Genome-wide association studies in Samoans give insight into the genetic architecture of fasting serum lipid levels

Jenna C. Carlson[1,2]*, Daniel E. Weeks[1,2¶], Nicola L. Hawley[3], Guangyun Sun[4], Hong Cheng[4], Take Naseri[5], Muagututi'a Sefuiva Reupena[6]; Ranjan Deka[4¶], Stephen T. McGarvey[7,8¶], Ryan L. Minster[1¶]

[1] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

[2] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

[3] Department of Epidemiology, School of Public Health, Yale University, New Haven, Connecticut, United States of America

[4] Department of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America

[5] Ministry of Health, Government of Samoa, Apia, Samoa

[6] Lutia i Puava ae Mapu i Fangalele, Apia, Samoa

[7] International Health Institute and Department of Epidemiology, School of Public Health, Brown University, Providence, Rhode Island, United States of America

[8] Department of Anthropology, Brown University, Providence, Rhode Island, United States of America

* Corresponding Author

Email: jnc35@pitt.edu

¶ DEW, RD, STM, RLM are Joint Senior Authors

Abstract

The current understanding of the genetic architecture of lipids has largely come from genome-wide association studies. To date, few studies have examined the genetic architecture of lipids in Polynesians, and none have in Samoans, whose unique population history, including many population bottlenecks, may provide insight into the biological foundations of variation in lipid levels. Here we performed a genome-wide association study of four fasting serum lipid levels: total cholesterol (TC), high-density lipoprotein (HDL), low-density lipoprotein (LDL), and triglycerides (TG) in a sample of 2,849 Samoans, with validation genotyping for associations in a replication cohort comprising 1,798 Samoans and American Samoans. We identified multiple genome-wide significant associations ($P < 5 \times 10^{-8}$) previously seen in other populations – *APOA1* with TG, *CETP* with HDL, and *APOE* with TC and LDL – and several suggestive associations ($P < 1 \times 10^{-5}$), including an association of variants downstream of *MGAT1* and *RAB21* with HDL. However, we observed different association signals for variants near *APOE* than what has been previously reported in non-Polynesian populations. The association with several known lipid loci combined with the newly-identified associations with variants near *MGAT1* and *RAB21* suggest that while some of the genetic architecture of lipids is shared between Samoans and other populations, part of the genetic architecture may be Polynesian-specific.

Key words: complex trait genetics; founder population genetics; lipids; genome-wide association study; Polynesia

Introduction

The Samoan Islands, comprising both the U.S. Territory of American Samoa (American Samoan) and the Independent State of Samoa (Samoa), have experienced a rise in prevalence of cardiovascular disease and other non-communicable diseases in the last 30 years partly due to economic modernization, rapid urbanization, and lifestyle changes such as increased caloric intake and sedentary behavior [1–3]. A 2010 population-based survey in Samoa, which gathered the "discovery cohort" studied further here, found that many Samoans are at elevated risk of cardiovascular disease based on known risk factors—increased total cholesterol (TC), low-density lipoprotein cholesterol (LDL), and triglycerides (TG), as well as decreased high-density lipoprotein (HDL) [4–6]—with 47% of the 2,938 adult Samoans examined in the study having elevated TC (≥ 5.2 mmol/L), 88% of men and 91% of women having elevated LDL (> 2.59 mmol/L), and 43% of women having low HDL (< 1.29 mmol/L) [1].

Our current understanding of the genetic component of serum lipid level variation has been largely due to genome-wide association studies (GWAS) [7]. The Global Lipids Genetics Consortium found strong evidence for 157 loci associated with one or more of these traits using a sample of 188,577 individuals of European, East Asian, South Asian, and African ancestry [8]. However, few GWAS of serum lipid levels have been conducted in Pacific Islanders [9, 10] and to our knowledge only one has included a small number of Polynesians [11]. Previous studies have estimated the heritability of serum lipid levels in Samoans, ranging from 16% for HDL to 23% for TG, and have identified genetic susceptibility loci via linkage analysis [12], warranting further study of

the genetic architecture of serum lipid levels in Samoans. Samoans are a genetically-isolated founder population, with unique evolutionary history, making them particularly useful in genomic studies [13, 14]. Thus, genomic studies of serum lipid levels could reveal novel lipid-altering loci specific to Pacific Islander populations, as well as highlight susceptibility loci shared with global populations.

Here we report the results of a GWAS of fasting TC, LDL, HDL, and TG in up to 2,849 individuals from independent Samoa followed by replication in up to 1,798 individuals from independent Samoa and American Samoa, as part of ongoing genome-wide association studies of cardiometabolic disease and adiposity-related traits in the Samoan Islands [1]. We identified multiple genome-wide significant associations previously seen in other populations – *APOA1* with TG, *CETP* with HDL, and *APOE* with TC and LDL – and several suggestive associations, including an association between variants downstream of *MGAT1* and *RAB21* with HDL.

Methods

Discovery Cohort and Genotyping

The discovery cohort data are available from dbGaP (accession number: phs000914.v1.p1). The discovery cohort of 2,849 individuals is drawn from a population-based sample recruited from Samoa in 2010 (Table 1). The sample selection, data collection methods, and phenotyping, including the laboratory assays for serum lipid and lipoprotein levels, have been previously reported [1, 13]. Briefly, serum lipid levels were derived from fasting whole blood samples collected after a minimum

10-hour overnight fast. Genotyping was performed using Genome-Wide Human SNP 6.0 arrays (Affymetrix). Extensive quality control was conducted on the basis of a pipeline developed by Laurie et al [15]. Additional details for sample genotyping and genotype quality control are described in Minster et al [13]. This study was approved by the institutional review board of Brown University and the Health Research Committee of the Samoa Ministry of Health. All participants gave informed consent. Imputation was not performed in this study because prior experience with this population using extant imputation panels such as the Phase 3 1000 Genomes panel showed that the resulting imputed genotypes did not correlate well with observed genotypes [13].

Table 1. Demographic, anthropometric, and blood biochemistry statistics of the genotyped discovery and replication cohorts.

	2010 Discovery Cohort Samoa, n = 2849		2002–2003 Replication Sample Set Samoa, n = 490		1994–1995 Replication Sample Set Samoa, n = 468		2002–2003 Replication Sample Set American Samoa, n = 592		1994–1995 Replication Sample Set American Samoa, n = 248	
	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
	n = 1703	n = 1146	n = 245	n = 245	n = 243	n = 225	n = 337	n = 255	n = 137	n = 111
age (years)	44.8 (11.10)	45.6 (11.10)	44 (17.00)	40.7 (16.30)	43.3 (8.60)	43.2 (8.90)	43 (16.00)	43.2 (16.60)	43.3 (9.10)	45.5 (10.60)
BMI (kg/m^2)	34.8 (6.70)	31.3 (5.90)	33.2 (7.70)	28.8 (5.40)	31.9 (5.70)	28.8 (5.00)	36.5 (8.40)	33.4 (7.60)	37.4 (6.90)	34.9 (6.00)
total cholesterol (mg/dL)	199.3 (36.10)	200.3 (38.70)	202.3 (35.90)	195.5 (40.60)	199.1 (37.20)	197.4 (35.00)	187.1 (38.60)	189.6 (37.80)	197.8 (31.20)	194.1 (31.00)
HDL (mg/dL)	46.5 (10.80)	43.7 (11.20)	47.1 (10.30)	46.3 (11.20)	43.4 (11.00)	42.5 (11.40)	40.9 (8.80)	40 (8.80)	34.8 (7.60)	32.8 (8.30)
LDL (mg/dL)	130 (32.60)	129.6 (35.30)	130.3 (32.80)	128.6 (37.50)	140.2 (34.80)	133.5 (33.00)	118.5 (33.70)	118.9 (35.10)	136.6 (32.40)	136 (37.50)
trigylcerides (mg/dL)	114.9 (80.50)	139.3 (113.10)	110.9 (58.90)	120.1 (91.30)	85.6 (72.80)	102.8 (97.20)	130.8 (78.10)	200.2 (206.50)	118.9 (63.40)	168.2 (109.00)
fasting glucose (mg/dL)	107.6 (56.20)	107.4 (54.80)	102.2 (46.40)	98.6 (49.80)	90.9 (38.00)	85.3 (18.90)	104.2 (50.30)	108.8 (54.70)	107.3 (49.80)	125.2 (61.70)

Replication Cohort and Genotyping

The replication cohort of 1,798 individuals contains two sample sets recruited from Samoa and American Samoa (Table 1). Although there is substantial economic variation across the two polities, with American Samoans generally having a higher standard of living, the Samoans from both territories form a single socio-cultural unit with frequent exchange of mates; genetically, they represent a single homogenous population [3, 16]. The first sample set, referred to as the 1990–95 replication sample set, contains 716 unrelated individuals derived from a longitudinal study of adiposity and cardiovascular disease risk factors among adults from American Samoa and Samoa (Table 1). Detailed descriptions of the sampling and recruitment have been reported previously [17–19]. Briefly, participants were recruited from 46 villages and worksites in American Samoa in 1990 and 9 villages in Samoa (Western Samoa, at the time of recruitment) in 1991 and followed up four years later in 1994 and 1995, respectively. All participants were, at baseline, free of self-reported history of heart disease, hypertension, or diabetes. This study was approved by the institutional review board of the Miriam Hospital, Providence, RI. All participants gave informed consent. The second sample set, referred to as the 2002–03 replication sample set, contains 1,082 individuals from American Samoa and Samoa and was drawn from an extended family-based genetic linkage analysis of cardiometabolic traits (Table 1) [12, 20–23]. Probands and relatives were unselected for obesity or related phenotypes. The recruitment process and criteria used for inclusion in this study have been described in detail previously [21, 23]. This study was approved by the institutional review board of Brown University and research ethics review committees in both Samoa and American Samoa.

All participants gave informed consent. Imputation was not performed in these studies for the same rationale as the discovery cohort above, but also because genome-wide marker data was not available for the samples in these studies.

In both replication sample sets, blood samples were collected in the morning after a minimum of 10 hours fasting, from which serum lipid levels were derived using assay methods published previously [12, 17]. Genotyping of variants selected for validation in the replication cohort (described below) was performed using custom-designed TaqMan OpenArray Real-Time PCR assays (Applied Biosystems). SNPs that could not be genotyped using OpenArray assays were genotyped individually using TaqMan SNP Genotyping assays (Applied Biosystems). Eight variants could not be genotyped due to technical difficulties.

Statistical Analyses

Prior to association analyses, residuals were generated for all four lipid traits. First, traits were transformed to normality with the Box–Cox power transformation; secondly, model selection was performed using step-wise linear regression with initial model covariates previously associated with serum lipid levels: age, age^2, sex, log-transformed BMI, fasting glucose, smoking status, farming status (as a measure of physical activity), and interactions between age, age^2, and sex. The final TC model adjusted for age, age^2, sex, age × sex, and age^2 × sex; the final LDL and TG models adjusted for age, age^2, sex, and age^2 × sex; the final HDL model adjusted for age and sex.

Preliminary associations were performed, and variants were selected for validation without consideration of hypolipidemic medication use, as it was not measured. However, participants did self-report use of heart disease medication. Sensitivity analysis revealed that this self-reported use of medication to treat heart disease was significantly associated with TC and LDL (results not shown); individuals reporting such medication use (n = 17) were excluded from analyses. The prioritization of variants for validation genotyping was updated using these analyses, but only after available resources were fully expended. Unfortunately, not all variants that should have been prioritized for validation genotyping were successfully genotyped. All results presented are those of the corrected analyses, removing the individuals with heart disease medication use.

Additional sensitivity analysis was performed for TG by excluding one outlying observation (i.e., TG > 4 standard deviations above mean); results did not change qualitatively, and, since the recorded value was within the range of plausible values for TG, the individual was retained for presented analyses.

Association between lipid residuals and autosomal genotypes of 659,492 SNPs with minor allele frequency (MAF) ≥ 0.05 and Hardy-Weinberg Equilibrium (HWE) test *P* value ≥ 5×10^{-5} was assessed using linear mixed modelling in GenABEL, including previously-derived empirical kinship estimates to adjust for subject relatedness [13, 24]. The association between X-chromosome genotypes and the lipid phenotypes were

186 calculated in GenABEL, without adjustment using the empirical kinship estimates.
187 Genomic inflation due to population stratification and cryptic relatedness was assessed
188 by estimating λ_{GC} using the lower 90% of the *P* value distribution [25]. GWAS *P* values
189 in the discovery cohort (P_D) were compared to a threshold for genome-wide significance
190 of $P_D < 5 \times 10^{-8}$ and a suggestive association threshold of $P_D < 1 \times 10^{-5}$. Statistical
191 power to detect signals at these thresholds was calculated using the Genetic Power
192 Calculator [26].
193
194 Gene-set enrichment analysis with MAGENTA was also performed to identify any
195 biological pathways enriched for discovery association signals [27]. Briefly, gene scores
196 were obtained from the most significant *P* value among SNPs located within each gene
197 using the association results from each lipid GWAS. Genes scores were adjusted for
198 confounding factors including gene size, number of variants, and linkage disequilibrium-
199 related properties by using step-wise multiple linear regression. The 95th percentile of
200 all gene scores was used as the enrichment cutoff for each trait [28]. Gene-set
201 enrichment *P* values were obtained for highly ranked gene scores. Gene sets were
202 obtained from Gene Ontology (April 2010), pathway information from the Ingenuity
203 (June 2008) and KEGG (June 2010), and biological processes and molecular function
204 from PANTHER (January 2010).
205
206 For each of the lipid traits, the INRICH program [29] was used to test for
207 enrichment of known genes (as constructed from Teslovich et al. [30] and Willer et al
208 [8]). INRICH tests if more known genes are contained in associated intervals than

expected by chance, using permutation based on 1 million replicates to generate experiment-wide empirical *P* values. For each lipid trait, we defined the associated intervals as 100 kb intervals centered on the most significant SNP within association peaks with $P_D < 1 \times 10^{-4}$.

We selected 21 regions demonstrating at least suggestive association for association validation in the replication cohort. An additional 10 regions which should have selected for validation were not followed-up because their exclusion was based on preliminary analyses that included 17 participants taking heart disease medication—participants who were ultimately excluded from these studies. The variant from each locus with smallest *P* value across the four lipid scans (defined as 1 Mb windows surrounding the peak SNP) or a proxy SNP in high linkage disequilibrium with the lowest-*P* value SNP was selected as representative of the locus for replication genotyping.

Statistical association was measured in the 1990–95 and 2002–03 replication sample sets independently, and results were combined using meta-analysis (see below). Association analyses for both sample sets were performed using GenABEL [31] in R [32], using the same regression models as in the discovery cohort but additionally adjusting for polity (American Samoa or Samoa); the 2002-03 sample set was additionally adjusted using expected kinship, as derived from familial pedigree information [33].

Prior to meta-analysis, quality control was performed using EasyQC to check for strand and allele frequency consistency [34]. *P* value-based meta-analysis using sample sizes as weights was performed using METAL [35] to generate two *P* values: one for the meta-analysis of the two replication cohorts (P_R) and one for the replication cohorts and discovery sample together (P_{DR}). Resulting meta-analysis signals were evaluated based on genome-wide significance and suggestive thresholds (as described above) and by the contribution of the replication sample to the signal. Effect directions for meta-analysis results of peak SNPs were qualitatively compared to those of previously reported lead SNPs.

For ease of reference, any locus identified here with a corresponding signal within 1 mega base pairs (Mb) in a prior lipid study is referred to by the previously prescribed locus name [8, 10]; for loci not previously associated with lipid traits, the symbol of the gene nearest the peak SNP in the locus or the hyphen-separated symbols of the nearest two genes is used as the locus label.

Results

The demographic, anthropometric, and biochemical characteristics of the 2,849 participants composing the discovery cohort for this GWAS of serum lipids levels and the 1,798 participants composing the replication cohort are presented in Table 1. A detailed description of the discovery cohort and its trends compared to the historical sample sets making up the replication cohort has been previously reported [1]. Briefly, the average age was similar for all cohorts; average BMI was higher among women

compared to men and in American Samoa compared to Samoa; average BMI for men and women in Samoa was higher in more recent studies; average lipid levels are largely similar across cohorts with minor exceptions.

We assessed 659,492 unique genome-wide markers for association with 4 traits—TC, HDL, LDL and TG—in up to 2,849 Samoans in the discovery cohort. Relatedness within the discovery cohort was well-controlled using the empirical kinship coefficients; λ_{GC} ranged between 1.03 and 1.07 for the four lipid traits (Figs S2, S4, S6, and S8 in S1 Appendix). We observed 38 genome-wide suggestive or significant associations across 31 loci from the four GWAS (Fig 1; Figs S1, S3, S5, and S7 in S1 Appendix; Tables S1, S7, S10, and S13 in S1 Appendix).

Fig 1. Manhattan plots for GWAS of four lipid traits in the discovery cohort of 2,849 Samoans. The dashed and solid lines denote genome-wide suggestive and genome-wide significant *P* value thresholds ($P < 1 \times 10^{-5}$ and $P < 5 \times 10^{-8}$, respectively). Peaks are labeled with the candidate gene or closest gene in the region if they have at least suggestive association in the discovery cohort for at least one trait and demonstrate evidence of replication or have been previously associated.

Genome-wide significant association was observed in the discovery cohort between all four traits and markers near *APOE*: TC and rs4420638 ($P_D = 2.67 \times 10^{-16}$, Fig 2E); HDL and rs4420638 ($P_D = 9.07 \times 10^{-9}$, Fig 2F); LDL and rs1160985 ($P_D = 2.61 \times 10^{-20}$; Fig 2G); and TG and rs4420638 ($P_D = 7.44 \times 10^{-10}$, Fig 2H). Additionally, HDL

was associated with markers near *CETP* (rs289708, $P_D = 1.19 \times 10^{-11}$), and TG, with *APOA1* (rs6589566, $P_D = 3.98 \times 10^{-18}$, Fig 2C). Suggestive associations were observed between lipid levels and markers at an additional 28 loci, including the *MGAT1* and *RAB21* loci and HDL (Fig 2A,D), and *APOA1* with TC (rs3741298, $P_D = 1.63 \times 10^{-7}$, Fig 2B). We had 80% power at $\alpha = 1 \times 10^{-5}$ and $\alpha = 5 \times 10^{-8}$ to detect SNPs that account for 1.0% and 1.5%, respectively, of the residual variance in a phenotype.

Fig 2. Regional association plots for selected loci. Regional association plots generated in LocusZoom [36] showing –log10(*P* values) for SNPs in the (A) *MGAT1* locus and HDL, (B) *APOA1* locus and TC, (C) *APOA1* locus and TG, (D) *RAB21* locus and HDL, and the *APOE* locus and (E) TC, (F) HDL, (G) LDL, and (H) TG. Points are color coded within each plot according to pairwise linkage disequilibrium (r^2) with the labeled SNPs; the saturation of the color of each plotted SNP measures the linkage disequilibrium (r^2) with the labeled SNP sharing the same color.

Gene-set enrichment analysis with MAGENTA highlighted, at a < 5% false-discovery rate (FDR), several lipid homeostasis pathways and gene ontologies for HDL and TG (Tables S8 and S14 in S1 Appendix). Four gene sets were below the FDR for both HDL and TG: HDL particle remodeling, reverse cholesterol transport, cholesterol efflux, and phospholipid efflux. An additional 12 gene sets were implicated for HDL and three gene sets for TG. The HDL particle remodeling and reverse cholesterol transport gene sets had significant enrichment for TC (Table S3 in S1 Appendix), and a single gene set was implicated with LDL, the amylase pathway (Table S11 in S1 Appendix). All

four traits had significant enrichment for known TC, HDL, LDL, and TG loci using the INRICH method (Tables S5, S9, S12, and S15 in S1 Appendix).

Validation of peak SNPs was attempted for 21 loci. At loci with multiple associated variants, the most significant variant was chosen as representative of the locus. For some loci, the exclusion of participants using self-reported heart disease medication resulted in a different peak SNP. Thus, for the *APOE* locus rs1160985 was genotyped instead of rs4420638 ($P_D = 2.67 \times 10^{-16}$ for TC, $P_D = 9.07\times 10^{-9}$ for HDL, and $P_D = 7.44 \times 10^{-10}$ for TG); for the *APOA1* locus rs964184 was genotyped instead of rs3741298 ($P_D = 1.63 \times 10^{-7}$ for TC) or rs6589566 ($P_D = 3.98 \times 10^{-18}$ for TG); for the *MGAT1* locus rs1038143 was genotyped instead of rs249356 ($P_D = 1.06 \times 10^{-6}$ for HDL); for the *APOB* locus rs754523 was genotyped instead of rs1469513 ($P_D = 2.71 \times 10^{-6}$ for LDL).

We successfully genotyped the peak SNP, or a proxy SNP, in the replication cohorts for 15 loci. Two loci (*APOA1* with TG, $P_{DR} = 1.81 \times 10^{-29}$; *APOE* with TC, $P_{DR} = 4.29 \times 10^{-21}$, and LDL, $P_{DR} = 1.53 \times 10^{-27}$) demonstrated genome-wide significant associations in the discovery-replication meta-analysis (Table 2 and Tables S1, S7, S10, and S13 in S1 Appendix). An additional four associations demonstrated evidence of replication with consistent directions of effect and suggestive joint P_{DR} values (*GCKR* with TG, $P_{DR} = 5.62 \times 10^{-8}$; *MGAT1* with HDL, $P_{DR} = 2.91 \times 10^{-7}$; *APOA1* with TC, $P_{DR} = 1.72 \times 10^{-6}$; *RAB21* with HDL, $P_{DR} = 5.92 \times 10^{-7}$). Three associations had suggestive joint P_{DR} values driven by the discovery associations only (*APOB* with LDL, $P_{DR} = 5.81 \times$

10^{-6}; *LIPC* with HDL, $P_{DR} = 9.15 \times 10^{-7}$; *CDH4* with HDL, $P_{DR} = 8.77 \times 10^{-6}$); associations at *APOB* and *CDH4* had consistent directions of effect. Among the remaining loci with at least suggestive association in the discovery sample, but not in the discovery-replication meta-analysis, consistent effect directions were also seen for TC and *APOB* and *ZHX2*; LDL and *ALG10* and *CPNE8* (Table 2 and Tables S1, S7, S10, and S13 in S1 Appendix).

330 **Table 2. Suggestive loci and replication genotyping**

331

Total Cholesterol																	
Locus	SNP	Chr	BP	EA	OA	Dir	P_D	P_R	P_{DR}	Known gene	Traits	SAM	EAS	SAS	EUR	AMR	AFR
APOB	rs754523	2	21311691	G	A	+++	*6.25E-06*	0.178	1.20E-05	*APOB*	C,H,L,T	0.247	0.265	0.140	0.309	0.306	0.201
PDE4D	rs7711093	5	59593138	G	A	++−	*3.01E-06*	0.747	5.37E-04			0.506	0.643	0.553	0.856	0.775	0.852
LUCAT1	rs10072084	5	90539203	C	T	+··	*9.48E-06*					0.541	0.559	0.422	0.232	0.429	0.822
*FILIP1***	rs2951921	6	76165524	T	C	+··	*9.04E-07*					0.073	0.022	0.063	0.015	0.030	0.293
ZHX2	rs7841763	8	123971081	T	C	+++	*4.82E-06*	0.631	1.03E-04			0.043	0.023	0.146	0.102	0.058	0.169
APOA1	rs964184*	11	116648917	C	G	+++	5.37E-05	0.009	*1.72E-06*	*APOA1*	C,H,L,T	0.560	0.760	0.771	0.838	0.723	0.779
SIRT2	rs10405150	19	39387919	C	T	+··	*6.34E-06*					0.056	0.147	0.144	0.081	0.117	0.774
ZNF283	rs16976816	19	44339377	G	A	+··	*9.78E-06*					0.977	0.970	0.994	0.987	0.976	0.864
APOE	rs1160985*	19	45403412	C	T	+++	**2.13E-13**	**3.36E-09**	**4.29E-21**	*APOE*	C,H,L,T	0.724	0.659	0.590	0.554	0.447	0.378
HDL																	
Locus	SNP	Chr	BP	EA	OA	Dir	P_D	P_R	P_{DR}	Known gene	Traits	SAM	EAS	SAS	EUR	AMR	AFR
STON1-GTF2A1L	rs6739536	2	48831901	A	G	−··	*1.58E-06*					0.762	0.676	0.837	0.918	0.842	0.609
MGAT1	rs1038143*	5	180213878	T	C	−−−	*3.72E-06*	0.016	*2.91E-07*			0.309	0.110	0.137	0.148	0.081	0.013
AKAP7	rs3777486	6	131584648	G	A	−+−	*3.09E-06*	0.808	4.37E-04			0.976	0.909	0.898	0.858	0.793	0.949
CSMD1	rs1626142	8	4345284	T	C	−··	*7.67E-06*					0.612	0.400	0.435	0.298	0.392	0.654
RAB21	rs328733	12	72197574	T	C	−−−	*2.57E-06*	0.036	*5.92E-07*			0.788	0.626	0.700	0.869	0.738	0.611
ZNF10	rs2292029	12	133734113	A	G	−··	*4.05E-06*					0.179	0.112	0.254	0.245	0.156	0.008
HS6ST3	rs16953620	13	97508453	A	G	−··	*8.48E-06*					0.968	0.950	0.949	0.848	0.911	0.836
LIPC	rs10438284	15	58629424	G	A	−+−	*4.00E-07*	0.133	*9.15E-07*	*LIPC*	C,H,T	0.286	0.326	0.173	0.272	0.199	0.047
CETP	rs289708	16	57038162	T	C	−··	**1.19E-11**			*CETP*	C,H,L,T	0.905	0.815	0.800	0.860	0.823	0.597
LIPG	rs16950739	18	47138509	C	T	−··	*1.07E-07*			*LIPG*	C,H	0.019	0.058	0.190	0.057	0.140	0.004
APOE	rs1160985*	19	45403412	C	T	−+−	0.003	0.342	0.004	*APOE*	C,H,L,T	0.724	0.659	0.590	0.554	0.447	0.378
CDH4	rs817687	20	59753355	A	G	−−−	*2.31E-06*	0.237	*8.77E-06*			0.986	0.924	0.945	0.967	0.855	0.610
LDL																	
Locus	SNP	Chr	BP	EA	OA	Dir	P_D	P_R	P_{DR}	Known gene	Traits	SAM	EAS	SAS	EUR	AMR	AFR
APOB	rs754523*	2	21311691	G	A	+++	*3.25E-06*	0.158	*5.81E-06*	*APOB*	C,H,L,T	0.247	0.265	0.140	0.309	0.306	0.201
KALRN	rs6789134	3	123942339	G	A	++−	*3.22E-06*	0.925	1.96E-04			0.078	0.161	0.120	0.046	0.118	0.212
ZHX2	rs7841763	8	123971081	T	C	++−	*1.80E-06*	0.557	3.78E-05			0.043	0.023	0.146	0.102	0.058	0.169
SH2D4B	rs10509415	10	82473065	A	C	+··	*7.96E-06*					0.706	0.510	0.735	0.758	0.808	0.708
ALG10	rs3912355	12	34079616	C	T	+++	*2.12E-06*	0.544	3.97E-05			0.855	0.872	0.651	0.605	0.732	0.863
ALG10B	rs10880642	12	38554152	A	G	+··	*5.56E-06*					0.802	0.715	0.555	0.481	0.532	0.299
CPNE8	rs11169807	12	39244161	C	T	+++	*4.77E-06*	0.418	4.13E-05			0.794	0.644	0.536	0.505	0.408	0.898
LINC02408	rs17104016	12	67969929	A	T	+··	*9.29E-06*					0.574	0.734	0.924	0.853	0.906	0.838
LINC00922	rs254371	16	65943650	T	C	+··	*9.04E-06*					0.555	0.698	0.650	0.599	0.693	0.884
ZNF283	rs16976816	19	44339377	G	A	+··	*1.78E-06*					0.977	0.970	0.994	0.987	0.976	0.864
APOE	rs1160985	19	45403412	C	T	+++	**2.61E-20**	**5.07E-09**	**1.53E-27**	*APOE*	C,H,L,T	0.724	0.659	0.590	0.554	0.447	0.378

Triglycerides

Locus	SNP	Chr	BP	EA	OA	Dir	P_D	P_R	P_{DR}	Known gene	Traits	SAM	EAS	SAS	EUR	AMR	AFR
GCKR	rs780094	2	27741237	T	C	+++	*9.84E-07*	0.01	*5.62E-08*	*GCKR*	C,T	0.334	0.476	0.198	0.411	0.360	0.132
CD200	rs2399416	3	112059213	A	G	++−	*5.12E-06*	0.668	9.29E-04			0.021	0.148	0.140	0.393	0.272	0.101
SPIN1	rs7861888	9	90886340	A	G	+··	*4.24E-06*					0.706	0.719	0.897	0.921	0.842	0.989
APOA1	rs964184*	11	116648917	C	G	+++	**2.37E-17**	**8.97E-14**	**1.81E-29**	*APOA1*	C,H,L,T	0.440	0.240	0.229	0.162	0.277	0.221
KIRREL3	rs3018434	11	126805881	G	A	+··	*4.16E-06*					0.916	0.803	0.920	0.866	0.857	0.974
APOE	rs1160985*	19	45403412	T	C	+−+	0.312	0.836	0.507	*APOE*	C,H,L,T	0.276	0.341	0.410	0.446	0.553	0.623

EA = effect allele
OA = other allele (reference allele)
Dir = direction of the effect in each of the four samples (+ indicates the effect allele is increasing the trait value on the raw scale)
P_D = discovery cohort GWAS *p*-value
P_R = replication cohort *p*-value
P_{DR} = joint discovery and replication cohorts *p*-value
Known gene = Loci observed in Teslovich et al. 2010 or Willer et al. 2013
Traits = Traits locus associated with in Teslovich et al. 2010 or Willer et al. 2013
C = TC
H = HDL
L = LDL
T = TG
SAM = Samoan effect allele frequency (EAF)
EAS = East Asian EAF
SAS = South Asian EAF
EUR = European EAF
AMR = Admixed American EAF
AFR = African EAF

*The SNP genotyped in the replication population was not the peak SNP at this locus for this trait.

We compared the directions of effect to those previously reported in Willer et al. [8] and Teslovich et al. [30] for *APOB*, *GCKR*, *APOA1*, *LIPC*, *LIPG*, and *APOE* (i.e., genome-wide suggestive loci that have been previously associated with lipid traits). We observed a consistent direction of effect for the representative SNP for all associations except for *LIPG* and HDL (Table 2).

The effect allele frequencies in the two samples—discovery and replication—were largely similar for each of the 15 successfully genotyped SNPs (Tables S1, S7, S10, and S13 Appendix). However, many loci had markedly different effect allele frequencies between Samoans and other 1000 Genomes populations (Table 2). For example, compared to 1000 Genomes populations, there were higher effect allele frequencies (EAFs) in Samoans for rs964184 near *APOA1* (G allele frequency: 0.440 in Samoans vs. < 0.277 in 1000 Genomes populations), rs1160985 near *APOE* (C allele frequency: 0.724 in Samoans vs. < 0.659 in 1000 Genomes populations), and rs1038143 near *MGAT1* (A allele frequency: 0.309 in Samoans vs. < 0.148 in 1000 Genomes populations).

Discussion

In this study, we examined four measures of fasting lipid levels—TC, HDL, LDL and TG—for associations with 659,492 SNPs from a genome-wide array in a discovery cohort of 2,849 Samoans, with follow-up genotyping of significant and suggestive findings in a replication cohort comprising 1,798 Samoans from Samoa and American

Samoa. Thirty-one loci had at least suggestive evidence of association with one or more lipid traits in the discovery cohort, of which eight have been reported to be associated with lipid levels previously: *APOB, GCKR, MGAT1, APOA1*, *LIPC*, *CETP, LIPG,* and *APOE* [8, 10, 30] although the direction of effect for the variant near LIPG was in the opposite direction from previous results. Enrichment analyses highlighted known lipid metabolism gene sets and previously associated lipid loci.

We observed a difference in the architecture of the statistical association signals between the four lipid traits and variants near *APOE* (Fig 2E-H). The peak SNP for TC and LDL was rs1160985, an intronic variant in *TOMM40* upstream of *APOE*; whereas the peak SNP for HDL and TG was rs4420638, an intergenic variant downstream of *APOC1* and *APOE*. rs1160985 demonstrated evidence of replication for TC and LDL but not for HDL and TG, consistent with the discovery findings. These markers are in low linkage disequilibrium with each other (r^2 = 0.093) and may represent distinct association signals. While this could support a shared genetic architecture for TC and LDL and for HDL and TG in Polynesians, this study was not positioned to adequately capture the association signal present at this locus. Future studies with sequencing or imputation of the 19q13.2 region will be necessary to dissect the genetic architecture of *APOE* and lipid levels in Polynesians.

We did not observe suggestive or genome-wide significant association with several loci which have figured prominently in multiple lipid GWAS (e.g., *LPL*, *LDLR*, *CILP2*, *FADS1*/*2*/*3*, *ANGPTL3*, *SORT1*, *PPP1R3B*, *MIXIPL*, *HNF4A*, *PCSK9*, *GALNT2*,

HMGCR) either because we lacked sufficient power to detect their effects, the effects are negligible in Samoans, or the allele frequencies of associated variants are different enough in Samoans to hinder detection. However, it is important to note that this study was not designed to evaluate the effect of known lipids loci in Samoans, nor were previously-associated loci examined specifically.

We detected and replicated a suggestive association between HDL and a variant on 5q35.3 (Fig 2A). While the peak SNP lies within an intron of *BTNL8*, the variant selected for follow-up genotyping is intergenic, downstream of *MGAT1*. A suggestive association between variants near *MGAT1* and HDL in a GWAS in the Micronesian population of Kosrae has been previously reported [10]. Although there is no evidence of association between *MGAT1* and lipids as reported in prior studies of non-Pacific Islanders, variation near *MGAT1* has been associated with BMI, serum fatty acid levels and composition, and glucose response in Europeans [37–39]. The encoded MGAT enzyme plays a major role in the absorption of dietary fat in the intestine [40]. Due to the greater frequency of the HDL-associated risk variant observed in Samoans compared to 1000 Genomes populations, it is plausible that variation near *MGAT1* may have a unique role in or a stronger effect on the lipid metabolism of Pacific Islanders.

We also detected and replicated a novel suggestive association between HDL and a variant downstream of *RAB21* (Fig 2D). Unlike the variant downstream of *MGAT1*, we observed similar allele frequencies between Samoans and 1000 Genomes populations (i.e., < 20% difference between Samoans and another 1000 Genomes

423 population) in the HDL-associated variant downstream of *RAB21*. This region, 12q21,
424 was previously seen in linkage analysis with both univariate and bivariate scans of TC
425 and LDL [12]. The individuals included in this linkage analysis are also included in our
426 2002–03 replication sample set, however, they do not appear to be driving the
427 association signal near *RAB21* (Table S7 in S1 Appendix). Variation near *RAB21* has
428 been previously associated with obesity [41]. *RAB21* belongs to the family of
429 monomeric GTPases involved in control of cellular membrane trafficking and is involved
430 in the targeted trafficking of integrins and the regulation of cell adhesion and migration
431 [42, 43].
432
433 This study is limited in drawing conclusions about the genetic architecture of
434 lipids in Samoans, as replication genotyping was unavailable for many loci and due to
435 the lack of genome-wide imputation. Future studies, evaluating the evidence of
436 association between associations seen here in separate cohorts as well fine-mapping
437 loci with genotype imputation (given the availability of a relevant reference panel), are
438 necessary to fully evaluate the genetic architecture of lipids in Samoans.
439
440 This is the first GWAS of lipid phenotypes in Samoans, and we observed
441 association with many known lipid loci, which was further supported by the gene-set
442 enrichment analysis highlighting lipid metabolism gene sets. However, the difference in
443 association results near *APOE*, coupled with evidence of Pacific-Islander–specific
444 associations with *MGAT1* and *RAB21* suggest that some, but not all, of the genetic
445 architecture of lipids is shared between Samoans and other populations. Given this

446 evidence of a partially distinct genetic architecture of lipids in Samoans, further
447 investigation and fine-mapping of lipid loci, especially that across multiple ethnicities, is
448 warranted.
449

450 # Acknowledgements

451 The authors would like to thank the Samoan participants of the study, local
452 village authorities, and the many Samoan and other field workers over the years. We
453 acknowledge the Samoan Ministry of Health, the Samoa Bureau of Statistics, and the
454 American Samoan Department of Health for their support of this research. We give
455 particular thanks to two research assistants, Melania Selu and Vaimoana Lupematasila,
456 who contributed to the 2010 recruitment and continue to assist us in our work in Samoa.
457

References

1. Hawley NL, Minster RL, Weeks DE, Viali S, Reupena MS, Sun G, Cheng H, Deka R, Mcgarvey ST (2014) Prevalence of adiposity and associated cardiometabolic risk factors in the samoan genome-wide association study. Am J Hum Biol 26:491–501
2. Keighley E., McGarvey ST, Quested C, Mccuddin C, Viali S, Maga UA (2007) Nutrition and health in modernizing Samoans: temporal trends and adaptive perspectives. In: Ohtsuka R, Ulijaszek SJ (eds) Heal. Chang. Asia-Pacific Reg. Biocultural Epidemiol. Approaches. Cambridge University Press, Cambridge, NY, pp 147–191
3. McGarvey ST (2001) Cardiovascular disease (CVD) risk factors in Samoa and American Samoa, 1990-95. Pac Health Dialog 8:157–62
4. Kannel WB, Castelli WP, Gordon T, McNamara PM (1971) Serum cholesterol, lipoproteins, and the risk of coronary heart disease. The Framingham study. Ann Intern Med 74:1–12
5. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR (1977) High density lipoprotein as a protective factor against coronary heart disease. The Framingham Study. Am J Med 62:707–14
6. Roeters van Lennep J, Westerveld HT, Erkelens DW, van der Wall EE (2002) Risk factors for coronary heart disease: implications of gender. Cardiovasc Res 53:538–549
7. Lange LA, Willer CJ, Rich SS (2015) Recent developments in genome and exome-wide analyses of plasma lipids. Curr Opin Lipidol 26:96–102
8. Willer CJ, Schmidt EM, Sengupta S, et al (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45:1274–1283
9. Burkhardt R, Kenny EE, Lowe JK, et al (2008) Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. Arterioscler Thromb Vasc Biol 28:2078–2084
10. Lowe JK, Maller JB, Pe'er I, et al (2009) Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. PLoS Genet 5:e1000365
11. Dumitrescu L, Carty CL, Taylor K, et al (2011) Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. PLoS Genet 7:e1002138
12. Åberg K, Dai F, Sun G, et al (2008) A genome-wide linkage scan identifies multiple chromosomal regions influencing serum lipid levels in the population on the Samoan islands. J Lipid Res 49:2169–2178
13. Minster RL, Hawley NL, Su C-T, et al (2016) A thrifty variant in CREBRF strongly influences body mass index in Samoans. Nat Genet 48:1049–1054
14. Kristiansson K, Naukkarinen J, Peltonen L (2008) Isolated populations and complex disease gene identification. Genome Biol 9:109
15. Laurie CC, Doheny KF, Mirel DB, et al (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34:591–602

16. Deka R, Mc Garvey ST, Ferrell RE, Kamboh MI, Yu LM, Aston CE, Jin L, Chakraborty R (1994) Genetic characterization of American and Western Samoans. Hum Biol 66:805–22
17. McGarvey ST, Levinson PD, Bausser-Man L, Galanis DJ, Hornick CA (1993) Population change in adult obesity and blood lipids in American Samoa from 1976–1978 to 1990. Am J Hum Biol 5:17–30
18. Chin-Hong P V, McGarvey ST (1996) Lifestyle incongruity and adult blood pressure in Western Samoa. Psychosom Med 58:131–7
19. Galanis DJ, McGarvey ST, Quested C, Sio B, Afele-Fa'amuli SA (1999) Dietary intake of modernizing Samoans: implications for risk of cardiovascular disease. J Am Diet Assoc 99:184–90
20. Åberg K, Dai F, Sun G, et al (2009) Susceptibility Loci for Adiposity Phenotypes on 8p, 9p, and 16q in American Samoa and Samoa. Obesity 17:518–524
21. Dai F, Keighley ED, Sun G, et al (2007) Genome-wide scan for adiposity-related phenotypes in adults from American Samoa. Int J Obes 31:1832–1842
22. Åberg K, Dai F, Viali S, Tuitele J, Sun G, Indugula SR, Deka R, Weeks DE, McGarvey ST (2009) Suggestive linkage detected for blood pressure related traits on 2q and 22q in the population on the Samoan islands. BMC Med Genet 10:107
23. Dai F, Sun G, Åberg K, et al (2008) A Whole Genome Linkage Scan Identifies Multiple Chromosomal Regions Influencing Adiposity-Related Traits among Samoans. Ann Hum Genet 72:780–792
24. Chen W-M, Abecasis GR (2007) Family-Based Association Tests for Genomewide Association Scans. Am J Hum Genet 81:913–926
25. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
26. Purcell S, Cherny SS, Sham PC (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19:149–150
27. Carithers LJ, Ardlie K, Barcus M, et al (2015) A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank 13:311–319
28. Segrè A V, Groop L, Mootha VK, Daly MJ, Altshuler D (2010) Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. PLoS Genet 6:e1001058
29. Lee PH, O'Dushlaine C, Thomas B, Purcell SM (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics 28:1797–1799
30. Teslovich TM, Musunuru K, Smith A V., et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713
31. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23:1294–1296
32. R Core Team (2014) R: A language and environment for statistical computing.
33. Therneau T, Atkinson E, Sinnwell J, Schaid D, McDonnell S (2014) kinship2: Pedigree functions.
34. Winkler TW, Day FR, Croteau-Chonka DC, et al (2014) Quality control and conduct of genome-wide association meta-analyses. Nat Protoc 9:1192–1212

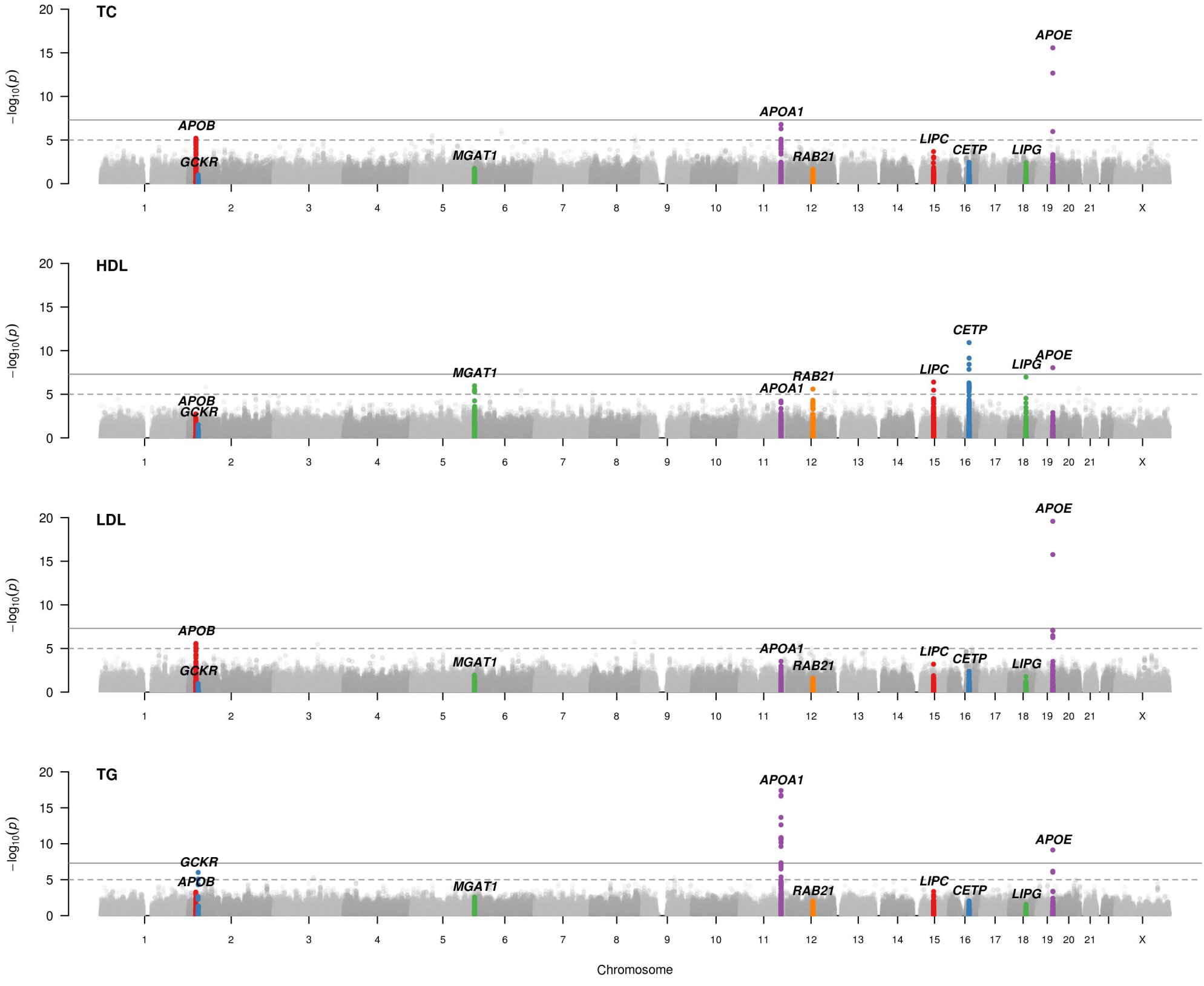
35. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191
36. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26:2336–2337
37. Johansson Å, Marroni F, Hayward C, et al (2010) Linkage and Genome-wide Association Analysis of Obesity-related Phenotypes: Association of Weight With the MGAT1 Gene. Obesity 18:803–808
38. Jacobsson JA, Rask-Andersen M, Risérus U, et al (2012) Genetic variants near the MGAT1 gene are associated with body weight, BMI and fatty acid metabolism among adults and children. Int J Obes 36:119–129
39. Del-Aguila JL, Beitelshees AL, Cooper-DeHoff RM, Chapman AB, Gums JG, Bailey K, Gong Y, Turner ST, Johnson JA, Boerwinkle E (2014) Genome-wide association analyses suggest NELL1 influences adverse metabolic response to HCTZ in African Americans. Pharmacogenomics J 14:35–40
40. Yen C-LE, Stone SJ, Cases S, Zhou P, Farese R V. (2002) Identification of a gene encoding MGAT1, a monoacylglycerol acyltransferase. Proc Natl Acad Sci 99:8512–8517
41. Turcot V, Lu Y, Highland HM, et al (2018) Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat Genet 50:26–41
42. Pellinen T, Ivaska J (2006) Integrin traffic. J Cell Sci 119:3723–31
43. Pellinen T, Tuomi S, Arjonen A, et al (2008) Integrin trafficking regulated by Rab21 is necessary for cytokinesis. Dev Cell 15:371–85

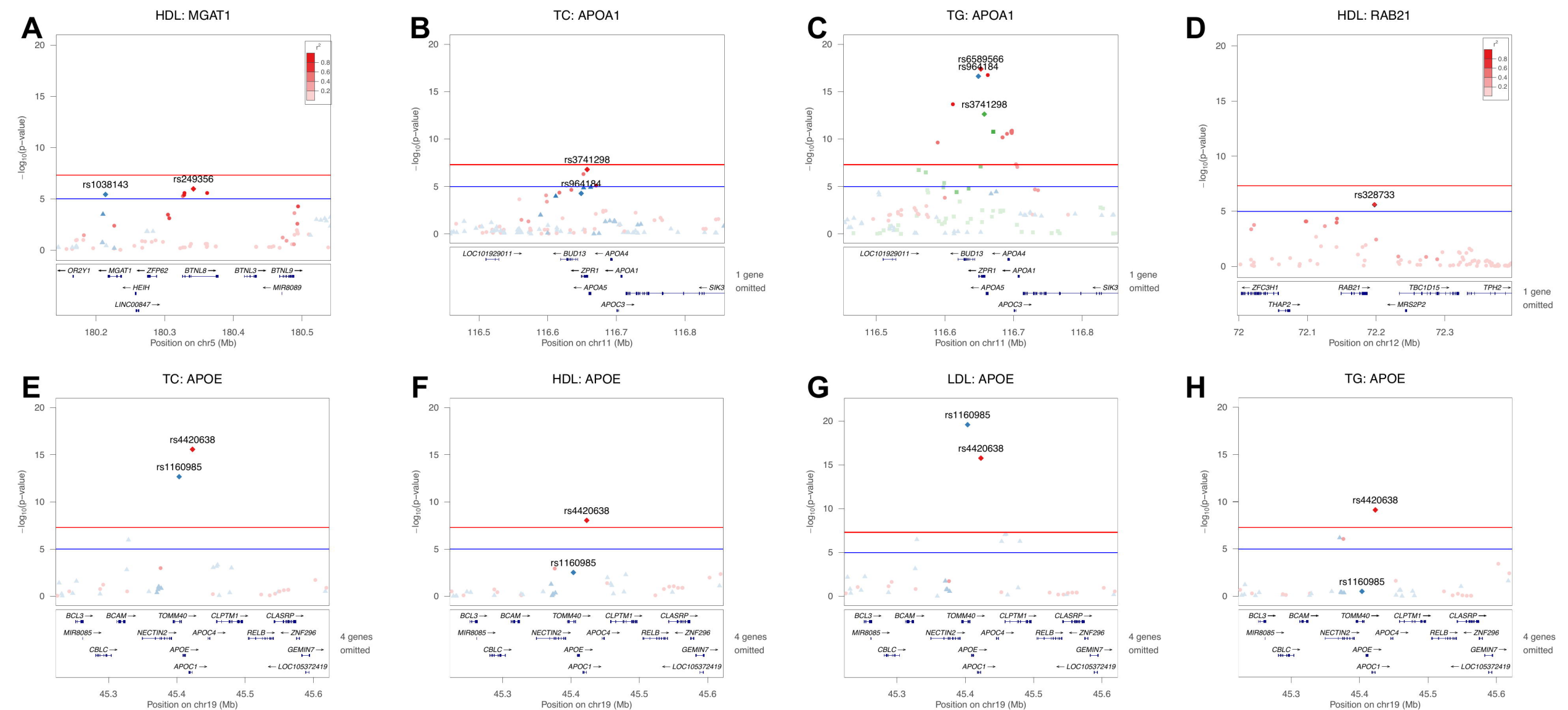
Supporting Information

S1 Appendix. Supporting Information.

Data Availability Statement

The discovery cohort data are available from dbGaP (accession number: phs000914.v1.p1).


TC
HDL
LDL
TG
−log10(p)
APOB
GCKR
MGAT1
APOA1
RAB21
LIPC
CETP
LIPG
APOE
0
5
10
15
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
X
Chromosome


A
HDL: MGAT1
rs1038143
rs249356
OR2Y1 MGAT1 ZFP62 BTNL8 BTNL3 BTNL9 HEIH MIR8089 LINC00847
Position on chr5 (Mb)
B
TC: APOA1
rs3741298
rs964184
LOC101929011 BUD13 APOA4 ZPR1 APOA1 APOA5 SIK3 APOC3
1 gene omitted
Position on chr11 (Mb)
C
TG: APOA1
rs6589566
rs964184
rs3741298
LOC101929011 BUD13 APOA4 ZPR1 APOA1 APOA5 SIK3 APOC3
1 gene omitted
Position on chr11 (Mb)
D
HDL: RAB21
rs328733
ZFC3H1 RAB21 TBC1D15 TPH2 THAP2 MRS2P2
1 gene omitted
Position on chr12 (Mb)
E
TC: APOE
rs4420638
rs1160985
BCL3 BCAM TOMM40 CLPTM1 CLASRP MIR8085 NECTIN2 APOC4 RELB ZNF296 CBLC APOE GEMIN7 APOC1 LOC105372419
4 genes omitted
Position on chr19 (Mb)
F
HDL: APOE
rs4420638
rs1160985
BCL3 BCAM TOMM40 CLPTM1 CLASRP MIR8085 NECTIN2 APOC4 RELB ZNF296 CBLC APOE GEMIN7 APOC1 LOC105372419
4 genes omitted
Position on chr19 (Mb)
G
LDL: APOE
rs1160985
rs4420638
BCL3 BCAM TOMM40 CLPTM1 CLASRP MIR8085 NECTIN2 APOC4 RELB ZNF296 CBLC APOE GEMIN7 APOC1 LOC105372419
4 genes omitted
Position on chr19 (Mb)
H
TG: APOE
rs4420638
rs1160985
BCL3 BCAM TOMM40 CLPTM1 CLASRP MIR8085 NECTIN2 APOC4 RELB ZNF296 CBLC APOE GEMIN7 APOC1 LOC105372419
4 genes omitted
Position on chr19 (Mb)
−log10(p-value)
r²