Human Molecular Genetics, 2011, Vol. 20, No. 18 **3710–3717**
doi:10.1093/hmg/ddr272
Advance Access published on June 10, 2011

Genome-wide association study identifies two loci strongly affecting transferrin glycosylation

Zoltán Kutalik[1,2,*,†], Beben Benyamin[3,†], Sven Bergmann[1,2], Vincent Mooser[4], Gérard Waeber[5], Grant W. Montgomery[6], Nicholas G. Martin[7], Pamela A.F. Madden[8], Andrew C. Heath[8], Jacques S. Beckmann[1,9], Peter Vollenweider[5], Pedro Marques-Vidal[10,‡] and John B. Whitfield[7,‡]

[1]Department of Medical Genetics, University of Lausanne, Lausanne, 1005, Switzerland, [2]Swiss Institute of Bioinformatics, Lausanne, 1005, Switzerland, [3]Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, Brisbane 4006, Australia, [4]GlaxoSmithKline, King of Prussia, PA 19406, USA, [5]Department of Medicine, Internal Medicine, CHUV, Lausanne 1011, Switzerland, [6]Molecular Epidemiology Laboratory and [7]Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Brisbane 4006, Australia, [8]Department of Psychiatry, Washington University, St Louis 63110, USA, [9]Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois (CHUV) University Hospital, 1011 Lausanne, Switzerland, and [10]Institute of Social and Preventive Medicine (IUMSP), University Hospital Center (CHUV) and University of Lausanne, Lausanne 1005, Switzerland

Received April 27, 2011; Revised and Accepted June 7, 2011

Polysaccharide sidechains attached to proteins play important roles in cell–cell and receptor–ligand interactions. Variation in the carbohydrate component has been extensively studied for the iron transport protein transferrin, because serum levels of the transferrin isoforms asialotransferrin + disialotransferrin (carbohydrate-deficient transferrin, CDT) are used as biomarkers of excessive alcohol intake. We conducted a genome-wide association study to assess whether genetic factors affect CDT concentration in serum. CDT was measured in three population-based studies: one in Switzerland (CoLaus study, $n = 5181$) and two in Australia ($n = 1509$, $n = 775$). The first cohort was used as the discovery panel and the latter ones served as replication. Genome-wide single-nucleotide polymorphism (SNP) typing data were used to identify loci with significant associations with CDT as a percentage of total transferrin (CDT%). The top three SNPs in the discovery panel (rs2749097 near *PGM1* on chromosome 1, and missense polymorphisms rs1049296, rs1799899 in *TF* on chromosome 3) were successfully replicated , yielding genome-wide significant combined association with CDT% ($P = 1.9 \times 10^{-9}$, 4×10^{-39}, 5.5×10^{-43}, respectively) and explain 5.8% of the variation in CDT%. These allelic effects are postulated to be caused by variation in availability of glucose-1-phosphate as a precursor of the glycan (*PGM1*), and variation in transferrin (*TF*) structure.

INTRODUCTION

Many proteins undergo N- or O-glycosylation during or immediately after peptide synthesis. Glycosylation of proteins modifies their physicochemical properties and plays an important role in receptor–ligand and cell–cell interactions (1). It has been estimated that half of all proteins are glycoproteins (2). Some proteins show variation in their glycosylation; the glycosylation sites may or may not be occupied (3), the carbohydrate structure can vary, and this variation can affect function. Conditions leading to variation in protein glycosylation include congenital disorders of glycosylation (4), infection, inflammation and cancers. In principle, variation in protein glycosylation could reflect genetic or acquired variation in (i) the enzymes which carry out the multiple steps of attachment, extension and modification of the glycan, (ii) the

*To whom correspondence should be addressed: Tel: +41 21 692 54 63; Email: zoltan.kutalik@unil.ch
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

Table 1. Studies and participants. All studies were population-based, but in the second Australian study subjects reporting high alcohol intake were prioritised for CDT measurement. In the table *R* stands for Pearson correlation, *P* for *P* value, *N* for sample size, and *SD* for standard deviation

CDT method	CoLaus Capillary electrophoresis	Australia, 1993–96 Mini-column (CDTect)	Australia, 2004–06 Immunoassay (N-Latex CDT)
n (male, female) with CDT in units/l	690 (338, 352)	1400 (539, 861)	2087 (1000, 1087)
n (male, female) with CDT in percent	5909 (2815, 3094)	1353 (514, 839)	2087 (1000, 1087)
n (male, female) with CDT% and genome-wide SNP data	5181 (2454, 2727)	745 (169, 576)	1509 (764, 745)
Drinks per week (Mean, SD, Range)	7.0 ± 9.1 (0–80)	6.8 ± 9.0 (0–70)	24.4 ± 23.1 (0–315)
CDT, percent (Mean, SD)	0.947 ± 0.783	0.614 ± 0.303	1.766 ± 0.649
CDT, units (Mean, SD)	31.2 ± 25.2	17.4 ± 8.9	48.9 ± 17.2
Total transferrin, g/l (Mean, SD)	3.80 ± 0.44	2.85 ± 0.52	2.80 ± 0.41
Correlation, drinks per week versus CDT percent (R, p, N)	0.407, 1.09×10^{-234}, 5401	0.241, 2.52×10^{-19}, 1353	0.373, 1.07×10^{-67}, 2020
Correlation, drinks per week versus CDT units (R, p, N)	0.367, 3.71×10^{-25}, 690	0.200, 4.15×10^{-14}, 1400	0.352, 8.07×10^{-60}, 2020

target or substrate protein which is to be glycosylated, (iii) the 'quality control' mechanisms which check and if necessary recycle proteins, or (iv) the turnover or repair of glycosylated proteins after their initial synthesis.

Variation in the glycosylation of transferrin has been extensively studied because of the role of plasma transferrin in iron transport, and because the concentration or proportion of less-glycosylated transferrin isoforms (asialotransferrin and disialotransferrin, collectively known as carbohydrate-deficient transferrin, CDT) can serve as a biomarker of excessive alcohol consumption. Circulating transferrin can vary between the asialo- and octasialo- isoforms, with tetrasialotransferrin the most abundant (5).

Up to 40% of genetic variation in total transferrin levels is due to variants in the transferrin *(TF)* and hemochromatosis (*HFE*) genes (6), but whether transferrin glycosylation is also gene dependent has never been assessed. In this study, we first tested whether *TF* gene polymorphisms have significant effects on CDT concentration or CDT as a percentage of total transferrin (CDT%). We then conducted a genome-wide search for polymorphisms affecting CDT, tested replication of significant or suggestive findings with independent cohorts and meta-analysed the combined dataset. In this work, we used data from three studies measuring CDT in different ways, which allows assessment of the robustness of allelic associations across methods.

Identification of significant allelic associations for transferrin glycosylation has the potential to improve clinical interpretation of CDT% as a diagnostic aid in the management of alcohol use disorders, through estimation of genotype-specific reference ranges. This would also provide proof of principle for potential allelic effects on other diagnostic glycoprotein measurements, such as tumour markers; and it has theoretical implications for genetic variation in the control of post-translational modification of proteins.

RESULTS

Relevant subject characteristics, for both the Swiss (CoLaus) and Australian participants, are summarized in Table 1.

Correlations between CDT results for two occasions 10 years apart were estimated from data for 66 Australian participants with both CDTect and N-Latex results. The repeatability of this measure was 0.59 for CDT concentration and 0.63 for CDT%, and these correlations were only slightly reduced by adjustment for sex and reported alcohol consumption. Heritability estimates, derived from sibling correlations in the second Australian (N-Latex) study, were 0.59 for CDT concentration and 0.60 for CDT%.

Genome-wide association results for CDT percentage

In our genome-wide association study (GWAS), two regions showed significant associations with CDT% (Supplementary Material, Fig. S1). These were on chromosomes 1 and 3, within or close to the *PGM1* and *TF* genes, respectively. Both loci fulfilled the selection criteria listed in Materials and Methods, with lead single-nucleotide polymorphisms (SNPs) rs2749097 ($P = 3.15 \times 10^{-6}$) and rs1534166 ($P = 1 \times 10^{-38}$) in the discovery set. In the latter region, both the association and the linkage disequilibrium (LD) were so strong that we performed step-wise model selection in order not to miss imperfect tagging or allelic heterogeneity in the region. This analysis led to the most plausible association model, which included three non-synonymous SNPs: rs1799899 (*TF* Gly277Ser, $P = 3.35 \times 10^{-35}$), rs1049296 (*TF* Pro589Ser, $P = 2.13 \times 10^{-32}$), rs8177318 (*TF* Ser55Arg, $P = 3.58 \times 10^{-4}$). This result indicates that the original strongest association with rs1534166 was tagging multiple missense variants in the transferrin gene. Therefore, we selected rs2749097 (*PGM1*, chromosome 1) and the three non-synonymous SNPs in *TF* for *in silico* replication. The hit on chromosome 1 (rs2749097) was confirmed in both replication cohorts ($P = 0.0068$ and $P = 0.0018$) in a direction-consistent manner, yielding a combined *P*-value of 1.9×10^{-9} (Fig. 1A).

Two of the three non-synonymous markers (rs1799899 and rs1049296) successfully replicated in the N-Latex-based Australian cohort ($P = 9.7 \times 10^{-6}$, $P = 7.1 \times 10^{-13}$, respectively), giving rise to very significant meta *P*-values ($P = 4 \times 10^{-39}$, $P = 5.5 \times 10^{-43}$). The third SNP (rs8177318) showed a direction-consistent replication signal, but—consistent with the weaker initial association and the smaller replication sample size—could not reach nominally significant replication *P*-value.

The association results for the significant chromosome 3 region are shown in Figure 1B, where initial GWAS-significant associations were found across a range of genes and intergenic regions between *BFSP2* and *RAB6B*, with the most significant SNP (rs1534166, $P = 4.3 \times 10^{-46}$) being in

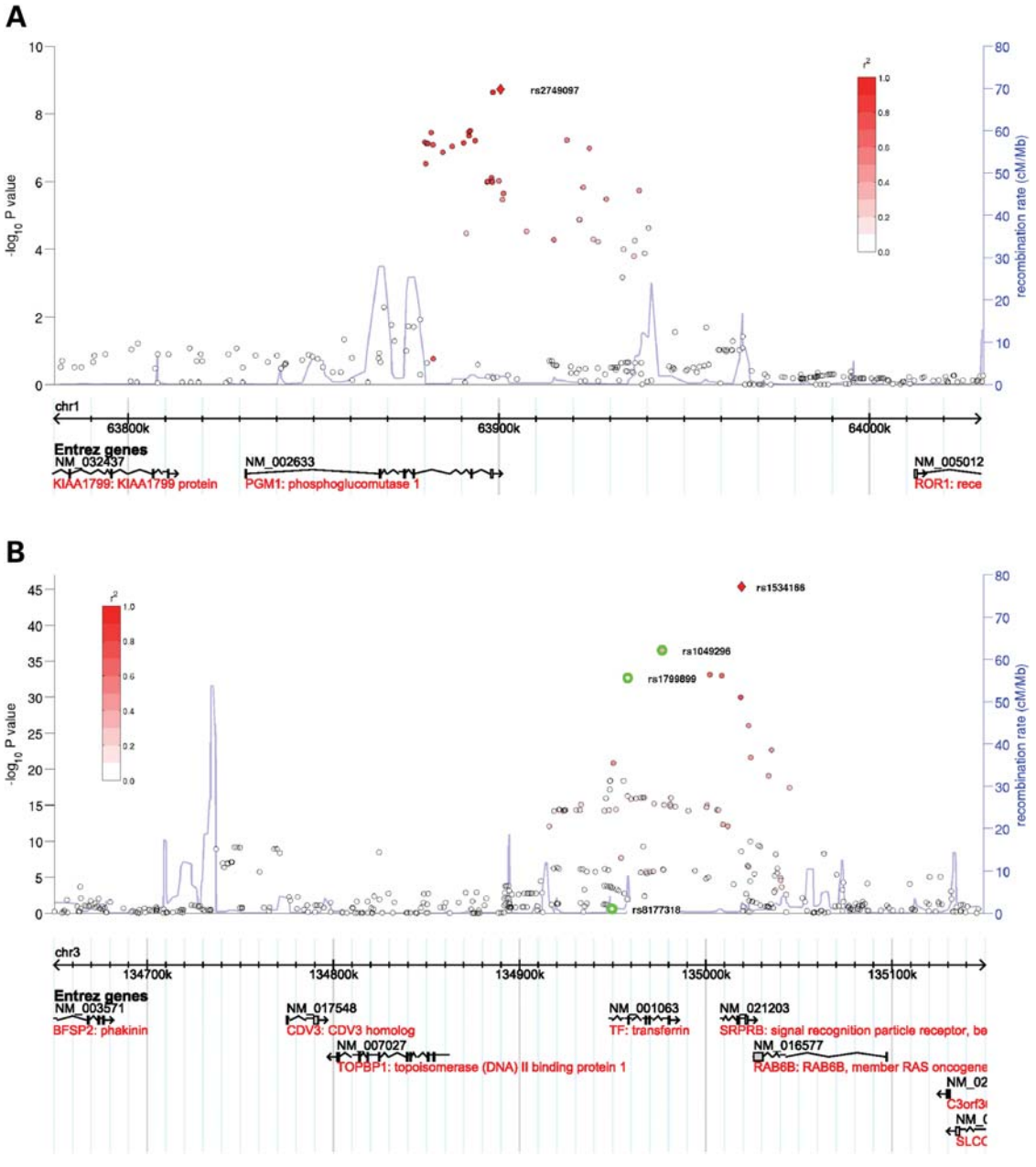

Figure 1. Local plots for the two regions associated with CDT%. The *y*-axis shows significance of the association as $-\log_{10}$(P), *x*-axis is physical distance (on Build 36). Shading of the data points indicates the strength of LD with the most significant SNP in the region. (**A**) SNP associations with CDT (as a percentage of total transferrin) near the *PGM1* locus on chromosome 1. The plotted combined *P*-values are from a meta-analysis of the discovery (CoLaus) and both (Australian) replication cohorts. (**B**) Association plot as described above, but for the region near the *TF* and *SRPRB* loci on chromosome 3. The plot shows combined *P*-values from a meta-analysis of the discovery (CoLaus) and the N-Latex-based (Australian) replication cohorts, showing univariate association results for all SNPs (red or uncoloured symbols). The three TF SNPs obtained by the multivariate model selection procedure (rs1049296, rs1799899 and rs8177318) are marked with green boundary circle.

SRPRB. In addition, the association results from the conditional analysis (on rs1799899 and rs1049296) clarified that the associations previously seen in the *BFSP2-CDV3-TOPBP1* and *RAB6B* regions could be ascribed to LD with the *TF* SNPs. Results are summarized in Table 2 and shown in more detail in Supplementary Material, Table S1.

Since CDT% is correlated with transferrin levels, we investigated whether SNPs associated with CDT% also show association when CDT% is conditioned on transferrin levels. Although associations became slightly less significant, effect sizes remained unchanged (*t*-test *P*-value > 0.75 for all SNPs). Thus, we consider that these genetic associations hold independently of the total transferrin level.

We also checked association *P*-values for these two *TF* SNPs (rs1534166 and rs1799899) in the other Australian replication cohort, but the results from the CDTect method showed an

Table 2. Summary of the most significant allelic associations with CDT and total transferrin

SNP	Chr	bp	Closest gene	Alleles Effect	Other	MAF CH	AU	Discovery Beta CH CE	P	Replication I Beta AU Latex	P	Replication II Beta AU CDTect	P	Combined Beta	SE	P	P_{het}	Comments
CDT percent																		
rs2749097	1	63 839 489	*PGM1*	G	C	0.187	0.184	−0.105	3.15E−06	−0.134	0.0068	−0.242	0.0018	−0.119	0.020	1.89E−09	2.24E−01	(a)
rs8177318	3	134 950 075	*TF*	A	T	0.030	0.025	−0.200	3.58E−04	−0.141	0.290			−0.187	0.051	2.10E−04	6.92E−01	(b, c)
rs1799899	3	134 958 510	*TF*	A	G	0.058	0.054	−0.480	3.35E−35	−0.364	9.70E−06			−0.461	0.035	3.96E−39	1.88E−01	(b, c)
rs1049296	3	134 977 052	*TF*	T	C	0.170	0.169	−0.270	2.13E−32	−0.365	7.10E−13			−0.288	0.021	5.45E−43	9.82E−02	(b, c)
CDT concentration																		
rs1799899	3	134 958 510	*TF*	A	G	0.058	0.054	−0.551	5.66E−07	−0.319	9.90E−05			−0.402	0.066	1.01E−09	9.12E−02	(b, d)
rs1534166	3	135 019 765	*SRPRB*	A	G	0.302	0.284	−0.282	3.03E−07	−0.293	5.80E−12			−0.289	0.034	1.56E−17	8.77E−01	(b)
Total transferrin																		
rs6439434	3	134 933 069	*TF*	T	G	0.491	0.473	0.199	9.13E−04	0.199	1.63E−08	0.133	9.98E−03	0.182	0.026	3.05E−12	5.47E−01	(a, f)
rs3811647	3	134 966 727	*TF*	A	G	0.318	0.338	0.362	1.27E−10	0.358	5.25E−19	0.353	8.80E−09	0.358	0.029	1.45E−35	9.94E−01	(a)
rs1800562	6	26 201 120	*HFE*	A	G	0.045	0.078	−0.466	0.00078	−0.647	8.67E−22	−0.687	9.46E−10	−0.629	0.053	2.13E−32	4.21E−01	(a, e)

(a) Meta-analysis based on all three methods.
(b) Meta-analysis based on CE and Latex methods only.
(c) Multivariate analysis of all three markers in TF exons (rs8177318, rs1799899, and rs1049296).
(d) Conditioned on the most significant SNP in the initial analysis (rs1534166).
(e) Conditioned on the most significant SNP in the initial analysis (rs3811647).
(f) Conditioned on the two most significant SNPs in the initial analysis (rs3811647 and rs1800562).
CH, CoLaus; AU, Australia. P_{het} values result from test of heterogeneity of allelic effects between the two or three datasets (CE, capillary electrophoresis, N-Latex and CDTect). A fuller list of SNP associations is given in Supplementary Material, Table S1.

opposite direction of the allelic effects compared with the other assessment methods ($P_{het} = 2.64 \times 10^{-9}$). This arises from the analytical principle of the CDTect method, and because changes in the amino acid sequence of transferrin can alter the isoelectric point (pI) of the isoforms. A more detailed explanation can be found in Discussion. For this reason, we excluded the CDTect replication sample and used only the N-Latex-based replication data for the chromosome 3 region.

Genome-wide association results for CDT concentration

For CDT concentration, the only significant SNPs were in the *TF-SRPRB* region of chromosome 3 (see Table 2). As for CDT%, the most significant SNP was rs1534166 in *SRPRB* ($P = 1.6 \times 10^{-17}$). Further, rs1799899 in *TF* was independently significant ($P = 1.0 \times 10^{-9}$) when the data were reanalysed after adjustment for the effects of rs1534166. These *P*-values were less significant than the equivalent ones for CDT%, because the number of Swiss participants who had total transferrin assessed (from which CDT concentration could be calculated) was only 690, but the effect sizes (betas in Table 2) were similar for CDT% and CDT concentration. Total CDT concentration was more weakly associated with the CDT%-associated *PGM1* polymorphism (rs2749097) yielding $P = 3.99 \times 10^{-4}$.

Genome-wide association results for total transferrin

As previously reported, SNPs in *TF* and *HFE* showed significant effects on total transferrin concentration (see Table 2). However, the *HFE* locus had no significant effects on either CDT% or CDT concentration. Even though the *TF* SNPs most strongly affecting total transferrin concentration showed only low-to-moderate LD with rs1799899 ($r^2 = 0.002$ for rs6439434 and 0.026 for rs3811647), the combined association *P*-value for this SNP was still highly significant ($P = 5.1 \times 10^{-12}$).

Genome-wide association results for total transferrin

The discovered variants explained 5.8% of the variance of CDT%, 4.9% for CDT levels and 11.2% for transferrin levels.

DISCUSSION

We have found that both the absolute and relative concentrations of 'carbohydrate-deficient' isoforms of transferrin are subject to genetic variation. Indeed, comparison of the heritability and repeatability estimates suggests that all or nearly all of the repeatable variation is genetic. As explained in Materials and Methods section, we tried to correct for all relevant environmental factors (first two principal components (PCs), sex, smoking status and alcohol consumption) to maximize our chances to tease out the most important genetic components determining CDT%. Other factors, such as pregnancy, haemodialysis, liver disease are either uncommon in the general population (in the CoLaus sample we had <11 pregnant women, 7 participants with liver disease and 9 participants with alcohol problems, <0.5% of the

total) or have negligible effect on CDT levels, thus could not have considerably influenced CDT levels.

In terms of specific SNPs and genes, there is evidence that variation near to the *PGM1* gene on chromosome 1, and in the *TF* gene on chromosome 3, contributes to the overall genetic effect. This leads us to consider how these genes fit into the processes of synthesis, release or circulation of the isoforms of transferrin.

The association with *PGM1* may be related to synthesis of the carbohydrate sidechains, which are attached to the protein and then modified by removal and addition of carbohydrate groups (7). The phosphoglucomutase enzyme converts glucose-6-phosphate into glucose-1-phosphate needed for synthesis of UDP-glucose, a precursor of the carbohydrate which is eventually transferred to the protein (8). However, phosphoglucomutase catalyses a very early step in this pathway and it is by no means clear how variation in the *PGM1* gene would affect the carbohydrate structure of the glycoprotein. Major defects in *PGM1* are known (http://www.ncbi.nlm.nih.gov/omim/171900), but they lead to a form of glycogen storage disease rather than a carbohydrate-deficient glycoprotein syndrome. However, *PGM1* expression in human liver (9) is significantly affected by rs4643 at 63 898 026 bp (expression $P = 1.0 \times 10^{-6}$), and this SNP is one of a group of three affecting CDT% at $P < 10^{-8}$. It is reasonable to accept that the SNPs on chromosome 1 showing significant association with CDT affect expression of *PGM1* in the liver (the site of transferrin synthesis), but an understanding of how phosphoglucomutase activity affects transferrin isoforms will require experiments which are beyond the scope of this study.

The associations on chromosome 3 are more readily related to relevant genes, but the details are complex. Although rs1534166 in *SRPRB* showed the strongest association ($P = 4.3 \times 10^{-46}$) with CDT% in the initial combined analysis, more detailed examination of this region showed two independent effects at neighbouring loci in different exons of *TF*, rs1799899 (Gly277Ser, $P = 3.96 \times 10^{-39}$), and rs1049296 (Pro589Ser, $P = 5.45 \times 10^{-43}$). These associations were identified using step-wise model selection; hence, each variant is associated when conditioned on the other one. The SNP rs1799899 had a substantially greater allelic effect than rs1049296, but was less common [minor allele frequency (MAF) = 5.8%]. A third non-synonymous SNP in *TF*, rs8177318 or Ser55Arg, may also contribute, but this could not be shown at the genome-wide significance level. Through this example we illustrate how computational methods can help in tracking down potentially causal variants, revealing allelic heterogeneity and substantially (from 3.6 to 5.8%) increasing the explained variance compared with univariate analysis.

The effects of these *TF* polymorphisms remain unchanged even in association with transferrin-corrected CDT%, so the significant effects on CDT% cannot be explained as being secondary to changes in total transferrin concentration. They may be due to changes in tertiary structure of transferrin because of the amino acid differences (Pro589Ser from rs1049296 and/or Gly277Ser from rs1799899), even though these changes are not at the N-glycosylation sites (which are asparagine residues at positions 413 and 611).

The heterogeneity of allelic effects across methods at the *TF* locus can be ascribed to these non-synonymous coding SNPs and the difference in analytical principles between the methods for measurement of CDT. The CDTect method relied on elution of isoforms with a pI >5.7 from an anion-exchange column followed by measurement of transferrin in the eluate. Any changes in the pI of transferrin isoforms resulting from changes in the amino acid sequence will either increase or decrease the amount of transferrin eluting from the column. The effects of the Pro589Ser (or C1/C2) polymorphism in the transferrin protein on the pI of transferrin and estimated CDT concentrations were examined by Stibler *et al.* (10). Although the effect was not significant with their limited number of samples, they found mean CDT concentrations of 51, 53 and 58 mg/l for C1, C1/C2 and C2 samples, respectively. The effects of this variant on the pI of disialotransferrin were illustrated by Helander *et al.* (11). They showed a shift in disialotransferrin towards the pI of the more usual forms of mono- and asialotransferrin in participants with the C2 variant. Because the CDTect assay measured the sum of asialo-monosialo- and part of disialotransferrin, the difference in pI leads to a greater proportion of the disialotransferrin being measured as CDT and an increase in the apparent CDT concentration. However, both the capillary electrophoresis and the N-Latex methods showed the effects of rs1049296 and rs1799899 on CDT which cannot be explained by alteration in isoelectric point. The N-Latex method is based on a monoclonal antibody, which specifically recognizes transferrin glycoforms that lack one or both of the complete N-glycans (disialo-, monosialo- and asialotransferrins) (12) and does not rely on a charge-based separation of the isoforms.

Apart from questions of how these polymorphisms affect serum CDT% or concentration, there may be practical implications for the diagnostic use of CDT measurement. CDT% is confirmed to be a robust marker of various aspects of alcohol consumption in our cohorts: the consumption of beer, wine and sprits strongly influence CDT% ($P < 10^{-58}$, $P < 10^{-57}$, $P < 10^{-6}$, respectively), so does alcohol consumption frequency ($P < 10^{-67}$) and total weekly alcohol consumption ($P < 10^{-134}$). As for any clinical test, it is necessary to establish a reference range against which patients' results can be compared. Problems with CDT as a marker for alcohol consumption have been reported, thus a definition of genotype-dependent reference intervals may improve the sensitivity or specificity of this test; this aspect will be examined in a further paper.

In summary, we have identified two loci (and three independent SNPs) affecting the degree of glycosylation of circulating transferrin through a genome-wide association approach. They account for 5.8% of phenotypic variation in CDT% or CDT, which is a high proportion compared with most GWAS outcomes. These loci require further studies to investigate the detailed mechanisms. The types of variation, affecting the process of oligosaccharide synthesis and showing the importance of protein structure, which we illustrate for transferrin, may well be relevant to other biologically important glycoproteins.

MATERIALS AND METHODS

Subjects and methods: CoLaus

The design of the CoLaus study has been described previously (13). Briefly, it is a population-based study conducted between

2003 and 2006 which recruited over 6000 subjects aged 35–75 years in Lausanne, Switzerland. The following inclusion criteria were applied: (a) voluntary participation in the examination, including blood sample, (b) aged 35–75 years and (c) Caucasian origin defined as having both parents and grand-parents Caucasian (determined by birth place). The Institutional Review Board of the Centre Hospitalier Universitaire Vaudois (CHUV) in Lausanne and the Cantonal Ethics Committee approved the study protocol and signed informed consent was obtained from participants. Participants were asked to attend the outpatient clinic at the CHUV, Lausanne, in the morning after an overnight fast. Data were collected by trained field interviewers during a single visit lasting ~60 min. Overall participation rate was 41%.

Most biological assays were performed by the CHUV Clinical Laboratory on fresh blood samples within 2 h of blood collection, and additional aliquots were stored at −80°C. Transferrin was measured in 690 participants by immunoassay with maximum inter- and intra-batch CVs of 1.8 and 1.0%, respectively. CDT measurement was done in a single batch for 5401 participants by capillary electrophoresis using the Ceofix-CDT reagent kits for the quantification of CDT in human serum (Analis, Belgium, kits #10-004760). Separation was performed on a Beckman P/ACE 5510 System (Beckman Coulter Instruments, Switzerland) equipped with an ultraviolet detector set at 200 nm. Maximum intra-batch CV was between 3.8% (highest CDT levels) and 15.3% (lowest CDT levels).

Nuclear DNA was extracted from whole blood for whole-genome scan analysis using Nucleon® genomic DNA extraction kit (Tepnel life sciences, Manchester, UK) according to the manufacturer's recommendations. Genotyping was performed using the Affymetrix GeneChip® Human Mapping 500K array set, as recommended by the manufacturer. Genotypes were called using BRLMM (http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf) (14). Duplicate individuals, and first and second degree relatives, were identified by computing genomic identity-by-descent coefficients, using PLINK (15). The younger individual from each duplicate or relative pair was removed. SNPs with call rate <70% and individuals with call rate <90% were excluded from further analysis.

A set of unmeasured HapMap SNPs was imputed. For imputation, only autosomal SNPs that were present in HapMap release 21 were used; the dataset used for imputation was 5435 unrelated CoLaus individuals and 390 631 measured SNPs (with Hardy–Weinberg P-value $>10^{-7}$ and MAF >1%). Imputation was performed using IMPUTE (16) version 0.2.0, and CEU haplotypes from HapMap release 21, all downloaded from IMPUTE (http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html). For any given SNP, IMPUTE computed genotype probabilities using information from all other measured SNPs except the focal SNP. Therefore, after running IMPUTE, measured SNP genotypes were used to replace the imputed genotype probabilities. Expected allele dosages were computed and plugged into the subsequent association analysis. SNPs with imputation quality (r^2-hat) <0.5 or MAF <1% were excluded from the analysis.

To account for possible population stratification, we computed PCs using the SMARTPCA (http://genepath.med.harvard.edu/~reich/) implementation (17), with default options except that no outlier removal iterations were performed. The first two PCs, sex, smoking status and alcohol consumption were used in all adjusted analyses.

Subjects and methods: Australia

Australian subjects participated in one or both of two studies evaluating the relationship between serum CDT concentration and alcohol consumption in the general population. All study participants gave informed consent, and the protocols were approved by appropriate institutional ethics review committees.

The first of these (18) was a twin study designed to assess the heritability of alcohol consumption and alcohol dependence, and to evaluate biochemical effects of excessive alcohol intake. Telephone interviews were conducted with a total of 7764 participants and blood was collected from 3375 of them (1134 men and 2241 women) between 1993 and 1996. Secondly, a twin-family study conducted between 2004 and 2006 (19) focused on identification of genetic loci contributing to susceptibility to alcohol or nicotine dependence, and to liver damage and biochemical abnormalities among excessive drinkers. Blood was collected from 9031 participants (3998 men and 5033 women) aged 18–92 years. In both studies, participants provided information about their alcohol use and symptoms associated with alcohol dependence as part of a telephone interview, and at the time of blood collection they completed a 7-day retrospective diary of alcohol use.

In the first study, CDT was measured on 1400 people aged 29–92 years (539 men and 861 women; the mean ages 44.8 and 46.5, respectively) using the Pharmacia CDTect method, in which CDT isoforms are eluted from an anion-exchange column and measured by immunoassay of transferrin. The CDTect method may be subject to bias due to polymorphisms in the *TF* gene that affect the pI of transferrin (for details, see Discussion), thus it can be unreliable for associations with *TF* SNPs. Therefore, we ignored association results for this cohort for the TF region. In the second study, CDT was measured in all available samples from participants who reported alcohol intake greater than recommended limits of 40 g/day for men and 20 g/day for women ($n = 1173$; 634 men and 539 women) and in 915 participants (367 men and 548 women) who reported lower or no alcohol intake. Serum CDT concentration was measured with a direct immunoassay (12) (N-Latex CDT method, Siemens Healthcare Diagnostics) on a Dade BN-II nephelometric analyser. CDT was also calculated as a percentage of total transferrin (CDT%). One result with very low total transferrin was excluded as a probable error or outlier.

Genome-wide SNP genotyping on DNA extracted from blood samples was performed using Illumina Human610-Quadv1 chips (~582 000 SNPs) or HumanCNV370-Quadv3 chips (~351 000 SNPs). SNPs were included in the analyses if they met the following conditions: Hardy–Weinberg equilibrium test $P \geq 10^{-6}$, MAF ≥ 1%, call-rate ≥ 0.95 and the mean value of GenCall score ≥0.7. Subjects found to be of non-European ancestry by PC analysis (EIGENSTRAT) (20) of the genotyping data were also excluded. The PC analysis used a set of 276 891 autosomal SNPs that were common to

Australian samples, HapMap 3 (11 global populations) and 5 Northern European (Denmark, Finland, the Netherlands, the UK and Sweden) populations from the GenomEUtwin Consortium. We excluded 277 individuals who were >6 SD from the mean of PCs 1 and 2 derived from the European populations.

Imputation of non-genotyped HapMap SNPs was performed. Because imputation is sensitive to both missingness and SNP density, to avoid introducing bias to the imputed data, a set of SNPs common to the genotyping chips ($n = 269\,840$) was used for imputation. Imputation was undertaken for 17 862 individuals participating in multiple twin and twin-family studies conducted at the Queensland Institute of Medical Research, after population outliers had been removed, as a single population. The imputation was run in two stages using Mach (21) (http://www.sph.umich.edu/csg/abecasis/MACH/index.html). Following imputation, data for 2.54 million Hapmap SNPs were available. The complete description of the initial quality controls of the GWAS data and imputation was given by Medland *et al*. (22). The best guess genotypes were used for the association analyses if they met the same conditions as mentioned for the genotyped SNPs as well as an imputation quality, r^2-hat of 0.3 or greater.

Association analyses taking account of family relationships were conducted using MERLIN (http://www.sph.umich.edu/csg/abecasis/Merlin/). CDT measurements and genome-wide SNP data were both available for 2284 subjects (775 with CDT measured by CDTect and 1509 by N-Latex). Prior to association analyses, the phenotypes were log-transformed, normalized to have zero mean and unit variance and adjusted for the effects of sex, age, self-reported smoking status, alcohol intake (number of drinks in the seven days preceding blood collection) and PCs 1 and 2 from the PC analysis. Visualization and annotation of the GWAS results were performed in WGAViewer (http://people.genome.duke.edu/~dg48/WGAViewer/download.php) (23).

Replication and meta-analysis

SNPs with association *P*-value $<5 \times 10^{-6}$ (in the Swiss discovery cohort), imputation quality (r^2-hat) >0.5 and MAF > 1% were selected, and evaluated in the replication cohorts. The selection procedure excluded all variants that had a stronger association signal in their 0.2 cM neighbourhood.

To dissect multiple independent association signals in regions with tight LD and strong association signal (e.g. the *TF* and *HFE* genes), we applied a step-wise model selection. The forward–backward step-wise model selection procedure started from the empty model (including non-genetic covariates only). In each step of the model selection, we tried to add (forward) or remove (backward) a SNP to maximize the Bayesian Information Criterion (BIC) (24) in a greedy fashion until (local) maximum is reached. If, in the procedure, only forward steps occurred, we called it *conditional analysis*, as simply the most associated variant(s) was/were included as covariate(s) in a conditional regression model. For associations with CDT and transferrin concentrations only forward steps were needed, while for CDT percentage it was a truly forward–backward model selection.

The results from the discovery and replication cohorts were combined into a fixed-effects meta-analysis using inverse variance weighting. Tests for heterogeneity were assessed using Cochran's *Q* statistic. As explained above, association results for the CDTect cohort for the *TF* region were ignored, thus the heterogeneity *P* values for that locus were computed based on two cohorts only. The details of the technical explanation for this artefact can be found in Discussion. Using only two cohorts for this region is not a limitation: the discovery *P* values for this locus were $<2 \times 10^{-32}$, so showing successful replication in only one replication cohort (N-Latex) is sufficiently convincing.

ACKNOWLEDGEMENTS

For the Australian samples, sample processing and biobank management was carried out under the leadership of Anjali Henders; genotype QC and imputation were conducted by the QIMR GWAS Group including Scott Gordon, Brian McEvoy, Sarah Medland, Dale Nyholt and Naomi Wray; and CDT measurements were performed by Veronica Dy (for N-Latex data) and Linda Fletcher and Theresa Murphy (for CDTect data).

Conflict of Interest statement. V.M. is a full-time employee of GlaxoSmithKline.

FUNDING

The CoLaus study was supported by research grants from GlaxoSmithKline. This work was supported by research grants from GlaxoSmithKline; Faculty of Biology and Medicine of Lausanne, Switzerland; Swiss National Science Foundation (33CSCO-122661); Australian National Health and Medical Research Council; EU 5th and 7th Framework Programmes (GenomEUtwin Project QLG2-CT-2002-01254); ENGAGE Consortium (HEALTH-F4-2007-201413) and U.S. National Institutes of Health (AA07535, AA10248, AA11998, AA13320, AA13321, AA13326, AA14041, AA17688, DA12854, MH66206). Swiss National Science Foundation (33CSCO-122661 to S.B., 3100AO-116323/1 to S.B., 310000-112552 to J.S.B.); Giorgi-Cavaglieri Foundation to S.B.; Swiss Institute of Bioinformatics (Service Grant to S.B.); European Framework Project 6 (AnEuploidy and EuroDia projects to S.B.); National Health and Medical Research Council (NHMRC) Fellowships (552498, 339446 and 619667 to B.B. and G.W.M.).

REFERENCES

1. Varki, A. and Lowe, J.B. (2009) Biological roles of glycans. In Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W. and Etzler, M.E., (eds), *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
2. Apweiler, R., Hermjakob, H. and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.
3. Jones, J., Krag, S.S. and Betenbaugh, M.J. (2005) Controlling N-linked glycan site occupancy. *Biochim. Biophys. Acta*, **1726**, 121–137.

4. Marquardt, T. and Freeze, H. (2001) Congenital disorders of glycosylation: glycosylation defects in man and biological models for their study. *Biol. Chem.*, **382**, 161–177.
5. Martensson, O., Harlin, A., Brandt, R., Seppa, K. and Sillanaukee, P. (1997) Transferrin isoform distribution: gender and alcohol consumption. *Alcohol Clin. Exp. Res.*, **21**, 1710–1715.
6. Benyamin, B., McRae, A.F., Zhu, G., Gordon, S., Henders, A.K., Palotie, A., Peltonen, L., Martin, N.G., Montgomery, G.W., Whitfield, J.B. *et al.* (2009) Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. *Am. J. Hum. Genet.*, **84**, 60–65.
7. Stanley, P., Schachter, H. and Taniguchi, N. (2009) N-Glycans. In Varki, A., Cummings, R.D., Esko, J.D., Freeze, H., Stanley, P., Bertozzi, C.R., Hart, G.W. and Etzler, M.E., (eds), *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
8. Freeze, H. and Elbein, A.D. (2009) Glycosylation precursors. In Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W. and Etzler, M.E., (eds), *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
9. 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.
10. Stibler, H., Borg, S. and Beckman, G. (1988) Transferrin phenotype and level of carbohydrate-deficient transferrin in healthy individuals. *Alcohol Clin. Exp. Res.*, **12**, 450–453.
11. Helander, A., Eriksson, G., Stibler, H. and Jeppsson, J.O. (2001) Interference of transferrin isoform types with carbohydrate-deficient transferrin quantification in the identification of alcohol abuse. *Clin. Chem.*, **47**, 1225–1233.
12. Delanghe, J.R., Helander, A., Wielders, J.P., Pekelharing, J.M., Roth, H.J., Schellenberg, F., Born, C., Yagmur, E., Gentzer, W. and Althaus, H. (2007) Development and multicenter evaluation of the N latex CDT direct immunonephelometric assay for serum carbohydrate-deficient transferrin. *Clin. Chem.*, **53**, 1115–1121.
13. Firmann, M., Mayor, V., Marques-Vidal, P., Bochud, M., Pécoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K.S., Yuan, X. *et al.* (2008) The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc. Disord.*, **8**, 6.
14. Affymetrix (2006) BRLMM: an improved genotype calling method for the GeneChip© Human Mapping 500K array set, pp 1–18.
15. 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.
16. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
17. Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
18. Whitfield, J.B., Fletcher, L.M., Murphy, T.L., Powell, L.W., Halliday, J., Heath, A.C. and Martin, N.G. (1998) Smoking, obesity, and hypertension alter the dose-response curve and test sensitivity of carbohydrate-deficient transferrin as a marker of alcohol intake. *Clin. Chem.*, **44**, 2480–2489.
19. Whitfield, J.B., Dy, V., Madden, P.A., Heath, A.C., Martin, N.G. and Montgomery, G.W. (2008) Measuring carbohydrate-deficient transferrin by direct immunoassay: factors affecting diagnostic sensitivity for excessive alcohol intake. *Clin. Chem.*, **54**, 1158–1165.
20. 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.
21. Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.
22. Medland, S.E., Nyholt, D.R., Painter, J.N., McEvoy, B.P., McRae, A.F., Zhu, G., Gordon, S.D., Ferreira, M.A., Wright, M.J., Henders, A.K. *et al.* (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.*, **85**, 750–755.
23. Ge, D., Zhang, K., Need, A.C., Martin, O., Fellay, J., Urban, T.J., Telenti, A. and Goldstein, D.B. (2008) WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res.*, **18**, 640–643.
24. Balding, D.J., Bishop, M.J. and Cannings, C. (2007) *Handbook of Statistical Genetics*. John Wiley and Sons Ltd, Chichester.