Large-Scale Genetic Analyses of Inflammatory Traits and Hematologic Parameters: Insights and Implications

Ursula Martine Schick

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Alexander P. Reiner, Chair

Paul L. Auer

Stephanie Malia Fullerton

Program Authorized to Offer Degree:

School of Public Health

University of Washington

Abstract

Large-Scale Genetic Analyses of Inflammatory Traits and Hematologic Parameters: Insights and Implications

Ursula Martine Schick

Chair of the Supervisory Committee:

Alexander P. Reiner, Research Professor

Department of Epidemiology

Genome-wide association studies (GWAS) have identified thousands of trait-associated common variants, but for most complex traits these polymorphisms have explained only a small proportion of the phenotypic variation. In an effort to assess the contribution of rare variation, sequencing based approaches targeting the exome and whole genome have now been applied to large numbers of individuals. The first aim of this dissertation seeks to identify novel associations in exome sequences from more than 9,000 participants with C-reactive protein (CRP) levels. This investigation revealed a putatively functional nonsynonymous variant in *CRP* and provided novel insights into variation in the apolipoprotein E gene (*APOE*). The second aim uses data from the exome array (primarily composed of rare variants identified through exome sequencing and a subset of common variants from GWAS) to investigate cross-phenotype associations in blood cell traits including red cell parameters (hematocrit and hemoglobin), platelet count, and

white blood cell count. This analysis confirmed previously identified cross-phenotype associations and provided evidence of novel cross-phenotype associations. The final aim of this study explores the issues surrounding somatic variants as incidental findings from genetic research. This aim seeks to summarize existing recommendations related to return of incidental findings to research participants and to evaluate the case for somatic mutations as returnable incidental findings using the Janus Kinase (*JAK2)* p.V617F mutation as a motivating example.

Acknowledgements

I wish to extend the sincerest thank you to the many individuals that assisted me in obtaining this degree. In particular:

- Alex Reiner
- Paul Auer
- Stephanie Malia Fullerton
- Tim Thornton
- Emily White and the National Cancer Institute
- Institute for Public Health Genetics faculty and students
- Immediate family and friends
- Tyler Jones

Table of Contents

Introduction

First applied in 2005, the Genome-Wide Association Study (GWAS) design has been successful at identifying single nucleotide polymorphisms (SNP) implicated in complex disease. This agnostic genotype-based approach captures common polymorphic loci distributed across the genome to assess association with a given phenotype. The nearly 2,000 published GWAS have identified thousands of associations, however these SNPs explain only a small fraction of the heritable variation of most complex traits.

In an effort to investigate the contribution of less common and rare variants to complex phenotypes, the National Heart, Lung and Blood Institute funded several exome sequencing projects including the Exome Sequencing Project (ESP) and Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE). Exome sequencing has previously been applied to discover the genetic basis of several Mendelian diseases, and results from ESP and CHARGE confirm that variants distributed across the allelic frequency spectrum contribute to complex traits. Specifically, we apply exome sequencing to identify novel associations and clarify previous associations with C-reactive protein (CRP) levels.

Beyond to rare variation, additional biological phenomenon may be important contributors to the etiology of complex traits. Moving into the post-GWAS era, detailed analyses are necessary to investigate the shared genetic architecture between correlated traits to better understand the shared pathophysiological mechanisms. The second aim of this dissertation seeks to identify and characterize cross-phenotype associations for blood cell phenotypes in an effort to explain direct and inverse correlation between these traits and better understand the shared genetic architecture underlying these traits.

In the course of identifying trait-associated loci, large-scale sequencing or genotyping studies may uncover variants with important health implications. As we observed in a recent study of blood traits, a rare somatic variant in janus kinase 2 was associated with our phenotypes of interest, but also confers information about development of myeloproliferative neoplasms. This work seeks to understand the implications of this finding and how the identification of this somatic mutation may contribute to the discussion of return of incidental findings from genomic research.

This dissertation project, covering a wide range of topics, performs genetic epidemiologic studies to evaluate the contribution of rare variation and to uncover pleiotropic effects in complex traits. Further, this work seeks to enumerate the complication of incidental identification of a somatic variant relevant to blood disorders.

Specific Aims

The Institute for Public Health Genetics doctoral program centers on two core areas of study: A) genomics in public health and B) implications of genetics for society. Consistently, this dissertation is divided according into these two areas of study. The A-side of this dissertation is composed of two related aims focusing on the genetic epidemiology of inflammatory and blood cell traits:

- Aim 1 focused on the identification of novel genes and variants associated with CRP levels in exome sequence data from the ESP and the CHARGE consortia.
- Aim 2 sought to identify cross-phenotype associations in blood cell traits in samples from the Women's Health Initiative.

The B-side aimed to explore the implications of the incidental identification of a somatic mutation in participants of genetic research studies. To explore the topic, a somatic mutation relevant for myeloproliferative neoplasms was utilized as a motivating example.

Part A, Aim I: Exome-based Analysis of CRP levels

Manuscript

Association of Exome Sequences with Plasma C-reactive Protein Levels in >9,000 Participants

Ursula M. Schick[1], Paul L. Auer[1,2], Joshua C. Bis[3], Honghuang Lin[4], Peng Wei[5], Nathan Pankratz[6], Leslie A. Lange[7], Jennifer Brody[3], Nathan O. Stitziel[8], Daniel S. Kim[9], Christopher S. Carlson[1], Myriam Fornage[5,10], Jeffery Haessler[1], Li Hsu[1,11], Rebecca D. Jackson[12], Charles Kooperberg[1], Suzanne M. Leal[13], Bruce M. Psaty[14,15], Eric Boerwinkle[5,16], Russell Tracy[17], Diego Ardissino[18], Svati Shah[19], Cristen Willer[20,21,22], Ruth Loos[23,24], Olle Melander[25], Ruth McPherson[26], Kees Hovingh[27,28], Muredach Reilly[29], Hugh Watkins[30,31], Domenico Girelli[32], Pierre Fontanillas[33], Daniel I. Chasman[34], Stacey B. Gabriel[33], Richard Gibbs[16], Deborah A. Nickerson[9], Sekar Kathiresan[33,35], Ulrike Peters[1], Josée Dupuis[36,37], James G. Wilson[38,*], Stephen S. Rich[39,*], Alanna C. Morrison[5,*], Emelia J. Benjamin[4, 36,40,*], Myron D. Gross[6,*], Alex P. Reiner[1, 41,*], on Behalf of the Cohorts for Heart and Aging Research in Genomic Epidemiology and the National Heart, Lung, and Blood Institute GO Exome Sequencing Project[#]

Affiliations

1. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, 2. School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA, 3. Cardiovascular Health Research Unit and Department of Medicine, University of Washington, Seattle, WA 98101, USA, 4. Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA, 5. Human Genetics Center, University of Texas Health Science Center, School of Public Health, Houston, TX 77030, USA, 6. Department of Lab Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455, USA, 7. Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC 27599, USA, 8. Cardiovascular Division, Department of Medicine and Division of Statistical Genomics, Washington University School of Medicine, Saint Louis, MO 63110, USA 9.Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA, 10. Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX, 77030, USA, 11. Department of Biostatistics, University of Washington, Seattle, WA 98105, USA, 12.Division of Endocrinology, Diabetes, and Metabolism, Ohio State University, Columbus, OH 43210, USA, 13. Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA, 14. Department of Epidemiology, Cardiovascular Health Research Unit, Department of Medicine, and Department of Health Services, University of Washington, Seattle, WA 98105, USA, 15. Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101, USA, 16. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA, 17. Departments of Biochemistry and Pathology, University of Vermont, Burlington, VT 05401, USA, 18. Division of Cardiology, Azienda Ospedaliero-Universitaria di Parma, Parma, Italy, 19. Division of Cardiology, Department of Medicine and Center for Human Genetics, Duke University, Durham, NC, 20. Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, Michigan, USA, 21. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA, 22. Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA, 23. The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, NY, USA, 24. The Mindich Child Health and Development Institute, The Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, 25. Department of Clinical Sciences, Diabetes and Endocrinology, Lund University, University Hospital Malmö, Malmö, Sweden, 26. Division of Cardiology, University of Ottawa Heart Institute, Ottawa, ON, Canada, 27. Department of Vascular Medicine, Academic Medical Center, Amsterdam, 1105 AZ, The Netherlands. 28. Department of Experimental Vascular Medicine, Academic Medical Center, Amsterdam, 1105 AZ, The Netherlands, 29. The Institute for Translational Medicine and Therapeutics and The Cardiovascular Institute, Perleman School of Medicine at the University of Pennsylvania, Philadelphia, PA, 19104, USA, 30. Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK, 31. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK, 32. University of Verona School of Medicine, Department of Medicine, Verona, Italy, 33. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, 34. Center for Cardiovascular Disease Prevention, Division of Preventative Medicine, Brigham and Women's Hospital, 900 Commonwealth Drive, Boston, Massachusetts 02115, USA 35. Department of Medicine, Harvard Medical School, Boston, MA 02115, USA 36. National Heart, Lung, and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA 01702, USA, 37.Department of Biostatistics, Boston University School of Public Health, MA 02118, USA, 38. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA, 39. Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA, 40. Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA, 41. Department of Epidemiology, University of Washington, Seattle, WA 98105, USA, *) These authors contributed equally to this work, [#])Authorship banner is included as supplemental materials.

Publication Details: This article has been accepted for publication in Human Molecular Genetics Published by Oxford University Press. (citation: 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. (2014). Association of exome sequences with plasma C-reactive protein levels in >9000 participants. Hum Mol Genet.)

Abstract

C-reactive protein (CRP) concentration is a heritable systemic marker of inflammation that is associated with cardiovascular disease risk. Genome Wide Association Studies have identified CRP-associated common variants associated in approximately 25 genes. The aims were to apply exome sequencing to 1) assess whether the candidate loci contain rare coding variants associated with CRP levels and; 2) perform an exome-wide search for rare variants in novel genes associated with CRP levels. We exome sequenced 6,050 European Americans and 3,109 African Americans from the NHLBI-ESP and the CHARGE consortia and performed association tests of sequence data with measured CRP levels. In single variant tests across candidate loci, a novel rare (MAF=0.16%) *CRP* coding variant (rs77832441-A; p.Thr59Met) was associated with 53% lower mean CRP levels ($P=2.9x10^{-6}$). We replicated the association of rs77832441 in an exome array genotype-based analysis of 11,414 European Americans ($P=3.0x10^{-15}$). Despite a strong effect on CRP levels, rs77832441 was not associated with inflammation-related phenotypes. In 19q13, we also found evidence for an African American-specific association of *APOE*-ε2 rs7214 with higher CRP levels. At the exome-wide significance level ($P<5.0x10^{-8}$), we confirmed associations for reported common variants of *HNF1A, CRP, IL6R,* and *TOMM40-APOE*. In gene-based tests, a burden of rare/low frequency variation in *CRP* in European Americans *(*$P\leq 6.8\ x10^{-4}$) and *RORA* in African Americans ($P=1.7\ x10^{-3}$) were associated with CRP levels at the candidate gene-level ($P<2.0\ x10^{-3}$). This inquiry did not elucidate novel genes, but instead demonstrated that variants distributed across the allele-frequency spectrum within candidate genes contribute to CRP levels.

Introduction

C-reactive protein (CRP) is an acute-phase protein reactant produced by the liver in response to proinflammatory stimuli. CRP is a sensitive, but nonspecific heritable biomarker of systemic inflammation that is associated with a variety of inflammation-mediated diseases[1-5]. Particular attention has been focused on characterizing the association between CRP and cardiovascular disease (CVD). Prospective epidemiologic studies suggest that basal CRP levels are predictive of risk of future CVD[2; 6-8], though the degree of association is dependent on levels of other conventional vascular risk factors[5; 9]. Current consensus recommendations support the clinical use of CRP to predict CVD risk among a subset of asymptomatic adults and in the selection of statin therapy[10]. Despite routine clinical use of CRP levels, data from Mendelian randomization studies suggests that CRP is unlikely to be causally related to CVD[11-17].

Genome Wide Association Studies (GWAS) have sought to characterize genetic determinates of CRP levels. This approach has been successful at identifying approximately 25 loci associated with CRP levels among individuals of European[17-22], Asian[23-25], and African[26; 27] descent (Table S1). To further interrogate known loci and to search for novel loci associated with CRP levels, we applied exome sequencing, which captures sequence variation in the protein-coding portion of the genome. Exome sequencing has proven useful for identifying rare causal variants for several Mendelian disorders [28-32]. Further, recent studies illustrate the application of exome sequencing to identify variation underlying complex traits [33-38].

In this study, we apply exome sequencing to a large sample of European American (EA) and African American (AA) ascertained through seven population-based cohorts (Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk Development in Young Adults (CARDIA), Cardiovascular Health Study (CHS), Framingham Heart Study (FHS),

Jackson Heart Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), and the Women's Health Initiative (WHI)) that compose the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Our specific aims were to 1) assess whether known CRP loci harboring common variants also contain rare coding variants associated with CRP levels, and; 2) perform an exome-wide search for rare variants in novel genes associated with CRP levels. Follow-up HumanExome BeadChip genotyping data from an independent sample derived from the WHI and JHS cohorts were used as replication for discovery findings.

Materials and Methods

Study Subjects and CRP Measurements

Our discovery sample consisted of exome sequence data from 3,360 individuals from the NHLBI-ESP and 5,799 individuals from the CHARGE project with valid CRP measures. In total, these 9,159 participants included 6,050 EA and 3,109 AA sampled from seven population-based cohorts: ARIC (N=4,827), CARDIA (N=190), CHS (N=946), FHS (N=1,144), JHS (N=346), MESA (N=399), and the WHI (N=1,307) as part of the NHLBI- ESP and an independent sample from three of the same population-based cohorts (ARIC, FHS, CHS) as part of the CHARGE consortium[39]. CRP levels were measured by high-sensitivity immunoassay in all 7 cohorts. Detailed descriptions of each of the seven cohorts and the techniques used to measure circulating CRP levels are provided in previous publications[40-45] and summarized in Table S2. Clinical information was collected through self-report and in-person examination. All participants provided written informed consent as approved by local human-subjects committees.

Sampling Design

The CHARGE cohorts' study participants included here were selected as part of a large random cohort sample or for extreme values for at least one of the following phenotypes: age at menopause, electrocardiogram QT interval, fasting blood glucose, fibrinogen level, renal function, Stamler-Kannel-like extremes of risk factors selected by principle components and waist-to-hip ratio. The sampling design of ESP included disease phenotypes (early onset myocardial infarction, ischemic stroke), and several quantitative cardiovascular risk factors that were sampled on the basis of phenotypic extremes (blood pressure, BMI and LDL cholesterol), as well as a deeply phenotyped random sample. Due to the extreme sampling of phenotypic extremes, we adjusted for sampling design to minimize bias.

Exome Sequencing and Variant Calling

In ESP, the processes of library construction, exome capture, sequencing, and mapping were performed as previously described[26; 37; 46]. Sequencing was performed at the University of Washington (UW) and the Broad Institute of MIT/Harvard (Broad). Briefly, exome capture was performed using Roche Nimblegen SeqCap EZ or Agilent SureSelect Human All Exon 50 Mb. Paired end sequencing (2 x 76bp) was performed on Illumina GAII and HiSeq instruments. Single Nucleotide Variants (SNVs) were called using a maximum likelihood approach[47] implemented in the UMAKE pipeline at University of Michigan, which allowed all samples to be analyzed simultaneously, both for variant calling and filtering. Binary Alignment/Map (BAM)[48] files summarizing Burrows-Wheeler Alignment (BWA)[47] alignments generated at the UW and the Broad were used as input. These BAM files summarized alignments mapped to the Genome Reference Consortium Human Build 37 (GRCh37), refined by duplicate removal, recalibration, and indel re-alignment using the Genome Analysis ToolKit (GATK)[49]. We excluded all reads that were not confidently mapped (Phred-scaled mapping quality < 20) from

further analysis. Mean depth was 127x in targeted regions. We then computed genotype likelihoods for exome targeted regions and 50 flanking bases, accounting for per base alignment quality using SAMtools[48]. Variable sites and their allele frequencies were identified using a maximum-likelihood model, implemented in glfMultiples[50]. These analyses assumed a uniform prior probability of polymorphism at each site. The final call-set was performed on 6,823 samples (referred to as the ESP6800 call-set).

In CHARGE, DNA samples were constructed into Illumina paired-end pre-capture libraries according to the manufacturer's protocol. The complete protocol and oligonucleotide sequences are accessible from the Human Genome Sequencing Center (HGSC) website. Either four or six pre-capture libraries were pooled together and then hybridized to Nimblegen exome capture array (HGSC VCRome 2.1 design[51] (42Mb, NimbleGen)) and sequenced in paired-end mode in a single lane on the Illumina HiSeq 2000 platform. Illumina sequence analysis was performed using the HGSC Mercury analysis pipeline. Pooled samples were de-multiplexed using the Consensus assessment of sequence and variation (CASAVA) software. Reads were then mapped to the GRCh37 human reference sequence using BWA producing BAM files. Aligned reads were then recalibrated using GATK[49] along with BAM sorting, duplicate read marking, and realignment near indels. The Atlas2[52] suite was used to call variants and produce high-quality variant call files (VCF)[53]. The VCF includes a filter indicating variants with apparent strand-bias, low allele fraction, low coverage, or low quality to produce a high-quality variant list. Specifically, the poor quality fields included variants with a posterior probability less than 0.95, less than 3 variant reads, variant read ratio less than 0.1, more than 99% variant reads in a single strand direction, total coverage less than 6, and homozygous reference alleles with less than 6X coverage.

Quality Control

ESP used a support vector machine classifier to separate likely true positive and false-positive variant sites, applying a series of variant-level filtering steps. Variant-level quality metrics included allelic balance (the proportional representation of each allele in likely heterozygotes), base quality distribution for sites supporting the reference and alternate alleles, and the distribution of supporting evidence between strands and sequencing cycle, among others. These steps were followed by quality control on individual samples within each study. We used as the positive training set variants identified by dbSNP[54] or 1000 Genomes[55] and we used variants that failed multiple filters as the negative training set. We found this method to be effective at removing sequencing artifacts while preserving good-quality data, as indicated by the transition-transversion (ti-tv) ratio for previously known and newly identified variant sites, the proportion of high frequency variants overlapping with dbSNP, and the ratio of synonymous to non-synonymous variants, as well as attempts at validation of a subset of sites. We excluded variants with read depth greater than 500, variants with more than 2 observed alleles in CHARGE and any genotype containing a copy of the less frequent alternate allele in ESP, or missing rates ≥ 10% for ESP and >20% for CHARGE, and HWE p-value < $5x10^{-8}$ and $P<5x10^{-6}$ for ESP and CHARGE, respectively.

In ESP, samples with discrepant self-reported race and ancestry derived from principal component analysis (PCA) performed on exome sequencing data in PLINK[56] as well as well as ancestry outliers by PCA were removed. Samples having very low concordance (<90%) with previously-obtained SNP array data were considered likely sample swaps and were also dropped from further analysis. In CHARGE, within each cohort, a sample was excluded if it fell beyond

6 standard deviations of any of 4 selected measures that were calculated by cohort and ancestry group: number of singletons, heterozygote to homozygote ratio, mean depth, or ti-tv ratio.

Variant Annotation

To facilitate meta-analysis between CHARGE and ESP, we created a combined variant annotation file including all quality-controlled variant sites observed in either study: 2,707,999 variants in CHARGE and 1,908,614 in ESP6800. We first annotated variants in the two studies separately using an in-house pipeline built on ANNOVAR[57] and dbNSFP v2.0[58] according to the reference genome GRCh37 and National Center for Biotechnology Information RefSeq. The majority of the exonic variants were annotated to a unique gene and functional category. A variant, however, can be annotated to multiple genes by ANNOVAR and can have more than one of the following functional categories: stop-gain, stop-loss, splicing, nonsynonymous, noncoding RNA splicing, synonymous, exonic, 5' untranslated region (UTR5), 3' untranslated region (UTR3), noncoding RNA exonic, upstream, intronic, noncoding RNA intronic, downstream, intergenic, where the first five categories are considered "functional" variants to be included in the rare variants burden tests. We then merged the CHARGE and ESP annotated variant lists to ensure that a variant that was present in both studies has the same reference allele and functional annotation. The combined CHARGE-ESP SNP info file that was used in the skatMeta package included a total of 3,494,971 unique autosomal sites present in either or both ESP and CHARGE.

Association Analyses

Samples with very high CRP values (>100 mg/L) were excluded from analyses and measured CRP values below the lower limit of detection were replaced with the assay lower limit value, leaving 3,109 AA and 6,050 EA available for association testing. CRP values were

natural log (ln) transformed to normalize the distribution of CRP levels. We performed two types of tests, single variant (common and lower frequency variants) and gene burden (rare and lower frequency variants only), as detailed further below and summarized in Table S3. Cohort-level analyses were carried out using the R[59] skatMeta package. In the CHARGE cohorts (ARIC, FHS and CHS) the "skatCohort" or "skatFamcohort" functions were used to create datasets for meta-analysis. In ESP, the six cohorts were pooled into a single dataset and were analyzed using the "skatCohort" function.

Meta-analyses of the "skatCohort" and "skatFamcohort" results were conducted at two independent sites to ensure concordance in findings. Both race-stratified meta-analysis considering a single ethnicity, and combined meta-analysis including both ethnicities were carried out because most CRP-associated loci identified to date have shown consistent patterns of association between EA and AA[26]. All meta-analyses were conducted in the skatMeta package using the "singlesnpMeta" function for single variant meta-analyses, and the "burdenMeta" and "skatMeta" functions for gene-level meta-analyses. We considered only variants on autosomal chromosomes in all analyses. Confirmatory analyses in METAL[60] evaluate between-study heterogeneity of significant results using the heterogeneity I^2 metric. Based on adjustment for 23 tests of significant CRP-associated SNV in the discovery sample, heterogeneity was considered to be statistically significant a p-value below 2.17×10^{-3}.

Single Variant Tests

Using the skatMeta we ran race-stratified study-specific objects (ESP, CHARGE-ARIC, CHARGE-FHS, CHARGE-CHS) for downstream meta-analyses. Within each meta-analysis group (EA, AA, combined) and for each variant site with 5 or more minor alleles detected, we tested for association with CRP levels via linear regression with an additive genetic model. We

included as covariates, race-specific principle components (PCs) as needed, age, sex, and BMI. In the ESP samples a dummy variable correcting for the sampling procedure, cohort and capture target was included in the model.

Gene Burden Tests

Using the skatMeta package we ran two different types of gene-level tests. The first was adopted from the T1 collapsing approach[61]. For the T1 tests, we only considered polymorphic variants having a within race minor allele frequency ≤0.01 that was calculated from the entire ESP-CHARGE call-set. The second gene-level test was adopted from the SKAT approach [62]. For the SKAT tests, we only considered variants having a MAF ≤ 0.05 that was calculated from the entire ESP-CHARGE call-set. Both T1 and SKAT tests considered only SNVs annotated as nonsynonymous, stop-gain, stop-loss, noncoding RNA splice or splice variants in the shared ESP/CHARGE annotation file. Burden tests considered only considered variants that passed quality control in either or both ESP and CHARGE. All gene-level tests were adjusted for age, sex, BMI, and principal components (as needed). Again, the ESP analyses included a dummy variable correcting for the sampling procedure, cohort and capture target.

Significance Thresholds

For analysis of rare variants (MAF<0.05) in the 25 known CRP-associated genes, we evaluated 267 coding variants and thus, we used a Bonferroni-corrected threshold of $P<1.87 \times 10^{-4}$ to declare significance at the single-variant level. In the burden test of candidate genes, we corrected for the number of candidate genes surveyed (N=25) to assign the threshold of significance at $P<2.0 \times 10^{-3}$. In the exome-wide exploratory analyses, an association was deemed to be statistically significant at $P\leq5.0 \times 10^{-8}$ for single variants (the standard GWAS common variant criteria for assessing significance based on a million test), which is a stringent threshold

considering that we only tested ~640,000 variant sites. Significance was assessed at $P<2.5 \times 10^{-6}$ for the gene-based rare variant association tests based on a correction for 19,230 genes. Table S3 summarizes the significance thresholds for all the reported tests.

Validation of CRP Association Signals in Additional Samples

Significant association findings from the discovery sample were followed-up in a sample 13,794 participants of the WHI with CRP measurement available (11,414 EA participants and a sub-sample of 2,380 AA participants from the WHI-SHARe project[26]) and 2,201 AA from JHS. Replication samples were independent from samples used in the ESP and CHARGE exome sequencing discovery sample. The JHS and WHI validation samples were genotyped using the Illumina HumanExome v1.0 BeadChip at Broad Institute or at the Translational Genomics Research Institute (Phoenix, AZ), respectively. Genotype calls were assigned using GenomeStudio v2010.3. We removed samples with call-rates less than 98%, SNPs with call-rates less than 95% or HWE p-values less than $5x10^{-6}$. We checked concordance of genotype calls across hundreds of duplicated samples and SNPs with concordance rates < 99% were excluded from analysis. High-sensitivity CRP was measured on with the use of a latex-particle enhanced immunoturbidimetric assay. Consistent with the discovery analysis, only samples with CRP values < 100 mg/L were included in the replication analyses. Single variant and gene burden tests for association tests were performed as described above.

Investigation of the association between rs77832441 and inflammation-related phenotypes

Samples from the WHI EA replication sample with available Illumina Human exome Array data were used to investigate the association between rs77832441 and inflammation-mediated phenotypes stroke, systolic blood pressure, waist-to-hip ratio, and BMI. To investigate the association to CHD, we used EA samples from the Myocardial Infarction Genetics Exome

Array Consortium with available Illumina HumanExome BeadChip data (N CHD cases=14,727, N controls=30,232). Availability of EA samples from 6,474 type 2 diabetes cases and 6,370 controls from the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples Consortium with available exome sequence data permitted us to evaluate whether rs77832441-A allele was associated with case status. Linear or logistic regression was used to assess association between the predictor (rs77832441- per A-allele) and the continuous or binary inflammation-mediated phenotype with adjustment for covariates as needed. Statistical significance was assessed at $P<0.05$.

Results

Participant Characteristics

Race-stratified characteristics of discovery and validation cohorts are summarized in Table S4. Overall, compared to EA, AA had a greater proportion of women, higher prevalence of hypertension and type 2 diabetes, higher body mass index (BMI), and higher median CRP levels.

Single Variant Test Results

Summary information for uniquely annotated variants included in the ESP-CHARGE meta-analyses is detailed in Table S5. The meta-analyzed exome-wide single variant association results for CRP levels in EA, AA, and combined races are summarized in Figure 1 and Figure S1A. Overall results for significant association signals were consistent with no significant between-study heterogeneity (Tables 1 and 2).

Confirmation that common coding variants are associated with CRP levels

Nine coding variants from four distinct chromosomal regions reached exome-wide significance in the combined sample (Table 1). All nine of these significant variants are common

(minor allele frequency (MAF)>10% in both EA and AA) and have been previously reported or are in linkage disequilibrium (LD) with known CRP-associated single nucleotide polymorphisms (SNP)[2; 19; 25; 26; 63-65]. The genes and their variants are hepatocyte nuclear factor 1 homeobox A (*HNF1A* (MIM 142410); rs2464196 (p.Ser487Asn), rs2259820 (p.Leu459=), rs1169288 (p.Ile27Leu) and rs1169289 (p.Leu17=), 12q24.2), interleukin 6 receptor (*IL6R* (MIM 147880)*;* rs2228145 (p.Asp358Ala), 1q21.31), leptin receptor (*LEPR* (MIM 601007)*;* rs1805096 (p.Pro1019=), 1p31.3), translocase of outer mitochondrial membrane 40 (*TOMM40* (MIM 608061)*;* rs157581 (p.Phe113=), rs11556505 (p.Phe131=), 19q13.32), and apolipoprotein E (*APOE* (MIM 107741)*;* rs429358 (p.Cys130Arg), 19q13.32). Consistent with the results of prior GWAS meta-analyses, common intronic variants in *CRP, LEPR, HNF1A* and *TOMM40* were also associated with CRP levels at an exome-wide significance level (Table S6). Table S7 summarizes effect estimates for all variants reaching significance by analytic group (ESP, CHARGE-ARIC, CHARGE-CHS or CHARGE-FHS).

A novel, rare missense variant of CRP is associated with CRP levels

Two lower frequency synonymous variants (*TOMM40* rs112849259 (p.Asp209=), MAF=3.1% and *CRP* rs1800947 (p.Leu184=), MAF=4.4%) reached exome-wide significance in the combined sample of EA and AA participants (Table 2). The *CRP* rs1800947 synonymous variant has been previously associated with lower CRP levels in EA, independently of the more common CRP-lowering haplotype at the chromosome 1q23 *CRP* locus[12; 20; 66]. *TOMM40* rs112849259 has not been previously been reported as associated with CRP levels. Since rs112849259 is not present on the Exome Array, we were unable to assess its association with CRP in our genotype-based validation sample.

In addition to the *CRP* rs1800947 synonymous variant, we identified a novel association of a rare nonsynonymous *CRP* variant rs77832441 (Thr59Met; MAF =0.16%) with lower CRP levels in the combined discovery sample (A-allele, β (SE)=-0.88(0.19); P = 2.90×10^{-6}) that was significant at the candidate gene-level (P<1.87×10^{-4}). We subsequently validated this finding among 11,414 independent EA Exome Array samples, where the A-allele (MAF=0.31%) was strongly associated with CRP levels (60% lower mean CRP, P=3.00×10^{-15}). The combined discovery and replication p-value was 3.86×10^{-16}.

Additional characterization of CRP p.Thr59Met rare variant association signal

To demonstrate that the *CRP* missense variant rs77832441 p.Thr59Met association signal is independent of other common and lower frequency *CRP* locus SNPs, we performed conditional regression analyses using either the ESP discovery exome sequenced samples or the genotype-based Exome Array validation samples. As shown in Table S8, the association of rs77832441 p.Thr59Met with CRP levels was independent of *CRP* variants known to be associated with CRP through prior GWAS (rs1417938, rs3093059 & rs1800947). Further, assessment of linkage disequilibrium (LD) in the 1000 Genomes European ancestry panel structure showed very weak LD between rs77832441 and common *CRP* variants (CEU; Table S9).

The rs77832441-A allele encoding the Threonine to Methionine amino acid substitution could reduce CRP levels by affecting mRNA splicing or stability, reducing CRP synthesis, or altering CRP monomer subunit structure or the ability for CRP monomers to associate into native circulating pentamers. The p.Thr59Met missense variant is located at residue 41 of the mature CRP polypeptide. According to the 3-dimensional x-ray crystallographic structure of CRP[67], this amino acid lies at the end of a β-sheet (residues 32-41) proximate to a 3/10 alpha helical domain

(residues 43-45). The 40-42 region of the CRP protein is involved in interprotomer interactions with residues 115-123 on the adjoining monomer (Figure 2). Therefore, the non-conservative Threonine to Methionine amino acid change (which is predicted to be functional by in silico protein conservation algorithms[68; 69]) has the potential to reduce CRP pentamerization or reduce native pentameric CRP stability.

Alternatively, since CRP levels were measured by immunoassay, the p.Thr59Met missense variant may alter an epitope recognized by the monoclonal antibody used for CRP capture or detection, leading to artificially low CRP values. We investigated the possibility that the p.Thr59Met amino acid substitution interferes with CRP detection by examining CRP measurements among a subset of 3,442 CARDIA participants for whom both polyclonal and monoclonal antibody assays were performed at study year 15. We hypothesized that the polyclonal assay would be less susceptible to artifactual bias due to the amino acid substitution at residue 59. Comparison of the results failed to demonstrate a large difference in CRP effect size between monoclonal and polyclonal assays in a subset of the discovery sample (Figure S2). CRP levels detected by monoclonal antibody and polyclonal antibody assays were highly correlated in this sample for p.Thr59Met carriers (r=0.97) and non-carriers (r=0.94). These results suggest that p.Thr59Met may confer a true biological reduction in circulating CRP levels, though further studies are warranted to fully characterize the functional significance of this variant.

To follow up on this association, we sought to test whether other inflammation-related phenotypes were associated with the CRP-lowering A-allele of rs77832441 p.Thr59Met. In data from 14,727 Coronary Heart Disease (CHD) cases and 30,232 controls from the Myocardial Infarction Genetics Exome Array Consortium, we failed to demonstrate an association between rs77832441 and decreased risk of CHD (Odds Ratio (95% Confidence Interval (CI))=0.99(0.65-

1.51); P=0.96) despite having an estimated 80% power to detect an odds ratio of 0.67 or less. Further, using data from the EA replication sample we were unable to demonstrate any association between the *CRP* rs77832441 variant and incident ischemic stroke (N_{cases}=2,378, $N_{controls}$=16,736; Odds Ratio (95% CI)= 1.28 (0.76-2.14); P=0.35), systolic blood pressure (N=19,111; β (SE)=-0.48(1.98) mm Hg; P=0.81), waist-to-hip ratio (N=19,111; β (SE)=-0.0014 (0.0071); P=0.84) and BMI (N=19,111; β (SE)=0.044(0.54) kg/m^2; P=0.93). Lastly, test for association of rs77832441 with type 2 diabetes in 6,474 cases and 6,370 controls from EA participants included in the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples Consortium failed to identify an association of the T59M variant on type 2 diabetes risk (Odds Ratio (95 % CI)=0.72 (0.22-2.37); P=0.59).

Additional analysis of common coding variants at the chromosome 19q13

Two common synonymous variants in *TOMM40* (rs157581-C allele, MAF=30%; and rs11556505-T allele, MAF=13%), were associated with ~12% lower mean CRP level ($P \leq 8.62 \times 10^{-9}$) in the combined EA and AA single variant meta-analysis. On the basis of previous studies[63; 70], we hypothesized that these *TOMM40* common variants and/or the low frequency (rs112849259) variant may be in LD with the functional *APOE* ε2, ε3, and ε4 alleles. We therefore assessed the relationship of the common and low frequency *TOMM40* variants to the *APOE* ε4-allele defining variant rs429358 p.Cys130Arg by performing conditional analysis using the ESP exome sequencing data. Conditional analysis in EA and AA demonstrated that all 3 *TOMM40* synonymous CRP-lowering variants (rs157581, rs11556505, rs112849259) are conditionally dependent on the *APOE* ε4 allele-defining SNP rs429358 (Table 3).

The APOE ε2-allele defining variant rs7412 p.Arg176Cys variant failed quality control due to poor sequencing coverage in both ESP and CHARGE discovery samples; however rs7412

genotype was available in 13,794 individuals from our validation sample genotyped on the Exome Array. Association analyses using the Exome Array samples suggested that the rs7412 minor allele (T) was associated with higher CRP levels in 2,379 AA (β (SE)= 0.17 (0.05), $P=6.79x10^{-4}$) from WHI; however there was no evidence of association of rs7412 with CRP in 11,404 EA (β (SE)=0. 018 (0.024), P=0.46). We subsequently validated the rs7412 association in an independent sample of 2,201 AA from JHS with Exome Array genotype data (β (SE)=0.20 (0.057); $p=4.44x10^{-4}$). The combined meta-analysis results for association of rs7412 in Exome Array validation sample with CRP (N=4,609) was β (SE)=0.18 (0.037); $P=1.09x10^{-6}$. Using allele dosage for the *APOE* ε4 variant rs429358 imputed from 1000 Genomes in WHI (imputation Rsq=0.64), we demonstrated in the combined AA Exome Array validation sample (N= 4,211) that the association of *APOE* ε2-defining variant rs7412 with higher CRP levels ($P= 3.74x10^{-5}$) and the association of *APOE* ε4-allele defining variant rs429358 with lower CRP levels ($P= 4.78x10^{-9}$) were conditionally independent.

Gene Burden Test Results

Results for collapsing approach (T1) and sequence kernel association test (SKAT) gene-based tests are summarized in Figures S1B & S1C, respectively. None of the genes reached exome-wide significance ($P<2.5 x10^{-6}$) in EA or AA alone or in the race-combined meta-analysis.

Among the 25 candidate genes from GWAS (significance threshold $P < 2.00 x 10^{-3}$), the most significant finding was that rare variation in the *CRP* locus was associated with CRP levels in EA with both the T1 test ($P=6.80x10^{-4}$) and SKAT ($P=1.71x10^{-4}$) (Tables S10 & S11). Results for both tests of EA on the Exome Array robustly replicated the *CRP* gene-based association ($P_{SKAT}=3.21x10^{-15;}$ $P_{T1}=2.54x10^{-14}$). Removal of the rare *CRP* variant rs77832441 (Thr59Met)

from the EA T1 and SKAT tests eliminated significant signal for the gene-based test (P_{SKAT}=0.96; P_{T1}=0.77), suggesting that the *CRP* gene-based association signal is driven solely by rs77832441. Gene-based results for CRP in AA did not reach statistical significance (P_{SKAT}=0.06; P_{T1}=1.05x10^{-3}).

Gene-based testing with SKAT additionally showed that rare putatively deleterious variation in the Retinoic Acid Receptor-related orphan receptor α (*RORA* (MIM 600825)) locus was associated with CRP levels in AA (P =1.73 x10^{-3}), but not in EA (P=0.83). Using HumanExome BeadChip genotype data from an independent African-American validation sample (p=0.06), the gene-based test did reach statistical significance. However, only one *RORA* rare variant was shared between exome sequencing discovery and HumanExome BeadChip validation platform, which may explain our inability to replicate the finding.

Discussion

By combining exome sequence data from two large consortia, we have identified and validated a novel association between lower CRP levels and the rs7732441 *CRP* missense variant. Though strongly associated with CRP levels, rs7732441 was not associated with other inflammation-mediated phenotypes including CHD, stroke, systolic blood pressure, waist-to-hip ratio and BMI. In addition to the previously described CRP-lowering association signal attributable to the canonical APOE ε4 allele, we demonstrate a second, independent association signal from the APOE ε2 allele that is specific to AA. Finally, a targeted gene-based analysis of known CRP-associated genes suggested possible associations between rare coding variants in *RORA* (in AA) on CRP levels.

Despite a comparatively large sample size relative to other exome sequencing studies, no novel CRP-associated genes were detected using a burden test of rare variants. One interpretation

of these results is that rare coding variants have negligible impact on CRP levels. Alternatively, much larger sample sizes may be required to detect associations of rare coding variants of large to moderate effect sizes with complex traits. For example, based on our discovery sample size of ~6,000 EAs, we had 80% power to detect a variant with ~2 mg/L effect on CRP at 3% MAF but considerably less power to detect the same effect size at MAF ≤1% (Figure S3). Previous studies have shown that tens of thousands of samples may be necessary to be sufficiently powered to detect associations between rare-variants, aggregated at the gene-level, and complex traits[71; 72].

Additional factors, such as heterogeneity of study sampling design (inclusive of both extreme sampling for other phenotypes and randomly selected individuals) or exome sequencing platform, and minor differences in variant calling and quality control procedures between ESP and CHARGE consortia, may have limited our ability to detect novel rare variants associated with CRP levels and/or limit generalizability of our findings.

CRP

Previous studies have identified and replicated associations between CRP levels and variants in the *CRP* locus in populations of European[17; 18; 20], African[26; 27], Asian [17; 23; 24; 63], and Hispanic ancestry[26]. These GWAS have identified three independently associated common alleles (rs33116653, rs12093699 & rs1205) and a single low frequency synonymous variant (rs1800947) [11] in the *CRP* gene associated with lower CRP levels in EA (Table S12). GWAS and candidate gene studies[12; 66] have identified an additional AA-specific allele (rs16827466) associated with higher CRP levels.

Extending these previous findings, we identify a novel rare *CRP* variant (rs77832441) that is both strongly associated with CRP level in single variant tests and drives the gene-based results for *CRP*. On the basis of conditional analyses (Table S8) and LD estimates (Table S9),

this variant appears to represent an independent signal from known GWAS *CRP* loci. Despite the magnitude of its effect, rs77832441 only contributed 0.5% of the overall ln(CRP) phenotypic variance. Our results extend the existence of allelic heterogeneity at the *CRP* locus and suggest that rare coding variation in the *CRP* gene contributes to CRP phenotypic variation.

Functional prediction algorithms taking into account conservation and protein structure, SIFT[69] and Polyphen-2[68], predict the p.Thr59Met variant to be deleterious (SIFT=0.03 (damaging) and Polyphen-2=1.0 (probably damaging)). Similarly, crystal structures of CRP indicate that the variant could disrupt the site involved in monomer subunit interactions[67], which may reduce CRP pentamer formation or stability. In vitro, the monomeric and native pentameric isoforms of CRP exhibit distinct physico-chemical and inflammatory properties[73; 74], which have potential implications for the role of CRP in athero-thrombotic disorders.

Despite the large, likely direct effect of rs77832441 on mean CRP levels, investigation of the relationship of the identified rare variant to other inflammation-mediated phenotypes failed to demonstrate an association. Similar to previous studies[11-17; 75-77], these findings provide additional evidence that CRP levels may not elicit direct effects on CHD, stroke, systolic blood pressure, hip-to-waist ratio, type 2 diabetes and BMI.

TOMM40-APOE

19q13 is a gene-dense region including *TOMM40, APOE* and apolipoprotein C-I (*APOC1;* MIM 107710). Variants in all three of these genes have been previously associated with CRP levels through GWAS or candidate gene analyses[17; 26; 63-65; 70]. In our exome sequencing analysis, the *APOE* ε4-tagging missense variant (rs429358) and three synonymous *TOMM40* variants were associated with CRP levels (rs157581, rs11556505, and rs112849259). Evidence from previous study and 1000 Genomes sequencing suggest that rs157581 and

rs11556505 are in moderate to strong LD with rs429358, the variant that defines the *APOE* ε4 allele[55; 78]. Similarly, we show in our dataset that the CRP-associated *TOMM40* variants (rs157581, rs11556505 and rs112849259) do not represent an independent signal from the *APOE* rs429358 variant. Further, we identify and replicate an additional AA-specific rs429358-independent novel association between *APOE* ε2- tagging variant rs7412 and higher CRP. APOE is a major constituent of lipoproteins and the *APOE* ε2 and ε4 alleles are important genetic determinants of CVD and Alzheimer's disease[79; 80]. The p.Arg158Cys missense variant encoded by ε2 exhibits reduced affinity for the low density lipoprotein receptor (LDLR) and reduced clearance of apoE-containing triglyceride-rich lipoprotein particles (such as very low density lipoprotein), and is associated with lower low density lipoprotein (LDL) cholesterol levels[65] and type III hyperlipoproteinemia[80]. By contrast, the p.Cys112Arg variant encoded by ε4 exhibits more rapid clearance of apoE-containing triglyceride-rich lipoproteins due to increased LDLR affinity and is associated with increased LDL cholesterol levels and increased cardiovascular risk[65; 79; 81; 82]. In addition to its role in cholesterol transport, APOE has anti-atherogenic and anti-inflammatory effects in experimental systems[80; 83]. However, the mechanisms for these pleiotropic effects as well as the paradoxical association of ε2 with lower LDL-c/higher CRP and ε4 with higher LDL-c/lower CRP are not well understood.

RORA

RORA encodes a nuclear receptor with strong homology to the retinoic acid receptor. Data from knockout *RORA* mice models suggest the importance of *RORA* in regulating immune and inflammatory responses, atherosclerosis susceptibility, and ischemia-induced angiogenesis[84; 85]. Previous GWAS indicate that common variation in *RORA* is associated with CRP[18] and liver enzyme levels[86; 87] at a genome-wide significance level in EA. The previous GWAS meta-

analysis of CRP identified a common variant rs340029-T allele as associated with increased CRP levels[18], however GWAS in AA and Hispanic Americans found no association between rs340029 and CRP levels[26]. Further characterization of rare variation of this locus through sequencing efforts is warranted to follow up on our discovery finding of a burden of rare variants at the locus.

Summary

Overall, our results suggest that variants distributed across the allele-frequency spectrum within biological candidate genes identified by GWAS contribute to CRP levels. As suggested by other studies, robustly associating rare coding variants with modest effects on complex traits will require extremely large sample sizes[26; 71]. Collaborative efforts involving meta-analysis of exome sequence or exome BeadChip genotype data will be necessary to amass the large sample sizes required to identify additional rare coding variants contributing to the phenotypic variance of CRP levels.

Acknowledgements

The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO).

Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). Data for "Building on GWAS for NHLBI-diseases: the U.S.

CHARGE consortium" was provided by Eric Boerwinkle on behalf of the Atherosclerosis Risk in Communities (ARIC) Study, L. Adrienne Cupples, principal investigator for the Framingham Heart Study, and Bruce Psaty, principal investigator for the Cardiovascular Health Study. Sequencing was carried out at the Baylor Genome Center (U54 HG003273).

The ARIC Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN2682011000010C, HHSN2682011000011C, and HHSN2682011000012C), R01HL087641, R01HL59367 and R01HL086694. The authors thank the staff and participants of the ARIC study for their important contributions. The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University (Contract No. N01-HC-25195), and its contract with Affymetrix, Inc., for genome-wide genotyping services (Contract No. N02-HL-6-4278), for quality control by Framingham Heart Study investigators using genotypes in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGA) computing resources at Boston University Medical Campus. This CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01 HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086 and grants HL080295, HL087652, HL105756 from the National Heart, Lung, and Blood Institute (NHLBI) with additional contribution from National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through AG023629 from the National Institutes on Aging (NIA). A full list of CHS principal investigators and institutions can be found at CHS-NHLBI.org.

Supported in part by grant R25CA094880 from the National Cancer Institute and by R01HL071862 from NHLBI.

The Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) project was supported by NIH grant U01DK085526.

References

1. Dehghan, A., Kardys, I., de Maat, M.P., Uitterlinden, A.G., Sijbrands, E.J., Bootsma, A.H., Stijnen, T., Hofman, A., Schram, M.T., and Witteman, J.C. (2007). Genetic variation, C-reactive protein levels, and incidence of diabetes. Diabetes 56, 872-878.
2. Ridker, P.M., Buring, J.E., Cook, N.R., and Rifai, N. (2003). C-reactive protein, the metabolic syndrome, and risk of incident cardiovascular events: an 8-year follow-up of 14 719 initially healthy American women. Circulation 107, 391-397.
3. Sesso, H.D., Buring, J.E., Rifai, N., Blake, G.J., Gaziano, J.M., and Ridker, P.M. (2003). C-reactive protein and the risk of developing hypertension. JAMA : the journal of the American Medical Association 290, 2945-2951.
4. Erlinger, T.P., Platz, E.A., Rifai, N., and Helzlsouer, K.J. (2004). C-reactive protein and the risk of incident colorectal cancer. JAMA : the journal of the American Medical Association 291, 585-590.
5. Kaptoge, S., Di Angelantonio, E., Lowe, G., Pepys, M.B., Thompson, S.G., Collins, R., and Danesh, J. (2010). C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet 375, 132-140.
6. Ridker, P.M., Hennekens, C.H., Buring, J.E., and Rifai, N. (2000). C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. The New England journal of medicine 342, 836-843.
7. Ridker, P.M., Rifai, N., Rose, L., Buring, J.E., and Cook, N.R. (2002). Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. The New England journal of medicine 347, 1557-1565.
8. Tracy, R.P., Lemaitre, R.N., Psaty, B.M., Ives, D.G., Evans, R.W., Cushman, M., Meilahn, E.N., and Kuller, L.H. (1997). Relationship of C-reactive protein to risk of cardiovascular disease in the elderly. Results from the Cardiovascular Health Study and the Rural Health Promotion Project. Arteriosclerosis, thrombosis, and vascular biology 17, 1121-1127.
9. Miller, M., Zhan, M., and Havas, S. (2005). High attributable risk of elevated C-reactive protein level to conventional coronary heart disease risk factors: the Third National Health and Nutrition Examination Survey. Archives of internal medicine 165, 2063-2068.
10. Greenland, P., Alpert, J.S., Beller, G.A., Benjamin, E.J., Budoff, M.J., Fayad, Z.A., Foster, E., Hlatky, M.A., Hodgson, J.M., Kushner, F.G., et al. (2010). 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 122, e584-636.
11. Kardys, I., de Maat, M.P., Uitterlinden, A.G., Hofman, A., and Witteman, J.C. (2006). C-reactive protein gene haplotypes and risk of coronary heart disease: the Rotterdam Study. European heart journal 27, 1331-1337.
12. Lange, L.A., Carlson, C.S., Hindorff, L.A., Lange, E.M., Walston, J., Durda, J.P., Cushman, M., Bis, J.C., Zeng, D., Lin, D., et al. (2006). Association of polymorphisms in the CRP gene with circulating C-reactive protein levels and cardiovascular events. JAMA : the journal of the American Medical Association 296, 2703-2711.
13. Casas, J.P., Shah, T., Cooper, J., Hawe, E., McMahon, A.D., Gaffney, D., Packard, C.J., O'Reilly, D.S., Juhan-Vague, I., Yudkin, J.S., et al. (2006). Insight into the nature of the CRP-coronary event association using Mendelian randomization. International journal of epidemiology 35, 922-931.

14. Pai, J.K., Mukamal, K.J., Rexrode, K.M., and Rimm, E.B. (2008). C-reactive protein (CRP) gene polymorphisms, CRP levels, and risk of incident coronary heart disease in two nested case-control studies. PloS one 3, e1395.
15. Lawlor, D.A., Harbord, R.M., Timpson, N.J., Lowe, G.D., Rumley, A., Gaunt, T.R., Baker, I., Yarnell, J.W., Kivimaki, M., Kumari, M., et al. (2008). The association of C-reactive protein and CRP genotype with coronary heart disease: findings from five studies with 4,610 cases amongst 18,637 participants. PloS one 3, e3011.
16. Zacho, J., Tybjaerg-Hansen, A., Jensen, J.S., Grande, P., Sillesen, H., and Nordestgaard, B.G. (2008). Genetically elevated C-reactive protein and ischemic vascular disease. The New England journal of medicine 359, 1897-1908.
17. Elliott, P., Chambers, J.C., Zhang, W., Clarke, R., Hopewell, J.C., Peden, J.F., Erdmann, J., Braund, P., Engert, J.C., Bennett, D., et al. (2009). Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA : the journal of the American Medical Association 302, 37-48.
18. Dehghan, A., Dupuis, J., Barbalic, M., Bis, J.C., Eiriksdottir, G., Lu, C., Pellikka, N., Wallaschofski, H., Kettunen, J., Henneman, P., et al. (2011). Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731-738.
19. Reiner, A.P., Barber, M.J., Guan, Y., Ridker, P.M., Lange, L.A., Chasman, D.I., Walston, J.D., Cooper, G.M., Jenny, N.S., Rieder, M.J., et al. (2008). Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein. American journal of human genetics 82, 1193-1201.
20. Ridker, P.M., Pare, G., Parker, A., Zee, R.Y., Danik, J.S., Buring, J.E., Kwiatkowski, D., Cook, N.R., Miletich, J.P., and Chasman, D.I. (2008). Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. American journal of human genetics 82, 1185-1192.
21. Sabatti, C., Service, S.K., Hartikainen, A.L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C.G., Zaitlen, N.A., Varilo, T., Kaakinen, M., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics 41, 35-46.
22. Benjamin, E.J., Dupuis, J., Larson, M.G., Lunetta, K.L., Booth, S.L., Govindaraju, D.R., Kathiresan, S., Keaney, J.F., Jr., Keyes, M.J., Lin, J.P., et al. (2007). Genome-wide association with select biomarker traits in the Framingham Heart Study. BMC medical genetics 8 Suppl 1, S11.
23. 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.
24. 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.
25. Kong, M., and Lee, C. (2013). Genetic associations with C-reactive protein level and white blood cell count in the KARE study. International journal of immunogenetics 40, 120-125.

26. 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. American journal of human genetics 91, 502-512.
27. 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.
28. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 106, 19096-19101.
29. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42, 30-35.
30. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272-276.
31. Bilguvar, K., Ozturk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoglu, D., Tuysuz, B., Caglayan, A.O., Gokben, S., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467, 207-210.
32. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics 42, 790-793.
33. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 237-241.
34. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics 43, 585-589.
35. Guerreiro, R.J., Lohmann, E., Kinsella, E., Bras, J.M., Luu, N., Gurunlian, N., Dursun, B., Bilgic, B., Santana, I., Hanagasi, H., et al. (2012). Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. Neurobiology of aging 33, 1008 e1017-1023.
36. Albrechtsen, A., Grarup, N., Li, Y., Sparso, T., Tian, G., Cao, H., Jiang, T., Kim, S.Y., Korneliussen, T., Li, Q., et al. (2013). Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. Diabetologia 56, 298-310.
37. Lange, L.A., Hu, Y., Zhang, H., Xue, C., Schmidt, E.M., Tang, Z.Z., Bizon, C., Lange, E.M., Smith, J.D., Turner, E.H., et al. (2014). Whole-Exome Sequencing Identifies Rare and Low-Frequency Coding Variants Associated with LDL Cholesterol. American journal of human genetics 94, 233-245.

38. Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. American journal of human genetics 94, 223-232.
39. Psaty, B.M., and Sitlani, C. (2013). The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium as a model of collaborative science. Epidemiology 24, 346-348.
40. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Jr., Liu, K., and Savage, P.J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. Journal of clinical epidemiology 41, 1105-1116.
41. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al. (1991). The Cardiovascular Health Study: design and rationale. Annals of epidemiology 1, 263-276.
42. Dawber, T.R., Meadors, G.F., and Moore, F.E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. American journal of public health and the nation's health 41, 279-281.
43. Sempos, C.T., Bild, D.E., and Manolio, T.A. (1999). Overview of the Jackson Heart Study: a study of cardiovascular diseases in African American men and women. The American journal of the medical sciences 317, 142-146.
44. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. American journal of epidemiology 156, 871-881.
45. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Controlled clinical trials 19, 61-109.
46. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.
47. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
48. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
49. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.
50. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research 18, 1851-1858.
51. Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L., and Gibbs, R.A. (2011). Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome biology 12, R68.
52. Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC bioinformatics 13, 8.

53. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.
54. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311.
55. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
56. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81, 559-575.
57. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164.
58. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Human mutation 34, E2393-2402.
59. R Development Core Team. (2011). R: A language and environment for statistical computing. In. (
60. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-2191.
61. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics 83, 311-321.
62. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American journal of human genetics 91, 224-237.
63. Curocichin, G., Wu, Y., McDade, T.W., Kuzawa, C.W., Borja, J.B., Qin, L., Lange, E.M., Adair, L.S., Lange, L.A., and Mohlke, K.L. (2011). Single-nucleotide polymorphisms at five loci are associated with C-reactive protein levels in a cohort of Filipino young adults. Journal of human genetics 56, 823-827.
64. Middelberg, R.P., Ferreira, M.A., Henders, A.K., Heath, A.C., Madden, P.A., Montgomery, G.W., Martin, N.G., and Whitfield, J.B. (2011). Genetic variants in LPL, OASL and TOMM40/APOE-C1-C2-C4 genes are associated with multiple cardiovascular-related traits. BMC medical genetics 12, 123.
65. Chasman, D.I., Kozlowski, P., Zee, R.Y., Kwiatkowski, D.J., and Ridker, P.M. (2006). Qualitative and quantitative effects of APOE genetic variation on plasma C-reactive protein, LDL-cholesterol, and apoE protein. Genes and immunity 7, 211-219.
66. Carlson, C.S., Aldred, S.F., Lee, P.K., Tracy, R.P., Schwartz, S.M., Rieder, M., Liu, K., Williams, O.D., Iribarren, C., Lewis, E.C., et al. (2005). Polymorphisms within the C-reactive protein (CRP) promoter region are associated with plasma CRP levels. American journal of human genetics 77, 64-77.

67. Shrive, A.K., Cheetham, G.M., Holden, D., Myles, D.A., Turnell, W.G., Volanakis, J.E., Pepys, M.B., Bloomer, A.C., and Greenhough, T.J. (1996). Three dimensional structure of human C-reactive protein. Nat Struct Biol 3, 346-354.
68. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics / editorial board, Jonathan L Haines [et al] Chapter 7, Unit7 20.
69. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4, 1073-1081.
70. Judson, R., Brain, C., Dain, B., Windemuth, A., Ruano, G., and Reed, C. (2004). New and confirmatory evidence of an association between APOE genotype and baseline C-reactive protein in dyslipidemic individuals. Atherosclerosis 177, 345-351.
71. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nature genetics 44, 623-630.
72. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proceedings of the National Academy of Sciences of the United States of America 106, 3871-3876.
73. Boncler, M., and Watala, C. (2009). Regulation of cell function by isoforms of C-reactive protein: a comparative analysis. Acta biochimica Polonica 56, 17-31.
74. Eisenhardt, S.U., Thiele, J.R., Bannasch, H., Stark, G.B., and Peter, K. (2009). C-reactive protein: how conformational changes influence inflammatory properties. Cell Cycle 8, 3885-3892.
75. Davey Smith, G., Lawlor, D.A., Harbord, R., Timpson, N., Rumley, A., Lowe, G.D., Day, I.N., and Ebrahim, S. (2005). Association of C-reactive protein with blood pressure and hypertension: life course confounding and mendelian randomization tests of causality. Arteriosclerosis, thrombosis, and vascular biology 25, 1051-1056.
76. Timpson, N.J., Lawlor, D.A., Harbord, R.M., Gaunt, T.R., Day, I.N., Palmer, L.J., Hattersley, A.T., Ebrahim, S., Lowe, G.D., Rumley, A., et al. (2005). C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. Lancet 366, 1954-1959.
77. Timpson, N.J., Nordestgaard, B.G., Harbord, R.M., Zacho, J., Frayling, T.M., Tybjaerg-Hansen, A., and Smith, G.D. (2011). C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. Int J Obes (Lond) 35, 300-308.
78. Yu, C.E., Seltman, H., Peskind, E.R., Galloway, N., Zhou, P.X., Rosenthal, E., Wijsman, E.M., Tsuang, D.W., Devlin, B., and Schellenberg, G.D. (2007). Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. Genomics 89, 655-665.
79. Bennet, A.M., Di Angelantonio, E., Ye, Z., Wensley, F., Dahlin, A., Ahlbom, A., Keavney, B., Collins, R., Wiman, B., de Faire, U., et al. (2007). Association of apolipoprotein E genotypes with lipid levels and coronary risk. JAMA : the journal of the American Medical Association 298, 1300-1311.
80. Mahley, R.W., and Rall, S.C., Jr. (2000). Apolipoprotein E: far more than a lipid transport protein. Annual review of genomics and human genetics 1, 507-537.

81. Rasmussen-Torvik, L.J., Pacheco, J.A., Wilke, R.A., Thompson, W.K., Ritchie, M.D., Kho, A.N., Muthalagu, A., Hayes, M.G., Armstrong, L.L., Scheftner, D.A., et al. (2012). High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. Clinical and translational science 5, 394-399.
82. Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS genetics 7, e1002198.
83. Curtiss, L.K. (2000). ApoE in atherosclerosis : a protein with multiple hats. Arteriosclerosis, thrombosis, and vascular biology 20, 1852-1853.
84. Besnard, S., Silvestre, J.S., Duriez, M., Bakouche, J., Lemaigre-Dubreuil, Y., Mariani, J., Levy, B.I., and Tedgui, A. (2001). Increased ischemia-induced angiogenesis in the staggerer mouse, a mutant of the nuclear receptor Roralpha. Circulation research 89, 1209-1215.
85. Chauvet, C., Vanhoutteghem, A., Duhem, C., Saint-Auret, G., Bois-Joyeux, B., Djian, P., Staels, B., and Danan, J.L. (2011). Control of gene expression by the retinoic acid-related orphan receptor alpha in HepG2 human hepatoma cells. PloS one 6, e22545.
86. Middelberg, R.P., Benyamin, B., de Moor, M.H., Warrington, N.M., Gordon, S., Henders, A.K., Medland, S.E., Nyholt, D.R., de Geus, E.J., Hottenga, J.J., et al. (2012). Loci affecting gamma-glutamyl transferase in adults and adolescents show age x SNP interaction and cardiometabolic disease associations. Hum Mol Genet 21, 446-455.
87. Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E., et al. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nature genetics 43, 1131-1138.

Figures & Tables

Figure 1. Manhattan plots of $-\log_{10}$(P-values) from single variant analyses. a) combined race, b)European American, c) African American. Variants in the 25 candidate loci identified through CRP GWAS are highlighted in red, candidate loci with variant reaching exome-wide significance are labeled in italics and rare variants rs77832441 and rs112849259 are also labeled.

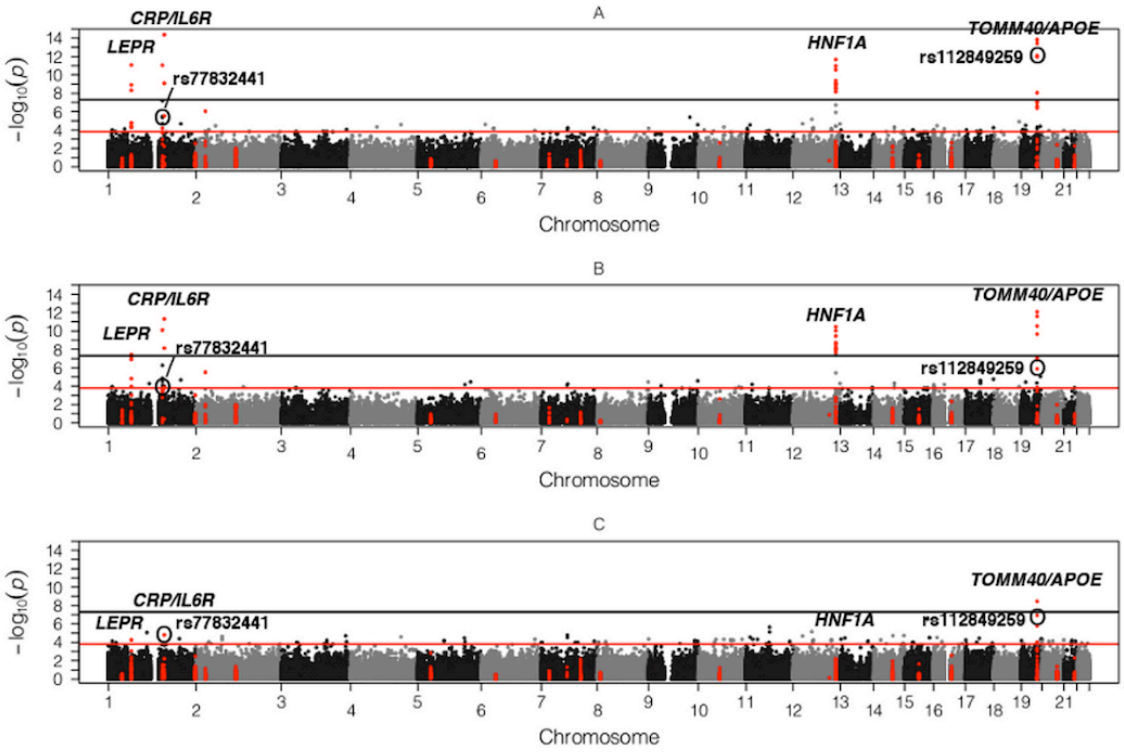


Figure 2. Protein structure model of C-reactive protein generated using Swiss PDB Viewer with the location of rs77832441 is indicated in yellow (Thr41, which corresponds the Thr59 in protein databank accession 1B09) and blue residues indicate regions involved with inter-monomer interactions.

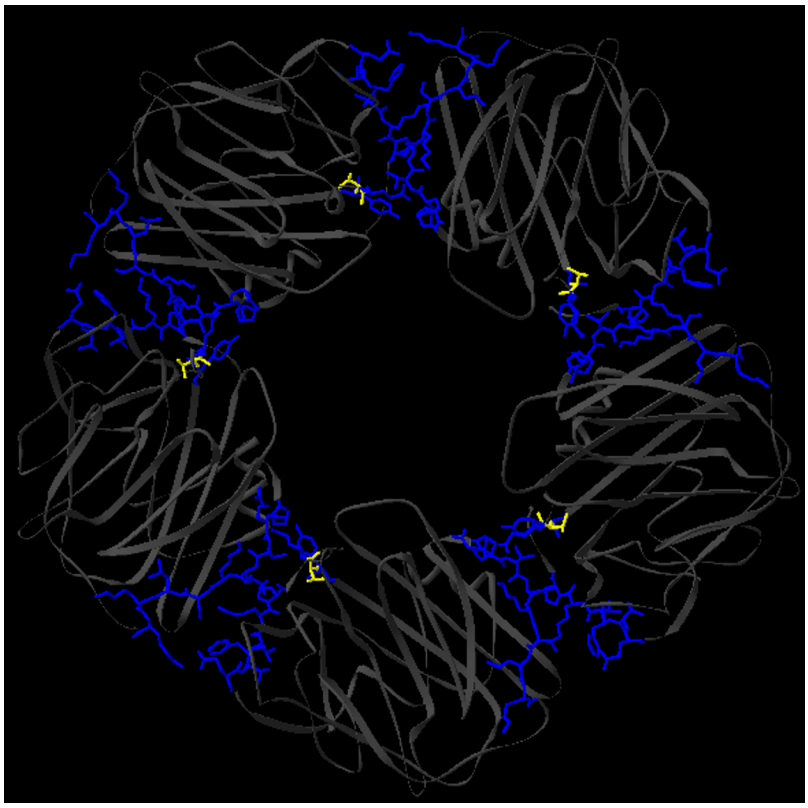


Table 1. Exome-wide significant single variant associations with CRP ($P<5.00x10^{-08}$) for common coding variants

Gene (chromosome: GRCh37 coordinate)	rs# (function; coded/non-coded allele)	Previous Report (GWAS variant if different than reported, LD measures)	Sample (N)	MAF	β (SE)	P	Het. I^2 (P)
LEPR (1:66102257)	rs1805096 (Syn; G/A)	Reiner et al.[26]	EA (6050)	0.37	-0.11 (0.019)	$3.56x10^{-08}$	70.4 (0.012)
			AA (3109)	0.46	-0.11 (0.027)	$5.37x10^{-05}$	0 (1)
			Combined (9159)	0.40	-0.11 (0.016)	$8.34x10^{-12}$	56.7 (0.04)
IL6R (1:154426970)	rs2228145 (nSyn; C/A)	Curocichin et al.[62]	EA (6050)	0.40	-0.12 (0.019)	$7.81x10^{-11}$	0 (0.93)
			AA (3109)	0.14	-0.089 (0.040)	$2.62x10^{-02}$	0 (0.38)
			Combined (9159)	0.31	-0.12 (0.017)	$8.81x10^{-12}$	0 (0.86)
HNF1A (12:121435342)	rs2259820 (Syn; T/C)	Reiner et al.[19] (rs2464196, r^2=1, D'=1)*	EA (6050)	0.31	-0.12 (0.020)	$1.83x10^{-09}$	0 (0.69)
			AA (3109)	0.12	-0.067 (0.040)	$9.62x10^{-02}$	0 (0.96)
			Combined (9159)	0.25	-0.11 (0.018)	$9.06x10^{-10}$	0 (0.72)
HNF1A (12:121435427)	rs2464196 (nSyn;A/G)	Reiner et al.[19]	EA (6050)	0.32	-0.12 (0.020)	$9.29x10^{-09}$	0 (0.71)
			AA (3109)	0.12	-0.07 (0.040)	$8.07x10^{-02}$	0 (0.80)
			Combined (9159)	0.25	-0.11 (0.018)	$3.27x10^{-09}$	0 (0.78)
HNF1A (12:121416622)	rs1169289 (Syn; G/C)	Kong et al.[25] (rs2393791, r^2=0.83, D'=0.93)*	EA (6050)	0.46	-0.12 (0.019)	$8.98x10^{-11}$	8.5 (0.34)
			AA (3109)	0.34	-0.06 (0.028)	$2.16x10^{-02}$	0 (1)
			Combined (9159)	0.42	-0.10 (0.016)	$2.64x10^{-11}$	20.9 (0.28)
HNF1A (12:121416650)	rs1169288 (nSyn; C/A)	Curocichin et al.[62]	EA (6050)	0.34	-0.11 (0.020)	$9.52x10^{-09}$	0 (0.73)
			AA (3109)	0.12	-0.08 (0.041)	$4.04x10^{-02}$	0 (1)
			Combined (9159)	0.26	-0.11 (0.018)	$1.36x10^{-09}$	0 (0.90)
TOMM40 (19:45395714)	rs157581 (Syn; C/T)	Middleburg et al.[63] (rs2075650, r^2=0.58, D'=1)*	EA (6050)	0.21	-0.16 (0.023)	$2.42x10^{-12}$	0 (0.68)
			AA (3109)	0.47	-0.09 (0.027)	$5.07x10^{-04}$	0 (0.88)
			Combined (9159)	0.30	-0.13 (0.018)	$3.73x10^{-14}$	7.9 (0.37)
TOMM40 (19:45396144)	rs11556505 (Syn; T/C)	Middleburg et al.[63] (rs2075650, r^2=1, D'=1)*	EA (6050)	0.14	-0.18 (0.027)	$2.86x10^{-11}$	0 (0.84)
			AA (3109)	0.12	-0.02 (0.040)	$6.33x10^{-01}$	0 (0.43)
			Combined (9159)	0.13	-0.13 (0.023)	$8.62x10^{-09}$	61 (0.025)
APOE (19:45411941)	rs429358 (nSyn; C/T)	Chasman et al.[64]	EA (1832)**	0.11	-0.31 (0.058)	$7.03x10^{-08}$	0 (1)
			AA (1528)**	0.19	-0.24 (0.049)	$1.52x10^{-06}$	0 (1)
			Combined (3360)**	0.14	-0.27 (0.038)	$8.05x10^{-13}$	0 (0.82)

Abbreviations: nonsynonymous (nSyn); synonymous (syn); minor allele frequency (MAF); Heterogeneity (Het. I^2); European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); Single Nucleotide Polymorphism (SNP); Genome Reference Consortium Human Build 37 (GRCh37); reference SNP ID number (rs#)

*D' and r^2 values are based on Broad SNAP proxy search using CEU 1000 Genomes Pilot 1 data.

**The reduced sample size for rs429358 is explained by the fact the variant passed quality control in ESP, but not in CHARGE.

Table 2. Single variant associations of rare and low frequency coding variants with CRP levels

Gene (GRCh37 coordinate)	rs# (function; coded/non-coded allele)	Previously Report	Sample (N)	MAF	β (SE)	P	Het. I^2 (P)
CRP (1:159683438)	rs1800947 (Syn; G/C)	Ridker et al.[2]	EA (6050)	0.0607	-0.27 (0.039)	$4.82x10^{-12}$	0 (0.56)
			AA (3109)	0.0101	-0.58 (0.134)	$1.54x10^{-05}$	0 (0.77)
			Combined (9159)	0.0435	-0.30 (0.038)	$4.22\ x10^{-15}$	28.6 (0.22)
CRP (1:159683814)	rs77832441 (nSyn; A/G)	Not Reported	EA (6050)	0.0022	-0.75 (0.197)	$1.39x10^{-04}$	0 (0.44)
			AA (3109)	0.0005	-2.06 (0.605)	$6.65x10^{-04}$	0 (0.82)
			Combined (9159)	0.0016	-0.88 (0.187)	$2.90x10^{-06}$	28.4 (0.22)
			Replication EA (11414)	0.0031	-0.90 (0.11)	$3.00x10^{-15}$	--
			Discovery + Replication (14573)	0.0034	--	$3.86x10^{-16}$	--
TOMM40 (19:45397307)	rs112849259 (Syn; C/T)	Not Reported	EA (6050)	0.0266	-0.28 (0.058)	$1.28x10^{-06}$	14.1 (0.32)
			AA (3109)	0.0391	-0.37 (0.069)	$1.13x10^{-07}$	0 (0.39)
			Combined (9159)	0.0308	-0.32 (0.044)	$1.06x10^{-12}$	16.2 (0.31)

Abbreviations: nonsynonymous (nSyn); synonymous (syn); minor allele frequency (MAF); Heterogeneity (Het. I^2); European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); Single Nucleotide Polymorphism (SNP); Genome Reference Consortium Human Build 37 (GRCh37); reference SNP ID number (rs#)

Table 3. Conditional Analysis of *TOMM40* synonymous variants on the APOE ε4-defining missense variant rs429358 in the ESP discovery sample

				Without adjustment for rs429358		With adjustment for rs429358	
rs# (function)	MAF (EA/AA/Combined)	Sample (N)	1000 Genomes Pairwise r^2 (rs112849259, rs157581, rs11556505)	β (SE)	P	β (SE)	P
rs112849259 (Synonymous)	0.026/0.039/0.031	EA (1832)	CEU (1, 0.07, 0.005)	-0.40 (0.11)	$1.29x10^{-06}$	-017 (0.12)	0.13
		AA (1528)	YRI (rs112849259 NA in YRI)	-0.30 (0.10)	$2.22x10^{-03}$	-0.15 (0.11)	0.15
rs157581 (Synonymous)	0.21/0.47/0.30	EA (1832)	CEU (0.07, 1, 0.58)	-0.16 (0.044)	$2.15x10^{-04}$	-0.030 (0.058)	0.60
		AA (1528)	YRI (NA, 1, 0.11)	-0.089 (0.038)	$1.99x10^{-02}$	-0.025 (0.043)	0.56
rs11556505 (Synonymous)	0.14/0.12/0.13	EA (1832)	CEU (0.005, 0.583, 1)	-0.18 (0.052)	$6.77x10^{-04}$	-0.014 (0.068)	0.84
		AA (1528)	YRI (NA, 0.11, 1)	-0.051 (0.056)	0.37	-0.032 (0.057)	0.57

Abbreviations: minor allele frequency (MAF); Heterogeneity (Het. I^2); European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); reference SNP ID number (rs#)

Part A, Aim II: Identifying Pleiotropy in Blood Cell Traits

Manuscript

Title

Identifying Pleiotropy in Blood Cell Traits in >20,000 Samples from the Women's Health Initiative

Abstract

Previous Genome-Wide Association Studies (GWAS) have provided evidence for overlapping genetic architecture for various complex traits, including blood cell traits; however detailed examination of the phenomenon is lacking. To investigate pleiotropy in blood cell traits, we used complementary multivariate and univariate frameworks to identify cross-phenotype associations in 21,663 European American women from the Women's Health Initiative with available phenotype and Illumina HumanExome Array genotype information. The primary aim of this study was to identify cross-phenotype associations for red cell traits (hematocrit, hemoglobin), platelet count and white blood cell count. We identified 65 genome-wide significant variants representing 11 loci that were associated with multiple blood cell lineages (red cell, white cell, platelet). Of these 11 loci, 4 signals corresponded to known cross-phenotype associations from GWAS and 7 signals were novel for one or more of the traits. Confirming previous associations and providing novel insights, these findings support the presence of genetic variants with pleiotropic effects on blood cell phenotypes.

Introduction

Many commonly studied traits are correlated with other related or seemingly unrelated traits. One possible explanation for this observation is pleiotropy. Pleiotropy describes the phenomenon by which a single gene influences multiple traits. Assessments of the extent of pleiotropy have revealed that this phenomenon is widespread[1; 2]. One such study suggests that 17% of genes and 4.6% of single nucleotide polymorphisms (SNPs) from published Genome Wide Association Studies (GWAS) have pleiotropic effects[1]. These estimates are likely to greatly underestimate the true extent of pleiotropy due to limitations of the data source (e.g. limited to published GWAS) and the requirement that variants reach a genome-wide significance level in two or more traits.

The observation of cross-phenotype associations, such as those revealed in GWAS, can be related to 1) biologic pleiotropy (single causal variant affects multiple phenotypes, or different causal variants in a gene are associated with each trait); 2) mediated pleiotropy (a variant is associated with a trait that modifies a second trait); or 3) spurious pleiotropy (causal variants in distinct genes in a region with high LD, or study design artifact related to misclassification or ascertainment bias)[3]. Particularly for GWAS, which frequently identify associations with intergenic polymorphisms, distinguishing the underlying relationship (biological, mediated or spurious) for the observed association is crucial. Of primary interest is to identify biologic pleiotropy, which can have implications for the shared etiology of complex diseases.

The aim of this study is to investigate pleiotropy in blood cell traits. These traits include red cell characteristics (hemoglobin levels, hematocrit levels), white blood cell count and platelet count. Significant direct and inverse correlation between several of these traits have been noted[4; 5]. Additionally, GWAS provide evidence of several cross-phenotype associations, revealing

shared genetic architecture of blood cell phenotypes (Figure 1)[5-7]. For example, chromosomal region 12q24.12 harbors the SH2B adaptor protein 3 (*SH2B3)* gene, which encodes an adaptor protein that is an important regulator of T and B cell development and formation of myeloid cells (white blood cell precursors)[6]. Previous GWAS have identified pleiotropic associations of this gene, reporting genome-wide significant *SH2B3* associations for platelet counts[8; 9], red blood cell traits[10; 11], and white blood cell count[6] as well as coronary artery disease[12; 13], diastolic blood pressure[14], and hypothyroidism[15]. To clarify the observation of correlation of blood cell parameters and to investigate the shared loci associated with multiple traits observed in GWAS[4], we conducted exome-wide analyses to explore the potential contribution of pleiotropy to blood cell phenotypes.

Materials and Methods

Study Participants

The study population consisted of European American (EA) participants from the Women's Health Initiative with available Illumina HumanExome BeadChip data and measured blood cell phenotypes (N=21,663). Research participants included in this study were eligible and consented to participate in genetic research studies.

Phenotype Data and Quality Control

The phenotype variables consisted of hematocrit (HCT), hemoglobin (HGB), platelet count (PLT) and white blood cell count (WBC). Blood samples were collected from participants at study baseline (1993-1998) and blood cell counts were performed using automated counters. Quality control was conducted on the phenotypes to limit phenotypic outliers. Specifically, phenotypic values below the 0.05 percentile or above the 99.5 percentile were replaced with the 0.05 percentile and 99.5 percentile values, respectively. Participants were restricted to those with

complete phenotype (HCT, HGB, PLT, WBC) and covariate (age, principle components, study selection) data.

Genotype Quality Control

We used standard quality control measures to remove poor quality samples from our dataset (e.g. samples with call-rates less than 98%, variants with call-rates less than 95% or Hardy Weinberg Equilibrium p-values less than $5x10^{-6}$, and checks of concordance of genotype calls with GWAS data (>99%)). SNPs with missing values were imputed to the mean value to facilitate use of the Multivariate Outcome Score Test (MOST)[16], which requires that genotype data be complete.

We created two datasets for 1) single variant and 2) gene-based association tests. For the single variant test, variants with a minimum of 5 minor alleles in the population were tested for association with the phenotypes. For gene-based tests, rare non-synonymous and splice variants below 1% or 5% Minor Allele Frequency (MAF) were collapsed across the gene. Similar to the Combined Multivariate and Collapsing Method[17], individuals harboring one or more qualifying variant were assigned a value of 1 for the gene, whereas individuals without a rare non-synonymous or splice variant in the gene were assigned a value of 0.

Software

To identify cross-phenotype associations we used a software program called MOST[16]. This program uses a regression-based framework to relate the marginal distribution of traits to the genetic variables. The method is capable of covariate adjustment and has the advantage that it outputs both univariate statistics for each individual trait and multivariate statistics for multiple traits.

Identification of cross-phenotype associations can be conducted under univariate or

multivariate approaches (summarized here[3]). Univariate approaches test for association between a single phenotype and a genetic variant then combine metrics across traits, whereas multivariate approaches seek to simultaneously test two or more phenotypes for association with a genetic variant. Multivariate analyses require individual-level data containing multiple phenotypes on a single individual, whereas univariate approaches can be applied to existing GWAS data by combining reported summary statistics. Multivariate approaches have the advantage that they have increased power relative to single trait analyses, and they may clarify the shared pathophysiological mechanisms underlying the set of traits.

Additional analyses including dataset generation, tests of mediation and summary of the results from MOST were carried out in the R software package[18].

Statistical Analysis

Two classes of statistical tests were used to evaluate multivariate associations: 1) single variant analyses considering variants with at least 5 minor alleles (MAF>0.014%) and 2) gene-based tests of qualifying variants below 1% and 5% MAF. Both single variant and gene-based test models included the following covariates: age at baseline, variables capturing study selection and the first principle component. For the purposes of these analyses, assessments of cross-phenotype associations were based on the Q-statistic. The Q-statistic tests the null hypothesis that the effect sizes of the genetic variable for all of the phenotypes are zero. This test makes no assumption about the relationship between phenotypes (effect estimates in the same or opposite direction).

Multivariate single variant tests considered 99,148 loci, therefore statistical significance was set at a p-value < 5.04×10^{-7}. The gene-based burden test considered 15,921 genes, thus statistical significance at 3.14×10^{-6}. To be considered a cross-phenotype association, we required

the Q-statistic to be less than the significance threshold and more than one phenotype to have a univariate p-value less than 0.05.

Mediation analyses were conducted to evaluate the relationship of blood cell variables to each other or to evaluate mediation by other non-blood cell phenotypes. To evaluate the relationship between blood cell phenotypes, analyses were conducting using MOST with inclusion of one or more additional blood cell phenotype as a covariate. Attenuation of univariate statistics upon adjustment provided evidence for mediation by the phenotype(s) included as covariate(s). Tests of mediation by non-study phenotypes were also conducted using MOST. In these analyses we included the non-study phenotype(s) as covariate(s) in the model to assess mediation of the observed association by an external phenotype.

For comparison to results from multivariate analyses from MOST, we calculated combined univariate p-values for HCT, PLT and WBC using Fisher's method. Due to the strong correlation of HCT and HGB (see below), we omitted HGB from the meta-analysis of univariate p-values.

Results

The study population consisted of 21,663 EA women aged 50-81 years. Demographic characteristics and the phenotypes of interest of the study population are summarized in Table 1. The four phenotypes showed varying degrees of pairwise correlation (Table 2). HCT and HGB are red blood cell traits that are highly correlated but only weakly correlated with WBC (Pearson's r (r) <0.17) and not correlated with PLT (r~0). WBC and PLT are moderately correlated (r=0.28).

Single Variant Findings

In total, we identified 65 cross phenotype associations (Table 3, Figure 2) that corresponded to 11 chromosomal regions (2q23.3, 4q12, 6q22.2, 6p21.3, 6q23.3, 9q34.2, 10q21.3, 12q24.12, 11q12.2 and 17q21.1). Our analysis both confirmed previously reported cross-phenotype associations from GWAS and identified novel signals in blood cell candidate genes (Table 4). The strongest signal from the multivariate analysis corresponded to 6q23.3, which harbors known pleiotropic loci HBS1-like protein (*HBS1L)* – V-MYB avian myelobastosis viral oncogene homolog (*MYB)*. Confirming previous reports, this locus was associated with HCT, HGB, PLT and WBC. Similarly, we confirmed the associations of 12q24.12 (*SH2B3*- ataxin 2 *(ATXN2))* with HCT, HGB, PLT and WBC, 6p21.3 (human leukocyte antigen (HLA) locus) with PLT and WBC, and 4q12 (platelet-derived growth factor receptor (*PDGRFA)* – v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog *(KIT))* with HCT, HGB and WBC.

We identified novel associations in 7 chromosomal regions (2q23.3, 4q12, 6q22.2, 6p21.3, 9q34.2, 10q21.3, and 11q12.2). In 2q23.3, common variants (rs1260326 (nonsynonymous/splice) and rs780094 (intronic)) in the glucokinase regulator (*GCKR*) gene (2p23.3) were associated with the traits in the multivariate model (Q.pval<4.9610^{-11}) with strong univariate signal for WBC (univariate P-value (Uni.pval)<$1.86x10^{-7}$) and PLT (Uni.pval<$5.39x10^{-7}$), as well as weaker signal for HCT (Uni.pval<$3.41x10^{-3}$) and HGB (Uni.pval<$3x10^{-3}$). *GCKR* is a PLT-associated locus from GWAS[8]; however *GCKR* has not previously been associated with WBC, HCT, and HGB. Analyses to assess mediation in the associated traits failed to demonstrate dependence of the association on other study traits, suggesting that the associations of the *GCKR* SNPs with WBC, RBC traits (HCT and HGB) and PLT are not attributable solely to correlation of the phenotypes (Table 5). Tests for possible

mediation by other lipid (total cholesterol, low density lipoprotein, high density lipoprotein and triglycerides) or metabolic phenotypes (fasting insulin, fasting glucose and body mass index) that have been previously associated with *GCKR* in GWAS did not show large attenuation of multivariate or univariate blood cell associations in *GCKR* (Table 6). In a subset of our sample with available complete blood cell counts (N=3,479), we tested for association of these *GCKR* variants with WBC subtypes (neutrophils, basophils, eosinophils, monocytes, and lymphocytes) and found that the variants were only associated with basophils and neutrophils (Table 7).

GCKR is located in a large linkage disequilibrium block that contains *GPN1,* a candidate gene for PLT. One *GPN1* variant (rs3749147) on the exome array was in moderate linkage disequilibrium with the two *GCKR* SNPs (r^2>0.44, D'=1), but failed quality control for low call rate (78%). To explore the relationship between *GPN1* and *GCKR*, we conducted exploratory association and conditional analyses with the three variants. The *GPN1* rs3749147 was associated with WBC (p=0.0003) and PLT (p=0.006), but conditioning on the *GCKR* SNPs removed the signal, which suggests that it is not an independent signal.

9q34.2 harbors ABO blood group (*ABO)*, a locus previously associated with red blood cell count[11] and red cell parameters (HCT and HGB)[19]. We identified four of the noncoding variants (rs507666 (MAF=17.4%), rs651007 (MAF=19.8%), rs579459 (MAF=22.0%) and rs635634 (MAF=17.4%)) that were in near complete LD (r^2=0.96, D'=1) and one variant (rs657152, MAF= 36.6%) that was in moderate LD with the other four variants (r^2<0.53, D'=1). All five significant variants were in or near to *ABO* (Q.pval<1.72 x10^{-7}), confirming the association with red cell parameters (HCT and HGB) and demonstrating a strong novel association with WBC (Uni.pval<3.37 x10^{-5}). In a subsample of our population with complete blood cell counts, the association between *ABO* and WBC appears to be driven by monocytes

and lymphocytes (Table 7). Additional analyses demonstrated that the association of 9q34.2 SNPs with WBC was not mediated by HCT or HGB (Table 5). In univariate models of HCT, HGB and WBC conditioning on rs651007, the signal for rs657152 did not remain significant, suggesting that the *ABO*-allele tagging variants do not represent independent signals. Lastly, lipid traits have previously been associated with *ABO*[20], however we did not see evidence for mediation of the ABO association by lipid traits (low density lipoprotein, high density lipoprotein, total cholesterol or triglyceride levels; Table 8).

We identified 6 noncoding variants at 11q12.2 (Q.pval<2.66 x10^{-7}) associated with HCT (Uni.pval<2.45 x10^{-4}), HGB (Uni.pval<1.38 x10^{-4}) and PLT (Uni.pval<5.04 x10^{-4}). 11q12.2 contains fatty acid desaturase 1 (*FADS1)* and fatty acid desaturase 2 (*FADS2),* which have previously been associated with PLT but appear to represent novel associations for the red cell traits (HCT and HGB). Mediation analyses suggest that the 11q12.2 association with PLT is not influenced by HCT and HGB and visa versa (Table 5).

In 6p22.2 we identified a novel association in and around the hemochromatosis (*HFE)* locus (Q.pval<2.24 x10^{-13}). This locus has been previously associated with red cell traits, but our results suggest that coding and non-coding variants in the region are also associated with WBC (Uni.pval<0.044). Nearby at 6p21.32 two common intronic variants in notch 4 (*NOTCH4)* were nominally significant in the multivariate analysis (Q.pval<4.90 x10^{-7}), showing association with HCT (Uni.pval<5.65 x10^{-3}), HGB (Uni.pval<7.70 x10^{-3}) and WBC (Uni.pval<1.27 x10^{-3}). Another significant association (Q.pval<6.72 x10^{-11}) was identified in or near jumonji domain containing 1C (*JMJD1C)* (10q21.3)*,* a locus previously associated with PLT[8; 9]. Beyond demonstrating strong univariate signal for PLT, multiple variants were weakly associated with HCT (Uni.pval<0.04) and WBC (Uni.pval<0.02). Lastly, 17q21.2 near gasdermin B *(GSDMB)* –

gasdermin A (*GSDMA)* showed signal in the multivariate analysis with strong univariate signal for WBC (Uni.pval<2.34x $x10^{-9}$; previously reported) and weak signal for HCT (Uni.pval< 0.018) and HGB (Uni.pval<0.04).

Gene-Based Findings

Gene-based tests failed to identify any genome-wide significant associations (Table 9 & 10). Top association signals for 1% MAF threshold test the four blood cell traits were *SH2B3* (Q.pval=$4.35x10^{-5}$) and tubulin polyglutamylase complex subunit 2 (*TPGS2;* Q.pval=$9.20x10^{-5}$). For the 5% threshold test, top signals were *SH2B3* (Q.pval=$4.35x10^{-5}$), chemokine receptor 2 *(CXCR2;* Q.pval=$1.20x10^{-4}$) and erythropoietin *(EPO;* Q.pval=$1.58x10^{-4}$).

Discussion

We replicated cross-phenotype associations for *HBS1L-MYB*, *SH2B3-ATXN2, HLA* and *PDGRFA-KIT*, which were previously identified through GWAS. In addition, we identified seven novel associations for one or more traits that reached genome-wide significance in the multivariate analysis (Figure 4). No gene-based results reached statistical significance, although several candidate loci including *SH2B3* and *EPO* showed association with several traits.

Of particular interest are association signals identified in chromosomal regions 2p23.3 (*GCKR*), 9q34.2 (*ABO*), and 11q12.2 (*FADS1-FADS2*). Two correlated SNPs (rs1260326 (nonsynonymous-near splice) and rs780094 (intronic); r^2=0.93, D'=1.0) in the *GCKR* locus (2p23.3) were strongly associated with PLT and white cells (total WBC, basophils and neutrophils), and to a lesser extent with HCT and HGB. Variant rs1260326 was previously reported to be associated with increased PLT (T-allele) in a large sample of EA[8] and with several other traits including cardiac/lipid[19; 21-23] and metabolic[24; 25] phenotypes. The relationship between *GCKR* variants and blood cell traits does not appear to be mediated by lipid or

metabolic phenotypes available in our sample. *GCKR* is an important regulator of glucokinase (GCK) activity, which plays an important role in controlling carbohydrate metabolism. The rs1260326 is thought to be a functional mutation that inhibits GCK activity[26].

GCKR is located in a long linkage disequilibrium block extending more than 450 kilobases (chr2:27422973-27873713) across more than 20 genes. Due to the genetic structure, the *GCKR* variants identified in this study may tag causal variants in another gene. A strong candidate is GPN-loop GTPase 1 (*GPN1),* located 100 kilobases away from *GCKR* but within the same linkage disequilibrium block. Functional evidence from *Drosophila* demonstrated that gene silencing of *Xab1* (human *GPN1*) led to increased plasmatocyte (similar to mammalian monocyte) and crystal cell counts[8], providing biological evidence for the association with blood cell parameters. In our dataset, the *GPN1* rs3749147 variant was associated with PLT and WBC, but was not an independent signal from *GCKR* polymorphisms. Evidence from the Encyclopedia of DNA Elements suggests that rs3749147 is located within an active promoter and the variant is predicted to disrupt transcription factor binding. Though there is compelling functional evidence for the association between *Xab1* (human *GPN1*) from *Drosophila,* further research is warranted to tease apart the observed cross-phenotype association at 2q23.3 to determine if one or multiple causal genes for blood cell traits are located within the region.

The second association signal composed of five correlated noncoding variants was observed at the *ABO* locus (at 9q34.2). Several previous GWAS have identified associations between *ABO* rs495828 and rs579459 polymorphisms and red cell traits (red blood cell count, mean corpuscular hemoglobin levels, HCT and HGB)[11; 19], which are in strong LD with the identified SNPs. *ABO* is a highly pleiotropic gene, associated with a wide array of phenotypes including lipid and inflammatory traits[20; 27-32], that determines the antigens present on red blood

cells, thereby assigning blood type (A/B/O). The rs657152-G is a marker of the O blood group, whereas rs507666-C (and linked rs651007, rs579459 and rs635634) perfectly tag the ABO A1 blood group[33]. Despite tagging separate ABO alleles, variants tagging the O and A1 alleles do not represent independent signals in our dataset. Previous reports have identified an association between lipid traits and *ABO*, however lipid traits do not appear to mediate the relationship of blood cell traits and ABO variants.

Beyond associations with lipid traits, GWAS have uncovered associations at the *ABO* locus with inflammatory biomarkers, including interleukin-6 (rs643434)[30], intercellular adhesion molecule 1 (rs507666 and rs649129)[28; 31; 32] and tumor necrosis factor alpha (rs505922)[29]. Similar to our observed association between *ABO* variants and lower WBC, these GWAS studies identified correlated variants to those identified in this study with a consistent effect of lower inflammatory biomarker levels[28-32]. Unfortunately due to limited availability of these inflammatory biomarkers in our sample, we were unable to conduct mediation analyses in WHI. Beyond the presence of red blood cell-specific expression *ABO* gene products, the mechanism by which this locus influences blood cell parameters is unknown and it remains possible that this relationship is mediated by inflammation.

Finally, noncoding variants in 11q12.2 (*FADS1-FADS2*) showed cross-phenotype associations for HCT, HGB and PLT. The *FADS1* and *FADS2* genes code for enzymes that mediate the synthesis of omega-3 and omega-6 long chain polyunsaturated fatty acids (PUFA), respectively. Dietary omega-3 fatty acids have been shown to decrease platelet aggregation[34; 35], however the association between *FADS1* and *FADS2* variants or PUFA levels and the phenotypes is less clear. The identified noncoding variants in *FADS2-FADS2* are in linkage disequilibrium (r^2>0.8, D'>0.96) with 3' untranslated region variant rs4246215 in adjacent gene

flap endonuclease 1 (*FEN1).* This variant was previously identified in GWAS as associated with PLT[8]. The rs4246215 G>T mutation is thought to be functional and is associated with decreased *FEN1* mRNA expression[36]. Strong evidence for pleiotropy comes from mouse models of mutant *FEN1* (knockout of *FEN1* interaction site with proliferating cell nuclear antigen (*PCNA)*) led to development of peripheral pancytopenia, characterized by a marked decrease in circulating red blood cells and near absence of white blood cells and platelets[37].

In addition to associations in 2p23.3, 9q34.2 and 11q12.2, we identified suggestive novel associations at 6p22.2, 6p21.32, 10q21.3 and 17q21.2. Region 6q22.2 has been previously associated with red blood cell parameters[7; 38-40] and PLT[8; 41] in GWAS. We report an association between the noncoding variants near *HFE* and WBC. Two candidate genes, *HFE* and Leucine rich repeat containing 16A (*LRRC16A),* are located in close proximity within the region. *HFE* is a well-known regulator of iron absorption, which is robustly associated with red blood cell phenotypes[7; 38-40]. Two of the SNPs (rs1800652 and rs1799945) we identified as cross-phenotype associations for red cell traits and PLT in the region are causally related to hereditary hemochromatosis, an iron storage disorder that leads to accumulation of iron in body tissues and organs. Owing to the known reciprocal relationship of PLT and iron stores, the observed *HFE* association with red blood cell traits (HCT and HGB) and PLT is biologically plausible. Another possibility is that the observed cross-phenotype association may relate to spurious pleiotropy whereby variants in separate (*HFE* and *LRRC16A*) genes have trait-specific effects. A previous GWAS identified an association between *LRRC16A* (rs12526480) and both PLT and mean platelet volume[8; 41]. However, the variants we identified in the region were not correlated with rs12526480 (r^2<0.065, D'=1), suggesting that the *LRRC16A* signal could be distinct from the identified variants in *HFE.*

A second region on chromosome 6 (6q21.3) showed genome-wide signal for two intronic variants in *NOTCH4*, showing association with HCT, HGB and WBC. *NOTCH4* is a novel association for these traits, but due to extensive linkage disequilibrium in the region it is unclear if *NOTCH4* is an independent signal from nearby HLA genes.

In region 10q21.3 one coding (rs1935) variant and five noncoding (rs10761731, rs12355784, rs2393967, rs2893923, rs10761779) variants in or near *JMJD1C* reached genome-wide signal and showed association with PLT, HGB and WBC. *JMJD1C* variants have previously been associated with PLT[8; 39; 41], but no associations with HGB and WBC have been reported. Interestingly, gene silencing of zebrafish *jmjd1c* resulted in ablation of erythropoiesis and thrombocyte formation[8]. This functional evidence may support the observed association of *JMJD1C* with HGB and PLT, however it does not explain the association with WBC.

Lastly, we identified a novel association of correlated coding and noncoding variants (r^2>0.84, D'>0.93) in 17q21.1 with HCT and HGB, and replicated associations of the variants with WBC[5; 19; 42-45]. 17q21.1 contains linkage disequilibrium block that encompasses several genes including *GSDMB, GSDMA, ORMDL3, PSMD3, CSF3,* and *MED24.* The colony stimulating factor 3 (*CSF3)* gene is a strong functional candidate for WBC as this protein is known to promote granulocyte production[46] and is used clinically to treat severe neutropenia. At present there is no functional evidence to suggest that *CSF3* is causally related to red blood cell traits, and it remains possible that the weak association we observed between 17q21.1 and red cell traits relates to causal variation in a distinct gene within the linkage disequilibrium block.

Comparison of Univariate and Multivariate Tests

Multivariate and univariate methodologies can be utilized to identify cross-phenotype associations. Previous reports indicate multivariate approaches are more powerful than univariate

association tests, capable of identifying variants with smaller effects on individual traits. Comparison of univariate and multivariate methods in our dataset indicated good concordance between methods with an overlapping variant set of 51 variants corresponding to 9 of the 11 loci identified in the multivariate analysis. Univariate analyses failed to identify the known *PDGRFA-KIT* and the novel *NOTCH4* associations, but identified additional variants in *ABO* and 6q21 (HLA region), and two novel cross-phenotype associations at 11q13.31 (patatin-like phospholipase domain containing 3 (*PNPLA3*) – sorting and assembly machine component 50 homolog (*SAMM50))* and 7q11.23 (bromodomain adjacent to zinc finger domain 1B (*BAZ1B*)). These results suggest that use of multivariate and univariate statistics in tandem may reveal additional associations.

Limitations

A limitation of our study was the small number of blood cell phenotypes. Our study phenotypes captured red cell traits (HCT and HGB), PLT and WBC, but it is likely that investigation of more blood cell phenotypes including red cell count and white cell subtypes could provide a more comprehensive picture of pleiotropy. Other potential limitations include power/sample size, variant selection and inclusion of only European ancestry samples. Our study was likely underpowered to detect multivariate associations, particularly for rare variants. Though we sought to maximize power through selection of phenotypes available in a relatively large population, it remains possible that we failed to identify cross-phenotype associations due to limited power. Further, owing to the composition of the exome array with primarily rare coding variants and a subset of trait-associated loci from GWAS, we may have had insufficient genomic coverage to identify additional pleiotropic loci. Additionally, our sample was comprised of individuals of European descent, which limits the generalizability of our findings to other

ancestral groups. Confirmation of these findings in diverse populations is warranted, as is replication of these results in independent samples of European descent. Lastly, in the interest of identifying biologic pleiotropy, functional analyses are warranted to rule out spurious and mediated pleiotropy as the cause for the observed cross-phenotype associations.

Summary

In summary, our results confirm known cross-phenotype associations and provide evidence for novel cross-phenotype associations for blood cell traits in EA. Our findings indicate that use of multivariate methods reveal shared genetic architecture of traits and perhaps identify true pleiotropic loci. Replication of these results is warranted to ensure the validity of these findings and additional functional analyses are necessary to better understand the cause for the cross-phenotype associations.

References

1. Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. American journal of human genetics 89, 607-618.
2. Kocarnik, J.M., and Fullerton, S.M. (2014). Returning pleiotropic results from genetic testing to patients and research participants. JAMA : the journal of the American Medical Association 311, 795-796.
3. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nature reviews Genetics 14, 483-495.
4. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Human molecular genetics 22, 2529-2538.
5. Okada, Y., Hirota, T., Kamatani, Y., Takahashi, A., Ohmiya, H., Kumasaka, N., Higasa, K., Yamaguchi-Kabata, Y., Hosono, N., Nalls, M.A., et al. (2011). Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. PLoS genetics 7, e1002067.
6. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dube, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nature genetics.
7. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nature genetics 41, 1191-1198.
8. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature 480, 201-208.
9. Shameer, K., Denny, J.C., Ding, K., Jouni, H., Crosslin, D.R., de Andrade, M., Chute, C.G., Peissig, P., Pacheco, J.A., Li, R., et al. (2014). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Human genetics 133, 95-109.
10. Gudbjartsson, D.F., Bjornsdottir, U.S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G.M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H., et al. (2009). Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nature genetics 41, 342-347.
11. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369-375.
12. Dichgans, M., Malik, R., Konig, I.R., Rosand, J., Clarke, R., Gretarsdottir, S., Thorleifsson, G., Mitchell, B.D., Assimes, T.L., Levi, C., et al. (2014). Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. Stroke 45, 24-36.
13. Schunkert, H., Konig, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics 43, 333-338.

14. International Consortium for Blood Pressure Genome-Wide Association, S., Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478, 103-109.
15. Eriksson, N., Tung, J.Y., Kiefer, A.K., Hinds, D.A., Francke, U., Mountain, J.L., and Do, C.B. (2012). Novel associations for hypothyroidism include known autoimmune risk loci. PloS one 7, e34442.
16. He, Q., Avery, C.L., and Lin, D.Y. (2013). A general framework for association tests with multivariate traits in large-scale genomics studies. Genet Epidemiol 37, 759-767.
17. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics 83, 311-321.
18. R Development Core Team. (2011). R: A language and environment for statistical computing. In. (
19. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nature genetics 42, 210-215.
20. Global Lipids Genetics, C., Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nature genetics 45, 1274-1283.
21. Kathiresan, S., Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., Schadt, E.E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., et al. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nature genetics 41, 56-65.
22. Chasman, D.I., Pare, G., Mora, S., Hopewell, J.C., Peloso, G., Clarke, R., Cupples, L.A., Hamsten, A., Kathiresan, S., Malarstig, A., et al. (2009). Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. PLoS genetics 5, e1000730.
23. 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. American journal of human genetics 91, 502-512.
24. Sabatti, C., Service, S.K., Hartikainen, A.L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C.G., Zaitlen, N.A., Varilo, T., Kaakinen, M., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics 41, 35-46.
25. Saxena, R., Hivert, M.F., Langenberg, C., Tanaka, T., Pankow, J.S., Vollenweider, P., Lyssenko, V., Bouatia-Naji, N., Dupuis, J., Jackson, A.U., et al. (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nature genetics 42, 142-148.
26. Beer, N.L., Tribble, N.D., McCulloch, L.J., Roos, C., Johnson, P.R., Orho-Melander, M., and Gloyn, A.L. (2009). The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. Hum Mol Genet 18, 4081-4088.
27. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research 42, D1001-1006.

28. Barbalic, M., Dupuis, J., Dehghan, A., Bis, J.C., Hoogeveen, R.C., Schnabel, R.B., Nambi, V., Bretler, M., Smith, N.L., Peters, A., et al. (2010). Large-scale genomic studies reveal central role of ABO in sP-selectin and sICAM-1 levels. Hum Mol Genet 19, 1863-1872.
29. Melzer, D., Perry, J.R., Hernandez, D., Corsi, A.M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J.R., Paolisso, G., et al. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS genetics 4, e1000072.
30. Naitza, S., Porcu, E., Steri, M., Taub, D.D., Mulas, A., Xiao, X., Strait, J., Dei, M., Lai, S., Busonero, F., et al. (2012). A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS genetics 8, e1002480.
31. Pare, G., Chasman, D.I., Kellogg, M., Zee, R.Y., Rifai, N., Badola, S., Miletich, J.P., and Ridker, P.M. (2008). Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. PLoS genetics 4, e1000118.
32. Pare, G., Ridker, P.M., Rose, L., Barbalic, M., Dupuis, J., Dehghan, A., Bis, J.C., Benjamin, E.J., Shiffman, D., Parker, A.N., et al. (2011). Genome-wide association analysis of soluble ICAM-1 concentration reveals novel associations at the NFKBIK, PNPLA3, RELA, and SH2B3 loci. PLoS genetics 7, e1001374.
33. Zhang, H., Mooney, C.J., and Reilly, M.P. (2012). ABO Blood Groups and Cardiovascular Diseases. Int J Vasc Med 2012, 641917.
34. Agren, J.J., Vaisanen, S., Hanninen, O., Muller, A.D., and Hornstra, G. (1997). Hemostatic factors and platelet aggregation after a fish-enriched diet or fish oil or docosahexaenoic acid supplementation. Prostaglandins Leukot Essent Fatty Acids 57, 419-421.
35. Mori, T.A., Beilin, L.J., Burke, V., Morris, J., and Ritchie, J. (1997). Interactions between dietary fat, fish, and fish oils and their effects on platelet function in men at risk of cardiovascular disease. Arteriosclerosis, thrombosis, and vascular biology 17, 279-286.
36. Yang, M., Guo, H., Wu, C., He, Y., Yu, D., Zhou, L., Wang, F., Xu, J., Tan, W., Wang, G., et al. (2009). Functional FEN1 polymorphisms are associated with DNA damage levels and lung cancer risk. Human mutation 30, 1320-1328.
37. Zheng, L., Dai, H., Qiu, J., Huang, Q., and Shen, B. (2007). Disruption of the FEN-1/PCNA interaction results in DNA replication defects, pulmonary hypoplasia, pancytopenia, and newborn lethality in mice. Mol Cell Biol 27, 3176-3186.
38. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum Mol Genet 22, 2529-2538.
39. Li, J., Glessner, J.T., Zhang, H., Hou, C., Wei, Z., Bradfield, J.P., Mentch, F.D., Guo, Y., Kim, C., Xia, Q., et al. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Hum Mol Genet 22, 1457-1464.
40. Kullo, I.J., Ding, K., Jouni, H., Smith, C.Y., and Chute, C.G. (2010). A genome-wide association study of red blood cell traits using the electronic medical record. PloS one 5.
41. Qayyum, R., Snively, B.M., Ziv, E., Nalls, M.A., Liu, Y., Tang, W., Yanek, L.R., Lange, L., Evans, M.K., Ganesh, S., et al. (2012). A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. PLoS genetics 8, e1002491.

42. Crosslin, D.R., McDavid, A., Weston, N., Nelson, S.C., Zheng, X., Hart, E., de Andrade, M., Kullo, I.J., McCarty, C.A., Doheny, K.F., et al. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Human genetics 131, 639-652.
43. Nalls, M.A., Couper, D.J., Tanaka, T., van Rooij, F.J., Chen, M.H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang, Q., Greinacher, A., et al. (2011). Multiple loci are associated with white blood cell phenotypes. PLoS genetics 7, e1002113.
44. Soranzo, N., Spector, T.D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature genetics 41, 1182-1190.
45. Kong, M., and Lee, C. (2013). Genetic associations with C-reactive protein level and white blood cell count in the KARE study. International journal of immunogenetics 40, 120-125.
46. Andrews, N.C. (2009). Genes determining blood cell traits. Nature genetics 41, 1161-1162.

Figures & Tables

Table 1. Characteristics of the population

	Women's Health Initiative (WHI)
Number of Individuals with genotype and phenotypes available	21,663
Age Mean (SD) in years	66.16 (6.73)
Hemoglobin Mean (SD) in g/dL	13.63 (1.07)
Hematocrit Mean (SD) in %	40.45 (3.25)
White Blood Cell Count Mean (SD) in (x10^9/L)	6.34 (11.70)
Platelet Count Mean (SD) in (x10^9/L)	246.43 (58.52)

Table 2. Pearson's r correlation between phenotypes

	Hematocrit	**Hemoglobin**	**Platelet**	**White Blood Cell**
Hematocrit	1.00	0.91	0.004	0.17
Hemoglobin	0.91	1.00	-0.04	0.16
Platelet	0.004	-0.04	1.00	0.28
White Blood Cell	0.17	0.16	0.28	1.00

Table 3. Cross-Phenotype Single Variant Association Results

chr.pos	dbSNPID	MAF	Function	Gene(s)	MV P-value	UV P-values (Direction of Effect)			
					Q	HCT	HGB	PLT	WBC
chr2:27730940	rs1260326	0.35	splicing; nonsynonymous	*GCKR*	4.96E-11	3.41E-03 (-)	3.01E-03 (-)	5.39E-07 (+)	1.86E-07 (+)
chr2:27741237	rs780094	0.35	intronic	*GCKR*	6.06E-12	2.78E-03 (-)	2.37E-03 (-)	2.52E-07 (+)	4.25E-08 (+)
chr4:55394172	rs218237	0.16	intergenic	*PDGFRA-KIT*	2.60E-08	3.54E-08 (-)	5.26E-06 (-)	5.78E-01 (+)	2.61E-02 (+)
chr4:55407762	rs172629	0.15	intergenic	*PDGFRA-KIT*	2.19E-08	2.86E-08 (-)	3.96E-06 (-)	5.66E-01 (+)	2.53E-02 (+)
chr6:25842951	rs1408272	0.05	intergenic	*PDGFRA-KIT*	1.52E-14	1.26E-07 (+)	1.01E-13 (+)	2.81E-02 (-)	3.61E-01 (+)
chr6:25997458	rs12216125	0.28	intergenic	*PDGFRA-KIT*	2.24E-13	1.31E-07 (+)	2.78E-13 (+)	9.73E-01 (-)	4.39E-02 (+)
chr6:26091179	rs1799945	0.12	nonsynonymous	*HFE*	4.03E-19	1.37E-07 (+)	2.10E-15 (+)	2.08E-02 (-)	7.47E-03 (+)
chr6:26093141	rs1800562	0.05	nonsynonymous	*HFE*	5.27E-17	7.26E-09 (+)	6.68E-16 (+)	2.54E-02 (-)	4.00E-01 (+)
chr6:26107463	rs198846	0.15	downstream	*HIST1H1T*	2.11E-18	8.83E-07 (+)	3.00E-14 (+)	2.07E-02 (-)	5.74E-03 (+)
chr6:31134888	rs3130931	0.27	intronic	*POU5F1*	2.90E-07	1.14E-01 (+)	2.43E-02 (+)	8.44E-04 (+)	1.21E-02 (-)
chr6:31139452	rs879882	0.38	intergenic	*HLA-B*	2.05E-07	8.08E-01 (+)	2.42E-01 (+)	6.73E-05 (+)	3.42E-02 (-)
chr6:31255541	rs2524050	0.11	intergenic	*HLA-B*	5.27E-10	2.13E-01 (-)	1.02E-01 (-)	1.42E-02 (-)	8.44E-07 (+)
chr6:31256753	rs2524044	0.15	intergenic	*HLA-B*	1.22E-09	1.22E-01 (-)	6.09E-02 (-)	2.13E-02 (-)	1.63E-06 (+)
chr6:31264922	rs2853925	0.11	intergenic	*HLA-B*	5.65E-10	2.12E-01 (-)	1.07E-01 (-)	1.59E-02 (-)	7.22E-07 (+)
chr6:31273224	rs364415	0.11	intergenic	*HLA-B*	4.76E-10	1.87E-01 (-)	8.71E-02 (-)	1.25E-02 (-)	1.09E-06 (+)
chr6:31440082	rs1055569	0.39	ncRNA_exonic	*HCG26*	1.33E-08	1.95E-01 (-)	8.22E-02 (-)	4.44E-04 (-)	5.79E-04 (+)
chr6:31440497	rs2516440	0.39	downstream	*HCG26*	1.41E-08	2.25E-01 (-)	1.04E-01 (-)	4.67E-04 (-)	4.41E-04 (+)
chr6:31441349	rs4413654	0.31	intergenic	*HCG26*	1.14E-07	3.32E-01 (-)	1.24E-01 (-)	3.97E-02 (-)	2.09E-05 (+)
chr6:31773821	rs2242668	0.11	intronic	*LSM2*	2.57E-07	1.73E-01 (-)	1.31E-01 (-)	4.86E-05 (-)	1.85E-02 (+)
chr6:31778529	rs2075799	0.12	synonymous	*HSPA1L*	1.27E-07	1.56E-01 (-)	1.56E-01 (-)	3.56E-05 (-)	1.35E-02 (+)
chr6:32166384	rs8192575	0.06	intronic	*NOTCH4*	3.51E-07	5.05E-03 (-)	6.12E-04 (-)	1.21E-01 (-)	1.11E-03 (+)
chr6:32185605	rs2854050	0.06	intronic	*NOTCH4*	4.90E-07	5.65E-03 (-)	7.70E-04 (-)	1.19E-01 (-)	1.27E-03 (+)
chr6:32219725	rs4959089	0.20	intergenic	*NOTCH4*	6.86E-08	4.06E-01 (-)	4.84E-01 (-)	1.43E-04 (-)	1.01E-03 (+)
chr6:33525680	rs210170	0.34	intergenic	*BAK1*	1.64E-07	9.35E-04 (-)	8.29E-03 (-)	4.98E-07 (-)	3.03E-01 (-)
chr6:135411228	rs9376090	0.20	intergenic	*HBS1L-MYB*	3.65E-41	2.72E-14 (-)	6.50E-09 (-)	1.73E-20 (+)	2.41E-05 (-)

chr6:135418635	rs7775698	0.24	intergenic	*HBS1L-MYB*	3.96E-40	1.76E-14 (-)	2.05E-09 (-)	1.25E-20 (+)	6.66E-05 (-)
chr6:135418916	rs7776054	0.24	intergenic	*HBS1L-MYB*	4.89E-40	1.21E-14 (-)	1.95E-09 (-)	3.40E-20 (+)	5.17E-05 (-)
chr6:135426573	rs4895441	0.22	intergenic	*HBS1L-MYB*	5.30E-41	1.28E-13 (-)	8.76E-09 (-)	1.66E-21 (+)	5.01E-05 (-)
chr6:135432552	rs9494145	0.19	intergenic	*HBS1L-MYB*	9.00E-37	8.66E-10 (-)	2.11E-06 (-)	2.91E-21 (+)	3.52E-05 (-)
chr6:135435501	rs9483788	0.21	intergenic	*HBS1L-MYB*	4.71E-31	4.26E-08 (-)	3.16E-05 (-)	7.19E-18 (+)	5.66E-05 (-)
chr6:135452152	rs6569992	0.18	intergenic	*HBS1L-MYB*	1.38E-26	1.17E-07 (-)	6.21E-05 (-)	3.34E-17 (+)	9.90E-03 (-)
chr6:135525396	rs3819409	0.44	intronic	*MYB*	1.17E-07	8.11E-03 (+)	1.06E-01 (+)	2.22E-06 (-)	6.50E-01 (+)
chr9:136139265	rs657152	0.38	intronic	*ABO*	1.03E-07	1.72E-07 (-)	2.40E-06 (-)	3.73E-01 (-)	3.37E-05 (-)
chr9:136149399	rs507666	0.17	intronic	*ABO*	5.35E-11	3.18E-09 (-)	3.25E-11 (-)	7.58E-01 (-)	3.01E-05 (-)
chr9:136153875	rs651007	0.20	intergenic	*ABO*	3.51E-10	2.65E-08 (-)	5.32E-10 (-)	4.65E-01 (-)	1.17E-05 (-)
chr9:136154168	rs579459	0.20	intergenic	*ABO*	4.14E-10	3.85E-08 (-)	7.79E-10 (-)	5.09E-01 (-)	9.69E-06 (-)
chr9:136155000	rs635634	0.17	intergenic	*ABO*	1.05E-10	5.40E-09 (-)	9.01E-11 (-)	6.17E-01 (-)	2.06E-05 (-)
chr10:64927823	rs1935	0.44	nonsynonymous	*JMJD1C*	2.07E-13	1.56E-01 (-)	4.00E-02 (-)	3.88E-10 (+)	2.17E-03 (-)
chr10:65027610	rs10761731	0.39	intronic	*JMJD1C*	9.79E-18	6.24E-01 (-)	1.83E-01 (-)	3.27E-15 (+)	2.00E-02 (-)
chr10:65121565	rs12355784	0.44	intronic	*JMJD1C*	1.17E-13	1.39E-01 (-)	3.32E-02 (-)	1.81E-10 (+)	2.79E-03 (-)
chr10:65133156	rs2393967	0.25	intronic	*JMJD1C*	2.90E-11	3.30E-01 (-)	3.50E-02 (-)	1.68E-09 (+)	5.79E-02 (-)
chr10:65261184	rs2893923	0.29	intergenic	*JMJD1C*	6.72E-11	2.70E-01 (-)	2.31E-02 (-)	4.63E-09 (+)	6.02E-02 (-)
chr10:65274927	rs10761779	0.43	intergenic	*JMJD1C*	3.90E-13	1.12E-01 (-)	2.77E-02 (-)	2.36E-10 (+)	5.06E-03 (-)
chr11:61557803	rs102275	0.43	intronic	*C11orf10*	2.66E-07	2.45E-04 (+)	1.38E-04 (+)	5.03E-04 (+)	2.98E-01 (-)
chr11:61569830	rs174546	0.28	UTR3	*FADS1*	7.48E-09	7.46E-05 (+)	6.84E-05 (+)	5.92E-05 (+)	2.69E-01 (-)
chr11:61570783	rs174547	0.28	intronic	*FADS1*	5.57E-09	6.05E-05 (+)	5.25E-05 (+)	6.86E-05 (+)	2.44E-01 (-)
chr11:61571478	rs174550	0.28	intronic	*FADS1*	6.66E-09	7.07E-05 (+)	6.53E-05 (+)	6.27E-05 (+)	2.53E-01 (-)
chr11:61597972	rs1535	0.30	intronic	*FADS2*	1.83E-08	8.07E-05 (+)	7.64E-05 (+)	9.52E-05 (+)	2.93E-01 (-)
chr11:61609750	rs174583	0.34	intronic	*FADS2*	1.66E-08	1.39E-04 (+)	9.68E-05 (+)	1.87E-04 (+)	1.61E-01 (-)
chr12:111884608	rs3184504	0.38	nonsynonymous	*SH2B3*	1.76E-30	7.37E-13 (+)	8.82E-17 (+)	5.32E-16 (+)	2.37E-08 (+)
chr12:111910219	rs10774625	0.39	intronic	*ATXN2*	4.37E-29	1.46E-11 (+)	2.27E-15 (+)	4.69E-16 (+)	5.37E-08 (+)
chr12:112007756	rs653178	0.38	intronic	*ATXN2*	3.22E-30	1.36E-12 (+)	1.81E-16 (+)	4.48E-16 (+)	2.17E-08 (+)

chr12:112072424	rs11065987	0.34	intergenic	*ATXN2*	4.96E-25	9.73E-10 (+)	3.17E-14 (+)	4.18E-13 (+)	2.26E-05 (+)
chr12:112211833	rs2238151	0.47	intronic	*ALDH2*	3.99E-07	2.87E-06 (-)	1.29E-07 (-)	3.99E-02 (-)	4.84E-03 (-)
chr12:112486818	rs17696736	0.35	intronic	*NAA25*	2.29E-25	1.23E-09 (+)	3.81E-14 (+)	1.01E-13 (+)	2.51E-04 (+)
chr12:113039943	rs233716	0.44	intergenic	*NAA25*	1.94E-07	7.76E-03 (+)	6.59E-04 (+)	5.33E-06 (+)	3.75E-03 (+)
chr17:37976469	rs9303277	0.48	intronic	*IKZF3*	8.16E-12	9.56E-02 (-)	4.01E-02 (-)	1.42E-01 (-)	2.34E-09 (+)
chr17:38028634	rs11557467	0.46	nonsynonymous	*ZPBP2*	2.28E-14	5.82E-02 (+)	1.47E-02 (+)	1.85E-01 (+)	4.05E-11 (-)
chr17:38040763	rs2872507	0.40	intergenic	*GSDMB*	2.90E-14	1.82E-02 (+)	6.26E-03 (+)	2.57E-01 (+)	4.97E-11 (-)
chr17:38051348	rs8067378	0.50	intergenic	*GSDMB*	8.36E-15	7.71E-02 (+)	1.69E-02 (+)	1.51E-01 (+)	2.50E-11 (-)
chr17:38062217	rs2305479	0.41	nonsynonymous	*GSDMB*	1.15E-15	5.18E-02 (+)	1.06E-02 (+)	1.69E-01 (+)	6.34E-12 (-)
chr17:38064469	rs11078928	0.37	splicing	*GSDMB*	1.22E-15	1.67E-02 (+)	4.48E-03 (+)	2.97E-01 (+)	4.16E-12 (-)
chr17:38066240	rs2290400	NA	intronic	*GSDMB*	1.42E-15	7.78E-02 (+)	1.53E-02 (+)	1.84E-01 (+)	4.86E-12 (-)
chr17:38069949	rs7216389	0.42	intronic	*GSDMB*	1.30E-14	5.31E-02 (+)	1.41E-02 (+)	1.39E-01 (+)	4.05E-11 (-)
chr17:38131187	rs56030650	0.43	nonsynonymous	*GSDMA*	1.46E-22	7.02E-01 (-)	4.80E-02 (-)	6.83E-01 (-)	1.23E-18 (+)
chr17:38156712	rs4794822	0.37	intergenic	*GSDMA*	9.52E-24	3.05E-01 (-)	3.19E-02 (-)	7.79E-01 (-)	1.21E-20 (+)

Table 4. Reported and Novel Regions

Region (gene)	Trait (minimum P-value)	Reported (citation) or Novel
2p23.3 (*GCKR*)	HCT (2.78E-03)	Novel
	HGB (2.37E-03)	Novel
	PLT (2.52E-07)	Reported (Gieger et al.)
	WBC (4.25E-08)	Novel
4q12 (*PDGFRA-KIT*)	HCT (2.86E-08)	Reported (Kamatami et al.)
	HGB (3.96E-06)	Reported
	WBC (0.025)	Reported (Kamatami et al.)
6p22.2 (*HFE*)	HCT (7.26E-09)	Reported (Ganesh et al.)
	HGB (6.68E-16)	Reported (Ganesh et al.)
	PLT (0.021)	Novel
	WBC (5.74E-03)	Novel
6p21.32 (*NOTCH4*)	HCT (5.05E-03)	Novel
	HGB (6.12E-04)	Novel
	WBC (1.11E-03)	Reported (Soranzo et al.)
6p21.3 (*HLA*)	PLT (4.86E-05)	Reported (Gieger et al.)
	WBC (7.22E-07)	Reported (Nalls et al.)
6q23.3 (*HBS1L-MYB*)	HCT (1.21E-14)	Reported (Kamatami et al.)
	HGB (1.95E-09)	Reported (Kamatami et al.)
	PLT (1.66E-21)	Reported (Kamatami et al.)
	WBC (2.41E-05)	Reported (Kamatami et al.)
9q34.2 (*ABO*)	HCT (5.40E-09)	Reported (Kamatami et al.)
	HGB (9.01E-11)	Reported (Kamatami et al.)
	WBC (9.69E-06)	Novel
10q21.3 (*JMJD1C*)	HGB (0.023)	Novel
	PLT (3.27E-15)	Reported (Gieger et al.)
	WBC (2.17E-03)	Novel
11q12.2 (*FADS1-FADS2*)	HCT (6.05E-05)	Novel - reported by Nathan/Santhi
	HGB (5.25E-05)	Novel - reported by Nathan/Santhi
	PLT (5.92E-05)	Reported (Gieger et al.)
12q24.12 (*SH2B3-ATXN2*)	HCT (7.37E-13)	Reported (Ganesh et al.)
	HGB (8.82E-17)	Reported (van der Harst et al.)
	PLT (4.48E-16)	Reported (Shameer et al.)
	WBC (2.17E-08)	Reported (Auer et al.)
17q21.1 (*GSDMB-GSDMA*)	HCT (0.017)	Novel
	HGB (4.48E-03)	Novel
	WBC (1.21E-20)	Reported (Kamatami et al.)

Table 5. Mediation Analyses of Study Phenotypes

Region	SNP	Trait	Adjustments	P.value Initial	P.value Adjusted	Evidence for Mediation?
2p23.3	rs780094	WBC	HCT, HGB	4.25E-08	1.13E-09	No mediation by HCT/HGB
2p23.3	rs1260326	WBC	HCT, HGB	1.86E-07	6.44E-09	No mediation by HCT/HGB
2p23.3	rs780094	WBC	PLT, HCT, HGB	4.25E-08	1.77E-06	Slight attenuation in signal with adjustment for PLT
2p23.3	rs1260326	WBC	PLT, HCT, HGB	1.86E-07	5.68E-06	Slight attenuation in signal with adjustment for PLT
2p23.3	rs780094	HCT	PLT, WBC	2.78E-03	1.38E-04	No evidence of mediation by PLT/WBC
2p23.3	rs1260326	HCT	PLT, WBC	3.41E-03	2.14E-04	No evidence of mediation by PLT/WBC
2p23.3	rs780094	HGB	PLT, WBC	2.37E-03	2.83E-04	No evidence of mediation by PLT/WBC
2p23.3	rs1260326	HGB	PLT, WBC	3.01E-03	4.35E-04	No evidence of mediation by PLT/WBC
9q34.2	rs657152	WBC	HCT, HGB	3.37E-05	9.75E-04	Slight attenuation with adjustment for HCT/HGB
9q34.2	rs507666	WBC	HCT, HGB	3.01E-05	1.48E-03	Slight attenuation with adjustment for HCT/HGB
9q34.2	rs651007	WBC	HCT, HGB	1.17E-05	5.43E-04	Slight attenuation with adjustment for HCT/HGB
9q34.2	rs579459	WBC	HCT, HGB	9.69E-06	4.46E-04	Slight attenuation with adjustment for HCT/HGB
9q34.2	rs635634	WBC	HCT, HGB	2.06E-05	1.03E-03	Slight attenuation with adjustment for HCT/HGB
9q34.2	rs657152	HCT	WBC	1.72E-07	4.53E-06	Slight attenuation with adjustment for WBC
9q34.2	rs507666	HCT	WBC	3.18E-09	1.19E-07	Slight attenuation with adjustment for WBC
9q34.2	rs651007	HCT	WBC	2.65E-08	9.98E-07	Slight attenuation with adjustment for WBC
9q34.2	rs579459	HCT	WBC	3.85E-08	1.45E-06	Slight attenuation with adjustment for WBC
9q34.2	rs635634	HCT	WBC	5.40E-09	2.14E-07	Slight attenuation with adjustment for WBC
9q34.2	rs657152	HGB	WBC	2.40E-06	3.99E-05	Slight attenuation with adjustment for WBC
9q34.2	rs507666	HGB	WBC	3.25E-11	1.39E-09	Slight attenuation with adjustment for WBC
9q34.2	rs651007	HGB	WBC	5.32E-10	2.25E-08	Slight attenuation with adjustment for WBC
9q34.2	rs579459	HGB	WBC	7.79E-10	3.33E-08	Slight attenuation with adjustment for WBC
9q34.2	rs635634	HGB	WBC	9.01E-11	3.98E-09	Slight attenuation with adjustment for WBC
11q12.2	rs102275	PLT	HCT, HGB	5.03E-04	2.99E-04	No evidence of mediation by HCT/HGB
11q12.2	rs174546	PLT	HCT, HGB	5.92E-05	3.61E-05	No evidence of mediation by HCT/HGB
11q12.2	rs174547	PLT	HCT, HGB	6.86E-05	4.12E-05	No evidence of mediation by HCT/HGB
11q12.2	rs174550	PLT	HCT, HGB	6.27E-05	3.83E-05	No evidence of mediation by HCT/HGB
11q12.2	rs1535	PLT	HCT, HGB	9.52E-05	5.97E-05	No evidence of mediation by HCT/HGB
11q12.2	rs174583	PLT	HCT, HGB	1.87E-04	1.13E-04	No evidence of mediation by HCT/HGB
11q12.2	rs102275	HCT	PLT	2.45E-04	2.53E-04	No evidence of mediation by PLT
11q12.2	rs174546	HCT	PLT	7.46E-05	7.77E-05	No evidence of mediation by PLT
11q12.2	rs174547	HCT	PLT	6.05E-05	6.30E-05	No evidence of mediation by PLT
11q12.2	rs174550	HCT	PLT	7.07E-05	7.36E-05	No evidence of mediation by PLT
11q12.2	rs1535	HCT	PLT	8.07E-05	8.40E-05	No evidence of mediation by PLT
11q12.2	rs174583	HCT	PLT	1.39E-04	1.44E-04	No evidence of mediation by PLT
11q12.2	rs102275	HGB	PLT	1.38E-04	7.29E-05	No evidence of mediation by PLT
11q12.2	rs174546	HGB	PLT	6.84E-05	3.18E-05	No evidence of mediation by PLT
11q12.2	rs174547	HGB	PLT	5.25E-05	2.43E-05	No evidence of mediation by PLT
11q12.2	rs174550	HGB	PLT	6.53E-05	3.03E-05	No evidence of mediation by PLT
11q12.2	rs1535	HGB	PLT	7.64E-05	3.65E-05	No evidence of mediation by PLT
11q12.2	rs174583	HGB	PLT	9.68E-05	4.81E-05	No evidence of mediation by PLT

Table 6. Mediation analysis of *GCKR* association by lipid (high density lipoprotein, low density lipoprotein, Triglycerides, total cholesterol) and metabolic traits (fasting insulin, fasting glucose, body mass index) in subsample of 9,918

Trait	**SNP**	**Initial Model**	**Initial +lipid and metabolic traits**
Multivariate	exm181733	1.92E-06	3.63E-05
	exm.rs780094	4.13E-07	4.78E-06
Univariate Hematocrit	exm181733	5.51E-02	2.08E-02
	exm.rs780094	3.72E-02	1.76E-02
Univariate Hemoglobin	exm181733	4.47E-02	8.30E-03
	exm.rs780094	3.05E-02	7.66E-03
Univariate Platelet	exm181733	2.07E-05	3.28E-03
	exm.rs780094	2.26E-05	2.18E-03
Univariate White blood cell	exm181733	1.29E-04	3.11E-04
	exm.rs780094	3.65E-05	4.84E-05

Table 7. Association of *GCKR* and *ABO* variants with WBC subtypes in subset of sample with available complete blood cell counts (N=3,479)

SNP (gene, region)	**Total White Cell beta(se);p**	**Neutrophil beta(se);p**	**Monocyte beta(se);p**	**Eosinophil beta(se);p**	**Basophil beta(se);p**	**Lymphocyte beta(se);p**
rs1260326 (*GCKR* 2p23.3)	0.12 (0.05); 0.014	-0.09 (0.03); 0.007	0.03 (0.04); 0.39	0.09 (0.09); 0.29	0.60 (0.20); 0.0032	0.04 (0.04); 0.40
rs780094 (*GCKR,* 2p23.3)	0.12 (0.05); 0.014	-0.08 (0.03); 0.02	0.03 (0.04); 0.30	0.11 (0.08); 0.16	0.62 (0.20); 0.0021	0.04 (0.04); 0.33
rs657152 (*ABO,* 9q34.2)	-0.08 (0.05); 0.08	-0.05 (0.03); 0.11	-0.05 (0.04); 0.17	-0.04 (0.08); 0.60	0.05 (0.20); 0.80	-0.04 (0.04); 0.30
rs507666 (*ABO,* 9q34.2)	-0.10 (0.04); 0.02	-0.04 (0.03); 0.13	-0.07 (0.03); 0.014	-0.11 (0.07); 0.10	-0.08 (0.16); 0.64	-0.08 (0.04); 0.032
rs651007 (*ABO,* 9q34.2)	-0.11 (0.04); 0.007	-0.05 (0.03); 0.11	-0.07 (0.03); 0.022	-0.09 (0.07); 0.21	0.02 (0.17); 0.92	-0.07 (0.04); 0.047
rs579459 (*ABO,* 9q34.2)	-0.11 (0.04); 0.008	-0.04 (0.03); 0.11	-0.07 (0.03); 0.026	-0.09 (0.07); 0.22	0.02 (0.17); 0.92	-0.07 (0.04); 0.057
rs635634 (*ABO,* 9q34.2)	-0.10 (0.04); 0.01	-0.04 (0.03); 0.13	-0.07 (0.0.3); 0.022	-0.10 (0.07); 0.14	-0.06 (0.16); 0.72	-0.08 (0.04); 0.028

Table 8. Mediation analysis of *ABO* association by lipid traits (N=10965)

Trait	SNP	Initial Model	Initial + LDL	Initial + HDL	Initial + Trig	Initial + Tchol	Initial + LDL + HDL + Trig + Tchol
Multivariate	rs657152	1.86E-04	8.50E-05	1.40E-03	7.99E-04	4.90E-05	1.53E-04
	rs507666	5.54E-04	1.79E-04	2.85E-03	2.58E-03	1.44E-04	4.59E-04
	rs651007	3.81E-04	8.50E-05	1.32E-03	1.25E-03	7.22E-05	3.08E-04
	rs579459	4.85E-04	1.09E-04	1.64E-03	1.52E-03	9.15E-05	4.16E-04
	rs635634	7.33E-04	2.39E-04	3.53E-03	3.06E-03	1.91E-04	6.01E-04
Univariate Hematocrit	rs657152	6.82E-06	2.99E-06	3.99E-05	2.49E-05	1.96E-06	5.99E-06
	rs507666	8.47E-05	3.74E-05	4.67E-04	4.81E-04	3.55E-05	7.71E-05
	rs651007	1.38E-04	5.69E-05	5.83E-04	6.49E-04	5.73E-05	1.30E-04
	rs579459	2.07E-04	8.96E-05	8.83E-04	9.78E-04	8.97E-05	2.05E-04
	rs635634	1.51E-04	7.11E-05	7.87E-04	7.98E-04	6.73E-05	1.39E-04
Univariate. Hemoglobin	rs657152	2.15E-05	9.23E-06	1.24E-04	7.34E-05	5.62E-06	1.93E-05
	rs507666	1.76E-05	6.52E-06	1.11E-04	1.13E-04	5.54E-06	1.60E-05
	rs651007	3.05E-05	1.02E-05	1.41E-04	1.59E-04	9.44E-06	2.86E-05
	rs579459	4.56E-05	1.60E-05	2.14E-04	2.41E-04	1.47E-05	4.50E-05
	rs635634	3.64E-05	1.46E-05	2.14E-04	2.15E-04	1.24E-05	3.35E-05
Univariate Platelet	rs657152	1.90E-01	1.83E-01	3.36E-01	3.53E-01	1.55E-01	1.70E-01
	rs507666	3.68E-01	3.02E-01	6.20E-01	6.53E-01	2.92E-01	3.24E-01
	rs651007	2.29E-01	1.76E-01	3.86E-01	4.20E-01	1.73E-01	1.96E-01
	rs579459	2.33E-01	1.81E-01	3.96E-01	4.32E-01	1.78E-01	2.02E-01
	rs635634	2.77E-01	2.25E-01	4.95E-01	5.22E-01	2.16E-01	2.41E-01
Univariate White Blood Cell	rs657152	2.99E-01	1.04E-01	2.81E-01	1.44E-01	9.46E-02	2.72E-01
	rs507666	1.33E-01	4.88E-02	8.48E-02	7.29E-02	4.29E-02	1.16E-01
	rs651007	2.92E-02	7.17E-03	1.60E-02	1.37E-02	6.22E-03	2.48E-02
	rs579459	2.37E-02	5.62E-03	1.33E-02	1.11E-02	4.85E-03	2.07E-02
	rs635634	7.80E-02	2.76E-02	4.91E-02	4.09E-02	2.41E-02	6.76E-02

Table 9. Top 10 Gene-Based Results (MAF threshold 1%)

	Multivariate	**Univariate**			
Gene	**Q.pval**	**HCT.pval**	**HGB.pval**	**PLT.pval**	**WBC.pval**
SH2B3	4.35E-05	1.64E-01	4.40E-02	1.91E-06	3.22E-02
TPGS2	9.20E-05	5.52E-02	6.71E-02	5.46E-01	1.37E-04
CXCR2	1.20E-04	5.53E-01	9.15E-01	1.45E-01	6.08E-06
CCBL1	2.76E-04	8.31E-01	3.05E-01	3.62E-04	5.49E-01
CR2	2.99E-04	6.62E-02	2.46E-01	2.63E-04	4.89E-01
VPS8	3.81E-04	1.31E-01	1.44E-01	5.89E-04	2.93E-01
ADIPOQ	5.85E-04	5.53E-02	1.29E-01	7.12E-01	1.28E-04
LMOD1	7.16E-04	3.47E-02	9.79E-02	7.80E-01	7.12E-03
MSL1	8.20E-04	6.72E-01	2.94E-01	7.09E-01	2.17E-03
AMFR	9.24E-04	2.46E-01	7.62E-01	9.36E-01	5.71E-02

Table 10. Top 10 Gene-Based Results (MAF threshold 5%)

	Multivariate	**Univariate**			
Gene	**Q.pval**	**HCT.pval**	**HGB.pval**	**PLT.pval**	**WBC.pval**
SH2B3	4.35E-05	1.64E-01	4.40E-02	1.91E-06	3.22E-02
CXCR2	1.20E-04	5.53E-01	9.15E-01	1.45E-01	6.08E-06
EPO	1.58E-04	2.57E-05	6.98E-05	2.50E-02	4.65E-01
MLKL	2.67E-04	8.81E-01	1.90E-01	1.39E-01	5.36E-01
CGN	3.75E-04	9.20E-04	3.38E-03	6.69E-01	2.57E-04
FERMT1	4.79E-04	9.59E-01	1.07E-01	1.23E-01	8.06E-02
ELL	5.45E-04	2.41E-01	1.43E-01	9.73E-03	5.67E-05
ADIPOQ	5.85E-04	5.53E-02	1.29E-01	7.12E-01	1.28E-04
KIAA1797	6.84E-04	2.12E-01	4.51E-03	3.16E-01	8.96E-01
LMOD1	7.16E-04	3.47E-02	9.79E-02	7.80E-01	7.12E-03

Figure 1. Pleiotropy in Blood Cell Traits

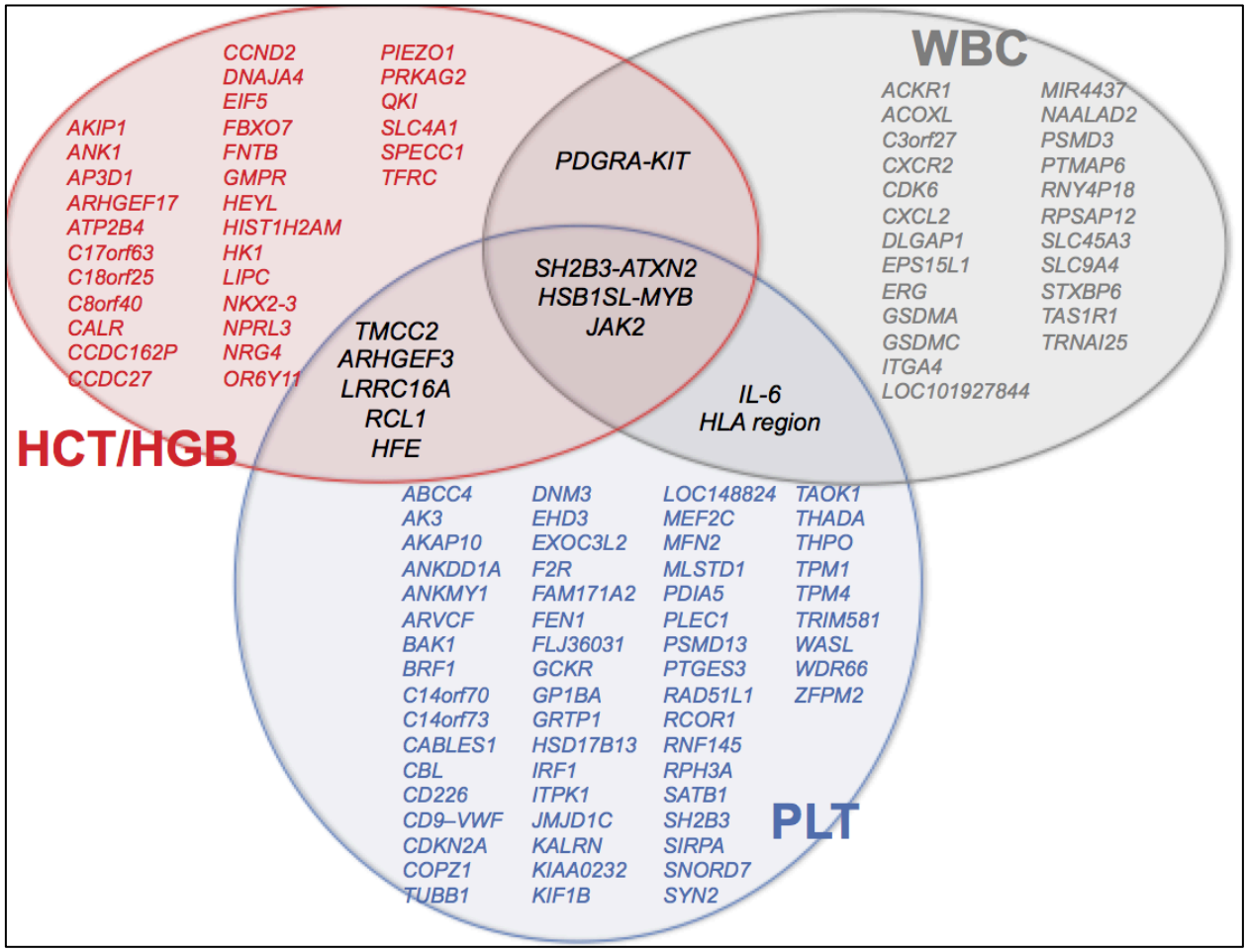


Figure 2. Manhattan Plot of Results from the Single Variant Analysis

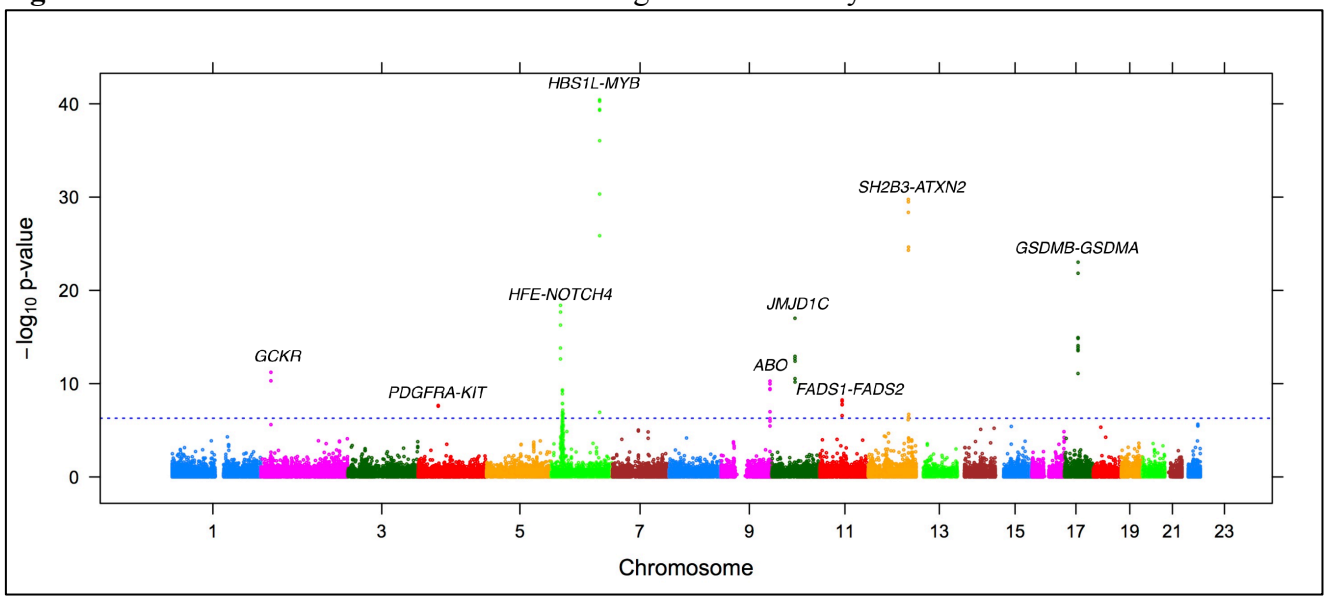
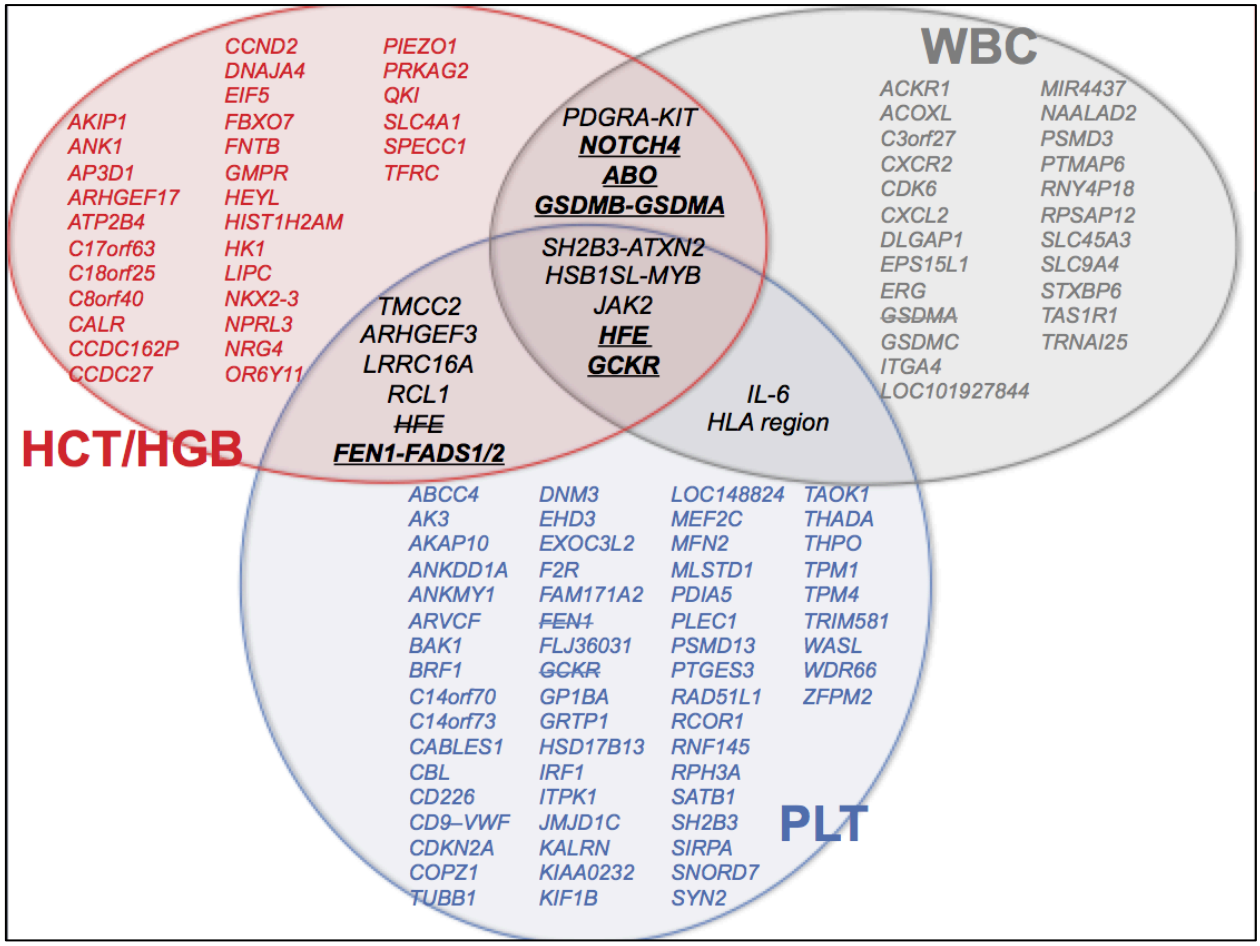


Figure 4. Pleiotropy in Blood Cell Traits with novel cross-phenotype associations



Part B: Somatic Variants of Clinical Relevance

Manuscript

Title: Somatic Variants of Clinical Relevance: expanding the incidentalome

Background

The completion of the Human Genome Project in 2001 heralded the beginning of the genomics era[1-2]. Along with creating a physical map of the human genome, this collaborative effort facilitated many large-scale projects, such as the International HapMap Project[3] and 1000 Genomes Project[4], which sought to characterize genetic variation within and between populations. Coupled with rapid technological advancement, these ambitious projects laid the foundation for widespread use of genome-wide sequencing and genotyping study designs.

More than a decade after the completion of the human genome, researchers are now able to generate vast quantities of genomic data on participants with relative ease. Interrogation at the genome-wide scale permits hypothesis-free or agnostic approaches to genetic research, which represent a significant opportunity to overcome challenges imposed by an incomplete understanding of the pathophysiology of the phenotype. Previous approaches relied on prior knowledge of genetic contribution to the phenotype (e.g. candidate genes), whereas genome-wide approaches permit investigators to interrogate a large number of loci of a given class (e.g. common variants, exonic variants) for association with the phenotype of interest. Capturing hundreds of thousands of genetic variants per genome, these agnostic approaches have been powerful methodologies to identify robust statistical associations with one or multiple phenotypes.

Owing to the large-scale of these genome-wide genotyping and, more recently, sequencing studies, these approaches also have the potential to uncover findings of potential health or reproductive salience to an individual research participant in the course of the study, but beyond the aims of the research study [5]. Termed ancillary, secondary, or most commonly, incidental findings, these results are distinguished from individual research results, which are uncovered in the course of research[5]. Along with the growing trend toward agnostic genome-wide scale projects, it can be argued that the distinction between incidental findings and individual research results ceases to be germane and henceforth, "incidental findings" will encompass both categories of findings. Further, in the context of this chapter, incidental findings will refer to genetic findings unless otherwise specified.

Existing Recommendations

The topic of return of results has been a contentious issue for genetics research. With opinions ranging from no result return to open source models (e.g. full disclosure of all genetic results), guidelines from experts in the field have sought to parameterize result return by striking a balance between providing salient (and presumed beneficial) clinical information and minimizing the burden of result return. Principally, these recommendations seek to give guidance on what should, may, and should not be returned to participants of genetic research.

Existing research ethics guidelines suggest researchers have a limited obligation to return genetic results of likely clinical significance[5-10]. Conditions surrounding the return of incidental findings to individual participants are largely uncontroversial. As summarized by Fabsitz et al.[7], these consensus recommendations generally support that individual incidental findings should be offered to study participant if all the following criteria are met:

1) The genetic finding has important health implications for the participant and the associated risks are established and substantial.

2) The genetic finding is actionable, that is, there are established therapeutic or preventive interventions or other available actions that have the potential to change the clinical course of the disease.

3) The test is analytically valid and the disclosure plan complies with all applicable laws.

4) During the informed consent process or subsequently, the study participant has opted to receive his/her individual genetic results.

These recommendations further suggest that results *may* be returned even if they don't meet criteria 1) and 2), if it is agreed that risks are outweighed by benefits associated with disclosure, the Institutional Review Board approves disclosure, and that conditions 3) and 4) are met. Findings that fail to meet the criteria for return (*should* or *may*) should not be returned to research participants.

The ethical justifications for existing consensus recommendations are derived from the principles of beneficence, non-maleficence and respect for autonomy[11]. As a compromise between these principles and practical considerations, these guidelines seek to maximize benefit and minimize risk to the participant while preventing undue burden on the research enterprise by limiting the number of results to consider for return. At the same time, these guidelines place a high value on personal autonomy to ensure that research participants' decisions and choices are respected.

Taken together, considerations 1) and 2) seek to ensure that the incidental finding is clinically useful and medically actionable, conferring a large and mitigable risk to the participant. Termed the "incidentalome"[12] by one set of authors, the researcher may have an

ethical obligation to warn the participant about such variants in an effort to avert risk of disease. As suggested previously, the result must have high clinical salience in order to ensure that the benefits of disclosure greatly outweigh the associated risks. Benefit is largely assessed on the clinical validity and clinical utility of the genetic finding. Clinical validity refers to the strength of the evidence linking the genetic variant to a phenotypic outcome, and clinical utility refers to the capacity of a genetic variant to confer information that can be useful in clinical interventions to mitigate a known phenotypic outcome[13]. Taken together, recommendations for high clinical utility and clinical validity serve to restrict the pool of potentially returnable variants to a small group of deleterious and actionable findings of immediate clinical salience.

In contrast to some clinical sequencing guidelines[14], most guidelines and legal analyses suggest that there is no duty for the researcher to actively look for actionable genomic findings beyond those uncovered in the normal process of their investigations[9; 15; 16]. The primary purpose of genetic research is to create generalizable knowledge, and consistent with this overarching purpose, attempts have been made to minimize the burden of result return for individual investigators and the research enterprise[17]. Arguments against individual incidental genetic result return focus on implications for resource allocation, and the potential for conflation of research and clinical practice. Considering this core mission of research to produce generalizable knowledge, return of individual genetic findings represents a significant departure. Some argue that this departure from the objective could promote the therapeutic misconception[17]. The therapeutic misconception describes the mistaken belief by a participant that the aim of research is partially or wholly therapeutic. Conflation of science and clinical medicine by participants may be mistakenly interpreted as an inducement to participate in scientific research. Further, use of resources allocated to scientific research to confer individual benefit through return of

incidental findings reduces funds available to create generalizable knowledge. Without specified budget for returning incidental genetic findings, costs to disclose results could significantly detract from available resources for advancement of scientific research, which taken together present compelling arguments in favor of no, or at least very limited, disclosure as suggested by existing guidelines.

The third criterion proposed by Fabsitz et al. maintains that research results considered for return must be analytically valid. Analytic validity in this context refers to the capacity of the test to correctly assign a genetic characteristic (e.g. nucleotide, copy number variant)[13]. Ensuring that results are analytically valid is a measure to reduce false negative or false positive findings, which have the potential to result in unnecessary medical interventions, psychosocial harm or a false sense of security[13]. Though an important consideration, it remains difficult to robustly demonstrate analytic validity of a genetic test result. The Clinical Laboratory Improvement Amendments (CLIA) of 1988 provides regulatory standards in the United States for laboratory tests of human samples for patient care. Administered by the Centers for Medicare & Medicaid Services (CMS), there is uncertainty as to whether research laboratories are exempted from CLIA standards, particularly when reporting clinically relevant incidental findings to participants[7]. Despite a lack consensus on the requirement for CLIA compliance, there is awareness that a plan for disclosure in prospective studies should include information on whether the variant will be genotyped or confirmed in a CLIA-certified laboratory[9]. Beyond Federal law, other applicable laws may provide additional governance surrounding the eligibility of incidental findings for return to participants. As recognized by Fabsitz et al., the third criterion is not straightforward and further research is likely required to provide more specific guidance on the parameters necessary to ensure analytic validity of research results.

The last criterion posits that research participants must opt to receive individual genetic results during the informed consent process. The informed consent process is intended to provide participants with a clear and accurate assessment of the risks and benefits associated with participation in research. Though some hold the opinion that the harm associated with withholding individual results may outweigh participant autonomy, most guidelines suggest that patient preference takes precedent[7; 9; 14]. Consistent with the ethical principle of respect for autonomy, investigators should respect participants' preferences as to whether to receive genetic results.

In addition to these four criteria, other factors are important in decisions surrounding result return. As suggested by Beskow and Burke, the research context, including factors such as "the scope of entrustment involved in the research, the intensity and duration of interactions with participants, and the vulnerability and dependence of the study population", is an important facet to consider in assessing researchers' potential obligations to participant[18]. Acknowledging that return of results does not fit nicely into a "one size fits all" approach, evaluation of the research context and the likelihood of discovery of incidental findings are important considerations for study design.

Due to the complexity of returning incidental findings, it is common for genetic research studies to consent participants under the condition that there will be no individual result return[19]. Though Institutional Review Boards can reconsider this position in the future, this feature of study design largely shuts the door on result return due to difficulties associated with re-contacting participants (e.g. de-identified data, losses to follow-up). Despite the many challenges associated with result return, a minor but growing proportion of research studies has elected to return incidental findings [20-22].

Selected examples of studies returning research results are as follows:

- Colon Cancer Family Registry [21]: return of deleterious germline mutations in mismatch repair genes to colorectal cancer patients and family members
- Costain et al.[20]: return of clinically significant copy number variants to schizophrenia patients and their families
- Harvard Personal Genomes Project [22]: return of "research-grade" whole genome or exome data

Classification Schemes for Genetic Variants

Though recommendations are fairly consistent in the criteria required for return of incidental findings[5-10], less clear guidance is available about what specific genetic findings meet the criteria for result return (e.g. Fabsitz criteria 1) and 2)). Recognizing this gap, Berg et al. devised a scheme to classify genes and variants to assess if a variant should be reported in a clinical context[23] (Figure 1). This scheme parses genes into 3 major bins designated as follows: 1) clinical utility (medically actionable incidental information), 2) clinical validity (A) low risk incidental information, B) medium risk incidental information, C) high risk incidental information) and 3) unknown clinical implications. Within each bin, this scheme defines whether known deleterious, presumed deleterious, unknown significance, presumed benign, and known benign variants should be reported.

Though providing a useful framework for classifying results for return, this publication did not explicitly enumerate all genes and variants fitting each bin. Two years following this publication, the American College of Medical Genetics and Genomics proposed recommendations for reporting incidental findings, which included a list of 56 genes that should

be evaluated and reported to patients who undergo clinical sequencing[14]. In a separate publication, Dorschner et al. recommended that the minimal set of actionable genes in the research setting be composed of 114 genes[24]. Some of the discrepancies relate to the onset of the associated condition (e.g. pediatric vs. adult), but largely the extension of the panel by Dorschner et al. is reflective of differing expert opinion. Though both geared toward clinical sequencing, these efforts to classify actionable genes will also have important implications for result return in the research setting.

Beyond a clear minimum set of actionable genes, there is currently no consensus as to the methodology to identify pathogenic variants within the candidate genes[24]. Attempts to characterize the burden of incidental findings (the "incidentalome") have suggested that ~1% of individuals will harbor an incidental finding in exome or whole genome sequencing[14; 24], while other research posits that a couple to several thousand incidental findings will be discovered per genome[25-27]. Much of the discrepancy in these figures relates to differences in the classification scheme for genes and variants considered to be incidental findings. Various resources, such as the Online Mendelian Inheritance of Man (OMIM[28]), University of California Santa Cruz Genome Browser[29], RegulomeDB[30], and database of Single Nucleotide Polymorphisms (dbSNP[31]) among many others are helpful resources to characterize loci, however these resources are not definitive tests of whether a variant is deleterious. In an attempt to standardize definitions of pathogenic mutations, efforts are currently underway to create databases of pathogenic mutations (e.g. Clinical Genomics Database[32], Human Gene Mutation Database[33], MutaDATABASE[34]).

Present guidelines appear to limit forms of variation considered for individual result return to germline mutations[14]. Germline mutations are heritable forms of variation present in

every cell of the body as a consequence of arising in a germ cell. These mutations are heritable, thus identification of a germline mutation has implications for the individual as well as their relatives. Germline mutations can be uncovered during traditional or tumor-normal subtractive (also referred to as paired tumor-normal) sequencing[25; 35; 36]. As recently pointed out, germline mutations frequently have pleiotropic effects, whereby a single variant or several variants within a single gene influence multiple phenotypes[37]. Existing guidelines largely ignore this well-recognized biological phenomenon, favoring mutually exclusive categories for genes based on clinical relevance for a single condition. Features of pleiotropic variants place them somewhat beyond the bounds of existing guidelines, and therefore Kocarnik and Fullerton suggest that guidelines need to be adapted to incorporate this class of genetic variants[37].

Somatic Mutations as a New Class of Incidental Finding

Discussion of the implications of the discovery of incidental findings of somatic mutations is largely absent from the literature. Current consensus recommendations regarding result return in research[5; 10; 38-40] are implicitly geared toward the discussion of germline mutations. Though germline mutations represent a much larger class of genetic mutations, somatic mutations can confer clinically useful and potentially actionable findings, which may warrant result return. Without clear guidance, it is uncertain how these research results should be handled.

Somatic mutations are acquired, arising after conception and present only in the cell lineage in which they initially arose. These non-heritable genetic changes accumulate throughout the lifespan of an individual, influencing wild-type functioning of individual cells and organ systems[41]. Though somatic mutations underlie many important processes including aging and

neurodegeneration, the clearest and most widely implemented clinical application of this class of mutations is to the field of cancer genomics. Touted as "personalized", "targeted" or "precision" medicine, there has been great enthusiasm to integrate genomic information into clinical settings, particularly oncology. Cancer genomes harbor hundreds to thousands of somatic mutations, frequently occurring in known tumor genes[42]. Though the majority of these mutations have unknown significance, a growing body of research has identified clinically relevant somatic mutations that can be useful in therapeutic selection and diagnosis, as well as in the subclassification of cancers [43-45].

The broadest application of somatic mutations to oncology is therapeutic selection. In addition to known germline mutations that influence drug metabolism, somatic mutations are important molecular markers for therapeutic decisions[44; 46] (Table 1). Based on our growing understanding of the underlying biology of many cancers, many chemotherapeutics have been developed to target proteins that are frequently activated by somatic mutations. Increasingly, identification of somatic mutations through prospective tumor sequencing has been utilized to assign the optimal chemotherapeutic agent to improve efficacy[46]. For example, somatic mutations in exons 12 and 13 of Kirsten rat sarcoma viral oncogene homolog (*KRAS)* of colorectal cancer tumors confer a lack of clinical benefit of Epidermal growth factor receptor (*EGFR*) inhibitors[44; 47]. Beyond drivers related to improving clinical outcomes, pressure to conduct molecular characterization of tumors is heightened by insurance companies that frequently require genetic tests prior to coverage for costly, targeted therapeutics[48]. Beyond assessment of drug effectiveness, somatic mutations can confer information relevant to potential adverse reactions or dosing of chemotherapeutics.

Traditionally relying on histology for diagnosis and subclassification of cancers, particular cancers have molecular signatures that provide additional information for biologic classification. An active and fruitful area of research, somatic mutations in Janus Kinase 2 (*JAK2*), Calrectulin (*CALR*) and Myeloproliferative Leukemia Virus Oncogene (*MPL*) can be useful molecular markers to subclassify Philadelphia chromosome negative myeloproliferative neoplasms (henceforth referred to as myeloproliferative neoplasms). Recently incorporated into the diagnostic criteria for myeloproliferative neoplasms, these somatic mutations serve as a useful case study for somatic mutations as a new class of incidental finding. Here somatic mutations relevant for Philadelphia chromosome negative myeloproliferative neoplasms will be explored as a motivating example for somatic mutations as a class of incidental finding.

Motivating Example: Somatic mutations in Myeloproliferative Neoplasms

In the research context, biospecimens that are easy to acquire are routinely used in genetic studies. Whole blood or fractions of blood (e.g. buffy coat) are the most frequently used biospecimens in population-based genetic research. Cellular constituents of blood include red blood cells, white blood cells, and platelets. Red blood cells as well as platelets are enucleated, whereas white blood cells are nucleated. Based on the absence of nuclei in platelets and red blood cells, white blood cells are the major contributors of genomic DNA. Owing to the widespread use of DNA from blood, genetic studies are capable of uncovering somatic mutations relevant to a range of blood disorders.

Somatic mutations are typically identified through comparison of DNA across multiple tissue types. In clinical sequencing applications, tumor-normal subtractive analyses seek to identify acquired (somatic) mutations specific to the tumor (e.g. not present in the reference

germline tissue). Though genetic studies do not typically have paired tissue DNA samples available for comparison, method development to call somatic mutations (point and larger chromosomal aberrations) is ongoing for both sequence and genotype data. Further, particularly for blood disorders such as myeloproliferative neoplasms (polycythemia vera, essential thrombocythemia and primary myelofibrosis) there are several well-characterized mutations that occur primarily or solely as somatic mutations. Identification of these mutations in routine analysis of stored blood-derived genomic DNA can be presumed to be somatic mutations. The ability to detect blood cell somatic mutations in sequence and genotyping has important implications for the identification of individuals with either increased susceptibility to develop a blood disorder or to identify individuals with an overt blood disorder.

Myeloproliferative Neoplasms

Myeloproliferative neoplasms are a class of chronic myeloproliferative disease that display aberrant hematopoietic proliferation as a result of acquired (somatic) genetic changes in hematopoietic stem cells. This class of neoplasms is typically composed of three major subclassifications: polycythemia vera (PV), essential thrombocythemia (ET) and primary myelofibrosis (PMF). PV is a disorder of the bone marrow that is characterized by high red blood cell counts, which may be accompanied by overproduction of white blood cells and platelets. ET is characterized by overproduction of platelets. Lastly, PMF is characterized by an accumulation of scar tissue in the bone marrow, which reduces the overall production of blood cells.

Landscape of Somatic Mutations Specific to Myeloproliferative Neoplasms

The role of somatic mutations in myeloproliferative neoplasm etiology is an active area of research. First recognized in 2005 with the parallel publication of findings on the gain of function *JAK2* p. Valine 617 Phenylalanine (V617F) somatic mutation, studies suggest that ~95% of PV, and 50% of ET and PMF cases have a detectable V617F acquired mutation[49-53]. *JAK2* is a regulator of blood cell development from hematopoietic stem cells, which when mutated at p.617 contributes to abnormal myeloproliferation with observed differences in phenotype depending on allelic burden (the ratio of mutant allele to total alleles)[54]. Germline variation in *JAK2* at the same amino acid (V617I) is associated with hereditary thrombocytosis, suggesting a causal role for the variant[55]. Additional studies have demonstrated the relevance of variants in exon 12 of *JAK2* among V617F negative patients[56-58]. Though V617F mutation is the predominant mutation in *JAK2,* other somatic mutations in exon 12 of the gene have been identified in about 2-5% of PV cases[59].

MPL mutations are present in a small number of myeloproliferative neoplasms (3% ET and 10% PMF). *MPL* mutations frequently occur in exon 10 of the gene (e.g. W515L, W515K, S505N) and typically carriers have clinical presentation of megakaryocytic myeloproliferation[60]. On the basis of some of the early findings related to the *JAK2* and *MPL* somatic mutations, the World Health Organization (WHO) updated their criteria for differential myeloproliferative neoplasm diagnosis (**Table 2**)[61; 62].

In addition, genetic studies have uncovered additional recurrent somatic mutations that can occur alone or co-occur with known *JAK2* or *MPL* mutations in two main classes of genes: signaling mutations influencing the JAK-STAT activation (lymphocyte-specific adapter protein (*LNK*)[63; 64]) and mutations affecting DNA structure through methylation (TET oncogene family member 2 (*TET2*)[60], DNA methyltransferase 3a (*DNMT3A*)[65; 66]) or other modifications to

chromatin structure (enhancer of zeste homolog 2 (*EZH2*) [67], additional sex combs-like 1 (*ASXL1)* [68]; **Table 3**). Taken together, these somatic mutations affect only a minority of patients and are not recognized as components of clinical diagnostic algorithms for myeloproliferative neoplasms.

Though genetic studies have uncovered the molecular basis of the majority of myeloproliferative neoplasms, until recently 30-45% of patients with myeloproliferative neoplasms (primarily ET and PMF) did not appear to harbor a known molecular marker[69]. However in late 2013 investigators identified somatic insertion/deletions (indels) in calreticulin (*CALR*) specific to ET and PMF that are mutually exclusive of *JAK2* and *MPL* mutations. The discovery of the *CALR* indels filled a major gap in existing knowledge of ET and PMF etiology. With the addition of the *CALR* indel, fewer than 10% of ET and PMF are triple-negative (*JAK2-, MPL- & CALR-*) for a molecular marker[69]. On the basis of these seminal findings, leading experts suggest that the WHO diagnostic criteria for myleoproliferative neoplasms be expanded to incorporative *CALR* mutational testing[70].

Application of Existing Recommendations

Somatic mutations in *JAK2, MPL* and *CALR* may be uncovered in the course of genetic research. For example, *JAK2* V617F is assayed on the widely available Illumina Human Exome and targeted cancer arrays. Application of these arrays has the potential to uncover carriers of the somatic mutation, particularly in the course of association analysis with blood cell traits. Recently, Auer et al. identified an association between rs77375493 (*JAK2* V617F) and platelet count, total white blood cell count and red blood cell parameters (hemoglobin and hematocrit)[71] in apparently hematologically normal individuals. This study identified 19 carriers of the somatic

mutation in the sample size of 24,868 participants, and noted that many carriers appeared to have clinical characteristics consistent with early-stage myeloproliferative neoplasms (**Table 4**). In addition to the *JAK2* V617F and other known somatic point mutations (e.g. *MPL* mutations), methods development in somatic indel calling may soon facilitate the detection of *CALR* indels in sequencing and genotyping data.

As noted above, *CALR, JAK2* and *MPL* mutations are clinically useful and recognized components of the diagnostic criteria for myeloproliferative neoplasms. Though these mutations clearly have relevance within the clinical setting, it is unclear if these mutations meet the criteria for return as an incidental finding in research. In an attempt to explore this topic in greater detail, this analysis will seek to apply criteria enumerated by Fabsitz et al. to assess whether a *JAK2* V617F somatic mutation could be considered for individual result return.

Fabsitz et al. first criterion states that results should be returned if "[t]he genetic finding has important health implications for the participant and the associated risks are established and substantial" and results could be considered for return if "[t]he investigator has concluded that the potential benefits of disclosure outweigh the risks from the participant's perspective"[7]. Of somatic mutations in the three genes, *JAK2* has been the most thoroughly characterized. Population-based studies suggest that carriers of the *JAK2* V617F somatic change (N=68) had clinical characteristics suggestive of myeloproliferative neoplasms (higher erythrocyte, thrombocyte and white blood cell counts) and had an excess risk of incident cancer (Hematologic Cancer Hazard Ratio (HR): 27.6 (95% CI: 12.0-63.4), Myeloproliferative Cancer HR: 97.1 (95% CI: 27.6-341.3)) and prevalent thrombotic events (deep venous thrombosis HR=4.6 (95% CI: 1.7-10.9))[72].

In addition to showing disease risk associated with presence of *JAK2* V617F somatic change, Nielson et al. demonstrated the diagnostic value of screening when applied to the general population[72]. The diagnostic value of *JAK2* V617F + test for prevalent myeloproliferative cancer does not reach clinically useful levels for screening of the general population (sensitivity=23%, specificity=100%, positive predictive value= 18%, negative predictive value = 100%), however, when combined with high blood cell counts, the test characteristics dramatically improved (**Table 5**). For example, use of hematocrit >50% in tandem with *JAK2* V617F status was able to correctly distinguish prevalent myeloproliferative cases from controls. This is relevant in the research context. As suggested previously, genome-wide association studies will frequently have access to complete blood cell parameters including hematocrit, particularly if blood cell measures are the main phenotypes of interest.

Taken together, *JAK2* V617F status, in tandem with other blood cell metrics (e.g. hematocrit), may have high diagnostic value. Further, the *JAK2* V617F somatic mutation confers substantial risk of incident hematologic and myeloproliferative cancer as well as thrombotic events. In particular settings, including association studies of blood cell counts, the identification of V617F could reach the threshold for "should" return and is likely to minimally reach the threshold for "may" return enumerated by Fabsitz et al.

Fabsitz et al. second criterion states that results should be returned if "[t] he genetic finding is actionable, that is, there are established therapeutic or preventive interventions or other available actions that have the potential to change the clinical course of the disease"[7]. For PV and ET patients, the largest cause of morbidity and mortality is from thrombotic events. Previous reports suggest that more than a third of PV cases[73] and an excess of ET[74-76] patients experience

thrombosis prior to diagnosis. Further, thrombotic risk in PV and ET patients appears to be higher among V617F-positive individuals[77-80]. A recent study demonstrated in mouse knock-in models that the V617F mutation has a causal role in thrombosis[14]. Specifically, the Hobbs et al. study demonstrated that the V617F mutation 1) causes intrinsic changes in megakaryocyte platelet formation; and 2) V617F platelets exhibit prothrombotic properties and have increased reactivity[14].

PV and ET are managed clinically as chronic conditions through treatments aimed at avoiding the first occurrence and/or reoccurrence of thrombotic and bleeding complications; reducing risk of transformation to acute leukemia or myelofibrosis; and managing related risk factors (e.g. classic cardiovascular risk factors, pregnancy, surgery)[81]. PV and ET are both characterized by proliferation of blood cells, thus treatment typically focuses on reducing blood viscosity. PV treatment strategies include phlebotomy[82], low-dose aspirin[83], and cytoreduction through hydroxyurea or interferon[84; 85] and these have been shown to be effective in reducing risk of complications. Reduction of thrombotic risk in ET is typically achieved through cytoreduction with hydroxyurea[74; 76; 86]. Timely treatment is essential for PV and ET to mitigate life-threatening complications including blood clots (stroke, myocardial infarction, deep vein thrombosis and pulmonary embolisms), enlarged spleen, skin problems and transformation to blood disorders (myelofibrosis, myelodysplastic syndrome or acute myelogenous leukemia)[87; 88]. Owing to the availability of treatments that reduce risk of thrombotic and myeloproliferative complications in PV and ET, the *JAK2* V617F somatic marker in tandem with blood count data is likely to meet the Fabsitz et al. criterion 2 for a marker that should be returned to participants.

The third criterion seeks to ensure that the test is analytically valid and that disclosure is in compliance with applicable laws. As discussed previously, analytic validity is difficult to

demonstrate in the research context. Somatic mutations, such as *JAK2* V617F, present additional challenges beyond those involved in the identification of germline mutations. Germline mutations are present as either as 100% allele burden (homozygotes) or 50% allele burden heterozygous (heterozygous) mutations, whereas somatic mutations are present on a continuum between the smallest detectable level to nearly 100% allele burden. Due to this feature of somatic mutations, available assays (e.g. genotyping or TaqMan) may fail to detect somatic mutations present at low allelic burden, which would result in reduced sensitivity and positive predictive value of the test[72]. Furthermore, genotype-calling software tends to be optimized for germline variants, frequently relying on clustering of homozygous wild type, heterozygous and homozygous alternative intensities to assign genotype. Because software is optimized for germline variation calls, genotypes falling outside of reference clusters may be discarded or misassigned to the nearest cluster. Due to these issues, there remains a strong potential for inaccuracies in detection of somatic mutations.

Despite difficulties relating to the assignment of somatic genotypes, once detected somatic mutations provide information that is just as relevant as other information that is routinely returned. Furthermore, when ancillary data (e.g. complete blood cell counts) are available, additional genotypic validation may be an unnecessarily high bar for result return. Abnormalities in blood cell counts are frequently reported as incidental findings in the research context due to their known clinical utility. For example, some studies within the Women's Health Initiative have returned test findings suggesting anemia (a measure of low red blood cell count) and complete blood cell counts as individual research results (disclosures summarized in **Table 6**). Though conferring clinically useful information that could supplement blood cell

counts, genetic individual results are not currently returned to participants of the Women's Health Initiative due to constrains in the initial consent language.

Genetic exceptionalism is a term that refers to idea that genetic information differs from other personal data including other forms of health information. Justifications for genetic exceptionalism typically center on the capacity of genetic information to: 1) predict medical future of an individual, 2) have implications for relatives, 3) lead to discrimination or stigmatization, and 4) cause psychosocial harm[89]. Though there are some features unique to genetic information (e.g implications for relatives), there are also many commonalities with non-genetic health-related data. In broad terms, genetic and non-genetic information both have the potential to harm or benefit participants. Consistently, evidence is sparse to suggest that genetic information is truly exceptional, requiring differential treatment in individual result return.

It is well recognized that blood cell parameters have a genetic basis[90]. Further, as demonstrated by the example of myeloproliferative neoplasms, acquired (somatic) genetic changes in hematopoietic stem cells can be causally related to aberrant blood cell phenotypes[14]. Both aberrant blood cell counts as well as molecular markers such as *JAK2* V617F are predictive of blood disorders as well as other complications; thus the justification for treating genetic versus non-genetic information differently ceases to be a relevant argument. Further, somatic mutations are acquired, which implies that this class of mutation has no relevance for related individuals. Hence, concerns about implications for relatives do not figure in the calculus of relative risks and benefits of disclosure.

Justifications related to the potential for discrimination/stigmatization or psychosocial harm similarly fail to support differential treatment of genetic information. Though use of

genetic information to discriminate against particular groups is well documented, non-genetic information such as human immunodeficiency status has a similarly unsavory history. Legal protections have largely been successful at protecting individuals from discrimination for genetic and non-genetic conditions. Federal protections against health-related and/or genetic discrimination come from the Affordable Care Act (ACA), the Health Insurance Portability and Accountability Act (HIPAA), the Americans with Disabilities Act (ADA) and Genetic Information Nondiscrimination Act (GINA).

Further, studies demonstrate that knowledge of disease status can lead to emotional distress[89; 91; 92], suggesting that psychosocial risks are not specific to genetics. Interestingly, despite little evidence of a need for special considerations, many studies categorically preclude genetic results from return as individual research results in consent documents (e.g. Women's Health Initiative Long Life Study).

Consent in Genetic Research

In general, application of the Fabsitz et al. criteria 1-3 appears to favor of return of *JAK2* V617F. However, there has thus far been no attempt (to my knowledge) to return this genetic finding in the WHI study or elsewhere. Beyond a lack of recognition of somatic mutations as a relevant class of incidental finding, this likely relates to a larger issue for individual result return: participants were not informed that they might be offered individual genetic findings during consent. Many studies, particularly those initiated well in the past, are 'silent' on return of genetic and non-genetic research results. In addition, studies such as the Women's Health Initiative Long Life Study consent form expressly forbid individual result return (**Table 6**). Under either consent protocol, re-consent would usually be required by the Institutional Review

Board in order to return genetic results. Though re-consent is necessary to respect the individual research participant, it presents challenges.

Re-consent involves re-contacting individuals enrolled in the study. Particularly in population-based genetic research, data are routinely de-identified to minimize risk of re-identification of participants, and to be classified as non-human subjects research (45 CFR 46.102 (f)). In the absence of participant identifiers, many genetic research studies are unable to re-contact study participants. In addition to limitations relating to re-contact, increasing losses to follow-up are expected as the duration from last contact increases.

Given the limitations with re-consent, it would be preferable if future studies would anticipate the generation of incidental findings (including incidental somatic findings) and allow for expression of preferences regarding individual result return in limited (highly actionable) circumstances. Recent attempts to consent participants prospectively for return of individual genetic findings[20-22] serve as leading examples for newer models of consent, indicating that these protocols are feasible to implement in the research setting.

Discussion

As demonstrated by the *JAK2* V617F mutation, somatic mutations can have clear clinical relevance and may merit return as incidental findings. Existing guidelines fail to recognize somatic mutations as a class of incidental finding, which may be attributable to relatively small number of somatic mutations that could be considered for return under existing criteria. An active area of research, it is likely that similarly deleterious somatic mutations could be identified with similar attributes to the *JAK2* V617F. For example, results from three population-based studies suggest that large-scale chromosomal mosaicism detected in blood or saliva is associated

with a large excess risk of hematologic cancer (e.g. leukemias and lymphomas)[93-95]. However, more research is necessary to confirm these findings and to identify specific somatic events underlying the association with hematologic cancer. Further, recent evidence of low-level somatic mutations present in an individual can be transmitted to their offspring[96]. This finding has huge implications for result return related to reproductive decision-making.

Limitations in detection, both with regard to the availability of multiple tissue types and the allelic burden of the somatic mutation, restrict the routine return of this class of mutations. However, methods development for somatic mutation detection may alleviate some of these issues in the future. For example, several groups are currently working on technology to capture circulating tumor cells in the bloodstream[97]. Coupled with development of single cell sequencing and genotyping assays, these technologies may promote the discovery of actionable somatic mutations.

Particularly in light of progress relating to the detection of somatic mutations, guidelines need to be adapted to recognize this class of incidental finding. As shown in the case study of myeloproliferative neoplasms, *JAK2* V617F arguably has characteristics consistent with return as an incidental finding. It is debatable if the level of evidence is consistent with *should* or *may* return, however it seems clear that reporting V617F confers a benefit that is larger than risks associated with disclosure.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science 291, 1304-1351.
3. (2003). The International HapMap Project. Nature 426, 789-796.
4. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
5. Wolf, S.M., Lawrenz, F.P., Nelson, C.A., Kahn, J.P., Cho, M.K., Clayton, E.W., Fletcher, J.G., Georgieff, M.K., Hammerschmidt, D., Hudson, K., et al. (2008). Managing incidental findings in human subjects research: analysis and recommendations. The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics 36, 219-248, 211.
6. Bookman, E.B., Langehorne, A.A., Eckfeldt, J.H., Glass, K.C., Jarvik, G.P., Klag, M., Koski, G., Motulsky, A., Wilfond, B., Manolio, T.A., et al. (2006). Reporting genetic results in research studies: summary and recommendations of an NHLBI working group. American journal of medical genetics Part A 140, 1033-1040.
7. Fabsitz, R.R., McGuire, A., Sharp, R.R., Puggal, M., Beskow, L.M., Biesecker, L.G., Bookman, E., Burke, W., Burchard, E.G., Church, G., et al. (2010). Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. Circulation Cardiovascular genetics 3, 574-580.
8. Dressler, L.G. (2009). Disclosure of research results from cancer genomic studies: state of the science. Clinical cancer research : an official journal of the American Association for Cancer Research 15, 4270-4276.
9. Jarvik, G.P., Amendola, L.M., Berg, J.S., Brothers, K., Clayton, E.W., Chung, W., Evans, B.J., Evans, J.P., Fullerton, S.M., Gallego, C.J., et al. (2014). Return of genomic results to research participants: the floor, the ceiling, and the choices in between. American journal of human genetics 94, 818-826.
10. Caulfield, T., McGuire, A.L., Cho, M., Buchanan, J.A., Burgess, M.M., Danilczyk, U., Diaz, C.M., Fryer-Edwards, K., Green, S.K., Hodosh, M.A., et al. (2008). Research ethics recommendations for whole-genome research: consensus statement. PLoS biology 6, e73.
11. Beauchamp, T.L., and Childress, J.F. (2001). Principles of biomedical ethics.(New York, N.Y.: Oxford University Press).
12. Kohane, I.S., Masys, D.R., and Altman, R.B. (2006). The incidentalome: a threat to genomic medicine. JAMA : the journal of the American Medical Association 296, 212-215.
13. Ravitsky, V., and Wilfond, B.S. (2006). Disclosing individual genetic results to research participants. The American journal of bioethics : AJOB 6, 8-17.
14. Hobbs, C.M., Manning, H., Bennett, C., Vasquez, L., Severin, S., Brain, L., Mazharian, A., Guerrero, J.A., Li, J., Soranzo, N., et al. (2013). JAK2V617F leads to intrinsic changes in platelet formation and reactivity in a knock-in mouse model of essential thrombocythemia. Blood 122, 3787-3797.

15. Ulrich, M. (2013). The duty to rescue in genomic research. The American journal of bioethics : AJOB 13, 50-51.
16. United States. Presidential Commission for the Study of Bioethical Issues. (2013). Anticipate and communicate : ethical management of incidental and secondary findings in the clinical, research, and direct-to-consumer contexts.(Washington, D.C.: Presidential Commission for the Study of Bioethical Issues).
17. Bredenoord, A.L., Kroes, H.Y., Cuppen, E., Parker, M., and van Delden, J.J. (2011). Disclosure of individual genetic data to research participants: the debate reconsidered. Trends in genetics : TIG 27, 41-47.
18. Beskow, L.M., and Burke, W. (2010). Offering individual genetic research results: context matters. Science translational medicine 2, 38cm20.
19. Cho, M.K. (2008). Understanding incidental findings in the context of genetics and genomics. The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics 36, 280-285, 212.
20. Costain, G., Lionel, A.C., Merico, D., Forsythe, P., Russell, K., Lowther, C., Yuen, T., Husted, J., Stavropoulos, D.J., Speevak, M., et al. (2013). Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays. Hum Mol Genet 22, 4485-4501.
21. Keogh, L.A., Fisher, D., Sheinfeld Gorin, S., Schully, S.D., Lowery, J.T., Ahnen, D.J., Maskiell, J.A., Lindor, N.M., Hopper, J.L., Burnett, T., et al. (2014). How do researchers manage genetic results in practice? The experience of the multinational Colon Cancer Family Registry. Journal of community genetics 5, 99-108.
22. Ball, M.P., Bobe, J.R., Chou, M.F., Clegg, T., Estep, P.W., Lunshof, J.E., Vandewege, W., Zaranek, A., and Church, G.M. (2014). Harvard Personal Genome Project: lessons from participatory public research. Genome medicine 6, 10.
23. Berg, J.S., Khoury, M.J., and Evans, J.P. (2011). Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. Genetics in medicine : official journal of the American College of Medical Genetics 13, 499-504.
24. Dorschner, M.O., Amendola, L.M., Turner, E.H., Robertson, P.D., Shirts, B.H., Gallego, C.J., Bennett, R.L., Jones, K.L., Tokita, M.J., Bennett, J.T., et al. (2013). Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. American journal of human genetics 93, 631-640.
25. Berg, J.S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C.P., Wilhelmsen, K.C., and Evans, J.P. (2013). An informatics approach to analyzing the incidentalome. Genetics in medicine : official journal of the American College of Medical Genetics 15, 36-44.
26. Cassa, C.A., Savage, S.K., Taylor, P.L., Green, R.C., McGuire, A.L., and Mandl, K.D. (2012). Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. Genome research 22, 421-428.
27. Johnston, J.J., Rubinstein, W.S., Facio, F.M., Ng, D., Singh, L.N., Teer, J.K., Mullikin, J.C., and Biesecker, L.G. (2012). Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. American journal of human genetics 91, 97-108.
28. McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. American journal of human genetics 80, 588-604.

29. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research 12, 996-1006.
30. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome research 22, 1790-1797.
31. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311.
32. Solomon, B.D., Nguyen, A.D., Bear, K.A., and Wolfsberg, T.G. (2013). Clinical genomic database. Proceedings of the National Academy of Sciences of the United States of America 110, 9851-9855.
33. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics 133, 1-9.
34. Bale, S., Devisscher, M., Van Criekinge, W., Rehm, H.L., Decouttere, F., Nussbaum, R., Dunnen, J.T., and Willems, P. (2011). MutaDATABASE: a centralized and standardized DNA variation database. Nature biotechnology 29, 117-118.
35. Bombard, Y., Robson, M., and Offit, K. (2013). Revealing the incidentalome when targeting the tumor genome. JAMA : the journal of the American Medical Association 310, 795-796.
36. Stadler, Z.K., Schrader, K.A., Vijai, J., Robson, M.E., and Offit, K. (2014). Cancer genomics and inherited risk. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 32, 687-698.
37. Kocarnik, J.M., and Fullerton, S.M. (2014). Returning pleiotropic results from genetic testing to patients and research participants. JAMA : the journal of the American Medical Association 311, 795-796.
38. (CDC), C.f.D.C.a.P. (2010). Genomic testing: ACCE model process for evaluating genetic tests. In. (
39. (NBAC), N.B.A.C. (1999). Research involving human biological materials: ethical issues and policy guidance, vol. 1.(Rockville, MD).
40. National Heart, L., Blood Institute working, g., Fabsitz, R.R., McGuire, A., Sharp, R.R., Puggal, M., Beskow, L.M., Biesecker, L.G., Bookman, E., Burke, W., et al. (2010). Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. Circulation Cardiovascular genetics 3, 574-580.
41. Kennedy, S.R., Loeb, L.A., and Herr, A.J. (2012). Somatic mutations in aging, cancer and neurodegeneration. Mechanisms of ageing and development 133, 118-126.
42. Dienstmann, R., Dong, F., Borger, D., Dias-Santagata, D., Ellisen, L.W., Le, L.P., and Iafrate, A.J. (2014). Standardized decision support in next generation sequencing reports of somatic cancer variants. Molecular oncology.
43. Tothova, Z., Steensma, D.P., and Ebert, B.L. (2013). New strategies in myelodysplastic syndromes: application of molecular diagnostics to clinical practice. Clinical cancer research : an official journal of the American Association for Cancer Research 19, 1637-1643.

44. McLeod, H.L. (2013). Cancer pharmacogenomics: early promise, but concerted effort needed. Science 339, 1563-1566.
45. McDermott, U., Downing, J.R., and Stratton, M.R. (2011). Genomics and the continuum of cancer care. The New England journal of medicine 364, 340-350.
46. Filipski, K.K., Mechanic, L.E., Long, R., and Freedman, A.N. (2014). Pharmacogenomics in oncology care. Frontiers in genetics 5, 73.
47. Amado, R.G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D.J., Juan, T., Sikorski, R., Suggs, S., Radinsky, R., et al. (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 26, 1626-1634.
48. Patel, J.N., Mandock, K., and McLeod, H.L. (2014). Clinically relevant cancer biomarkers and pharmacogenetic assays. Journal of oncology pharmacy practice : official publication of the International Society of Oncology Pharmacy Practitioners 20, 65-72.
49. Baxter, E.J., Scott, L.M., Campbell, P.J., East, C., Fourouclas, N., Swanton, S., Vassiliou, G.S., Bench, A.J., Boyd, E.M., Curtin, N., et al. (2005). Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. Lancet 365, 1054-1061.
50. James, C., Ugo, V., Le Couédic, J.P., Staerk, J., Delhommeau, F., Lacout, C., Garçon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., et al. (2005). A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature 434, 1144-1148.
51. Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.S., Tiedt, R., Passweg, J.R., Tichelli, A., Cazzola, M., and Skoda, R.C. (2005). A gain-of-function mutation of JAK2 in myeloproliferative disorders. The New England journal of medicine 352, 1779-1790.
52. Levine, R.L., Wadleigh, M., Cools, J., Ebert, B.L., Wernig, G., Huntly, B.J., Boggon, T.J., Wlodarska, I., Clark, J.J., Moore, S., et al. (2005). Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell 7, 387-397.
53. Jones, A.V., Chase, A., Silver, R.T., Oscier, D., Zoi, K., Wang, Y.L., Cario, H., Pahl, H.L., Collins, A., Reiter, A., et al. (2009). JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nature genetics 41, 446-449.
54. Tefferi, A., Noel, P., and Hanson, C.A. (2011). Uses and abuses of JAK2 and MPL mutation tests in myeloproliferative neoplasms a paper from the 2010 William Beaumont hospital symposium on molecular pathology. The Journal of molecular diagnostics : JMD 13, 461-466.
55. Mead, A.J., Rugless, M.J., Jacobsen, S.E., and Schuh, A. (2012). Germline JAK2 mutation in a family with hereditary thrombocytosis. The New England journal of medicine 366, 967-969.
56. Pardanani, A., Lasho, T.L., Finke, C., Hanson, C.A., and Tefferi, A. (2007). Prevalence and clinicopathologic correlates of JAK2 exon 12 mutations in JAK2V617F-negative polycythemia vera. Leukemia 21, 1960-1963.
57. Scott, L.M., Tong, W., Levine, R.L., Scott, M.A., Beer, P.A., Stratton, M.R., Futreal, P.A., Erber, W.N., McMullin, M.F., Harrison, C.N., et al. (2007). JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. The New England journal of medicine 356, 459-468.

58. Williams, D.M., Kim, A.H., Rogers, O., Spivak, J.L., and Moliterno, A.R. (2007). Phenotypic variations and new mutations in JAK2 V617F-negative polycythemia vera, erythrocytosis, and idiopathic myelofibrosis. Exp Hematol 35, 1641-1646.
59. Kiladjian, J.J. (2012). The spectrum of JAK2-positive myeloproliferative neoplasms. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program 2012, 561-566.
60. Tefferi, A. (2010). Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1. Leukemia 24, 1128-1138.
61. Tefferi, A., Thiele, J., and Vardiman, J.W. (2009). The 2008 World Health Organization classification system for myeloproliferative neoplasms: order out of chaos. Cancer 115, 3842-3847.
62. Thiele, J., and Kvasnicka, H.M. (2009). The 2008 WHO diagnostic criteria for polycythemia vera, essential thrombocythemia, and primary myelofibrosis. Current hematologic malignancy reports 4, 33-40.
63. Lasho, T.L., Pardanani, A., and Tefferi, A. (2010). LNK mutations in JAK2 mutation-negative erythrocytosis. The New England journal of medicine 363, 1189-1190.
64. Pardanani, A., Lasho, T., Finke, C., Oh, S.T., Gotlib, J., and Tefferi, A. (2010). LNK mutation studies in blast-phase myeloproliferative neoplasms, and in chronic-phase disease with TET2, IDH, JAK2 or MPL mutations. Leukemia 24, 1713-1718.
65. Abdel-Wahab, O., Pardanani, A., Rampal, R., Lasho, T.L., Levine, R.L., and Tefferi, A. (2011). DNMT3A mutational analysis in primary myelofibrosis, chronic myelomonocytic leukemia and advanced phases of myeloproliferative neoplasms. Leukemia 25, 1219-1220.
66. Stegelmann, F., Bullinger, L., Schlenk, R.F., Paschka, P., Griesshammer, M., Blersch, C., Kuhn, S., Schauer, S., Dohner, H., and Dohner, K. (2011). DNMT3A mutations in myeloproliferative neoplasms. Leukemia 25, 1217-1219.
67. Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A.V., Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., et al. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nature genetics 42, 722-726.
68. Carbuccia, N., Murati, A., Trouplin, V., Brecqueville, M., Adelaide, J., Rey, J., Vainchenker, W., Bernard, O.A., Chaffanet, M., Vey, N., et al. (2009). Mutations of ASXL1 gene in myeloproliferative neoplasms. Leukemia 23, 2183-2186.
69. Klampfl, T., Gisslinger, H., Harutyunyan, A.S., Nivarthi, H., Rumi, E., Milosevic, J.D., Them, N.C., Berg, T., Gisslinger, B., Pietra, D., et al. (2013). Somatic mutations of calreticulin in myeloproliferative neoplasms. The New England journal of medicine 369, 2379-2390.
70. Tefferi, A., and Pardanani, A. (2014). Genetics: CALR mutations and a new diagnostic algorithm for MPN. Nature reviews Clinical oncology 11, 125-126.
71. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dube, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nature genetics.
72. Nielsen, C., Birgens, H.S., Nordestgaard, B.G., and Bojesen, S.E. (2013). Diagnostic value of JAK2 V617F somatic mutation for myeloproliferative cancer in 49 488 individuals from the general population. British journal of haematology 160, 70-79.

73. Landolfi, R., Marchioli, R., Kutti, J., Gisslinger, H., Tognoni, G., Patrono, C., Barbui, T., and Investigators, E.C.o.L.-D.A.i.P.V. (2004). Efficacy and safety of low-dose aspirin in polycythemia vera. The New England journal of medicine 350, 114-124.
74. Cortelazzo, S., Finazzi, G., Ruggeri, M., Vestri, O., Galli, M., Rodeghiero, F., and Barbui, T. (1995). Hydroxyurea for patients with essential thrombocythemia and a high risk of thrombosis. The New England journal of medicine 332, 1132-1136.
75. Elliott, M.A., and Tefferi, A. (2005). Thrombosis and haemorrhage in polycythaemia vera and essential thrombocythaemia. British journal of haematology 128, 275-290.
76. Harrison, C.N., Campbell, P.J., Buck, G., Wheatley, K., East, C.L., Bareford, D., Wilkins, B.S., van der Walt, J.D., Reilly, J.T., Grigg, A.P., et al. (2005). Hydroxyurea compared with anagrelide in high-risk essential thrombocythemia. The New England journal of medicine 353, 33-45.
77. Campbell, P.J., Scott, L.M., Buck, G., Wheatley, K., East, C.L., Marsden, J.T., Duffy, A., Boyd, E.M., Bench, A.J., Scott, M.A., et al. (2005). Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. Lancet 366, 1945-1953.
78. Dahabreh, I.J., Zoi, K., Giannouli, S., Zoi, C., Loukopoulos, D., and Voulgarelis, M. (2009). Is JAK2 V617F mutation more than a diagnostic index? A meta-analysis of clinical outcomes in essential thrombocythemia. Leukemia research 33, 67-73.
79. Lussana, F., Caberlon, S., Pagani, C., Kamphuisen, P.W., Buller, H.R., and Cattaneo, M. (2009). Association of V617F Jak2 mutation with the risk of thrombosis among patients with essential thrombocythaemia or idiopathic myelofibrosis: a systematic review. Thrombosis research 124, 409-417.
80. Vannucchi, A.M., Antonioli, E., Guglielmelli, P., Rambaldi, A., Barosi, G., Marchioli, R., Marfisi, R.M., Finazzi, G., Guerini, V., Fabris, F., et al. (2007). Clinical profile of homozygous JAK2 617V>F mutation in patients with polycythemia vera or essential thrombocythemia. Blood 110, 840-846.
81. Barbui, T., Barosi, G., Birgegard, G., Cervantes, F., Finazzi, G., Griesshammer, M., Harrison, C., Hasselbalch, H.C., Hehlmann, R., Hoffman, R., et al. (2011). Philadelphia-negative classical myeloproliferative neoplasms: critical concepts and management recommendations from European LeukemiaNet. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 29, 761-770.
82. Marchioli, R., Finazzi, G., Specchia, G., Cacciola, R., Cavazzina, R., Cilloni, D., De Stefano, V., Elli, E., Iurlo, A., Latagliata, R., et al. (2013). Cardiovascular events and intensity of treatment in polycythemia vera. The New England journal of medicine 368, 22-33.
83. Landolfi, R., Marchioli, R., Kutti, J., Gisslinger, H., Tognoni, G., Patrono, C., Barbui, T., and European Collaboration on Low-Dose Aspirin in Polycythemia Vera, I. (2004). Efficacy and safety of low-dose aspirin in polycythemia vera. The New England journal of medicine 350, 114-124.
84. Larsen, T.S., Bjerrum, O.W., Pallisgaard, N., Andersen, M.T., Moller, M.B., and Hasselbalch, H.C. (2008). Sustained major molecular response on interferon alpha-2b in two patients with polycythemia vera. Annals of hematology 87, 847-850.
85. Kiladjian, J.J., Cassinat, B., Chevret, S., Turlure, P., Cambier, N., Roussel, M., Bellucci, S., Grandchamp, B., Chomienne, C., and Fenaux, P. (2008). Pegylated interferon-alfa-2a induces complete hematologic and molecular responses with low toxicity in polycythemia vera. Blood 112, 3065-3072.

86. Tefferi, A., Gangat, N., and Wolanskyj, A.P. (2006). Management of extreme thrombocytosis in otherwise low-risk essential thrombocythemia; does number matter? Blood 108, 2493-2494.
87. Finazzi, G., and Barbui, T. (2007). How I treat patients with polycythemia vera. Blood 109, 5104-5111.
88. Nielsen, C., Birgens, H.S., Nordestgaard, B.G., Kjaer, L., and Bojesen, S.E. (2011). The JAK2 V617F somatic mutation, mortality and cancer risk in the general population. Haematologica 96, 450-453.
89. Green, M.J., and Botkin, J.R. (2003). "Genetic exceptionalism" in medicine: clarifying the differences between genetic and nongenetic tests. Annals of internal medicine 138, 571-575.
90. Pilia, G., Chen, W.M., Scuteri, A., Orru, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS genetics 2, e132.
91. Alonzo, A.A., and Reynolds, N.R. (1995). Stigma, HIV and AIDS: an exploration and elaboration of a stigma trajectory. Social science & medicine 41, 303-315.
92. Johnston, M.E., Gibson, E.S., Terry, C.W., Haynes, R.B., Taylor, D.W., Gafni, A., Sicurella, J.I., and Sackett, D.L. (1984). Effects of labelling on income, work and social function among hypertensive employees. Journal of chronic diseases 37, 417-423.
93. Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J., et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nature genetics 44, 651-658.
94. Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nature genetics 44, 642-650.
95. Schick, U.M., McDavid, A., Crane, P.K., Weston, N., Ehrlich, K., Newton, K.M., Wallace, R., Bookman, E., Harrison, T., Aragaki, A., et al. (2013). Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. PloS one 8, e59823.
96. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M.E., Nagamani, S.C., Erez, A., Bartnik, M., Wisniowiecka-Kowalnik, B., et al. (2014). Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. American journal of human genetics.
97. Marx, V. (2013). Tracking metastasis and tricking cancer. Nature 494, 133-136.
98. Tefferi, A., Lasho, T.L., Abdel-Wahab, O., Guglielmelli, P., Patel, J., Caramazza, D., Pieri, L., Finke, C.M., Kilpivaara, O., Wadleigh, M., et al. (2010). IDH1 and IDH2 mutation studies in 1473 patients with chronic-, fibrotic- or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis. Leukemia 24, 1302-1309.
99. Grand, F.H., Hidalgo-Curtis, C.E., Ernst, T., Zoi, K., Zoi, C., McGuire, C., Kreil, S., Jones, A., Score, J., Metzgeroth, G., et al. (2009). Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. Blood 113, 6182-6192.
100. Jager, R., Gisslinger, H., Passamonti, F., Rumi, E., Berg, T., Gisslinger, B., Pietra, D., Harutyunyan, A., Klampfl, T., Olcaydu, D., et al. (2010). Deletions of the transcription factor Ikaros in myeloproliferative neoplasms. Leukemia 24, 1290-1298.

Figures & Tables

Figure 1. Berg et al.[23] Classification Scheme for Genes and Variants

	Criteria:	*Clinical Utility*	*Clinical Validity*			*Unknown Clinical Implications*
Genes	Bins:	**Bin 1 Medically actionable incidental information**	**Bin 2A Low risk incidental information**	**Bin 2B Medium risk incidental information**	**Bin 2C High risk incidental information**	**Bin 3**
	Examples:	*BRCA1/2 MLH1, MSH2 FBN1 NF1*	PGx variants and common risk SNPs	*APOE* Carrier status for recessive Mendelian disorders	Huntington Prion diseases ALS (SOD1)	**All other loci**
	Estimated number of genes/loci:	10s	10s (eventually 100s – 1000s)	1000s	10s	~20,000
Alleles that would be reportable (YES) or not reportable (NO) in a clinical context						
Variants	Known deleterious	YES	YES/NO [1]	YES/NO [1]	YES/NO [1]	N/A [2]
	Presumed deleterious	YES	N/A [3]	YES/NO [1]	YES/NO [1]	NO [4]
	VUS	NO	N/A [3]	NO	NO	NO [4]
	Presumed benign	NO	N/A [3]	NO	NO	NO
	Known benign	NO	NO	NO	NO	NO

N/A: not applicable; VUS: Variant of uncertain significance

[1] Reporting through decision making with an appropriate provider if elected by the patient.

[2] By definition, variants in genes with unknown implications could not be considered deleterious.

[3] By definition, SNPs or PGx variants will either be present or absent.

[4] Variants in genes with unknown clinical implications would not be reported; however, they may serve as an important substrate for research, potentially uncovering new disease genes.

Table 1. Pharmacogenomic DNA markers for chemotherapy (adapted from[44])

Germline	Somatic	Drug	Effect
Thiopurine methyltransferase	_	Mercaptopurine, thioguanine	Neutropenia risk
UDP-glucuronosyltransferase 1A1	_	Irinotecan, nilotinib	Neutropenia risk, underdosing risk
Glucose-6-phosphate dehydrogenase	_	Rasburicase	Anemia
Cytochrome P450 2D6	_	Codeine, oxycodone; tamoxifen	Altered pain control; altered drug dose
_	Janus kinase 2 (JAK2)	Ruxolitinib	Altered drug activity
_	Human epidermal growth factor receptor 1 (EGFR)	Cetuximab Erlotinib, Gefitinib, Panitumumab	Altered drug activity
_	Kirsten rat sarcoma viral oncogene homolog (KRAS)	Cetuximab, Pantumumab	Lack of drug activity
_	Abelson murine leukemia viral oncogene homolog (ABL)	Imatinib, dastinib, nilotinib	Altered drug activity
_	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (KIT)	Imatinib	Altered drug activity
_	Human epidermal growth factor receptor 2 (HER2)	Lapatinib Trastuzumab	Enhanced drug activity
_	v-Raf murine sarcoma viral oncogene homolog B1 (BRAF)	Vemurafenib	Enhanced drug activity
_	Anaplastic lymphoma receptor tyrosine kinase (ALK)	Crizontinib	Altered drug activity

Table 2. World Health Organization Diagnostic Criteria for Myeloproliferative Disorders from 2008[61]

Criteria	**PV***	**ET****	**PMF*****
Major 1	Hgb >18.5 g/dL (men) >16.5 g/dL (women) or Hgb >17 g/dL (men), or >15 g/dL (women) if associated with a sustained increase of ≥ 2 g/dL from baseline that can not be attributed to correction of iron deficiency or§	Platelet count ≥ 450 $x10^9$/L	Megakaryocyte proliferation and atypia‖ accompanied by either reticulin and/or collagen fibrosis, or in the absence of reticulin fibroisis, the megakaryocyte changes must be accompanied by increased bonemarrow cellularity, granulocytic proliferation and often decreased erythropoiesis (ie, prefibrotic PMF)
Major 2	Presence of *JAK2*V617F or similar mutation	Megakaryocyte proliferation with large and mature morphology. No or little granulocyte or erythroid proliferation	Not meeting WHO criteria for CML, PV, MDS, or other myeloid neoplasm
Major 3	NA	Not meeting WHO criteria for 3 CML, PV, PMF, MDS or other myeloid neoplasm	Demonstration of *JAK2*V617F or other clonal marker or no evidence of reactive bone marrow fibrosis
Major 4	NA	Demonstration of *JAK2*V617F or other clonal marker or no evidence of reactive thrombocytosis	
Minor 1	BM trilineage myeloproliferation	NA	Leukoerythroblastosis
Minor 2	Subnormal serum Epo level	NA	Increased serum LDH
Minor 3	ECC growth	NA	Anemia
Minor 4	NA	NA	Palpable splenomegaly

WHO indicates World Health Organization; PV, polycythemia vera; ET, essential thrombocythemia; PMF, primary myelofibrosis; Hgb, hemoglobin; CML, chronic myelogenous leukemia; MDS, myelodysplastic syndrome; BM, bone marrow; Epo, erythropoietin; LDH, lactate dehydrogenase; EEC, endogenous erythroid colony. * The diagnosis of PV requires meeting both major criteria and 1 minor criterion or the first major criterion and 2 minor criteria. ** The diagnosis of ET requires meeting all 4 major criteria. *** The diagnosis of PMF requires meeting all 3 major criteria and 2 minor criteria. § Or Hgb or hematocrit greater than the 99th percentile of reference range for age, sex, or altitude of residence or red cell mass >25% above the mean normal predicted. || Small to large megakaryocytes with an aberrant nuclear/cytoplasmic ratio and hyperchromatic and irregularly folded nuclei and dense clustering.

Table 3. Known Recurrent Mutations in *BCR-ABL1*-Negative Myeloproliferative Neoplasms (adapted from[54])

Mutations	Chromosome location	Mutational frequency, %
JAK2 V617F	9p24	
PV		~96[60]
ET		~55[60]
PMF		~65[60]
BP-MPN		~50[60]
JAK2 exon 12 mutation	9p24	
PV		~3[60]
CALR indel	19p13	
PV		Rare[69]
ET		~67[69]
PMF		~88[69]
MPL	1p34	
ET		~3[60]
PMF		~10[60]
BP-MPN		~5[60]
LNK	12q24.12	
PV		Rare[63; 64]
ET		Rare[67; 68]
PMF		Rare[67; 68]
BP-MPN		~10[63]
TET2	4q24	
PV		~16[60]
ET		~5[60]
PMF		~17[60]
BP-MPN		~17[60]
ASXL1	20q11.1	
ET		~3[68]
PMF		~13[68]
BP-MPN		~18[68]
IDH1/*IDH2*	2q33.3/15q26.1	
PV		~2[98]
ET		~1[98]
PMF		~4[98]
BP-MPN		~20[98]
EZH2	7q36.1	
PV		~3[67]
PMF		~7
DNMT3A	2p23	
PV		~7[66]
PMF		~7[65; 66]
BP-MPN		~14[65; 66]
CBL	11q23.3	
PV		Rare[99]
ET		Rare[99]
MF		~6[99]
IKZF1	7p12	
CP-MPN		Rare[100]
BP-MPN		~19[100]

BP-MPN, blast-phase MPN; CP-MPN, chronic phase MPN; MF, both PMF and post-ET/PV myelofibrosis

Table 4. Characteristics of WHI *JAK2* somatic 617F>V variant carriers (Adapted from Auer et al.[71])

ID	Baseline				Year 3 (follow-up)				CHD	Stroke	Leukemia	Cause of death	Number of *JAK2* p.Val617Phe alleles
	HCT%	HGB g/dl	PLT 10^9/l	WBC 10^9/l	HCT %	HGB g/dl	PLT 10^9/l	WBC 10^9/l					
1	46	15.9	658	9	50.6	16.8	662	12.7	No	No	No	NA	1
2	40	12.5	502	5.9	NA	NA	NA	NA	No	Yes	No	Other Cancer	1
3	46	13	796	23.4	54	15.4	647	67.5	No	Yes	No	Cerebrovascular	2
4	40.7	12.9	482	10.7	NA	NA	NA	NA	No	Yes	Yes	NA	1
5	41.1	14	437	3.9	NA	NA	NA	NA	Yes	Yes	No	NA	1
6	46.6	16.1	310	8.3	47.2	15.4	259	11.5	No	Yes	No	NA	1
7	48	15.9	357	6.9	NA	NA	NA	NA	No	No	No	NA	1
8	45.4	14.8	411	8.6	NA	NA	NA	NA	Yes	No	No	Other Known Cause	1
9	39.9	13.5	227	5.4	NA	NA	NA	NA	No	Yes	No	NA	1
10	43.8	14.8	365	5.8	NA	NA	NA	NA	Yes	No	No	NA	1
11	42.5	14.7	279	5.8	NA	NA	NA	NA	Yes	No	No	NA	1
12	49.5	16.5	529	7.3	NA	NA	NA	NA	No	No	No	NA	1
13	42.5	14.2	411	7	NA	NA	NA	NA	No	No	No	NA	1
14	49.3	16.7	266	9.1	NA	NA	NA	NA	No	No	No	NA	1
15	49.4	16.4	697	8	NA	NA	NA	NA	Yes	No	No	NA	1
16	45.1	14.2	412	9.5	NA	NA	NA	NA	No	Yes	No	NA	1
17	37.5	13.1	239	11.4	NA	NA	NA	NA	Yes	No	No	NA	1
18	42.9	14.4	201	11.7	NA	NA	NA	NA	Yes	Yes	No	NA	1
19	42.3	14.6	291	14.5	NA	NA	NA	NA	Yes	No	No	NA	1

Table 5. Diagnostic value of V617F somatic mutation test for myeloproliferative cancer from Neilson et al[72]

	N	Sensitivity (%)	Specificity (%)	Positive Predictive Value (%)	Negative Predictive Value (%)
All participants	49,488	23	100	18	100
Age >70 years	7,958	5	100	6	100
Erythrocyte Count >5.5 $x10^{12}$/l	665	67	99	43	100
Platelet Count >450$x10^{9}$/l	856	47	98	22	99
Leukocyte Count >15 $x10^{9}$/l	144	60	98	50	99
Hematocrit >50%	141	100	100	100	100
Included are prevalent myeloproliferative cancers and such cancer diagnosed within 2 years after blood sampling The diagnostic value for the JAK2 V617F somatic mutation test was examined individually for each substratum, while ignoring the other parameters.					

Table 6. Summary of WHI consent documents pertaining to phenotypic and genetic individual result return

Document	Phenotype	Genetic
Initial Consent (4/01/98)	"Abnormal findings of the following clinic tests will be reported to you, your doctor or your clinic: e.g. high blood pressure or blood test for anemia done at your Clinical Center."	Consent form does not address return of genetic findings
HT Trial Consent (4/01/98)	"Abnormal findings of the clinic tests will be reported to you or your doctor or clinic: blood pressure, blood test for anemia done at your Clinical Center; mammogram, pelvic exam, Pap smear, and electrocardiogram (ECG)." and "You will be informed if definite benefits or harmful results are found during the study"	"You will be informed if definite benefits or harmful results are found during the study."
WHI CaD Consent (4/01/98)	"You will be informed if definite benefits or harmful results are found during the study"	"You will be informed if definite benefits or harmful results are found during the study."
WHI OS Consent (4/01/98)	" Abnormal findings of the clinic tests will be reported to you or your doctor or clinic: blood pressure, blood test for anemia done at your Clinical Center" and " These blood tests will not replace your usual medical care, and results will not be available for your medical care (for example, your cholesterol level will not be reported to you or your doctor). Research studies require only looking at all lab results together, and individual results will not be available."	" …individual results will not be available."
WHI DM Consent (4/01/98)	"Some of the blood drawn will be stored for tests at a later date, including possible genetic studies. This stored blood will be used whether you are in the Clinical Trial or Observational Study. These blood tests will not replace your usual medical care, and results will not be available for your medical care (for example, your cholesterol level will not be reported to you or your doctor). Research studies require only looking at all lab results together, and individual results will not be available."	"…individual results will not be available."
Supplemental Consent (8/1/04)	"All results in the WHI will be kept confidential and no results of genetic or **blood** studies done on your samples will be provided to you, your family or your doctor."	"All results in the WHI will be kept confidential and no results of **genetic** or blood studies done on your samples will be provided to you, your family or your doctor."
Long Life Consent (5/17/12)	"4. What will you do with my blood? Soon after it is collected, we will send one tube of your blood to a hospital lab for a complete blood count (CBC). We will store the rest of your blood sample, and the genetic material in your blood (DNA and RNA), for future research testing. We will give you the results of your CBC, but we will not give you the results of the further research testing of your blood."	"…we will not give you the results of the further research testing of your blood."

Appendix

Supplemental material Part A, Aim I
Further acknowledgements

HeartGO:

Atherosclerosis Risk in Communities (ARIC): NHLBI (N01 HC-55015, N01 HC-55016, N01HC-55017, N01 HC-55018, N01 HC-55019, N01 HC-55020, N01 HC-55021); **Cardiovascular Health Study (CHS)**: NHLBI (HHSN268201200036C, HHSN268200800007C, N01-HC-85239, N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, and grant HL080295), with additional support from NINDS and from NIA (AG-023629, AG-15928, AG-20098, and AG-027058); **Coronary Artery Risk Development in Young Adults (CARDIA)**: NHLBI (N01-HC95095 & N01-HC48047, N01-HC48048, N01-HC48049, and N01-HC48050); **Framingham Heart Study (FHS)**: NHLBI (N01-HC-25195 and grant R01 NS17950) with additional support from NIA (AG08122 and AG033193); **Jackson Heart Study (JHS)**: NHLBI and the National Institute on Minority Health and Health Disparities (N01 HC-95170, N01 HC-95171 and N01 HC-95172); **Multi-Ethnic Study of Atherosclerosis (MESA):** NHLBI (N01-HC-95159 through N01-HC-95169 and RR-024156).

Lung GO:

Cystic Fibrosis (CF): Cystic Fibrosis Foundation (GIBSON07K0, KNOWLE00A0, OBSERV04K0, RDP R026), the NHLBI (R01 HL-068890, R02 HL-095396), NIH National Center for Research Resources (UL1 RR-025014), and the National Human Genome Research Institute (NHGRI) (5R00 HG-004316). **Chronic Obstructive Pulmonary Disease (COPDGene):** NHLBI (U01 HL-089897, U01 HL-089856), and the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, and Sunovian. The COPDGene clinical centers and investigators are available at www.copdgene.org. **Acute Lung Injury (ALI)**: NHLBI (RC2 HL-101779). **Lung Health Study (LHS)**: NHLBI (RC2 HL-066583), the NHGRI (HG-004738), and the NHLBI Division of Lung Diseases (HR-46002). **Pulmonary Arterial Hypertension (PAH):** NIH (P50 HL-084946, K23 AR-52742), and the NHLBI (F32 HL-083714). **Asthma:** NHLBI (RC2 HL-101651), and the NIH (HL-077916, HL-69197, HL-76285, M01 RR-07122).

SWISS and ISGS:

Siblings with Ischemic Stroke Study (SWISS): National Institute of Neurological Disorders and Stroke (NINDS) (R01 NS039987); Ischemic Stroke Genetics Study (ISGS): NINDS (R01 NS042733)

WHISP:

Women's Health Initiative (WHI): The WHI Sequencing Project is funded by the National Heart, Lung, and Blood Institute (HL-102924) as well as the National Institutes of Health (NIH), U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: *https://cleo.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf*

NHLBI GO Exome Sequencing Project

BroadGO

Stacey B. Gabriel (Broad Institute)[4, 5, 11, 16, 17], David M. Altshuler (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[1, 5, 7, 17], Gonçalo R. Abecasis (University of Michigan)[3, 5, 9, 13, 15, 17], Hooman Allayee (University of Southern California)[5], Sharon Cresci (Washington University School of Medicine)[5], Mark J. Daly (Broad Institute, Massachusetts General Hospital), Paul I. W. de Bakker (Broad Institute, Harvard Medical School, University Medical Center Utrecht)[3, 15], Mark A. DePristo (Broad Institute)[4, 13, 15, 16], Ron Do (Broad Institute)[5, 9, 13, 15], Peter Donnelly (University of Oxford)[5], Deborah N. Farlow (Broad Institute)[3, 4, 5, 12, 14, 16, 17], Tim Fennell (Broad Institute), Kiran Garimella (University of Oxford)[4, 16], Stanley L. Hazen (Cleveland Clinic)[5], Youna Hu (University of Michigan)[3, 9, 15], Daniel M. Jordan (Harvard Medical School, Harvard University)[13], Goo Jun (University of Michigan)[13], Sekar Kathiresan (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[5, 8, 9, 12, 14, 15, 17, 20], Hyun Min Kang (University of Michigan)[9, 13, 16], Adam Kiezun (Broad Institute)[5, 13, 15], Guillaume Lettre (Broad Institute, Montreal Heart Institute, Université de Montréal)[1, 2, 13, 15], Bingshan Li (University of Michigan)[3], Mingyao Li (University of Pennsylvania)[5], Christopher H. Newton-Cheh (Broad Institute, Massachusetts General Hospital, Harvard Medical School)[3, 8, 15], Sandosh Padmanabhan (University of Glasgow School of Medicine)[3, 12, 15], Gina Peloso (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[5], Sara Pulit (Broad Institute)[3, 15], Daniel J. Rader (University of Pennsylvania)[5], David Reich (Broad Institute, Harvard Medical School)[15], Muredach P. Reilly (University of Pennsylvania)[5], Manuel A. Rivas (Broad Institute, Massachusetts General Hospital)[5], Steve Schwartz (Fred Hutchinson Cancer Research Center)[5, 12], Laura Scott (University of Michigan)[1], David S. Siscovick (University of Washington)[5, 1, 25], John A. Spertus (University of Missouri Kansas City)[5], Nathan O. Stitziel (Brigham and Women's Hospital)[5, 15], Nina Stoletzki (Brigham and Women's Hospital, Broad Institute, Harvard Medical School)[13], Shamil R. Sunyaev (Brigham and Women's Hospital, Broad Institute, Harvard Medical School)[1, 3, 5, 13, 15], Benjamin F. Voight (Broad Institute, Massachusetts General Hospital), Cristen J. Willer (University of Michigan)[1, 9, 13, 15]

HeartGO

Stephen S. Rich (University of Virginia)[2, 4, 7, 8, 9, 11,14, 15, 17, 18, 31], Ermeg Akylbekova (Jackson State University, University of Mississippi Medical Center)[29], Larry D. Atwood* (Boston University)[1, 11, 28], Christie M. Ballantyne (Baylor College of Medicine, Methodist DeBakey Heart Center)[9, 22], Maja Barbalic (University of Texas Health Science Center Houston)[9, 14, 15, 17, 22], R. Graham Barr (Columbia University Medical Center)[10, 31], Emelia J. Benjamin (Boston University)[14, 20, 28], Joshua Bis (University of Washington)[15, 23], Eric Boerwinkle (University of Texas Health Science Center Houston)[3, 5, 9, 13, 15, 17, 22], Donald W. Bowden (Wake Forest University)[1, 31], Jennifer Brody (University of Washington)[3, 5, 15, 23], Matthew Budoff (Harbor-UCLA Medical Center)[31], Greg Burke (Wake Forest University)[5, 31], Sarah Buxbaum (Jackson State University)[3, 13, 15, 29], Jeff Carr (Wake Forest University)[25, 29, 31], Donna T. Chen (University of Virginia)[6, 11], Ida Y. Chen (Cedars-Sinai Medical Center)[1, 31], Wei-Min Chen (University of Virginia)[13, 15, 18], Pat Concannon (University of Virginia)[11], Jacy Crosby (University of Texas Health Science Center Houston)[22], L. Adrienne Cupples (Boston University)[1, 3, 5, 9, 13, 15, 18, 28], Ralph D'Agostino (Boston University)[28], Anita L. DeStefano (Boston University)[13, 18, 28], Albert Dreisbach (University of Mississippi Medical Center)[3, 29], Josée Dupuis (Boston University)[1, 28], J. Peter Durda (University of Vermont)[15, 23], Jaclyn Ellis (University of North Carolina Chapel Hill)[1], Aaron R. Folsom (University of Minnesota)[5, 22], Myriam Fornage (University of Texas Health Science Center Houston)[3, 18, 25], Caroline S. Fox (National Heart, Lung, and Blood Institute)[1, 28], Ervin Fox (University of Mississippi Medical Center)[3, 9, 29], Vincent Funari (Cedars-Sinai Medical Center)[1, 11, 31], Santhi K. Ganesh (University of Michigan)[2, 22], Julius Gardin (Hackensack University Medical Center)[25], David Goff (Wake Forest University)[25], Ora Gordon (Cedars-Sinai Medical Center)[11, 31], Wayne Grody (University of California Los Angeles)[11, 31], Myron Gross (University of Minnesota)[1, 5, 14, 25], Xiuqing Guo (Cedars-Sinai Medical Center)[3, 15, 31], Ira M. Hall (University of Virginia), Nancy L. Heard-Costa (Boston University)[1, 11, 28], Susan R. Heckbert (University of Washington)[10, 14, 20, 23], Nicholas Heintz (University of Vermont), David M. Herrington (Wake Forest University)[5, 31], DeMarc Hickson (Jackson State University, University of Mississippi Medical Center)[29], Jie Huang (National Heart, Lung, and Blood Institute)[5, 28], Shih-Jen Hwang (Boston University, National Heart, Lung, and Blood Institute)[3, 28], David R. Jacobs (University of Minnesota)[25], Nancy S. Jenny (University of Vermont)[1, 2, 23], Andrew D. Johnson (National Heart, Lung, and Blood Institute)[2, 5, 11, 28], Craig W. Johnson (University of Washington)[15, 31], Steven Kawut (University of Pennsylvania)[10,31], Richard Kronmal (University of Washington)[31], Raluca Kurz (Cedars-Sinai Medical Center)[11, 31], Ethan M. Lange (University of North Carolina Chapel Hill)[3, 5, 9, 13, 34], Leslie A. Lange (University of North Carolina Chapel Hill)[1, 2, 3, 5, 9, 12, 13, 15, 17, 18, 20, 25, 34], Martin G. Larson (Boston University)[3, 15, 28], Mark Lawson (University of Virginia), Cora E. Lewis (University of Alabama at Birmingham)[25,34], Daniel Levy (National Heart, Lung, and Blood Institute)[3, 15, 17, 28], Dalin Li (Cedars-Sinai Medical Center)[11, 15, 31], Honghuang Lin (Boston University)[20, 28], Chunyu Liu (National Heart, Lung, and Blood Institute)[3, 28], Jiankang Liu (University of Mississippi Medical Center)[1, 29], Kiang Liu (Northwestern University)[25], Xiaoming Liu (University of Texas Health Science Center Houston)[15, 22], Yongmei Liu (Wake Forest University)[2, 5, 31], William T. Longstreth (University of Washington)[18, 23], Cay Loria (National Heart, Lung, and Blood Institute)[25], Thomas Lumley (University of Auckland)[9, 23], Kathryn Lunetta (Boston University)[28], Aaron J.

Mackey (University of Virginia)[16, 18], Rachel Mackey (University of Pittsburgh)[1, 23, 31], Ani Manichaikul (University of Virginia)[8, 15, 18, 31], Taylor Maxwell (University of Texas Health Science Center Houston)[22], Barbara McKnight (University of Washington)[15, 23], James B. Meigs (Brigham and Women's Hospital, Harvard Medical School, Massachusetts General Hospital)[1, 28], Alanna C. Morrison (University of Texas Health Science Center Houston)[3, 15, 17], Solomon K. Musani (University of Mississippi Medical Center)[3, 29], Josyf C. Mychaleckyj (University of Virginia)[13, 15, 31], Jennifer A. Nettleton (University of Texas Health Science Center Houston)[9, 22], Kari North (University of North Carolina Chapel Hill)[1, 3, 9, 10, 13, 15, 17, 34], Christopher J. O'Donnell (Massachusetts General Hospital, National Heart, Lung, and Blood Institute)[2, 5, 9, 11, 12, 14, 15, 17, 20, 28], Daniel O'Leary (Tufts University School of Medicine)[25, 31], Frank S. Ong (Cedars-Sinai Medical Center)[3, 11, 31], Walter Palmas (Columbia University)[3, 15, 31], James S. Pankow (University of Minnesota)[1, 22], Nathan D. Pankratz (Indiana University School of Medicine)[15, 25], Shom Paul (University of Virginia), Marco Perez (Stanford University School of Medicine), Sharina D. Person (University of Alabama at Birmingham, University of Alabama at Tuscaloosa)[25], Joseph Polak (Tufts University School of Medicine)[31], Wendy S. Post (Johns Hopkins University)[3, 9, 11, 14, 20, 31], Bruce M. Psaty (Group Health Research Institute, University of Washington)[3, 5, 9, 11, 14, 15, 23], Aaron R. Quinlan (University of Virginia)[18, 19], Leslie J. Raffel (Cedars-Sinai Medical Center)[6, 11, 31], Vasan S. Ramachandran (Boston University)[3, 28], Alexander P. Reiner (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 9, 11, 12, 13, 14, 15, 20, 25, 34], Kenneth Rice (University of Washington)[15, 23], Jerome I. Rotter (Cedars-Sinai Medical Center)[1, 3, 6, 8, 11, 15, 31], Jill P. Sanders (University of Vermont)[23], Pamela Schreiner (University of Minnesota)[25], Sudha Seshadri (Boston University)[18, 28], Steve Shea (Brigham and Women's Hospital, Harvard University)[28], Stephen Sidney (Kaiser Permanente Division of Research, Oakland, CA)[25], Kevin Silverstein (University of Minnesota)[25], David S. Siscovick (University of Washington)[5, 1, 25], Nicholas L. Smith (University of Washington)[2, 15, 20, 23], Nona Sotoodehnia (University of Washington)[3, 15, 23], Asoke Srinivasan (Tougaloo College)[29], Herman A. Taylor (Jackson State University, Tougaloo College, University of Mississippi Medical Center)[5, 29], Kent Taylor (Cedars-Sinai Medical Center)[31], Fridtjof Thomas (University of Texas Health Science Center Houston)[3, 22], Russell P. Tracy (University of Vermont)[5, 9, 11, 12, 14, 15, 17, 20, 23], Michael Y. Tsai (University of Minnesota)[9, 31], Kelly A. Volcik (University of Texas Health Science Center Houston)[22], Chrstina L Wassel (University of California San Diego)[9, 15, 31], Karol Watson (University of California Los Angeles)[31], Gina Wei (National Heart, Lung, and Blood Institute)[25], Wendy White (Tougaloo College)[29], Kerri L. Wiggins (University of Vermont)[23], Jemma B. Wilk (Boston University)[10, 28], O. Dale Williams (Florida International University)[25], Gregory Wilson (Jackson State University)[29], James G. Wilson (University of Mississippi Medical Center)[1, 2, 5, 8, 9, 11, 12, 14, 17, 20, 29], Phillip Wolf (Boston University)[28], Neil A. Zakai (University of Vermont)[2, 23]

ISGS and SWISS

John Hardy (Reta Lila Weston Research Laboratories, Institute of Neurology, University College London)[18], James F. Meschia (Mayo Clinic)[18], Michael Nalls (National Institute on Aging)[2, 18], Stephen S. Rich (University of Virginia)[2, 4, 7, 8, 9, 11, 14, 15, 17, 18, 31], Andrew Singleton (National Institute on Aging)[18], Brad Worrall (University of Virginia)[18]

LungGO

Michael J. Bamshad (Seattle Children's Hospital, University of Washington)[4, 6, 7, 8, 10, 11, 13, 15, 17, 27], Kathleen C. Barnes (Johns Hopkins University)[2, 10, 12, 14, 15, 17, 20, 24, 30, 32], Ibrahim Abdulhamid (Children's Hospital of Michigan)[27], Frank Accurso (University of Colorado)[27], Ran Anbar (Upstate Medical University)[27], Terri Beaty (Johns Hopkins University)[24, 30], Abigail Bigham (University of Washington)[13, 15, 27], Phillip Black (Children's Mercy Hospital)[27], Eugene Bleecker (Wake Forest University)[33], Kati Buckingham (University of Washington)[27], Anne Marie Cairns (Maine Medical Center)[27], Wei-Min Chen (University of Virginia)[13, 15, 18], Daniel Caplan (Emory University)[27], Barbara Chatfield (University of Utah)[27], Aaron Chidekel (A.I. Dupont Institute Medical Center)[27], Michael Cho (Brigham and Women's Hospital, Harvard Medical School)[13, 15, 24], David C. Christiani (Massachusetts General Hospital)[21], James D. Crapo (National Jewish Health)[24, 30], Julia Crouch (Seattle Children's Hospital)6, Denise Daley (University of British Columbia)[30], Anthony Dang (University of North Carolina Chapel Hill)[26], Hong Dang (University of North Carolina Chapel Hill)[26], Alicia De Paula (Ochsner Health System)[27], Joan DeCelie-Germana (Schneider Children's Hospital)[27], Allen Dozor (New York Medical College, Westchester Medical Center)[27,] Mitch Drumm (University of North Carolina Chapel Hill)[26], Maynard Dyson (Cook Children's Med. Center)[27], Julia Emerson (Seattle Children's Hospital, University of Washington)[27], Mary J. Emond (University of Washington)[10, 13, 15, 17, 27], Thomas Ferkol (St. Louis Children's Hospital, Washington University School of Medicine)[27], Robert Fink (Children's Medical Center of Dayton)[27], Cassandra Foster (Johns Hopkins University)[30], Deborah Froh (University of Virginia)[27], Li Gao (Johns Hopkins University)[24, 30, 32], William Gershan (Children's Hospital of Wisconsin)[27], Ronald L. Gibson (Seattle Children's Hospital, University of Washington)[10, 27], Elizabeth Godwin (University of North Carolina Chapel Hill)[26], Magdalen Gondor (All Children's Hospital Cystic Fibrosis Center)[27], Hector Gutierrez (University of Alabama at Birmingham)[27], Nadia N. Hansel (Johns Hopkins University, Johns Hopkins University School of Public Health)[10, 15, 30], Paul M. Hassoun (Johns Hopkins University)[10, 14, 32], Peter Hiatt (Texas Children's Hospital)[27], John E. Hokanson (University of Colorado)[24], Michelle Howenstine (Indiana University, Riley Hospital for Children)[27], Laura K. Hummer (Johns Hopkins University)[32], Seema M. Jamal (University of Washington)[11], Jamshed Kanga (University of Kentucky)[27], Yoonhee Kim (National Human Genome Research Institute)[24, 32], Michael R. Knowles (University of North Carolina Chapel Hill)[10, 26], Michael Konstan (Rainbow Babies & Children's Hospital)[27], Thomas Lahiri (Vermont Children's Hospital at Fletcher Allen Health Care)[27], Nan Laird (Harvard School of Public Health)[24], Christoph Lange (Harvard School of Public Health)[24], Lin Lin (Harvard Medical School)[21], Xihong Lin (Harvard School of Public Health)[21], Tin L. Louie (University of Washington)[13, 15, 27], David Lynch (National Jewish Health)[24], Barry Make (National Jewish Health)[24], Thomas R. Martin (University of Washington, VA Puget Sound Medical Center)[10, 21], Steve C. Mathai (Johns Hopkins University)[32], Rasika A. Mathias (Johns Hopkins University)[10, 13, 15, 30, 32], John McNamara (Children's Hospitals and Clinics of Minnesota)[27], Sharon McNamara (Seattle Children's Hospital)[27], Deborah Meyers (Wake Forest University)[33], Susan Millard (DeVos Children's Butterworth Hospital, Spectrum Health Systems)[27], Peter Mogayzel (Johns Hopkins University)[27], Richard Moss (Stanford University)[27], Tanda Murray (Johns Hopkins University)[30],

Dennis Nielson (University of California at San Francisco)[27], Blakeslee Noyes (Cardinal Glennon Children's Hospital)[27], Wanda O'Neal (University of North Carolina Chapel Hill)[26], David Orenstein (Children's Hospital of Pittsburgh)[27], Brian O'Sullivan (University of Massachusetts Memorial Health Care)[27], Rhonda Pace (University of North Carolina Chapel Hill)[26], Peter Pare (St. Paul's Hospital)[30], H. Worth Parker (Dartmouth-Hitchcock Medical Center, New Hampshire Cystic Fibrosis Center)[27], Mary Ann Passero (Rhode Island Hospital)[27], Elizabeth Perkett (Vanderbilt University)[27,] Adrienne Prestridge (Children's Memorial Hospital)[27], Nicholas M. Rafaels (Johns Hopkins University)[30], Bonnie Ramsey (Seattle Children's Hospital, University of Washington)[27], Elizabeth Regan (National Jewish Health)[24], Clement Ren (University of Rochester)[27], George Retsch-Bogart (University of North Carolina Chapel Hill)[27], Michael Rock (University of Wisconsin Hospital and Clinics)[27], Antony Rosen (Johns Hopkins University)[32], Margaret Rosenfeld (Seattle Children's Hospital, University of Washington)[27], Ingo Ruczinski (Johns Hopkins University School of Public Health)[13, 15, 30], Andrew Sanford (University of British Columbia)[30], David Schaeffer (Nemours Children's Clinic)[27], Cindy Sell (University of North Carolina Chapel Hill)[26], Daniel Sheehan (Children's Hospital of Buffalo)[27], Edwin K. Silverman (Brigham and Women's Hospital, Harvard Medical School)[24, 30], Don Sin (Children's Medical Center of Dayton)[30], Terry Spencer (Elliot Health System)[27], Jackie Stonebraker (University of North Carolina Chapel Hill)[26], Holly K. Tabor (Seattle Children's Hospital, University of Washington)[6, 10, 11, 17, 27], Laurie Varlotta (St. Christopher's Hospital for Children)[27], Candelaria I. Vergara (Johns Hopkins University)[30], Robert Weiss [30], Fred Wigley (Johns Hopkins University)[32], Robert A. Wise (Johns Hopkins University)[30], Fred A. Wright (University of North Carolina Chapel Hill)[26], Mark M. Wurfel (University of Washington)[10, 14, 21], Robert Zanni (Monmouth Medical Center)[27], Fei Zou (University of North Carolina Chapel Hill)[26]

SeattleGO

Deborah A. Nickerson (University of Washington)[3, 4, 5, 7, 8, 9, 11, 15, 17, 18, 19], Mark J. Rieder (University of Washington)[4, 11, 13, 15, 16, 17, 19], Phil Green (University of Washington), Jay Shendure (University of Washington)[1, 8, 14, 16, 17], Joshua M. Akey (University of Washington)[13, 14, 15], Michael J. Bamshad (Seattle Children's Hospital, University of Washington)[4, 6, 7, 8, 10, 11, 13, 15, 17, 27], Kristine L. Bucasas (Baylor College of Medicine)[15], Carlos D. Bustamante (Stanford University School of Medicine)[3, 13, 15], David R. Crosslin (University of Washington)[2, 9], Evan E. Eichler (University of Washington)[19], P. Keolu Fox[2], Wenqing Fu (University of Washington)[13], Adam Gordon (University of Washington)[11], Simon Gravel (Stanford University School of Medicine)[13, 15], Gail P. Jarvik (University of Washington)[9, 15], Jill M. Johnsen (Puget Sound Blood Center, University of Washington)[2], Mengyuan Kan (Baylor College of Medicine)[13], Eimear E. Kenny (Stanford University School of Medicine)[3, 13, 15], Jeffrey M. Kidd (Stanford University School of Medicine)[13, 15], Fremiet Lara-Garduno (Baylor College of Medicine)[15], Suzanne M. Leal (Baylor College of Medicine)[1, 13, 15, 16, 17, 19, 20], Dajiang J. Liu (Baylor College of Medicine)[13, 15], Sean McGee (University of Washington)[13, 15, 19], Timothy D. O'Connor (University of Washington)[13], Bryan Paeper (University of

Washington)[16], Peggy D. Robertson (University of Washington)[4], Joshua D. Smith (University of Washington)[4, 16, 19], Jeffrey C. Staples (University of Washington), Jacob A. Tennessen (University of Washington)[13], Emily H. Turner (University of Washington)[4, 16], Gao Wang (Baylor College of Medicine)[1,13,20], Qian Yi (University of Washington)[4]

WHISP

Rebecca Jackson (Ohio State University)[1, 2, 4, 5, 8, 12, 14, 15, 17, 18, 20, 34], Kari North (University of North Carolina Chapel Hill)[1, 3, 9, 10, 13, 15, 17, 34,] Ulrike Peters (Fred Hutchinson Cancer Research Center)[1, 3, 11, 12, 13, 15, 17, 18, 34], Christopher S. Carlson (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 12, 13, 14, 15, 16, 17, 18, 19, 34], Garnet Anderson (Fred Hutchinson Cancer Research Center)[34], Hoda Anton-Culver (University of California at Irvine)[34], Themistocles L. Assimes (Stanford University School of Medicine)[5, 9, 11, 34], Paul L. Auer (Fred Hutchinson Cancer Research Center)[1, 2, 3, 5, 11, 12, 13, 15, 16, 18, 34], Shirley Beresford (Fred Hutchinson Cancer Research Center)[34], Chris Bizon (University of North Carolina Chapel Hill)[3, 9, 13, 15, 34], Henry Black (Rush Medical Center)[34], Robert Brunner (University of Nevada)[34], Robert Brzyski (University of Texas Health Science Center San Antonio)[34], Dale Burwen (National Heart, Lung, and Blood Institute WHI Project Office)[34], Bette Caan (Kaiser Permanente Division of Research, Oakland, CA)[34], Cara L. Carty (Fred Hutchinson Cancer Research Center)[18, 34], Rowan Chlebowski (Los Angeles Biomedical Research Institute)[34], Steven Cummings (University of California at San Francisco)[34], J. David Curb* (University of Hawaii)[9, 18, 34], Charles B. Eaton (Brown University, Memorial Hospital of Rhode Island)[12, 34], Leslie Ford (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[34], Nora Franceschini (University of North Carolina Chapel Hill)[2, 3, 9, 10, 15, 34], Stephanie M. Fullerton (University of Washington)[6, 11, 34], Margery Gass (University of Cincinnati)[34], Nancy Geller (National Heart, Lung, and Blood Institute WHI Project Office)[34], Gerardo Heiss (University of North Carolina Chapel Hill)[5, 34], Barbara V. Howard (Howard University, MedStar Research Institute)34, Li Hsu (Fred Hutchinson Cancer Research Center)[1, 13, 15, 18, 34], Carolyn M. Hutter (Fred Hutchinson Cancer Research Center)[13, 15, 18, 34], John Ioannidis (Stanford University School of Medicine)[11, 34], Shuo Jiao (Fred Hutchinson Cancer Research Center)[34], Karen C. Johnson (University of Tennessee Health Science Center)[3, 34], Charles Kooperberg (Fred Hutchinson Cancer Research Center)[1, 5, 9, 13, 14, 15, 17, 18, 34], Lewis Kuller (University of Pittsburgh)[34], Andrea LaCroix (Fred Hutchinson Cancer Research Center)[34], Kamakshi Lakshminarayan (University of Minnesota)[18, 34], Dorothy Lane (State University of New York at Stony Brook)[34], Ethan M. Lange (University of North Carolina Chapel Hill)[3, 5, 9, 13, 34], Leslie A. Lange (University of North Carolina Chapel Hill)[1, 2, 3, 5, 9, 12, 13, 15, 17, 18, 20, 25, 34], Norman Lasser (University of Medicine and Dentistry of New Jersey)[34], Erin LeBlanc (Kaiser Permanente Center for Health Research, Portland, OR)[34], Cora E. Lewis (University of Alabama at Birmingham)[25,34], Kuo-Ping Li (University of North Carolina Chapel Hill)[9, 34], Marian Limacher (University of Florida)[34], Dan-Yu Lin (University of North Carolina Chapel Hill)[1, 3, 9, 13, 15, 34], Benjamin A. Logsdon (Fred Hutchinson Cancer Research Center)[2, 34], Shari Ludlam (National Heart, Lung, and Blood Institute WHI Project Office)[34], JoAnn E. Manson (Brigham and Women's Hospital, Harvard School of Public Health)[34], Karen Margolis (University of Minnesota)[34], Lisa Martin (George

Washington University Medical Center)[9, 34], Joan McGowan (National Heart, Lung, and Blood Institute WHI Project Office)[34], Keri L. Monda (Amgen, Inc.)[1, 15, 34], Jane Morley Kotchen (Medical College of Wisconsin)[34], Lauren Nathan (University of California Los Angeles)[34], Judith Ockene (Fallon Clinic, University of Massachusetts)[34], Mary Jo O'Sullivan (University of Miami)[34], Lawrence S. Phillips (Emory University)[34], Ross L. Prentice (Fred Hutchinson Cancer Research Center)[34], Alexander P. Reiner (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 9, 11, 12, 13, 14, 15, 20, 25, 34], John Robbins (University of California at Davis)[34], Jennifer G. Robinson (University of Iowa)[9, 11, 18, 34], Jacques E. Rossouw (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[5, 14, 17, 20, 34], Haleh Sangi-Haghpeykar (Baylor College of Medicine)[34], Gloria E. Sarto (University of Wisconsin)[34], Sally Shumaker (Wake Forest University)[34], Michael S. Simon (Wayne State University)[34], Marcia L. Stefanick (Stanford University School of Medicine)[34], Evan Stein (Medical Research Labs)[34], Hua Tang (Stanford University)[2, 34], Kira C. Taylor (University of Louisville)[1, 3, 13, 15, 20, 34], Cynthia A. Thomson (University of Arizona)[34], Timothy A. Thornton (University of Washington)[13, 15, 18, 34], Linda Van Horn (Northwestern University)[34], Mara Vitolins (Wake Forest University)[34], Jean Wactawski-Wende (University of Buffalo)[34], Robert Wallace (University of Iowa)[2, 34], Sylvia Wassertheil-Smoller (Boston University)[18, 34], Donglin Zeng (University of North Carolina Chapel Hill)[9, 34]

*deceased

NHLBI GO ESP Project Team

Deborah Applebaum-Bowden (National Heart, Lung, and Blood Institute)[4, 7, 12, 17], Michael Feolo (National Center for Biotechnology Information)[12], Weiniu Gan (National Heart, Lung, and Blood Institute)[7, 8, 16, 17], Dina N. Paltoo (National Heart, Lung, and Blood Institute)[4, 6, 11, 17], Jacques E. Rossouw (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[5, 14, 17, 20, 34], Phyliss Sholinsky (National Heart, Lung, and Blood Institute)[4, 12, 17], Anne Sturcke (National Center for Biotechnology Information)[12]

ESP Groups

[1]Anthropometry Project Team, [2]Blood Count/Hematology Project Team, [3]Blood Pressure Project Team, [4]Data Flow Working Group, [5]Early MI Project Team, [6]ELSI Working Group, [7]Executive Committee, [8]Family Study Project Team, [9]Lipids Project Team, [10]Lung Project Team, [11]Personal Genomics Project Team, [12]Phenotype and Harmonization Working Group, [13]Population Genetics and Statistical Analysis Working Group, [14]Publications and Presentations Working Group, [15]Quantitative Analysis Ad Hoc Task Group, [16]Sequencing and Genotyping Working Group, [17]Steering Committee, [18]Stroke Project Team, [19]Structural Variation Working Group, [20]Subclinical/Quantitative Project Team

ESP Cohorts

[21]Acute Lung Injury (ALI), [22]Atherosclerosis Risk in Communities (ARIC), [23]Cardiovascular Health Study (CHS), [24]Chronic Obstructive Pulmonary Disease (COPDGene), [25]Coronary Artery Risk Development in Young Adults (CARDIA), [26]Cystic Fibrosis (CF), [27]Early Pseudomonas Infection Control (EPIC), [28]Framingham Heart Study (FHS), [29]Jackson Heart Study (JHS), [30]Lung Health Study (LHS), [31]Multi-Ethnic Study of Atherosclerosis (MESA), [32]Pulmonary Arterial Hypertension (PAH), [33]Severe Asthma Research Program (SARP), [34]Women's Health Initiative (WHI)

Supplemental Tables

Table S1. Reported CRP-associated candidate genes from published Genome Wide Association Studies

Gene/Region	Study Reporting Association	Population(s) Included in Studies
APOC1/APOE	Dehghan et al.[1], Elliott et al.[2], Ridker et al.[3]	European American & Asian Indian
ASCL1	Dehghan et al.[1], Ridker et al.[3]	European American
BCL7B	Dehghan et al.[1]	European American
CRP	Dehghan et al.[1], Reiner et al.[4], Reiner et al.[5], Doumatey et al.[6], Wu et al.[7], Okada et al.[8], Elliott et al.[2], Ridker et al.[3], Curocichin et al.[9]	European American, African American, Hispanic American, South Asian, Japanese, Asian Indian, Filipino
DOCK4	Kong et al.[10]	Korean
FAM13C	Reiner et al.[4]	European American
GCKR	Ridker et al.[3]	European American
HNF1A	Dehghan et al.,[1] Reiner et al.[4], Reiner et al.[5], Wu et al.[7], Okada et al.[8], Elliott et al.[2], Ridker et al.[3], Kong et al.[10]	European American, African American, Hispanic American, South Asian, Japanese, Asian Indian, Korean
HNF4A	Dehghan et al.[1]	European American
IL1F10/ IL1RN	Dehghan et al.[1], Reiner et al.[5]	European American, African American
IL6	Okada et al.[8]	Japanese
IL6R	Dehghan et al.[1], Elliott et al.[2], Ridker et al.[3], Curocichin et al.[9]	European American, Asian Indian, Filipino
LEPR	Dehghan et al.[1], Reiner et al.[5], Elliott et al.[2], Ridker et al.[3]	European American, Hispanic American, Asian Indian
NLPR3	Dehghan et al.[1]	European American
PABPC4	Dehghan et al.[1]	European American
PPP1R3B	Dehghan et al.[1]	European American
PSMG1	Dehghan et al.[1]	European American
RGS6	Dehghan et al.[1], Kong et al.[10]	European American, Korean
RORA	Dehghan et al.[1]	European American
SALL1	Dehghan et al.[1]	European American
SLC1A3	Reiner et al.[4]	European American
TOMM40	Reiner et al.[5]	African American
TREM2	Reiner et al. [5]	African American

Table S2. Description of the 7 contributing cohort studies

Study	Race	N	Age Range at Baseline	CRP Assay
ARIC	EA, AA	4,827	45- 64 years	immunoturbidimetric CRP-Latex (II) high-sensitivity assay from Denka Seiken (Tokyo, Japan) on a Hitachi 911 analyzer (Roche Diagnostics, Indianapolis, Indiana)
CARDIA	EA	190	18-30 years	high-sensitivity nephelometry-based methods (BNII Nephelometer 100 Analyzer, Dade Behring)
CHS	EA	946	≥65 years	high-sensitivity enzyme-linked immunosorbent assay
FHS	EA	1,144	12-75 years	high-sensitivity assay (Dade Behring BN100)
JHS	AA	346	21-84 years	immunoturbidimetric CRP-Latex assay (Kamiya Biomedical Company, Seattle, Washington) on a Hitachi 911 analyzer (Roche Diagnostics, Indianapolis, Indiana)
MESA	EA, AA	399	45-84 years	BNII nephelometer (N High Sensitivity CRP, Dade Behring Inc, Deerfield, IL)
WHI	EA, AA	1,307	50-79 years	latex-particle enhanced immunoturbidimetric assay kit (Roche Diagnostics, Indianapolis, IN)

Abbreviations: European American (EA); African American (AA); Atherosclerosis risk in Communities (ARIC); Framingham Heart Study (FHS), Coronary Heart Study (CHS); Women's Health Initiative (WHI); Jackson Heart Study (JHS)

Table S3. Summary of reported tests and significance levels.

Test (variants included)	Analysis	Variant allele frequency	Number of Tests Performed	Correction	Threshold for Significance
Single-variant (target sequence variants)	Exome-wide	≥5 minor alleles	639,770*	GWAS threshold	5.00×10^{-8}
	Candidate Gene	MAF <0.05	267*	Bonferroni	1.87×10^{-4}
SKAT (stop-gain, stop-loss, splicing, nonsynonymous, noncoding RNA splicing)	Exome-wide	MAF <0.05	19,230	Bonferroni	2.50×10^{-6}
	Candidate Gene	MAF <0.05	25	Bonferroni	2.00×10^{-3}
T1 (stop-gain, stop-loss, splicing, nonsynonymous, noncoding RNA splicing)	Exome-wide	MAF <0.01	19,230	Bonferroni	2.50×10^{-6}
	Candidate Gene	MAF<0.01	25	Bonferroni	2.00×10^{-3}

*Number of tests from the combined sample; race-specific estimates are a subset of the tests performed in the combined sample.
Abbreviations: Sequence Kernel Association Test (SKAT); burden (T1); minor allele frequency (MAF); Genome Wide Association Study (GWAS)

Table S4. Characteristics of Discovery and Replication Samples

	Exome Sequencing Discovery						**Exome Chip Replication**		
	ESP**		**CHARGE ARIC**		**CHARGE FHS**	**CHARGE CHS**	**WHI**		**JHS**
	EA	**AA**	**EA**	**AA**	**EA**	**AA**	**EA**	**AA**	**AA**
N	1,832	1,528	2,718	1,581	757	743	11,414	2,380	2,201
Mean Age, years (SD)	61.8 (11.1)	58.9 (10.2)	63.0 (5.6)	61.6 (5.6)	60.2 (12.0)	72.9 (5.7)	67.5 (6.3)	66.7 (5.7)	52.8 (12.7)
Female %	57.4%	76.2%	54.3%	66.0%	51.1%	53%	100%	100%	62.8%
Mean BMI kg/m^2 (SD)	27.5 (5.3)	34.1 (10.5)	28.4 (5.2)	30.7 (6.5)	28.7 (5.6)	26.8 (4.5)	28.5 (5.7)	30.1 (5.4)	31.4 (6.4)
Smoker N (%)	412 (22.5%)	276 (18.2%)	389 (14.3%)	262 (16.9%)	109 (14.4%)	85 (11.0%)	888 (7.9%)	222 (9.6%)	706 (32.2%)
Diabetes N (%)	170 (8.4%)	349 (22.9%)	341 (12.6%)	385 (24.7%)	115 (15.2%)	100 (13.0%)	559 (4.9%)	262 (11.0%)	354 (16.3%)
Hypertension N %	752 (41.1%)	843 (55.3%)	1,143 (42.3%)	1,033 (65.6%)	359 (47.4%)	416 (56.0%)	3829 (33.5%)	1041 (43.7%)	1302 (59.5%)
Median CRP mg/L (SD)	2.4 (6.4)	4.0 (8.5)	2.2 (5.8)	3.8 (7.4)	2.6 (7.5)	2.5 (8.7)	2.4 (6.0)	3.2 (6.8)	2.6 (8.0)

*Missing values may lead to percents that do not sum to expected values.

**Composed of samples from ARIC, CHS, FHS, JHS, MESA, and WHI.

Abbreviations: number (N); standard deviation(SD); body mass index (BMI); C-reactive Protein (CRP); European American (EA); African American (AA); Exome Sequencing Project (ESP); Cohorts for Hearth and Aging Research in Genomic Epidemiology (CHARGE); Framingham Heart Study (FHS), Coronary Heart Study (CHS); Women's Health Initiative (WHI); Jackson Heart Study (JHS)

Table S5. Characteristics of uniquely annotated variants (>=5 minor alleles represented in EA and/or AA) included in CHARGE-ESP single variant discovery analyses

		AA	**EA**	**AA+EA**
Variants	N	489,466	330,620	639,770
Minor Allele Frequency	<1% N	269,989	209,511	468,421
	1-5% N	112,498	45,433	86,976
	>5% N	106,979	75,676	84,373
Annotation	downstream N	2,950	1,983	3,726
	exonic N	2,134	1,715	2,995
	intergenic N	747	1,229	1,678
	intronic N	239,117	152,721	304,241
	nonsynonymous N	106,698	81,325	150,761
	splicing N	649	572	1,010
	stopgain N	1,182	1,122	1,959
	stoploss N	97	60	129
	synonymous N	98,528	65,133	125,741
	upstream N	3,610	2,341	4,486
	UTR 3' N	18,264	12,141	23,841
	UTR 5' N	8,746	5,551	11,350
	ncRNA_intronic N	4,306	3,021	4,969
	ncRNA_exonic N	2,427	1,696	2,870
	ncRNA_splicing N	11	10	14

Abbreviations: Number (N); African American (AA); European American; untranslated region (UTR); noncoding RNA (ncRNA)

Table S6. Intronic Variants reaching exome wide significance ($P<5x10^{-8}$) in the discovery sample

Gene (hg19 location); rsID	Study Reported (GWAS variant)	N (all, EA, AA)	MAF (all, EA, AA)	β (SE)	P	Het. I2 (P)	β (SE)	P	Het. I2 (P)	β (SE)	P	Het. I2 (P)
CRP (chr1:159684186); rs1417938	Ridker et al.[3] (same SNP)	9159, 6050, 3109	0.250, 0.310, 0.134	0.11 (0.018)	$8.15x10^{-10}$	0 (0.50)	0.12 (0.020)	$7.27x10^{-09}$	0 (0.56)	0.086 (0.039)	$2.92x10^{-02}$	45.3 (0.18)
LEPR (chr1:66085574); rs3790439	Elliott et al.[2] (rs6700896, r^2=0.97, D'=1)	3159, 6050, 3109	0.431, 0.371, 0.451	-0.095 (0.016)	$1.22x10^{-09}$	62.8 (0.020)	-0.10 (0.019)	$5.91x10^{-08}$	74.4 (0.0084)	-0.077 (0.027)	$4.04x10^{-03}$	0 (1)
LEPR (chr1:66088701); rs12067936	Elliott et al.[2] (proxy rs6700896, r^2=0.97, D'=1	3159, 6050, 3109	0.432, 0.367, 0.450	-0.092 (0.016)	$4.75x10^{-09}$	63.7 (0.017)	-0.10 (0.019)	$1.20x10^{-07}$	74.4 (0.0084)	-0.071 (0.027)	$7.54x10^{-03}$	0 (1)
HNF1A (chr12:121426594); rs1169294	Dehghan et al.[1] (rs1183910, r^2=0.923, D'=0.961)	9159, 6050, 3109	0.266, 0.321, 0.157	-0.111 (0.018)	$2.15x10^{-10}$	0 (0.79)	-0.121 (0.020)	$2.46x10^{-09}$	0 (0.84)	-0.078 (0.036)	$3.16x10^{-02}$	0 (0.53)
HNF1A (chr12:121431300); rs1169301	Ridker et al.[3] (rs1169300 , r^2=1, D'=1)	9159, 6050, 3109	0.246, 0.307, 0.126	-0.111 (0.018)	$9.2x10^{-10}$	0 (0.79)	-0.118 (0.020)	$5.10x10^{-09}$	0 (0.62)	-0.086 (0.04)	$4.21x10^{-02}$	0 (0.79)
HNF1A (chr12:121434833); rs2259852	Reiner et al.[5] (rs2259816 , r^2=1, D'=1)	5799, 4218, 1581	0.307, 0.369, 0.143	-0.123 (0.021)	$6.64x10^{-09}$	0 (0.46)	-0.129 (0.023)	$2.07x10^{-08}$	2.6 (0.36)	-0.088 (0.053)	$9.87x10^{-02}$	0 (1)
HNF1A (chr12:121435475); rs2464195	Reiner et al.[5] (rs2259816 , r^2=1, D'=1)	5799, 4218, 1581	0.300, 0.371, 0.141	-0.118 (0.017)	$4.54x10^{-09}$	0 (0.76)	-0.121 (0.019)	$3.48x10^{-10}$	0 (0.37)	-0.094 (0.053)	$7.70x10^{-02}$	0 (1)
HNF1A (chr12:121435587); rs2259816	Reiner et al.[5] (same SNP)	5799, 4218, 1581	0.310, 0.373, 0.143	-0.126 (0.021)	$2.56x10^{-09}$	0 (0.51)	-0.132 (0.023)	$9.99x10^{-09}$	0 (0.39)	-0.094 (0.053)	$7.94x10^{-02}$	0 (1)
HNF1A (chr12:121438844); rs735396	Reiner et al.[5] (rs2259816 proxy, r^2=1, D'=1)	9159, 6050, 3109	0.282, 0.354, 0.141	-0.124 (0.018)	$2.00x10^{-12}$	0 (0.53)	-0.131 (0.020)	$3.45x10^{-11}$	7.3 (0.36)	-0.095 (0.038)	$1.28x10^{-02}$	0 (0.71)
TOMM40 (chr19:45396219); rs157582	Not Reported	9159, 6050, 3109	0.299, 0.209, 0.474	-0.136 (0.018)	$8.76x10^{-14}$	9 (0.36)	-0.167 (0.023)	$8.04x10^{-13}$	0 (0.73)	-0.089 (0.027)	$4.84x10^{-04}$	0 (0.99)
TOMM40 (chr19:45404431); rs741780	Reiner et al[5] (rs1160985, r^2=1, D'=1)	9159, 6050, 3109	0.493, 0.441, 0.367	0.083 (0.016)	$1.31x10^{-07}$	65.5 (0.013)	0.046 (0.019)	$1.43x10^{-02}$	0 (0.70)	0.166 (0.028)	$3.48x10^{-09}$	0 (0.50)

Abbreviations: nonsynonymous (nSyn); synonymous (syn);minor allele frequency (MAF); Heterogeneity (Het. I^2); European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); European American (EA); African American (AA);Standard Error(SE);P-value (P); Beta (β);reference SNP ID number (rsID)

Table S7. Cohort-specific estimates of effect variants reported in Tables 1, 2 and S4

		ARIC (N_{AA}=1581;N_{EA}=2718)				ESP (N_{AA}=1528;N_{EA}=1832)				FHS (N_{EA}=757)		CHS (N_{EA}=743)	
Name	**Gene (function)**	**AA β (SE)**	**AA P**	**EA β (SE)**	**EA P**	**AA β (SE)**	**AA P**	**EA β (SE)**	**EA P**	**EA β (SE)**	**EA P**	**EA β (SE)**	**EA P**
chr1:154426970	*IL6R* (nSyn)	-0.052 (0.058)	3.70E-01	-0.131 (0.028)	2.49E-06	-0.122 (0.055)	2.65E-02	-0.129 (0.037)	4.43E-04	-0.117 (0.056)	3.68E-02	-0.094 (0.052)	7.21E-02
chr1:159683438	*CRP* (Syn)	-0.617 (0.19)	1.15E-03	-0.286 (0.058)	7.12E-07	-0.539 (0.188)	4.18E-03	-0.219 (0.073)	2.86E-03	-0.202 (0.116)	8.19E-02	-0.385 (0.105)	2.62E-04
chr1:159683814	*CRP* (nSyn)	-2.247 (1.036)	3.02E-02	-0.577 (0.243)	1.78E-02	-1.96 (0.744)	8.45E-03	-1.459 (0.478)	2.26E-03	-0.715 (0.473)	1.30E-01	<NA>	<NA>
chr1:159684186	*CRP* (noncoding)	0.138 (0.055)	1.23E-02	0.133 (0.029)	4.48E-06	0.032 (0.056)	5.69E-01	0.121 (0.038)	1.57E-03	0.114 (0.06)	5.57E-02	0.039 (0.059)	5.16E-01
chr1:66085574	*LEPR* (noncoding)	-0.05 (0.038)	1.80E-01	-0.073 (0.028)	9.07E-03	<NA>	<NA>	-0.194 (0.037)	1.71E-07	-0.139 (0.056)	1.28E-02	0.007 (0.054)	9.00E-01
chr1:66088701	*LEPR* (noncoding)	-0.038 (0.038)	3.09E-01	-0.07 (0.028)	1.39E-02	<NA>	<NA>	-0.193 (0.037)	2.25E-07	-0.139 (0.056)	1.27E-02	0.005 (0.054)	9.24E-01
chr1:66102257	*LEPR* (Syn)	-0.084 (0.037)	2.46E-02	-0.073 (0.028)	9.79E-03	<NA>	<NA>	-0.183 (0.037)	5.58E-07	-0.152 (0.056)	6.42E-03	-0.003 (0.054)	9.49E-01
chr12:121416622	*HNF1A* (Syn)	-0.033 (0.04)	4.04E-01	-0.102 (0.027)	1.81E-04	<NA>	<NA>	<NA>	<NA>	-0.183 (0.055)	7.92E-04	-0.156 (0.053)	3.15E-03
chr12:121416650	*HNF1A* (nSyn)	-0.052 (0.058)	3.67E-01	-0.107 (0.029)	2.19E-04	<NA>	<NA>	<NA>	<NA>	-0.155 (0.057)	6.45E-03	-0.101 (0.055)	6.85E-02
chr12:121426594	*HNF1A* (noncoding)	-0.055 (0.052)	2.87E-01	-0.108 (0.03)	3.30E-04	-0.101 (0.051)	4.71E-02	-0.123 (0.038)	1.11E-03	-0.16 (0.058)	6.14E-03	-0.149 (0.058)	9.78E-03
chr12:121431300	*HNF1A* (noncoding)	-0.075(0.056)	1.77E-01	-0.101 (0.03)	8.10E-04	-0.097 (0.057)	8.93E-02	-0.102 (0.038)	6.79E-03	-0.16 (0.058)	5.91E-03	-0.166 (0.055)	2.73E-03
chr12:121434833	*HNF1A* (noncoding)	-0.088 (0.053)	9.87E-02	-0.105 (0.028)	2.17E-04	<NA>	<NA>	<NA>	<NA>	-0.177 (0.058)	2.21E-03	-0.173 (0.054)	1.24E-03
chr12:121435342	*HNF1A* (Syn)	-0.065(0.056)	2.44E-01	-0.101 (0.03)	6.26E-04	-0.069 (0.059)	2.41E-01	-0.121 (0.038)	1.56E-03	-0.168 (0.058)	3.77E-03	-0.153 (0.055)	5.09E-03
chr12:121435427	*HNF1A* (nSyn)	-0.061 (0.056)	2.76E-01	-0.098 (0.03)	9.46E-04	-0.081 (0.058)	1.64E-01	-0.108 (0.039)	4.93E-03	-0.158 (0.058)	6.46E-03	-0.152 (0.055)	5.30E-03
chr12:121435475	*HNF1A* (noncoding)	-0.094 (0.053)	7.94E-02	-0.106 (0.028)	2.00E-04	<NA>	<NA>	<NA>	<NA>	-0.176 (0.057)	2.01E-03	-0.171 (0.053)	1.33E-03
chr12:121435587	*HNF1A* (noncoding)	-0.094 (0.053)	7.94E-02	-0.109 (0.028)	1.28E-04	<NA>	<NA>	<NA>	<NA>	-0.176 (0.057)	2.01E-03	-0.173 (0.053)	1.20E-03
chr12:121438844	*HNF1A* (noncoding)	-0.081 (0.053)	1.30E-01	-0.115 (0.029)	7.18E-05	-0.109 (0.054)	4.46E-02	-0.107 (0.036)	3.26E-03	-0.188 (0.057)	1.04E-03	-0.203 (0.058)	5.10E-04
chr19:45395714	*TOMM40* (Syn)	-0.097 (0.038)	9.68E-03	-0.186 (0.033)	1.57E-08	-0.089 (0.038)	1.99E-02	-0.162 (0.044)	2.20E-04	-0.115 (0.07)	9.81E-02	-0.11 (0.073)	1.32E-01
chr19:45396144	*TOMM40* (Syn)	0.013 (0.056)	8.21E-01	-0.202 (0.038)	9.47E-08	-0.051 (0.056)	3.68E-01	-0.176 (0.052)	6.77E-04	-0.13 (0.086)	1.31E-01	-0.145 (0.087)	9.36E-02
chr19:45396219	*TOMM40* (noncoding)	-0.089 (0.038)	1.80E-02	-0.189 (0.033)	1.15E-08	-0.09 (0.039)	2.14E-02	-0.15 (0.045)	8.59E-04	-0.111 (0.07)	1.14E-01	-0.145 (0.07)	3.80E-02
chr19:45397307	*TOMM40* (Syn)	-0.424 (0.096)	9.69E-06	-0.159 (0.085)	6.03E-02	-0.304 (0.102)	2.84E-03	-0.365 (0.11)	9.36E-04	-0.445 (0.175)	1.10E-02	-0.303 (0.17)	7.44E-02
chr19:45404431	*TOMM40* (noncoding)	0.185 (0.04)	3.90E-06	0.025 (0.028)	3.63E-01	0.147 (0.04)	2.48E-04	0.064 (0.035)	6.96E-02	0.087 (0.055)	1.09E-01	0.044 (0.052)	4.01E-01
chr19:45411941	*APOE* (nSyn)	<NA>	<NA>	<NA>	<NA>	-0.24 (0.049)	7.03E-08	-0.31 (0.058)	1.52E-06	<NA>	<NA>	<NA>	<NA>

Abbreviations: nonsynonymous (nSyn); synonymous (syn);minor allele frequency (MAF); Heterogeneity (Het. I^2); European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); reference SNP ID number (rsID)

Table S8. Conditional Analysis of rs77832441 with reported GWAS variants in ESP and Exome Array EA samples

			rs77832441 without adjustment for reported CRP variants		rs77832441 with adjustment for reported CRP variants	
Sample	**CRP-associated variants in dataset**	**N**	**β (SE)**	**P**	**β (SE)**	**P**
ESP EA	rs1417938, rs180947	1,832	-1.46 (0.48)	$2.24x10^{-03}$	-1.57 (0.47)	$9.11x10^{-04}$
Exome Array EA	rs3093059	11,414	-0.90 (0.11)	$3.00x10^{-15}$	-0.88 (0.11)	$5.97x10^{-15}$

Abbreviations: European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β)

Table S9. CEU 1000 genomes panel estimates of linkage disequilibrium between reported Genome Wide Association Study common polymorphisms and rs77832441 from SNAP pairwise LD query

Novel variant	**Reported common variant**	**Distance**	**r^2**	**D'**
rs77832441	rs3116636	2669	0.02	1
rs77832441	rs3122012	5509	0.019	1
rs77832441	rs1417938	372	0.018	1
rs77832441	rs3091244	851	0.017	1
rs77832441	rs3116656	8558	0.017	1
rs77832441	rs3116653	13096	0.017	1
rs77832441	rs3116651	14671	0.017	1
rs77832441	rs12093699	35826	0.017	1
rs77832441	rs2592887	30875	0.006	1
rs77832441	rs2027471	5574	0.004	1
rs77832441	rs1205	1581	0.003	1
rs77832441	rs2794520	4998	0.003	1
rs77832441	rs7553007	14735	0.003	1
rs77832441	rs12744244	36158	0.002	1
rs77832441	rs3093059	1322	0.001	1
rs77832441	rs3093068	2450	0.001	1
rs77832441	rs3093075	3901	0.001	1
rs77832441	rs12068753	8723	0.001	1
rs77832441	rs11265260	16225	0.001	1
rs77832441	rs1800947	376	0	1

Table S10. Gene-based test results reaching candidate gene level significance ($P< 2.0x10^{-3}$)

			Discovery			Replication	
Test (MAF)	**Analysis**	**Gene**	**P (EA)**	**P (AA)**	**P (EA+AA)**	**P (EA)**	**P (AA)**
T1 (≤1%)	Candidate Gene	*CRP*	$6.80x10^{-04}$	$1.05x10^{-03}$	$2.36x10^{-06}$	$2.54x10^{-14}$	$5.07x10^{-01}$
SKAT (≤5%)	Candidate Gene	*CRP*	$1.71x10^{-04}$	$5.97x10^{-02}$	$3.37x10^{-06}$	$3.21x10^{-15}$	$6.46x10^{-01}$
SKAT (≤5%)	Candidate Gene	*RORA*	$8.34x10^{-01}$	$1.73x10^{-03}$	$7.76x10^{-02}$	$2.79x10^{-01}$	$6.22x10^{-02}$

Abbreviations: European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); Sequence Kernel Association Test (SKAT); burden (T1)

Table S11. Gene-based test results for the CRP Genome Wide Association Study loci

	SKAT (MAF<5%)			T1 (MAF<1%)		
Gene	**P (EA)**	**P (AA)**	**P (EA+AA)**	**P (EA)**	**P (AA)**	**P (EA+AA)**
APOC1	0.768	0.242	0.648	0.667	0.463	0.645
APOE	0.088	0.236	0.109	0.018	0.074	0.004
ASCL1	0.725	0.1	0.318	0.545	0.875	0.572
BCL7B	0.977	0.412	0.871	0.509	0.321	0.925
CRP	$1.71 x10^{-4}$	0.06	$3.37 x10^{-6}$	$6.80 x10^{-4}$	$1.05 x10^{-5}$	$2.36 x10^{-6}$
DOCK4	0.369	0.554	0.322	0.823	0.346	0.492
FAM13C	0.795	0.895	0.96	0.327	0.748	0.805
GCKR	0.169	0.37	0.114	0.563	0.371	0.948
HNF1A	0.04	0.013	0.015	0.661	0.035	0.316
HNF4A	0.008	0.144	0.003	0.886	0.344	0.932
IL1F10	0.579	0.214	0.61	0.344	0.028	0.065
IL1RN	0.71	0.132	0.201	0.854	0.302	0.394
IL6	0.706	0.59	0.371	0.653	0.506	0.241
IL6R	0.283	0.45	0.739	0.929	0.709	0.501
LEPR	0.647	0.614	0.69	0.509	0.466	0.284
NLRP3	0.682	0.274	0.597	0.489	0.112	0.147
PABPC4	0.887	0.324	0.547	0.913	0.098	0.171
PPP1R3B	0.854	0.925	0.854	0.923	0.141	0.411
PSMG1	0.864	0.085	0.239	0.524	0.019	0.181
RGS6	0.439	0.299	0.465	0.314	0.23	0.936
RORA	0.834	$1.73 x10^{-3}$	0.078	0.688	0.15	0.618
SALL1	0.427	0.006	0.043	0.308	0.749	0.042
SLC1A3	0.733	0.056	0.135	0.216	0.419	0.178
TOMM40	0.28	0.566	0.535	0.678	0.261	0.264
TREM2	0.284	0.566	0.448	0.456	0.868	0.406

Abbreviations: European American (EA); African American (AA); Standard Error (SE); P-value (P); Beta (β); Sequence Kernel Association Test (SKAT); burden (T1)

Table S12. Summary of previously reported *CRP* alleles

SNP	Genomic Context	A1/ A2	Location (hg19)	CEU tagSNP (r^2)	YRI tagSNP (r^2)	1000 Genomes MAF CEU/YRI (Allele)	Effect EA	Effect AA
rs3116653	intergenic	G*/C	1: 159698485	rs3116653 (Same SNP)	rs3116653 (Same SNP)	33%(G) / 1.7%(G)	↑	NR
rs3116656	intergenic	C*/T	1: 159692372	rs3116653 (0.96)	rs3116653 (1)	33%(C) / 1.7%(C)	↑	NR
rs3116651	intergenic	A*/G	1: 159698485	rs3116653 (0.96)	NA	33%(A) / NA	↑	NR
rs1417938	intronic	A*/T	1: 159684186	rs3116653 (0.93)	rs1417938 (Same SNP)	32%(A) /6.8%(A)	↑	NR
rs3122012	intergenic	C*/T	1: 159689323	rs3116653 (0.89)	rs1417938 (0.71)	31%(C)/ 9.3%(C)	↑	NR
rs3116636	intergenic	T*/C	1: 159686483	rs3116653 (0.86)	rs1417938 (1)	30%(T) / 6.8%(T)	↑	NR
rs12744244	Intergenic	A*/C	1: 159647656	rs3116653 (0.47)	NA	19.2%(A) / NA	↑	NR
rs12093699	Intergenic	A*/G	1: 159647988	rs12093699 (Same SNP)	rs12068753 (0.37)	33%(A) / 34%(A)	↑	NR
rs12068753	Intergenic	A*/T	1: 159692537	rs12068753 (Same SNP)	rs12068753 (Same SNP)	6.7%(A) / 42%(A)	↑	NR
rs3093075	intergenic	A*/C	1: 159679913	rs12068753 (1)	rs12068753 (0.66)	6.7%(A) / 34%(A)	↑	NR
rs3093068	downstream	G*/C	1: 159681364	rs12068753 (1)	rs16827466 (1)	6.7%(G) / 28%(G)	↑	NR
rs3093059	upstream	C*/T	1: 159685136	rs12068753 (1)	rs12068753 (0.70)	6.7%(C) / 33%(C)	↑	NR
rs11265260	Intergenic	G*/A	1: 159700039	rs12068753 (1)	NA	6.7%(G) / 7.6%(G)	↑	NR
rs1205	Intergenic	T*/C	1: 159682233	rs1205 (Same SNP)	rs1205 (Same SNP)	29%(T) / 15%(T)	↓	NR
rs2794520	intergenic	T*/C	1: 159678816	rs1205 (1)	rs1205 (0.74)	29%(T) / 20%(T)	↓	↓
rs7553007	intergenic	A*/G	1: 159698549	rs1205 (1)	rs1205 (0.67)	29%(A) / 21%(A)	↓	NR
rs2027471	Intergenic	A*/T	1: 159689388	rs1205 (0.96)	rs1205 (0.71)	30%(A) / 20%(A)	↓	NR
rs2592887	Intergenic	A*/G	1: 159652939	rs1205 (0.62)	rs12093699 (0.41)	40%(A) / 53%(A)	↓	NR
rs16827466 (merged with rs10494326)	intergenic	T*/C	1: 159649700	NA	rs16827466 (Same SNP)	NA / 18.6% (T)	NR	↑
rs1800947	synonymous	C*/G	1: 159683438	NA	NA	4.2%(C) / 0%(C)	↓	NR
rs3091244	upstream	A*/G/ T*	1: 159684665	NA	rs12068753 (0.74)	33%(T) / 33%(A)	↑	NR

*=minor allele, NA= not in 1000 Genomes for proxy or not in HapMap
Abbreviations: Minor Allele Frequency (MAF); European American (EA); African American (AA); Single Nucleotide Polymorphism (SNP)

Supplementary Figures

Figure S1. Quantile-quantile Plots of observed and expected $-\log_{10}(P)$ for all variants included in: A) single variant, B) T1, and C) SKAT tests.

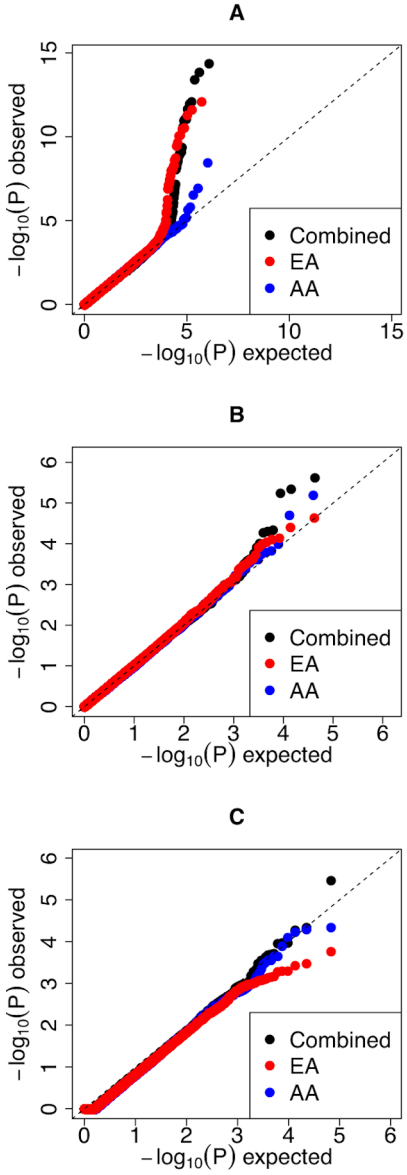


Figure S2. Comparison of Coronary Artery Risk Development in Young Adults CRP levels (mg/L) among Thr59Met carriers of the reference (blue) and carriers of the alternative allele (red) with polyclonal and monoclonal detection assays.

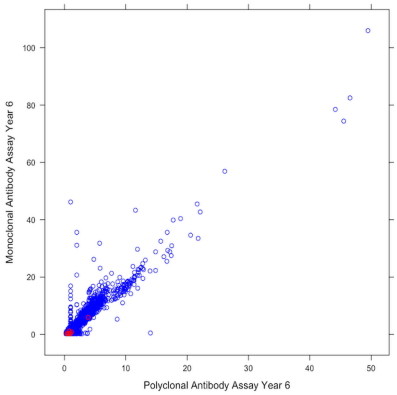
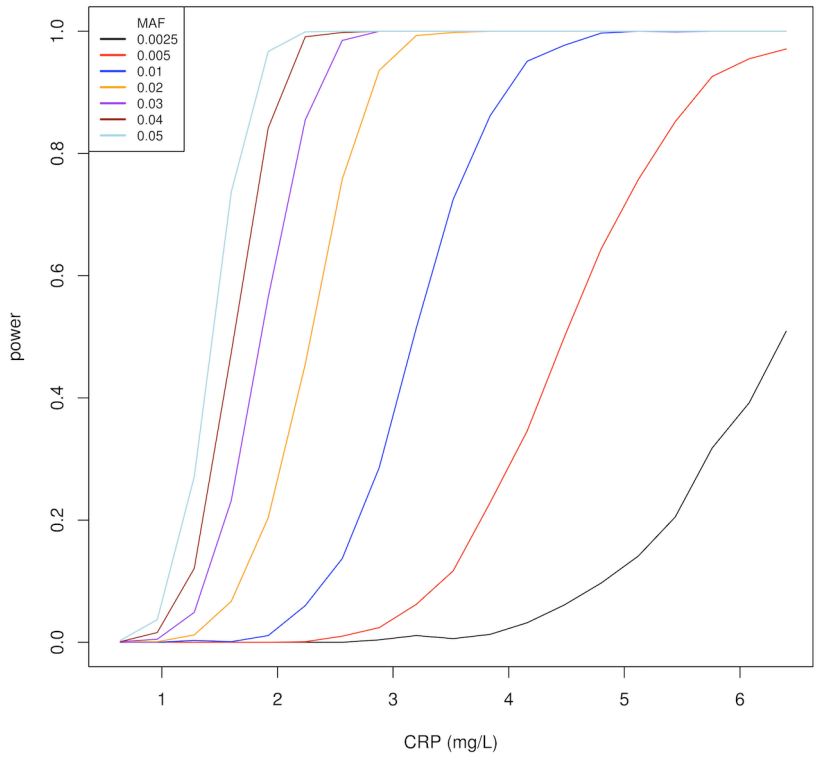


Figure S3. Graphical display of discovery sample power calculations based on N= 6050, the sample size of EA included in our study.



References

1. Dehghan, A., Dupuis, J., Barbalic, M., Bis, J.C., Eiriksdottir, G., Lu, C., Pellikka, N., Wallaschofski, H., Kettunen, J., Henneman, P., et al. (2011). Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731-738.
2. Elliott, P., Chambers, J.C., Zhang, W., Clarke, R., Hopewell, J.C., Peden, J.F., Erdmann, J., Braund, P., Engert, J.C., Bennett, D., et al. (2009). Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA : the journal of the American Medical Association 302, 37-48.
3. Ridker, P.M., Pare, G., Parker, A., Zee, R.Y., Danik, J.S., Buring, J.E., Kwiatkowski, D., Cook, N.R., Miletich, J.P., and Chasman, D.I. (2008). Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. American journal of human genetics 82, 1185-1192.
4. Reiner, A.P., Barber, M.J., Guan, Y., Ridker, P.M., Lange, L.A., Chasman, D.I., Walston, J.D., Cooper, G.M., Jenny, N.S., Rieder, M.J., et al. (2008). Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein. American journal of human genetics 82, 1193-1201.
5. 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. American journal of human genetics 91, 502-512.
6. 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.
7. 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.
8. 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.
9. Curocichin, G., Wu, Y., McDade, T.W., Kuzawa, C.W., Borja, J.B., Qin, L., Lange, E.M., Adair, L.S., Lange, L.A., and Mohlke, K.L. (2011). Single-nucleotide polymorphisms at five loci are associated with C-reactive protein levels in a cohort of Filipino young adults. Journal of human genetics 56, 823-827.
10. Kong, M., and Lee, C. (2013). Genetic associations with C-reactive protein level and white blood cell count in the KARE study. International journal of immunogenetics 40, 120-125.