Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits

Yuetiva Deming1,4, Jian Xia1,2,*, Yefei Cai1, Jenny Lord3,3, Jorge L. Del-Aguila1, Maria Victoria Fernandez1, David Carrell1, Kathleen Black1, John Budde1, ShengMei Ma1, Benjamin Saeedi1, Bill Howells1, Sarah Bertelsen1, Matthew Bailey4, Perry G. Ridge4, Alzheimer’s Disease Neuroimaging Initiative (ADNI)5,6,7,8, John C. Morris5,6,7,8, Kelly Bales9, Eve H. Pickering9, Jin-Moo Lee5, Laura Heitsch5, John Kauwe4, Alison Goate1,7,8, Laura Piccio5 & Carlos Cruchaga1,8

Genome-wide association studies of 146 plasma protein levels in 818 individuals revealed 56 genome-wide significant associations (28 novel) with 47 analytes. Loci associated with plasma levels of 39 proteins tested have been previously associated with various complex traits such as heart disease, inflammatory bowel disease, Type 2 diabetes, and multiple sclerosis. These data suggest that these plasma protein levels may constitute informative endophenotypes for these complex traits. We found three potential pleiotropic genes: ABO for plasma SELE and ACE levels, FUT2 for CA19-9 and CEA plasma levels, and APOE for ApoE and CRP levels. We also found multiple independent signals in loci associated with plasma levels of ApoH, CA19-9, FetuinA, IL6r, and LPa. Our study highlights the power of biological traits for genetic studies to identify genetic variants influencing clinically relevant traits, potential pleiotropic effects, and complex disease associations in the same locus.

Plasma proteins play important roles in numerous biological pathways, contribute to risk for many diseases, and have long been used for clinical risk assessment, diagnosis, prognosis and evaluation of treatment efficacy. Protein levels used as a quantitative trait in genome-wide association studies (GWAS) can act as an intermediate phenotype that functionally links genetic variation to disease-predisposing factors and then to complex disease end points1,2. Therefore, studies that link genetic variants with protein traits may provide a means to reveal the underlying mechanisms of the GWAS findings.

Previous case-control studies have associated many loci with various complex diseases. Unfortunately the effect sizes of genetic associations with complex disorders are generally small and the functional information on the underlying biological processes is often unclear or absent, which complicates the interpretation of the results. As a result, the focus of GWAS is now shifting increasingly away from studying associations with disease end points and toward associations with intermediate traits that are known risk factors for disease3–5.

A previous study used GWAS data and various commercially available enzyme-linked immunosorbent assay (ELISA) kits to find genetic variants associated with plasma or serum levels of 42 different proteins (such as interleukin 18, insulin, and leptin) implicated in various complex diseases (such as lupus, diabetes, and obesity)6. They identified several GWAS hits that could help in understanding the biology of those complex traits6. Recent

1Department of Psychiatry, Washington University School of Medicine, 660 S. Euclid Ave. B8134, St. Louis, MO 63110, USA. 2Department of Neurology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, P.R. China. 3Human Genetics Programme, Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK. 4Department of Biology, Brigham Young University, Provo, UT, USA. 5Department of Neurology, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110, USA. 6Department of Developmental Biology, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110, USA. 7Knight Alzheimer’s Disease Research Center, Washington University School of Medicine, 4488 Forest Park Ave., St. Louis, MO 63108, USA. 8Hope Center for Neurological Disorders. Washington University School of Medicine, 660 S. Euclid Ave. B8111, St. Louis, MO 63110, USA. 9Neuroscience Research Unit, Worldwide Research and Development, Pfizer, Inc., Groton, CT, USA.

*These authors contributed equally to this work. ¶A full list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to C.C. (email: cruchagc@psychiatry.wustl.edu).
technological developments have made possible the quantification of multiple proteins in a single analytical procedure, allowing both broader and deeper molecular profiling of large cohorts. Genetic analyses of these data have discovered numerous genomic regions associated with clinically relevant proteins, with recent large-scale proteome analyses having identified many loci associated with serum and plasma concentrations of individual proteins. Nevertheless, our understanding of the genetic basis and pathophysiological impact of variations in protein levels remains far from complete. Most of these studies limited analyses to cis variants or focused on candidate regions rather than genome-wide scans. Recent research suggests the importance of investigating protein phenotypes beyond those used in traditional genetic studies.

Here we present the results of an unbiased large genetic investigation of protein phenotypes in 818 unrelated individuals from the Washington University Knight Alzheimer’s Disease Research Center (KADRC) and Alzheimer’s Disease Neuroimaging Initiative (ADNI) who were analyzed for both genome-wide SNP genotypes and for 146 phenotypic measures obtained from multi-analyte panels (Human DiscoveryMAP) of human plasma samples.

Results
Before any genetic analyses we performed extensive quality control (QC) in the genotype and phenotype data. After log transformation and standardization (see materials and methods) we confirmed that the protein levels followed a normal distribution. We also tested the correlation between the analyte values and covariates such as age, gender, and Alzheimer’s disease (AD) status (Supplementary Tables S1 and S2). Age, gender, disease status, study, and principal components factors (PCs) from population stratification were included as covariates.

We decided to perform a one-stage GWAS rather than a two-stage GWAS because 1) we have GWAS for all the samples, and 2) it has been shown that combining data from both stages of a two-stage GWAS to perform a single analysis almost always has increased power to identify genetic association than analyzing the groups separately even though a lower statistical threshold is required to determine significance. To maximize our statistical power, we combined the two datasets to perform a joint one-stage GWAS with all 818 individuals from ADNI and KADRC (characteristics shown in Table 1). To verify our results, we followed up with additional analyses stratified by study and performed meta-analyses of the results from each dataset for each analyte, and we found that the p-values from the meta-analyses were similar to the joint GWAS p-values (Supplementary Table S3). In order to avoid spurious association and consider a single nucleotide polymorphism (SNP) as a real signal, we required each genome-wide significant association from the joint analysis to meet additional criteria: 1) the SNP association had to be consistent between the two series, in the same direction and with similar effect size, which represents an internal replication (Supplementary Table S3) and 2) since we were using cohorts from AD studies, we wanted to be sure our results were not confounded by AD status. In addition to using AD status as a covariate in our initial analyses, we performed separate GWAS on cases and controls and found no difference in effect size or direction indicating the associations found in the combined GWAS were not confounded by AD status (Supplementary Table S4).

We decided to use the common threshold for genome-wide significance \(p < 5.0 \times 10^{-8}\) instead of \(p < 3.42 \times 10^{-10}\) (Bonferroni multiple test correction taking into account SNPs and phenotypes) because the latter would consider that all the analytes are independent and not correlated. However there is extensive evidence that this is not the case and in a recent study we demonstrated that some analytes are highly correlated. Additionally five of the associations in this study in the \(p = 5 \times 10^{-8} - 3.42 \times 10^{-10}\) range have been previously reported and others are located in receptors and genes known to regulate levels of the analyte (Table 2 and Supplementary Table S5) which indicate that these are real signals. We also found complex loci and potential pleiotropic effects that support the evidence that not all of the SNPs and analytes act independently of others. These findings suggest that a multiple test correction threshold of \(p < 3.42 \times 10^{-10}\) would be too stringent. For this reason we decided to report all the loci with a \(p < 5.0 \times 10^{-8}\), but we also highlight on Table 2 those that pass the \(p < 3.42 \times 10^{-10}\) threshold.

### Genome-wide association study results
After performing the linear regression with each analyte as a phenotype, there were 56 genome-wide significant loci for 47 analytes (Table 2). Twenty-eight of the 56 associations (9 novel) pass the \(p < 3.42 \times 10^{-10}\) threshold.

### Previously reported findings
Twenty-eight of our genome-wide signals replicated associations reported by 14 different genetic studies of plasma or serum protein levels in humans (Table 2 and Supplementary Table S5). Six of our most significant SNPs were the same SNP reported previously and the remaining SNPs were in linkage disequilibrium (LD) with reported SNPs (Supplementary Table S5). Fifteen of these 28 genome-wide loci had \(p < 3.42 \times 10^{-10}\) in our study and five others were in the \(p = 5 \times 10^{-8} - 3.42 \times 10^{-10}\) range, indicating that signals in this range in our study constitute strong associations. Twenty-three of these previously reported loci are...
| SNP    | Gene  | Effect  | MAF     | Joint p-value | beta | Previously reported |
|--------|-------|---------|---------|---------------|------|---------------------|
| rs2275501 | CD5L  | cis     | 0.395   | 9.97E-10      | 0.309 | known 9             |
| rs763105 | LPA   | cis     | 0.025   | 1.96E-09      | 0.290 | known 20            |
| rs646776 | CELSR2| cis     | 0.231   | 2.20E-09      | 0.301 | novel               |
| rs7541151 | DDAH1 | trans    | 0.028   | 6.64E-09      | 0.070 | novel               |
| rs4244229 | MIR4990 | intergenic | 0.247  | 8.37E-09   | 0.311 | novel               |
| rs2031368 | GLRX3 | trans    | 0.399   | 1.04E-08      | 0.250 | novel               |
| rs409336 | CXCL5 | cis     | 0.143   | 1.11E-08      | 0.405 | novel               |
| rs4785768 | GLRX3 | trans    | 0.399   | 1.04E-08      | 0.250 | novel               |
| rs3301394 | CST3  | cis     | 0.147   | 1.23E-08      | 0.347 | known 9,14,23       |
| rs2775281 | CDH6  | trans    | 0.031   | 1.37E-08      | 0.767 | novel               |
| rs4953649 | FSHR  | trans    | 0.347   | 1.66E-08      | 0.088 | novel               |
| rs2519093 | ABO   | trans    | 0.072   | 1.90E-08      | 0.352 | **known **26        |
| rs7374126 | CTNN2D2 | trans   | 0.028   | 2.08E-08      | 0.841 | novel               |
| rs1263549 | PTPRN2 | trans    | 0.039   | 2.16E-08      | 0.685 | novel               |
| rs3742714 | PARVG | trans    | 0.411   | 2.21E-08      | 0.277 | novel               |
| rs7292754 | ST6SIA3 | trans   | 0.038   | 2.26E-08      | 0.679 | novel               |
| rs111494896 | GFRA2 | trans    | 0.041   | 2.77E-08      | 0.695 | novel               |
| rs46243736 | MIR169 | trans    | 0.062   | 3.06E-08      | 0.559 | novel               |
| rs11594693 | WDR11 | trans    | 0.265   | 3.72E-08      | 0.239 | novel               |
| rs2279191 | DHHX | trans    | 0.242   | 3.83E-08      | 0.239 | novel               |
| rs13027473 | RAMPI | trans    | 0.149   | 4.53E-08      | 0.345 | novel               |
| rs11889675 | TAF1B | trans    | 0.152   | 4.60E-08      | 0.364 | novel               |

SNPs below this line have p < 5E-8 and > 3.42E-10
in cis (within 1MB of the gene that encodes the protein) and five are in trans (Table 2, Fig. 1, and Supplementary Figs 1–26). Twelve (52%) of the cis effects are coding variants (nine missense) and four of the trans effects are coding variants (three missense; Table 2). None of the trans effects are located in untranslated regions (UTR) but two of the analytes had cis effects that are in the UTR (CD40: 5′ UTR of CD40 and HCC4: 3′ UTR of CCL16). All of the trans effects are located within genes (four coding, one intronic) that have interactions with the analyte that are not known or well understood (Table 2). However our results and the previous published studies suggest that these loci in trans proteins play an important role in regulating the levels of CA19-9, CEA, CRP, SELE, and ACE in plasma\(^{5,7,21,24}\). More interestingly, some of these loci, like ABO or FUT2, are genome-wide for more than one analyte, which also indicates that these may constitute master regulatory signals (see pleiotropic section).

**Novel findings.** We found 28 loci associated with 25 analytes that have not been reported previously (Table 2, Fig. 1, and Supplementary Figs S8, S12, S21, S22, S27–36). Of these novel associations nine pass the \(p < 3.42 \times 10^{-7}\) threshold (Table 2). All of the associations were highly consistent (same effect size or beta) between the two datasets which represents an internal replication (Supplementary Table S3) and were not confirmed by AD status (Supplementary Table S4).

Five of our 28 novel findings were cis effects (one coding variant and four intergenic; Table 2): 1) rs926144 which is 29.7 KB from SERPINA1 was significantly associated with plasma levels of AAT (\(p = 4.71 \times 10^{-12}\); Supplementary Fig. S27), 2) rs2015086 within 2 KB upstream of CCL18 was significantly associated with MIPIa levels in plasma (\(p = 2.56 \times 10^{-11}\); Supplementary Fig. S38), 3) a missense variant in AGER (rs2070600) was associated with plasma RAGE levels (\(p = 1.86 \times 10^{-11}\); Supplementary Table S40), 4) rs646776 which is located 33.7 KB from SORT1 was significantly associated with Sortilin plasma levels (\(p = 2.20 \times 10^{-9}\); Supplementary Fig. S41), and 5) rs409336 located 3.7 KB from CXCL5 was significantly associated with ENA78 plasma levels (\(p = 1.11 \times 10^{-8}\); Supplementary Fig. S30).

Twenty-three of our 28 novel findings were trans effects. Twelve analytes were associated with loci that contained only intragenic SNPs and eleven analytes (ANG2, BLC, CEA, F7, FG4, GROa, MIPIb, MMP7, RAGE, THPO, TNC) were associated with SNPs on intronic regions in gene-rich areas. Interestingly some of these loci contain intronic SNPs that are likely to be regulatory based on RegulomeDB\(^{25}\): SCARA5 (associated with RAGE levels) and PARV (associated with ANG2 levels) contain SNPs with RegulomeDB\(^{25}\) scores lower than 3 (Supplementary Table S6). Plasma MIPIb levels were also associated with a locus that contains SNPs that are likely to be regulatory. We found that rs145617407, located in the intron of CCR3, was significantly associated with MIPIb levels in plasma (\(p = 2.58 \times 10^{-9}\)) and this SNP is located less than 119 KB from CCR5 which is the receptor for CCL4/MIPIb (Supplementary Fig. S21).

**GWAS Conditional on top hits revealed additional signals within same loci.** We then performed conditional analyses to determine whether more than one signal in the same loci exists. When we added the most significant SNP to the linear regression model, five analytes (ApoH, CA19-9, FetuinA, IL6r, and LPa) still showed independent and genome-wide significant SNPs at the same locus (Fig. 1, Table 3 and Supplementary Fig. S5, S12, S17 and S19). It is interesting to note that three of four of the complex loci we found were cis with the respective protein whereas the FUT2/FUT6/FUT7 locus was associated with CA19-9 plasma levels. Since we decided to use the traditional genome-wide p-value threshold (\(p < 5 \times 10^{-8}\)) for the conditional analyses, we may be missing some additional independent signals.

After conditioning on rs2070633, located in an AHSG intron, we found that rs4917, a missense variant also located in AHSG, was still significantly associated with plasma levels of FETUa (\(p = 7.27 \times 10^{-9}\), original \(p = 2.61 \times 10^{-42}\); Table 3 and Supplementary Fig. S13). After conditioning on both SNPs no additional signals were found. An intronic variant in IL6R, rs7526131, was still significantly associated with IL6r plasma levels after conditioning on rs12126142, also located in an intron of IL6R (\(p = 1.43 \times 10^{-10}\), original \(p = 4.47 \times 10^{-7}\); Table 3 and Supplementary Fig. S17). Plasma levels of LPa were significantly associated with rs783147, located in an intron of PLG 506 KB from LPA, and after conditioning on this SNP an intronic variant of SLC22A1 approximately 0.4 MB from LPA (rs783147), was still significantly associated with LPa levels (\(p = 1.64 \times 10^{-5}\), original \(p = 9.86 \times 10^{-9}\); Table 3 and Supplementary Fig. S19).

We found two analytes (ApoH and CA19-9) that the genome-wide locus contained up to three independent signals (Fig. 1, Table 3 and Supplementary Fig. S5). All three signals in the ApoH analyses contained missense variants located in APOH (rs2597880: I141T, \(p = 1.57 \times 10^{-12}\); rs1801690: W335S, \(p = 5.15 \times 10^{-3}\), original \(p = 2.77 \times 10^{-11}\); Table 3 and Supplementary Fig. S17). As reported above, the initial signal in the CA19-9 GWAS contained a missense variant in FUT2. After conditioning on the most significant SNP (rs485073, \(p = 2.12 \times 10^{-22}\)) from the CA19-9 GWAS, the new signal contained a synonymous variant located in FUT6 (rs112313064, \(p = 3.79 \times 10^{-26}\), original \(p = 7.46 \times 10^{-22}\)), and conditioning on the two SNPs resulted in a separate signal upstream of FUT3 (rs230696, \(p = 2.78 \times 10^{-9}\), original \(p = 6.11 \times 10^{-21}\); Supplementary Fig. S5).

All of these results indicate these protein levels are highly regulated and that different and independent regulation mechanisms, even at the same locus, are in place: some mechanisms may act by affecting cleavage or receptor binding (non-synonymous variants) and others by regulating gene expression (non-coding variants).
Potential pleiotropy. In addition to finding that some proteins have complex regulation within the structural gene (or a different gene in the case of CA19-9), we also found potentially pleiotropic effects with one gene affecting more than one protein. Potential pleiotropic effects were found for three groups of analyte/associations even though the analyte levels were not correlated: ABO associated with ACE activity and SELE plasma and serum levels, ABO has also been associated with vWF plasma levels, and although the locus did not reach genome-wide significance in our analysis it was very close. 

Figure 1. Manhattan and regional plots for associations with plasma levels of ApoH. (a) Manhattan plot of $-\log_{10} p$-values for association with plasma levels of ApoH; (b) Regional plot for genome-wide significant association on chromosome 17 with ApoH plasma levels; (c) Regional plot for genome-wide significant association on chromosome 18 with ApoH plasma levels; (d) Regional plot for genome-wide significant association on chromosome 9 with ApoH plasma levels; (e) Regional plot for genome-wide significant association on chromosome 17 after conditioning for rs2873966; (f) Regional plot for genome-wide significant association on chromosome 17 after conditioning for rs2873966 and rs17690171.
### Table 3. Plasma analyte levels associated with multiple loci. Chr = chromosome, LD = linkage disequilibrium.

| Analyte | SNP  | Gene | Chr | Position | p-value | beta | p-value | beta | r² | D' |
|---------|------|------|-----|----------|---------|------|---------|------|----|----|
| ApoH    | rs2873966 | APOH | 17  | 64211973 | 3.37E-15 | 0.428 | reference SNP | 0.000 | 1.000 |
|         | rs17690171 | APOH | 64202857 | rs2873966 | 5.53E-09 | 0.389 | 1.004 | 0.981 |
|         | rs8178828 | APOH | 64223183 | rs2873966, rs17690171 | 6.64E-13 | 0.700 | 0.030 | 1.000 |
| CA19-9  | rs485073  | FUT2 | 19  | 4902755  | 2.12E-23 | 0.509 | reference SNP | 1.000 | 1.000 |
|         | rs485073  | FUT6 | 5831724 | rs485073 | 3.79E-26 | 0.509 | 0.000 | 0.012 |
| FetuinA  | rs2306969 | FUT3 | 3   | 186335941 | 2.88E-44 | 0.629 | reference SNP | 1.000 | 1.000 |
| IL6r    | rs12126142 | IL6R | 1   | 154425456 | 1.81E-106 | 0.850 | reference SNP | 1.000 | 1.000 |
| Lpa     | rs2873966 | PLG  | 6   | 161137990 | 9.86E-09 | 0.590 | reference SNP | 0.000 | 0.149 |

### Table 4. Potential pleiotropic associations. Chr = chromosome.

| Analyte | SNP  | Gene | Chr | Position | Joint p-value | Correlation of analyte levels |
|---------|------|------|-----|----------|---------------|-------------------------------|
| SELE    | rs2519093 | ABO  | 9   | 136141870 | 7.62E-51 | ACE/SELE: r = 0.118, p = 7.17E-4 |
| ACE     | rs2519093 | ABO  | 9   | 136141870 | 1.90E-08 | vWF/ACE: r = -0.059, p = 0.09 |
| vWF     | rs687289 | ABO  | 9   | 136137106 | 8.87E-08 | SELE/vWF: r = -0.097, p = 5.61E-3 |
| CA19-9  | rs485073 | FUT2 | 9   | 4920755  | 2.12E-23 | rs = 0.166, p = 2.98E-6 |
| CEA     | rs485073 | FUT2 | 19  | 4920755  | 4.07E-16 | r = 0.204, p = 4.82E-9 |
| ApoE    | rs769449 | APOE | 19  | 45410002 | 2.76E-26 | n/a |
| CRP     | rs769449 | APOE | 19  | 45410002 | 6.69E-09 | n/a |

*FUT2* was associated with plasma levels of CA19-9 and CEA (*p* = 2.12 × 10⁻²³, *beta* = −0.509; *p* = 4.07 × 10⁻¹⁶, *beta* = −0.406) respectively; Table 4 and Supplementary Fig. S5, S8); and the *APOE* region was associated not only with plasma levels of ApoE but also CRP (*p* = 2.76 × 10⁻²⁶, *beta* = −0.594; *p* = 6.69 × 10⁻⁹, *beta* = −0.354) respectively; Table 4 and Supplementary Fig. S4, S10).

Interestingly none of these analyte pairs or trios are highly correlated (*r* < 0.25; Table 4), which again supports the idea that these loci (*ABO, FUT2, and APOE-TOMM40 region*) are truly master-regulatory regions, that protein levels are highly and complexly regulated, and that studying the genetic architecture of biological traits can lead to a deeper knowledge of the biological processes.

#### Impact of these findings with complex diseases.

Of the 56 loci that we found associated with plasma protein levels, 46 loci have also been reported to be associated with complex traits and diseases including coronary artery disease (ACE and SELE), stroke (ACE and SELE), various cancers (ACE, CA19-9, CEA, RAGE, and SELE), age-related macular degeneration (APOE, CFHR1, and CRP), periodontitis (APOH), multiple sclerosis (BLC and CD40), inflammatory bowel disease (CD40 and ENA78), and Type 2 diabetes (IL13, MC5F, and RAGE) (Table 5; see supplementary results for a complete description). As an example, the *AGER* variant rs2070633, which in our study was associated with plasma RAGE levels (*p* = 1.86 × 10⁻¹¹), has been reported to be associated with pulmonary function⁸⁻¹⁸. A recent study of RAGE plasma levels suggests they are a promising biomarker for acute respiratory distress syndrome, supporting our hypothesis⁸⁻¹⁸.

Similarly our genetic analysis for BLG revealed a significant association with SNPs located in *DDAH1* (rs7541151, *p* = 6.44 × 10⁻⁹; Table 2), a gene that has been associated with multiple sclerosis (MS). Interestingly BLC levels have recently been reported to be different between patients with MS and controls³⁰, which further supports BLC as a potential biomarker.

Since levels of CD40 in plasma were associated with the CD40 locus and CD40 variants have been associated with MS in three independent GWAS⁶⁻⁻³², we hypothesized that plasma levels of CD40 may also be associated with MS status. As a proof of concept, we used a Quantikine sandwich ELISA kit (R&D Systems cat #DCCD40) to measure plasma levels of CD40 in 20 individuals with relapsing remitting MS in remission at time of plasma collection (8 male, 12 female; mean age = 44.45 ± 15.51 years) and 20 healthy controls (8 male, 12 female; mean age = 41.84 ± 11.52 years; Supplementary Table S7). We used linear regression to determine if log values of plasma CD40 levels were significantly different between MS cases and controls, with age and gender as covariates.
We found plasma levels of CD40 were significantly higher in MS cases (753.26 ± 235.71 pg/mL) than controls (603.02 ± 139.01 pg/mL; p = 0.041, beta = −1.837; Fig. 3), supporting our hypothesis. More than half of the loci associated with plasma protein levels in our study have been previously reported to be associated with various complex diseases. Based on the current knowledge for RAGE and BLC, and in the concept of Mendelian randomization, we hypothesize that these protein levels constitute informative biomarkers for these complex traits although additional studies would be necessary to validate this hypothesis. More detailed information about potential novel biomarkers for complex traits is included in Supplementary Results and analyte abbreviations with full names are in Supplementary Table S8.
| Analyte | SNP   | Gene | Predicted function | Joint p-value (based on NHGRI catalog) | Other traits associated with loci or gene |
|---------|-------|------|--------------------|----------------------------------------|------------------------------------------|
| AAT     | rs926144 | SERPINA6 | intergenic         | 4.71E-12                               | Metabolome levels; Metabolic traits; Angiotensin-converting enzyme activity |
| ACE     | rs4343  | ACE  | coding             | 6.66E-44                               | Lipid traits; Coronary artery disease; Ischemic stroke; Large artery stroke; Serum alkaline phosphatase levels; Malaria; Venous thromboembolism; Graves' disease; Thyroid hormone levels; Tumor biomarkers; End-stage coagulation; Coagulation factor levels; Red blood cell traits; Obesity-related traits; Activated partial thromboplastin time; Duodenal ulcer; Inflammatory biomarkers; Liver enzyme levels; Metabolic traits; Soluble ICAM-1; D-dimer levels; Phystosterol levels; E-selectin levels; Soluble levels of adhesion molecules; Hematological and biochemical traits; mean corpuscular hemoglobin concentration; Angiotensin-converting enzyme activity; Pancreatic cancer; vWF and FVIII levels |
| AGT     | rs55837081 | AGT  | intergenic         | 4.45E-34                               | AGT levels |
| ANG2    | rs3747214 | PARVG | inronic            | 2.21E-08                               | — |
| ApoA4   | rs1263167 | APOA4 | intergenic         | 2.64E-54                               | Hypertriglyceridemia; Coronary heart disease; HDL cholesterol; LDL cholesterol; Triglycerides; total cholesterol; Hematological and biochemical traits |
| ApoE    | rs769449 | APOE | inronic            | 2.76E-26                               | LDL cholesterol; Alzheimer's disease; HDL cholesterol; C-reactive protein; Age-related macular degeneration; Cholesterol, total; Alzheimer's disease biomarkers; Brain imaging; Triglycerides; Quantitative traits; Apolipoprotein Levels; Metabolite levels; Cardiovascular disease risk factors; Lipid traits; Response to statin therapy (LDL-C); Lipid metabolism phenotypes |
| ApoH    | rs2873966 | APOH | inronic            | 3.37E-15                               | Blood pressure measurement (high sodium and potassium intervention); LDL cholesterol; B2-Glycoprotein I plasma levels |
| ApoH    | rs17690171 | APOH | intergenic         | 1.02E-15                               | Blood pressure measurement (high sodium and potassium intervention); LDL cholesterol; B2-Glycoprotein I plasma levels |
| ApoH    | rs8178828 | APOH | inronic            | 6.84E-13                               | Blood pressure measurement (high sodium and potassium intervention); LDL cholesterol; B2-Glycoprotein I plasma levels |
| ApoH    | rs17842936 | MB2D | intergenic         | 1.03E-10                               | Periodontitis |
| ApoH    | rs48521708 | TIP2 | intergenic         | 2.64E-10                               | Refractive error; Renal sinus fat |
| BLC     | rs7541151 | DDAH1 | inronic            | 6.44E-09                               | Serum dimethylarginine levels; multiple sclerosis |
| CA19-9  | rs485073 | FUT2  | 3' UTR             | 2.12E-23                               | Tumor biomarkers; Vitamin B12 levels; Bipolar disorder; Retinal vascular caliber; Liver enzyme levels (alkaline phosphatase); C-reactive protein; Age-related macular degeneration; Cholesterol, total; Alzheimer's disease biomarkers; Brain imaging; Triglycerides; Quantitative traits; Apolipoprotein Levels; Metabolite levels; Cardiovascular disease risk factors; Lipid traits; Response to statin therapy (LDL-C); Lipid metabolism phenotypes |
| CA19-9  | rs12313064 | FUT5  | coding             | 7.46E-23                               | Tumor biomarkers; Vitamin B12 levels; N-glycan levels |
| CA19-9  | rs2306969 | FUT3  | upstream 2KB       | 6.11E-23                               | Elevated serum carcinoembryonic antigen levels; N-glycan levels |
| CD40    | rs6032660 | CD40  | intergenic         | 9.81E-21                               | Inflammatory bowel disease; Kawasaki disease; Rheumatoid arthritis; multiple sclerosis |
| C5L     | rs2765501 | C5L   | inronic            | 9.97E-10                               | CD6; CD5; PTGDR2 |
| CEA     | rs570794 | FUT2  | 3' UTR             | 2.84E-17                               | Tumor biomarkers; Vitamin B12 levels; Bipolar disorder; Retinal vascular caliber; Liver enzyme levels (alkaline phosphatase); C-reactive protein; Age-related macular degeneration; Cholesterol, total; Alzheimer's disease biomarkers; Brain imaging; Triglycerides; Quantitative traits; Apolipoprotein Levels; Metabolite levels; Cardiovascular disease risk factors; Lipid traits; Response to statin therapy (LDL-C); Lipid metabolism phenotypes |
| CEA     | rs12468845 | AFF3  | inronic            | 3.88E-08                               | — |
| CFHR1   | rs12144939 | CFH  | inronic            | 8.99E-143                              | Age-related macular degeneration; Circulating myeloperoxidase levels (serum); Complement C3 and C4 levels; Nephropathy Meningococcal disease |
| CRP     | rs12972156 | PVR2  | inronic            | 9.93E-11                               | Alzheimer's disease biomarkers; Age-related macular degeneration; HDL cholesterol; Alzheimer's disease |
| CystC   | rs13039144 | CST3  | intergenic         | 1.23E-08                               | Chronic kidney disease; Cystatin C |
| ENA78   | rs409336 | CXCL5 | intergenic         | 1.11E-08                               | Inflammatory bowel disease; Metabolite levels |
| F7      | rs100655 | MCF2L | 3' UTR             | 1.44E-26                               | Osteoarthritis; Factor VII |
| F1      | rs11594693 | WDRJ1 | inronic            | 3.72E-08                               | — |
| FetuinA | rs207633 | AHSG  | inronic            | 2.88E-44                               | Fetuin-A levels; Activated partial thromboplastin time |
| FetuinA | rs2593813 | AHSG  | inronic            | 6.74E-44                               | Fetuin-A levels; Activated partial thromboplastin time |
| FGF4    | rs3117858 | GALNT6 | intergenic       | 3.12E-10                               | Myopia (pathological); Response to amphetamines; Bipolar disorder and schizophrenia; Obesity-related traits |
| GROa    | rs1263549 | PTPRN2 | inronic           | 2.16E-08                               | — |
| GSTa    | rs9395826 | GSTA1 | intergenic         | 8.19E-11                               | GSTa levels |
| HCC4    | rs80329614 | CCL16 | downstream 500B    | 1.16E-27                               | HCC4 levels |
| Continued |       |      |                    |                                        | — |
| Analyte SNP | Gene | Predicted function | Joint p-value | Other traits associated with loci or gene (based on NHGRI catalog) |
|------------|------|--------------------|---------------|---------------------------------------------------------------|
| HP rs72787038 | DHODH | intergenic | 9.69E-14 | Attention deficit hyperactivity disorder and conduct disorder |
| IL13 rs7433647 | UBE2E2 | intergenic | 1.21E-08 | Psychosis (atypical); Type 2 diabetes |
| IL16 rs11556218 | IL16 | missense | 6.02E-32 | Inattentive symptoms |
| IL18 rs146245376 | MIR3169 | intergenic | 3.06E-08 | — |
| IL6 rs12126142 | IL6R | intronic | 1.81E-106 | — |
| IL6 rs7526131 | IL6R | intronic | 4.47E-72 | — |
| IL8 rs11889675 | TAF1B | intergenic | 4.60E-08 | — |
| Leptin rs203468 | GLRX3 | intergenic | 1.04E-08 | HIV-1 susceptibility |
| LPa rs783147 | PLG | intronic | 1.96E-09 | Aging; LP(a) levels |
| LPa rs4646272 | SLC22A1 | intergenic | 9.86E-09 | — |
| MCP2 rs1133763 | CCL8 | missense | 3.54E-13 | — |
| MCSF rs3741236 | CTNND2 | intergenic | 2.08E-08 | Amyotrophic lateral sclerosis (sporadic); Myopia (pathological) |
| MCSF rs111494896 | GFRα2 | intergenic | 2.77E-08 | Neurophatic pain in type 2 diabetes; Migraine with aura |
| MIP1a rs2015086 | CCL18 | upstream 2KB | 2.56E-15 | Higher macrophage expression of CCL18 in human carotid atherosclerotic plaques |
| MIP1β rs145617407 | CCR3 | intronic | 2.58E-10 | MCP1 levels; Obesity-related traits; Monocyte chemoattractant protein-1; Celiac disease |
| MIP1β rs4796217 | CCL4L2 | intergenic | 1.19E-08 | — |
| MIP7 rs9753755 | MACROD2 | intronic | 8.87E-11 | — |
| MIPF1 rs861273 | CCL23 | upstream 2KB | 6.37E-14 | — |
| MIPF1 rs27252381 | CDH6 | intergenic | 1.37E-08 | Response to methotrexate in juvenile idiopathic arthritis; Liver enzyme levels (gamma-glutamyl transferase); Emphysema-related traits |
| NrCAM rs10487851 | NRCAM | intronic | 3.01E-10 | — |
| RAGE rs2070600 | AGER | missense | 1.86E-11 | Normal glucose metabolism; Impaired glucose metabolism; Type 2 diabetes mellitus; Chronic obstructive pulmonary disease; Prostate cancer; Pulmonary function |
| RAGE rs4953649 | FSHR | intronic | 1.66E-08 | Response to anti-retroviral therapy (ddI/ddT) in HIV-1 infection (Grade 3 peripheral neuropathy); Adverse response to chemotherapy (neuropenia/leucopenia) (etoposide); Polycystic ovary syndrome; Erectile dysfunction and prostate cancer treatment; Radiation response |
| SELE rs507666 | ABO | intronic | 1.01E-52 | Lipid traits; Coronary artery disease; Ischemic stroke; Large artery stroke; Serum alkaline phosphatase levels; Malaria; Venous thromboembolism; Graves' disease; Thyroid hormone levels; Tumor biomarkers; End-stage coagulation; Coagulation factor levels; Red blood cell traits; Obesity-related traits; Activated partial thromboplastin time; Duodenal ulcer; Inflammatory biomarkers; Liver enzyme levels; Metabolic traits; Soluble ICAM-1; D-dimer levels; Phystosterol levels; E-selectin levels; Soluble levels of adhesion molecules; Hematological and biochemical traits; mean corpuscular hemoglobin concentration; Angiotensin-converting enzyme activity; Pancreatic cancer; vWF and FVIII levels |
| Sortilin rs646776 | CELSR2 | downstream 500B | 2.20E-09 | — |
| TECK rs72927542 | ST8SIA3 | intergenic | 2.26E-08 | — |
| TF rs6762415 | TF | intronic | 1.17E-14 | Transferin levels; Alcohol consumption (transferin glycosylation); Iron status biomarkers; Hepcidin levels; Iron levels |
| TFF3 rs2444229 | MIR4790 | intergenic | 8.33E-09 | — |
| THP rs12934455 | UMOD | intronic | 2.80E-42 | Femoral neck bone geometry; Hypertension; Chronic kidney disease and serum creatinine levels; Chronic kidney disease; Renal function and chronic kidney disease |
| THPO rs2279191 | DHX8 | intronic | 3.83E-08 | — |
| TNC rs2685421 | SCARA5 | intronic | 5.27E-11 | Adverse response to chemotherapy (neuropenia/leucopenia) (cisplatin) |
| VCAM1 rs13027473 | RAMP1 | intergenic | 4.53E-08 | — |

Table 5. Joint GWAS top SNPs/genes related to disease based on NHGRI catalog.
Discussion

GWAS of complex traits have been very successful in identifying novel loci associated with those traits, but these studies require extremely large sample sizes, and in some cases it is difficult to interpret the results because the associations are with surrogate tag SNPs which may not be the causal SNPs. Many loci contain multiple genes which also makes it difficult to determine the causal gene or variant. Additionally some loci are located in non-protein coding regions where functional effects are poorly understood. Genetic analyses of biological traits may provide more power than traditional GWAS and may be more informative about the biological effects for specific loci. Using a more unbiased approach than previous genetic studies, we were able to replicate many previously reported associations with various plasma protein levels and uncover several novel associations that could warrant further research. The results from our careful analyses suggest that even though we utilized two datasets from Alzheimer's disease studies there was no confounding effect due to disease status or dataset. Combining datasets from high-throughput technologies that deliver genome-wide genetic data and quantification of protein levels in a single procedure provides a great deal of power to analyses that may help researchers understand the biology of complex traits including the complex loci involved and pleiotropic effects.

Our results clearly indicate that the protein levels are highly and complexly regulated. We found master regulatory regions (pleiotropic; Table 4, Fig. 2, and Supplementary Fig. S4, S10) as well as several independent regulatory elements in the same locus for the same proteins (Table 3, Fig. 1, and Supplementary Fig. S5, S13, S17 and S19). We found protein levels associated with variants in or near the gene coding that protein (cis effects) as well as variants located elsewhere in the genome (trans effects) demonstrating that protein levels are not only affected by the genes that encode the protein but also by interaction with other proteins as in the case of ABO or FUT2 (Table 4).

Interestingly, we found that for almost half of the cis effects (13 out of 28), the association could be explained by a coding variant but for the trans effects most of the loci (24 out of 28) only contain regulatory variants (Table 2). Although these non-coding signals could be synthetic association and are being driven by low frequency variants, our results and those recently published by ENCODE and the GTeX consortium would suggest that those associations are likely to affect gene expression^{33,34}. For this same reason, it is more likely that the association in cis (more frequently due to a non-synonymous variant) will present a higher effect size and are easier to identify in a genetic study than a trans signal, which is more likely to affect gene expression through regulation.

Table 2 shows that most of the trans effects associated with plasma protein levels had less significant p-values and lower betas than most of the cis effects. This could explain why only three of the trans effects we found were previously reported while the other 24 were novel. It is of vital importance to identify trans effects because that will help us to identify novel biological interactions and pathways. Of the 28 trans effects we found in our study, only one corresponded to a protein that constituted the receptor of the studied analyte or a gene known to interact directly with the analyte (rs145617407 located less than 119 KB from CCR5 which is the receptor for CCL4/MIP1b^{35}). However, the fact that the associations of SELE, ACE, and vWF with the ABO locus or CA19-9 and CEA with FUT2 have been identified in other studies, indicates that these signals are real and some of these novel loci may be implicated in regulating the levels of one or more proteins. Additional work is needed because currently it is not clear how ABO regulates plasma levels of SELE, ACE, and vWF or how FUT2 regulates CEA and CA19-9 levels. For the novel loci this can be more complicated because several signals are located in very gene-rich regions and several genes could drive the association (Fig. 1 and Supplementary Fig. S1, S6, S8, S10, S21, S24, S28, 29, S33, S36-S37, S42, S44, 46).

Another important finding related to this study is its implication on complex traits. Proteins play a key role in many complex traits, so understanding the genetic variations associated with protein levels is important in understanding the biological basis of these traits. We used the concepts of Mendelian randomization, our data, and the data from the NHGRI GWAS catalog to identify genetic regions that are genome-wide significant for various analyte levels as well as previously associated with complex traits. While most of these loci have been associated with complex traits, the associations of most of the plasma analytes with the complex traits have not been previously reported. Our results suggest that some of these plasma protein levels could be novel biomarkers or even endophenotypes for these complex traits.

As an example of our approach providing information useful for identifying potential pleiotropic effects in promising biomarkers for complex diseases that has been supported by previous research: rs485073 in FUT2 was associated in our study with plasma levels of both CEA and CA19-9, which are only weakly correlated in plasma (r = 0.166, p = 2.98 × 10^{-4}). This potential pleiotropy strongly suggests that rs485073 is part of a master regulatory region. In this case this means that plasma levels of CEA and CA19-9 could be important for understanding gastric cancer because FUT2 variants have also been associated with gastric cancer risk^{37}. This is further supported by the fact that both CEA and CA19-9 have been reported as FDA approved biomarkers for other types of cancer^{37}.

We found several promising plasma biomarkers for complex traits including IL13, ENA78, BLG, and CD40. Based on our results, plasma levels of IL13 may be informative in Type 2 diabetes research. We found rs7433647, located near UBE2E2, was associated with IL13 plasma levels (p = 1.21 × 10^{-10}). UBE2E2 has previously been associated with Type 2 diabetes in a large GWAS meta-analysis of more than 26,000 cases and 83,000 controls with varied ancestry^{38}. A recent study using a mouse model for Type 2 diabetes suggests that expression of IL13 plays a key role in adipose tissue inflammation and insulin resistance, further supporting the idea that IL13 levels may be important in studying Type 2 diabetes^{39}. ENA78/CXCL5 expression is elevated in the inflamed tissues of patients with rheumatoid arthritis, ulcerative colitis and Crohn's disease^{40,41}. Several studies have reported association of CXCL5 variants with inflammatory bowel disease and metabolite levels^{42,43}. In our study rs409336, near the CXCL5 gene, showed the strongest effect on plasma ENA78/CXCL5 levels. Because of the similarity in genetic influences on ENA78/CXCL5 levels and inflammatory bowel disease, it is possible that these traits share a common pathophysiological pathway and our findings support further investigation of the involvement of ENA78/CXCL5 in the etiology of inflammatory bowel disease.
We found two promising plasma protein biomarkers for MS: BLC and CD40. In our study rs7541151 in DDAH1 was associated with plasma BLC levels. DDAH1 is responsible for the degradation of ADMA into citrulline and dimethylamine, and previous studies showed an association of DDAH1 variants with MS and ADMA levels\(^{30,44}\). Previous studies indicate that CSF levels of BLC/CXCL13 may be an informative biomarker for studying treatment effects in MS\(^{45-47}\). Our results indicate plasma BLC/CXCL13 levels may be informative as well. The CD40 locus has been associated with MS\(^{30-32}\) but our study appears to be the first to associate CD40 plasma levels with CD40 variants. Plasma levels of CD40 have not been reported as a potential biomarker for MS, but our preliminary data suggests they may be a biomarker for MS. Although we did find a significant difference in CD40 levels in plasma between MS cases and controls, our sample size was small and only contained patients in remission so it would be prudent to evaluate a larger, more varied cohort to determine the possible utility of plasma levels of CD40 as an MS biomarker.

**Methods**

**Ethics Statement.** The Institutional Review Board (IRB) at the Washington University School of Medicine in Saint Louis approved the study. Research was carried out in accordance with the approved protocol. A written informed consent was obtained from participants and their family members by the Clinical Core of the Charles F. and Joanne Knight Alzheimer’s Disease Research Center (Knight-ADRC). The approval number for the Knight-ADRC Genetics Core family studies is 93-0006. The MS and control patients have signed the consent for the MS repository, approval number 201104379.

**Cohort descriptions.** Demographics of the samples included in this manuscript are reported in Table 1.

**Washington University Knight Alzheimer’s Disease Research Center (KADRC) cohort.** The KADRC sample included 124 AD cases and 188 cognitively normal controls. These individuals were evaluated by Clinical Core personnel of Washington University. Cases received a clinical diagnosis of Alzheimer’s disease in accordance with standard criteria and dementia severity was determined using the Clinical Dementia Rating (CDR)\(^{48}\). Plasma from all KADRC samples was collected in the morning after an overnight fast, immediately centrifuged, and stored at \(-80^\circ\text{C}\) until assayed according to standard procedures\(^{49}\).

**Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort.** The ADNI sample included 434 AD cases and 72 cognitively normal controls. Data used in the preparation of this article were obtained from the ADNI database (http://adni.loni.usc.edu/). See Supplementary Methods for further information about ADNI’s methods and for up-to-date information see http://www.adni-info.org/. Plasma was collected in the morning after an overnight fast, immediately centrifuged, and stored at \(-80^\circ\text{C}\) until assayed as described previously\(^{5}\). Genetic and phenotypic data for 506 samples was available for this study.

**Genotyping and Quality Control.** The ADNI protocol for collecting genomic DNA samples has been previously described\(^{30}\). All ADNI samples were genotyped using the Illumina Human610-Quad BeadChip, which contains over 600,000 SNP markers. KADRC samples were genotyped with the Human610-Quad BeadChip or the Omnixpress chip\(^{51}\). Prior to association analysis, all samples and genotypes underwent stringent QC. Genotype data was cleaned using PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/)\(^{32}\) by applying a minimum call rate for SNPs and individuals (98%) and minimum minor allele frequencies (MAF\(=0.02\)). SNPs not in Hardy-Weinberg equilibrium (\(P<1 \times 10^{-6}\)) were excluded. Gender identification was verified by analysis of X-chromosome SNPs. We tested for unanticipated duplicates and cryptic relatedness (Pihat \(\geq 0.5\)) using pairwise genome-wide estimates of proportion identity-by-descent using PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/)\(^{32}\). When a pair of identical samples or a pair of samples with cryptic relatedness was identified, the sample with a higher number of SNPs that passed QC was prioritized. EIGENSTRAT\(^{33}\) was used for each cohort separately to calculate principal component factors for each sample and confirm the ethnicity of the samples. The 1000 genomes data (June 2011 release) and BEAGLE v3.3.1\(^{14}\) were used to impute up to 6 million SNPs. SNPs with a BEAGLE \(R^2 < 0.3\), a minor allele frequency (MAF) \(< 0.025\), a call rate lower than 95%, a Gprobs score lower than 0.90 and those out of Hardy-Weinberg equilibrium (\(P<1 \times 10^{-8}\)) were removed. After imputation, 5,815,690 SNPs passed the QC process.

**Assessment of Analyte Profiles and Quality Control.** A set of 0.5 mL EDTA plasma samples from ADNI and KADRC participants was selected and shipped to Myriad Rules Based Medicine, Inc. (Myriad RBM, Austin, TX). A set of 190 protein levels from plasma for each selected individual was measured by multiplex immunoassay on the Human DiscoveryMAP panel v1.0 (https://rbm.myriad.com/products/services/humannmap-services/human-discoverymap/) using the LumineX100 platform by RBM. Samples with more than 10% of missing data across analytes were removed, then analytes were excluded if they had missing data for 10% of the samples or values were below the detection limit, in either of the studies. After the QC step, a total of 146 metabolites were included in each dataset of the present study.

**Statistical analyses.** For each study, prior to the analyses, all analyte values were log-transformed, standardized so the mean for each analyte was equal to zero, and outliers were removed as previously described\(^{12,51,55-59}\). Log-transformed, standardized values were tested for significant deviations from a normal distribution using the Shapiro-Wilk test. We performed a single variant analysis for each analyte using PLINK v1.9 (http://pngu.mgh.harvard.edu/purcell/plink/)\(^{32}\), including age, gender, AD status, and the first 2 principal components as covariates. The significance threshold for the joint analyses was defined as \(p < 5.0 \times 10^{-8}\) based on the commonly used threshold thought to be appropriate for the likely number of independent tests with Bonferroni correction. To approximate an internal replication, all SNPs that passed the genome-wide significance threshold had to pass...
the threshold $p < 0.05$ in single variant analyses of the individual datasets and had to have similar effect sizes in the same direction. To ensure that results were not confounded by AD status, single variant analyses were performed on all of the AD cases from both datasets separately from all of the controls from both datasets. All genome-wide significant SNPs from the joint analyses also had to have similar effect sizes in the same direction in the case-control stratified analyses. QQ plots were generated for each analysis to illustrate the distribution of the observed and expected $p$-values for all eligible SNPs. Regional plots showing LD and the location of nearby genes were generated for the top ranking SNPs for each metabolite using LocusZoom v1.1, build hg19/1000 Genomes Mar 2012 EUR (http://csg.sph.umich.edu/locuszoom/i). If more than one significant SNP clustered at a locus, the SNP with the smallest $p$-value was reported as the sentinel marker. All analyses were performed using BEAGLE v3.3.14, EIGENSTRAT53, SAS v9.2 for Linux (copyright © 2008 by SAS Institute Inc) and PLINK v1.07 and v1.9 (http://pngu.mgh.harvard.edu/purcell/plink/)52 software.

Meta-analyses. We performed the single variant analyses as described above for ADNI and KADRC separately. We used METAL (version released 2011-03-25, http://www.sph.umich.edu/csg/abecasis/Metal/index.html)62 to perform meta-analyses of the two datasets for each analyte by combining $p$-values across studies, weighting each study by its sample size.

Conditional analyses. To identify additional independent signals in a locus we conducted conditional analyses. We performed a series of sequential conditional analyses by adding the most strongly associated SNP into the regression model as a covariate and testing all remaining regional SNPs for association. This approach was used to determine additional secondary signals and was performed by adding SNPs one at a time until no significance was seen. Consistent with the locus-specific analysis statistical significance for the conditional analysis was defined at $p < 5.0 \times 10^{-8}$.

Annotation of GWAS hits. All significant GWAS SNPs were taken forward for functional annotation. We used SNPnexus (http://www.snp-nexus.org), build GRCh37/hg1963 and ANNOVAR version 2015-03-2264 to perform meta-analyses of the two datasets for each analyte by combining $p$-values across studies, weighting each study by its sample size.

Figure 3. Plasma levels of CD40 in MS cases versus controls. Levels of CD40 in plasma was significantly higher in MS cases than controls ($p = 0.041$, beta $= -1.837$).

References
1. Garge, N. et al. Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. Mol Cell Proteomics 9, 1383–1399, doi: 10.1074/mcp.M900378-MCP200 (2010).
2. Lourdusamy, A. et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum Mol Genet 21, 3719–3726, doi: 10.1093/hmg/ddt186 (2012).
3. Civelek, M. & Luisi, A. J. Systems genetics approaches to understand complex traits. Nat Rev Genet 15, 34–48, doi: 10.1038/nrg3575 (2014).
4. Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. Nat Genet 46, 543–550, doi: 10.1038/ng.2982 (2014).
5. Li, Y., Huang, J. & Amos, C. I. Genetic association analysis of complex diseases incorporating intermediate phenotype information. PLoS One 7, e46612, doi: 10.1371/journal.pone.0046612 (2012).
6. Melzer, D. et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet 4, e1000072, doi: 10.1371/journal.pgen.1000072 (2008).
7. Gieger, C. et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* **4**, e1000282, doi: 10.1371/journal.pgen.1000282 (2008).

8. Kettunen, J. et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269–276, doi: 10.1038/ng.1073 (2012).

9. Kim, S. et al. Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PLoS One* **8**, e70269, doi: 10.1371/journal.pone.0070269 (2013).

10. Yu, B. et al. Genetic determinants influencing human serum metabolome among African Americans. *PLoS Genet* **10**, e1004212, doi: 10.1371/journal.pgen.1004212 (2014).

11. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209–213, doi: 10.1038/ng1706 (2006).

12. Harari, O. et al. Phosphorylated tau-Abeta2 ratio as a continuous trait for biomarker discovery for early-stage Alzheimer’s disease in multiplex immunoassay panels of cerebrospinal fluid. *Brain Res Psychi**5**75**, 723–731, doi: 10.1016/j.brps.2013.11.032 (2014).

13. Yang, Q., Kathiresan, S., Lin, J. P., Tofler, G. H. & O’Donnell, C. J. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *BM**C Med Genet* **8** Suppl 1, S12, doi: 10.1186/1471-2350-8-S1-S12 (2007).

14. He, M. Genome-wide association study on dimethylarginines reveals novel AGXT2 variants associated with heart rate variability but not with overall mortality. *Eur Heart J* **35**, 524–531, doi: 10.1093/eurheartj/eht447 (2014).

15. Krumholz, M. H. et al. Chemokines in multiple sclerosis: CXCL12 and CXCL13 up-regulation is differentially linked to CNS immune cell recruitment. *Brain* **129**, 200–211, doi: 10.1093/brain/awh680 (2006).
Supplementary information accompanies this paper at http://www.nature.com/srep

Acknowledgements

This work was supported by Pfizer and grants from the National Institutes of Health (R01-NS085419; R01-AG044546, P01-AG003991, and R01-AG035083), and the Alzheimer Association (NIRG-11-200110). Dr. Piccio is a Harry Weaver Neuroscience Scholar of the National MS Society (JF 2144A2/1). This research was conducted while CC was a recipient of a New Investigator Award in Alzheimer’s disease from the American Federation for Aging Research. CC is a recipient of a BrightFocus Foundation Alzheimer’s Disease Research Grant (A2013359S). The recruitment and clinical characterization of research participants at Washington University were supported by NIH P50 AG05681, P01 AG003991, and P01 AG026276. Some of the samples used in this study were genotyped by the ADGC and GERAD. ADGC is supported by grants from the NIH (#U01AG032984) and GERAD from the Wellcome Trust (GR082604MA) and the Medical Research Council (G0300429). Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG049904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Euroimmun; F Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; and IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Rev December 5, 2013 Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders and the Departments of Neurology and Psychiatry at Washington University School of Medicine.

Author Contributions

Y.D. and J.X. contributed equally to this work: analyzed data and wrote the manuscript. Y.C., J.L., J.L-D.-A., M.V.F., D.C., K.B., J.B., S.M., B.S., B.H., S.B., M.B. and P.G.R. prepared genetic data: performed imputation, cleaning, and calculated principal components. A.D.N.I., D.H., J.C.M., K.B., E.H.P., J.M.L., L.H., J.K., A.G. and L.P. provided data. C.C. prepared manuscript and supervised the project.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep
How to cite this article: Deming, Y. et al. Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits. Sci. Rep. 6, 18092; doi: 10.1038/srep18092 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
21. Cleveland Clinic  
22. Columbia University Medical Center  
23. Cornell University  
24. Dartmouth Hitchcock Medical Center  
25. Dent Neurologic Institute  
26. Department Of Defense  
27. Duke University Medical Center  
28. Eli Lilly  
29. Emory University  
30. Georgetown University Medical Center  
31. Hartford Hospital  
32. Harvard Medical School  
33. Howard University  
34. Indiana University  
35. Johns Hopkins University  
36. Mayo Clinic, Jacksonville  
37. Mayo Clinic, Rochester  
38. McGill University  
39. Medical University South Carolina  
40. Mount Sinai School of Medicine  
41. Nathan Kline Institute  
42. National Institute of Mental Health  
43. National Institute on Aging  
44. Neurological Care of CNY  
45. New York University  
46. Northwestern University  
47. Ohio State University  
48. Oregon Health and Science University  
49. Parkwood Hospital  
50. Premiere Research Institute  
51. Prevent Alzheimer's Disease 2020 Inc.  
52. Rhode Island Hospital  
53. Richard Frank Consulting  
54. Rush University Medical Center  
55. Siemens  
56. St Joseph's Health Care Ontario  
57. Stanford University  
58. Sunnybrook Health Sciences Ontario  
59. U.B.C. Clinic for AD & Related Disorders  
60. University College London  
61. University Iowa College Medicine  
62. University Kansas Medical Center  
63. University of Alabama at Birmingham  
64. University of California, Davis - Sacramento  
65. University of California, Irvine  
66. University of California, Berkeley  
67. University of California, Los Angeles  
68. University of California, San Diego  
69. University of California, San Francisco  
70. University of Kentucky  
71. University of Michigan  
72. University of Pennsylvania  
73. University of Pennsylvania School of Medicine  
74. University of Southern California  
75. University of Texas Southwestern Medical School  
76. University of Utah  
77. University of Washington  
78. University of Wisconsin  
79. University Rochester Medical Center  
80. University South Florida  
81. University of Pittsburgh  
82. Wake Forest University  
83. Washington University, St. Louis  
84. Wien Center  
85. Yale School of Medicine  
86. Yeshiva University