# Original Research

# Molecular Characterization of the Lipid Genome-Wide Association Study Signal on Chromosome 18q11.2 Implicates HNF4A-Mediated Regulation of the *TMEM241* Gene

Alejandra Rodríguez, Luis Gonzalez, Arthur Ko, Marcus Alvarez, Zong Miao, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas, Päivi Pajukanta

***Objective***—We recently identified a locus on chromosome 18q11.2 for high serum triglycerides in Mexicans. We hypothesize that the lead genome-wide association study single-nucleotide polymorphism rs9949617, or its linkage disequilibrium proxies, regulates 1 of the 5 genes in the triglyceride-associated region.

***Approach and Results***—We performed a linkage disequilibrium analysis and found 9 additional variants in linkage disequilibrium ($r^2>0.7$) with the lead single-nucleotide polymorphism. To prioritize the variants for functional analyses, we annotated the 10 variants using DNase I hypersensitive sites, transcription factor and chromatin states and identified rs17259126 as the lead candidate variant for functional in vitro validation. Using luciferase transcriptional reporter assay in liver HepG2 cells, we found that the G allele exhibits a significantly lower effect on transcription ($P<0.05$). The electrophoretic mobility shift and ChIPqPCR assays confirmed that the minor G allele of rs17259126 disrupts an hepatocyte nuclear factor 4 α–binding site. To find the regional candidate gene, we performed a local expression quantitative trait locus analysis and found that rs17259126 and its linkage disequilibrium proxies alter expression of the regional transmembrane protein 241 (*TMEM241*) gene in 795 adipose RNAs from the METSIM cohort ($P=6.11\times10^{-07}$–$5.80\times10^{-04}$). These results were replicated in expression profiles of *TMEM241* from the MuTHER resource (n=856).

***Conclusions***—The Mexican genome-wide association study signal for high serum triglycerides on chromosome 18q11.2 harbors a regulatory single-nucleotide polymorphism, rs17259126, which disrupts normal hepatocyte nuclear factor 4 α binding and decreases the expression of the regional TMEM241 gene. Our data suggest that decreased transcript levels of *TMEM241* contribute to increased triglyceride levels in Mexicans. (***Arterioscler Thromb Vasc Biol***. 2016;36:00-00. DOI: 10.1161/ATVBAHA.116.307182.)

**Key Words:** cell line ■ chromatin ■ dyslipidemias ■ genome-wide association study ■ triglycerides

Serum triglyceride levels are heritable and environmentally modifiable risk factor for cardiovascular disease.[1] Several groups have successfully used genome-wide association studies (GWAS) to identify signals for triglycerides and other lipid traits, including high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, and total cholesterol.[2] However, the lead GWAS signals may not themselves be functional rather in linkage disequilibrium (LD) with the actual underlying susceptibility variant. This limitation in GWAS derives from the fact that the human genome is only relatively superficially screened in GWAS using common tag single-nucleotide polymorphisms (SNPs). Furthermore, the functional variant often acts through a regional gene. Therefore, GWASs are only a starting point and require subsequent fine mapping and functional validation studies to identify the actual susceptibility variants and genes.

According to a recent survey, both the US Hispanic men and women have higher levels of serum triglycerides than non-Hispanic whites or blacks,[3] a result consistently reported for the past 2 decades.[4] Recent studies using Latino cohorts have successfully narrowed European lipid loci.[5] Moreover, because of the higher incidence of metabolic disease in the Amerindian origin populations, the investigation of their admixed genomes provides an opportunity to

| Nonstandard Abbreviations and Acronyms | |
| --- | --- |
| *cis*-eQTL | *cis*-expression quantitative trait locus |
| ENCODE | encyclopedia of DNA elements |
| GWAS | genome-wide association study |
| HNF4A | hepatocyte nuclear factor 4 α |
| LD | linkage disequilibrium |
| SNP | single-nucleotide polymorphism |
| TFBS | transcription factor–binding sites |
| TMEM241 | transmembrane protein 241 |

identify Amerindian-specific susceptibility variants for complex cardiovascular traits.[6] Despite their high predisposition to dyslipidemias, Hispanics remain underinvestigated as the discovery study stage in genomic cardiovascular studies. Previously, we identified a locus on chromosome 18q11.2 associated with high serum triglycerides in Mexicans using GWAS.[5] However, similar to other GWAS, the functional variants and the underlying gene(s) through which these variants exert their effects in the triglyceride phenotype remain to be elucidated. To find the actual functional risk variant(s), we systematically annotated the SNPs in the triglyceride-associated LD block with chromatin state marks and transcription factor–binding events which nominated rs17259126 as the top candidate functional variant. Its genomic landscape harbors regulatory sites and is predicted to disrupt an hepatocyte nuclear factor 4 α (HNF4A)–binding site. We show that the G allele of rs7259126 reduces expression of the luciferase reporter gene in a human liver cell line. Consistent with this result, the mobility shift and ChIPqPCR assays confirmed that the same allele disrupts an HNF4A-binding site. Replicated *cis*-expression quantitative trait locus (*cis*-eQTL) analyses also implicate the minor G allele of rs17259126 for reduced expression of transmembrane protein 241 (*TMEM241*), suggesting *TMEM241* as the regional candidate gene. Taken together, we found that the triglyceride locus on chromosome 18q11.2 harbors at least one functional variant, rs17259126, associated with a decreased expression of the regional *TMEM241* gene, a novel gene for triglycerides in the rapidly growing Hispanic population with a high predisposition to dyslipidemias.

## Materials and Methods

Materials and Methods are available in the online-only Data Supplement.

## Results

### Pairwise LD Analysis to Identify LD Proxies

In our original GWAS,[5] conditional association analyses at the top 12 genotyped loci did not reveal additional independent SNPs with $P \leq 2.5 \times 10^{-3}$. To identify the full set of variants in LD with the lead GWAS SNP rs9949617, we first performed a regional LD analysis in the triglyceride-associated LD block. The LD block was determined in our previous study as the region spanning SNPs in LD of $r^2 \geq 0.5$ with the lead SNP rs9949617.[5] For the LD analysis, we used our genotyped and imputed GWAS data,[5] and we also verified using the 1000

Genomes Project data that no additional SNP(s) inside or outside this LD block (±500 kb from the block borders) have emerged to be in LD with the lead SNP rs9949617 since our previous study.[5] We found 3 genotyped and 6 imputed SNPs in LD ($r^2 \geq 0.7$) with the lead SNP rs9949617 (Table). Two of these 10 SNPs (rs9949617 and rs4800467) were genotyped in stages 1 and 2 of our original GWAS scan,[5] both resulting in $P$ values $<5 \times 10^{-8}$. Because any of these 10 SNPs in LD can be the functional variant underlying the triglyceride association on chromosome 18q11.2, we first performed functional annotation followed by hypothesis-driven functional assays to uncover the functional variant in the triglyceride-associated LD block. We also tested the candidate variant and its LD proxies for regional effects on gene expression among the 5 genes in the triglyceride-associated LD block using a *cis*-eQTL analysis to investigate whether the variant changes expression of a particular regional gene.

### Functional Genomics Analysis Using Encyclopedia of DNA Elements Data

*Cis*-eQTL variants often reside in regulatory elements such as transcription factor–binding sites (TFBS) and interrupt transcription factor occupancy, leading to transcriptional changes. However, functional variants may also act through multiple other mechanisms making functional validation studies challenging. To facilitate the identification of suitable functional assays, we used the encyclopedia of DNA elements (ENCODE) data sets to give biological interpretation to the variants, and based on their predicted functionality, we conducted hypothesis-driven functional assays. TFBS often coincide with regions of open chromatin; hence, we annotated the chromatin state using ENCODE DNase I hypersensitive sites and histone marks in disease-relevant cell lines and control cell lines. In addition to the ENCODE biochemical annotations, we looked for transcription factor motif disruptions using HaploReg. We hypothesized that variants with the greatest amount of regulatory evidence from experimental data sets and bioinformatic predictions are more likely to be functional. Using this approach, we screened all 10 SNPs (the lead SNP and its 9 LD proxies) and selected rs17259126 as a top candidate for functional validation because it resides in a TFBS and a likely regulatory element defined by the co-occurrence of H3K27ac and H3K4me1. The G allele of rs17259126 is also predicted to disrupt a HNF4A regulatory motif (Figure I in the online-only Data Supplement). HNF4A is a known regulator of several metabolic genes.[7] On the basis of these annotations, we hypothesized that rs17259126 resides in a TFBS and regulates expression of one of the regional genes on chromosome 18q11.

### Functional Validation of Candidate Variants

We sought to validate our predicted functional variant rs17259126. We performed luciferase reporter assays using engineered vectors containing a 600-bp sequence around the SNP. At 48 hours post transfection of HepG2 cells, we found that the minor allele G displays a 1.5-fold decreased reporter expression ($P < 0.05$) compared with the major A allele in 3 biological replicates (Figure 1). These results are consistent with the observed direction of the *cis*-eQTL effect (β=−0.149;

**Table.  The lead SNP rs9949617 and its LD proxies are *cis*-eQTLs for the *TMEM241* gene**

| SNP | Minor Allele | $r^2$ | MAF (AMR) | MAF (EUR) | MAF (FIN) | *cis*-eQTL *P* value* (β) |
|---|---|---|---|---|---|---|
| rs9949617† | T | 1.00 | 0.34 | 0.17 | 0.14 | $5.96×10^{-06}$ (−0.129) |
| rs9962573† | T | 0.97 | 0.32 | 0.17 | 0.14 | $5.96×10^{-06}$ (−0.129) |
| rs4800467† | G | 0.92 | 0.35 | 0.2 | 0.18 | $6.11×10^{-07}$ (−0.131) |
| rs1276322† | G | 0.82 | 0.33 | 0.17 | 0.14 | $1.7×10^{-04}$ (−0.109) |
| rs17259126 | G | 0.77 | 0.25 | 0.06 | 0.08 | $1.1×10^{-04}$ (−0.149) |
| rs9954334 | G | 1.00 | 0.34 | 0.17 | 0.14 | $5.96×10^{-06}$ (−0.128) |
| rs67124903 | G | 0.97 | 0.34 | 0.17 | 0.14 | $5.96×10^{-06}$ (−0.128) |
| rs71360517 | A | 0.88 | 0.29 | 0.14 | 0.10 | $5.80×10^{-04}$ (−0.112) |
| rs77127070 | A | 0.74 | 0.22 | 0.03 | 0.04 | N/A |
| rs4800154 | A | 0.74 | 0.23 | 0.03 | 0.04 | N/A |

The regional LD analysis uncovered 3 genotyped and 6 imputed variants in LD ($r^2≥0.7$) with the lead SNP, rs9949617 in Mexicans. *cis*-eQTL indicates *cis*-expression quantitative trait locus; LD, linkage disequilibrium; MAF, minor allele frequency in the 1000 Genomes Project on the admixed American (AMR) individuals, European ancestry (EUR) individuals, and Finns (FIN); NA, not available; SNP, single-nucleotide polymorphism; and TMEM241, transmembrane protein 241.

*The cis-eQTL *P* values obtained in the analysis of the Finnish METSIM RNA-seq data (n=795) pass the Bonferroni correction for 50 tests (10 SNPs and 5 regional genes in the triglyceride-associated LD block; *P*<0.001). The β is shown for the minor allele.

†Genotyped SNPs.

Table; see below). Similar assays for the lead SNP rs9949617 and rs4800467 did not reveal significant expression changes in the luciferase assay.

To further investigate whether the variant disrupts an HNF4A motif, we performed EMSA assays using isolated HNF4A protein (Figure 2A) or HepG2 cell nuclear extracts (Figure 2B) and found evidence that HNF4A preferentially binds the major A allele of rs17259126 in 4 biological replicates. We also performed EMSA assays for the 9 other LD proxies variants. No allele-specific shifts were observed (Figure II in the online-only Data Supplement). Together, the luciferase (Figure 1) and EMSA (Figure 2) assays suggest that HNF4A may regulate expression of a target gene by directly binding to the rs17259126 regulatory site.

To confirm that HNF4A interacts with the variant site in HepG2 cells, we performed chromatin immunoprecipitation followed by qPCR targeting a 71-bp (site 1) or 151-bp (site 2) sequence surrounding rs17259126 (Figure 3). We found an average enrichment of 4.23 and 2.29 for the sequences, respectively, when compared with an unbound control site. Our functional studies provide converging evidence that the sequence underlying rs17259126 is an HNF4A-binding site and that the G minor allele significantly inhibits this interaction in vitro.
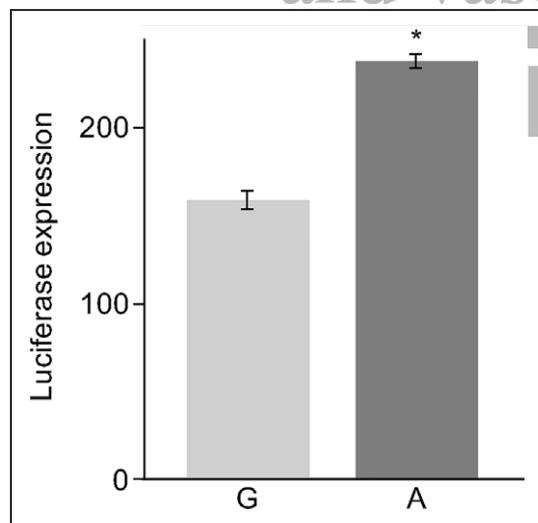
### *cis*-eQTL Analysis

GWAS variants residing in regulatory elements such as TFBS can lead to gene expression changes and contribute to disease susceptibility. We investigated whether the lead GWAS SNP may affect expression of the regional genes in the ≈300-kb region defining the triglyceride-associated window on chromosome 18q11.2 (LD $r^2$>0.5 with the lead SNP).[5] We performed a *cis*-eQTL analysis for the 5 genes within this triglyceride-associated LD block using adipose RNA-seq samples (n=795) from the METSIM cohort and discovered that the lead SNP rs9949617 (ie, the SNP with the strongest triglyceride association signal[4]) and its LD proxies are a *cis*-eQTL, regulating the expression of one regional gene, the *TMEM241* ($P=6.11×10^{-07}$–$5.80×10^{-04}$; Table). These results pass the Bonferroni correction for the 50 performed tests (10 SNPs tested for 5 regional genes; *P*<0.001; Table), and the 10 triglyceride-associated SNPs did not regulate expression of any of the 4 other genes within the LD block (Bonferroni corrected *P*>0.05).

To validate and replicate these regional *cis*-eQTL analysis results, we used expression data from 856 publicly available human adipose, skin, and lymphocyte RNA
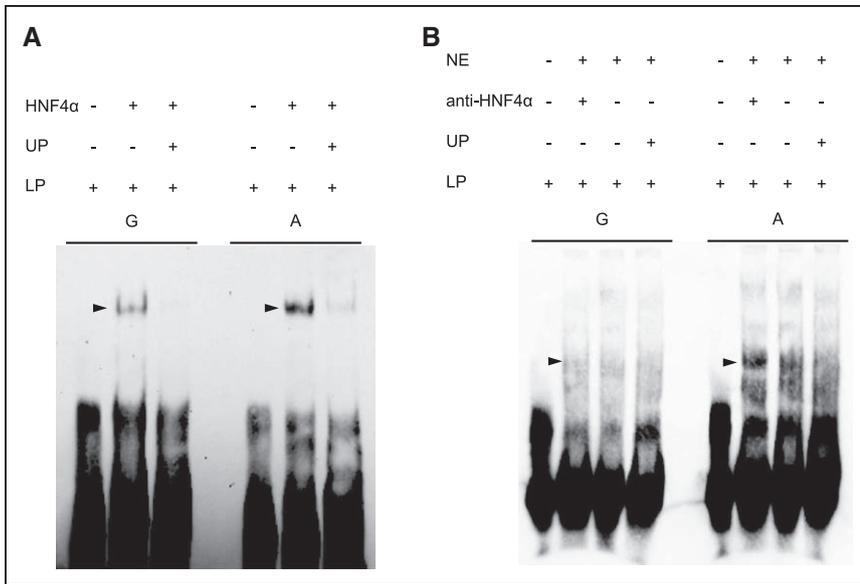


**Figure 1.** Luciferase expression of rs17259126 alleles in HepG2 cells. The reference allele A exhibits a 1.5-fold increased reporter expression ($P=2.97×10^{-4}$) when compared with the minor allele G. Luciferase assays were read at 48 hours post transfection. Statistical analysis was performed using the *t* test function in R. The Y axis represents the percent change in reporter expression relative to the empty pGL4.23[luc2/minP] vector. Error bars indicate the SE for the means of 3 independent biological replicates each done in triplicate.

**Figure 2.** As predicted by motif analysis (Figure I in the online-only Data Supplement), EMSA assays revealed that hepatocyte nuclear factor 4 α (HNF4A) recombinant protein directly interacts and displays a higher affinity for the reference A allele of the transmembrane protein 241 expression quantitative trait locus SNP rs17259126 (**A**), although no super shift was obtained when using HepG2 nuclear extract and anti-HNF4A (**B**). **A**, HNF4A isolate. **B**, HepG2 hepatoma nuclear extract. HNF4A (HNF4A isolated protein), unlabeled probe (UP), labeled probe (LP), nuclear extract (NE), and anti-HNF4A (anti-HNF4A). SNP indicates single-nucleotide polymorphism.

microarray samples from the MuTHER resource[8] and similarly discovered that the lead SNP rs9949617 is a *cis*-eQTL (Figure III in the online-only Data Supplement), regulating the expression of *TMEM241* ($P<1\times10^{-5}$ across all 3 tissues; $\beta=-0.107$ for adipose). These replication data are consistent, including the direction of the effect, with our *cis*-eQTL signal in Finns and our luciferase assays in which the minor G allele results in a decreased expression (Table and Figure 1). We also found comparable *cis*-eQTL results for the lead SNP rs9949617 in the HapMap3 data sets for the CEU ($P=0.0010$), CHB ($P=0.0019$), and JPT ($P=6.0\times10^{-4}$) samples in lymphoblastoid cells. Although there was a trend toward significance, this relationship did not hold for the MEX HapMap sample ($P=0.20$), perhaps because of the low number of Mexican-American samples (n=45) included in the HapMap project. These results implicate *TMEM241* as a likely regional gene underlying the GWAS association because the lead SNP and/or its LD proxies robustly regulate *TMEM241* expression through multiple cohorts. Taken together, these data suggest that rs17259126 is at least one of the functional SNPs underlying the original triglyceride GWAS signal[5] on chromosome 18q11.2 in Amerindian origin populations.

## Discussion

We recently identified a locus on chromosome 18q11.2 associated with high serum triglycerides in Mexicans using GWAS.[5] However, GWAS typically do not conclusively identify a functional regulatory variant and candidate gene, rather they require statistical and biochemical follow-up studies.[9,10] We used statistical fine mapping to first identify variants in the triglyceride-associated LD block. Because all variants represent 3′UTR or noncoding variants, we annotated their biological function using available regulatory data sets and bioinformatic tools and subsequently validated our recorded annotations using appropriate molecular assays.

Our LD analyses uncovered 9 variants in LD with the lead GWAS SNP rs9949617. Functional annotations using HaploReg[11] found that rs17259126 is predicted to disrupt an HNF4A-binding site, the minor G allele exhibiting a lower enrichment score. Furthermore, the ENCODE TF ChIP-seq data in HepG2 showed evidence of HNF4A enrichment
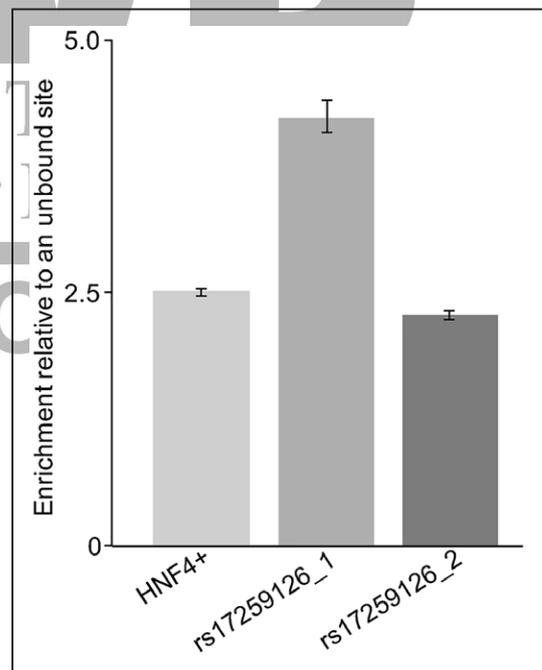


**Figure 3.** ChIP DNA analyzed by real-time polymerase chain reaction showed that hepatocyte nuclear factor 4 α (HNF4A) binds the transmembrane protein 241 expression quantitative trait locus SNP rs17259126 site in hepatoma cells. The y axis represents the enrichment of HNF4A at the rs17259126 site relative to an unbound control site. Error bars represent the SE for the means of 3 experimental replicates each done in triplicate. The region surrounding rs17259126 was targeted using 2 different pairs of oligos, indicated as rs17259126_1 and rs17259126_2. The sequences of the oligos are given in Table I in the online-only Data Supplement. SNP indicates single-nucleotide polymorphism.

around rs17259126. These findings prompted us to nominate rs17259126 as the lead candidate for molecular validation. We performed HNF4A ChIPqPCR targeting the SNP region and confirmed that HNF4A indeed binds the SNP site. HNF4A is a well-known, central regulator of hepatocyte development, differentiation, and gene expression[7,12] associated with type 2 diabetes mellitus, consistent with the triglyceride association. In line with our bioinformatics prediction, we also show that the G allele of rs17259126 reduces transcription of the luciferase reporter and significantly inhibits HNF4A binding in mobility shift assays. It is worth noting that Amerindian origin populations have >3-fold higher frequency of the minor allele G of rs17259126 when compared with Europeans (minor allele frequencies for AMR=0.22, EUR=0.06, AFR=0.08, and ASN=0.20, respectively).

To identify the regional gene, we performed cis-eQTL analyses using expression data from multiple cohorts, tissues, and platforms. We provide replicated evidence that the minor G allele of rs17259126 and its LD proxies are a robust *cis*-eQTL decreasing expression of the regional *TMEM241* gene across many cohorts. Our results suggest that HNF4A binds the A allele of rs17259126 site and increases expression of the *TMEM241* gene, 1 of the 5 regional genes in the LD block. We hypothesize that individuals with the G allele have decreased *TMEM241* expression which affects the normal triglyceride synthesis and/or secretory pathways through an unknown mechanism.

The *TMEM241* gene is a yeast VRG4 homolog, a Golgi-localized GDP-mannose transporter. Yeast VRG4 is pleiotropically required for a range of Golgi functions, including N-linked glycosylation, secretion, protein sorting, and the maintenance of a normal endomembrane system.[13,14] In the mammalian Golgi, carbohydrate processing is a highly diverse process. Carbohydrate chains may contain galactose, sialic acid, fucose, xylose, *N*-acetylglucosamine, and *N*-acetylgalactosamine unlike in the yeast *Saccharomyces cerevisiae*, where glycosylation is restricted to mannosylation. Thus, human *TMEM241* may function in the transport of other nucleotide sugars required in mammalian systems. In addition to glycoproteins, sphingolipids are also modified in the Golgi and have been implicated in metabolic disease.[15] *TMEM241* is believed to function as a nucleotide sugar transporter and, when defective, may lead to underglycosylation of glycoproteins and sphingolipids, potentially resulting in dysregulation of triglyceride synthesis.

Together, our results provide converging evidence suggesting rs17259126 as one of the functional variants underlying the GWAS association signal on 18q11.2,[5] and *TMEM241* as the underlying gene for triglycerides in Amerindian origin populations. However, because not all individuals of Mexican ancestry share the same composition of Amerindian DNA, additional cohorts may or may not replicate this particular association.

Future studies focusing on characterizing the role of *TMEM241* in triglyceride metabolism could include CRISPR/Cas9,[16] an emerging technology for targeted genomic modification. This technology allows a site-specific genetic engineering in disease-relevant cell lines to interrogate the function of specific genes and single-nucleotide variants in their native chromatin state. Elucidation of the role of *TMEM241* in triglyceride metabolism may help guide future research and development of new therapies for effective triglyceride management and prevention of heart disease in the rapidly growing Hispanic populations, currently underinvestigated in genomic cardiovascular studies despite their high predisposition to dyslipidemias.

## Disclosures

None.

## References

1. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45:1274–1283.
2. Go AS, Mozaffarian D, Roger VL, et al; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Executive summary: heart disease and stroke statistics–2014 update: a report from the American Heart Association. *Circulation*. 2014;129:399–410. doi: 10.1161/01.cir.0000442015.53336.12.
3. Aguilar-Salinas CA, Canizales-Quinteros S, Rojas-Martínez R, Mehta R, Villarreal-Molina MT, Arellano-Campos O, Riba L, Gómez-Pérez FJ, Tusié-Luna MT. Hypoalphalipoproteinemia in populations of Native American ancestry: an opportunity to assess the interaction of genes and the environment. *Curr Opin Lipidol*. 2009;20:92–97.
4. Aguilar-Salinas CA, Tusie-Luna T, Pajukanta P. Genetic and environmental determinants of the susceptibility of Amerindian derived populations for having hypertriglyceridemia. *Metabolism*. 2014;63:887–894. doi: 10.1016/j.metabol.2014.03.012.
5. Weissglas-Volkov D, Aguilar-Salinas CA, Nikkola E, et al. Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *J Med Genet*. 2013;50:298–308. doi: 10.1136/jmedgenet-2012-101461.
6. Ko A, Cantor RM, Weissglas-Volkov D, et al. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat Commun*. 2014;5:3983. doi: 10.1038/ncomms4983.
7. Hayhurst GP, Lee YH, Lambert G, Ward JM, Gonzalez FJ. Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. *Mol Cell Biol*. 2001;21:1393–1403. doi: 10.1128/MCB.21.4.1393-1403.2001.
8. Grundberg E, Small KS, Hedman ÅK, et al; Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*. 2012;44:1084–1089. doi: 10.1038/ng.2394.
9. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012;22:1748–1759. doi: 10.1101/gr.136127.111.

10. Lo KS, Vadlamudi S, Fogarty MP, Mohlke KL, Lettre G. Strategies to fine-map genetic associations with lipid levels by combining epigenomic annotations and liver-specific transcription profiles. *Genomics*. 2014;104:105–112. doi: 10.1016/j.ygeno.2014.04.006.

11. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930–D934. doi: 10.1093/nar/gkr917.

12. Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, Ning G, Kaestner KH, Rossi JM, Zaret KS, Duncan SA. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet*. 2003;34:292–296. doi: 10.1038/ng1175.

13. Hansen HG, Schmidt JD, Søltoft CL, Ramming T, Geertz-Hansen HM, Christensen B, Sørensen ES, Juncker AS, Appenzeller-Herzog C, Ellgaard L. Hyperactivity of the Ero1α oxidase elicits endoplasmic reticulum stress but no broad antioxidant response. *J Biol Chem*. 2012;287:39513–39523. doi: 10.1074/jbc.M112.405050.

14. Dean N, Zhang YB, Poster JB. The VRG4 gene is required for GDP-mannose transport into the lumen of the Golgi in the yeast, Saccharomyces cerevisiae. *J Biol Chem*. 1997;272:31908–31914.

15. Brice SE, Cowart LA. Sphingolipid metabolism and analysis in metabolic disease. *Adv Exp Med Biol*. 2011;721:1–17. doi: 10.1007/978-1-4614-0650-1_1.

16. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*. 2014;157:1262–1278. doi: 10.1016/j.cell.2014.05.010.

## Highlights

- The triglyceride locus on chromosome 18q11.2 harbors at least one functional variant, rs17259126, associated with a decreased expression of the regional *TMEM241* gene, a novel gene for triglycerides in the Hispanic population.
- HNF4A may regulate the expression of the TMEM241 gene by directly binding the rs17259126 regulatory site.
- Our findings suggest that decreased transcript levels of *TMEM241* contribute to increased triglyceride levels in Mexicans.

# Arteriosclerosis, Thrombosis, and Vascular Biology

**Molecular Characterization of the Lipid Genome-Wide Association Study Signal on Chromosome 18q11.2 Implicates HNF4A-Mediated Regulation of the *TMEM241* Gene**

Alejandra Rodríguez, Luis Gonzalez, Arthur Ko, Marcus Alvarez, Zong Miao, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas and Päivi Pajukanta

The online version of this article, along with updated information and services, is located on the World Wide Web at:

http://atvb.ahajournals.org/content/early/2016/05/19/ATVBAHA.116.307182

Data Supplement (unedited) at:

http://atvb.ahajournals.org/content/suppl/2016/05/19/ATVBAHA.116.307182.DC1.html
http://atvb.ahajournals.org/content/suppl/2016/05/19/ATVBAHA.116.307182.DC2.html

**Detailed Materials and Methods**
Materials and Methods are available in the online-only Data Supplement

**Study cohorts**
All Mexican participants were recruited at the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ), Mexico City, as described in detail previously[1]. The study design was approved by the local ethics committee and all participants gave an informed consent. Measurements of fasting lipid levels were performed with commercially available standardized methods[1].

A total of 795 participants from the Finnish METabolic Syndrome in Men (METSIM, total n=10,197)[2] were included in the regional *cis*-eQTL analysis of this study. The METSIM participants who underwent a subcutaneous abdominal adipose biopsy for RNA-seq were recruited at the University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. All 10,197 METSIM participants are male with a median age at examination 57 years (range: 45-74 years)[2]. The study design was approved by the local ethics committee and all participants gave an informed consent.

**Linkage disequilibrium analysis**
To search for variants in linkage disequilibrium (LD) ($r^2 \geq 0.7$) with the lead SNP rs9949617, we performed an LD analysis using PLINK[3] in the ~300-kb region on chr18q11.2 associated with TGs[1]. For the LD analysis, we used our genotyped and imputed GWAS data[1], and we also verified utilizing the 1000 Genomes Project data that no additional SNP(s) inside or outside this LD block (+/-500 kb from the block borders) has emerged to be in LD with the lead SNP rs9949617 since our previous study[1]. To maintain the strength of the association evidence while searching for a maximum number of variants in relatively tight LD with the lead SNP, we arbitrarily set the LD cut-point to $r^2=0.7$.

**Genotyping rs17259126 in 200 Mexican GWAS controls**
We genotyped the variant rs17259126 in 200 Mexican GWAS controls to validate the imputation results. This analysis confirmed that rs17259126 is in LD ($r^2=0.77$) with the lead GWAS SNP rs9949617. The genotype concordance between the genotyped and imputed genotypes of rs17259126 was 98.8%.

**Prioritizing functional variation through epigenomic annotation**
To prioritize the variants for experimental analysis, we filtered the variants through systematic data mining, including enrichment of signatures for regulatory elements at the variation site in disease relevant cell lines, including adipose and liver when available and utilizing ENCODE and Roadmap Epigenomics Project data[4-5]. The functional potential was evaluated based on whether the SNP resided in a region containing DNase I hypersensitive site, transcription factor binding site and by evaluating its chromatin state using and histone mark ChIP-seq data.

**Plasmid DNA constructs and luciferase assays**

To experimentally validate the lead candidate variants, rs9949617, rs4800467 and rs17259126, for enhancer activity, a sequence of 500-bp surrounding the SNPs was amplified from genomic DNA and cloned upstream of a minimal promoter in the pGL4.23[luc2/minP] vector (Promega). To obtain constructs with the alternative allele, we utilized the GeneArt® Site-Directed Mutagenesis PLUS System (Invitrogen A14604) following the manufacturer's protocol. All plasmid constructs were verified by sequencing using the RVprimer3 (Promega) and other appropriate custom designed sequencing primers.

Reporter constructs were transiently transfected using Lipofectamine 2000 (invitrogen 11668-027) into human liver hepatocellular carcinoma cells (HepG2) (ATCC HB-8065). Luciferase assays were read at 48 hrs post transfection in triplicates and the results were reproduced in three independent biological replicates each with three technical replicates. To normalize the activity of the pGL4.23[luc2/minP] engineered vector, we used the pGL4.74[hRluc/TK] internal control, which minimizes the experimental variability caused by differences in cell viability or transfection efficiency. All experiments were performed following the manufacturer's recommendations with minor modifications. Statistical analysis was performed using the R t test function. Expression values were normalized and expressed as a percent change from the un-engineered control pGL4.23[luc2/minP] vector.

## Mobility shift assays

To detect DNA-protein interactions, we performed electrophoretic mobility shift assays (EMSA) (Active Motif, 37341) for the 10 variants in the LD TG block. HepG2 cell extracts (Active motif, 36011) or purified HNF4A factor (OriGene, TP316588) were incubated with a biotin labeled DNA probe (IDT) containing the SNP of interest. Samples were then resolved by electrophoresis on a 6% DNA retardation gels (ThermoFisher, EC6365BOX) and transferred to a nylon membrane. The biotin end-labeled DNA probe was detected using streptavidin conjugated to horseradish peroxidase (HRP) and a chemiluminescent substrate (Active Motif, 37341).

## ChIPqPCR

HepG2 cells were grown to confluence and cross-linked using 1% formaldehyde. Chromatin was sheared by sonication to an average size of 500 bp and incubated with anti-HNF4A (PPMX, PP-H1415-00) antibodies overnight at 4C. Immunoprecipitated complexes were captured using magnetic protein A beads (ThermoFisher, 10001D). DNA was eluted from the beads and incubated over night at 65°C to reverse the cross-links. After purification (Active Motif, 58002), DNA was analyzed by qPCR with primers corresponding to the regions around the rs17259126 site. Sites corresponding to peaks for transcriptional repression mark, H3K27me3 or HNF4A peaks in HepG2 ChIP-seq from ENCODE/Broad promoter sites were included as negative and positive controls, respectively (suppl table I). For each primer pair, the average cycle threshold values of the triplicates (variability <0.2) were calculated. The relative DNA amount was calculated using the $\Delta\Delta CT$ method for relative quantification.

## METSIM RNA-seq

We isolated total RNA from abdominal subcutaneous adipose needle biopsy using Qiagen miRNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Polyadenylated mRNA was prepared using Illumina TruSeq RNA Sample Preparation Kit v2. Paired end, 50-bp reads were generated using the Illumina HiSeq 2000 platform. We used STAR[6] to align the reads to the hg19 reference genome, allowing up to 2 mismatches per read-pair. To quantify gene expression, we first cleaned the Gencode v19 annotation data using the following steps mimicking closely the GTEx RNA-seq analysis pipeline[7]: 1) use only genes annotated as "protein_coding" or "lincRNA"; 2) remove transcripts annotated as "retained_intron"; 3) collapse overlapping exons from all transcripts of a gene; and 4) remove exonic intervals belonging to more than 1 gene. This generated non-redundant exons from all transcripts for each gene. For transcript assembly, we employed Cufflinks[8] v2.2.1 using the -G option and --overhang-tolerance set to 0 in order to use only reads that fully overlap exons. To normalize gene expression, we analyzed genes with expression FPKM > 0 in more than 90% of individuals. FPKM values were rank transformed to a standard normal distribution for each gene. We corrected for technical and environmental confounders using PEER[9] with 40 factors. As an additional quality control step, we tested the RNA-seq data for genotype concordance between the DNA and RNA samples using the VerifyBamID[10], MixupMapper[11], and GATK best practice guidelines to call variants from RNA-seq data, and matched the data accordingly.

**METSIM genotypes and Imputation**
METSIM genotypes were generated from the Illumina OmniExpress and Illumina ExomeChip arrays. All imputation analyses in the study were performed first pre-phasing SNPs with SHAPEIT[12] v2.r727 and then performing imputations using IMPUTE2[13-15] v2.3.0 with default parameters. The 1000 Genomes phase 1 version 3 (March 2012 release) was utilized for both pre-phasing and imputation. We removed variants with genotype posterior probabilities less than 0.9, minor allele frequency less than 0.01, and Hardy-Weinberg Equilibrium p-value less than $1\times10^{-5}$.

**METSIM *cis*-eQTL analysis**
We used linear regression employing an additive model implemented in Matrix-eQTL[16] to map *cis*-eQTLs from genotype/imputation data and normalized gene expression in the chr18q11.2 region. The ten TG-associated SNPs in LD ($r^2 \geq 0.7$) were tested for *cis*-eQTL effects for the 5 genes within the TG-associated ~300-kb LD block. P-values passing the Bonferroni correction for 50 tests (10 SNPs tested for 5 genes; p<0.001) were considered as statistically significant.

**MuTHER *cis*-eQTL analysis**
To replicate the *cis*-eQTL found in METSIM, we performed a *cis*-eQTL analysis utilizing the Multiple Tissue Human Expression Resource (MuTHER) study, a publicly available microarray data set of 856 samples from the TwinsUK registry[17]. We analyzed differential gene expression in the 18q11.2 GWAS locus across three different tissues: human adipose, skin, and transformed lymphocytes.

## Supplemental References

1. Weissglas-Volkov D, Aguilar-Salinas CA, Deere K, et al. Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. J Med Genet. 2013;50:298-308.
2. Stancáková A, Civelek M, Saleem NK, et al. Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. Diabetes. 2012;61:1895-1902.
3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81:559-575.
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
5. Bernstein BE, Stamatoyannopouos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol*. 2010;28:1045-1048.
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
7. Melé M, Ferreira PG, Reverter F, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348:660-665.
8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511-115.
9. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7:500-507.
10. Jun G1, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. Am J Hum Genet. 2012;91:839-848.
11. Westra HJ1, Jansen RC, Fehrmann RS, te Meerman GJ, van Heel D, Wijmenga C, Franke L. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics. 2011;27:2104-2111.
12. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9:179-181.
13. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
14. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3. 2011;1:457-470.

15. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44:955-959.
16. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353-1358.
17. Grundberg E, Small KS, Hedman ÅK, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44:1084-1089.
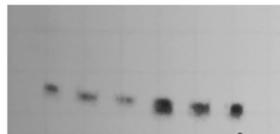
SUPPLEMENTAL MATERIAL
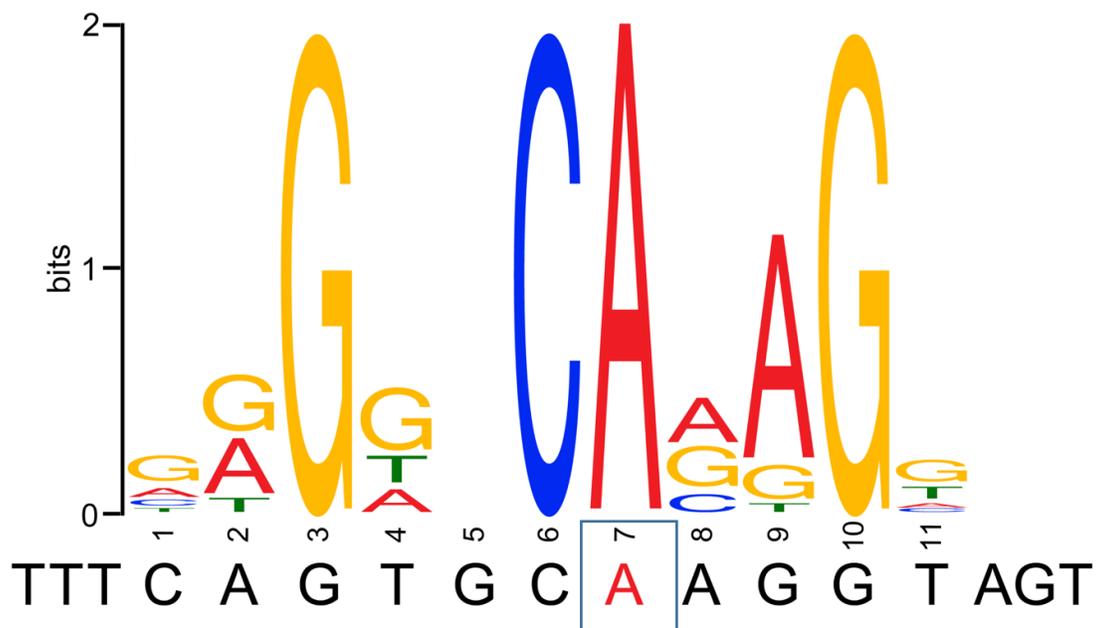
Supplemental Figure I. Motif discovery analyses using publicly available HNF4A ChIP-seq data sets (Kheradpour and Kellis 2013) show that the sequence underlying rs17259126 is an HNF4A binding motif (HNF4_disc3) and that the A major allele has a higher enrichment score (14 vs. 2) when compared to the G allele.
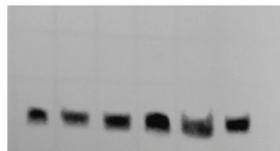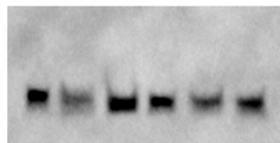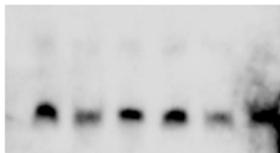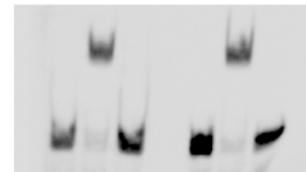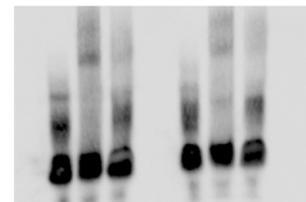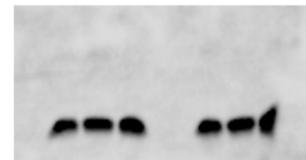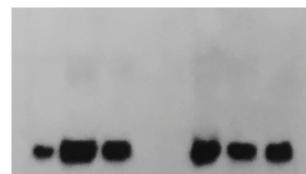
Supplemental Figure II. Representative images for EMSAs for the 9 SNPs in the TG-associated LD block that did not result in allele-specific shifts. Variants rs71360517 and rs4800154 display a TFB shift but no allele-specific differences were observed. No shift was observed for the remaining 7 SNPs (rs9949617, rs67124903, rs9962573, rs4800467, rs1276322, rs9954334, and rs77127070). Three different experimental conditions were tested for each allele: 1) LP (labeled probe); 2) LP and HepG2 nuclear extract (NE); and 3) LP, NE, and unlabeled probe (UP) to act as a competitor.
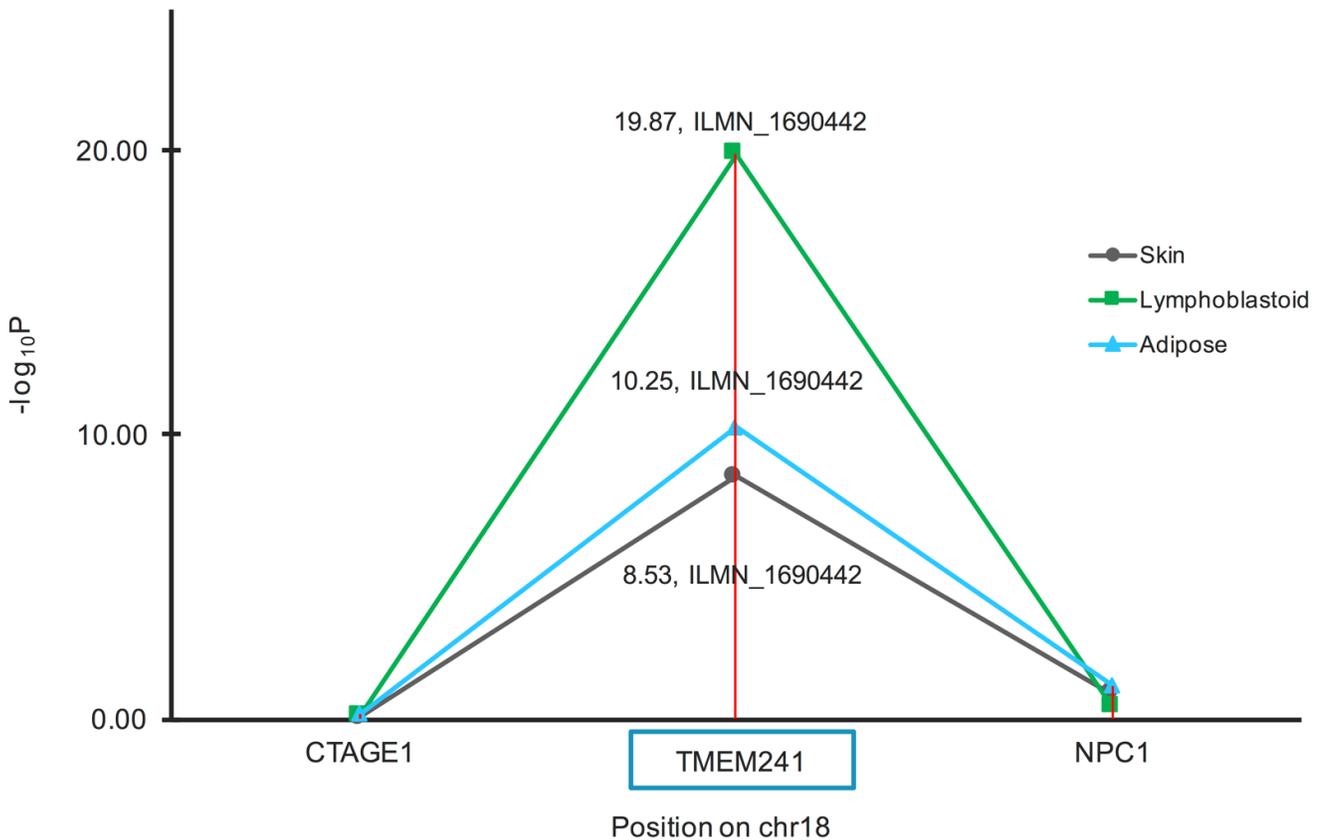
Supplemental Figure III. The lead TG-associated GWAS SNP rs9949617 is a cis-eQTL for one of the 5 regional genes, transmembrane protein 241 (TMEM241), on chromosome 18q11.2 ($p < 1 \times 10^{-5}$ for all 3 tissues). Expression analyses were performed for the regional genes in the ~300 kb region defining the TG-associated window (LD $r^2 > 0.5$ with the lead SNP rs9949617)[5] using experimental data from 856 publicly available human adipose, skin, and lymphocyte RNA microarray samples from the MuTHER resource[7].

Supplemental table I. Primers used for ChIPqPCR.

| site | sequence |
|---|---|
| rs17259126_1 | CCCCTTAGGCCCACTACC |
| | CTTCACACCTCTGGATGCCA |
| rs17259126_2 | TTAGGCCCACTACCTTGCAC |
| | ACCTCTCTCAACCGTGCTGT |
| H3k27me3-unbound site | GGACAAGAGAAGAGAGCACAG |
| | AGGACAGGTAGAAGCCCTAATA |
| | CTTCTACCTGTCCTGGAAACAG |
| | GAACCGTGGATGCTTAACAAAG |
| HNF4 bound site | GGCAAGACTCCTCTGAAAGGA |
| | TAGGACAGTAGTGGGGCATCA |