Human Molecular Genetics, 2012, Vol. 21, No. 16 **3719–3726**
doi:10.1093/hmg/dds186
Advance Access published on May 16, 2012

Identification of *cis*-regulatory variation influencing protein abundance levels in human plasma

Anbarasu Lourdusamy[1,*], Stephan Newhouse[1], Katie Lunnon[1], Petra Proitsi[2], John Powell[2], Angela Hodges[1], Sally K. Nelson[3], Alex Stewart[3], Stephen Williams[3], Iwona Kloszewska[4], Patrizia Mecocci[5], Hilkka Soininen[6], Magda Tsolaki[7], Bruno Vellas[8], Simon Lovestone[1,2], on behalf of the AddNeuroMed Consortium, and Richard Dobson[1], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Foundation Trust and Institute of Psychiatry and [2]MRC Centre for Neurodegeneration, Kings College London, London, UK [3]SomaLogic, Inc., 2945 Wilderness Place, Boulder, CO 80301, USA, [4]Department of Old Age Psychiatry and Psychotic Disorders, Medical University of Lodz, Lodz, Poland, [5]Section of Gerontology and Geriatrics, Department of Clinical and Experimental Medicine, University of Perugia, Perugia, Italy, [6]Department of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland, [7]Third Department of Neurology, Aristotle University of Thessaloniki, Thessaloniki, Greece and [8]Department of Internal and Geriatrics Medicine, Hôpitaux de Toulouse, Toulouse, France

Received December 14, 2011; Revised May 1, 2012; Accepted May 11, 2012

Proteins are central to almost all cellular processes, and dysregulation of expression and function is associated with a range of disorders. A number of studies in human have recently shown that genetic factors significantly contribute gene expression variation. In contrast, very little is known about the genetic basis of variation in protein abundance in man. Here, we assayed the abundance levels of proteins in plasma from 96 elderly Europeans using a new aptamer-based proteomic technology and performed genome-wide local (*cis*-) regulatory association analysis to identify protein quantitative trait loci (pQTL). We detected robust *cis*-associations for 60 proteins at a false discovery rate of 5%. The most highly significant single nucleotide polymorphism detected was rs7021589 (false discovery rate, 2.5×10^{-12}), mapped within the gene coding sequence of Tenascin C (TNC). Importantly, we identified evidence of *cis*-regulatory variation for 20 previously disease-associated genes encoding protein, including variants with strong evidence of disease association show significant association with protein abundance levels. These results demonstrate that common genetic variants contribute to the differences in protein abundance levels in human plasma. Identification of pQTLs will significantly enhance our ability to discover and comprehend the biological and functional consequences of loci identified from genome-wide association study of complex traits. This is the first large-scale genetic association study of proteins in plasma measured using a novel, highly multiplexed slow off-rate modified aptamer (SOMAmer) proteomic platform.

*To whom correspondence should be addressed at: NIHR Biomedical Research Centre for Mental Health, Institute of Psychiatry, Kings College London, Box P092, De Crespigny Park, London SE5 8AF, UK. Tel: +44 2078485495; Email: anbarasu.lourdusamy@kcl.ac.uk

†Data used in preparation of this article were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (www. loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

INTRODUCTION

The Genome-wide association studies (GWAS) have successfully identified several common susceptibility variants associated with complex traits. However, multiple variants of small to intermediate effect sizes identified by GWAS often explain small proportion of trait's estimated heritability (1,2). For example, the estimated genetic heritability of type 2 diabetes (T2D) is 40%, but 40 loci identified by T2D GWAS to date collectively explain <10% of the variation in T2D (3–5). Furthermore, GWAS provides a limited insight into the mechanisms through which common genetic variant influences trait variations. Studies that link genetic variants with gene expression, protein abundance and other intermediate phenotypes that manifest as trait variation may provide a means to reveal the underlying mechanisms.

Recent research in human and model organisms has shown that genetic variants influence gene expression. Several studies have identified expression quantitative trait loci (eQTL), showing local variations contribute to the transcript expression levels, in yeast and mice using recombinant inbred strains (6–11). Gene expression values as quantitative traits have been reported in humans using lymphocytes, Epstein–Barr virus-transformed lymphoblastoid cell lines (LCLs) and tissues including the adipose, blood, brain and liver (12–18). These studies have demonstrated that mRNA expressions are highly heritable and genetic factors account for a substantial proportion of the variation in human gene expression. In addition, these studies showed that variants mapping primarily in the vicinity of the gene (*cis*-effect) have stronger effects than variants mapping at distances >1 mb or located on different chromosomes (*trans*-effect). In some cases, eQTLs have been shown to be associated with diseases, but in general the functional significance of eQTLs remains unknown (19–22).

Most genes encode for proteins; therefore, functionally important changes in mRNA expression are expected to be reflected in changes in the levels of corresponding proteins. However, a weak correlation between transcript and protein levels in yeast shows that various other mechanisms of post-transcriptional regulation can lead to changes in protein abundance in the absence of a corresponding transcript effect (23). Thus, as a more proximal component of trait variations, direct examination of the proteome in relation to common genetic variants may provide biological insights and disease biomarkers that cannot be captured through evaluation of the transcriptome alone. Efforts to identify genetic variants associated with protein abundance in human have lagged behind when compared with many published studies on eQTL identification. To our knowledge, only two studies have so far explored an association between single nucleotide polymorphisms (SNPs) and protein abundance levels in human (24,25). One study described experiments associating SNPs with 42 serum proteins measured in a population of 1200 individuals (24). This study resulted in the identification of protein QTL associating with eight serum proteins (*cis*-effects) including interleukin-6 receptor (IL6R) and one *trans*-effect between ABO blood group and tumor necrosis factor-alpha. Using 2D difference gel electrophoresis approach, the second study measured 544 proteins in a population of 24 human LCLs and identified protein quantitative trait loci (pQTLs) for 15 proteins (25).

The aim of the present study was to identify polymorphic variants significantly associated with protein abundance levels in human plasma. We conducted a genome-wide association analysis using protein levels as the primary phenotype. We used an aptamer-based proteomic technology to quantify protein abundances of 813 proteins measured from the plasma of 96 elderly healthy individuals, at risk for age-related chronic disease.

RESULTS

The study cohort of normal elderly individuals consisted of 42 males (43.8%) and 54 females (56.2%) with a mean age of 72.1 years (ranging from 52 to 87). A total of 778 proteins were used for the association analysis after removing three proteins for which there were no SNPs found in Illumina Hap610 SNP array. These 778 proteins represent a wide range of protein families including signaling proteins, cytokines, growth factors and kinases (Supplementary Material, Fig. S1). We imputed genotypes with reference to genotypes from 1000G EUR (see 'MATERIALS AND METHODS'). We chose imputed SNPs with minor allele frequencies (MAFs) of >5% and MACH RSQ values of ≥0.3 for further analyses. (The MACH RSQ value is a post-imputation quality score that indicates the correlation between true and estimated allele counts of imputed SNPs.) Applying these criteria, 776 864 genetic variants were successfully imputed for regions of genes encoding for 778 proteins.

We identified *cis*-SNPs, by testing for association between protein abundance and SNP genotypes, within 300 kb of the gene encoding the relevant protein. SNPs were considered to be significantly associated with protein abundance levels if they passed the threshold of genome-wide false discovery rate (FDR), Q <0.05. A total of 2016 SNPs representing 60 of the 778 proteins tested show significant evidence for *cis*-effect variation (Supplementary Material, Table S4). The most significantly associated SNP for each protein is reported in Table 1 (top SNPs in Table 1). Examples of these *cis*-SNPs are shown in Figure 1 (Supplementary Material, Fig. S2). The most highly significant SNP detected was rs7021589 (FDR, 2.5×10^{-12}), mapped within the gene coding sequence of Tenascin C (TNC).

Genetic variants in the protein-coding region were found to be enriched in significant *cis*-SNPs when compared with variants in non-coding regions (OR = 1.88, $P < 6.9 \times 10^{-04}$; Fisher's exact test). A total of 35 SNPs were detected in the coding regions of 20 proteins involved in axon guidance (EPHA1), complement and coagulation cascades pathway [complement factor H (CFH), kininogen 1 (KNG1)], IL10 signaling [FCGR2A, IL1RL1, lipopolysaccharide-binding protein (LBP), FCGR2B], immune response [cathepsin S (CTSS)], serine-type endopeptidase inhibitor activity (CD109, PRTN3) and T-cell receptor signaling (SIGLEC9) (Supplementary Material, Table S1). The non-synonymous coding SNP rs1065489 in CFH alters the amino acid Glu to Asp at position 936, whereas the SNP rs2304456 in KNG1 causes Ile to Met alteration at position 197. The amino acid substitution in KNG1 is predicted to be damaging (Supplementary Material, Table S1). The rs267738 in CTSS is a

Table 1. Summary of significant *cis*-SNPs detected for protein abundance variation

SwissProt	Gene	Top SNP	Chr	Position	Alleles (R/E)	β (SE)	Unadjusted *P*-value	FDR
P24821	TNC	rs7021589	9	117 804 667	T/C	2.05 (0.24)	1.50×10^{-17}	2.52×10^{-12}
Q9Y336	SIGLEC9	rs2075803	19	51 628 529	A/G	1.50 (0.181)	9.63×10^{-17}	8.53×10^{-12}
P09619	PDGFRB	rs2240781	5	149 516 480	C/T	−0.45 (0.06)	4.42×10^{-15}	2.40×10^{-10}
P12318	FCGR2A	rs1801274	1	161 479 745	A/G	−1.35 (0.17)	5.17×10^{-15}	2.64×10^{-10}
P20138	CD33	rs12985029	19	51 713 365	G/A	−0.57 (0.08)	2.65×10^{-13}	3.74×10^{-09}
P13385	TDGF1	chr3:46619238	3	46 619 238	T/A	0.22 (0.03)	7.10×10^{-13}	8.22×10^{-09}
Q9NPH3	IL1RAP	rs724608	3	190 348 810	A/G	0.64 (0.09)	8.81×10^{-13}	9.89×10^{-09}
P08887	IL6R	rs11265613	1	154 418 415	T/C	0.31 (0.04)	1.43×10^{-12}	1.54×10^{-08}
P31994	FCGR2B	rs6665610	1	161 641 384	G/A	1.06 (0.16)	1.27×10^{-11}	9.03×10^{-08}
O15467	CCL16	chr17:34306470	17	34 306 470	G/C	−1.27 (0.19)	2.52×10^{-11}	1.67×10^{-07}
P02735	SAA1	chr11:18290906	11	18 290 906	G/A	1.78 (0.28)	1.16×10^{-10}	6.69×10^{-07}
Q6YHK3	CD109	rs9447004	6	74 458 737	A/G	0.40 (0.07)	8.61×10^{-10}	4.29×10^{-06}
P11226	MBL2	rs7899547	10	54 536 839	T/G	−0.83 (0.14)	1.60×10^{-09}	7.59×10^{-06}
P24666	ACP1	rs17713879	2	254 215	G/A	0.27 (0.05)	2.46×10^{-09}	9.96×10^{-06}
P08709	F7	rs474671	13	113 776 218	T/C	−0.68 (0.12)	2.75×10^{-09}	1.09×10^{-05}
P06681	C2	chr6:31981247	6	31 981 247	T/C	−0.48 (0.08)	3.70×10^{-09}	1.39×10^{-05}
P08603	CFH	rs1048663	1	196 674 982	G/A	−0.11 (0.02)	2.94×10^{-08}	7.38×10^{-05}
Q13231	CHIT1	rs2486950	1	203 174 670	G/C	−0.49 (0.09)	3.15×10^{-08}	7.68×10^{-05}
P21709	EPHA1	rs1804527	7	143 088 823	T/C	−0.52 (0.09)	4.34×10^{-08}	9.93×10^{-05}
Q76LX8	ADAMTS13	rs28647808	9	136 305 530	C/G	−0.45 (0.08)	5.45×10^{-08}	1.17×10^{-04}
P02765	AHSG	rs2070632	3	186 334 004	A/C	−0.20 (0.04)	7.31×10^{-08}	1.48×10^{-04}
Q9GZX6	IL22	rs7133527	12	68 357 125	G/A	0.50 (0.09)	4.60×10^{-07}	5.31×10^{-04}
Q96F46	IL17RA	rs2241047	22	17 586 583	C/G	0.37 (0.07)	6.35×10^{-07}	7.03×10^{-04}
P51665	PSMD7	rs7359422	16	74 592 483	T/C	0.19 (0.04)	8.49×10^{-07}	9.14×10^{-04}
Q01638	IL1RL1	rs12470864	2	102 926 362	G/A	−0.27 (0.06)	8.85×10^{-07}	9.49×10^{-04}
P24158	PRTN3	rs7251804	19	839 948	G/C	0.23 (0.05)	9.49×10^{-07}	1.01×10^{-03}
Q99665	IL12RB2	chr1:67489003	1	67 489 003	T/C	0.76 (0.16)	9.82×10^{-07}	1.03×10^{-03}
Q9BX67	JAM3	rs11223763	11	134 214 549	C/T	0.39 (0.08)	1.60×10^{-06}	1.61×10^{-03}
P18428	LBP	rs11536949	20	36 985 885	G/T	−0.72 (0.15)	2.17×10^{-06}	2.10×10^{-03}
P09758	TACSTD2	rs232815	1	59 080 879	T/C	0.22 (0.05)	2.18×10^{-06}	2.10×10^{-03}
Q12904	AIMP1	chr4:107298048	4	107 298 048	G/A	0.29 (0.06)	3.56×10^{-06}	3.17×10^{-03}
Q9BXR6	CFHR5	rs6695321	1	196 675 861	A/G	−0.30 (0.07)	4.13×10^{-06}	3.61×10^{-03}
P22749	GNLY	rs7577293	2	85 935 282	C/T	0.35 (0.08)	4.85×10^{-06}	4.13×10^{-03}
P49961	ENTPD1	rs72822592	10	97 180 509	G/C	0.38 (0.09)	5.54×10^{-06}	4.62×10^{-03}
P42830	CXCL5	chr4:74596624	4	74 596 624	A/T	0.50 (0.11)	7.29×10^{-06}	5.80×10^{-03}
Q13478	IL18R1	chr2:102951851	2	102 951 851	G/A	0.14 (0.03)	8.26×10^{-06}	6.42×10^{-03}
P01042	KNG1	rs5030049	3	186 450 863	T/C	−0.33 (0.08)	1.11×10^{-05}	8.18×10^{-03}
P47992	XCL1	rs1933112	1	168 521 417	A/G	0.14 (0.031)	1.33×10^{-05}	9.36×10^{-03}
P55285	CDH6	rs13165280	5	31 578 552	G/A	0.19 (0.04)	1.39×10^{-05}	9.71×10^{-03}
P14555	PLA2G2A	rs11573156	1	20 306 146	G/C	0.39 (0.09)	1.95×10^{-05}	1.29×10^{-02}
Q16663	CCL15	rs854624	17	34 327 923	G/T	0.40 (0.09)	2.00×10^{-05}	1.32×10^{-02}
P10636	MAPT	rs73317026	17	44 081 268	G/A	0.12 (0.03)	2.12×10^{-05}	1.39×10^{-02}
P01374	LTA	rs9267054	6	31 397 367	T/C	0.43 (0.1)	2.21×10^{-05}	1.44×10^{-02}
P05771	PRKCB	rs7499480	16	23 825 207	A/G	−0.15 (0.04)	2.95×10^{-05}	1.69×10^{-02}
Q9H293	IL25	chr14:23770877	14	23 770 877	T/C	0.62 (0.15)	3.00×10^{-05}	1.71×10^{-02}
P02649	APOE	rs2967668	19	45 302 951	A/G	−0.30 (0.07)	4.15×10^{-05}	2.21×10^{-02}
Q9Y337	KLK5	rs266863	19	51 355 650	T/C	0.134 (0.03)	4.50×10^{-05}	2.35×10^{-02}
Q01973	ROR1	rs71499308	1	64 666 320	G/A	−0.14 (0.033)	4.68×10^{-05}	2.43×10^{-02}
O95711	LY86	rs9405302	6	6 477 849	G/A	0.19 (0.05)	5.10×10^{-05}	2.57×10^{-02}
P22301	IL10	chr1:207234422	1	207 234 422	T/A	0.32 (0.08)	6.04×10^{-05}	2.91×10^{-02}
P12272	PTHLH	chr12:27848777	12	27 848 777	T/C	0.20 (0.05)	6.25×10^{-05}	3.00×10^{-02}
P0C0L4	C4A	rs2844452	6	31 882 024	A/G	−0.16 (0.04)	6.72×10^{-05}	3.10×10^{-02}
P25774	CTSS	chr1:150868102	1	150 868 102	C/A	−0.10 (0.03)	7.65×10^{-05}	3.41×10^{-02}
P05362	ICAM1	chr19:10082034	19	10 082 034	G/A	0.26 (0.07)	9.87×10^{-05}	4.11×10^{-02}
Q15056	EIF4H	rs12538827	7	73 792 436	G/A	0.16 (0.04)	9.99×10^{-05}	4.16×10^{-02}
Q96GD0	PDXP	rs5995497	22	38 182 706	A/G	−0.27 (0.07)	1.00×10^{-04}	4.16×10^{-02}
P35625	TIMP3	rs2413151	22	33 167 869	C/T	0.56 (0.15)	1.07×10^{-04}	4.40×10^{-02}
P16435	POR	rs11770797	7	75 855 511	A/G	0.14 (0.04)	1.09×10^{-04}	4.48×10^{-02}
Q96IY4	CPB2	rs2094247	13	46 602 162	A/G	0.10 (0.03)	1.17×10^{-04}	4.74×10^{-02}
Q9ULZ9	MMP17	rs10751701	12	132 330 735	C/G	0.28 (0.07)	1.21×10^{-04}	4.87×10^{-02}

Chr, chromosome; Alleles (R/E), alleles are given as the reference (R)/effect (E) allele; β (SE), beta, change in phenotype per allele copy and its standard error estimated from the linear model.

non-synonymous SNP that causes Glu to Ala, Gly and Val alterations at position 115 and the change Glu115Ala is predicted to be conservative and deleterious (Supplementary Material, Table S1). The papain family cysteine protease protein CTSS is localized in lysosome and interacts with KNG1 (26). Alterations caused by coding variants

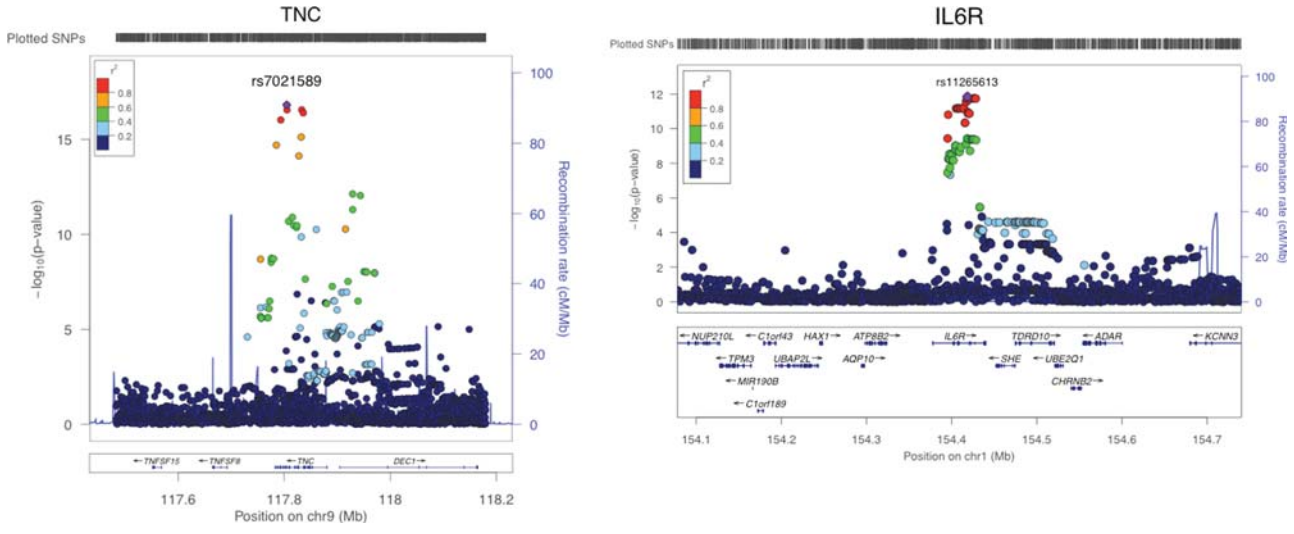

Figure 1. *Cis*-association results. The most significant *cis*-association was found for rs7021589 in TNC (left panel). The SNP rs11265613 in IL6R was significantly associated with the IL6R protein abundance (right panel), Here, *x*-axis shows chromosomal position of genomic region, left *y*-axis shows the −log10 *P*-values and the *y*-axis at the right shows the recombination rate calculated from HapMap phase II CEU.

rs2228145 (Asp358Ala) in IL6R, rs28647808 (Pro618Ala) in ADAM metallopeptidase with thrombospondin type 1 motif, 13 (ADAMTS13), rs4917 (Met248Thr) in alpha-2-HS-glycoprotein (AHSG), rs1800450 (Gly54Asp) in mannose-binding lectin (protein C) 2, soluble (MBL2) and rs2232613 (Pro333Leu) in LBP are predicted to be damaging.

More than 30% of the identified SNPs mapped onto the intronic region of the gene coding for the proteins, which includes platelet-derived growth factor receptor, beta polypeptide (PDGFRB), teratocarcinoma-derived growth factor 1 (TDGF1), IL1R accessory protein (IL1RAP), TNC, chemokine (C-C motif) ligand 16 (CCL16), coagulation factor VII (F7), ADAMTS13 and microtubule-associated protein tau (MAPT). One of the significant SNPs detected for IL6R protein on chromosome 1 was rs4129267 (FDR, 1.86×10^{-08}). The association of rs4129267 with IL6R protein level has been previously reported in an ELISA-based study (24).

Twenty of the 60 genes encoding protein have previously been associated with diseases and traits by GWAS studies. In particular, 16 *cis*-SNPs have been associated with different disease including Alzheimer's disease (Supplementary Material, Table S2). The SNP rs1801274 identified in FCGR2A is associated with ulcerative colitis (27,28), rs4129267 in IL6R is associated with asthma and C-reactive protein levels (29,30), rs1420101 in IL1RL1 is associated with plasma eosinophil count and asthma (31), and most significantly in a study of elderly people, two SNPs rs3826656 and rs3865444 in CD33 have been associated with Alzheimer's disease (32,33). We detected cases where identified *cis*-SNPs are associated with more than two diseases and traits (Supplementary Material, Table S2). For example, the SNP rs4129267 is associated with IL6R serum protein levels, pulmonary function traits, asthma and C-reactive protein (34); and the IL6R SNP rs4129267 is in strong linkage disequilibrium (LD) with rs8192284 ($r^2 = 0.98$) that is associated Fibrinogen levels (35). In addition, we found evidence for genetic associations of gene encoding proteins in our study with different phenotypes available in the GWAS catalogue (Supplementary Material, Table S2). Genetic variants in the gene IL1RAP are associated with lung cancer and waist circumference (36,37); variants in F7 are associated with hematological phenotypes (38); variant in EPHA1 is associated with Alzheimer's disease (32,33); SNPs in CFH are associated with nephropathy and age-related macular degeneration (39,40); and variants in MAPT are associated with progressive supranuclear palsy and Parkinson's disease (41,42).

We assessed whether identified *cis*-SNPs in our study were also the SNPs most strongly associated with the gene expression using publically available resources (43,44). A total of 32 *cis*-SNPs have been associated with the expression levels of 23 genes in different tissue types. The SNP, rs723177 is associated with FCGR2A gene expression in Epstein–Barr virus-transformed LCLs ($\beta = 0.47$, $P < 1.10 \times 10^{-07}$) derived from children and the expression of FCGR2A is highly heritable ($h^2 = 9.09$) (Supplementary Material, Table S3a). It is remarkable that the effect allele C of rs723177 increases both gene expression and protein abundance levels ($\beta = 0.68$, FDR $< 7.41 \times 10^{-05}$). We found evidences of strong SNP-gene expression association for several *cis*-SNPs from the meta-analysis of HapMap human LCLs (Supplementary Material, Table S3b). Interestingly, we found six SNP-gene expression associations for IL16 in LCL and three associations for SURF6 in the brain frontal cortex from the genotype-tissue expression (GTEx) expression QTL browser (Supplementary Material, Table S3c).

DISCUSSION

The results of the present study show that common genetic variants influence the abundance of plasma proteins. This study provides the first assessment of pQTL on a large scale. Earlier candidate protein studies have merely explored genetic associations with single-protein measures, or small

numbers of proteins as measures by ELISA (24). We used aptamer-based technology [slow off-rate modified aptamer (SOMAmers)] to quantify the abundance levels of proteins. SOMAmers are single-stranded nucleic acids that form well-defined three-dimensional shapes, allowing them to bind target molecules in a manner that is conceptually similar to antibodies. SOMAmers combine the optimal characteristics of small molecules and antibodies, including high specificity and affinity, chemical stability and the ability to target protein–protein interactions (44,45).

Large-scale proteomic and high-density genotyping in plasma from elderly normal individuals has allowed us to test 776 864 SNP–protein associations representing 778 proteins. Given the sample size of 96 individuals, our study is not well powered to detect *trans*-effects and hence we tested only *cis*-associations between each proteins and SNPs within 300 kb of the gene encoding the relevant protein. The Bonferroni corrected *P*-value for testing 776 864 SNPs is $P < 6.64 \times 10^{-08}$. LD exists between SNPs but they are assumed to be completely independent in the Bonferroni method of adjusting for multiple comparisons. LD-based pruning may used to estimate the number of independent SNPs. However, the estimates obtained with LD pruning may be more or less conservative under varying levels of LD (at $r^2 = 0.8$, there are 172 248 independent SNPs in our study, at $r^2 = 0.5$, 103 517 independent SNPs and at $r^2 = 0.2$, there are only 43 300 independent SNPs). Given that the Bonferroni method is likely to be conservative and the LD pruning is based on the choice of the LD level, we used the FDR to identify significant SNP–protein associations.

A total of 2106 unique SNP–protein associations representing 60 proteins were identified at FDR <0.05 (Supplementary Material, Table S4). Eight of 60 proteins involved in LXR/RXR activation (KNG1, APOE, IL1RL1, AHSG, C4B, LBP, SAA1 and IL1RAP), five proteins in cytokine–cytokine receptor interaction pathway (CCL16, IL18R1, IL6R, PDGFRB and XCL1), four proteins involved in complement and coagulation cascades (C2, CFH, F7, KNG1 and MBL2), and two in Fc gamma R-mediated phagocytosis pathway (FCGR2A, FCGR2B). One of the *cis*-SNPs, rs4129267 is significantly associated with the abundance of IL6R protein. This association was reported in a previous pQTL study, where 42 proteins were measured in human serum using ELISA (24). It is noteworthy that the direction of the effect is in the same direction as reported in Melzer *et al*. study although the protein abundance was measured by two different approaches. To confirm the robustness of our *cis*-SNPs finding, we used the proteomic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) plasma-based biomarker study, which includes 58 controls. A total of 190 proteins were measured on the Luminex xMAP platform by Rules-Based Medicine in ADNI and six of them overlapped with our data. Five *cis*-SNPs representing two genes (IL6R and TNC) were marginally associated with protein abundance levels (Supplementary Material, Table S5). Further studies with large sample sizes would be required to assess the robustness of our *cis*-SNP findings.

Our protein QTL analyses provide several insights regarding the role of genetic variants identified by GWA studies of common disease/traits. In several instances, disease-associated variant influences the protein abundance levels. For example, rs3865444 on chromosome 19q13.3 is strongly associated with AD in a meta-analysis of several case–control studies (OR = 0.91, $P < 1.9 \times 10^{-09}$). The effect allele A of rs3865444 reduces the protein abundance of CD33 ($\beta = -0.45$, FDR $<5.06 \times 10^{-09}$) indicating that the variant might influence AD susceptibility through a mechanism of altered protein abundance. CD33 is a member of sialic acid-binding immunoglobulin-like lectin (Siglec) family, which regulates functions of cell in the innate and adaptive immune systems (46). SIGLEC9 is one of the CD33-related Siglecs located 100.2 kb proximal to CD33 and involved in down-regulating innate immune responses (47). Interestingly, we detected a non-synonymous *cis*-SNP rs2075803 (Lys100Glu) that modulates SIGLEC9 protein abundance levels indicating SIGLEC9 is a plausible biological candidate for AD. In some instances, *cis*-SNPs are associated with susceptibility to more than one trait indicating genetic pleiotropy (Supplementary Material, Table S2). We also note several examples of disease-associated variants from the same gene encoding the protein that are different from *cis*-SNPs. For example, the SNP rs11767557 is strongly associated with AD (OR = 0.90, $P < 6.0 \times 10^{-10}$) and is not associated with protein abundance levels ($P = 0.038$), whereas the non-synonymous SNP rs4725617 located 12.04 kb proximal to rs11767557 ($r^2 = 0.005$ and $D' = 0.490$) is strongly associated with protein abundance levels.

As observed in several eQTL studies, our study showed that protein QTLs can be mapped by GWAS approaches and *cis*-SNPs are within the proximity of the gene encoding the protein. For the *cis*-SNP rs723177 in FCGR2A, we found an evidence of gene expression variation in LCLs indicating *cis*-regulatory genetic variant modulating both PDGRFB transcript and protein levels. For 23 *cis*-associated proteins, we have found several SNPs in the same gene encoding the protein associated with gene expression levels indicating possible simultaneous genetic perturbations on transcript and protein levels. Although it is assumed that genetic perturbations of transcript levels are correlated with protein levels, molecular events such as alternative splicing, transport and localization, translational efficiency and degradation, all of which influence protein levels independently of transcripts. It is important to note that the plasma protein levels measured in our elderly individuals may be affected by the constellation of changes that occur with advanced age including the environment. Although we remove the effect of age in our SNP–protein association testing, the altered protein levels induced by genetic variants and its temporal interaction with environment remains to be investigated.

In summary, using a population-based genetic mapping of proteomics variation, we have identified common genetic variants that influence abundance levels of 60 proteins in human plasma. pQTLs identified in this study represent a subset of total genetic variation that are potential candidates for the involvement in human phenotypes including diseases. Genetic variants that influence the protein abundance have a clear impact at the cellular level and result in phenotypic differences. The subsequent identification of pQTLs across various cell types and tissues could facilitate identification of all regulatory variation relevant in complex traits and diseases.

MATERIALS AND METHODS

Samples and diagnostic criteria

A total of 100 normal elderly individuals were selected from the AddNeuroMed study, a European multi-center study for the discovery and validation of biomarkers for Alzheimer's disease. These samples were collected from six centers: Aristotle University of Thessaloniki (Greece), King's College London (UK), Medical University of Lodz (Poland), University of Kuopio (Finland), University of Perugia (Italy) and University of Toulouse (France). Informed consent was obtained for all subjects, and the relevant institutional review board at each data acquisition site approved protocols and procedures.

Protein abundance data

Plasma samples from the remaining 96 cognitively normal individuals were analyzed using an aptamer-based proteomic technology, which is capable of simultaneously measuring thousands of proteins (44,45). This technology uses a new class of DNA-based aptamer, the SOMAmer that contains chemically modified nucleotides. Using a multiplexed assay, the quantity of each targeted protein is transformed into a corresponding quantity of aptamer, which is quantified on a custom DNA microarray. Protein quantities are recorded as relative fluorescent units, which can be converted to concentrations by comparison with standard curves. For each sample, the protein levels of 813 proteins were assayed. Thirty-two proteins that are highly correlated ($r > 0.5$) with kidney function marker proteins cystatin-C and beta-2 microglobin were removed from further analyses.

Genotyping and imputation

The AddNeuroMed samples were genotyped using the Illumnia HumanHap610—Qaud Beadchip at CNG (Centre National de Génotypage CNG, France). Genotype quality control (QC) was performed with software packages PLINK (48) and EIGENSTRAT (49,50). Four individuals were excluded owing to the low genotyping call rate (<98% chip—wide genotyping success). We further removed SNPs with call rates <98%, MAF <0.05 and Hardy–Weinberg equilibrium, $P < 10^{-5}$. All individuals were screened for relatedness using PLINK by calculating the pairwise estimated proportion of alleles shared identical-by-descent (referred as Pi-hat in PLINK) on 113 554 randomly distributed markers throughout the genome. Pairwise Pihat values in excess of 0.01 were indicative of relatedness. This step did not eliminate any individuals, indicating un-relatedness in the study participants. Following QC measures, we assessed the population structure of AddNeuroMed participants using principle component analysis implemented in EIGENSTRAT. No further samples were excluded from further analyses. Imputation was performed on the cleaned data using Minimac software, which uses MaCH algorithm for genotype imputation (51,52). Imputation included a 1000 Genomes (1000G) imputation reference panel for European population (EUR) in NCBI Build 37 from the 1000G Interim Phase I data release (June 2011).

Genome-wide association analysis with protein measures

The abundance of each protein was log transformed and then adjusted for age, gender and study site using linear regression (LR) models in R (53). The residuals from the regression were used as the phenotype values for all subsequent analyses. For association testing, the allelic dosages, which represent the expected number of copies of a distinct allele rather than the best-guess imputed genotype of each SNP, were analyzed in an LR framework in order to account for imputation uncertainty. Adjustment for the population structure was performed with MACH2QTL software by including the first five principal components derived from an EIGENSTRAT analysis of genotype data as covariates. For the *cis*-analysis, the association of SNPs with abundance levels was calculated for each protein within a 300-kb window around the gene coding for that protein. Regression *P*-values were adjusted to FDR by the q-value procedure implemented in R Bioconductor package Q-value (54). We used FDR 5% as a threshold to identify significant *cis*-genetic associations.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank all the research participants for their generous involvement in this research. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of 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 California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

Conflict of Interest statement. S.K.N., A.S. and S.W. are employees of SomaLogic, Inc. and the proteomic assay

reported in this manuscript was performed using reagents supplied by SomaLogic, Inc.

FUNDING

This work was supported by InnoMed (Innovative Medicines in Europe), an Integrated Project funded by the European Union of the Sixth Framework program priority (FP6-2004-LIFESCIHEALTH-5); the Alzheimer's Research Trust; The John and Lucille van Geest Foundation and the NIHR Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and (Institute of Psychiatry) Kings College London; Kuopio University Hospital (H.S.) and funding from UEFBRAIN (H.S.).

REFERENCES

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
3. Poulsen, P., Kyvik, K.O., Vaag, A. and Beck-Nielsen, H. (1999) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. *Diabetologia*, **42**, 139–145.
4. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G. *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638–645.
5. Hakonarson, H. and Grant, S.F. (2011) GWAS and its impact on elucidating the etiology of diabetes. *Diabetes Metab. Res. Rev.*, 10.1002/dmrr.1221.
6. Brem, R.B. and Kruglyak, L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 1572–1577.
7. Ronald, J., Brem, R.B., Whittle, J. and Kruglyak, L. (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.*, **1**, e25.
8. Ghazalpour, A., Doss, S., Sheth, S.S., Ingram-Drake, L.A., Schadt, E.E., Lusis, A.J. and Drake, T.A. (2005) Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol.*, **6**, R59.
9. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
10. Quigley, D.A., To, M.D., Perez-Losada, J., Pelorosso, F.G., Mao, J.H., Nagase, H., Ginzinger, D.G. and Balmain, A. (2009) Genetic architecture of mouse skin inflammation and tumour susceptibility. *Nature*, **458**, 505–508.
11. Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005) Cis-acting expression quantitative trait loci in mice. *Genome Res.*, **15**, 681–691.
12. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
13. Goring, H.H., Curran, J.E., Johnson, M.P., Dyer, T.D., Charlesworth, J., Cole, S.A., Jowett, J.B., Abraham, L.J., Rainwater, D.L., Comuzzie, A.G. *et al.* (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.*, **39**, 1208–1216.
14. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
15. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C.C., Taylor, J., Burnett, E., Gut, I., Farrall, M. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.
16. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.-Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.
17. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
18. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.
19. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
20. Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.
21. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
22. Schadt, E.E. (2005) Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr. Opin. Biotechnol.*, **16**, 647–654.
23. Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, **19**, 1720–1730.
24. 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 Genet.*, **4**, e1000072.
25. Garge, N., Pan, H., Rowland, M.D., Cargile, B.J., Zhang, X., Cooley, P.C., Page, G.P. and Bunger, M.K. (2010) Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol. Cell. Proteomics*, **9**, 1383–1399.
26. Turk, B., Stoka, V., Turk, V., Johansson, G., Cazzulo, J.J. and Bjork, I. (1996) High-molecular-weight kininogen binds two molecules of cysteine proteinases with different rate constants. *FEBS Lett.*, **391**, 109–112.
27. Asano, K., Matsushita, T., Umeno, J., Hosono, N., Takahashi, A., Kawaguchi, T., Matsumoto, T., Matsui, T., Kakuta, Y., Kinouchi, Y. *et al.* (2009) A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat. Genet.*, **41**, 1325–1329.
28. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A. *et al.* (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.*, **43**, 246–252.
29. Ferreira, M.A.R., Matheson, M.C., Duffy, D.L., Marks, G.B., Hui, J., Le Souef, P., Danoy, P., Baltic, S., Nyholt, D.R., Jenkins, M. *et al.* (2011) Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet*, **378**, 1006–1014.
30. 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.
31. 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. *Nat. Genet.*, **41**, 342–347.
32. Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V. *et al.* (2011) Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.*, **43**, 429–435.
33. Naj, A.C., Jun, G., Beecham, G.W., Wang, L.-S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K. *et al.* (2011) Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.*, **43**, 436–441.

34. Marzi, C., Albrecht, E., Hysi, P.G., Lagou, V., Waldenberger, M., Tonjes, A., Prokopenko, I., Heim, K., Blackburn, H., Ried, J.S. *et al.* (2010) Genome-wide association study identifies two novel regions at 11p15.5-p13 and 1p31 with major impact on acute-phase serum amyloid A. *PLoS Genet.*, **6**, e1001213.
35. Danik, J.S., Pare, G., Chasman, D.I., Zee, R.Y., Kwiatkowski, D.J., Parker, A., Miletich, J.P. and Ridker, P.M. (2009) Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. *Circ. Cardiovasc. Genet.*, **2**, 134–141.
36. Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J. *et al.* (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.*, **40**, 616–622.
37. Croteau-Chonka, D.C., Marvelle, A.F., Lange, E.M., Lee, N.R., Adair, L.S., Lange, L.A. and Mohlke, K.L. (2011) Genome-wide association study of anthropometric traits and evidence of interactions with age and study year in Filipino women. *Obesity (Silver Spring)*, **19**, 1019–1027.
38. Yang, Q., Kathiresan, S., Lin, J.-P., Tofler, G.H. and O'Donnell, C.J. (2007) Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *BMC Med. Genet.*, **8**(Suppl. 1), S12.
39. Gharavi, A.G., Kiryluk, K., Choi, M., Li, Y., Hou, P., Xie, J., Sanna-Cherchi, S., Men, C.J., Julian, B.A., Wyatt, R.J. *et al.* (2011) Genome-wide association study identifies susceptibility loci for IgA nephropathy. *Nat. Genet.*, **43**, 321–327.
40. Neale, B.M., Fagerness, J., Reynolds, R., Sobrin, L., Parker, M., Raychaudhuri, S., Tan, P.L., Oh, E.C., Merriam, J.E., Souied, E. *et al.* (2010) Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). *Proc. Natl Acad. Sci. USA*, **107**, 7395–7400.
41. Hoglinger, G.U., Melhem, N.M., Dickson, D.W., Sleiman, P.M., Wang, L.S., Klei, L., Rademakers, R., de Silva, R., Litvan, I., Riley, D.E. *et al.* (2011) Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.*, **43**, 699–705.
42. Spencer, C.C., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., Band, G., Barker, R.A., Bellenguez, C., Bhatia, K., Blackburn, H. *et al.* (2011) Dissection of the genetics of Parkinson's disease identifies an additional association 5′ of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.*, **20**, 345–353.
43. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.
44. Keeney, TR., Bock, C., Gold, L., Kraemer, S., Lollo, B., Nikrad, M., Stanton, M., Stewart, A., Vaught, JD. and Walker, JJ. (2009) Automation of the somalogic proteomics assay: a platform for biomarker discovery. *J. Assoc. Lab. Automat.*, **14**, 360–366.
45. Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E.N., Carter, J., Dalby, A.B., Eaton, B.E., Fitzwater, T. *et al.* (2010) Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE*, **5**, e15004.
46. Crocker, P.R., Paulson, J.C. and Varki, A. (2007) Siglecs and their roles in the immune system. *Nat. Rev. Immunol.*, **7**, 255–266.
47. Carlin, A.F., Uchiyama, S., Chang, Y.C., Lewis, A.L., Nizet, V. and Varki, A. (2009) Molecular mimicry of host sialylated glycans allows a bacterial pathogen to engage neutrophil Siglec-9 and dampen the innate immune response. *Blood*, **113**, 3333–3336.
48. 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. *Am. J. Hum. Genet.*, **81**, 559–575.
49. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
50. Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N. (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.
51. Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.
52. Li, Y., Willer, C.J., Sanna, S. and Abecasis, G.R. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.
53. Gentleman, R. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.
54. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.