

University of Texas Rio Grande Valley

**ScholarWorks @ UTRGV**

---

School of Medicine Publications and  
Presentations

School of Medicine

---

4-4-2018

## **Exome Sequencing Identifies Genetic Variants Associated with Circulating Lipid Levels in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study (IRASFS)**

Chuan Gao

Keri L. Tabb

Latchezar Dimitrov

Kent D. Taylor

Nan Wang

*See next page for additional authors*

Follow this and additional works at: [https://scholarworks.utrgv.edu/som\\_pub](https://scholarworks.utrgv.edu/som_pub)



Part of the [Medicine and Health Sciences Commons](#)


---

---

**Authors**


Chuan Gao, Keri L. Tabb, Latchezar Dimitrov, Kent D. Taylor, Nan Wang, Xiuqing Guo, Jirong Long, Jerome I. Rotter, Richard M. Watanabe, Joanne E. Curran, and John Blangero

# SCIENTIFIC REPORTS



OPEN

## Exome Sequencing Identifies Genetic Variants Associated with Circulating Lipid Levels in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study (IRASFS)

Chuan Gao<sup>1,2</sup>, Keri L. Tabb<sup>2,3</sup>, Latchezar M. Dimitrov<sup>2</sup>, Kent D. Taylor<sup>5</sup>, Nan Wang<sup>6</sup>, Xiuqing Guo<sup>5</sup>, Jirong Long<sup>7</sup>, Jerome I. Rotter<sup>5</sup>, Richard M. Watanabe<sup>6</sup>, Joanne E. Curran<sup>8</sup>, John Blangero<sup>8</sup>, Carl D. Langefeld<sup>4</sup>, Donald W. Bowden<sup>2,3</sup> & Nicholette D. Palmer<sup>2,3</sup> 

Genome-wide association studies have identified numerous variants associated with lipid levels; yet, the majority are located in non-coding regions with unclear mechanisms. In the Insulin Resistance Atherosclerosis Family Study (IRASFS), heritability estimates suggest a strong genetic basis: low-density lipoprotein (LDL,  $h^2 = 0.50$ ), high-density lipoprotein (HDL,  $h^2 = 0.57$ ), total cholesterol (TC,  $h^2 = 0.53$ ), and triglyceride (TG,  $h^2 = 0.42$ ) levels. Exome sequencing of 1,205 Mexican Americans (90 pedigrees) from the IRASFS identified 548,889 variants and association and linkage analyses with lipid levels were performed. One genome-wide significant signal was detected in *APOA5* with TG (rs651821,  $P_{TG} = 3.67 \times 10^{-10}$ ,  $LOD_{TG} = 2.36$ ,  $MAF = 14.2\%$ ). In addition, two correlated SNPs ( $r^2 = 1.0$ ) rs189547099 ( $P_{TG} = 6.31 \times 10^{-08}$ ,  $LOD_{TG} = 3.13$ ,  $MAF = 0.50\%$ ) and chr4:157997598 ( $P_{TG} = 6.31 \times 10^{-08}$ ,  $LOD_{TG} = 3.13$ ,  $MAF = 0.50\%$ ) reached exome-wide significance ( $P < 9.11 \times 10^{-08}$ ). rs189547099 is an intronic SNP in *FNIP2* and SNP chr4:157997598 is intronic in *GLRB*. Linkage analysis revealed 46 SNPs with a  $LOD > 3$  with the strongest signal at rs1141070 ( $LOD_{LDL} = 4.30$ ,  $P_{LDL} = 0.33$ ,  $MAF = 21.6\%$ ) in *DFFB*. A total of 53 nominally associated variants ( $P < 5.00 \times 10^{-05}$ ,  $MAF \geq 1.0\%$ ) were selected for replication in six Mexican-American cohorts ( $N = 3,280$ ). The strongest signal observed was a synonymous variant (rs1160983,  $P_{LDL} = 4.44 \times 10^{-17}$ ,  $MAF = 2.7\%$ ) in *TOMM40*. Beyond primary findings, previously reported lipid loci were fine-mapped using exome sequencing in IRASFS. These results support that exome sequencing complements and extends insights into the genetics of lipid levels.

Cardiovascular disease (CVD) is the leading cause of death worldwide<sup>1</sup>. In the United States, CVD accounts for more deaths than any other major cause and, on average, 2,200 Americans die of CVD each day<sup>2</sup>. While the exact mechanism of disease remains unclear, lipid concentrations are well-accepted as a major risk factor as well

<sup>1</sup>Molecular Genetics and Genomics Program, Winston-Salem, NC, USA. <sup>2</sup>Center for Genomics and Personalized Medicine Research, Winston-Salem, NC, USA. <sup>3</sup>Department of Biochemistry, Winston-Salem, NC, USA. <sup>4</sup>Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA. <sup>5</sup>Institute for Translational Genomics and Population Sciences and Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. <sup>6</sup>Department of Preventive Medicine and Physiology and Biophysics, University of Southern California Keck School of Medicine, Los Angeles, CA, USA. <sup>7</sup>Department of Medicine and Vanderbilt Epidemiology Center Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA. <sup>8</sup>South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. Correspondence and requests for materials should be addressed to N.D.P. (email: [nallred@wakehealth.edu](mailto:nallred@wakehealth.edu))

	Discovery	Replication					
	IRASFS	IRAS	TRIPOD	BetaGene	HTN-IR	MACAD	NIDDM-Athero
n	1205	181	125	1218	763	749	244
Female (%)	58.5	41.1	100.0	27.5	41.1	42.6	38.9
Age (years) <sup>b</sup>	42.7 ± 14.5	54.1 ± 8.2	34.9 ± 6.4	34.7 ± 7.9	39.3 ± 15.1	34.7 ± 9.1	38.1 ± 14.9
Body Mass Index (BMI; kg/m <sup>2</sup> )	28.9 ± 6.2	28.2 ± 5.1	30.8 ± 5.6	29.6 ± 6.1	29.3 ± 5.7	29.0 ± 5.2	29.1 ± 6.3
High-density Lipoprotein (HDL; mg/dl)	43.61 ± 12.86	43.18 ± 14.63	37.39 ± 9.52	46.89 ± 11.08	48.13 ± 13.24	46.14 ± 12.12	47.5 ± 11.7
Low-density Lipoprotein (LDL; mg/dl)	109.41 ± 31.04	140.04 ± 36.75	109.01 ± 27.30	102.80 ± 28.65	106.03 ± 31.38	108.23 ± 31.3	106.63 ± 32.01
Total Cholesterol (TC; mg/dl)	178.06 ± 37.45	210.90 ± 43.74	173.15 ± 32.59	172.52 ± 33.25	179.01 ± 35.41	180.53 ± 36.43	181.47 ± 39.23
Triglycerides (TG; mg/dl)	124.80 ± 84.00	157.24 ± 99.21	133.79 ± 78.91	114.12 ± 90.27	124.23 ± 78.96	137.25 ± 102.82	145.19 ± 166

**Table 1.** Demographic information of the study individuals.

as clinical indicators for CVD. Genetic studies have suggested a strong heritability for circulating lipid levels, i.e. total cholesterol (TC), LDL cholesterol (LDL), HDL cholesterol (HDL), and triglycerides (TG). Based on a European twin-pairs study, it is estimated that circulating lipid heritability ranges from 0.58 to 0.66 ( $h^2_{HDL} = 0.61$ ,  $h^2_{LDL} = 0.59$ ,  $h^2_{TC} = 0.58$ ,  $h^2_{TG} = 0.66$ )<sup>3</sup>.

Given the public health relevance as well as the strong genetic component, numerous genome-wide association studies (GWAS) have been performed to investigate the genetic architecture of circulating lipid levels. The most recent Global Lipid Genetics Consortium (GLGC) analyzed 188,577 individuals from four ethnicities (Europeans, East Asians, South Asians, and Africans) and identified 157 loci associated with plasma lipid traits<sup>4</sup>. While well-powered, these efforts have largely overlooked the fastest growing US minority population, Hispanic Americans. Compared to non-Hispanic whites, Hispanics suffer an even higher risk for CVD, i.e. 32.1% versus 23.8%<sup>5,6</sup>. Until now, the largest lipid GWAS in Hispanics was performed by Below *et al.* in 2015 with 4,383 Mexican ancestry individuals. However, all genome-wide significant regions identified in the Mexican meta-analysis were previously identified<sup>4,7–9</sup>.

GWAS were designed with a focus on common variants, partly supported by the “common disease, common variant” hypothesis<sup>10</sup>. However, despite the large number of genetic signals identified by GWAS, over 80% fall outside of protein coding regions, which complicates causal inference<sup>10</sup>. Among genes with evidence of association, only a small proportion of the variance is explained, providing limited information for disease risk prediction<sup>4,11</sup>. Sequencing of the exome, rather than the entire human genome, has been shown to be an efficient strategy to search for novel variants with a clear biological mechanism<sup>12</sup>. Previous exome sequencing studies have identified multiple rare variants associated with CVD in non-Hispanic populations<sup>13,14</sup>.

To search for functional coding variants regulating circulating lipid levels in a Mexican-ancestry population, whole exome sequencing was performed in 1,205 Mexican Americans from the Insulin Resistance Atherosclerosis Family Study (IRASFS). Association and family-based linkage analyses were performed for 548,889 variants. We hypothesized that a family cohort would have increased power to detect rare variants due to their transmission in multigenerational pedigrees. With the complimentary approaches of linkage and association, exome sequencing has the potential to identify ethnic-specific variants regulating lipid levels.

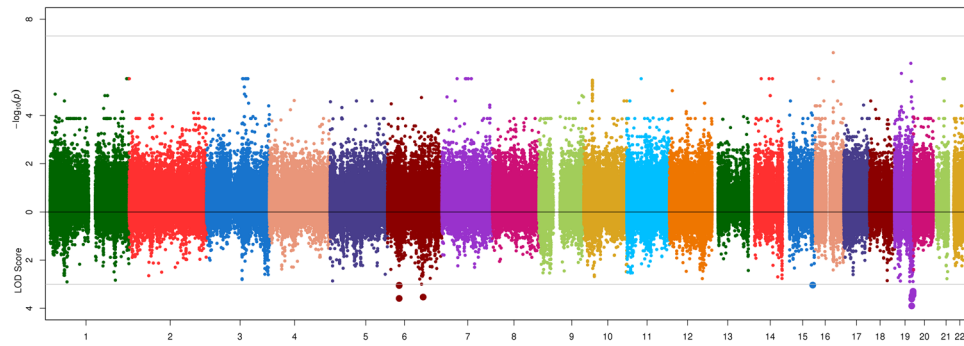
## Results

A total of 1,205 individuals were included in association and linkage analyses. Characteristics of the study individuals are shown in Table 1. Overall, individuals were predominantly female (59%) and were overweight with an average BMI of 28.9 kg/m<sup>2</sup>. Since the recruitment was based on family size rather than diagnosis, e.g. CVD was not required for participation, the participants were metabolically normal, with an average HDL (43.61 mg/dl), LDL (109.41 mg/dl), TC (178.06 mg/dl), and TG (124.80 mg/dl) within desirable or near-desirable ranges<sup>15</sup>. According to the National Cholesterol Education Program (NCEP)<sup>15</sup>, 183 individuals (15%) within the study had undesirable high TG levels (TG > 150 mg/dl), 121 (10%) individuals had an LDL level greater than 160 mg/dl, 521 (43%) individuals had an HDL level less than 40 mg/dl, and 290 (24%) individuals had a TC level greater than 200 mg/dl.

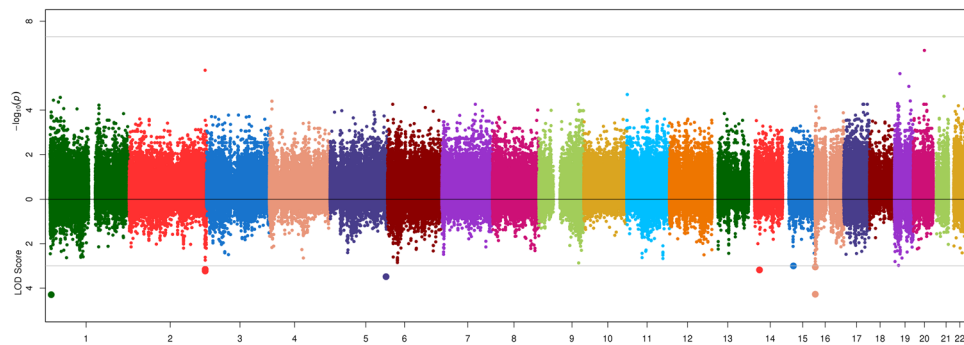
Heritability analysis in IRASFS suggested a strong genetic component for lipid levels. HDL had the strongest heritability ( $h^2_{HDL} = 0.57$ ) with 17.2% of the variance explained by covariates (age, sex, center, BMI;  $P = 5.87 \times 10^{-26}$ ), the heritability of TC was 0.53 with 10.3% of the variance explained by covariates ( $P = 3.12 \times 10^{-23}$ ), the heritability of LDL was 0.50 with 7.3% of the variance explained by covariates ( $P = 1.22 \times 10^{-21}$ ), and TG had the lowest heritability ( $h^2_{TG} = 0.42$ ) with 15.9% of the variance explained by covariates ( $P = 5.37 \times 10^{-14}$ ) (Table S1).

From exome sequencing data, a total of 548,889 variants were successfully analyzed for association and linkage. Among them, 30.0% (164,591) were extremely rare variants with only one or two observations and 82.5% (452,807) were low frequency variants as defined by a MAF < 5% (Figure S1); 6.8% (N = 37,157) of the variants were insertions/deletions; 33.1% (N = 182,020) and 15% (N = 83,874) of the variants marked a non-synonymous and synonymous amino acid change in coding genes, respectively.

Association and linkage results are shown in Figs 1–4. The strongest evidence of association attaining genome-wide significance ( $P < 5.00 \times 10^{-8}$ ) was observed at the apolipoprotein A-V gene (*APOA5*) on chromosome 11 with two highly correlated SNPs ( $r^2 = 0.92$ ) rs651821 ( $P = 3.67 \times 10^{-10}$ , LOD = 2.36, MAF = 14%) and rs2072560 ( $P = 5.14 \times 10^{-10}$ , LOD = 2.05, MAF = 13%) and TG (Fig. 4; Table 2). Conditional analysis on one



**Figure 1.** Opposed plots for association and linkage analysis of HDL. Association results are plotted on the positive y-axis. The line at  $-\log_{10}(\text{PVAL}) = 7.30$  represents genome-wide significance ( $P = 5.00 \times 10^{-08}$ ). The linkage results are plotted on the negative y-axis. The line at  $\text{LOD} = 3$  represents a significant linkage signal.



**Figure 2.** Opposed plots for association and linkage analysis of LDL. Association results are plotted on the positive y-axis. The line at  $-\log_{10}(\text{PVAL}) = 7.30$  represents genome-wide significance ( $P = 5.00 \times 10^{-08}$ ). The linkage results are plotted on the negative y-axis. The line at  $\text{LOD} = 3$  represents a significant linkage signal.

variant was able to abolish the signal limiting the ability to statistically implicate a functional variant. In addition, both SNPs were also nominally associated with HDL ( $P_{\text{rs651821}} = 2.63 \times 10^{-3}$ ;  $P_{\text{rs2072560}} = 9.42 \times 10^{-3}$ ), while no signal was detected for LDL ( $P > 0.90$ ) or TC ( $P > 0.18$ ). On average, individuals had 23% more TG for each risk allele carried at rs651821 (TT: 117.59 mg/dl, TC: 142.62 mg/dl, CC: 174.41 mg/dl). Nominally associated and linked signals are provided in Table S2.

Additional evidence of association which attained exome-wide significance ( $P < 9.11 \times 10^{-08}$ ; based on a conservative Bonferroni correction for 548,889 variants) included two correlated SNPs ( $r^2 = 1.0$ ) on chromosome 4: rs189547099 ( $P_{\text{TG}} = 6.31 \times 10^{-08}$ ,  $\text{LOD}_{\text{TG}} = 3.13$ ,  $\text{MAF} = 0.5\%$ ) and chr4:157997598 ( $P_{\text{TG}} = 6.31 \times 10^{-08}$ ,  $\text{LOD}_{\text{TG}} = 3.13$ ,  $\text{MAF} = 0.5\%$ ) (Table 2). Notably, these variants were both significantly associated and linked with TG. SNP rs189547099 is an intronic SNP located in the folliculin interacting protein 2 gene (*FNIP2*) and the chromosome 4 open reading frame 45 (*C4orf25*). SNP chr4:157997598 is located 1,817 kb upstream of SNP rs189547099 in the first intron of the glycine receptor beta gene (*GLRB*). On average, risk allele (T) heterozygous carriers for chr4:157997598 had 2.9 times TG compared to non-carriers (CC: 124 mg/dl vs. TC: 358 mg/dl). No risk allele homozygotes were found. Burden testing failed to identify additional genes significantly associated with lipid phenotypes after correction for the number of test performed, i.e.  $P < 7.23 \times 10^{-06}$ , 0.05/139,173; Table S3.

Two-point linkage analysis was performed for 548,889 variants. Of these, two SNPs had a LOD score greater than 4 (Table 2), 46 SNPs had a LOD score greater than 3 (Table S4). Among these, 14 variants were significantly linked ( $\text{LOD} > 3$ ) with TC, 13 variants were significantly linked with HDL, 13 variants were significantly linked with TG, and 8 variants were significantly linked with LDL with two variants overlapping between TC and LDL. The strongest linkage signal was observed at SNP rs1141070 ( $\text{LOD}_{\text{LDL}} = 4.30$ ,  $\text{LOD}_{\text{TC}} = 3.93$ ,  $\text{MAF} = 22\%$ ) with LDL and TC levels. rs1141070 is located in exon 5 of the DNA fragmentation factor gene (*DFFB*) on chromosome 1 and marks a synonymous amino acid substitution. In addition, SNP rs11648905 ( $\text{LOD}_{\text{LDL}} = 4.28$ ,  $\text{LOD}_{\text{TC}} = 3.77$ ,  $\text{MAF} = 41\%$ ) was strongly linked with LDL and TC levels. This variant is an intronic SNP located between exon 7 and 8 of the transmembrane protein 8 A gene (*TMEM8A*). No significant association signals were observed for the two linked signals: rs1141070,  $P_{\text{LDL}} = 0.33$ ,  $P_{\text{TC}} = 0.74$ ; rs11648905,  $P_{\text{LDL}} = 0.78$ ,  $P_{\text{TC}} = 0.84$  (Table 2).

**Meta-analysis.** Meta-analysis with six additional independent cohorts was computed for the 53 selected SNPs. Overall, six variants reached genome-wide significance after meta-analysis (Table 3). The strongest signal was observed at rs1160983 ( $P_{\text{LDL}} = 4.44 \times 10^{-17}$ ,  $\text{MAF} = 2.7\%$ ) with LDL. rs1160983 is a synonymous coding variant in the translocase of outer mitochondrial membrane 40 gene (*TOMM40*). The two *APOA5* variants,

SNP	Chr:Pos (hg19)	Gene	Annotation	Alleles <sup>a</sup>	RAF <sup>b</sup>	P <sub>HDL</sub>	LOD <sub>HDL</sub>	P <sub>LDL</sub>	LOG <sub>LDL</sub>	P <sub>TC</sub>	LOD <sub>TC</sub>	P <sub>TG</sub>	LOD <sub>TG</sub>
rs1141070	1:3786189	<i>DFFB</i>	coding-synon	A/G	0.22	0.98	0.01	0.33	<b>4.30</b>	0.74	<b>3.93</b>	0.87	0.27
chr4:157997598	4:157997598	<i>GLRB</i>	intron	T/C	0.0050	2.88E-04	1.87	0.71	0	0.24	0	<b>6.31E-08</b>	<b>3.13</b>
rs189547099	4:159814881	<i>C4orf45</i>	intron	C/G	0.0050	2.88E-04	1.87	0.71	0	0.24	0	<b>6.31E-08</b>	<b>3.13</b>
rs2072560	11:116661826	<i>APOA5</i>	intron	T/C	0.13	2.03E-03	0.32	0.87	0	0.014	0.010	<b>5.14E-10</b>	2.06
rs651821	11:116662579	<i>APOA5</i>	intron	C/T	0.14	8.93E-04	0.30	0.99	0	0.018	0	<b>3.67E-10</b>	2.36
rs11648905	16:425298	<i>TMEM8A</i>	intron	T/G	0.41	0.74	0	0.78	<b>4.28</b>	0.84	<b>3.77</b>	0.18	0.82

**Table 2.** Top association ( $P < 9.11 \times 10^{-08}$ ) and linkage signals ( $LOD > 4$ ). <sup>a</sup>Reference (minor)/Other allele; <sup>b</sup>Reference allele frequency based on the entire population.

SNP	Chr:Pos (hg19)	RAF <sup>a</sup>	Trait	Gene	Annotation	IRASFS		Meta-analysis	
						N	P	N	P
rs2072560	11:116661826	0.13	TG	<i>APOA5</i>	intron	1205	5.14E-10	4241	5.67E-16
rs651821	11:116662579	0.14	TG	<i>APOA5</i>	intron	1205	3.67E-10	4241	2.66E-15
rs2070665	11:116707684	0.17	TG	<i>APOA1</i>	intron	1205	4.10E-05	4241	7.03E-09
rs1532625	16:57005301	0.38	HDL	<i>CETP</i>	intron	1205	2.46E-07	4235	7.72E-14
rs11076176	16:57007446	0.28	HDL	<i>CETP</i>	intron	1205	3.87E-06	4235	2.15E-08
rs1160983	19:45397229	0.027	LDL	<i>TOMM40</i>	synonymous	1205	8.61E-06	4177	4.44E-17

**Table 3.** Top association signals ( $P < 5 \times 10^{-8}$ ) from meta-analysis. <sup>a</sup>Reference allele frequency based on the IRASFS cohort.

which attained genome-wide significance in IRASFS, were also successfully replicated (meta-analysis p-values: rs2072560,  $P_{TG} = 5.67 \times 10^{-16}$ ; rs651821,  $P_{TG} = 2.66 \times 10^{-15}$ ) with a consistent direction of effect across all cohorts. In addition, strong meta-analysis signals were detected for *APOA1* (rs2070665,  $P_{TG} = 7.03 \times 10^{-09}$ ) and *CETP* (rs1532625,  $P_{HDL} = 7.72 \times 10^{-14}$ , rs11076176,  $P_{HDL} = 2.15 \times 10^{-08}$ ) with TG and HDL, respectively. SNP rs72685601 was selected as the proxy SNP ( $r^2 = 0.59$ ) for the two variants that reached exome-wide significance (chr4:157997598, rs189547099). It was nominally associated with TG ( $P_{TG} = 3.69 \times 10^{-03}$ ) with consistent direction of effect across six of the seven cohorts. A complete list of meta-analysis results can be found in Table S5.

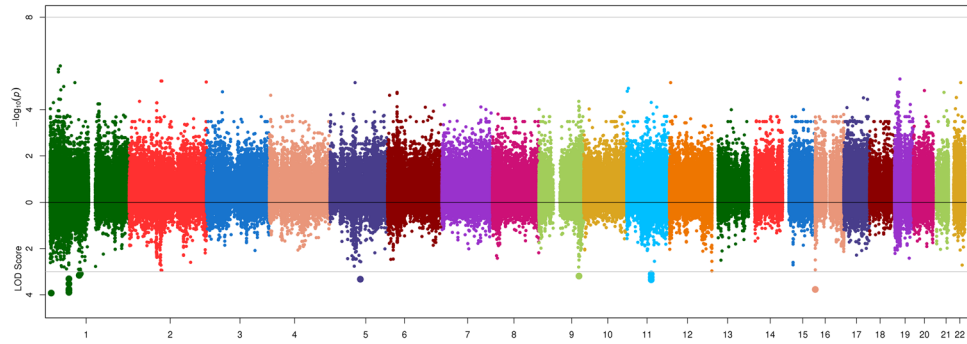
**Previously identified lipid loci.** Fine-mapping of the previously identified 157 loci  $\pm 100$  kb resulted in a total of 14,232 exome sequencing variants which were analyzed for association and linkage in IRASFS. In addition, conditional analyses based on the locus-specific GLGC index SNP were performed (Table S6). For association, 1,231 of the variants were identified with a P-value less than 0.05 with at least one of the reported lipid traits. Among the 1,231 significant variants, 646 remained to be significant ( $P < 0.05$ ) after conditional analysis based on the locus-specific index SNP. The strongest association signal was observed at rs651821 ( $P_{TG} = 3.67 \times 10^{-10}$ ,  $LOD_{TG} = 2.36$ ) in *APOA5*. After conditioning on SNP rs964184, it remained nominally associated and linked with TG ( $P_{TG|rs964184} = 1.76 \times 10^{-04}$ ,  $LOD_{TG|rs964184} = 0.99$ ). In addition, SNP rs1532625 also survived the stringent Bonferroni correction ( $P < 3.51 \times 10^{-06}$  for 14,232 variants):  $P_{HDL} = 2.46 \times 10^{-07}$ ,  $LOD_{HDL} = 1.42$ . This SNP is an intronic variant located in the cholesteryl ester transfer protein gene (*CETP*). However, conditional analysis with the index SNP rs3764261 totally abolished the signal ( $P_{HDL|rs3764261} = 0.61$ ,  $LOD_{HDL|rs3764261} = 0.02$ ). For linkage analysis, 14 and 127 SNPs reached a LOD score greater than two and one, respectively, for at least one reported lipid trait. Among them, one (rs1134760,  $LOD_{HDL|rs16942887} = 2.46$ ) and 28 ( $LOD > 1$ ) remained to be linked after conditional analysis, respectively.

## Discussion

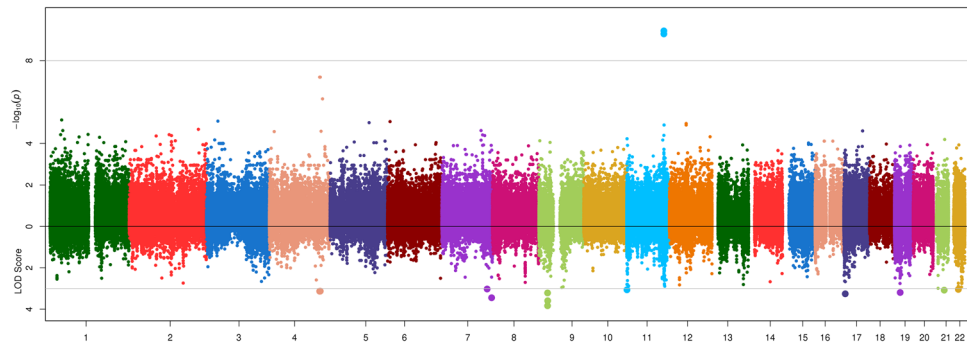
An individual's lipid profile represents a well-accepted major risk factor and clinical indicator of CVD. In this study, heritability estimates for four lipid phenotypes were reported in Mexican Americans from IRASFS and demonstrated a strong genetic component. Subsequently, exome-wide association analysis was performed using exome sequencing data derived from 1,205 Mexican Americans from IRASFS. Multiple significant association signals were identified and top signals were evaluated in six additional independent cohorts ( $n = 3,280$ ). As a complementary approach to association, linkage analysis was performed to identify rare variant signals. In addition, 157 previously identified lipid loci were fine-mapped using exome sequencing data in IRASFS.

Strong association and suggestive linkage signals were observed with two SNPs in *APOA5*: rs651821 ( $P_{TG} = 3.67 \times 10^{-10}$ ,  $LOD_{TG} = 2.36$ ,  $MAF = 14\%$ ) and rs2072560 ( $P_{TG} = 5.14 \times 10^{-10}$ ,  $LOD_{TG} = 2.05$ ,  $MAF = 13\%$ ). *APOA5* encodes an apolipoprotein that plays an important role in regulating plasma triglyceride levels, which is a strong risk factor for CVD<sup>16</sup>. This gene is located within the apolipoprotein gene cluster on chromosome 11q23.3, which contains multiple lipid-related genes including *APOA1*, *APOA3*, *APOA4*, *APOA5*, and *PCSK7*. Multiple strong association signals have been identified in the region with HDL, TG, and TC<sup>4,9,17</sup>. In 2015, Do *et al.*<sup>14</sup> described a large exome sequencing study in 9,793 European and African Americans and identified strong associations between *APOA5* functional variants and myocardial infarction (MI). A recent Hispanic GWAS of lipid





**Figure 3.** Opposed plots for association and linkage analysis of TC. Association results are plotted on the positive y-axis. The line at  $-\log_{10}(\text{PVAL}) = 7.30$  represents genome-wide significance ( $P = 5.00 \times 10^{-08}$ ). The linkage results are plotted on the negative y-axis. The line at  $\text{LOD} = 3$  represents a significant linkage signal.



**Figure 4.** Opposed plots for association and linkage analysis of TG. Association results are plotted on the positive y-axis. The line at  $-\log_{10}(\text{PVAL}) = 7.30$  represents genome-wide significance ( $P = 5.00 \times 10^{-08}$ ). The linkage results are plotted on the negative y-axis. The line at  $\text{LOD} = 3$  represents a significant linkage signal.

phenotypes identified SNP rs964184, 359 bases downstream of zinc finger protein 259 (*ZNF259*) and 11 kb downstream of *APOA5*, as significantly associated with TG<sup>7</sup>. In addition, Parra *et al.* presented robust association for rs964184 and no comparable signals were identified in the 5' UTR for *APOA5*<sup>18</sup>. In IRASFS, SNP rs964184 was also associated with TG ( $P = 4.79 \times 10^{-07}$ ). However, including SNP rs964184 as a covariant failed to completely abolish the genetic signal of rs651821 ( $P_{\text{before}} = 3.67 \times 10^{-10}$ ,  $P_{\text{after}} = 1.76 \times 10^{-04}$ ), suggesting that the two signals were likely independent ( $r^2 = 0.37$ ) (Figure S2).

In IRASFS, two common variants (rs651821, rs2072560) within *APOA5* were identified with strong association signals with TG in a Mexican-American family cohort. While SNP rs651821 and rs2072560 have been previously identified to be strongly associated with TG in multiple ethnicities (Europeans, East Asians, and North Africans)<sup>19–23</sup>, this is the first reported evidence in a Mexican-ancestry population. SNP rs651821 is a 5'-UTR variant that is three bases upstream of the coding exon. Worth mentioning, rs651821 is also located in the binding site of transcription factor POLR2A as suggested in HepG2 cells by ENCODE<sup>24</sup>. Previous expression quantitative trait loci (eQTL) studies revealed associations between rs651821 and transgelin (*TAGLN*) gene expression levels<sup>25</sup> yet no *APOA5* expression regulation effect was found. SNP rs2072560 is located intronically between exons 3 and 4 of *APOA5*. Interestingly, rs2072560 is also a missense variant of an alternative transcript of *APOA5* (NM\_001166598) which contains two exons. This variant marks a glutamic acid to glycine amino acid change (E66G) in exon 2 (Figure S5). To further explore the potential function of the alternative transcript, four ENCODE primary hepatocyte RNA sequencing experiments were analyzed and plotted using the UCSC genome browser<sup>24,26</sup>. However, no RNA sequencing evidence was found to support existence or function of the alternative transcript (Figure S5). Taken together, strong association, linkage, and replication signals were identified for the two *APOA5* SNPs with TG in Mexican Americans. While not enough biological evidence was found to support their causality, the results refined the scope of the *APOA5* association signals and provided information for future efforts to locate the causal variant in the region.

While not attaining strict genome-wide significance, two correlated SNPs ( $r^2 = 1.0$ ) rs189547099 ( $P_{\text{TG}} = 6.31 \times 10^{-08}$ ,  $\text{LOD}_{\text{TG}} = 3.13$ ,  $\text{MAF} = 0.5\%$ ) and chr4:157997598 ( $P_{\text{TG}} = 6.31 \times 10^{-08}$ ,  $\text{LOD}_{\text{TG}} = 3.13$ ,  $\text{MAF} = 0.5\%$ ) were detected with exome-wide significance. These are two rare SNPs with 12 heterozygous and no homozygous carriers. SNP rs189547099 is an intronic variant for both the chromosome 4 open reading frame 45 gene (*C4orf45*) and folliculin interacting protein 2 gene (*FNIP2*). *C4orf45* is an uncharacterized gene with unknown biological function. GTEx<sup>27</sup> has detected that *C4orf45* is strongly expressed in testis, yet there was almost no expression in other tissues. *FNIP2* is a tumor suppressor gene that has been shown to be involved

in regulating the apoptosis signaling pathway in tumors and is responsible for cellular metabolism and nutrient sensing<sup>28,29</sup>. SNP chr4:157997598 is an intronic variant in the glycine receptor beta gene (*GLRB*). This gene encodes the beta subunit of the glycine receptor and has been shown to function as a neurotransmitter-gated ion channel. Mutations in this gene have been shown to cause startle disease<sup>30,31</sup>. Interestingly, SNP chr4:157997598 is located in a CpG island and modifies the binding consensus sequence for a transcription factor zinc finger protein 263 (*ZNF263*) (Figure S6). Although no biological mechanism was found between *GLRB* and lipid metabolism, it is possible that SNP chr4:157997598 regulates TG levels through *ZNF263*.

Among the 53 variants identified for replication, synonymous SNP rs1160983 in exon 5 of the translocase of outer mitochondrial membrane 40 gene (*TOMM40*) exhibited the strongest association signal after replication and meta-analysis with six independent cohorts ( $P_{LDL} = 4.44 \times 10^{-17}$ , MAF = 2.7%). Before meta-analysis, this variant was nominally associated in IRASFS ( $P_{LDL} = 8.61 \times 10^{-06}$ ). *TOMM40* has been shown to be the forming subunit of the translocase of the mitochondrial outer membrane complex and is essential for the import of protein precursors into mitochondria<sup>32</sup>. Genetic studies have identified two adjacent genes (*APOE* and *TOMM40*) in this region to be highly associated with circulating lipid levels<sup>4,33</sup>. After reviewing previously identified *APOE* variants, SNP rs7412 has the highest LD with rs1160983 ( $r^2 = 0.49$ ,  $D' = 0.94$ ). In IRASFS, rs7412 was nominally associated with LDL ( $P_{LDL} = 6.61 \times 10^{-06}$ ). Interestingly, after adjusting for the *APOE* variant (rs7412), rs1160983 remained nominally significant ( $P = 9.10 \times 10^{-03}$ ) with LDL. This suggests that the known *APOE* signals do not fully explain the rs1160983 signal in *TOMM40*. It is possible that *TOMM40* may directly contribute to the regulation of LDL levels or SNP rs1160983 may influence *APOE* expression.

An interesting observation from this study is the lack of overlap between the majority of linkage and association signals, even with exome sequencing data. One explanation is that association and linkage capture different mechanisms of phenotypic contributions. Association analysis detects signals that statistically associate with phenotypic variability either directly or through linkage disequilibrium (LD) and thus targets more proximal effects. Linkage detects the co-segregation of an allele with the phenotype in families and therefore can detect long-range effects due to limited recombination events across successive generations. Therefore, each approach has its advantages and limitations. For example, association analysis has gained much success in common variant analysis while often suffering from reduced power to detect rare variants, e.g. statistical power is largely affected by inadequate sample size and limited LD with proximal common variants. In contrast, linkage analysis performance is largely dependent on family structures as well as the number of segregation events, e.g. when family structures are incomplete or allele segregation information is incomplete (only two generations or the parental generation allele information is missing), linkage analysis performance is largely dampened. On the other hand, rare variants (in populations) that are missed by association can be relatively common in a given family with segregation across multiple generations. In this scenario, linkage has increased power over association. Taken together, both association and linkage analysis are valuable approaches for analyzing sequencing data in genetic studies, providing potentially independent information.

Despite multiple strong signals identified, study limitations exist. First, the modest sample size in IRASFS ( $n = 1,205$ ) limits the power for the association analysis of rare variants, especially given the fact that 82% of the variants analyzed had a MAF < 5%. As a complementary approach, linkage analysis in 90 pedigrees was performed and identified signals that were likely missed by association. Unfortunately, linkage analysis was unavailable in the replication cohorts, and therefore meta-analysis of linkage signals was not performed. Dyslipidemia was not required for participation in IRASFS, thus the majority of individuals were metabolically normal, potentially providing limited enrichment of genetic risk alleles. Third, exome sequencing was not available in replication cohorts, and therefore replication was limited to GWAS imputed or proxy variants only. This approach modestly limited the number of SNPs for replication. Also, while all cohorts were of Mexican ancestry, different ascertainment criteria were used. For example, BetaGene recruited participants at high risk of gestational diabetes while HTN-IR recruited participants at high risk of hypertension. This differs from IRASFS which is a population-based study recruited for large family size.

In summary, exome-wide association and linkage analyses were performed using exome sequencing data in 1,205 Mexican Americans from IRASFS. Multiple signals were detected with circulating lipid levels and top signals were analyzed in six additional independent cohorts for replication. Our results suggested multiple lipid genetic signals in *APOA5*, *TOMM40*, and *GLRB/C4orf45*, fine-mapped known lipid genes with exome sequencing data in IRASFS, and explored a combined approach of association and linkage analyzing sequencing data. These results confirm that exome sequencing is a powerful tool to screen for functional genetic variants in the population.

## Methods

**Insulin Resistance Atherosclerosis Family Study (IRASFS).** The study design, recruitment, and phenotyping for the IRASFS has been previously described<sup>34</sup>. In brief, the IRASFS was designed to investigate the genetic and environmental basis of insulin resistance and adiposity. Mexican Americans included in this cohort ( $n = 1,205$  individuals, 90 pedigrees) were recruited from clinical centers in San Antonio, TX and San Luis Valley, CO. Recruitment was based on reported family size and not on health status. Phenotype acquisition and variable calculations have been previously described<sup>34,35</sup>. In brief, TC, TG, and HDL were measured from fasting plasma with standards and LDL was calculated using the Friedewald formula. The study protocol was approved by the Institutional Review Board of each participating clinical and analysis site and all participants provided their written informed consent. All methods in this study were carried out in accordance with the principles of the Declaration of Helsinki.

**Exome Sequencing.** Exome sequencing was performed at Texas Biomedical Research Institute using the Illumina Nextera Exome Enrichment System in conjunction with the Illumina HiSeq 2500 sequencer. All



sequence reads passed through the Illumina Data Analysis Pipeline, and those from samples passing QC criteria were mapped to the human genome reference sequence (hg19). A detailed description of the sequencing platform and analysis pipeline has been published<sup>36</sup>. Of note, multi-sample recalibration was performed prior to variant calling. The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Statistical Analysis.** To ensure normality, high density lipoprotein (HDL), total cholesterol (TC), and triglycerides (TG) were natural log-transformed; low density lipoprotein (LDL) was square root-transformed. While lipid medications were carefully evaluated, the majority of the participants in IRASFS were metabolically healthy with a lipid medication rate of 3.9%. Therefore, only a single regression model was performed without accounting for lipid medication status. Heritability of lipid levels was estimated using Sequential Oligogenic Linkage Analysis Routines (SOLAR)<sup>37</sup> adjusting for age, sex, body mass index (BMI), and recruitment center. Variants with Mendelian inconsistencies were removed ( $n = 4,024$ ), resulting in a final number of 548,889 variants<sup>38</sup>. Each variant was coded to an additive model based on the minor allele (reference allele). Genetic models of association were calculated adjusting for age, sex, BMI, recruitment center, and admixture estimates. Admixture estimates were calculated as described previously<sup>39</sup> using maximum likelihood estimation of individual ancestries as implemented in ADMIXTURE<sup>40</sup>. Tests of association between individual variants and quantitative traits were computed using the Wald test from the variance component model implemented in SOLAR. Burden tests were computed using famSKAT<sup>41</sup>. Gene units were defined using the UCSC gene definition file from NCBI genome build 37 (hg19) whereby each alternatively spliced transcript is included for a total of 39,173 genes. For family-based linkage analysis, variant-specific identity-by-descent (IBD) probabilities were computed using the Monte Carlo method implemented in SOLAR<sup>42</sup>. Two-point linkage was performed using the variance components method implemented in SOLAR, with adjustment of age, gender, BMI, and recruitment center<sup>42</sup>. Variant annotation was performed using ANNOVAR<sup>43</sup>.

**Replication and Meta-analysis.** Six cohorts participating in the Genetics Underlying Diabetes in Hispanics (GUARDIAN) Consortium<sup>44</sup> provided in silico replication data: the Insulin Resistance Atherosclerosis Study (IRAS<sup>45</sup>), BetaGene<sup>46–49</sup>, the Troglitazone in Prevention of Diabetes Study (TRIPOD<sup>50,51</sup>), the Hypertension-Insulin Resistance Family Study (HTN-IR<sup>52,53</sup>), the Mexican-American Coronary Artery Disease Study (MACAD<sup>54–56</sup>) and the NIDDM-Atherosclerosis Study (NIDDM-Athero<sup>57</sup>). All study protocols were approved by the local institutional review committees and all participants gave their informed consent. GWAS genotyping was supported through the GUARDIAN Consortium<sup>44</sup> using the Illumina OmniExpress array (Illumina Inc.; San Diego, CA, USA) and imputation was performed centrally using IMPUTE2<sup>58</sup> and the 1000 Genomes phase I integrated reference panel (March 2012). Variants included in analysis had confidence scores  $> 0.90$  and information scores  $> 0.50$ . A detailed description of quality control has been described previously<sup>39</sup>.

A total of 56 nominally associated SNPs ( $P < 5.00 \times 10^{-05}$ ) with minor allele frequency (MAF)  $\geq 1.0\%$  as well as two rare SNPs that reached exome-wide significance ( $P < 9.11 \times 10^{-08}$ , MAF  $< 1.0\%$ ) were selected for replication in the GUARDIAN Consortium. Meta-analysis was performed among IRASFS ( $n_{\max} = 1,205$ ), IRAS ( $n_{\max} = 181$ ), BetaGene ( $n_{\max} = 1,218$ ), TRIPOD ( $n_{\max} = 125$ ), HTN-IR ( $n_{\max} = 763$ ), MACAD ( $n_{\max} = 749$ ), and NIDDM-Athero ( $n_{\max} = 244$ ) using the 1000 Genomes imputation dataset. Overall, 49 SNPs were directly tagged in replication cohorts and four SNPs were tagged by a proxy SNP ( $r^2 > 0.6$ ). However, no available proxies were found for the remaining five SNPs, which were excluded, resulting in a total of 53 SNPs in the meta-analysis. The meta-analysis was computed using METAL (<http://csg.sph.umich.edu/abecasis/metal>). Considering the differential study designs, a weighted meta-analysis of the p-values and samples sizes accounting for direction of effect was performed.

**Previously identified signals.** Previously identified lipid loci ( $N = 157$ ) from the recent GLGC<sup>4</sup> were extracted and all exome sequencing variants within  $\pm 100$  kb of the reported index SNPs were selected for fine-mapping. Conditional association and linkage analyses were performed with the GLGC index SNP as an adjusting covariate.

**ENCODE RNA sequencing data.** Four human primary hepatocytes RNA sequencing results were plotted using the UCSC genome browser<sup>24,26</sup>. The four liver biopsy samples included were derived from four European individuals: GSM2072386 (20-week female), GSM2072387 (22-week male), GSM2072372 (32-year male), and GSM2072373 (6-year female).

## References

- Libby, P. Inflammation in atherosclerosis. *Nature* **420**, 868–874, <https://doi.org/10.1038/nature01323> (2002).
- Mozaffarian, D. *et al.* Heart Disease and Stroke Statistics–2016 Update: A Report From the American Heart Association. *Circulation* **133**, e38–60, <https://doi.org/10.1161/CIR.0000000000000350> (2016).
- Knoblauch, H. *et al.* Heritability analysis of lipids and three gene loci in twins link the macrophage scavenger receptor to HDL cholesterol concentrations. *Arterioscler Thromb Vasc Biol* **17**, 2054–2060 (1997).
- Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283, <https://doi.org/10.1038/ng.2797> (2013).
- Go, A. S. *et al.* Heart disease and stroke statistics–2013 update: a report from the American Heart Association. *Circulation* **127**, e6–e245, <https://doi.org/10.1161/CIR.0b013e31828124ad> (2013).
- CDC. Deaths, percent of total deaths, and death rates for the 15 leading causes of death in 10-year age groups, by race and sex: United States, 2013. (2013).
- Below, J. E. *et al.* Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. *Sci Rep* **6**, 19429, <https://doi.org/10.1038/srep19429> (2016).

8. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161–169 (2008).
9. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet* **47**, 589–597, <https://doi.org/10.1038/ng.3300> (2015).
10. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753, <https://doi.org/10.1038/nature08494> (2009).
11. Choquet, H. & Meyre, D. Genetics of Obesity: What have we Learned? *Curr Genomics* **12**, 169–179, <https://doi.org/10.2174/138920211795677895> (2011).
12. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30–35, <https://doi.org/10.1038/ng.499> (2010).
13. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264–1272, <https://doi.org/10.1056/NEJMoa054013> (2006).
14. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106, <https://doi.org/10.1038/nature13917> (2015).
15. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**, 3143–3421 (2002).
16. Miller, M. *et al.* Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. *Circulation* **123**, 2292–2333, <https://doi.org/10.1161/CIR.0b013e3182160726> (2011).
17. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713, <https://doi.org/10.1038/nature09270> (2010).
18. Parra, E. J. *et al.* Admixture mapping in two Mexican samples identifies significant associations of locus ancestry with triglyceride levels in the BUD13/ZNF259/APOA5 region and fine mapping points to rs964184 as the main driver of the association signal. *PLoS One* **12**, e0172880, <https://doi.org/10.1371/journal.pone.0172880> (2017).
19. Zhou, L. *et al.* A genome wide association study identifies common variants associated with lipid levels in the Chinese population. *PLoS One* **8**, e82420, <https://doi.org/10.1371/journal.pone.0082420> (2013).
20. Tan, A. *et al.* A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population. *Hum Mol Genet* **21**, 1658–1664, <https://doi.org/10.1093/hmg/ddr587> (2012).
21. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269–276, <https://doi.org/10.1038/ng.1073> (2012).
22. Costanza, M. C., Beer-Borst, S., James, R. W., Gaspoz, J. M. & Morabia, A. Consistency between cross-sectional and longitudinal SNP: blood lipid associations. *Eur J Epidemiol* **27**, 131–138, <https://doi.org/10.1007/s10654-012-9670-1> (2012).
23. Ken-Dror, G., Goldbourt, U. & Dankner, R. Different effects of apolipoprotein A5 SNPs and haplotypes on triglyceride concentration in three ethnic origins. *J Hum Genet* **55**, 300–307, <https://doi.org/10.1038/jhg.2010.27> (2010).
24. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, <https://doi.org/10.1038/nature11247> (2012).
25. Jeong, S. W., Chung, M., Park, S. J., Cho, S. B. & Hong, K. W. Genome-wide association study of metabolic syndrome in Koreans. *Genomics Inform* **12**, 187–194, <https://doi.org/10.5808/GI.2014.12.4.187> (2014).
26. Bashliev, I. Temporary loss of work capacity in myocardial infarct patients who underwent rehabilitation. *Vutr Boles* **26**, 45–50 (1987).
27. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* **13**, 307–308, <https://doi.org/10.1089/bio.2015.29031.hmm> (2015).
28. Hasumi, H. *et al.* Folliculin-interacting proteins Flnp1 and Flnp2 play critical roles in kidney tumor suppression in cooperation with Flcn. *Proc Natl Acad Sci USA* **112**, E1624–1631, <https://doi.org/10.1073/pnas.1419502112> (2015).
29. Linehan, W. M., Srinivasan, R. & Schmidt, L. S. The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol* **7**, 277–285, <https://doi.org/10.1038/nrurol.2010.47> (2010).
30. James, V. M. *et al.* Novel missense mutations in the glycine receptor beta subunit gene (GLRB) in startle disease. *Neurobiol Dis* **52**, 137–149, <https://doi.org/10.1016/j.nbd.2012.12.001> (2013).
31. Al-Owain, M. *et al.* Novel mutation in GLRB in a large family with hereditary hyperekplexia. *Clin Genet* **81**, 479–484, <https://doi.org/10.1111/j.1399-0004.2011.01661.x> (2012).
32. Humphries, A. D. *et al.* Dissection of the mitochondrial import and assembly pathway for human Tom40. *J Biol Chem* **280**, 11535–11543, <https://doi.org/10.1074/jbc.M413816200> (2005).
33. Salakhov, R. R. *et al.* TOMM40 gene polymorphism association with lipid profile. *Genetika* **50**, 222–229 (2014).
34. Henkin, L. *et al.* Genetic epidemiology of insulin resistance and visceral adiposity. The IRAS Family Study design and methods. *Ann Epidemiol* **13**, 211–217, [https://doi.org/10.1016/S1047-2797\(02\)00412-X](https://doi.org/10.1016/S1047-2797(02)00412-X) (2003).
35. Wing, M. R. *et al.* Analysis of FTO gene variants with obesity and glucose homeostasis measures in the multiethnic Insulin Resistance Atherosclerosis Study cohort. *Int J Obes (Lond)* **35**, 1173–1182, <https://doi.org/10.1038/ijo.2010.244> (2011).
36. Tabb, K. L. *et al.* Analysis of whole exome sequencing with cardiometabolic traits using family-based linkage and association in the IRAS Family Study. *Annals of Human Genetics*. <https://doi.org/10.1111/ahg.12184> (2016).
37. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**, 1198–1211, <https://doi.org/10.1086/301844> (1998).
38. O'Connell, J. R. & Weeks, D. E. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* **63**, 259–266 (1998).
39. Gao, C. *et al.* A Comprehensive Analysis of Common and Rare Variants to Identify Adiposity Loci in Hispanic Americans: The IRAS Family Study (IRASFS). *PLoS One* **10**, e0134649, <https://doi.org/10.1371/journal.pone.0134649> (2015).
40. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664, <https://doi.org/10.1101/gr.094052.109> (2009).
41. Chen, H., Meigs, J. B. & Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* **37**, 196–204, <https://doi.org/10.1002/gepi.21703> (2013).
42. Hellwege, J. N. *et al.* Genome-wide family-based linkage analysis of exome chip variants and cardiometabolic risk. *Genet Epidemiol* **38**, 345–352, <https://doi.org/10.1002/gepi.21801> (2014).
43. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, <https://doi.org/10.1093/nar/gkq603> (2010).
44. Goodarzi, M. O. *et al.* Insulin sensitivity and insulin clearance are heritable and have strong genetic correlation in Mexican Americans. *Obesity (Silver Spring)* **22**, 1157–1164, <https://doi.org/10.1002/oby.20639> (2014).
45. Wagenknecht, L. E. *et al.* The insulin resistance atherosclerosis study (IRAS) objectives, design, and recruitment results. *Ann Epidemiol* **5**, 464–472, [https://doi.org/10.1016/1047-2797\(95\)00062-3](https://doi.org/10.1016/1047-2797(95)00062-3) (1995).
46. Watanabe, R. M. *et al.* Transcription factor 7-like 2 (TCF7L2) is associated with gestational diabetes mellitus and interacts with adiposity to alter insulin secretion in Mexican Americans. *Diabetes* **56**, 1481–1485, <https://doi.org/10.2337/db06-1682> (2007).
47. Black, M. H. *et al.* Evidence of interaction between PPARG2 and HNF4A contributing to variation in insulin sensitivity in Mexican Americans. *Diabetes* **57**, 1048–1056, <https://doi.org/10.2337/db07-0848> (2008).

48. Li, X. *et al.* Variation in IGF2BP2 interacts with adiposity to alter insulin sensitivity in Mexican Americans. *Obesity (Silver Spring)* **17**, 729–736, <https://doi.org/10.1038/oby.2008.593> (2009).
49. Shu, Y. H. *et al.* Evidence for sex-specific associations between variation in acid phosphatase locus 1 (ACP1) and insulin sensitivity in Mexican-Americans. *J Clin Endocrinol Metab* **94**, 4094–4102, <https://doi.org/10.1210/jc.2008-2751> (2009).
50. Buchanan, T. A. *et al.* Preservation of pancreatic beta-cell function and prevention of type 2 diabetes by pharmacological treatment of insulin resistance in high-risk hispanic women. *Diabetes* **51**, 2796–2803 (2002).
51. Buchanan, T. A. *et al.* Response of pancreatic beta-cells to improved insulin sensitivity in women at high risk for type 2 diabetes. *Diabetes* **49**, 782–788 (2000).
52. Xiang, A. H. *et al.* Evidence for joint genetic control of insulin sensitivity and systolic blood pressure in hispanic families with a hypertensive proband. *Circulation* **103**, 78–83 (2001).
53. Cheng, L. S. *et al.* Coincident linkage of fasting plasma insulin and blood pressure to chromosome 7q in hypertensive hispanic families. *Circulation* **104**, 1255–1260 (2001).
54. Goodarzi, M. O. *et al.* Determination and use of haplotypes: ethnic comparison and association of the lipoprotein lipase gene and coronary artery disease in Mexican-Americans. *Genet Med* **5**, 322–327, <https://doi.org/10.1097/01.GIM.0000076971.55421.AD> (2003).
55. Goodarzi, M. O. *et al.* Lipoprotein lipase is a gene for insulin resistance in Mexican Americans. *Diabetes* **53**, 214–220 (2004).
56. Goodarzi, M. O. *et al.* Variation in the gene for muscle-specific AMP deaminase is associated with insulin clearance, a highly heritable trait. *Diabetes* **54**, 1222–1227 (2005).
57. Wang, Y.-P. *et al.* Insulin and blood pressure are linked to the LDL receptor-related protein locus on chromosome 12q (Abstract). *Diabetes* **49**(Supp 1), A204 (2000).
58. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, <https://doi.org/10.1371/journal.pgen.1000529> (2009).

## Acknowledgements

This research was jointly supported by HG007112 from the national Human Genome Research Institute (NHGRI) and DK097524 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Computational resources were provided, in part, by the Wake Forest School of Medicine Center for Public Health Genomics. The authors would like to acknowledge the members of the GUARDIAN Consortium with research supported by DK085175 from NIDDK and from the following grants: IRAS Classic (HL047887, HL047889, HL047890, and HL47902), IRAS Family Study (HL060944 and HL061019), BetaGene (DK061628), MACAD (HL088457), HTN-IR (HL069794), and NIDDM (HL055798). The authors thank the other investigators, the staff, and the participants of the studies for their valuable contributions. The provision of genotyping data was supported in part by UL1TR000124 (CTSI), DK063491 (DRC), DK081350, HG007112 and DK087914.

## Author Contributions

C.G. researched the data and wrote the manuscript; K.L.T. and N.D.P. researched the data and reviewed/edited the manuscript; L.M.D., K.D.T., N.W., X.G., J.L., J.C. performed analysis; J.I.R., R.M.W., J.B., C.D.L., D.W.B. contributed to discussion and reviewed/edited the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-23727-2>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018