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

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1038/ncomms7065

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:15034781

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels


Varicetals 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 (rs77542162, P = 6.61 x 10^-4), including a rare missense variant in ABCA6 (rs177542162, 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 (βLDLC = 0.135, βTC = 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 discovered \(^1\)–\(^3\). Despite the fact that rare functional variants are known to play a major role in lipid metabolism \(^1\)–\(^3\), 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 effects \(^4\), 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 interest \(^5\)–\(^6\). 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 interest \(^7\)–\(^9\).

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 \(13/4 \) coverage\(^5\)\(^6\). 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 (\(\sim 19.5\) million single-nucleotide polymorphisms (SNPs)) on an identical protocol\(^5\)\(^6\). 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 GCTA\(^11\) 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;
We were able to replicate five out of the six variants using the consortium. Therefore, we set out replication in independent populations of Europe, Asia and Africa (Methods and Supplementary Table 9). Conditional analysis revealed the presence of linkage disequilibrium (LD) with previously described GWAS variants in well-known lipid genes, such as CETP and ABC1, member 6 (ABCA6) (Methods and Supplementary Table 11).

To identify novel loci associated with blood lipid levels, we selected from the list of variants identified by GCTA, those variants located more than 1 Mb away from previously identified loci. This resulted in six novel associations at five loci (Methods, Tables 1 and 2 and Supplementary Table 8). The five loci are not in linkage disequilibrium (LD) with previously described GWAS loci (Methods and Supplementary Table 9). Conditional analysis in 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 those for 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 \(\times\) 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 (ABCA1), 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), MutationTaster\(^{13}\) and LRT\(^{14}\). The effect size of rs77542162 (\(\beta_{\text{LDLC}} = 0.135\) and \(\beta_{\text{TC}} = 0.140\)) is very similar to those observed for other single-loci 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, palmitoylated 3). Two replicated variants have similar frequencies in the GoNL and 1-kG reference sets: rs4752801 (Supplementary Fig. 8), an new intergenic variant with a high frequency (MAF = 0.355) that is located in a region previously identified, and rs17162033 (Supplementary Fig. 9), 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 rs17162033 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 sequence and best-guess 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 variants.
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 low-frequency and rare variants associated with blood lipid levels, this effort is ongoing.

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 ancestries (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-fast. 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 age2 in all cohorts. Various methods were used to account for family relationships: in ERF dictionary, GenABEL version 1.7.6 (refs 17,18) was used; in the Croatia-Korcula, Croatia-Split, Croatia-Vis and Generation Scotland cohorts mscnscore (GenABEL)20 was used; and in LLS, q-tassoc was used. In CHS the clinic was used as extra covariate (in Lifelines PC1 and PC2, in FamHS the field center, the genotyping array (Illunina 550k, 610k and 1 M), PCs only for TC and PC1 only for LDL, in FHS the cohort (offspring and third generation) and PCs, in MESA Whites 2 PCs and study site, in NTR-NEFDA PCs and chip effect, in ORCADES the genotyping array and PC1, PC2 and PC3, in PROSPER-Dutch only PC1 and in both PROSPER-Scottish and PROSPER-Irish PC1-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 IMPUTE29 or Minimac20 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 can 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 λ per trait per cohort and Supplementary Figs 10–13 show the λ per MAF bin per trait per cohort. We therefore used only the SNPs with a R2>0.3, R2<1.1 and expected minor allele count (expMAC = \(2 \times MAF \times R^2\) 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 METAL21. 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<R<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 guaranteed22. 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 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.23 and Willer et al.24 identified 157 loci associated with one of more of the lipids. Teslovich et al. identified 47, 37, 52 and 32 loci to be associated with HDL-C, LDL-C, TC and TG, respectively. The locations of these loci were reported on human genome build 36; we therefore lifted these positions over to human genome build 57 and checked the association results after the meta of the different traits. The effect size of these loci was reported in mg dl−1, whereas in this study we use mmol−1. 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 resolution of the meta-analysis in ref. 1. We did the same for the replication cohorts 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 age2 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 tool21. 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 (R2<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 METAL21. The Bonferroni correction for multiple testing was 8.33 × 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 variants 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 results were similar, we conclude that the newly identified variants are independent of the known loci.

References


Acknowledgements

We especially thank all volunteers who participated in our study. This study made use of data generated by the ‘Genome of the Netherlands’ project, which is funded by the Netherlands Organization for Scientific Research (grant no. 184020107). The data were made available as a Rainbow Project of BBMRI-NL. Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.health-aging.nl), and the Wellcome Trust Clinical Research Facility in Edinburgh.

The CHARGE consortium (http://www.chargeconsortium.org) and the Wellcome Trust Clinical Research Facility are supported by the Department of Health (grant number R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071252, R01HL071258 and R01HL071259). We thank the participants of the MESA study, the Coordinating Center, MESA investigators and staff for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

Netherlands Twin Register (NTR) and Netherland Studies on Depression and Anxiety (NDS): Funding was obtained from the Netherlands Organization for Scientific Research (NWO) and MagW/ZonMW grants Middelgroen 911-09-032, Spinazoomerij 56-464-14492, Geestkracht programme of the Netherlands Organization for Health and Development (Zon-MW, grant number 10-000-1002), Center for Medical Studies (CSMB, NWO Genomics), NIBC/BioAssist/RK (2008.024), Biobanking and Bioresources Resources Research Infrastructure (BBMRI-NL, 184.021.007), VU University’s Institute for Health and Care Research (EMGO+) and Neuroscience Genome Project (NCGP). The MESA Heart study (MESA) was supported by contracts N01-HC-95159 through N01-HC-95169 and RR-024156 from the NHLBI. Funding for MESA SHARE genotyping was provided by NHLBI Contract N02-HL-64278. MESA Family is conducted and supported in collaboration with MESA investigators; support is provided by grants and contracts R01HL71051, R01HL71052, R01HL71250, R01HL71251, R01HL71252, R01HL71258 and R01HL71259. We thank the participants of the MESA study, the Coordinating Center, MESA investigators and staff for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.
The Genome of the Netherlands Consortium

Pieter B.T. Neerincx, Clara C. Elbers, Pier Francesco Palamara, Itsik Pe’er, Abdel Abdellahoui, Wigard P. Kloosterman, Mannis van Oven, Martijn Vermaat, Mingkun Li, Jeroen F.J. Laros, J.B.O., P.I.W.d.B. E.M.v.L. performed the meta-analysis. Biological association of loci and bioinformatics were carried out by E.M.v.L. and C.M.v.D.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

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

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: van Leeuwen, E. M. et al. Genome of the Netherlands population-specific imputations identify a ABCA6 variant associated with cholesterol levels. Nat. Commun. 6:6065 doi: 10.1038/ncomms7065 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are 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/

The Genome of the Netherlands Consortium

Pieter B.T. Neerincx, Clara C. Elbers, Pier Francesco Palamara, Itsik Pe’er, Abdel Abdellahoui, Wigard P. Kloosterman, Mannis van Oven, Martijn Vermaat, Mingkun Li, Jeroen F.J. Laros, J.B.O., P.I.W.d.B. E.M.v.L. performed the meta-analysis. Biological association of loci and bioinformatics were carried out by E.M.v.L. and C.M.v.D.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

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

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: van Leeuwen, E. M. et al. Genome of the Netherlands population-specific imputations identify a ABCA6 variant associated with cholesterol levels. Nat. Commun. 6:6065 doi: 10.1038/ncomms7065 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are 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/