Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels

Mary F. Feitosa  
*Washington University School of Medicine in St. Louis*

Kai Ye  
*Washington University School of Medicine in St. Louis*

Ingrid B. Borecki  
*Washington University School of Medicine in St. Louis*

et al

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

Please let us know how this document benefits you.

**Recommended Citation**

[https://digitalcommons.wustl.edu/open_access_pubs/4234](https://digitalcommons.wustl.edu/open_access_pubs/4234)
Variants associated with blood lipid levels may be population-specific. To identify low-frequency variants associated with this phenotype, population-specific reference panels may be used. Here we impute nine large Dutch biobanks (~35,000 samples) with the population-specific reference panel created by the Genome of the Netherlands Project and perform association testing with blood lipid levels. We report the discovery of five novel associations at four loci (P value < 6.61 × 10^{-4}), including a rare missense variant in ABCA6 (rs77542162, p.Cys1359Arg, frequency 0.034), which is predicted to be deleterious. The frequency of this ABCA6 variant is 3.65-fold increased in the Dutch and its effect (β_{LDL-C} = 0.135, β_{TC} = 0.140) is estimated to be very similar to those observed for single variants in well-known lipid genes, such as LDLR.
Genome-wide association studies (GWAS) have identified a large number of loci associated with blood lipid levels and analysis suggest there are additional susceptibility loci that have not yet been discovered1–3. Despite the fact that rare functional variants are known to play a major role in lipid metabolism1–3, there has been limited success in finding such variants in population-based studies using next-generation sequencing. Even if the effect of these variants is expected to be larger than that of common variants, the sample size needed to detect these rare or low-frequency variants increases dramatically with variant rarity. As the frequency of rare variants may increase in certain populations because of drift and founder effects4, the power of searches for rare functional variants may improve by the use of reference sets specific to distinct populations. Such references allow for better quality imputation of rare variants especially those with increased frequency in the population of interest3,5,6. Previous studies have successfully detected rare variants by imputation into larger sets of individuals in isolated populations followed by association testing to detect variants associated with the trait of interest7–9.

Here we describe an imputation-based GWAS for circulating lipid levels using a custom-built reference panel for the Dutch population (Genome of the Netherlands, GoNL, http://www.nlgenome.nl/), in which the whole genomes of 250 parent–offspring trios were sequenced at \( \frac{B_1}{C_2} \) coverage5,6. Owing to the trio design, the phasing quality of the reference panel was better than that of the 1000 Genomes (1-kG) Phase 1 panel. In this study we show that using this population-specific reference panel we were able to identify five novel associations at four loci.

Results
Nine large Dutch epidemiological cohorts (comprising 36,000 samples in total) were imputed with the GoNL reference panel (~19.5 million single-nucleotide polymorphisms (SNPs)) on an identical protocol8,10. All cohorts conducted association analysis on the imputed variants assuming an additive genetic effect on high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglyceride (TG) levels (Methods, Supplementary Methods and Supplementary Table 1), and the results were meta-analysed. We used conditional analysis implemented in GCTA11 to identify variants associated independently with lipid levels.

Both rare (minor allele frequency (MAF) <0.01), low (0.01<MAF<0.05) and common variants (MAF>0.05) were associated with HDL-C (\( N = 60 \) variants), LDL-C (\( N = 142 \) variants), TC (\( N = 134 \) variants) and TG (\( N = 16 \) variants) in both known and novel loci (Methods, Supplementary Tables 2–5 and Supplementary Fig. 1). In Fig. 1 we compare the allele frequencies that reach genome-wide significance in the GCTA analysis (\( P \) value <5\( \times 10^{-8} \)) to those reported in refs 1,2 (Fig. 1). The majority of the known HDL-C (31 of 45, 68.9%), LDL-C (24 of 34, 70.6%), TC (33 of 48, 68.6%) and TG (13 of 30, 43.3%) loci described in ref. 1 replicated at a \( P \) value <3.18\( \times 10^{-4} \) (Bonferroni correction based on 157 variants;
Methods, Supplementary Figs 2 and 3 and Supplementary Tables 6–7). We also confirmed several of the HDL-C (6 of 27, 22.2%), LDL-C (7 of 21, 33.3%), TC (4 of 23, 17.4%) and TG (1 of 12, 8.3%) loci described in ref. 2 at a value threshold of 8.33 (Table 1, Methods and Supplementary Table 4). Analysis of the discovery cohorts showed that these new variants were independent from previously identified loci (Supplementary Table 10 and Supplementary Fig. 4). Of the five loci, three (rs149580368, rs77542162 and rs144984216) have an increased frequency in GoNL compared with 1-kG (Phase 1 integrated release v3, April 2012, all ancestries; Table 1), suggesting that there may have been genetic drift in the Dutch population for these loci. Yet, as each of these loci has a MAF > 0.005, we assumed that these alleles also segregate in other populations of European descent, such as the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Therefore, we set out replication in independent samples from the CHARGE cohorts using the 1-kG reference panel (Phase 1 integrated release v3, April 2012, all ancestries; Table 1). We were able to replicate five out of the six variants using the Bonferroni-corrected P value threshold of 8.33 × 10⁻³ (Table 2, Methods and Supplementary Table 11).

Of the replicated variants, rs77542162 is the most interesting variant. This missense variant is associated with both LDL-C and TC (Supplementary Figs 5 and 6) and is located on chromosome 17 within the ABCA6 gene (ATP-binding cassette, subfamily A (ABCC), member 6). The frequency of this variant is 1.31-fold higher in the discovery cohorts than in the replication cohorts and even 3.65-fold higher in the GoNL population than in the 1-kG population. This missense variant changes the amino acid cysteine into arginine at position 1359 (Cys1359Arg) and is predicted to be damaging for the structure and function of the protein by PolyPhen2 (ref. 12), MutationTaster13 and LRT14. The effect size of rs77542162 (βLDL-C = 0.135 and βTC = 0.140) is very similar to those observed for other single variants in well-known lipid genes, such as LDLR and CETP, as reported in ref. 1. The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters that transport various molecules across extra- and intracellular membranes. This protein is a member of the ABC1 subfamily, which is the only major ABC subfamily found exclusively in multicellular eukaryotes. ABCA6 is clustered with four other ABC1 family members on chromosome 17q24 and appears to play a role in macrophage lipid homeostasis.

One other replicated variant, rs149580368, is also enriched with a 1.92-fold increase in frequency in the Dutch population compared with the 1-kG population. This intergenic variant (Supplementary Fig. 7), without a significant cis-eQTL effect, is located between the protein-coding genes C17orf105 and MPP3 (membrane protein, patimoylated 3). Two replicated variants have similar frequencies in the GoNL and 1-kG reference sets: rs7452801 (Supplementary Fig. 8), an new intergenic variant with a high frequency (MAF = 0.355) that is located in a region previously identified1, and rs117162033 (Supplementary Fig. 9), an an intronic variant in the myosin F (MYOIF)-coding gene. C17orf15, MPP3 and MYOIF have no known impact on lipid levels. As the imputation quality of rs117162033 is lower than the other variants, we validated the imputation of this variant using the same approach as published in ref. 15. We compared in a random sample of 65 participants of the GoNL reference panel their GoNL-imputed genotypes and found that the concordance was 100% (all participants were correctly imputed). The association between TG and the intronic variant in the MYOIF gene is remarkable because of the low frequency of the variant. This confirms the conclusions as published before about the GoNL reference panel, that the trio-based phasing contributed significantly to the imputation quality of rare variants5.
In this current study, the GoNL reference panel was used for imputations of the discovery cohorts and the 1-kG reference panel for the imputation of the replication cohorts. Although it would be interesting to impute with a combined reference panel of both the GoNL data, the 1-kG data and other sequence data, this effort is ongoing.

This study shows that the imputation of a population-specific reference panel into large epidemiological cohorts can reveal both low-frequency and rare variants associated with blood lipid levels using classical association testing approaches. The three variants with increased frequency in the Dutch population as compared with the 1-kG population include a rare, predicted to be deleterious missense variant in ABCA6, which has increased frequency 3.63 times larger in the Dutch population. The effect of this variant is comparable to that of variants in the LDLR gene, a gene for which several population-based screening programmes have been initiated. Our findings suggest that next-generation-sequencing effort may yield clinically relevant findings. Our paper further shows that next-generation-sequencing efforts in specific homogeneous populations as the Dutch may yield clinically relevant findings worldwide.

Methods

Study descriptions. The descriptions of the including cohorts can be found in the Supplementary Methods. A written informed consent was obtained from all study participants for all cohorts and local ethical committees at participating institutions approved individual study protocols.

Study samples and phenotypes. A summary of the details of both the discovery and replication cohorts participating in this study can be found in Supplementary Tables 1 and 12.

Only samples of Dutch ancestry were used in the discovery cohorts; the samples in the replication cohorts are from various ancestry (see Supplementary Table 12). In all studies, except MESA Whites, all individuals who used lipid-lowering medication at the time the lipid levels were measured, were excluded. In MESA Whites, the total cholesterol values for individuals on lipid-lowering medication were divided by 0.8. In all studies except for LLS and PREVEND, the subjects were fasting when the lipid levels were measured. In LLS all samples were non-fasted, and in PREVEND 2.99% were non-fasted. The LDL-C levels were measured within the ERF, Croatia-Korcula, Croatia-Split, Croatia-Vis, FamHS and Lifelines cohorts, within the other cohorts the Friedewald equation was used to calculate the LDL-C levels.

The lipid measurements were adjusted for sex, age and age<sup>2</sup> in all cohorts. Various methods were used to account for family relationships: in ERF grammar, GenABEL version 1.7.6 (refs 17,18) was used; in the Croatia-Korcula, Croatia-Split, Croatia-Vis and Generation Scotland cohorts mmscore (GenABEL)<sup>23</sup> was used; and in LLS, qtpassoc was used. In CHS the clinic was used as extra covariate (in Lifelines PC1 and PC2, in FamHS the field centre, the genotyping array (Illunina 550 k, 610 k and 1 M), PCS only for TC and PCI only for LDL, in FHS the cohort (offspring and third generation) and PCs, in MESA Whites 2 PCs and study site, in NTR-NESDA PCs and chip effect, in ORCADES the genotyping array and PCI, PC1 and PC3, in PROSPER-Dutch only PCI and in both PROSPER-Scottish and PROSPER-Irish PCI-PC4.

Genotyping and imputations. Detailed information about genotyping and imputations per cohort can be found in the Supplementary Methods. In summary, all cohorts were genotyped using commercially available Affymetrix or Illumina genotyping arrays, or custom Perlegen arrays. Quality control was performed independently for each study. To facilitate meta-analysis, each replication cohort performed genotype imputation using IMPUTE<sup>21</sup> or Minimac<sup>20</sup> with reference to the GoNL project data for the discovery cohorts and with reference to the 1-kG project data for the replication cohorts.

GWAS in all discovery cohorts. All nine discovery cohorts ran separate the genome-wide association study for each of the four traits: HDL-C, LDL-C, TC and TG. Supplementary Table 13 shows the genomic control factor \( \lambda \) per trait per cohort and Supplementary Figs 10–13 show the \( \lambda \) per MAF bin per trait per cohort. We therefore used only the SNPs with a \( R^2 > 0.3 \), \( R^2 < 1.1 \) and expected minor allele count (expMAC = \( \eta^{2} \times \text{MAF} \times \text{R}^{2} \times \text{sample size} \) > 10. Most inflations are observed within the ERF study, especially in the lowest-frequency variants, which is probably caused by the family structure in this cohort.

Meta-analysis of discovery cohorts. The association results of all studies were combined and the s.e.-based weights were calculated using METAL<sup>22</sup>. This tool also applies genomic control by automatically correcting the test statistics to account for small amounts of population stratification or unaccounted relatedness. METAL also allows for heterogeneity. We used the following filters: \( 0.3 < \chi^{2} < 1.1 \) and expMAC > 10.

After meta-analyses of all available variants, we excluded the variants that are not present in at least six of the nine cohorts. We also excluded all variants that are labelled as being in the inaccessible genome, since the quality of those SNPs cannot be guaranteed<sup>22</sup>. The remaining variants per trait, see Supplementary Table 14, were used to create Manhattan plots and QQ plots, see Supplementary Figs 14 and 15. The meta-analysis resulted in 1,905 SNPs with a \( P \) value less than \( 10^{-8 \pm 4} \) for HDL-C, 2,626 SNPs for LDL-C, 3,133 SNPs for TC and 1,310 for TG.

Confirmation of known loci. Previously, Teslovich et al.<sup>1</sup> and Willer et al.<sup>2</sup> identified 157 loci associated with one of more of the lipids. Teslovich et al.<sup>1</sup> identified 47, 37, 52 and 32 loci to be associated with HDL-C, LDL-C, TC and TG, respectively. The positions of these loci were reported on human genome build 36; we therefore lifted these positions over to human genome build 37 and checked the association results after the meta-analysis of all the different traits. The effect size of these loci was reported in mg dl<sup>-1</sup>, whereas in this study we use mmol<sup>-1</sup>. We therefore multiplied the effect size for the loci associated with TG with 0.0259 and the other loci with 0.011. Supplementary Fig. 2 and Supplementary Table 6 show the comparison per trait of our meta-analysis of all discovery cohorts with the results of the meta-analysis in ref. 1. We did the same for the results described in ref. 2, see Supplementary Fig. 3 and Supplementary Table 7. The effect size of these loci could not be compared with our results, since trait residuals within each study participating in the meta-analysis in ref. 2 were adjusted for sex and age<sup>2</sup> and subsequently quantile normalized. Their GWAS was performed with the inverse normal transformed trait values.

Selection of independent variants. In order to select only associated variants that were independent of previous findings, we used the GCTA tool<sup>23</sup>. This tool performs a stepwise selection procedure to select multiple associated SNPs by a conditional and joint analysis approach using summary-level statistics from a meta-analysis and LD corrections between SNPs estimated from the GoNL reference panel, release 4. This analysis revealed 60 independent variants associated with HDL-C, 142 independent variants associated with LDL-C, 134 independent variants associated with TC and 16 independent variants associated with TG. By using this approach, we were able to identify additional independent variants in known loci. Figure 1 shows that we identified both common and rare variants and more rare variants compared with refs 1,2. There is an overlap between the genome-wide significant SNPs of the different traits, and also between the independent SNPs of the different traits, as shown in Supplementary Fig. 1.

Identification of potential novel variants. To identify potential novel variants, we first excluded all variants within 1 Mb of a known loci from refs 1,2. Since the number of loci associated with the four traits differ, we end up with 7,946,245 SNPs for HDL-C, 8,014,693 SNPs for LDL-C, 7,923,530 SNPs for TC and 7,468,790 SNPs for TG. For all traits we do find some genome-wide significant loci, see Supplementary Figs 16 and 17. We used the GCTA tool to select only those variants that are independently associated with the lipid trait. This analysis revealed two novel independent variants associated with HDL-C, one novel independent variant associated with LDL-C, two novel independent variants associated with TC and one novel independent variants associated with TG, see Supplementary Table 8 and Supplementary Fig. 18. We used PLINK to test whether these six variants are in LD with the known loci from refs 1,2. None of the six variants are in LD with known loci associated with the same trait on the same chromosome (\( R^2 < 0.14 \)).

Replication of potential novel variants. The six potential novel loci were replicated in 11 cohorts: CHS, Croatia-Korcula, Croatia-Split, Croatia-Vis, FamHS, FHS, Generation Scotland, MESA Whites, ORCADES, PROSPER-Scottish and PROSPER-Irish. The association results of all cohorts were combined and the s.e.-based weights were calculated using METAL<sup>22</sup>. The Bonferroni correction for multiple testing was \( 8.33 	imes 10^{-3} \). This resulted in the significant replication of five out of the six variants, see Supplementary Fig. 19 and Supplementary Table 11.

Conditional analysis. Within the discovery cohorts we performed a conditional analysis to see whether the novel loci identified are independent of the known loci from refs 1,2. Supplementary Table 10 shows the results within these cohorts with and without adjusting for the known loci for the trait in question, if available in the GoNL reference panel. Since the adjusted and unadjusted result are very similar, we conclude that the newly identified variants are independent of the known loci.

References


© 2015 Macmillan Publishers Limited. All rights reserved.


The Genome of the Netherlands Consortium


Additional information

Competing financial interests: M.P. served on the DSMB of a clinical trial funded by the manufacturer (Zoll LifeCore) and on the Yale Open Data Access Project funded by Johnson & Johnson. The remaining authors declare no competing financial interests.

Reprints and permission: This article is included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/