REPORT

Allelic Heterogeneity at the *CRP* Locus Identified by Whole-Genome Sequencing in Multi-ancestry Cohorts

Laura M. Raffield,[1] Apoorva K. Iyengar,[1] Biqi Wang,[2] Sheila M. Gaynor,[3] Cassandra N. Spracklen,[1] Xue Zhong,[4] Madeline H. Kowalski,[5] Shabnam Salimi,[6] Linda M. Polfus,[7] Emelia J. Benjamin,[8,9,10] Joshua C. Bis,[11] Russell Bowler,[12] Brian E. Cade,[13,14] Won Jung Choi,[15] Alejandro P. Comellas,[16] Adolfo Correa,[17] Pedro Cruz,[18] Harsha Doddapaneni,[19] Peter Durda,[20] Stephanie M. Gogarten,[21] Deepti Jain,[21] Ryan W. Kim,[15] Brian G. Kral,[22,23] Leslie A. Lange,[24] Martin G. Larson,[2,10] Cecelia Laurie,[21] Jiwon Lee,[13] Seonwook Lee,[15] Joshua P. Lewis,[25] Ginger A. Metcalf,[19] Braxton D. Mitchell,[25,26] Zeineen Momin,[19] Donna M. Muzny,[19] Nathan Pankratz,[27] Cheol Joo Park,[15] Stephen S. Rich,[28] Jerome I. Rotter,[29] Kathleen Ryan,[25] Daekwan Seo,[15] Russell P. Tracy,[20,30] Karine A. Viaud-Martinez,[18] Lisa R. Yanek,[22] Lue Ping Zhao,[31,32] Xihong Lin,[3,33,34] Bingshan Li,[35] Yun Li,[1,5,36] Josée Dupuis,[2,10] Alexander P. Reiner,[37] Karen L. Mohlke,[1] Paul L. Auer,[38,*] TOPMed Inflammation Working Group, and NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Whole-genome sequencing (WGS) can improve assessment of low-frequency and rare variants, particularly in non-European populations that have been underrepresented in existing genomic studies. The genetic determinants of C-reactive protein (CRP), a biomarker of chronic inflammation, have been extensively studied, with existing genome-wide association studies (GWASs) conducted in >200,000 individuals of European ancestry. In order to discover novel loci associated with CRP levels, we examined a multi-ancestry population (n = 23,279) with WGS (~38× coverage) from the Trans-Omics for Precision Medicine (TOPMed) program. We found evidence for eight distinct associations at the *CRP* locus, including two variants that have not been identified previously (rs11265259 and rs181704186), both of which are non-coding and more common in individuals of African ancestry (~10% and ~1% minor allele frequency, respectively, and rare or monomorphic in 1000 Genomes populations of East Asian, South Asian, and European ancestry). We show that the minor (G) allele of rs181704186 is associated with lower CRP levels and decreased transcriptional activity and protein binding *in vitro*, providing a plausible molecular mechanism for this African ancestry-specific signal. The individuals homozygous for rs181704186-G have a mean CRP level of 0.23 mg/L, in contrast to individuals heterozygous for rs181704186 with mean CRP of 2.97 mg/L and major allele homozygotes with mean CRP of 4.11 mg/L. This study demonstrates the utility of WGS in multi-ethnic populations to drive discovery of complex trait associations of large effect and to identify functional alleles in noncoding regulatory regions.

[1]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; [2]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA; [3]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA; [4]Department of Medicine, Division of Genetic Medicine, Vanderbilt University, Nashville, TN 37232, USA; [5]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA; [6]Department of Epidemiology and Public Health, School of Medicine, University of Maryland, Baltimore, MD 21201, USA; [7]Department of Preventive Medicine, Center for Genetic Epidemiology, University of Southern California, Los Angeles, CA 90089, USA; [8]Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA; [9]Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA; [10]National Heart, Lung, and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA 01702, USA; [11]Department of Medicine, Cardiovascular Health Research Unit, University of Washington, Seattle, WA 98101, USA; [12]Department of Medicine, Division of Pulmonary, Critical Care & Sleep Medicine, National Jewish Health, Denver, CO 80206, USA; [13]Department of Medicine, Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA 02115, USA; [14]Department of Medicine, Division of Sleep Medicine, Harvard Medical School, Boston, MA 02115, USA; [15]Macrogen USA, Rockville, MD 20850, USA; [16]Department of Medicine, Division of Pulmonary and Critical Care, University of Iowa, Iowa City, IA 52242, USA; [17]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; [18]Illumina Laboratory Services, Illumina Inc., San Diego, CA 92122, USA; [19]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [20]Department of Pathology & Laboratory Medicine, Larner College of Medicine, University of Vermont, Burlington, VT 05446, USA; [21]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [22]GeneSTAR Research Program, Division of General Internal Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [23]Division of Cardiology, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [24]Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO 80045, USA; [25]Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [26]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD 21201, USA; [27]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455, USA; [28]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA; [29]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [30]Department of Biochemistry, Larner College of Medicine, University of Vermont, Burlington, VT 05446, USA; [31]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [32]School of Public Health, University of Washington, Seattle, WA 98195, USA; [33]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [34]Department of Statistics, Harvard University, Cambridge, MA 02138, USA; [35]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37232, USA; [36]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA; [37]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [38]Joseph J. Zilber School of Public Health, University of Wisconsin Milwaukee, Milwaukee, WI 53205, USA
*Correspondence: pauer@uwm.edu
https://doi.org/10.1016/j.ajhg.2019.12.002.

Whole-genome sequencing (WGS) data are being rapidly generated in deeply phenotyped cohorts or case-referent samples of complex disorders by projects such as the United Kingdom's 100,000 Genomes Project,[1] the National Institute of Mental Health's Whole Genome Sequencing for Psychiatric Disorders Consortium,[2] the National Human Genome Research Institute's Centers for Common Disease Genomics (CCDG) project (see Web Resources), and the National Heart, Lung, and Blood Institute's Trans-Omics for Precision Medicine (TOPMed) Program.[3] WGS resources can improve interrogation of low-frequency and rare variation associated with quantitative traits or clinical outcomes[4] compared to genotyping array-based studies. However, sample sizes remain modest compared to large-scale genome-wide association studies (GWASs).

WGS-based analysis may offer particular advantages for non-European populations currently underrepresented in GWASs, with ~95% of GWAS participants being of European or East Asian ancestry.[5] WGS can assess population-specific variants which are at very low frequency or absent in large European GWASs, including variants that are often poorly imputed with standard reference panels and genotyping arrays. Current imputation reference panels for non-European populations (notably 1000 Genomes phase 3, n = 5,008 haplotypes across 26 mostly non-European populations[6]) are also much smaller than resources like the Haplotype Reference Consortium (HRC) for European populations (n = 64,976 haplotypes),[7] making imputation of low-frequency variants more difficult. Along with discrepancies in imputation reference panel size, many genotyping arrays have poor genomic coverage in non-European populations.[8] Because WGS assesses the entire genome of each individual, the limitations of genotyping arrays and imputation reference panels are easily overcome, allowing better understanding of the genetic architecture of complex traits in non-European populations. Based on previous success in identifying novel coding low-frequency or population-specific variants for inflammatory biomarkers in sequencing-based analyses,[9,10] we evaluated the ability of WGS to identify additional high-impact non-coding variation for commonly assessed inflammation biomarker C-reactive protein (CRP).

CRP is an acute-phase protein synthesized in the liver and is often used as a biomarker for chronic low-grade inflammation. As such, its relationship to cardiovascular disease (CVD) has been well established by numerous epidemiological studies, though current analyses do not point to a causal relationship with CVD.[11,12] CRP has also been associated with inflammatory disorders,[13,14] type 2 diabetes,[15] and overall mortality,[16] and recent Mendelian randomization studies have pointed to a potential causal role in bipolar disorder and schizophrenia.[12]

CRP demonstrates substantial heritability in family-based studies (~30% in East Asians,[17] ~30%–40% in Europeans,[18–20] ~45% in African Americans[21]). CRP levels vary by race/ethnicity group with higher levels observed in individuals of African ancestry compared to European or East Asian ancestry.[22,23] The genetic architecture of CRP has been investigated in diverse populations by whole-exome sequencing (WES),[10] genome-wide association,[24–26] and fine-mapping studies imputed to various reference panels[27,28] in tens of thousands of samples. Most recently, the largest GWAS was conducted in up to 204,402 individuals of European ancestry, identifying 58 loci and explaining 7% of the trait variance.[12] Some studies have also reported population-specific variants associated with CRP levels.[27] Among reported loci, the locus surrounding the *CRP* (MIM: 123260) gene itself on chromosome 1 explains the largest portion of phenotypic variance (1.4%[12]), with multiple distinct signals reported and clear evidence of allelic heterogeneity across populations.[27,28] For example, using approximate conditional analysis, the most recent European GWAS analysis reported 13 signals at the *CRP* locus (including rs149520992, an intergenic variant with a minor allele frequency [MAF] of 1% in Europeans and rare in other populations),[12] and four distinct signals (shared across ancestry groups) were reported in the multi-ethnic fine-mapping effort from the Population Architecture using Genomics and Epidemiology (PAGE) study.[28] African-specific variant rs726640 or variants in linkage disequilibrium (LD) with it have also been reported in several previous studies.[26,27,29]

Using data from the NHLBI TOPMed WGS project, we sought to investigate the additional value of WGS (beyond whole-exome sequencing and imputed GWAS) for single-variant analysis in a set of 23,279 individuals predominantly of self-reported European, African American, East Asian, and Hispanic/Latino ancestry with measured CRP levels (Table S1). We identified association with CRP levels at eight known loci (*CRP, APOE* [MIM: 107741], *HNF1A* [MIM: 142410], *LEPR* [MIM: 601007], *GCKR* [MIM: 600842], *IL6R* [MIM: 147880], *IL1F10* [MIM: 615296], and *NLRP3* [MIM: 606416]) with $p < 1 \times 10^{-9}$ in an ancestry-pooled genome-wide single-variant analysis (Table 1, Figure S1). We also examined these eight CRP-associated loci separately in African American (n = 6,545) and European American (n = 15,065) participants (Table S2). In the European American analysis, at least one variant at each locus met the locus-wide significance threshold for association with CRP levels with the exception of the *NLRP3* locus. The African American analysis also demonstrated at least one locus-wide significant variant at all loci except *GCKR* and *LEPR*.

We performed stepwise conditional analyses at each of the eight loci by conditioning on the lead variant at each locus and then sequentially conditioning on each new lead variant until no variants met our locus-wide significance thresholds (Table 1). Stepwise conditional analyses were performed in ancestry pooled and stratified (self-reported European American- and African American-specific) analyses. We identified two conditionally distinct signals at *HNF1A* and eight at the *CRP* locus (Table 2, Figures 1, S2, and S3). The presence of multiple association signals

Table 1. Eight Loci Significantly Associated ($p < 1 \times 10^{-9}$) with C-Reactive Protein Levels in TOPMed

Locus	Lead Variant	Annotation	p Value	Beta	Effect Allele	TOPMed EAF Overall	TOPMed African American EAF	TOPMed European American EAF	After Conditioning on Lead Variant			
									New Lead Variant	p Value	2nd Signal Threshold	Total # Signals
LEPR	rs7516341	intronic	1.9E−19	−0.09	C	0.43	0.54	0.37	rs72683129	4.7E−05	4.7E−06	1
IL6R	rs4129267	intronic	5.0E−12	−0.07	T	0.33	0.14	0.40	rs149417774	2.7E−04	6.3E−06	1
CRP	rs7551731	intergenic	1.1E−65	−0.18	C	0.30	0.22	0.33	rs73024795	1.2E−42	2.4E−06	8
NLRP3	rs56188865	intronic	2.6E−11	−0.06	C	0.42	0.52	0.38	rs115695052	1.6E−05	4.5E−06	1
GCKR	rs1260326	missense, p.Leu446Pro (*GCKR*)	1.9E−13	−0.08	C	0.66	0.85	0.58	rs183628627	4.7E−04	6.7E−06	1
IL1F10	rs6734238	intergenic	8.4E−12	0.07	G	0.41	0.45	0.41	rs148498391	4.1E−04	6.2E−06	1
HNF1A	rs2243458	intronic	1.5E−33	−0.13	T	0.27	0.12	0.33	rs544759708	3.3E−06	4.3E−06	2
APOE	rs429358	missense, p.Cys130Arg (*APOE4*)	1.1E−65	−0.22	C	0.15	0.21	0.13	rs186472069	1.6E−05	4.7E−06	1

Significance threshold for identification of second signals calculated as p = (0.05/tested variants). EAF, effect allele frequency, for those in TOPMed CRP analysis.

at both *CRP* and *HNF1A* has been reported in previous studies, with at least two signals identified at both loci in a recent multi-ethnic fine-mapping effort (four signals at *CRP*, two signals at *HNF1A*)[28] and in the largest European meta-analysis (13 approximate conditional signals at *CRP* and 2 at *HNF1A*).[12] The eight identified signals at the *CRP* locus include low-frequency, exonic variants (rs1800947 [p.Leu184Leu] and rs553202904, a noncoding proxy for rs77832441 [p.Thr59Met]) and noncoding variants with much higher MAF in African ancestry individuals. These African American-driven signals include both known (rs73024795) and previously unreported (rs11265259, rs181704186) associations. In an unrelated subset (n = 17,371), these eight conditionally distinct signals explained 4.2% of variance in natural log transformed CRP (2.6% in European Americans, 6.0% in African Americans). When performing stepwise conditional analyses at the *CRP* locus separately by ancestry, five conditionally distinct signals were identified in African Americans alone and four conditionally distinct signals were identified in European Americans. Based on these results and with consideration of population-specific allele frequencies, four signals at *CRP* were driven primarily by African American individuals (rs73024795, rs11265259, rs181704186, rs2211321) and two by European Americans (rs553202904, rs12734907) (Table S3). The other two signals (rs7551731 and rs1800947) were shared between African Americans and European Americans.

To determine whether the association signals we observed at the *CRP* or *HNF1A* loci were tagging previously reported associations, we performed a separate conditional analysis by which we adjusted for all variants associated with CRP levels at the *CRP* or *HNF1A* loci in prior GWAS, fine-mapping, or exome-sequencing efforts (Tables S4 and S5). In this analysis, two African American-driven signals at *CRP* remained locus-wide significant including rs11265259 (signal "E"; $\beta = -0.32$, $p = 7.3 \times 10^{-18}$; African American MAF = 0.10) and rs181704186 (signal "H"; $\beta = -0.46$, $p = 3.0 \times 10^{-7}$; African American MAF = 0.01); both are rare or monomorphic in other ancestry populations, with no copies of the minor allele for either variant found in 1000 Genomes European, East Asian, or South Asian populations. We also note the unusually large effect size for rs181704186, with major allele homozygotes having mean CRP levels of 4.11 mg/L (similar to the overall TOPMed mean of 4.10 mg/L), heterozygotes, 2.97 mg/L, and minor allele homozygotes, 0.23 mg/L, respectively (Figure 2A). By contrast, the more common variant, rs11265259, has mean CRP levels of 4.10, 4.36, and 3.04 mg/L, respectively. LD in African Americans from TOPMed between rs11265259 and rs181704186 and known signals is listed in Table S6. After adjusting for known variants at the *HNF1A* locus (Table S5), both association signals were attenuated below the locus-wide significance threshold. We thus carried forward the two conditionally distinct *CRP* signals, and not the secondary signal at *HNF1A*, for further follow-up.

As both remaining *CRP* variant associations appeared to be distinct from any previously identified *CRP* locus variant association, we attempted to replicate these two signals using CRP measurements in African American women from the Women's Health Initiative (WHI) study (n = 7,108). The WHI participants had genotype data from an Affymetrix 6.0 array imputed to the TOPMed reference panel (freeze 5b, Michigan Imputation Server) but were not whole genome sequenced through TOPMed at the time of freeze 5b's release. Both variants were locus-wide significant (using the same $p = 2.47 \times 10^{-6}$ locus-wide threshold used in our TOPMed analysis in Table 2) in our independent WHI replication sample of African Americans (Table S7, rs11265259, $p = 6.1 \times 10^{-9}$, rs181704186, $p = 9.2 \times 10^{-11}$) with consistent direction

Table 2. Eight Conditionally Distinct Signals Associated with C-Reactive Protein Were Identified at the *CRP* Locus in TOPMed

Signal	Variant	Annotation	Beta	p Value	Effect Allele	TOPMed Overall EAF	TOPMed African American EAF	TOPMed European American EAF	1000 Genomes AFR EAF	1000 Genomes EUR EAF	Sequential Conditional p Value
A	rs7551731	intergenic	−0.18	1.1E−65	C	0.30	0.22	0.33	0.20	0.31	–
B	rs73024795	intergenic	0.36	5.0E−54	T	0.05	0.16	4.98E−04	0.18	N/A	1.2E−42
C	rs2211321	intergenic	−0.02	0.05	C	0.70	0.65	0.71	0.64	0.71	3.1E−27
D	rs553202904[a]	intergenic	−0.70	1.4E−12	G	0.002	3.82E-04	0.003	N/A	0.003	8.8E−17
E	rs11265259	intergenic	−0.18	8.9E−09	C	0.03	0.09	4.31E−04	0.10	N/A	9.3E−12
F	rs1800947	synonymous, p.Leu184Leu	−0.24	5.8E−26	G	0.05	0.01	0.06	0.002	0.05	9.2E−09
G	rs12734907	intergenic	0.08	1.5E−12	T	0.26	0.08	0.34	0.02	0.37	7.9E−10
H	rs181704186	intergenic	−0.61	3.9E−12	G	0.003	0.009	9.96E−05	0.01	N/A	1.0E−07

Abbreviations: AFR, African; EUR, European; N/A, not applicable (monomorphic). Letters correspond to the signals displayed in the LocusZoom plot in Figure 1. Beta, p value, and overall effect allele frequency are from TOPMed pooled ancestry analysis. EAF, effect allele frequency, for those in TOPMed CRP analysis.
[a]Proxy variant is missense, Thr59Met ($r^2 = 0.98$ in analyzed TOPMed samples)

of effect. This remained true when conditioning on all known variants from prior GWASs and exome-sequencing studies in Table S4 (rs11265259, $p = 8.7 \times 10^{-12}$, rs181704186, $p = 9.7 \times 10^{-6}$). These replication results in WHI provide evidence to the validity of these variants and show the utility of the TOPMed reference panel for imputation in non-European ancestry individuals.

We performed several *in silico* analyses to further characterize the putative functional regulatory mechanisms of these two variants. Both rs11265259 (located ~6 kb downstream of *CRP*, signal E) and rs181704186 (located ~37 kb upstream of *CRP*, signal H) have high Genomic Evolutionary Rate Profiling (GERP)[31] scores (7.08 for rs11265259, 7.45 for rs181704186), indicating sequence conservation across species. In addition, both variants are located in predicted enhancer regions based on ChromHMM[32] models in liver (Figures 2B and S4), where CRP is produced. Neither is in strong LD (defined as $r^2 > 0.8$) with any other variant sequenced in the TOPMed African American samples. Integrated functional annotation scores from FUN-LDA comparing all Roadmap Epigenomics project tissues were highest in adult liver for both variants (Table S8a), suggesting that liver is a likely tissue in which these variants play a functional role. The annotation score for rs181704186 was 1.0 in liver, the highest possible score. The highest score for rs11265259 was more modest (0.0746), suggesting weaker evidence of enhancer function for this variant. Concordant with these results, our cross-tissue annotation principal components analysis (see Supplemental Material and Methods) found that both rs181704186 and rs11265259 were in the top 10% for conservation (scores of 18.8 and 16.3, respectively), with rs181704186 also having high epigenetics and transcription factor binding scores (Table S8b). Neither *CRP* locus variant E nor H was colocalized with eQTLs from any tissue available in GTEx,[33] whole blood (eQTLGen browser[34]), or in a recent large adult liver eQTL analysis.[35] Curiously, however, the latter liver eQTL mega-analysis identified no *cis*-eQTL for *CRP*, despite the very high expression of *CRP* in the liver.[35] We do note, however, that existing eQTL datasets that include some African Americans (such as GTEx) are fairly small; greater sample sizes and increased genetic diversity of included participants are needed to better explore eQTL effects for ancestry specific or low frequency variants like rs181704186 and rs11265259. However, GeneHancer[36] did link the enhancer region containing rs181704186 to the *CRP* gene ("elite" enhancer-gene connection [interaction confidence score 10.61], reflecting both a high-likelihood enhancer and strong enhancer-gene link). In summary, rs181704186 in particular had strong functional annotation scores in a relevant tissue for CRP levels (liver), as well as a large effect size, making it an attractive candidate for functional follow-up.

Finally, because we observed multiple independent signals at the CRP locus, we attempted to jointly model these effects with the FINEMAP statistical fine-mapping approach. We ran FINEMAP separately on the African American (AA) and European American (EA) samples, assuming a maximum of 5 causal variants in AAs and 4 causal variants in EAs (based on the results from the ancestry-specific conditional analyses). The FINEMAP method identified 7 variants in the 95% credible set in AAs (see Table S9 for all variants in the credible sets, including AA conditional analysis lead rs11265259) and 26 variants in EAs, including conditional analysis lead variants rs2211320 and rs1800947. Interestingly, while rs11265259 was included in the 95% credible set in AAs, rs181704186 was not ($r^2 < 0.03$ with all 7 credible set variants). Nevertheless, we nominated the rs181704186 variant for experimental follow up based on the preponderance of annotation-based evidence detailed above.

We performed further *in vitro* functional assays to characterize the regulatory role of rs181704186. We cloned a 1141-bp element designed to capture the surrounding

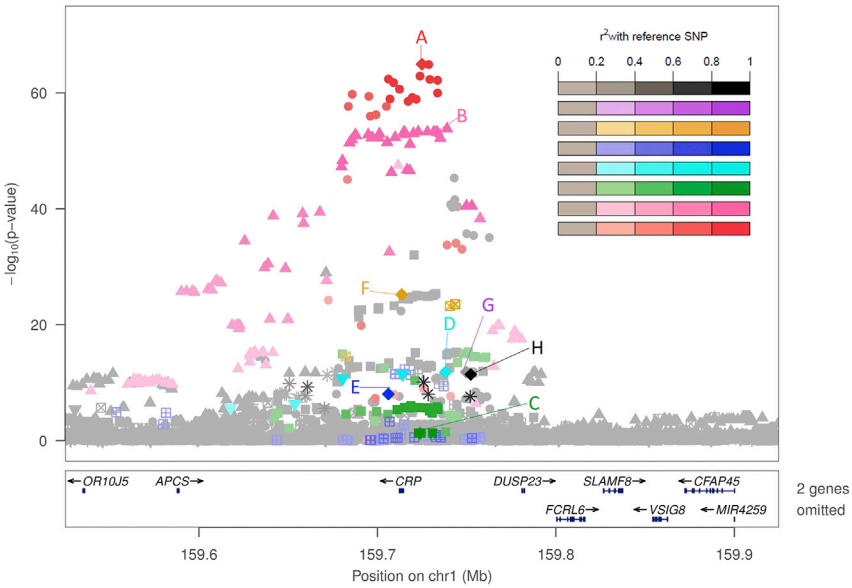

Figure 1. Eight Conditionally Distinct Signals Associated with C-Reactive Protein Were Identified at the *CRP* Locus in TOPMed

LocusZoom plot of −log10(p value) versus genomic location for all distinct signals at the *CRP* locus. Letters correspond to the list of conditionally distinct signals in Table 2. The lead variant for each conditionally distinct signal is indicated with a diamond, with other variants in linkage disequilibrium $r^2 > 0.2$ indicated in the colors used for each letter label and displayed on the legend at right, each with a different shape (for example, variants in close linkage disequilibrium with signal A (rs7551731) are displayed as red circles). Linkage disequilibrium is calculated using the same TOPMed samples included in our pooled ancestry C-reactive protein analyses.

regions of accessible chromatin and of cross-species conservation and containing each allele into a luciferase reporter vector in both orientations with respect to a minimal promoter (Table S10). Allele-specific clones of the reporter vector were transfected into the HepG2 hepatocyte/liver carcinoma cell line. Consistent with the GWAS direction of effect, the G allele associated with lower CRP levels was also associated with lower transcriptional activity in both the forward and reverse orientations (Figures 2C and S5A) than the A allele. *In vivo*, this likely reflects lower transcription of *CRP*, based on proximity and the GeneHancer links between this enhancer and the *CRP* transcription start site.[36] The cloned regulatory element appears to be a repressor, as the levels of transcriptional activity are lower than empty vector controls (Figure 2C).

We next performed an electrophoretic mobility shift assay (EMSA) to test the alleles of rs181704186 for differences in transcription factor binding (Figures 2E and S5B–S5D). We observed an allele-specific band at rs181704186-A (as indicated with an arrow; comparing lane 2 versus 7) that is competed away by a 40× excess of a probe containing the A allele (lane 3), but unaffected by probes containing the G allele (lane 4). The rs181704186 variant overlaps a CCAAT Enhancer Binding Protein Beta (CEBPB) binding site in ENCODE ChIP-seq experiments from HepG2 and HeLa cells, along with several other transcription factor binding proteins (Figure 2B). The rs181704186-G allele is predicted to disrupt the CEBPB motif, changing the position weight matrix log of the odds score from 14.8 to 2.9[17,18] (Figure 2D). CEBPB is a transcription factor known to be important for production of CRP in liver[37,38] and a strong candidate for contributing to the observed allelic differences in transcriptional activity. We attempted to supershift the EMSA DNA-protein complexes with antibodies to CEBPB. Incubation with an antibody targeting CEBPB showed a weaker band, which may represent a partially disrupted the A-allele-specific protein-DNA complex (lane 5). These allele-specific differences in protein binding are concordant with the transcriptional reporter assay and are suggestive that disruption of transcription factor binding at least partially mediates these regulatory effects, although further evidence is needed to determine the role of CEBPB and/or other transcription factors.

Using data from the TOPMed program, we report two low-frequency, population-specific variants that are associated with circulating CRP levels. Prior studies of genotypes imputed to the 1000 Genomes reference panels have not detected these associations. The best powered CRP GWAS to date included only individuals of European ancestry,[12] a population for which these variants would not have been detectable given their very low frequency. Notably, a recent study from the PAGE consortium included CRP as an exemplary quantitative trait, with data from 8,349 African Americans with CRP, genotyped on the Multi-Ethnic Genotyping Array (MEGA) and imputed to 1000 Genomes Phase 3. Neither variant was observed to be associated with CRP, despite detailed examination of secondary signals in a larger pooled sample size than available here for African Americans (and in a sample including some of the same African American participants, notably from WHI, as in our discovery and replication cohorts). This suggests that the use of a genotyping array developed to more equitably capture global genetic variation and subsequent imputation to the 1000 Genomes reference panel may still miss some population-specific variant associations that can be identified using WGS. In WHI our CRP-associated variants can be well imputed using TOPMed as a reference panel (imputation quality $r^2 \geq 0.9$); the TOPMed reference panel has ~20× larger sample size than 1000 Genomes Phase 3, and increased imputation quality is expected in African Americans based on previous work.[39] Imputation quality is only modestly attenuated in WHI using 1000 Genomes

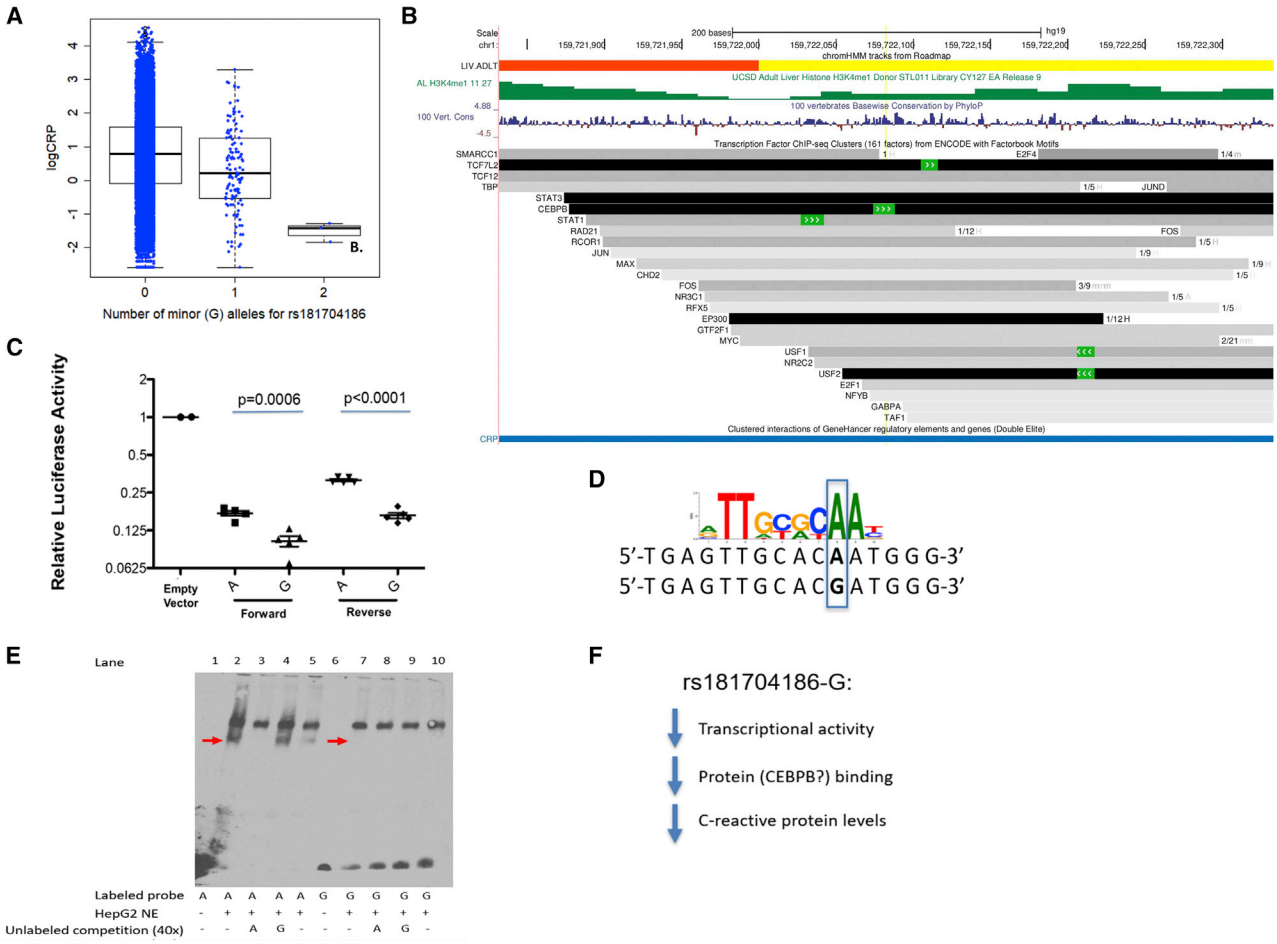

Figure 2. Regulatory Role of Low-Frequency, African Ancestry-Specific Variant rs181704186
(A) Boxplot of natural log-transformed CRP values by allele for rs181704186 (for 23,157 major allele homozygotes, 119 heterozygotes, and 3 minor allele homozygotes).
(B) Genome browser plot for rs181704186, chromHMM annotation in adult liver (yellow, enhancer; yellow, enhancer; red, transcription start site) from RoadMap Epigenomics, H3K4me1 signal from adult liver, 100 vertebrates basewise conservation by PhyloP, transcription factor ChIP-seq clusters from ENCODE (161 factor version, motifs highlighted in green, proportion cell types detected/total number of cell types assayed displayed). We also display GeneHancer's connection of the region containing this variant to *CRP*. No other variants have linkage disequilibrium $r^2 \geq 0.8$ with lead variant rs181704186.
(C) Luciferase assay demonstrating reduced transcriptional activity for the G allele, which is also associated with lower CRP levels. Blue lines indicate the groups compared for each listed p value.
(D) Disrupted CEBPB transcription factor binding motif position weight matrix from Kheradpour and Kellis[30] (CEBPB-disc1, with blue box highlighting position changed by rs181704186).
(E) Differential protein binding for A and G allele in EMSA assay. EMSA with biotin-labeled probes containing the A or G allele of rs181704186 shows an allele-specific band (lane 2 versus 7, indicated with red arrows) that is competed away by 40-fold excess of unlabeled probe containing the A allele (lane 3), but unaffected by a 40-fold excess of probe containing the G allele (lane 4). Incubation with an antibody targeting CEBPB partially disrupts the A-allele-specific protein-DNA complex (lane 5). NE, nuclear extract.
(F) Summary of direction of effect of rs181704186-G.

Phase 3 as a reference panel (imputation quality $r^2 \geq 0.75$), but this still leads to weaker association for rs11265259 in particular using 1000 Genomes imputation, likely due to a reduction in effective sample size (product of sample size and r^2). Concurrent association analysis in both sequenced and imputed data (using the largest relevant sequencing dataset, such as TOPMed, as a reference panel) may be a powerful strategy for discovering low-frequency and rare variant associations with many complex traits, particularly in non-European populations.[39]

Our results using WGS and replicated with TOPMed imputed data exemplify the value of WGS in individuals of diverse genetic ancestry. Despite having only 10% of the sample size of the largest European GWAS meta-analysis to date, the genetic diversity and accurate genotype calls for low frequency and rare variants in our multi-ancestry study afforded us the ability to detect additional population-specific association signals, including a low-frequency variant with a large effect size. These association signals add to our knowledge of the extensive allelic heterogeneity and diversity of the *CRP* genomic region, which

contains a number of shared and population-specific coding and regulatory alleles.[10,12,28] Ultimately, finer dissection of the functional alleles at the *CRP* locus may have consequences for understanding the biology of acute or chronic inflammation or the causal role of CRP in inflammation-related complex disorders. To determine whether the two replicated African-specific CRP-associated variants (rs11265259 and rs181704186) have downstream clinical consequences, we performed a phenome-wide association study (pheWAS) in the BioVU biobank. No phenotype associations were statistically significant at a Bonferroni adjusted level. Though this result may be a consequence of small sample size or sub-optimal imputation quality, it is largely consistent with previous studies that have failed to find a large number of clinical outcomes that correlate with CRP-associated variants.[12]

A primary goal of many human genetics studies is to identify the causal allele that underlies the association with a human trait or disease. As such, the value of deep sequencing data on hundreds of thousands of individuals from diverse genetic backgrounds should not be understated. Our results demonstrate the potential for WGS analysis to discover genetic signals, including conditionally distinct, low-frequency signals at known loci. Limitations of our current analysis include the modest sample size, particularly for ancestry groups other than European and African Americans, and the focus on single-variant tests only. As larger sample sizes become available, further study of aggregate tests for very rare variants and structural variation is warranted. Future studies from TOPMed and other large WGS efforts integrating both sequencing data and dense imputation, along with interrogation of rich functional annotation databases and higher-throughput cellular assays, will continue to clarify the role of genetic variation on complex traits.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.12.002.

Acknowledgments

Analysis of CRP variants was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (RO1 DK072193 and U01 DK105561). A.K.I. and K.L.M. were supported by RO1 DK072193. L.M.R. was supported by T32 HL129982. C.N.S. was supported by American Heart Association Postdoctoral Fellowship 15POST24470131 and 17POST33650016. E.J.B. was supported by HHSN2682015000001I, N01-HC 25195, RO1 HL64753, R01 HL076784, and R01 AG028321. B.E.C. was supported by K01 HL135405. M.H.K., Y.L., and A.P.R. were supported by R01 HL129132. A.C. was supported by HHSN268201800010, HHSN268201800011, HHSN268201800012, HHSN268201800013, HHSN268201800015, and HHSN268201800015. J.P.L. was supported by R01 HL137922. R.P.T. was supported by R01 HL120854. J.D. was supported by R01 HL128914. P.L.A. was supported by R01 HL132947. B.L. was supported by U01HG009086. S.S. was supported by K01AG059898.

Declaration of Interests

The authors declare no competing interests.

Received: October 4, 2019
Accepted: December 2, 2019
Published: December 26, 2019

Web Resources

AuthorArranger, https://authorarranger.nci.nih.gov/#/
Centers for Common Disease Genomics (CCDG), https://ccdg.rutgers.edu/
ENCORE, https://encore.sph.umich.edu/
eQTLGen, https://www.eqtlgen.org/index.html
GTEx, https://www.gtexportal.org/home/
OASIS, https://edn.som.umaryland.edu/OASIS/
OMIM, https://www.omim.org/
TOPMed Methods, https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2

References

1. The NIHR BioResource on behalf of the 100000 Genomes Project. (2019). Whole-genome sequencing of rare disease patients in a national healthcare system. bioRxiv. https://doi.org/10.1101/507244.
2. Sanders, S.J., Neale, B.M., Huang, H., Werling, D.M., An, J.-Y., Dong, S., Abecasis, G., Arguello, P.A., Blangero, J., Boehnke, M., et al.; Whole Genome Sequencing for Psychiatric Disorders (WGSPD) (2017). Whole genome sequencing in psychiatric disorders: the WGSPD consortium. Nat. Neurosci. *20*, 1661–1668.
3. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv. https://doi.org/10.1101/563866.
4. Lappalainen, T., Scott, A.J., Brandt, M., and Hall, I.M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. Cell *177*, 70–84.
5. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
6. The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
7. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
8. Wojcik, G.L., Fuchsberger, C., Taliun, D., Welch, R., Martin, A.R., Shringarpure, S., Carlson, C.S., Abecasis, G., Kang, H.M., Boehnke, M., et al. (2018). Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. G3 (Bethesda) *8*, 3255–3267.
9. Polfus, L.M., Raffield, L.M., Wheeler, M.M., Tracy, R.P., Lange, L.A., Lettre, G., Miller, A., Correa, A., Bowler, R.P., Bis, J.C., et al. (2019). Whole genome sequence association with E-

selectin levels reveals Loss-of-function variant in African Americans. Hum. Mol. Genet. *28*, 515–523.
10. Schick, U.M., Auer, P.L., Bis, J.C., Lin, H., Wei, P., Pankratz, N., Lange, L.A., Brody, J., Stitziel, N.O., Kim, D.S., et al.; Cohorts for Heart and Aging Research in Genomic Epidemiology; and National Heart, Lung, and Blood Institute GO Exome Sequencing Project (2015). Association of exome sequences with plasma C-reactive protein levels in >9000 participants. Hum. Mol. Genet. *24*, 559–571.
11. Prins, B.P., Abbasi, A., Wong, A., Vaez, A., Nolte, I., Franceschini, N., Stuart, P.E., Guterriez Achury, J., Mistry, V., Bradfield, J.P., et al.; PAGE Consortium; International Stroke Genetics Consortium; Systemic Sclerosis consortium; Treat OA consortium; DIAGRAM Consortium; CARDIoGRAMplusC4D Consortium; ALS consortium; International Parkinson's Disease Genomics Consortium; Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium; CKDGen consortium; GERAD1 Consortium; International Consortium for Blood Pressure; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and Inflammation Working Group of the CHARGE Consortium (2016). Investigating the Causal Relationship of C-Reactive Protein with 32 Complex Somatic and Psychiatric Outcomes: A Large-Scale Cross-Consortium Mendelian Randomization Study. PLoS Med. *13*, e1001976.
12. Ligthart, S., Vaez, A., Võsa, U., Stathopoulou, M.G., de Vries, P.S., Prins, B.P., Van der Most, P.J., Tanaka, T., Naderi, E., Rose, L.M., et al.; LifeLines Cohort Study; and CHARGE Inflammation Working Group (2018). Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. Am. J. Hum. Genet. *103*, 691–706.
13. Markatseli, T.E., Voulgari, P.V., Alamanos, Y., and Drosos, A.A. (2011). Prognostic factors of radiological damage in rheumatoid arthritis: a 10-year retrospective study. J. Rheumatol. *38*, 44–52.
14. Gaitonde, S., Samols, D., and Kushner, I. (2008). C-reactive protein and systemic lupus erythematosus. Arthritis Rheum. *59*, 1814–1820.
15. Wang, X., Bao, W., Liu, J., Ouyang, Y.-Y., Wang, D., Rong, S., Xiao, X., Shan, Z.-L., Zhang, Y., Yao, P., and Liu, L.G. (2013). Inflammatory markers and risk of type 2 diabetes: a systematic review and meta-analysis. Diabetes Care *36*, 166–175.
16. Zacho, J., Tybjaerg-Hansen, A., and Nordestgaard, B.G. (2010). C-reactive protein and all-cause mortality–the Copenhagen City Heart Study. Eur. Heart J. *31*, 1624–1632.
17. Austin, M.A., Zhang, C., Humphries, S.E., Chandler, W.L., Talmud, P.J., Edwards, K.L., Leonetti, D.L., McNeely, M.J., and Fujimoto, W.Y. (2004). Heritability of C-reactive protein and association with apolipoprotein E genotypes in Japanese Americans. Ann. Hum. Genet. *68*, 179–188.
18. Pankow, J.S., Folsom, A.R., Cushman, M., Borecki, I.B., Hopkins, P.N., Eckfeldt, J.H., and Tracy, R.P. (2001). Familial and genetic determinants of systemic markers of inflammation: the NHLBI family heart study. Atherosclerosis *154*, 681–689.
19. Vickers, M.A., Green, F.R., Terry, C., Mayosi, B.M., Julier, C., Lathrop, M., Ratcliffe, P.J., Watkins, H.C., and Keavney, B. (2002). Genotype at a promoter polymorphism of the interleukin-6 gene is associated with baseline levels of plasma C-reactive protein. Cardiovasc. Res. *53*, 1029–1034.
20. Schnabel, R.B., Lunetta, K.L., Larson, M.G., Dupuis, J., Lipinska, I., Rong, J., Chen, M.-H., Zhao, Z., Yamamoto, J.F., Meigs, J.B., et al. (2009). The relation of genetic and environmental factors to systemic inflammatory biomarker concentrations. Circ Cardiovasc Genet *2*, 229–237.
21. Fox, E.R., Benjamin, E.J., Sarpong, D.F., Rotimi, C.N., Wilson, J.G., Steffes, M.W., Chen, G., Adeyemo, A., Taylor, J.K., Samdarshi, T.E., and Taylor, H.A., Jr. (2008). Epidemiology, heritability, and genetic linkage of C-reactive protein in African Americans (from the Jackson Heart Study). Am. J. Cardiol. *102*, 835–841.
22. Khera, A., McGuire, D.K., Murphy, S.A., Stanek, H.G., Das, S.R., Vongpatanasin, W., Wians, F.H., Jr., Grundy, S.M., and de Lemos, J.A. (2005). Race and gender differences in C-reactive protein levels. J. Am. Coll. Cardiol. *46*, 464–469.
23. Lakoski, S.G., Cushman, M., Palmas, W., Blumenthal, R., D'Agostino, R.B., Jr., and Herrington, D.M. (2005). The relationship between blood pressure and C-reactive protein in the Multi-Ethnic Study of Atherosclerosis (MESA). J. Am. Coll. Cardiol. *46*, 1869–1874.
24. Wu, Y., McDade, T.W., Kuzawa, C.W., Borja, J., Li, Y., Adair, L.S., Mohlke, K.L., and Lange, L.A. (2012). Genome-wide association with C-reactive protein levels in CLHNS: evidence for the CRP and HNF1A loci and their interaction with exposure to a pathogenic environment. Inflammation *35*, 574–583.
25. Okada, Y., Takahashi, A., Ohmiya, H., Kumasaka, N., Kamatani, Y., Hosono, N., Tsunoda, T., Matsuda, K., Tanaka, T., Kubo, M., et al. (2011). Genome-wide association study for C-reactive protein levels identified pleiotropic associations in the IL6 locus. Hum. Mol. Genet. *20*, 1224–1231.
26. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
27. Reiner, A.P., Beleza, S., Franceschini, N., Auer, P.L., Robinson, J.G., Kooperberg, C., Peters, U., and Tang, H. (2012). Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. Am. J. Hum. Genet. *91*, 502–512.
28. Kocarnik, J.M., Richard, M., Graff, M., Haessler, J., Bien, S., Carlson, C., Carty, C.L., Reiner, A.P., Avery, C.L., Ballantyne, C.M., et al. (2018). Discovery, fine-mapping, and conditional analyses of genetic variants associated with C-reactive protein in multiethnic populations using the Metabochip in the Population Architecture using Genomics and Epidemiology (PAGE) study. Hum. Mol. Genet. *27*, 2940–2953.
29. Doumatey, A.P., Chen, G., Tekola Ayele, F., Zhou, J., Erdos, M., Shriner, D., Huang, H., Adeleye, J., Balogun, W., Fasanmade, O., et al. (2012). C-reactive protein (CRP) promoter polymorphisms influence circulating CRP levels in a genome-wide association study of African Americans. Hum. Mol. Genet. *21*, 3063–3072.
30. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.
31. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., Sidow, A.; and NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
32. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930–D934.

33. Gamazon, E.R., Segrè, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., et al.; GTEx Consortium (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. *50*, 956–967.
34. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv. https://doi.org/10.1101/447367.
35. Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci. Rep. *8*, 5865.
36. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*, bax028.
37. Wang, T.M., Hsieh, S.C., Chen, J.W., and Chiang, A.N. (2013). Docosahexaenoic acid and eicosapentaenoic acid reduce C-reactive protein expression and STAT3 activation in IL-6-treated HepG2 cells. Mol. Cell. Biochem. *377*, 97–106.
38. Tsukada, J., Yoshida, Y., Kominato, Y., and Auron, P.E. (2011). The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. Cytokine *54*, 6–19.
39. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al. (2019). 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. bioRxiv. https://doi.org/10.1101/683201.