wang<sup>24</sup>, April P. Carson <sup>25</sup>, Arnita F. Norwood<sup>25</sup>, Richard A. Gibbs<sup>26</sup>, Charles Kooperberg <sup>27</sup>, Jessica Lundin <sup>27</sup>, Ulrike Peters<sup>27</sup>, Josée Dupuis <sup>28,29</sup>, Lifang Hou<sup>30</sup>, Myriam Fornage <sup>31</sup>, Emelia J. Benjamin <sup>32,33,34</sup>, Alexander P. Reiner<sup>35</sup>, Russell P. Bowler <sup>18</sup>, Xihong Lin <sup>2</sup>, Paul L. Auer <sup>36,†</sup>, Laura M. Raffield <sup>1,†,\*</sup>, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Inflammation Working Group

<sup>1</sup>Department of Genetics, University of North Carolina at Chapel Hill, 120 Mason Farm Road, Chapel Hill, NC 27599, United States

<sup>2</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA 02115, United States

<sup>3</sup>Regeneron Genetics Center, 777 Old Saw Mill River Road, Tarrytown, NY 10591, United States

<sup>4</sup>Department of Biostatistics, 135 Dauer Drive, 4115D McGavran-Greenberg Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

<sup>5</sup>Department of Biostatistics, 4333 Brooklyn Ave NE, University of Washington, Seattle, WA 98105, United States

<sup>6</sup>Centre for Health Economics Research & Modelling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO) WHO Collaborating Centre, University of Antwerp, Campus Drie Eiken - Building S; Universiteitsplein 1 2610 Antwerpen, Belgium

<sup>7</sup>School of Mathematics and Statistics, Northeast Normal University, 5268 Renmin Street, Changchun, JL 130024, China

<sup>8</sup>Advanced Analytics, Ambry Genetics, 1 Enterprise, Aliso Viejo, CA 92656, United States

<sup>9</sup>Department of Epidemiology and Public Health, Division of Gerontology, University of Maryland School of Medicine, 655 W. Baltimore Street, Baltimore, MD 21201, United States

<sup>10</sup>Cardiovascular Health Research Unit, Department of Medicine, University of Washington, 4333 Brooklyn Ave NE, Box 359458, Seattle, WA 98195, United States

<sup>11</sup>Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, 420 Delaware Street SE, Minneapolis, MN 55455, United States

<sup>12</sup>Department of Medicine, General Internal Medicine, Johns Hopkins University School of Medicine, 1830 E Monument St Rm 8024, Baltimore, MD 21287, United States

<sup>13</sup>Department of Pathology & Laboratory Medicine, University of Vermont Larner College of Medicine, 360 South Park Drive, Colchester, VT 05446, United States

<sup>14</sup>Center for Public Health Genomics, University of Virginia School of Medicine, 200 Jeanette Lancaster Way, Charlottesville, VA 22903, United States

<sup>15</sup>The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, 1124 W. Carson Street, Torrance, CA 90502, United States

<sup>16</sup>Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland School of Medicine, 670 W. Baltimore St., Baltimore, MD 21201, United States

<sup>17</sup>Departments of Epidemiology and Health Systems and Population Health, University of Washington, 4333 Brooklyn Ave NE, Seattle, WA 98101, United States

<sup>18</sup>Department of Medicine, Division of Pulmonary, Critical Care, and Sleep Medicine, National Jewish Health, 1400 Jackson Street, Denver, CO 80206, United States

<sup>19</sup>Department of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115, United States

<sup>20</sup>Department of Epidemiology and Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, United States

<sup>21</sup>Department of Epidemiology, University of North Carolina at Chapel Hill, 137 East Franklin Street, Chapel Hill, NC 27599, United States

<sup>22</sup>Department of Medicine, Allergy and Clinical Immunology, Johns Hopkins University School of Medicine, 5501 Hopkins Bayview Cir JHAAC Room 3B53, Baltimore, MD 21287, United States

<sup>23</sup>Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, 600 N Wolfe St, Baltimore, MD 21287, United States

<sup>24</sup>Department of Medicine, University of Massachusetts Chan Medical School, 55 Lake Ave North, Worcester, MA 01655, United States

<sup>25</sup>Department of Medicine, University of Mississippi Medical Center, 350 W. Woodrow Wilson Avenue, Suite 701, Jackson, MS 39213, United States

<sup>26</sup>Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, United States

<sup>27</sup>Division of Public Health Sciences, Fred Hutchinson Cancer Center, 1100 Fairview Avenue N, Seattle, WA 98109, United States

<sup>28</sup>Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 2001 McGill College Avenue, Montreal, QC H3A 1G1, Canada

<sup>29</sup>Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118, United States

<sup>30</sup>Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N Lake Shore Drive, Chicago, IL 60611, United States

<sup>31</sup>Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, 1825 Pressler Street, Houston, TX 77030, United States

<sup>32</sup>Department of Medicine, Cardiovascular Medicine, Boston Medical Center, Boston University Chobanian and Avedisian School of Medicine, 72 East Newton Street, Boston, MA 02118, United States

<sup>33</sup>Department of Epidemiology, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118, United States

<sup>34</sup>Boston University and National Heart, Lung, and Blood Institute's Framingham Heart Study, 73 Mount Wayte Ave #2, Framingham, MA 01702, United States

<sup>35</sup>Department of Epidemiology, University of Washington, 4333 Brooklyn Ave NE, Seattle, WA 98105, United States

<sup>36</sup>Division of Biostatistics, Institute for Health and Equity, and Cancer Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, United States

\*Corresponding author. 120 Mason Farm Road, 5042 Genetic Medicine Building, Chapel Hill NC 27599, United States. E-mail: laura\_raffield@unc.edu

†Min-Zhi Jiang, Sheila M. Gaynor, Paul L. Auer and Laura M. Raffield contributed equally.

## Abstract

Inflammation biomarkers can provide valuable insight into the role of inflammatory processes in many diseases and conditions. Sequencing based analyses of such biomarkers can also serve as an exemplar of the genetic architecture of quantitative traits. To evaluate the biological insight, which can be provided by a multi-ancestry, whole-genome based association study, we performed a comprehensive analysis of 21 inflammation biomarkers from up to 38465 individuals with whole-genome sequencing from the Trans-Omics for Precision Medicine (TOPMed) program (with varying sample size by trait, where the minimum sample size was  $n = 737$  for MMP-1). We identified 22 distinct single-variant associations across 6 traits—E-selectin, intercellular adhesion molecule 1, interleukin-6, lipoprotein-associated phospholipase A2 activity and mass, and P-selectin—that remained significant after conditioning on previously identified associations for these inflammatory biomarkers. We further expanded upon known biomarker associations by pairing the single-variant analysis with a rare variant set-based analysis that further identified 19 significant rare variant set-based associations with 5 traits. These signals were distinct from both significant single variant association signals within TOPMed and genetic signals observed in prior studies, demonstrating the complementary value of performing both single and rare variant analyses when analyzing quantitative traits. We also confirm several previously reported signals from semi-quantitative proteomics platforms. Many of these signals demonstrate the extensive allelic heterogeneity and ancestry-differentiated variant-trait associations common for inflammation biomarkers, a characteristic we hypothesize will be increasingly observed with well-powered, large-scale analyses of complex traits.

**Keywords:** inflammation biomarkers; whole genome sequencing data; rare variant aggregate test; genome-wide association studies; Trans-Omics for Precision Medicine (TOPMed) consortium

## Introduction

Chronic inflammation is a risk factor for many diseases including cardiovascular disease, asthma, cancer and diabetes [1–3]. Chronic inflammation has been assessed in human cohorts using a variety of immunoassay measured biomarker traits, particularly markers of innate immune system activation such as C-reactive protein (CRP) and interleukin 6 (IL-6) [2]. Though there is a strong influence of social and environmental factors, previous analyses, including genome-wide association studies (GWAS), have demonstrated an underlying genetic component to variance in these traits [4, 5]. Heritability of biomarkers of inflammation have been estimated, for instance, to be 25%–60% [6, 7] for IL-6 and 30%–45% [8–12] for CRP. However, most studies have only analyzed relatively small and ancestrally homogenous (mostly European ancestry) populations and as such have not fully elucidated the genetic influence on these traits [4, 13–16].

The National Heart Lung and Blood Institute's Trans-Omics for Precision Medicine (TOPMed) initiative has now generated whole genome sequencing data on > 150 000 individuals from diverse population-based cohorts enriched for heart, lung, and blood relevant disease traits. Novel ancestry-differentiated variant associations for CRP [17] (including confirmation of regulatory impacts *in vitro*) and E-selectin [18] reported in earlier TOPMed publications demonstrated the potential for genetic discovery for inflammation traits in these diverse cohorts. Thus, analysis of more biomarkers across a larger, more diverse set of samples with the addition of rare variant aggregate tests may identify additional associated individual variants and genomic regions. Here, we perform single variant and aggregate rare variant analyses across 21 inflammation-related biomarkers, some of which are in moderate to low correlation (Fig. S1), assessed in TOPMed cohort studies (Table 1), including performing detailed conditional analyses to identify distinct genetic association signals. Our results both inform our understanding of inflammation trait biology and of the expected findings for sequencing-based analyses of complex traits, particularly protein quantitative biomarkers.

## Results

Our analyses of 21 inflammation biomarkers, generally measured by ELISA, included 12 cohorts from the TOPMed Program (Table S1, Table S2); phenotype availability and sample characteristics varied by trait (Table S3). For example, for CRP as an example trait, average age was 57.6, and the participants analyzed were 64.8% female, 1.5% Asian, 22.9% non-Hispanic Black, 22.7% Hispanic/Latino, 52.9% non-Hispanic White. By contrast, due to different cohorts contributing, analysis of interleukin-6 included 70.9% non-Hispanic White participants. In brief, we performed single variant analysis to identify trait-associated loci, followed by stepwise conditional analysis to identify the total number of statistically distinct signals. We also conditioned on previously associated variants to identify distinct signals not identified in prior papers. We performed genetic region and gene centric rare variant set-based analyses for each trait and likewise conditioned on previously identified signals.

Of the 21 traits tested, CRP, E-selectin, intercellular adhesion molecule 1 (ICAM-1), interleukin 18 (IL-18), IL-6, lipoprotein-associated phospholipase A2 (Lp-PLA2) activity and mass, monocyte chemoattractant protein-1 (MCP-1), matrix metalloproteinase-9 (MMP-9), P-selectin, and tumor necrosis factor  $\alpha$  receptor 2 (TNFR2) had at least 1 genome-wide significant locus in single variant analyses. Across these 11 traits there were a total of 30 genome-wide significant loci ( $P < 1.0 \times 10^{-9}$  [21]) (Table S4, Figs S2–S32), for which stepwise conditional analysis revealed a total of 67 distinct signals (Table S5). After conditioning on previously identified associations (Table S6), 22 conditionally distinct variants across 8 loci remained locus-wide significant for 6 traits (Table S7 and Table 2, Fig. 1, significance thresholds listed in Table S7), and 1 trait (MMP-9) had a locus not reported in the GWAS catalog (Table 2, Fig. 1). We focus on these 7 traits with findings distinct from those already reported in the GWAS catalog below.

In aggregate rare variant analyses, we detected 51 significant gene-centric sets associated with 6 traits (Table S10A) and 214

**Table 1.** Overview of 21 inflammation-related biomarkers. Mean and (standard deviation) are reported for quantitative values, total N and % for dichotomous variables.

| Trait (unit)                                                               | N      | Age         | Female         | Measurement  | Cohort <sup>a</sup>                                                       |
|----------------------------------------------------------------------------|--------|-------------|----------------|--------------|---------------------------------------------------------------------------|
| Cluster of Differentiation 40 (CD40) (ng/ml)                               | 2381   | 60.4 (8.9)  | 1330 (55.9%)   | 0.59 (1.10)  | MESA, FHS                                                                 |
| C-Reactive Protein (CRP) (mg/l)                                            | 38 465 | 57.6 (15.6) | 24 912 (64.8%) | 0.70 (1.14)  | JHS, CARDIA, COPDGene, WHI, SOL, CHS, CFS, ARIC, OOA, GeneSTAR, MESA, FHS |
| E-selectin (ng/ml)                                                         | 5489   | 64.5 (10.1) | 4128 (75.2%)   | 3.46 (0.69)  | JHS, MESA, WHI, COPDGene                                                  |
| Intercellular Adhesion Molecule 1 (ICAM-1) (ng/ml)                         | 9268   | 50.5 (17.5) | 5107 (55.1%)   | 5.42 (0.42)  | CARDIA, CHS, CFS, MESA, FHS                                               |
| Interleukin-10 (IL-10) (pg/ml)                                             | 5533   | 60.1 (13.4) | 3130 (56.6%)   | 1.28 (1.24)  | WHI, COPDGene, CFS, GeneSTAR, MESA                                        |
| Interleukin-18 (IL-18) (pg/ml)                                             | 2151   | 61.9 (8.6)  | 1153 (53.6%)   | 5.44 (0.40)  | FHS, COPDGene                                                             |
| Interleukin-1 $\beta$ (IL-1 $\beta$ ) (pg/ml)                              | 1638   | 55.4 (18.7) | 1264 (77.2%)   | -0.02 (1.46) | GeneSTAR, WHI, CFS                                                        |
| Interleukin-6 (IL-6) (pg/ml)                                               | 18 844 | 60.6 (14.1) | 11 803 (62.6%) | 0.63 (0.80)  | CARDIA, COPDGene, WHI, CHS, CFS, GeneSTAR, MESA, FHS                      |
| Interleukin-8 (IL-8) (pg/ml)                                               | 2826   | 65.9 (8.6)  | 1584 (56.1%)   | 2.44 (0.61)  | WHI, COPDGene                                                             |
| 8-iso Prostaglandin F2 $\alpha$ (isoprostane-8-epi-pgf2 $\alpha$ ) (pg/ml) | 2778   | 50.4 (13.8) | 1523 (54.8%)   | 6.73 (1.00)  | FHS                                                                       |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Activity (nmol/min/ml)   | 10 210 | 62.0 (13.2) | 5611 (55.0%)   | 4.52 (0.70)  | FHS, MESA, CHS                                                            |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Mass (ng/ml)             | 10 132 | 61.9 (13.3) | 5566 (54.9%)   | 5.47 (0.38)  | FHS, MESA, CHS                                                            |
| Monocyte Chemoattractant Protein-1 (MCP-1) (pg/ml)                         | 3125   | 51.8 (13.9) | 1700 (54.4%)   | 5.77 (0.32)  | FHS                                                                       |
| Matrix Metalloproteinase-1 (MMP-1) (pg/ml)                                 | 737    | 63.5 (8.8)  | 348 (47.2%)    | 6.23 (0.91)  | COPDGene                                                                  |
| Matrix metalloproteinase-9 (MMP-9) (ng/ml)                                 | 5191   | 56.4 (13.4) | 3195 (61.5%)   | 8.55 (2.68)  | WHI, MESA, FHS, COPDGene                                                  |
| Myeloperoxidase (MPO) (ng/ml)                                              | 1582   | 61.3 (8.6)  | 878 (55.5%)    | 3.68 (0.53)  | FHS                                                                       |
| Osteoprotegerin (OPG) (pmol/l)                                             | 3131   | 54.9 (16.3) | 1699 (54.3%)   | 1.49 (0.34)  | FHS                                                                       |
| P-selectin (ng/ml)                                                         | 5032   | 55.7 (14.3) | 2902 (57.7%)   | 3.60 (0.38)  | JHS, FHS                                                                  |
| Tumor Necrosis Factor- $\alpha$ Receptor 1 (TNFR1) (pg/ml)                 | 3400   | 63.0 (9.5)  | 2157 (63.4%)   | 7.22 (0.30)  | WHI, MESA, COPDGene                                                       |
| Tumor Necrosis Factor- $\alpha$ (TNF- $\alpha$ ) (pg/ml)                   | 7591   | 62.5 (12.5) | 4738 (62.4%)   | 1.37 (1.09)  | COPDGene, WHI, CFS, GeneSTAR, MESA, FHS                                   |
| Tumor Necrosis Factor Receptor 2 (TNFR2) (pg/ml)                           | 3071   | 51.2 (13.9) | 1672 (54.4%)   | 7.65 (0.27)  | FHS                                                                       |

<sup>a</sup>Cohort studies involved in analysis. ARIC: Atherosclerosis Risk in Communities Study. CARDIA: Coronary Artery Risk Development in Young Adults Study. CFS: Cleveland Family Study. CHS: Cardiovascular Health Study. COPDGene: Genetic epidemiology of COPD Study. FHS: Framingham Heart Study. GeneSTAR: Genetic Study of Atherosclerosis Risk Study. JHS: Jackson Heart Study. MESA: Multi-Ethnic Study of Atherosclerosis Study. OOA: Old Order Amish Study. SOL: The Hispanic Community Health Study/Study of Latinos. WHI: Women's Health Initiative Study.

significant 2-kb sliding windows associated with 7 traits (Table S11A). We observed 19 significant rare variant aggregate test associations (some in overlapping or adjoining regions) after conditioning on known variants from the GWAS catalog and single-variant signals in the present analysis (Table 3), with traits P-selectin, ICAM-1, CRP, Lp-PLA2 activity and mass, all of which also had conditionally distinct single variant results (Tables S10C and S11C). If possible, we attempted to replicate distinct single variant findings using semiquantitative inflammation biomarker measures (i.e. measures which do not give an exact protein concentration in their results, for example in mg/dl) from the SomaScan or Olink platforms in independent samples (Table 4), as well as, for CRP, with rare variant testing in UK Biobank. While not all rare variant signals could be replicated due to lack of data availability, we note that all distinct rare

variant aggregation signals were in known regions, increasing the plausibility of their association with inflammation traits.

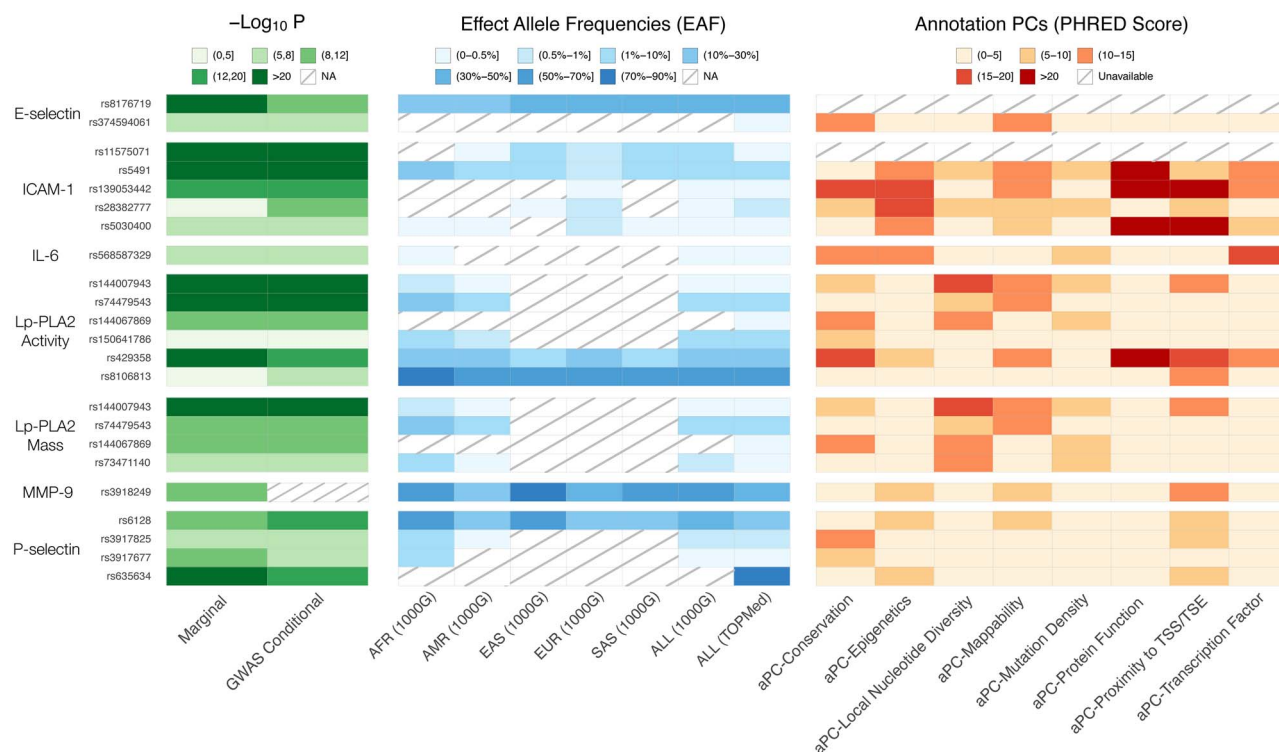
## C-reactive protein

We identified genetic variants associated with CRP consistent with and expanding upon our previous analysis of CRP in 23 279 TOPMed participants [17]. All 8 distinct single variant signals at the CRP locus previously known in TOPMed [17] (in partially overlapping samples) were also found here. We identified 1 additional distinct signal, rare variant rs370370301 (TOPMed Effect Allele Frequency (EAF): 0.2%, 1000G EUR EAF: 0.1%, 1000G SAS EAF: 0.1%, and not available in all other populations in 1000G), which was still significant after all conditional analyses. This non-coding variant did not reach genome-wide significance in the previous TOPMed analysis ( $P = 5.0 \times 10^{-6}$ ) but was significant

**Table 2.** Lead single variant signals at distinct locus (MMP9) and signals distinct from prior GWAS identified variants at known loci.

| Trait                                                      | Locus Name | rsID        | CHR | POS (hg38)  | Effect | Other | Unconditional |       | Conditional |       | EAF  | Distal/Local |
|------------------------------------------------------------|------------|-------------|-----|-------------|--------|-------|---------------|-------|-------------|-------|------|--------------|
|                                                            |            |             |     |             |        |       | P-value       | Beta  | P-value     | Beta  |      |              |
| E-selectin                                                 | ABO        | rs8176719   | 9   | 133 257 521 | TC     | T     | 4.3E-141      | -0.24 | 7.7E-12     | -0.07 | 0.01 | 33.9% Distal |
|                                                            | ABO        | rs374594061 | 9   | 132 553 865 | A      | G     | 2.6E-06       | 0.71  | 1.8E-07     | 0.71  | 0.14 | 0.1% Distal  |
| Intercellular Adhesion Molecule 1 (ICAM-1)                 | ICAM1      | rs11575071  | 19  | 10272 168   | G      | C     | 2.3E-45       | -0.49 | 3.4E-52     | -0.54 | 0.04 | 0.3% Local   |
|                                                            | ICAM1      | rs5491      | 19  | 10274 864   | T      | A     | 2.5E-36       | -0.14 | 1.7E-17     | -0.10 | 0.01 | 4.5% Local   |
|                                                            | ICAM1      | rs139053442 | 19  | 10283 720   | C      | G     | 9.1E-17       | -0.54 | 1.9E-17     | -0.53 | 0.06 | 0.1% Local   |
|                                                            | ICAM1      | rs28382777  | 19  | 10400 963   | G      | T     | 6.6E-04       | -0.08 | 1.4E-09     | -0.13 | 0.02 | 0.6% Local   |
|                                                            | ICAM1      | rs5030400   | 19  | 10285 120   | T      | C     | 4.7E-07       | 0.15  | 2.0E-07     | 0.14  | 0.03 | 0.4% Local   |
| Interleukin-6 (IL-6)                                       | IL6        | rs568587329 | 1   | 154 730 517 | T      | C     | 5.4E-06       | -0.93 | 1.2E-06     | -0.99 | 0.20 | 0.0% Distal  |
|                                                            | PLA2G7     | rs144007943 | 6   | 46662 909   | G      | T     | 8.0E-34       | -0.46 | 5.8E-36     | -0.48 | 0.04 | 0.2% Local   |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Activity | PLA2G7     | rs74479543  | 6   | 46784 401   | A      | G     | 2.3E-22       | -0.13 | 1.8E-24     | -0.13 | 0.01 | 2.4% Local   |
|                                                            | PLA2G7     | rs144067869 | 6   | 46709 433   | G      | A     | 1.7E-10       | -0.35 | 8.7E-11     | -0.36 | 0.06 | 0.1% Local   |
|                                                            | PLA2G7     | rs150641786 | 6   | 46774 942   | A      | C     | 3.6E-03       | 0.05  | 1.1E-06     | 0.07  | 0.02 | 1.4% Local   |
|                                                            | APOE       | rs429358    | 19  | 44908 684   | C      | T     | 4.2E-37       | 0.06  | 1.0E-13     | 0.06  | 0.01 | 13.7% Distal |
|                                                            | APOE       | rs8106813   | 19  | 44928 401   | G      | A     | 1.2E-02       | 0.01  | 1.2E-07     | 0.02  | 0.00 | 54.6% Distal |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Mass     | PLA2G7     | rs144007943 | 6   | 46662 909   | G      | T     | 9.0E-25       | -0.39 | 5.1E-25     | -0.39 | 0.04 | 0.2% Local   |
|                                                            | PLA2G7     | rs74479543  | 6   | 46784 401   | A      | G     | 2.6E-10       | -0.08 | 9.0E-13     | -0.09 | 0.01 | 2.4% Local   |
|                                                            | PLA2G7     | rs144067869 | 6   | 46709 433   | G      | A     | 5.3E-11       | -0.39 | 2.5E-10     | -0.37 | 0.06 | 0.1% Local   |
|                                                            | PLA2G7     | rs73471140  | 6   | 46641 939   | C      | T     | 3.1E-07       | -0.17 | 8.2E-09     | -0.19 | 0.03 | 0.3% Local   |
|                                                            | SELP       | rs6128      | 1   | 169 593 666 | T      | C     | 2.3E-10       | -0.05 | 5.8E-17     | -0.07 | 0.01 | 28.9% Local  |
| P-selectin                                                 | SELP       | rs3917825   | 1   | 169 595 320 | G      | A     | 3.9E-07       | -0.19 | 2.9E-10     | -0.23 | 0.04 | 0.9% Local   |
|                                                            | SELP       | rs3917677   | 1   | 169 622 970 | C      | A     | 4.7E-09       | -0.31 | 5.1E-08     | -0.28 | 0.05 | 0.4% Local   |
|                                                            | ABO        | rs635634    | 9   | 133 279 427 | C      | T     | 1.0E-55       | 0.16  | 2.0E-15     | 0.19  | 0.02 | 84.7% Distal |
| Matrix metalloproteinase-9 (MMP-9)                         | MMP9       | rs3918249   | 20  | 46009 497   | C      | T     | 1.6E-11       | 0.07  | 0.01        | 0.01  | 0.01 | 35.5% Local  |

Trait: trait name. Locus Name: significant loci identified by single variant analysis. rsID: rsID of lead signal. Unconditional: summary statistics of marginal analysis including P-value, beta coefficient, and standard error (SE). Conditional: summary statistics of conditional analysis including P-value, beta coefficient, and standard error (SE). Variants conditioned on are listed in Table S6. EAF: Effect Allele Frequency (TOPMed). frequency of effect allele of the lead signal within TOPMed. Distal/Local: Distal means that the lead signal is more than 1 Mb from the locus while local is in the 1 Mb region on either side of the center of the locus.



**Figure 1.** Single variant findings conditionally distinct from GWAS catalog variants. We report *P*-values for association for marginal and conditional results, reference population effect allele frequencies (EAF) by continental ancestry group as defined by 1000 genomes project (1000G) [19]—African (AFR), admixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS), as well as all participants in 1000G (ALL)—and the overall effect allele frequency for all participants included in our TOPMed analyses, and annotation principal components (aPCs) from FAVOR [20]. NA means the variant is not reported in the reference panel. We note that this information is available for all variants in close linkage disequilibrium with these lead variants in Table S8.

( $P = 1.4 \times 10^{-11}$ ) in the present analysis. This increase in significance is likely due to the larger number of participants, which is now 38465. This variant was not previously identified in an analysis of CRP in UK Biobank (UKB) [22], likely due to the rare frequency. Rare variant analysis yielded 1 significantly associated gene-centric set of 54 missense rare variants ( $P = 3.6 \times 10^{-22}$ ) on CRP locus driven in part by rs77832441 ( $P = 7.8 \times 10^{-16}$  for analysis of individual variant in TOPMed) (Table S10B). We also tested a similar gene-centric missense rare variant set for association in UKB ( $P = 6.4 \times 10^{-34}$  based on 116 variants, details in Table S15). rs77832441 (MAC = 153, EAF = 0.2%) was previously identified in Schick et al. [23]. We note that rs77832441 was pruned from the conditional analysis list based on linkage disequilibrium (LD) (see Materials and Methods) but a variant in close LD, rs553202904 ( $r^2 = 0.97$ ), was included (Tables S6, S10C and S11C), and the significance of the gene centric test was attenuated but still significant (missense set,  $P = 1.3 \times 10^{-8}$ , Tables 3 and S10A) when this signal was adjusted for, suggesting additional subthreshold CRP missense variants in particular remain to be identified as individually significant in larger analyses.

In addition to signals at the CRP locus, we also identified multiple loci in the single variant association analyses not previously detected in prior TOPMed analysis, including 3 with multiple distinct signals (LEPR, SALL1, APOE) (Table S5). Each of these signals were attenuated below the genome-wide significance threshold after adjusting for known associations from the GWAS catalog and other prior publications [17, 22].

## E-selectin

There are 9 distinct signals at the *SELL/SELE*, *FUT6*, and *ABO* loci associated with E-selectin, and 2 distinct signals remaining at

the *ABO* locus after conditioning on previously identified signals, including single variant signals from previous TOPMed analysis. This pair of signals, rs8176719 and rs374594061, were the second and third distinct signals in our marginal analysis. Variant rs8176719 is a frameshift insertion exonic variant common across all populations that tags blood group O [28]. We do note that in our prior work from TOPMed [18], while this was not captured as an independent genome-wide signal, associations of differential E-selectin levels across blood groups (with O treated as reference) were also observed. This variant's association with E-selectin further illustrates the extensive pleiotropy of the *ABO* locus, which has been previously associated with many diseases and traits. E-selectin associated distinct variant rs374594061 is rare across all populations (TOPMed EAF: 0.9%, and not available in 1000G) and, likely as a consequence, has no previously reported associations in the GWAS catalog and was also not tested in available replication cohorts.

## Intercellular adhesion molecule 1

For ICAM-1, we identified 9 distinct single variant signals at the *ICAM1* and *ABO* loci; 5 distinct signals at *ICAM1* remained (Table S7) after conditioning on known associations (Table S6). The GWAS conditionally significant association at rs5491, the fourth distinct signal in unconditional results at the *ICAM1* locus, is an exonic variant (TOPMed EAF: 4.5%, 1000G AFR EAF: 25.0%, 1000G AMR EAF: 1.7%, 1000G EAS EAF: 5.3%, 1000G EUR EAF: 0.7%, 1000G SAS EAF: 2.0%) that is low frequency in most populations but common among African ancestry populations. There are 4 other conditionally distinct noncoding variants—rs11575071, rs139053442, rs28382777, rs5030400—at the *ICAM1* locus (Table 2); most have low or rare frequency across all populations. As displayed in

**Table 3.** Significant gene-centric and genetic region rare variant set analysis (after conditioning on known variants from the GWAS catalog and single-variant signals in the present analysis).

| Gene Centric Analysis                                      |     |           |           |            |      |                 |             |            |
|------------------------------------------------------------|-----|-----------|-----------|------------|------|-----------------|-------------|------------|
| Trait                                                      | CHR | Symbol    | category  | # variants | cMAC | STAAR-O p-value |             |            |
|                                                            |     |           |           |            |      | unconditional   | conditional | cond2round |
| C-Reactive Protein (CRP)                                   | 1   | CRP       | missense  | 54         | 336  | 3.6E-22         | 1.3E-08     |            |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Activity | 6   | PLA2G7    | pLOF      | 5          | 14   | 1.3E-13         | 1.1E-13     | 1.6E-06    |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Mass     | 6   | PLA2G7    | missense  | 56         | 323  | 6.4E-78         | 8.5E-28     | 3.9E-23    |
| Intercellular Adhesion Molecule 1 (ICAM-1)                 | 6   | PLA2G7    | pLOF      | 5          | 13   | 1.7E-10         | 1.1E-10     |            |
|                                                            | 6   | PLA2G7    | missense  | 55         | 326  | 1.8E-75         | 1.1E-18     |            |
|                                                            | 19  | ICAM1     | missense  | 69         | 451  | 7.8E-15         | 5.0E-08     | 3.9E-05    |
|                                                            | 19  | ZNF653    | enhancer  | 126        | 577  | 2.3E-11         | 2.3E-11     | 8.7E-01    |
| Region-Based Analysis                                      |     |           |           |            |      |                 |             |            |
| Trait                                                      | CHR | pos_min   | pos_max   | # variants | cMAC | STAAR-O p-value |             |            |
|                                                            |     |           |           |            |      | unconditional   | conditional | cond2round |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Activity | 6   | 46707812  | 46709811  | 103        | 526  | 2.2E-46         | 1.7E-20     | 9.1E-10    |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Mass     | 6   | 46708812  | 46710811  | 95         | 389  | 6.9E-74         | 1.3E-21     | 4.2E-11    |
| Intercellular Adhesion Molecule 1 (ICAM-1)                 | 6   | 46707812  | 46709811  | 103        | 532  | 5.3E-44         | 9.3E-14     |            |
|                                                            | 6   | 46708812  | 46710811  | 94         | 394  | 1.3E-64         | 2.0E-14     |            |
|                                                            | 19  | 11282547  | 11284546  | 68         | 591  | 6.1E-12         | 4.5E-10     | 6.8E-01    |
|                                                            | 19  | 11283547  | 11285546  | 91         | 892  | 7.7E-12         | 5.6E-10     | 9.7E-01    |
|                                                            | 19  | 11284547  | 11286546  | 96         | 1337 | 1.8E-09         | 7.0E-09     | 6.5E-01    |
|                                                            | 19  | 11285547  | 11287546  | 85         | 871  | 1.1E-09         | 1.8E-08     | 7.4E-01    |
|                                                            | 19  | 11503547  | 11505546  | 119        | 729  | 2.8E-11         | 2.8E-11     | 8.3E-01    |
|                                                            | 19  | 11504547  | 11506546  | 154        | 818  | 3.6E-11         | 3.6E-11     | 6.5E-01    |
| P-selectin                                                 | 1   | 169615464 | 169617463 | 65         | 433  | 7.5E-12         | 3.6E-12     | 4.8E-07    |
|                                                            | 1   | 169616464 | 169618463 | 66         | 363  | 4.6E-12         | 5.1E-12     | 5.6E-07    |

Trait: trait name. CHR: chromosome where the gene is located. Symbol: gene symbols. Category: category of gene-based test; pLOF means putative loss of function. pos\_min: starting position of the region tested, hg38. pos\_max: ending position of the region tested, hg38. # variants: number of variants tested in the aggregate test. cMAC: cumulative minor allele count. STAAR-O P-value: P-values of aggregate tests in 3 cases. unconditional: P-value of unconditional analysis. conditional: P-value of conditional analysis conditioning on (1) variants reported on previous literature (Table S6), (2) lead signals from our conditional single variant association analysis. cond2round: P-value of second round of conditional analysis; conditional list for second round conditional analysis consists of 2 parts: (1) conditional list for the first round conditional analysis; (2) additionally, variants included in the aggregate test which had nominally significant individual variant P-values ( $P < 1.0 \times 10^{-6}$ ) (Tables S10A and S11A). Note that not all gene sets have such remaining significant variants, so we do not further perform the second round conditional analysis in these cases, and leave the column blank.

Fig. S17, there is some long-range LD for variants identified in the ICAM1 locus, notably for rs5491 (displayed in turquoise) in Fig. S17.

We also identified multiple conditionally significant rare variant set-based associations with ICAM-1 including 2 gene-centric sets (Table S10A) and 6 2-kb sliding windows (Table S11A, individual variants included in tests included in Table S11B), and 2 of them overlap the ICAM1 locus. We identify a set of missense rare variants at ICAM1, whose most significant variant was the identified rs139053442 association but which remains significant after conditioning on rs139053442 and other single variant findings from TOPMed and other studies (Table S6).

Matrix Metalloproteinase-9

We identified the MMP9 encoding gene for association with MMP-9 levels in single variant analysis. This cis pQTL locus included 1 distinct signal at intronic variant rs3918249 that was common in all populations, and it has repressed regulatory function with high H3K27me3 score 48 according to FAVOR [20]. Our identified variant rs3918249 (Fig. 1, TOPMed EAF 35.5%) is highly linked ( $r^2 = 0.938$ ) with coding variant rs17576 (Table S8).

P-selectin

For P-selectin, we identified 5 distinct single variant signals at the SELP locus (Table S5), and 3 of them remain significant (Table S7) after conditioning on known associations (Table S6), and 1 distinct

single variant signal at the ABO locus that is significant conditional on known associations (Tables S5 and S7). At the SELP locus, 2 of 3 conditionally significant signals are intronic (rs3917677, rs3917825). rs3917825 is relatively conserved (top 9.1% genome-wide aPC-conservation score) [20]. Both of these variants have low frequency in AFR ancestry participants (1.7% for rs3917677 in 1000G, 2.8% for rs3917825 in 1000G) and are not observed in EUR ancestry participants (from reference panels). The remaining significant signal in the SELP locus is the synonymous variant rs6128, which is more common in AFR ancestry (53.3%) than in EUR ancestry (16.6%) participants from 1000G. Variant rs6128 is a platelet splice QTL that alters SELP exon 14 skipping and soluble versus transmembrane P-selectin protein production [33].

For aggregate tests of rare variants, lead signals were detected at 2 consecutive 2-kb sliding windows in the SELP locus located at chr1:169615464-169617463 and chr1:169616464-169618463 (Table S11A), which are driven in part by rs7529463. This coding variant is highly conserved (top 1.6% genome-wide aPC-conservation score), very rare (TOPMed AF 0.1%), and has high aPC protein function scores (top 0.2% genome wide) [20].

At the ABO locus, the distinct signal (rs635634, which tags blood group A) remained significant after conditioning on known variants (Table S6); however, the P-value is significantly attenuated (from  $P = 1.0 \times 10^{-55}$  to  $P = 2.0 \times 10^{-15}$ , Table S7) when adjusting for known GWAS catalog variants.

**Table 4.** Replication of newly identified signals in previous semiquantitative platform pQTL analysis.<sup>a</sup>

| Trait                                                      | Locus Name                                        | rsID        | CHR         | POS (hg38) | Allele   | TOPMed |         | Folkersen et al. (2020), PMID: 33067605 | Pietzner et al. (2021), PMID: 34648354 | Sun et al. (2018), PMID: 29875488 |                    | Ferklingstad et al. (2021), PMID: 34857953 |                    | Zhang, et al. (2022), PMID: 35501419 <sup>b</sup> |              |           |              |  |  |  |  |  |
|------------------------------------------------------------|---------------------------------------------------|-------------|-------------|------------|----------|--------|---------|-----------------------------------------|----------------------------------------|-----------------------------------|--------------------|--------------------------------------------|--------------------|---------------------------------------------------|--------------|-----------|--------------|--|--|--|--|--|
|                                                            |                                                   |             |             |            |          | Beta   | p-value |                                         |                                        | Beta                              | p-value            | Beta                                       | p-value            | Beta (AA)                                         | p-value (AA) | Beta (EA) | p-value (EA) |  |  |  |  |  |
| E-selectin                                                 | ABO                                               | rs8176719   | 9           | 133257521  | TC       | T      | −0.238  | 4.3E-141                                | −0.118 (+) 9.1E-20                     | −0.592 (+) 0.0E+00                |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | ABO                                               | rs374594061 | 9           | 132553865  | A        | G      | 0.710   | 2.6E-06                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | Intercellular Adhesion Molecule 1 (ICAM-1)        | ICAM1       | rs11575071  | 19         | 10272168 | G      | C       | −0.488                                  |                                        |                                   |                    |                                            |                    |                                                   |              | 2.3E-45   |              |  |  |  |  |  |
|                                                            |                                                   | ICAM1       | rs5491      | 19         | 10274864 | T      | A       | −0.141                                  |                                        |                                   |                    |                                            |                    |                                                   |              | 2.5E-36   |              |  |  |  |  |  |
|                                                            |                                                   | ICAM1       | rs139053442 | 19         | 10283720 | C      | G       | −0.542                                  |                                        |                                   |                    |                                            |                    |                                                   |              | 9.1E-17   |              |  |  |  |  |  |
| Interleukin-6 (IL-6)                                       | ICAM1                                             | rs28382777  | 19          | 10400963   | G        | T      | −0.076  | 6.6E-04                                 | 1.237 (+) 4.3E-14                      | −0.570 (+) 5.3E-13                | −0.452 (+) 8.6E-11 | 0.353 (−) 1.0E-20                          |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | ICAM1                                             | rs5030400   | 19          | 10285120   | T        | C      | 0.147   | 4.7E-07                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | IL6R                                              | rs568587329 | 1           | 154730517  | T        | C      | −0.929  | 5.4E-06                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | Lipoprotein-associated phospholipase A2 (Lp-PLA2) | PLA2G7      | rs144007943 | 6          | 46662909 | G      | T       | −0.463                                  |                                        |                                   |                    |                                            |                    |                                                   |              | 8.0E-34   |              |  |  |  |  |  |
|                                                            | PLA2G7                                            | rs74479543  | 6           | 46784401   | A        | G      | −0.127  | 2.3E-22                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
| Lipoprotein-associated phospholipase A2 (Lp-PLA2) Activity | PLA2G7                                            | rs144067869 | 6           | 46709433   | G        | A      | −0.354  | 1.7E-10                                 | −1.439 (+) 1.4E-28                     | 0.253 (+) 7.0E-112                | 0.003 (+) 7.0E-01  |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | PLA2G7                                            | rs150641786 | 6           | 46774942   | A        | C      | 0.045   | 3.6E-03                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | APOE                                              | rs8106813   | 19          | 44928401   | G        | A      | 0.009   | 4.2E-37                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | PLA2G7                                            | rs144007943 | 6           | 46662909   | G        | T      | −0.389  | 1.2E-02                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | PLA2G7                                            | rs74479543  | 6           | 46784401   | A        | G      | −0.083  | 9.0E-25                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
| Mass                                                       | PLA2G7                                            | rs144067869 | 6           | 46709433   | G        | A      | −0.388  | 2.6E-10                                 | −1.439 (+) 1.4E-28                     | −1.702 (+) 6.6E-41                | −0.374 (+) 3.8E-16 |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | PLA2G7                                            | rs73471140  | 6           | 46641939   | C        | T      | −0.171  | 5.3E-11                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | SELP                                              | rs6128      | 1           | 169593666  | T        | C      | −0.054  | 3.1E-07                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | SELP                                              | rs3917825   | 1           | 169595320  | G        | A      | −0.188  | 2.3E-10                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | SELP                                              | rs3917677   | 1           | 169622970  | C        | A      | −0.306  | 3.9E-07                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
| P-selectin                                                 | ABO                                               | rs635634    | 9           | 133279427  | C        | T      | 0.163   | 4.7E-09                                 | −0.059 (+) 5.4E-04                     | −0.095 (+) 2.3E-16                | −0.148 (+) 4.2E-06 | −0.549 (+) 1.6E-06                         | −1.070 (+) 2.2E-19 | −0.223 (+) 4.6E-22                                |              |           |              |  |  |  |  |  |
|                                                            | CRP                                               | rs370370301 | 1           | 159712228  | A        | G      | −0.625  | 1.0E-55                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            | MMP9                                              | rs3918249   | 20          | 46009497   | C        | T      | 0.070   | 1.4E-11                                 |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            |                                                   |             |             |            |          |        |         |                                         |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |
|                                                            |                                                   |             |             |            |          |        |         |                                         |                                        |                                   |                    |                                            |                    |                                                   |              |           |              |  |  |  |  |  |

<sup>a</sup>FAVOR annotation of each variant and variants in close LD in Table S8. <sup>b</sup>Note that some ARIC participants are also included in our analyses, so Zhang et al. (2022) [27] cannot be considered a truly independent replication cohort. The (−) and (+) notations after reported beta values indicate if results are directionally concordant with TOPMed.

## Interleukin 6

We identified the *IL6R* locus in the marginal single variant analysis (lead variant rs61812598,  $P = 1.1 \times 10^{-49}$ , Table S5). After conditioning on previous GWAS-identified variants (Table S6), the lead variant was rs568587329 (Table S7) ( $P = 1.2 \times 10^{-6}$ , when adjusted for known variants at the *IL6R* locus (GWAS catalog, [35]), which met our locus-wide significance threshold. This variant is rare in all populations (TOPMed EAF: 0.03%, 1000G AFR EAF: 0.4%, and not available in all other populations in 1000G) and has a high aPC-Transcription-Factor score 17.29 (top 1.87% genome wide) [20].

## Lipoprotein-associated phospholipase A2 activity and mass

For the Lp-PLA2 activity trait, we identified 11 distinct single variant signals at the *CELSR2*, *APOE*, *LDLR*, and *PLA2G7* loci (Table S5). After conditioning on previous GWAS identified variants (Table S6), 2 GWAS conditional distinct signals remain at the *APOE* locus (Table S7), and 4 GWAS conditional distinct signals remain at the *PLA2G7* locus (Table S7).

At the *APOE* locus, the GWAS conditional distinct signals rs429358 (representing the well-known *APOE-ε4* allele) and rs8106813 are the second and the third distinct signals of our stepwise analysis.

We observe 4 low-frequency distinct signals, rs144007943, rs74479543, rs144067869, and rs150641786, at the *PLA2G7* locus significant upon conditioning on prior GWAS identified signals. In addition to these single variant associations, we observe 2 gene centric and 2 2-kb sliding windows significantly associated at the *PLA2G7* locus. We observe a set of putative loss-of-function (pLOF) rare variants and missense rare variants. The pLOF set is partly driven by rs140020965, whereas the missense set is partly driven by rs200303358 (though the set is still quite significant after conditioning on this variant ( $P = 3.9 \times 10^{-23}$ ) (Table S10A). We also observe a 2-kb sliding window spanning chr6:46707812-46709811 and another 2-kb sliding window spanning chr6:46708812-46710811 both partially driven by rs140020965 and rs200303358 (Table S11A).

For Lp-PLA2 mass, we identified 6 distinct signals at the *PLA2G7* locus (Table S5). After conditioning on previous GWAS-identified variants (Table S6), 4 signals remained significant (Table S7)—rs144007943, rs74479543, rs144067869, and rs73471140—3 of which were identified in our analysis of Lp-PLA2 activity, unsurprisingly given the high correlation between the traits. The additional signal at rs73471140 is rare across all populations and in very low LD with all Lp-PLA2 activity lead variants ( $r^2 < 0.01$ ). We again observe associations with pLOF rare variants and missense rare variants at the *PLA2G7* locus (Table S10A), and the same 2 significant 2-kb sliding windows as Lp-PLA2 activity (spanning chr6:46707812-46709811 and chr6:46708812-46710811) are also significant.

## Discussion

We sought to evaluate the genetic determinants of 21 inflammation biomarkers using data from the TOPMed Program. Previous efforts in TOPMed with E-selectin [18] and CRP [17] demonstrated that inclusion of diverse cohorts yielded further insights into the genetic determinants of these biomarkers. Our work extends these findings by incorporating both larger samples for these previously analyzed traits and expanding the scope to include 19 additional traits and rare variant aggregate tests. We identified

significant associations with 6 traits in single variant analysis and 5 traits in aggregate rare variant analysis that remained significant after conditioning on known associations.

Our findings demonstrate the complementary value of performing both single and rare variant analyses when analyzing quantitative traits. Recent analyses of quantitative lipid traits from TOPMed also combined single and rare variant analyses, similarly finding both common signals and conditionally distinct aggregate rare variant signals, mostly at known genes, for both coding and noncoding variant sets [36], similar to our findings here. Several exome sequencing efforts for diverse traits and diseases, for example waist hip ratio [37] and schizophrenia [38], have similarly identified joint impacts from common noncoding variants and rare coding variants at the same loci (including at Mendelian genes), but similar findings in the noncoding space have been less widely reported. Previous analysis [17] of CRP in TOPMed identified variants in enhancer regions (including 1 whose impact on transcription and protein binding to the enhancer region was validated *in vitro*) that were more common in AFR versus EUR ancestry individuals, demonstrating the contributions of ancestry differentiated variants in noncoding regions to the genetic architecture of the trait. That analysis did not include aggregate tests of rare variants, and in the present analysis we observe that even after conditioning on known single variant associations additional signals are identified by performing aggregate analyses. This aggregate test replicated in UK Biobank. This is concordant with prior reports of other subthreshold CRP associated missense variants identified in the CARDIA study [26]. We identify a similar joint contribution of common, rare, and low frequency variants for multiple traits, including P-selectin and ICAM-1. We do note that in some cases our rare variant signals are consecutive or overlapping, suggesting that multiple rare signals within a broad region may contribute to gene regulation (Lp-PLA2 and ICAM-1). We note that it remains an outstanding challenge to completely disentangle whether a common or rare variant signal is driving biological processes, and continued large-scale analysis will likely provide further insight.

Our analysis yielded more distinct signals than previously detected for inflammation biomarkers, primarily at known loci. This finding points to the extensive allelic heterogeneity at, in particular, encoding gene loci, as reflected by the increased number of statistically distinct cis pQTL [24] and cis eQTL [31] distinct signals observed with increasing sample size. Studies of populations with different ancestry often observe different cis eQTL and pQTL signals due to ancestry differentiated allele frequencies for such variants [39, 40], including our own analyses of CRP within TOPMed [17]. Prior work suggested that such distinct signals can have different molecular mechanisms (even acting through distinct transcripts, as at the adiponectin encoding gene locus [41]), with variants in different distinct signals often impacting different regulatory regions (including distinct enhancer and promoter regions). We anticipate that expanded efforts to understand such “secondary” distinct signals at known GWAS identified loci for quantitative traits in expanded sample sizes will identify many additional loci with significant allelic heterogeneity and ancestry differentiated QTLs. Such analyses would be completed ideally with individual level data to avoid issues with approximate conditional analysis with poor matching between the LD reference panel and the GWAS or WGS analysis population. Both individual level sequence data and improved imputation reference panels [42–44] may help increase discovery in the low frequency/rare variant space. We note that, where possible, we have attempted to replicate putative novel single variant

findings using semi-quantitative proteomics platforms. For variants tested in these external datasets, 16 out of 18 variants are both significant and have effects in the same direction between previous semi-quantitative pQTL analysis and our TOPMed analysis (Table 4). Such replication results are described in additional detail in the supplemental material. We also replicate many distinct signals from prior semiquantitative high throughput platform publications in our own immunoassay-based findings—for variants tested in both datasets, 217 out of 431 variants are both significant ( $P < 0.05$ ) and in the same direction between previous semi-quantitative pQTL analyses and our TOPMed analysis (Table S13). Using CRP as an example trait, this includes 10 out of 11 available CRP lead signals from Ferkingstad *et al.* (2021) [26], 5 out of 5 CRP lead signals from Pietzner *et al.* (2021) [27], 2 out of 2 available CRP lead signals from Sun *et al.* (2018) [28], 4 out of 5 CRP lead signals from African American Atherosclerosis Risk in Communities (ARIC) participants, and 5 out of 5 CRP lead signals from European American ARIC participants from Zhang *et al.* (2022) [29]. Note that some ARIC participants are also included in our analyses, so this is not an independent replication sample for Zhang *et al.* (2022) [29] findings. Similar look-ups were performed for all other overlapping traits and are noted in Table S13.

One locus identified by our analysis, the *cis* region around MMP-9 for its encoded protein, had not been reported in prior GWAS for quantitative immunoassays. This variant, rs3918249, was also identified by Ferkingstad *et al.* (2021) [26], Pietzner *et al.* (2021) [27], and Sun *et al.* (2018) [28], but to our knowledge this is the first report using a quantitative immunoassay. Our identified variant rs3918249 is highly linked with a coding variant rs17576. It is possible such a coding variant signal may tag an antibody binding effect without true impact on protein abundance. However, we note that rs3918249 is also highly linked with rs6017721 ( $r^2 = 0.86$ ) and rs4810482 ( $r^2 = 0.92$ ), both of which are significant conditionally distinct lead variants in GTEx V8 *cis*-eQTL results for MMP-9 (Table S9). The finding suggests that this variant influences transcript and likely protein abundance, not just antibody binding to the MMP-9 target protein. The MMP-9 coding variant rs3918249 we identified is in moderate LD ( $r^2 = 0.664$ ) with the intronic variant rs3918253. rs3918253 is associated with liver enzyme levels; this close LD suggests MMP-9 abundance could mediate this liver-related signal rs3918253 [34]. In contrast to the supporting evidence for this signal at MMP-9, prior work has found assay-binding artifacts for coding variants in ICAM1 [31]; similarly, the coding variant we identified at ICAM1 (rs5491) and its LD proxies were not ICAM1 eQTLs in eQTLGen phase I [32] and Genotype-Tissue Expression (GTEx) V8 [33] look-ups (as described in Materials and Methods) and we suspect it may be an assay interference effect.

Our analysis further highlights the value of including study populations inclusive of multiple ancestry groups. Using a larger sample size, we confirmed findings from previous TOPMed analyses driven by variants common only in AFR reference populations including rs3917422 and rs17855739 for E-selectin [18], as well as rs11265259 and rs181704186 for CRP [17]. Given the diversity of our sample, we were able to additionally identify associations with Lp-PLA2 traits, P-selectin, and ICAM1 that were exclusively or disproportionately observed in AFR reference populations (Fig. 1). Many previous large-scale analyses have been conducted primarily in European ancestry individuals.

Coding *cis* pQTLs present particular challenges for biomarker traits. Such QTLs often have large effect sizes, but it is unclear whether these effects represent a true impact on protein abundance versus interference with antibody/aptamer binding. Such

issues have also been identified in previous work from TOPMed, notably for the E-selectin signal rs3917422 identified by Polfus *et al.* (2019) [18], as well as in prior genetic analyses for other antibody measured biomarker traits [29, 45, 46]. As a supplemental analysis, we assessed coincidence of our identified coding pQTL signals with distinct eQTL signals in GTEx V8 [31] and eQTLGen phase I [30], and found that our MMP-9 coding variant signal, but not the signal at ICAM-1, coincided with an eQTL. When such coding pQTL variants also influence transcription, it is less likely they are an aptamer or antibody effect. This should be carefully evaluated in future pQTL efforts, using both quantitative and semiquantitative platforms.

Our analysis provides significant novelty to the literature in multiple ways. First, as noted in prior publications for high throughput semiquantitative proteomic QTL findings [47–49], it is important that such results be corroborated with quantitative orthogonal assays, as done here. Our results also provide a model for integrated single common and rare variant analysis for quantitative traits, with a vital role for conditional analysis on known variants, including common ones, in interpretation of identified rare variant sets. Of our original 51 significant gene-centric sets and 214 significant 2-kb sliding windows, only 19 significant rare variant aggregate test associations (some in overlapping or adjoining regions) were still significant after conditioning on nearby common variants. This highlights how essential adjustment for common, mostly noncoding GWAS identified-signals is, even in gene-centric or missense variant only tests. Notably, our work demonstrates the significant allelic heterogeneity, including continued discovery of additional signals at known loci, for quantitative traits, with likely distinct mechanisms as discussed in prior work for similar circulating biomarkers [41]. This includes signals that are rare in EUR populations, highlighting the importance of diversity in genomic analyses. Finally, our work provides a still fairly unusual replication of an aggregate test rare variant signal, utilizing UK Biobank data for CRP and strongly replicating a similarly filtered missense variant set to the one identified here in TOPMed.

There are multiple limitations of our present analysis. While the TOPMed program provides a rich sequencing data source, there are a relatively limited set of cohorts within TOPMed that have measured inflammation biomarkers in their participants. Similarly, few other large scale studies have incorporated inflammation biomarker measurement, and most of those have primarily limited their measurements to CRP [22]. This limits our ability to perform a well-powered analysis among some traits in TOPMed, and to replicate all of our findings in external datasets. However, along with the single variant replication discussed above, we also replicate our significant CRP-associated rare variant set identified in TOPMed using similar variant filters in UK Biobank sequencing data (Table S15), with both single variant and aggregate test replication results supporting the validity of our results. Correlation both between ELISAs themselves and between ELISA and aptamer assays (as well as between Olink and SomaScan) varies, and will impact expected replication rates [48, 50, 51]. However, such information is unfortunately not available for the vast majority of the specific immunoassays used here. We also note that many of our biomarkers are still mostly measured in non-Hispanic White participants; future efforts should focus on further increasing the inclusion of additional populations.

Through our analysis of 21 inflammation biomarkers, we identified additional signals and highlighted features of such large-scale analyses. Across this set of traits, consistently observed features included a combination of contributing common and

rare variant signal, extensive allelic heterogeneity, and ancestry specificity of some identified variants. Such features have been observed in other efforts, such as the analysis of lipids and blood cell traits in the TOPMed program [52]. We anticipate that with continually increasing sample sizes (and thereby statistical power) that these key aspects of our study would be observed in similar sequencing-based analyses of complex traits.

## Materials and methods

### Whole genome sequencing

We analyzed variants with whole genome sequencing from blood in samples from the NHLBI TOPMed program. All participants had deep coverage sequencing, with harmonization, variant discovery, and genotype calling previously described [52]. We specifically leveraged data from Freeze 8, which was aligned to GRCh38 reads [53]. All positions in this manuscript are reported based on GRCh38. Samples were processed by the TOPMed Data Coordinating Center, resulting in 1.02B variants for 138 K samples. For all Freeze 8 samples, population principal components of genetic ancestry were calculated using PC-AiR [54], genetic relatedness was calculated using PC-Relate [55], and race/ethnicity was reported by each study (mostly from participant self-report). No individuals were removed based on genetic ancestry cluster from these race/ethnicity groupings. Full single variant and aggregate test summary statistics will be provided at time of publication to the TOPMed genomic summary result dbGaP accession (phs001974).

### Phenotype harmonization and study sample

Phenotype harmonization for 21 inflammation biomarkers was primarily performed by the TOPMed Data Coordinating Center [56] as previously described. COPDGene, GeneSTAR, and WHI were harmonized directly from study-provided data. Methods of inflammation biomarker measurement are listed in Table S1. We note that not all cohorts used the same platform, and samples run on multiple platforms are not available for assay recalibration. This is unfortunately a common limitation for cross-cohort analyses of inflammation biomarker traits. Study participants were included based on informed consent restrictions (excluding some individuals with consent for only disease specific analyses), duplicates were removed to retain observations with the highest frequency assay type where applicable, trait measurements exceeding 3 standard deviations from the mean were removed, and individuals with missing data were excluded. CRP was natural log-transformed to address non-normality in distribution. All traits were analyzed after rank-based inverse normal transformation, performed by study-race/ethnicity strata, with variance rescaled within each strata. The present analysis of inflammation biomarkers included sample sizes ranging from 737–38 465 individuals from 12 cohorts in Freeze 8 of the NHLBI TOPMed program. Across all traits, the sample is primarily non-Hispanic White, though efforts were made to include a multi-ethnic population wherever possible. The sample is described in Tables S1 and S3.

### Single variant analysis

We performed single variant analyses across ancestry groups as was done in several previous studies in TOPMed [17, 57–60]. We tested PASS variants (based on support vector machine variant classifier, as previously described in TOPMed sequencing methods [52] with a minor allele count (MAC) of at least 10 in

our pooled sample, resulting in a test of between 11 793 614–57 072 499 variants for each biomarker trait. We used linear mixed effects models [61] as implemented in GENetic Estimation and Inference in Structured samples (GENESIS 2.19.1 [62]) on the BioData Catalyst Seven Bridges platform [63], adjusting for age, sex, variables combining study and race and ethnicity, an empirical kinship matrix for relatedness and population structure, 11 ancestry principal components [54, 55] and permitting heterogeneous variance across the strata of the combination of study and race and ethnicity. Differences in ancestry were accounted for by our principal components and kinship matrix adjustment, and we also adjusted for race/ethnicity as a self or study reported variable, given previously reported impacts of these social constructs on levels of inflammatory biomarkers [64, 65]. All efforts to examine whether identified genetic variants differed in frequency by genetic ancestry or similarity cluster used public resources such as 1000 Genomes, and did not use these study- or self-reported race/ethnicity labels within TOPMed results. Loci were defined as statistically significant according to a genome-wide threshold given as  $1 \times 10^{-9}$  [21].

We next performed stepwise conditional analysis at significant loci to identify the total number of conditionally distinct signals within a  $\pm 1$  Mb ( $\pm 3$  Mb for ICAM1 chr19) window. Conditional analysis was performed by running the association analysis conditioning on the lead variant defined by *P*-value, and repeating this process until no variants were significant at the locus. Significance was defined at  $\alpha=0.05$  using a Bonferroni correction for the number of variants tested within the locus, for example  $0.05/39488 = 1.3 \times 10^{-6}$  for CRP at the CRP locus. The threshold for conditional analysis of each trait conditioning on distinct signals and known variants are listed in Tables S5 and S7.

### Identification of distinct signals through conditional analysis

Many previous studies of inflammatory biomarkers have identified genome-wide significant signals for the inflammation biomarkers tested here (Table S6). To identify which single variant signals in our analysis were distinct from previously identified GWAS variants, we performed stepwise conditional analysis at significant loci for each trait, conditioning on the reported associations from the GWAS Catalog, Raffield et al. (2020) [17], Sinnott-Armstrong et al. (2021) [22], Ahluwalia et al. (2021) [35], Folkersen et al. (2017) [66], and Polfus et al. (2019) [18] as covariates in our null model to determine which associations in our TOPMed analysis are distinct from those previously identified. We mapped published associations within a  $\pm 1$  Mb window ( $\pm 3$  Mb window for ICAM1 chr19 due to very long range LD) of the TOPMed identified loci (i.e. GWAS conditional distinct signals at Table S7) to TOPMed Freeze 8 variants by positions and alleles. To avoid collinearity, we pruned the previous GWAS identified variant set with the linkage disequilibrium threshold  $r^2=0.9$  to obtain a list of previously identified distinct signals at each locus. All known variants were included as fixed effects in the null model. If any variants were still significant using a locus-wide threshold after this adjustment for known variants, we proceeded to perform stepwise conditional analysis again, to identify the total number of distinct signals after adjustment for known variants from prior GWAS.

### Rare variant analysis

We performed rare variant analysis for both gene-centric and genetic region aggregation units. We tested PASS variants with

MAC at least 1 and minor allele frequency (MAF) less than 1.0% in our pooled sample. We used linear mixed effects models with weighting by functional annotation as implemented in STAAR [67–69], adjusting for age, sex, race/ethnicity-study, and 11 population ancestry principal components and permitting heterogeneous variance across race-study strata and empirical kinship for relatedness and population structure. Gene-centric units were defined for all protein-coding genes using coding annotations based on GENCODE consequences as (a) putative loss of function (stop gain, stop loss, splicing), (b) missense, and (c) synonymous variants; non-coding variants were captured via masks characterized by (a) promoters if within  $\pm 3$  kb of a transcription start site overlaid with DHS signal, or (b) enhancers if identified by GeneHancer overlaid with DHS signal. Genetic region analysis used 2-kb sliding windows with a 1 kb skip length.

The STAAR-O P-value, incorporating 2 weighting schemes using the beta distribution based on MAF (with  $\alpha_1 = 1, \alpha_2 = 25$  to upweight rarer variants or with  $\alpha_1 = \alpha_2 = 1$  treat all equally) in addition to annotation-based weights using CADD, LINSIGHT, FATHMM-XF, aPC-Protein-Function, aPC-Conservation, aPC-Epigenetics-Active, aPC-Epigenetics-Repressed, aPC-Epigenetics-Transcription, aPC-Local-Nucleotide-Diversity, aPC-Mutation-Density, aPC-Transcription-Factor, aPC-Mappability, aPC-Proximity-To-TSS-TES, was considered. Sets were defined as statistically significant according to a Bonferroni-corrected significance threshold separately for gene-centric, correcting for all 5 masks, and genetic region analysis, correcting for all windows (Table S12). We performed conditional analysis to identify signals by obtaining trait-specific associations from the GWAS catalog and the single-variant analysis in a 1 Mb window from the start and end of the positions spanned by the set.

## Annotation

We used multiple resources to obtain functional annotations for inclusion in the rare variant analysis and to describe identified variants, including FAVOR, GTEx, and ANNOVAR. We obtain aPCs from FAVOR [20, 67], providing summarized functional categories by aggregating correlated individual functional annotations. These aPCs provide variant-level measures as a PHRED score yielding the interpretation that scores greater than 10 within a given functional category are in the top 10% for all TOPMed variants.

## Replication

Many genetic loci and distinct signals have been identified in previous pQTL studies using untargeted semiquantitative platforms (SomaScan and Olink) [24–27, 70]. For our conditionally distinct signals (GWAS conditional distinct signals at known loci, and rs3918249 for MMP-9), we pulled results from summary statistics of these prior published studies and compared their direction of effect and level of significance with our findings in TOPMed (Table 4). Conversely, we also attempted to replicate all previously reported distinct pQTL signals for overlapping traits in our summary statistics (Table S13).

For the CRP phenotype, we replicated our results using 188 912 samples with whole genome sequencing data from UKB [43, 71]. The null model was constructed using the same methods as the TOPMed analyses, and both single variant and variant set analyses were conducted using STAARPipeline app (<https://github.com/xihaoli/staarpipeline-rap>) [67, 68] on the UKB Research Analysis Platform (RAP).

## eQTL coincidence

We also checked the coincidence of eQTL signals from cis-eQTLGen phase I [30] and GTEx V8 [31] for the distinct signals we detected on the corresponding coding region of the inflammation biomarker traits. For cis-eQTLGen, we performed GCTA-COJO [72] on the summary-based Mendelian randomization [73] formatted cis-eQTLGen results to identify statistically distinct lead signals. For GTEx V8 [31], conditionally distinct signals were already reported (details in Table S9).

## Acknowledgements

Molecular data for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). Study-specific omics support information can be found in the supplement. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination, were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. TOPMed specific acknowledgments for studies are included in Table S2. Additional study specific acknowledgments are included under cohort descriptions in Table S1.

Support for this work was provided by the National Institutes of Health, National Heart, Lung, and Blood Institute, through the BioData Catalyst program (awards 1OT3HL142479-01, 1OT3HL142478-01, 1OT3HL142481-01, 1OT3HL142480-01, and 1OT3HL147154).

The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst® (BDC) ecosystem. L.M.R., S.G., and Z.L. were supported by NHLBI BioData Catalyst Fellowship program. X.Li was supported by NHLBI TOPMed Fellowship program. L.M.R. was also supported by R01HG010297. X.Lin was supported by NHLBI R01HL163560 and NHGRI U01 HG009088 and U01HG012064. E.J.B. was supported by 75N92019D00031, 1R01 HL64753, R01 HL076784, 1R01 AG028321, and R01HL092577. H.L. was supported by U01AG068221. B.M.P. was supported by R01HL105756. J.C.B. was supported by R01HL105756. The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant KL2TR002490 (L.M.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 03/31/2020.

## Supplementary data

Supplementary data is available at HMG Journal online.

**Conflict of interest statement:** LMR is a consultant for the TOPMed Administrative Coordinating Center (through Westat). SG is a current employee and stockholder of Regeneron Genetics Center or Regeneron Pharmaceuticals. All other authors declare no conflict of interest.

In the past three years, Edwin K. Silverman received grant support from Bayer and Northpond Laboratories.

## References

- Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat Rev Cardiol* 2018;**15**:505–22.
- Pickup JC. Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes. *Diabetes Care* 2004;**27**:813–23.
- Lambrecht BN, Hammad H, Fahy JV. The cytokines of asthma. *Immunity* 2019;**50**:975–91.
- Ahola-Olli AV, Würtz P, Havulinna AS. et al. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *Am J Hum Genet* 2017;**100**:40–50.
- Ligthart S, Vaez A, Vösa U. et al. Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and high-light pathways that link inflammation and complex disorders. *Am J Hum Genet* 2018;**103**:691–706.
- Amaral WZ, Krueger RF, Ryff CD. et al. Genetic and environmental determinants of population variation in interleukin-6, its soluble receptor and C-reactive protein: insights from identical and fraternal twins. *Brain Behav Immun* 2015;**49**:171–81.
- Shah T, Zabaneh D, Gaunt T. et al. Gene-centric analysis identifies variants associated with interleukin-6 levels and shared pathways with other inflammation markers. *Circ Cardiovasc Genet* 2013;**6**:163–70.
- Austin MA, Zhang C, Humphries SE. et al. Heritability of C-reactive protein and association with apolipoprotein E genotypes in Japanese Americans. *Ann Hum Genet* 2004;**68**:179–88.
- Pankow JS, Folsom AR, Cushman M. et al. Familial and genetic determinants of systemic markers of inflammation: the NHLBI family heart study. *Atherosclerosis* 2001;**154**:681–9.
- Vickers MA, Green FR, Terry C. et al. Genotype at a promoter polymorphism of the interleukin-6 gene is associated with baseline levels of plasma C-reactive protein. *Cardiovasc Res* 2002;**53**:1029–34.
- Schnabel RB, Lunetta KL, Larson MG. et al. The relation of genetic and environmental factors to systemic inflammatory biomarker concentrations. *Circ Cardiovasc Genet* 2009;**2**:229–37.
- Fox ER, Benjamin EJ, Sarpong DF. et al. Epidemiology, heritability, and genetic linkage of C-reactive protein in African Americans (from the Jackson heart study). *Am J Cardiol* 2008;**102**:835–41.
- Suhre K, Arnold M, Bhagwat AM. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun* 2017;**8**:14357.
- Voruganti VS, Laston S, Haack K. et al. Genome-wide association replicates the association of Duffy antigen receptor for chemokines (DARC) polymorphisms with serum monocyte chemoattractant protein-1 (MCP-1) levels in Hispanic children. *Cytokine* 2012;**60**:634–8.
- Benjamin EJ, Dupuis J, Larson MG. et al. Genome-wide association with select biomarker traits in the Framingham heart study. *BMC Med Genet* 2007;**8**:S11.
- Naitza S, Porcu E, Steri M. et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS Genet* 2012;**8**:e1002480.
- Raffield LM, Iyengar AK, Wang B. et al. Allelic heterogeneity at the CRP locus identified by whole-genome sequencing in multi-ancestry cohorts. *Am J Hum Genet* 2020;**106**:112–20.
- Polfus LM, Raffield LM, Wheeler MM. et al. Whole genome sequence association with E-selectin levels reveals loss-of-function variant in African Americans. *Hum Mol Genet* 2019;**28**:515–23.
- 1000 Genomes Project Consortium, Auton A, Brooks LD. et al. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
- Zhou H, Arapoglou T, Li X. et al. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res* 2023;**51**:D1300–11.
- Pulit SL, de With SAJ, de Bakker PIW. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. *Genet Epidemiol* 2017;**41**:145–51.
- Sinnott-Armstrong N, Tanigawa Y, Amar D. et al. Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat Genet* 2021;**53**:185–94.
- Schick UM, Auer PL, Bis JC. et al. Association of exome sequences with plasma C-reactive protein levels in >9000 participants. *Hum Mol Genet* 2015;**24**:559–71.
- Ferkingstad E, Sulem P, Atlason BA. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 2021;**53**:1712–21.
- Pietzner M, Wheeler E, Carrasco-Zanini J. et al. Mapping the proteo-genomic convergence of human diseases. *Science* 2021;**374**:eabj1541.
- Sun BB, Maranville JC, Peters JE. et al. Genomic atlas of the human plasma proteome. *Nature* 2018;**558**:73–9.
- Zhang J, Dutta D, Köttgen A. et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat Genet* 2022;**54**:593–602.
- Wang M, Gao J, Liu J. et al. Genomic association vs. serological determination of ABO blood types in a Chinese cohort, with application in Mendelian randomization. *Genes* 2021;**12**:959.
- Paré G, Chasman DI, Kellogg M. et al. Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet* 2008;**4**:e1000118.
- Vösa U, Claringbould A, Westra H-J. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 2021;**53**:1300–10.
- GTEX Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30.
- Pazoki R, Vujkovic M, Elliott J. et al. Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nat Commun* 2021;**12**:2579.
- Rondina MT, Voora D, Simon LM. et al. Longitudinal RNA-Seq analysis of the repeatability of gene expression and splicing in human platelets identifies a platelet SELP splice QTL. *Circ Res* 2020;**126**:501–16.
- Wojcik GL, Graff M, Nishimura KK. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 2019;**570**:514–8.
- Ahluwalia TS, Prins BP, Abdollahi M. et al. Genome-wide association study of circulating interleukin 6 levels identifies novel loci. *Hum Mol Genet* 2021;**30**:393–409.
- Selvaraj MS, Li X, Li Z. et al. Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nat Commun* 2022;**13**:5995.
- Akbari P, Sosina OA, Bovijn J. et al. Multi-ancestry exome sequencing reveals INHBE mutations associated with favorable fat

- distribution and protection from diabetes. *Nat Commun* 2022;**13**:4844.
38. Singh T, Poterba T, Curtis D. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* 2022;**604**:509–16.
  39. Katz DH, Tahir UA, Bick AG. et al. Whole genome sequence analysis of the plasma proteome in black adults provides novel insights into cardiovascular disease. *Circulation* 2022;**145**:357–70.
  40. Kachuri L, Mak ACY, Hu D. et al. Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nat Genet* 2023;**55**:952–63.
  41. Spracklen CN, Iyengar AK, Vadlamudi S. et al. Adiponectin GWAS loci harboring extensive allelic heterogeneity exhibit distinct molecular consequences. *PLoS Genet* 2020;**16**:e1009019.
  42. Hanks SC, Forer L, Schönherr S. et al. Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. *Am J Hum Genet* 2022;**109**:1653–66.
  43. Halldorsson BV, Eggertsson HP, Moore KHS. et al. The sequences of 150,119 genomes in the UK biobank. *Nature* 2022;**607**:732–40.
  44. Kowalski MH, Qian H, Hou Z. et al. Use of >100,000 NHLBI trans-omics for precision medicine (TOPMed) consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* 2019;**15**:e1008500.
  45. Croteau-Chonka DC, Wu Y, Li Y. et al. Population-specific coding variant underlies genome-wide association with adiponectin level. *Hum Mol Genet* 2012;**21**:463–71.
  46. Sun W, Kechris K, Jacobson S. et al. Common genetic polymorphisms influence blood biomarker measurements in COPD. *PLoS Genet* 2016;**12**:e1006011.
  47. Sun BB, Chiou J, Traylor M. et al. Plasma proteomic associations with genetics and health in the UK biobank. *Nature* 2023;**622**:329–38.
  48. Pietzner M, Wheeler E, Carrasco-Zanini J. et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat Commun* 2021;**12**:6822.
  49. Joshi A, Mayr M. In Aptamers they trust: the caveats of the SOMAscan biomarker discovery platform from SomaLogic. In Aptamers they trust: the caveats of the SOMAscan biomarker discovery platform from SomaLogic. *Circulation* 2018, 2018;**138**:2482–5.
  50. Katz DH, Robbins JM, Deng S. et al. Proteomic profiling platforms head to head: leveraging genetics and clinical traits to compare aptamer- and antibody-based methods. *Sci Adv* 2022;**8**:eabm5164.
  51. Raffield LM, Dang H, Pratte KA. et al. Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* 2020;**20**:e1900278.
  52. Taliun D, Harris DN, Kessler MD. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* 2021;**590**:290–9.
  53. TOPMed whole genome sequencing methods: Freeze 8 <https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8> (accessed Mar 2, 2022).
  54. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* 2015;**39**:276–93.
  55. Conomos MP, Reiner AP, Weir BS. et al. Model-free estimation of recent genetic relatedness. *Am J Hum Genet* 2016;**98**:127–48.
  56. Stilp AM, Emery LS, Broome JG. et al. A system for phenotype harmonization in the National Heart, Lung, and Blood Institute trans-omics for precision medicine (TOPMed) program. *Am J Epidemiol* 2021;**190**:1977–92.
  57. Little A, Hu Y, Sun Q. et al. Whole genome sequence analysis of platelet traits in the NHLBI trans-omics for precision medicine (TOPMed) initiative. *Hum Mol Genet* 2022;**31**:347–61.
  58. Mikhaylova AV, McHugh CP, Polfus LM. et al. Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: the NHLBI TOPMed program. *Am J Hum Genet* 2021;**108**:1836–51.
  59. Selvaraj MS, Li X, Li Z. et al. Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nat Commun* 2022;**13**:5995.
  60. Hu Y, Stilp AM, McHugh CP. et al. Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program. *Am J Hum Genet* 2021;**108**:874–93.
  61. Sofer T, Zheng X, Gogarten SM. et al. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol* 2019;**43**:263–75.
  62. Gogarten SM, Sofer T, Chen H. et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* 2019;**35**:5346–8.
  63. National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services. *The NHLBI BioData Catalyst*. Bethesda, MD: The NHLBI BioData Catalyst, 2020.
  64. Lam PH, Chiang JJ, Chen E. et al. Race, socioeconomic status, and low-grade inflammatory biomarkers across the life-course: a pooled analysis of seven studies. *Psychoneuroendocrinology* 2021;**123**:104917.
  65. Farmer HR, Wray LA, Haas SA. Race, gender, and socioeconomic variations in C-reactive protein using the health and retirement study. *J Gerontol B Psychol Sci Soc Sci* 2021;**76**:583–95.
  66. Folkersen L, Fauman E, Sabater-Lleal M. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet* 2017;**13**:e1006706.
  67. Li X, Li Z, Zhou H. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* 2020;**52**:969–83.
  68. Li Z, Li X, Zhou H. et al. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat Methods* 2022;**19**:1599–611.
  69. Gaynor SM, Westerman KE, Ackovic LL. et al. STAAR workflow: a cloud-based workflow for scalable and reproducible rare variant analysis. *Bioinformatics* 2022;**38**:3116–7.
  70. Folkersen L, Gustafsson S, Wang Q. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* 2020;**2**:1135–48.
  71. Sudlow C, Gallacher J, Allen N. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.
  72. Yang J, Lee SH, Goddard ME. et al. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.
  73. Zhu Z, Zhang F, Hu H. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016;**48**:481–7.