

# WEB MATERIAL

## Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions

Paul S. de Vries et al.

### Contents

Web Appendix.....	2
Adjustment of LDL-C for statin use .....	2
Study-specific analyses: software .....	2
Quality control .....	2
Genetic main effect.....	3
Variance explained .....	3
Gene Prioritization using DEPICT.....	4
References .....	5
Web Figure 1 .....	6
Web Figure 2 .....	9
Web Figure 3 .....	14
Web Figure 4 .....	23
Web Figure 5 .....	33

## Web Appendix

### Adjustment of LDL-C for statin use

If information on statin-specific use was unavailable, LDL-C values were adjusted for use of unspecified lipid-lowering medication, but only when lipid measurements were performed after 1994. When LDL-C values were directly assayed, LDL-C values were adjusted for lipid-lowering medication use by dividing the original values by 0.7. When LDL-C values were not directly assayed, LDL-C values were adjusted by first dividing total cholesterol by 0.8, and then using the corrected total cholesterol value in the Friedewald equation.

### Study-specific analyses: software

To obtain robust estimates of covariance matrices and robust standard errors (1, 2), studies of unrelated subjects used either the sandwich R package or ProbABEL (3, 4). To account for relatedness in families, family studies used the generalized estimating equations (GEE) approach, treating each family as a cluster, with the geepack package in R (5), or a linear mixed effect model approach with a random polygenic component (for which the covariance matrix depends on the kinship matrix) with GenABEL or R (6).

### Quality control

Extensive quality control (QC) using the R package EasyQC was performed for all study-specific GWAS results (7). We performed QC at two levels. The “study-level” QC reviewed result files from each study individually, which includes checking the provided allele frequencies against the ancestry-specific 1000 Genomes reference panel and harmonizing marker names to ensure consistencies across studies. To exclude unstable study-specific results that reflect low minor allele count (MAC) within alcohol consumption exposure categories or low imputation quality measures, variants were excluded from study-level results if  $\min(\text{MAC}_{\text{exposed}}, \text{MAC}_{\text{unexposed}}) \times \text{imputation quality measure} < 20$ . Variants were further excluded if the imputation quality measure was  $< 0.5$  (8). The “meta-level” QC reviewed result files from each specific analysis across all studies and this included: 1) visually comparing summary statistics (mean, median, standard deviation, inter-quartile range, minimum, maximum) on all effect estimates, standard errors and *P*-values; and 2) examining SE-N and QQ plots to reveal any issues with trait transformation or other analytical problems (7). A SE-N plot examines the ratio of a study’s standard deviation with the median standard error on one axis with the square root of the

study's sample size on the other axis for deviations from the 45-degree line relative to the other studies in the meta-analysis.

After meta-analysis was performed, a variant was excluded if the sample size was below 5000 or if the number of studies with the variant measured was less than 3. Given the limited availability of data on individuals of Hispanic ancestry, we excluded variants from the Hispanic ancestry analysis if the sample size was below 3000 or the numbers of studies with the variant measured was less than 2.

### Genetic main effect

The study-specific analyses were not adjusted for alcohol intake or for the interaction term, and were performed only for the lipid trait that each variant was most significantly associated to in the main analysis. METAL was used to meta-analyse results across studies within each ancestry group, and subsequently to meta-analyse results across ancestry groups (9).

### Variance explained

The percent of variance explained in HDL-C, LDL-C, and TG by all previously known and novel variants was evaluated using participant-level data in ten studies from multiple ancestries. These analyses were performed stratified by ancestry group. The 314 variants (Web Table 8) previously identified for lipid traits in any ancestry were considered as “known” variants (10-16), while the index variants at the 18 novel loci were considered “novel” variants.

We calculated percent variance using a series of five nested standard linear regression models. In hierarchical order, these include Model 1, adjusting for study-specific covariates; Model 2, adjusting for Model 1 and additionally adjusting for alcohol exposure; Model 3, adjusting for Model 2 and additionally adjusting for known variants; Model 4 adjusting for Model 3 and additionally adjusting for novel variants; and finally Model 5, adjusting for Model 4 and additionally adjusting for gene-alcohol interaction terms of novel variants. To select a subset of variants and interaction terms to include in Models 3, 4, and 5, for each lipid trait, we performed a stepwise regression procedure with significance tests for inclusion of one variant at a time and for backward elimination of redundant variants. From one model to the next, we kept the significant variants from the previous model before adding the next nested model's variants or interaction terms. The  $r^2$  values obtained at each nested level were used as measures of the percent variance explained by the respective models. Through sequential subtraction of appropriate  $r^2$  values, we determined the additional percent variance explained

by a given set of variants. For example, for a given lipid trait, the variance explained by known variants was calculated as the Model 3  $r^2$  subtracted by the Model 2  $r^2$ , and the variance explained by novel variants was calculated as the Model 4  $r^2$  subtracted by the Model 3  $r^2$ .

Variance explained estimates obtained using this method may be overestimates due to bias arising from the use of backwards elimination and from the estimation of variance explained in samples that were also included in the genetic association study itself.

### Gene Prioritization using DEPICT

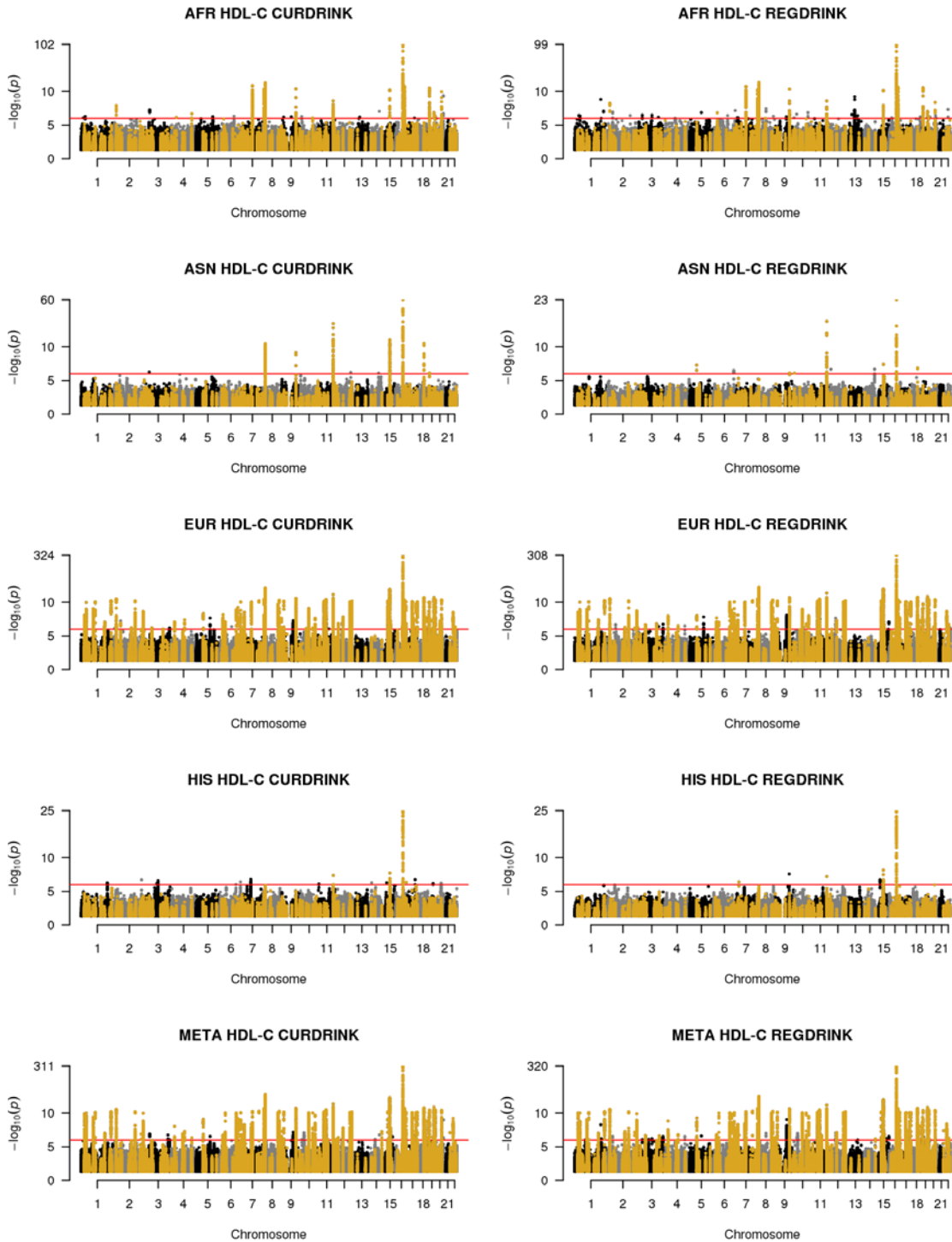
We used the Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) software to prioritize genes at the 147 loci associated in the combined analysis of Stage 1 and 2.

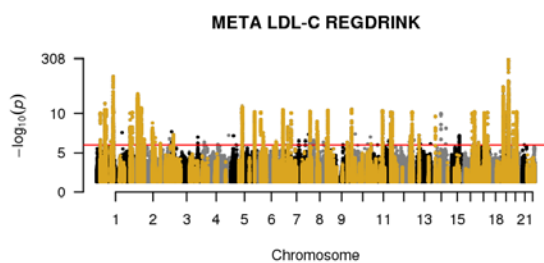
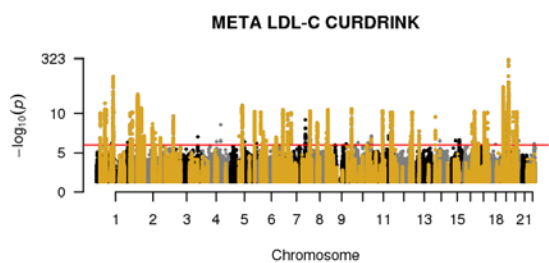
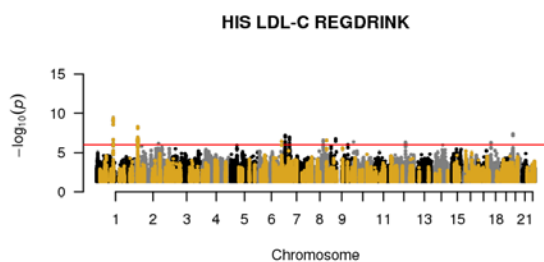
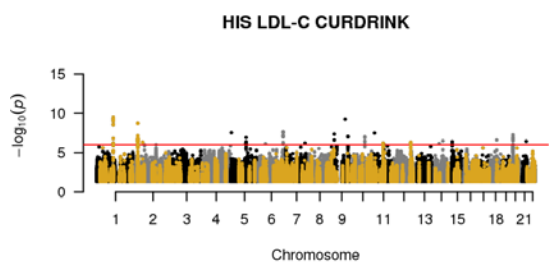
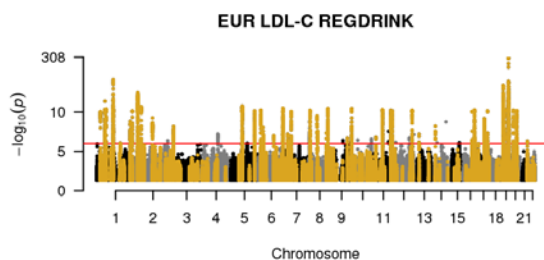
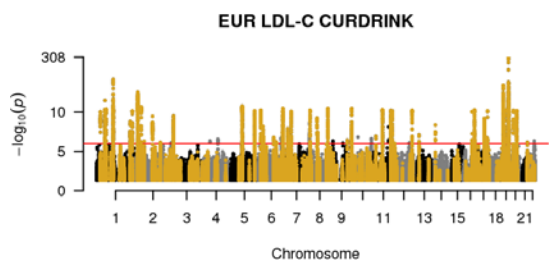
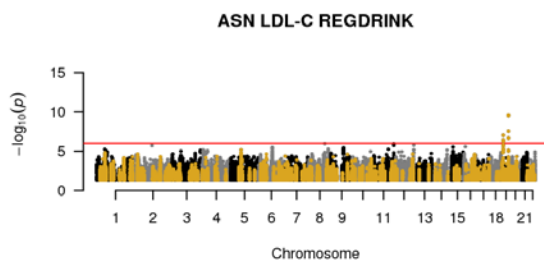
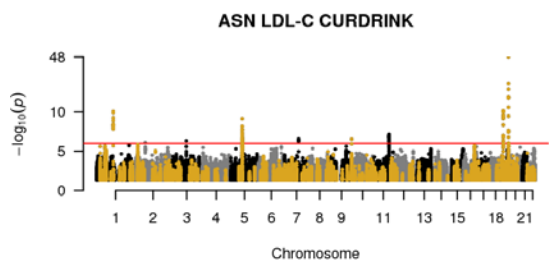
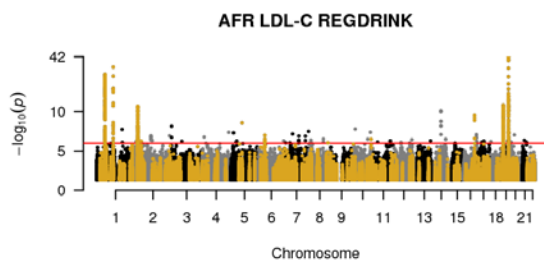
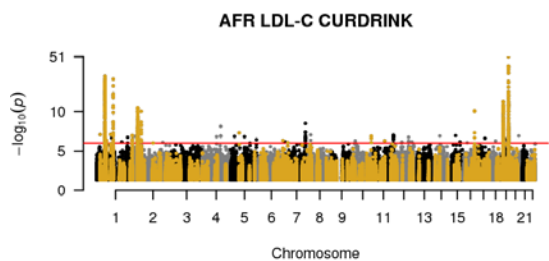
Separate DEPICT analyses were performed for variants associated at genome-wide significance with each of the three lipid traits (HDL-C, LDL-C and TG) in the combined analysis of Stages 1 and 2 (17). For each trait, we first created a subset of non-overlapping variants (>500 kb flanking regions and LD  $r^2 > 0.1$ ) between the associated variants and the 1000 Genomes reference data (18). Next, DEPICT obtained lists of overlapping genes by applying an LD based threshold ( $r^2 > 0.5$ ) between the non-overlapping variants and known functional coding or cis-acting regulatory variants (of the respective genes). The major histocompatibility complex region on chromosome 6 (base position 25,000,000 - 35,000,000) was removed for further analyses. Using DEPICT, we compared the functional similarity of overlapping genes across loci using a gene score that was adjusted for varying confounders, such as gene length. To derive an experiment-wide False-Discovery-Rate (FDR) for the gene prioritization, the scoring step was repeated 50 times utilizing index variants from 500 pre-compiled null GWAS. Our DEPICT gene-set enrichment analyses were based on a total of 14,461 pre-compiled reconstituted gene sets - including 737 Reactome database pathways (19), 2,473 phenotypic gene sets (derived from the Mouse Genetics Initiative (20)), 184 Kyoto Encyclopedia of Genes and Genomes (KEGG) database pathways (21), 5,083 Gene Ontology database terms (22), and 5,984 protein molecular pathways (derived from protein-protein interactions) (23). Our DEPICT tissue, cell type and physiological system specificity analyses utilize expression data in any of the 209 MeSH annotations for 37,427 microarrays of the Affymetrix U133 Plus 2.0 Array platform.

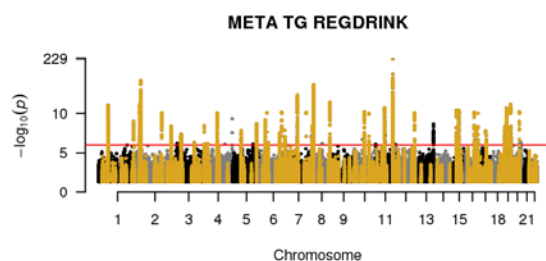
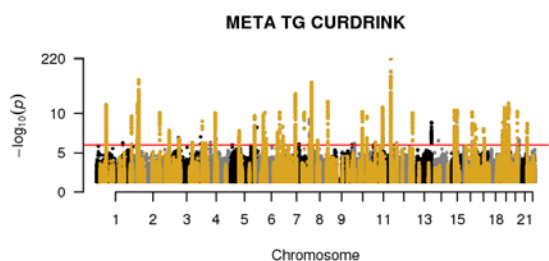
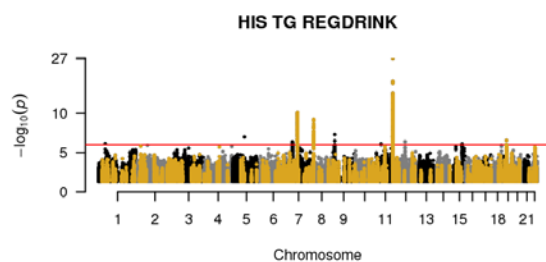
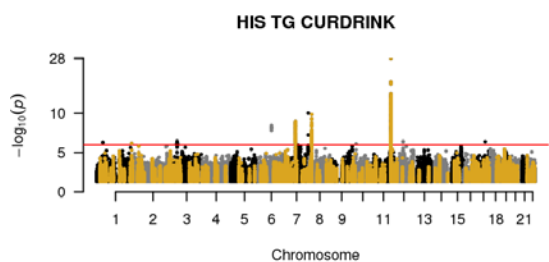
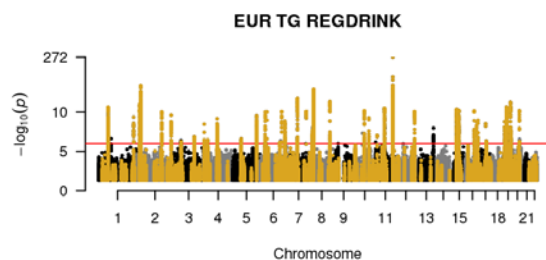
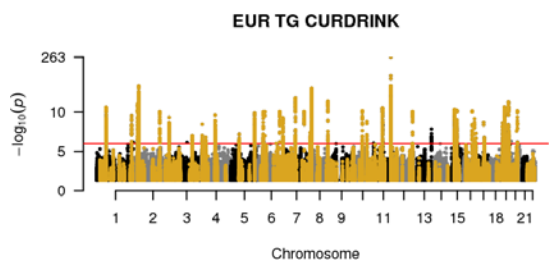
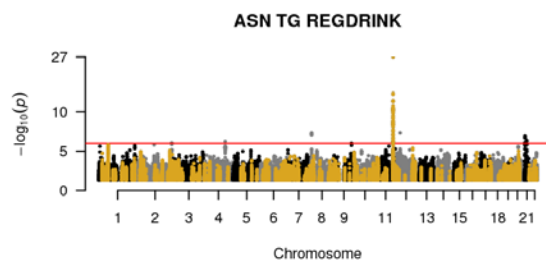
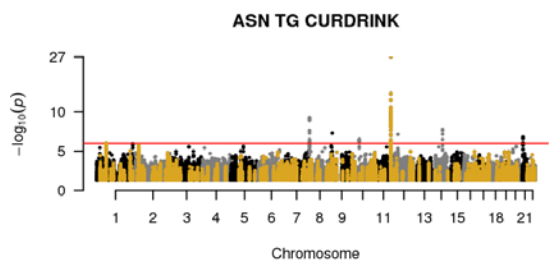
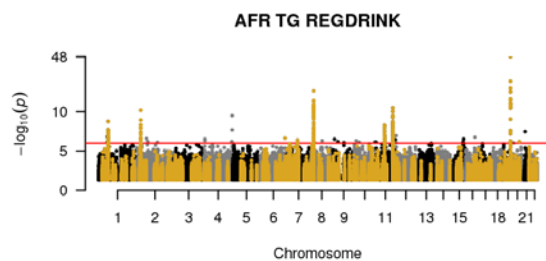
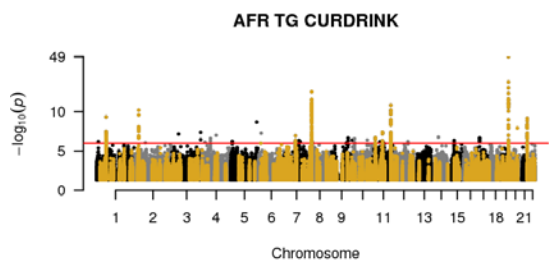
## References

1. Voorman A, Lumley T, McKnight B, et al. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One* 2011;6(5):e19416.
2. Tchetgen Tchetgen EJ, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiology* 2011;22(2):257-61.
3. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 2010;11:134.
4. Zeileis A. Object-oriented computation of sandwich estimators. *J Stat Softw* 2006;16(9).
5. Halekoh U, Hojsgaard S, Yan J. The R Package geepack for Generalized Estimating Equations. *J Stat Softw* 2006;15(2).
6. Aulchenko YS, Ripke S, Isaacs A, et al. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007;23(10):1294-6.
7. Winkler TW, Day FR, Croteau-Chonka DC, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* 2014;9(5):1192-212.
8. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11(7):499-511.
9. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26(17):2190-1.
10. Asselbergs FW, Guo Y, van Iperen EP, et al. Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am J Hum Genet* 2012;91(5):823-38.
11. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;45(11):1274-83.
12. Peloso GM, Auer PL, Bis JC, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* 2014;94(2):223-32.
13. Spracklen CN, Chen P, Kim YJ, et al. Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Hum Mol Genet* 2017;26(9):1770-84.
14. Surakka I, Horikoshi M, Magi R, et al. The impact of low-frequency and rare variants on lipid levels. *Nat Genet* 2015;47(6):589-97.
15. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;466(7307):707-13.
16. van Leeuwen EM, Sabo A, Bis JC, et al. Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. *J Med Genet* 2016;53(7):441-9.
17. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* 2015;6:5890.
18. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010;467(7319):1061-73.
19. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;39(Database issue):D691-7.
20. Blake JA, Bult CJ, Eppig JT, et al. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* 2014;42(Database issue):D810-7.
21. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40(Database issue):D109-14.
22. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25-9.
23. Lage K, Karlberg EO, Storling ZM, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;25(3):309-16.

**Web Figure 1.** Manhattan plots for each of the lipid-alcohol consumption combinations by ethnicity for Stage 1. The  $-\log_{10}(P\text{-values})$  for each variant are plotted against its chromosomal position. The line indicating suggestive significance at  $1 \times 10^{-6}$  is shown in red. Variants within  $\pm 1$  Mbp of a published known locus are highlighted in gold. AFR, African ancestry meta-analysis; ASN, Asian ancestry meta-analysis; CURDRINK, current drinkers; EUR, European ancestry meta-analysis; HDL-C, high-density lipoprotein cholesterol; HIS, Hispanic ancestry meta-analysis; LDL-C, low-density lipoprotein cholesterol; META, trans-ethnic meta-analysis; REGDRINK, regular drinkers; TG, triglycerides.

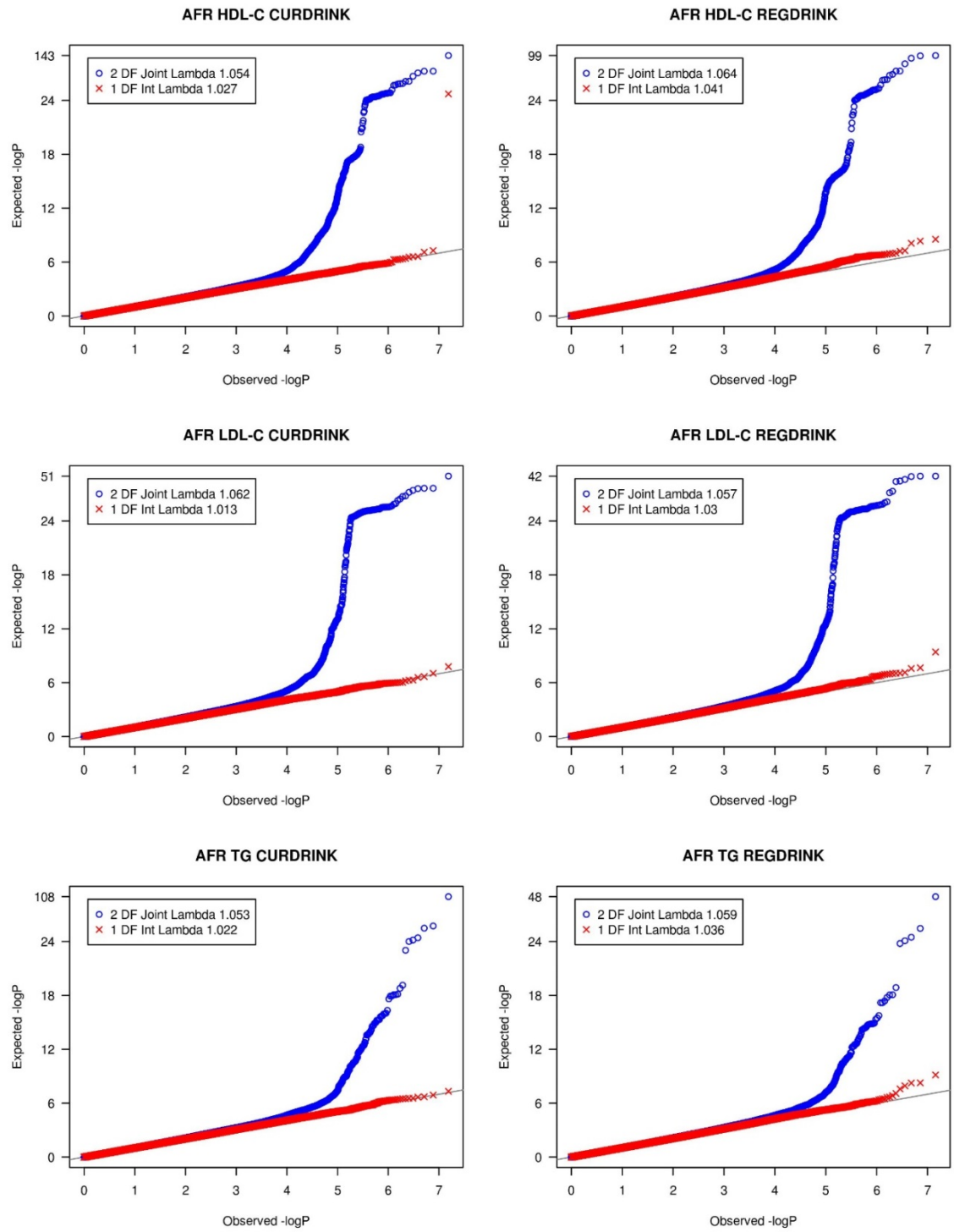




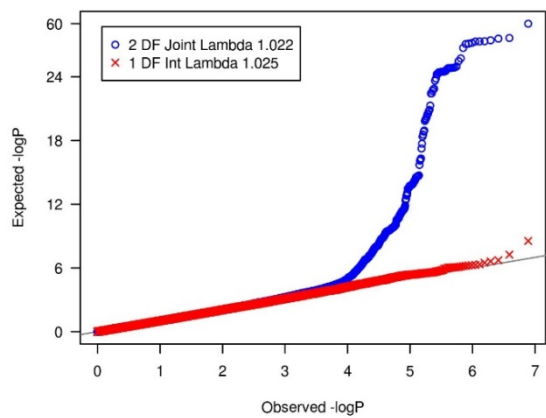




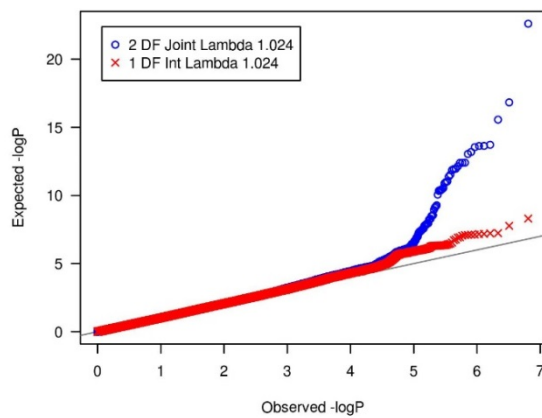
**Web Figure 2.** Quantile-quantile (QQ) plots for each of the lipid-alcohol consumption combinations by ethnicity for Stage 1. The observed  $-\log_{10}(P\text{-values})$  are plotted against the expected  $-\log_{10}(P\text{-values})$ . Here the blue circles denote the joint 2 DF test  $P$ -values. The red “X”s denote the 1 DF interaction  $P$ -values. AFR, African ancestry meta-analysis; ASN, Asian ancestry meta-analysis; CURDRINK, current drinkers; DF, degrees of freedom; EUR, European ancestry meta-analysis; HDL-C, high-density lipoprotein cholesterol; HIS, Hispanic ancestry meta-analysis; LDL-C, low-density lipoprotein cholesterol; META, trans-ethnic meta-analysis; REGDRINK, regular drinkers; TG, triglycerides.



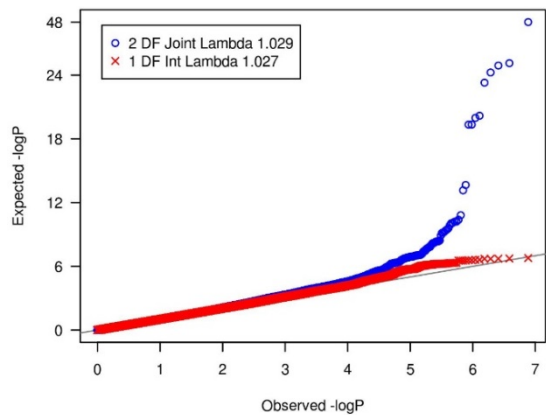
ASN HDL-C CURDRINK



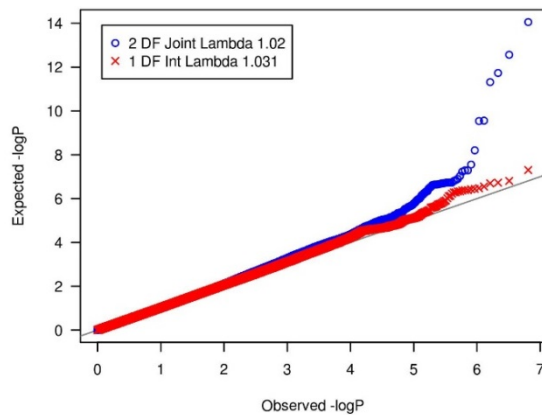
ASN HDL-C REGDRINK



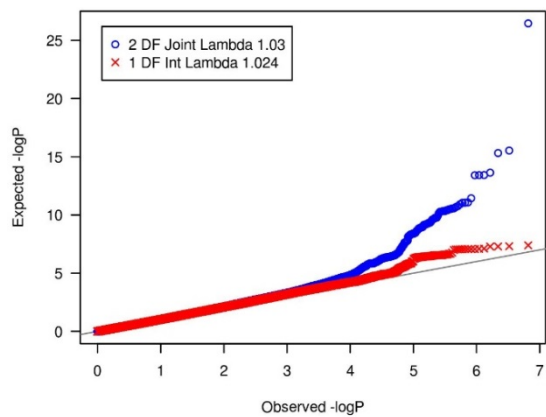
ASN LDL-C CURDRINK



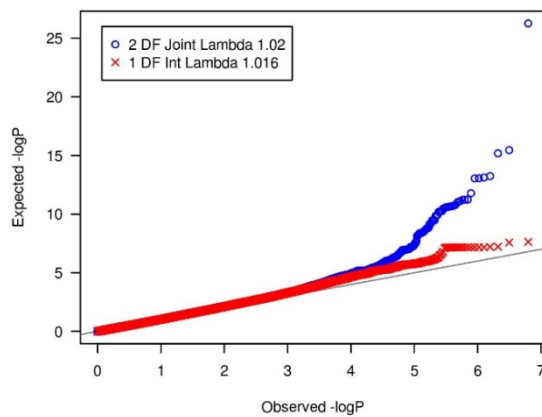
ASN LDL-C REGDRINK

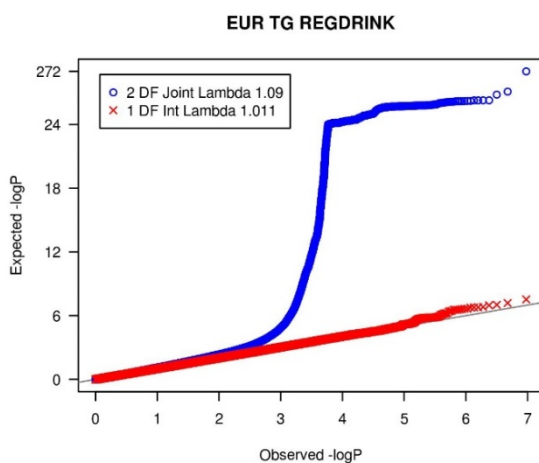
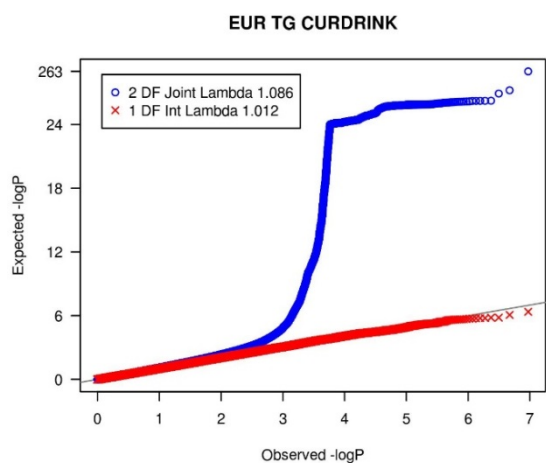
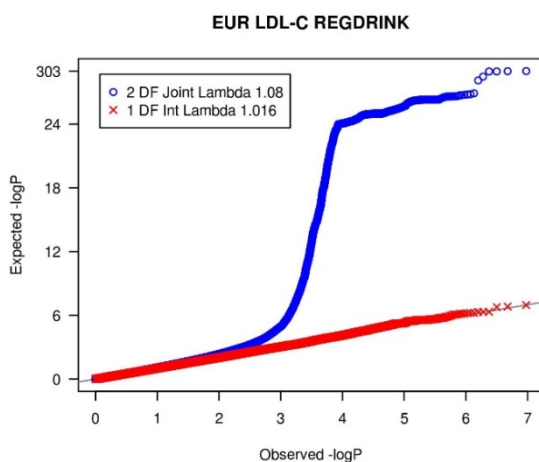
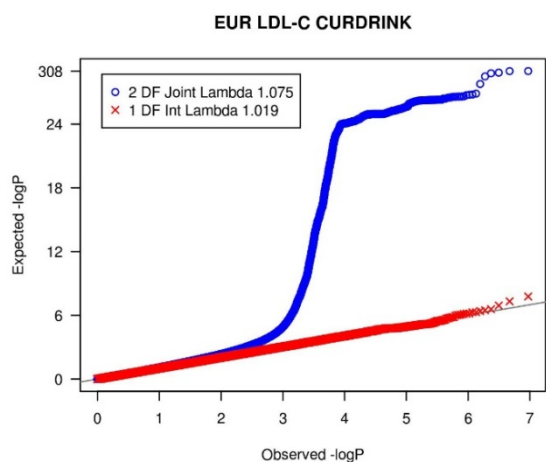
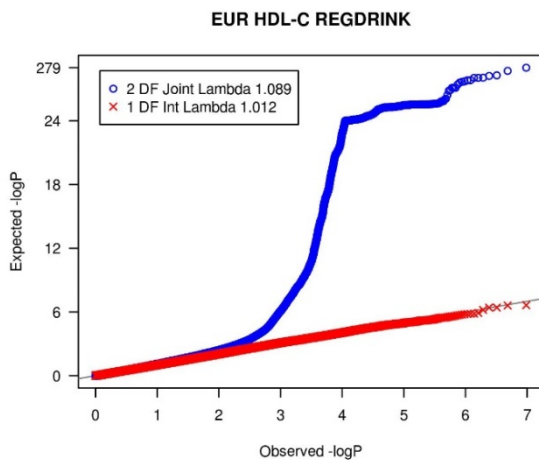
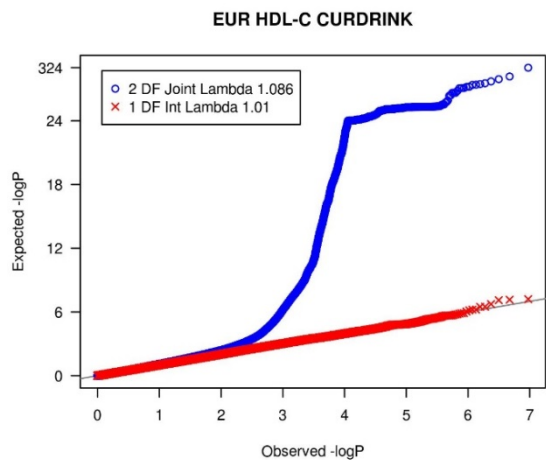


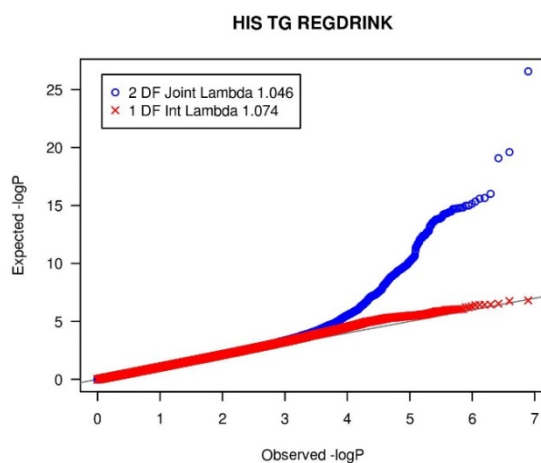
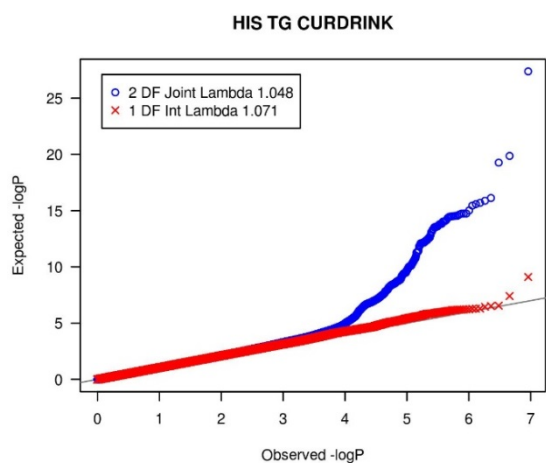
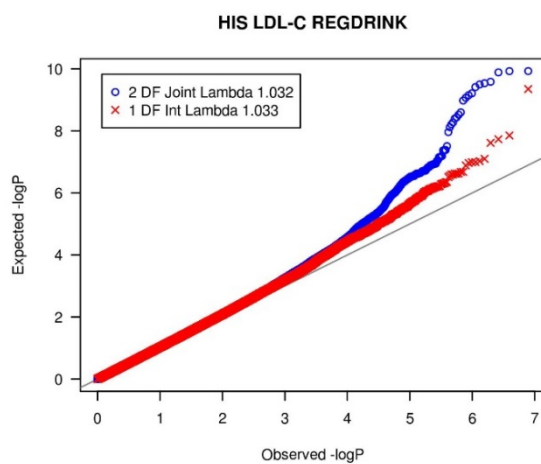
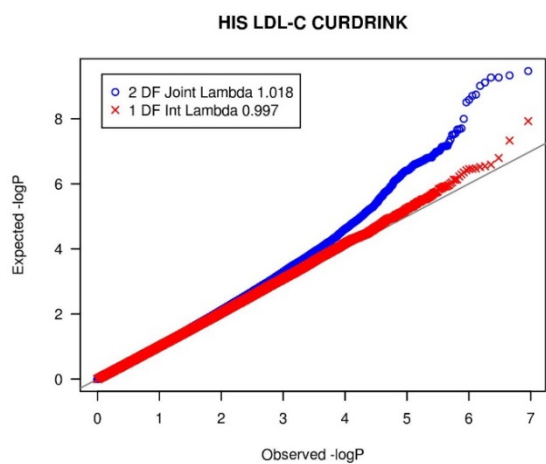
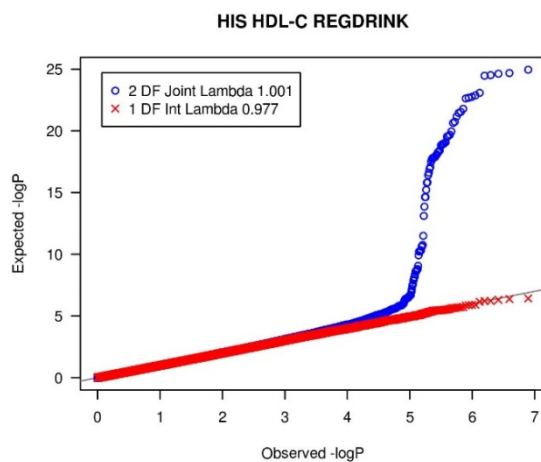
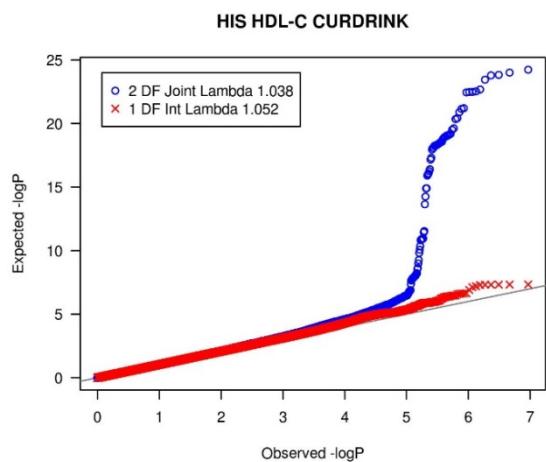
ASN TG CURDRINK

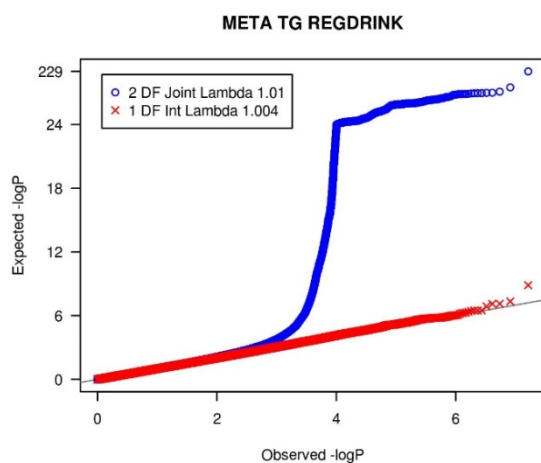
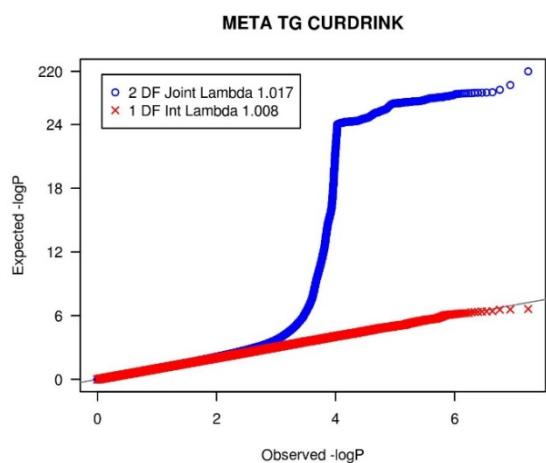
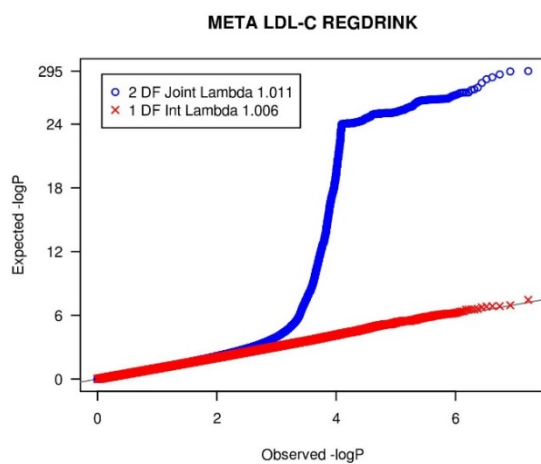
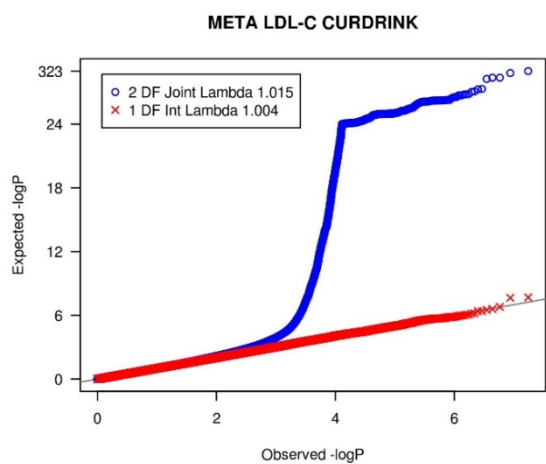
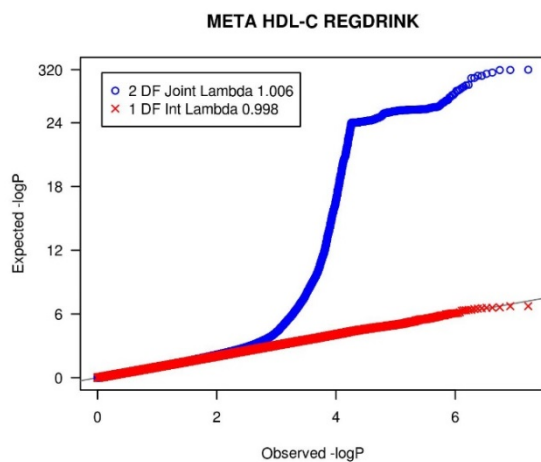
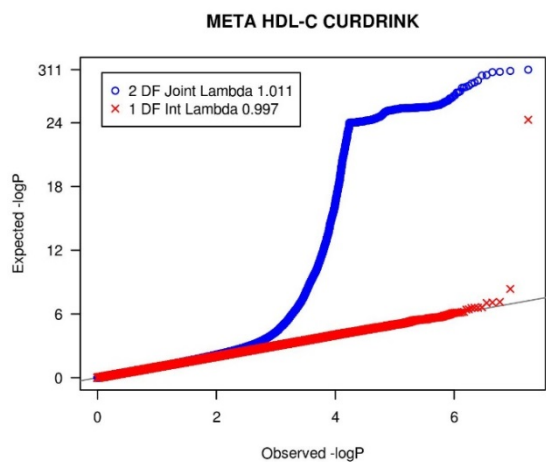


ASN TG REGDRINK



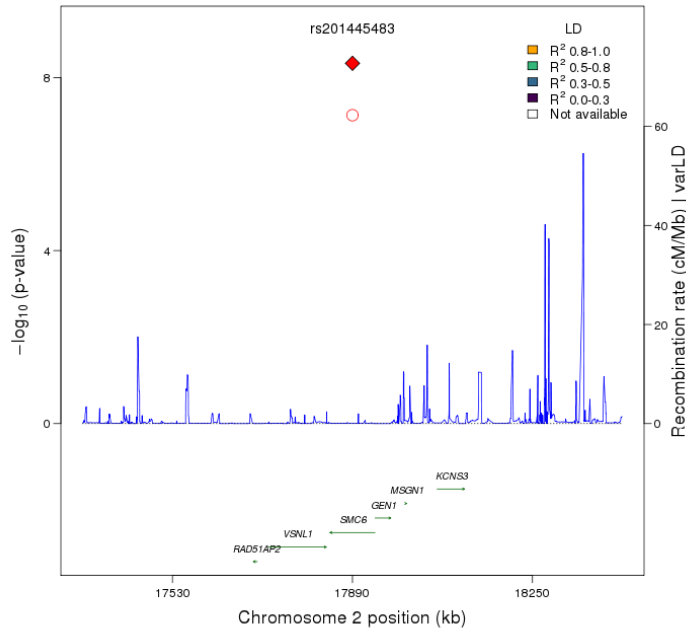




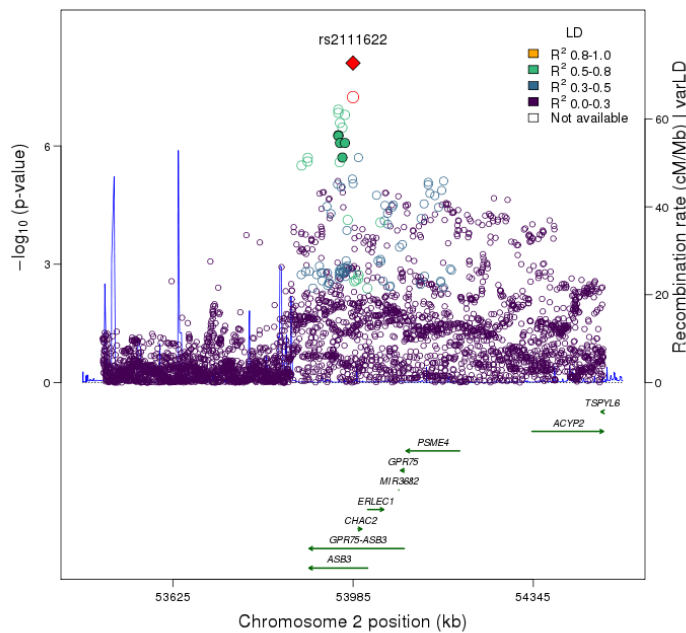


**Web Figure 3.** Regional plots for 18 novel loci, with  $P$ -values from the joint 2 DF test. Variants are identified from trans-ethnic meta-analyses (META), as well as European (EUR) and African (AFR) ancestry-specific meta-analyses. Unfilled circles represent variants that were included only Stage 1 and not Stage 2. Filled circles represent variants that were included in the combined analysis of Stage 1 and Stage 2. HDL-C, high-density lipoprotein cholesterol; LD, linkage disequilibrium; LDL-C, low-density lipoprotein cholesterol; TG, triglycerides.

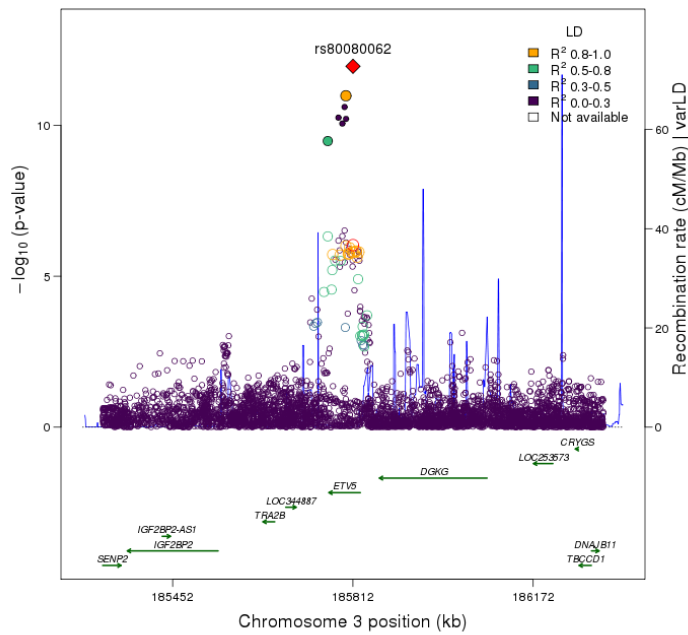
3a. Variant rs201445483 at chromosome 2, position 17890087. LDL-C Current Drinker (META)



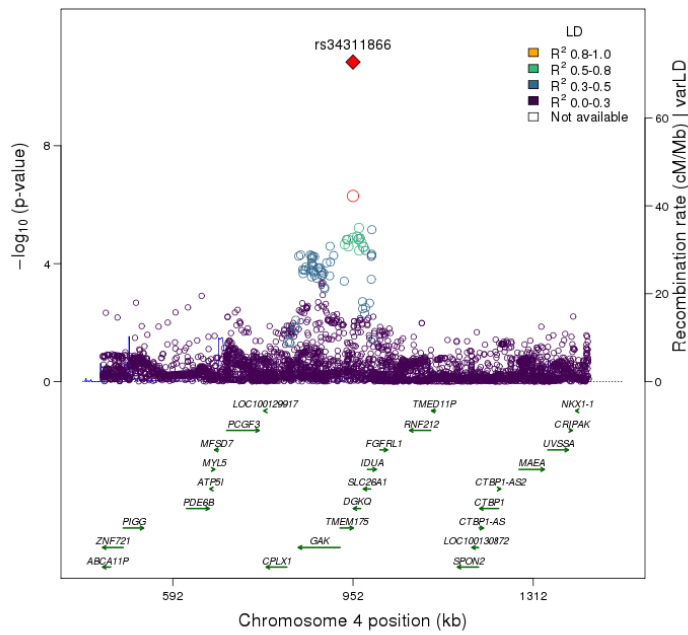
3b. Variant rs2111622 at chromosome 2, position 53984823. HDL-C Current Drinker (EUR)



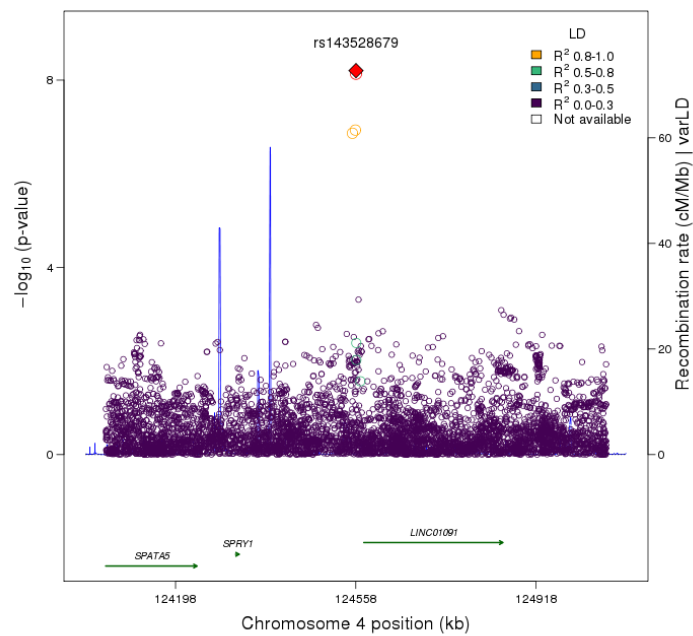
3c. Variant rs80080062 at chromosome 3, position 185812169. HDL-C Regular Drinker (META)



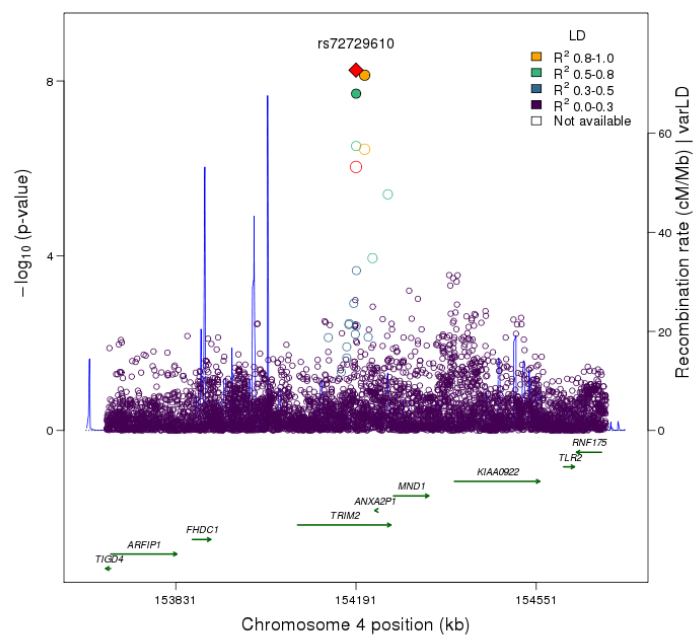
3d. Variant rs34311866 at chromosome 4, position 951947. TG Current Drinker (EUR)



3e. Variant rs143528679 at chromosome 4, position 124558378. LDL-C Current Drinker (AFR)

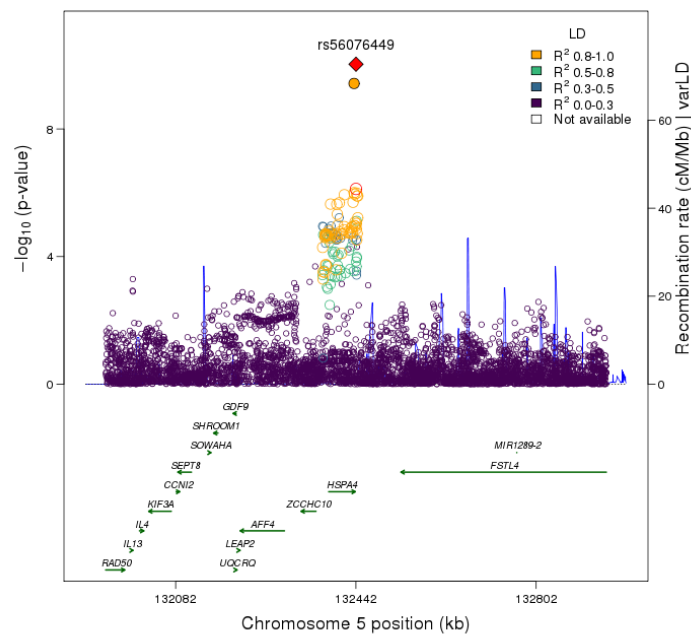


3f. Variant rs72729610 at chromosome 4, position 154190965. HDL-C Regular Drinker (META)

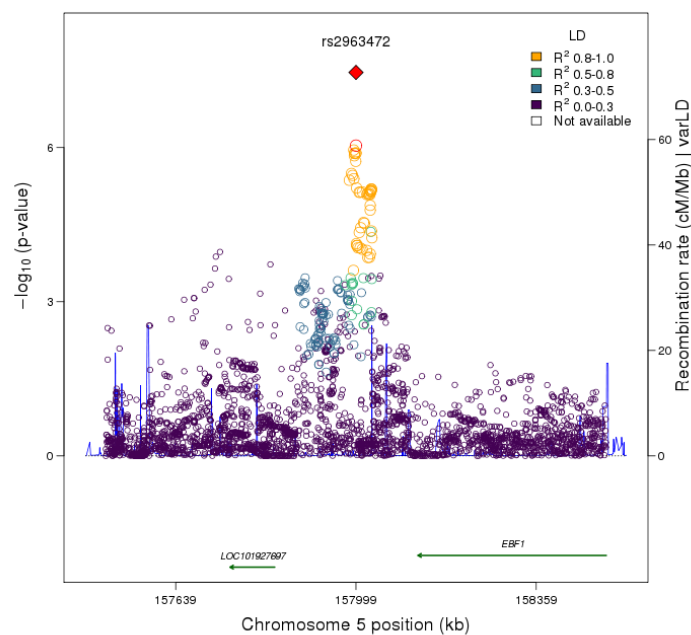




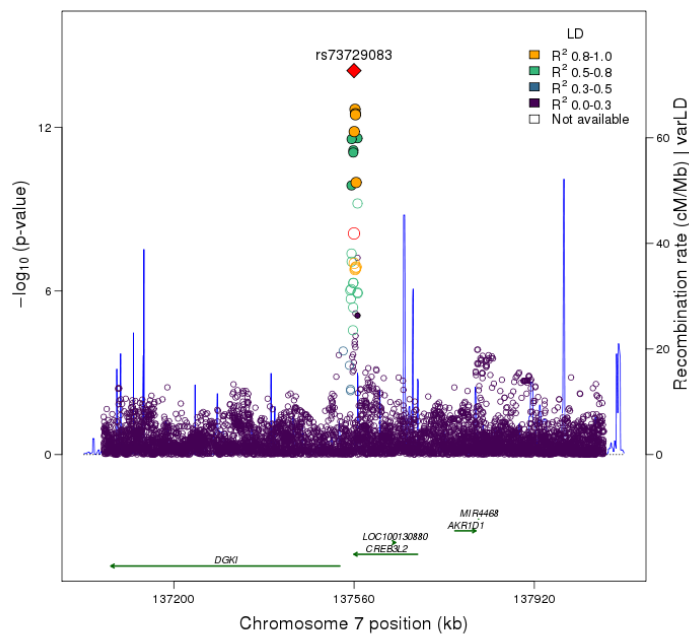
3g. Variant rs56076449 at chromosome 5, position 132442190. TG Regular Drinker (META)



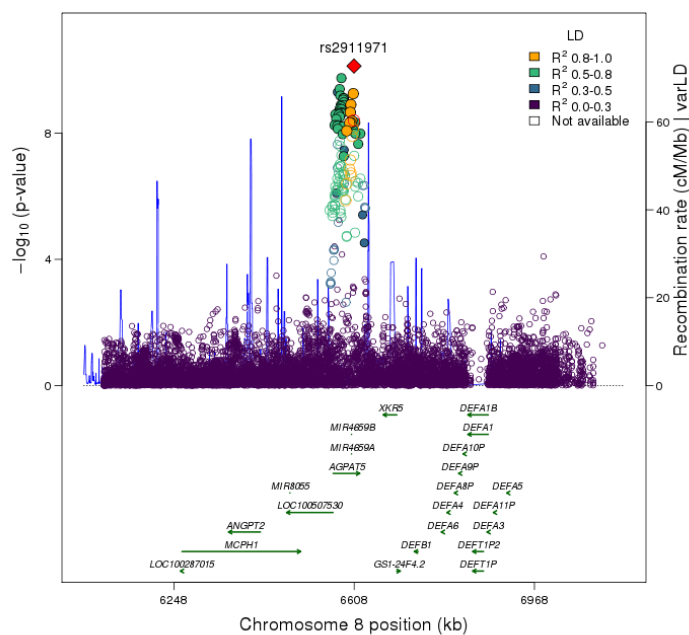
3h. Variant rs2963472 at chromosome 5, position 157999022. TG Regular Drinker (EUR)



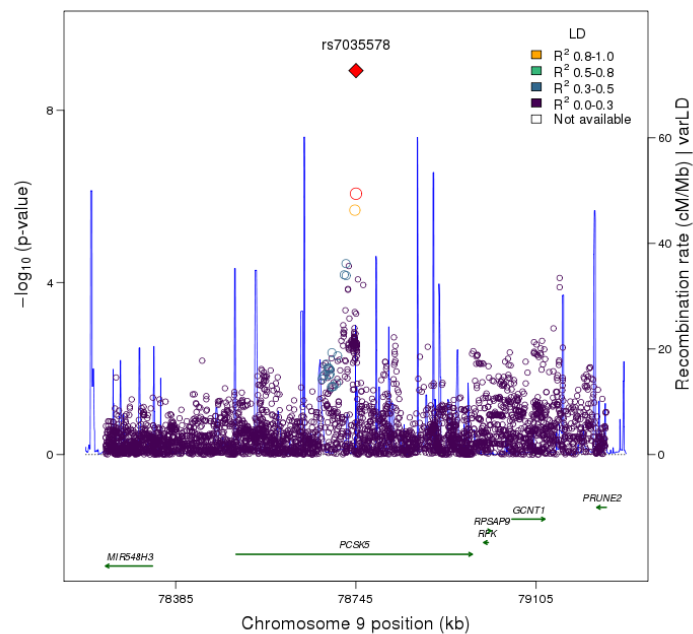
3i. Variant rs73729083 at chromosome 7, position 137559799. LDL-C Current Drinker (META)



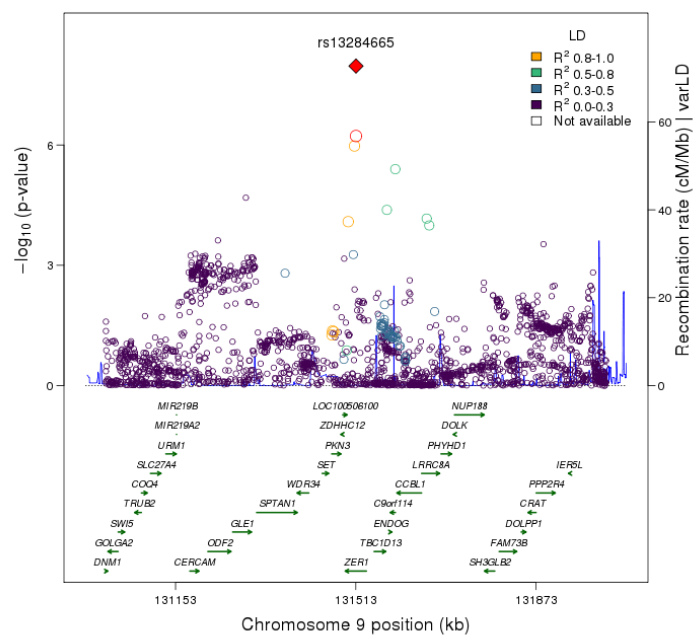
3j. Variant rs2911971 at chromosome 8, position 6607634. LDL-C Current Drinker (META)



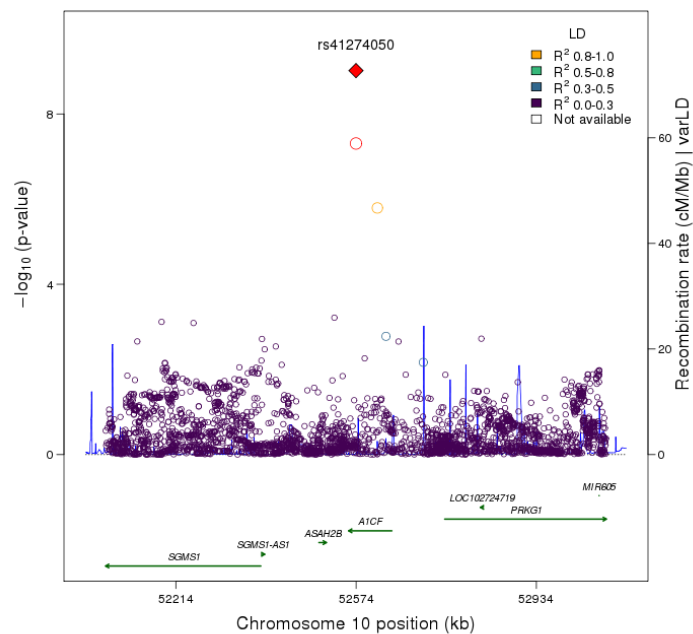
3k. Variant rs7035578 at chromosome 9, position 78745177. LDL-C Current Drinker (EUR)



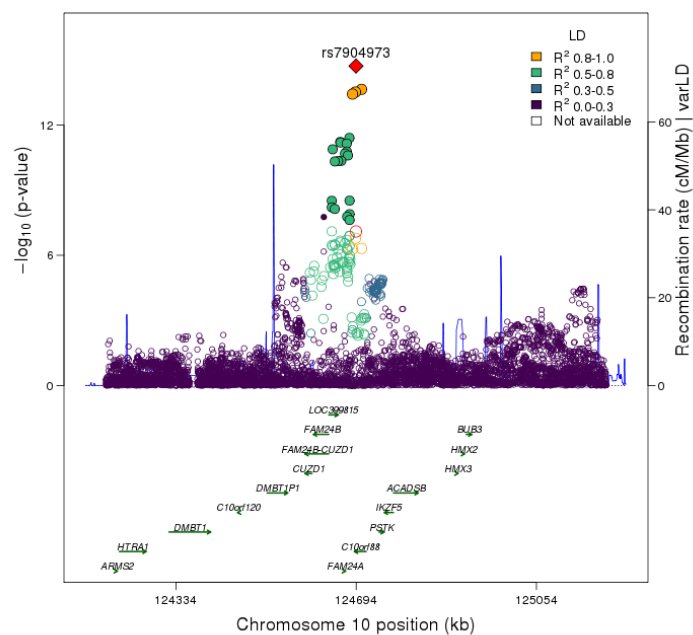
3l. Variant rs13284665 at chromosome 9, position 131513370. LDL-C Current Drinker (EUR)



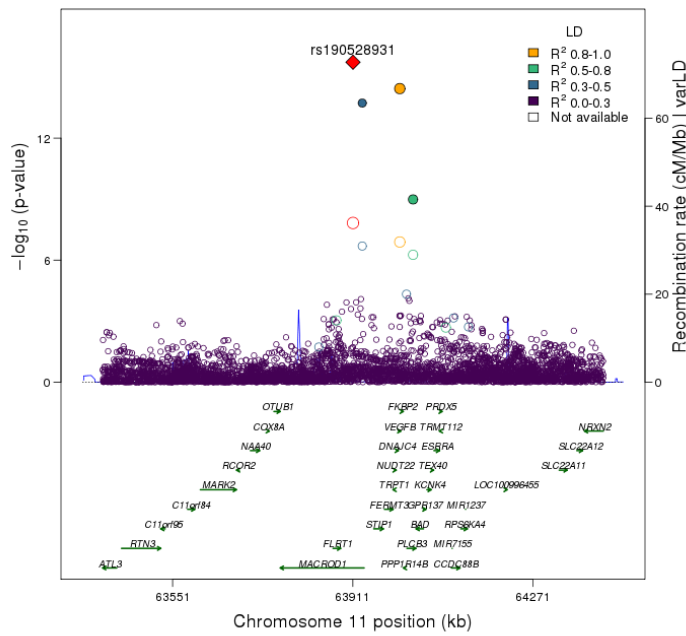
3m. Variant rs41274050 at chromosome 10, position 52573772. TG Regular Drinker (EUR)



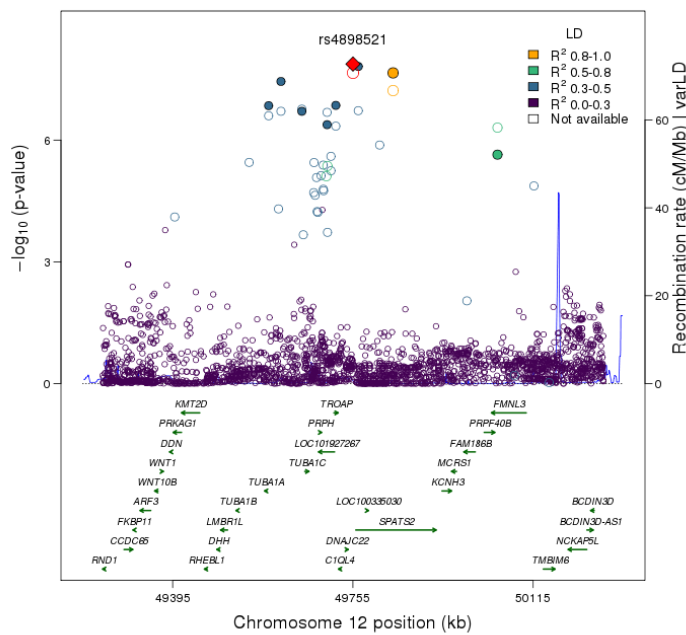
3n. Variant rs7904973 at chromosome 10, position 124693587. LDL-C Current Drinker (META)



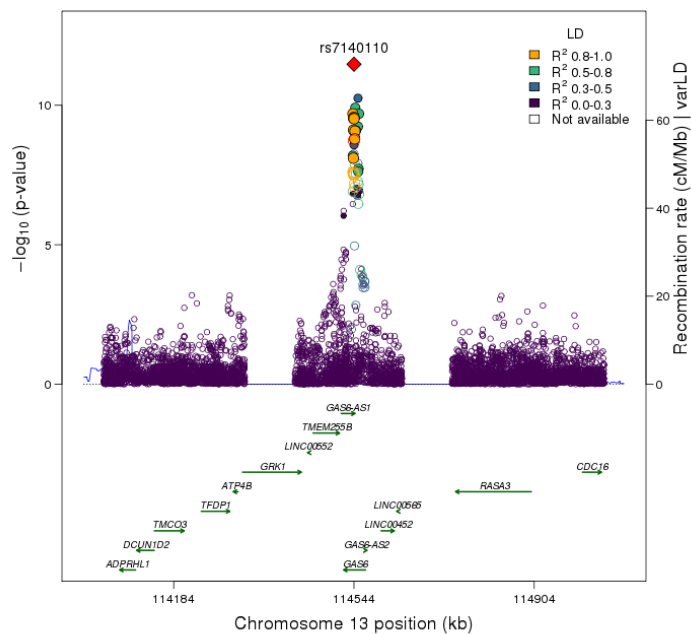
3o. Variant rs190528931 at chromosome 11, position 63911273. HDL-C Current Drinker (META)



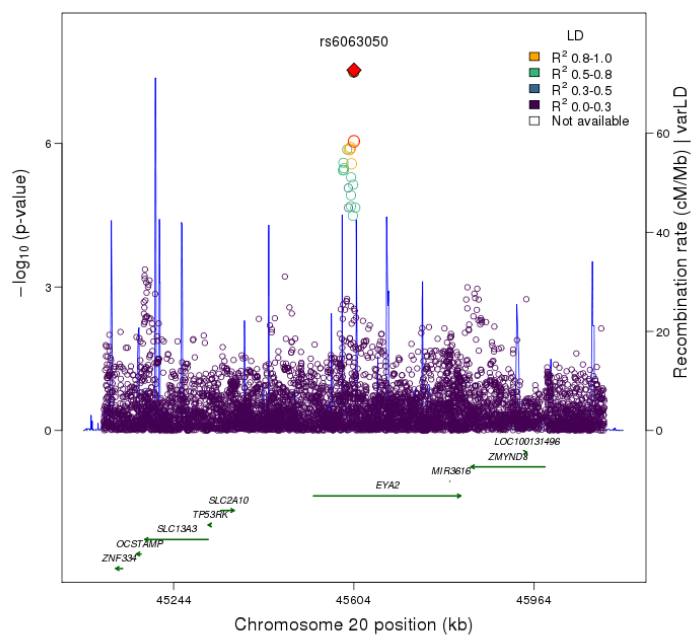
3p. Variant rs4898521 at chromosome 12, position 49755162. HDL-C Regular Drinker (EUR)



3q. Variant rs7140110 at chromosome 13, position 114544024. TG Current Drinker (META)



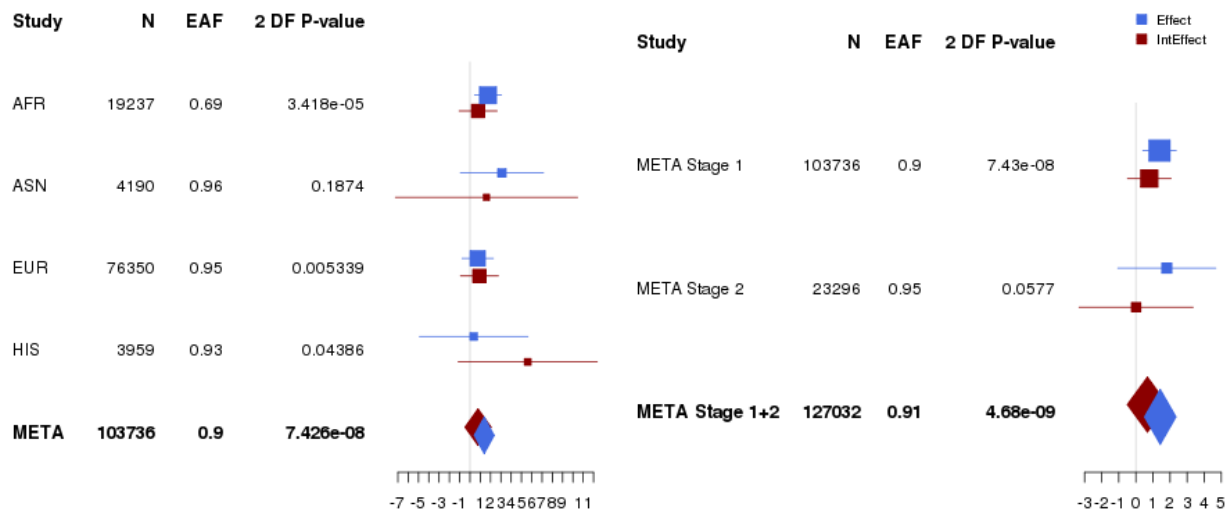
3r. Variant rs6063050 at chromosome 20, position 45604240. TG Current Drinker (META)



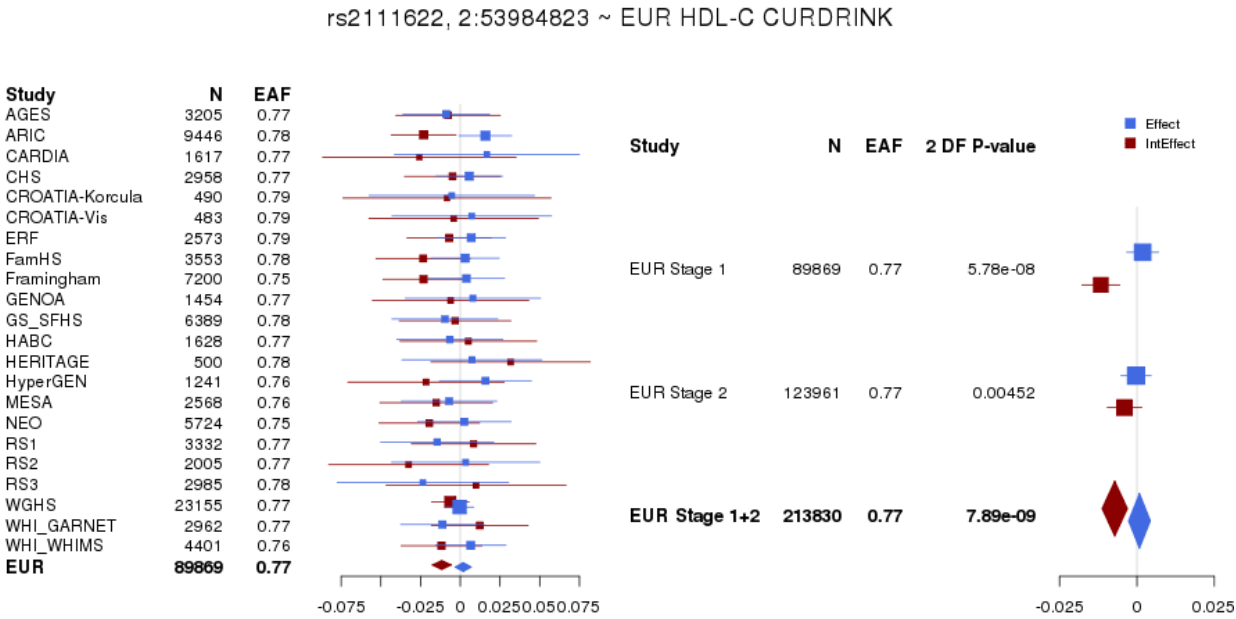
**Web Figure 4.** Forest plots of the top hits of the 18 novel loci. The left panel shows the studies used in Stage 1, while the right panel compares the effects sizes of the Stage 1 meta-analysis with the Stage 2 meta-analysis, used in the combined Stage 1 + Stage 2 analysis. Point estimates of the main effect are shown in blue with 95% confidence interval bands, and those for the interaction effect are shown in red. For variants identified through trans-ethnic meta-analysis (META), the left panel shows the constituent ancestry-specific meta-analyses instead of individual studies' effects. AFR, African ancestry meta-analysis; AGES, Age Gene/Environment Susceptibility Reykjavik Study; ARIC, Atherosclerosis Risk in Communities Study; ASN, Asian ancestry meta-analysis; CARDIA, Coronary Artery Risk Development in Young Adults; CHS, Cardiovascular Health Study; CURDRINK, current drinkers; DF, degrees of freedom; EAF, effect allele frequency; Effect, SNP main effect; ERF, Erasmus Rucphen Family Study; EUR, European ancestry meta-analysis; FamHS, Family Heart Study; Framingham, Framingham Heart Study; GENOA, Genetic Epidemiology Network of Arteriopathy; GS\_SFHS, Generation Scotland: Scottish Family Health Study; HABC, Health, Aging, and Body Composition; HANDLS, Healthy Aging in Neighborhoods of Diversity across the Life Span; HDL-C, high-density lipoprotein cholesterol; HERITAGE, Health, Risk Factors, Exercise Training and Genetics Family Study; HIS, Hispanic ancestry meta-analysis; HUFS, Howard University Family Study; HyperGEN, Hypertension Genetic Epidemiology Network; IntEffect, SNP interaction effect; JHS, Jackson Heart Study; LDL-C, low-density lipoprotein cholesterol; MESA, Multi-Ethnic Study of Atherosclerosis; N, number of samples; NEO, Netherlands Epidemiology of Obesity Study; REGDRINK, regular drinkers; rs, reference SNP; RS1, Rotterdam Study, first cohort; RS2, Rotterdam Study, second cohort; RS3, Rotterdam Study, third cohort; SNP, single nucleotide polymorphism; TG, triglycerides; WGHS, Women's Genome Health Study; WHI, Women's Health Initiative; WHI\_GARNET, Women's Health Initiative, Genomics and Randomized Trials Network cohort; WHI\_WHIMS, Women's Health Initiative, Memory Study cohort.

#### 4a. Variant rs201445483

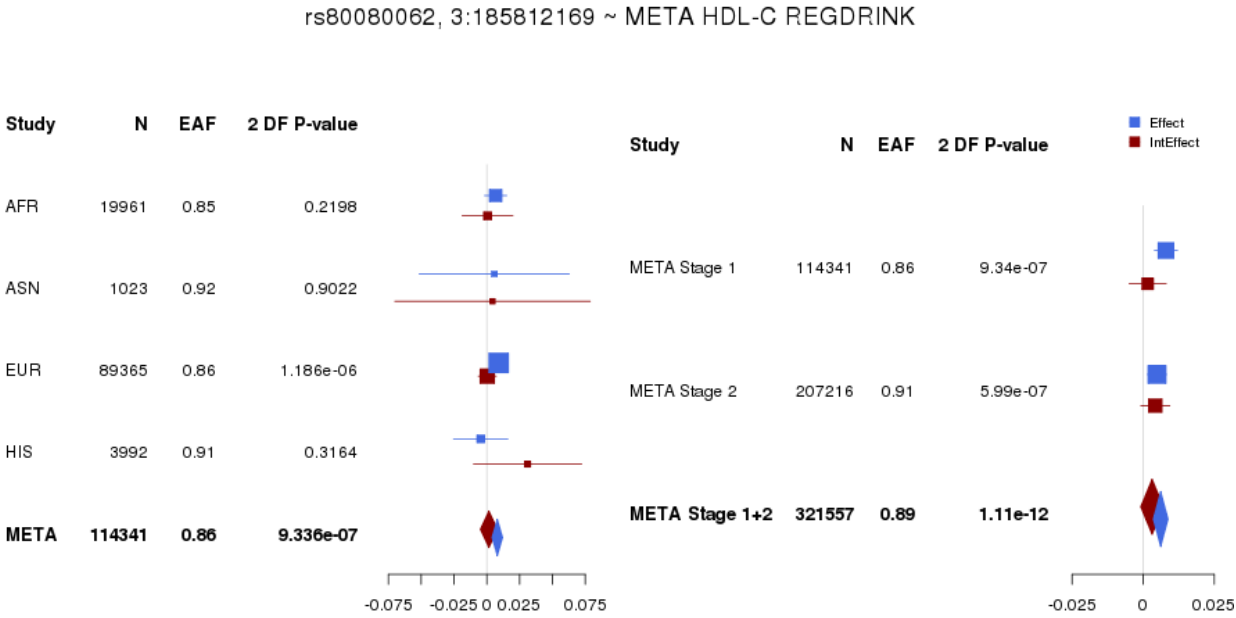
rs201445483, 2:17890087:ID ~ META LDL-C CURDRINK



4b. Variant rs2111622



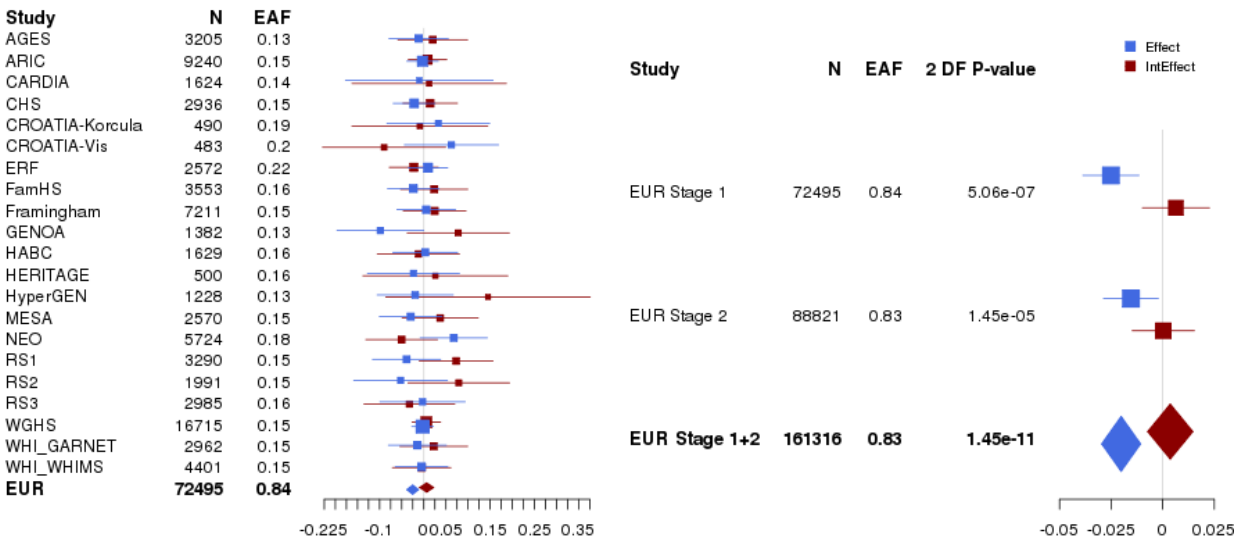
4c. Variant rs80080062





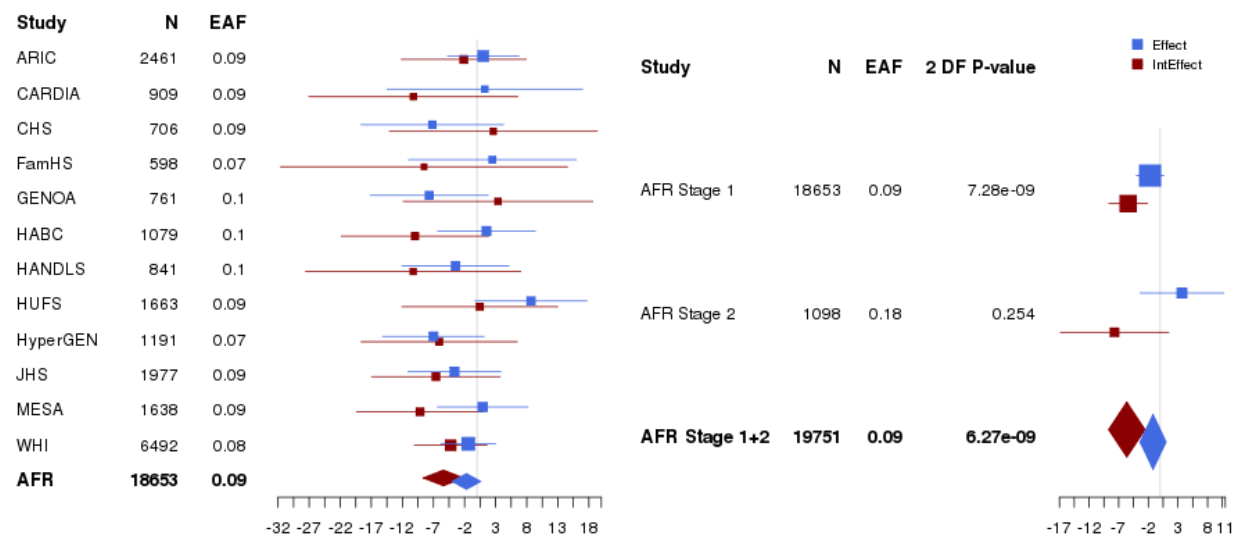
4d. Variant rs34311866

rs34311866, 4:951947 ~ EUR TG CURDRINK



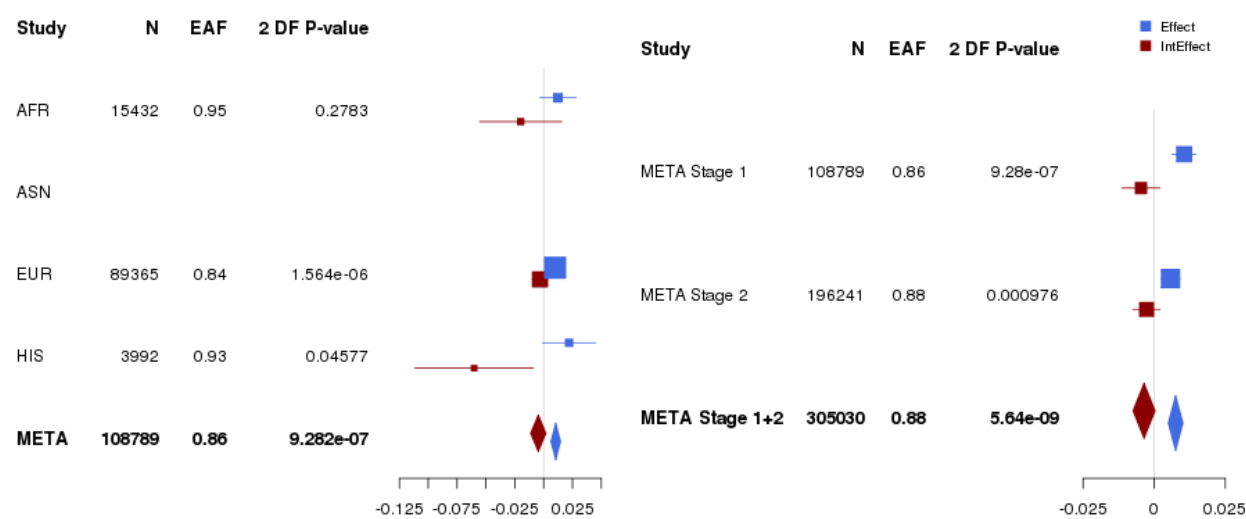
4e. Variant rs143528679

rs143528679, 4:124558378 ~ AFR LDL-C CURDRINK



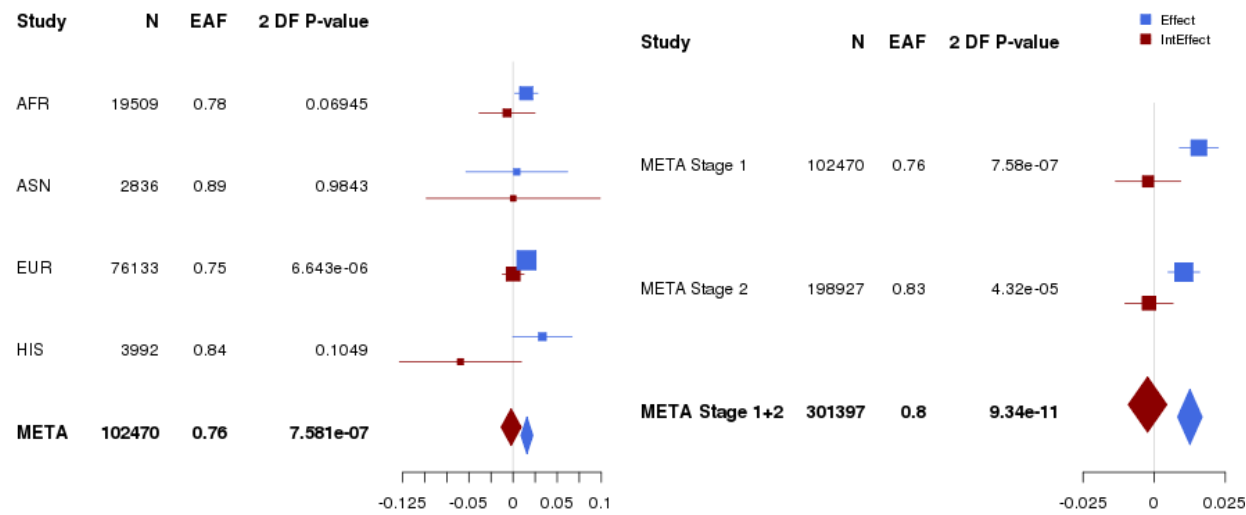
4f. Variant rs72729610

rs72729610, 4:154190965 ~ META HDL-C REGDRINK



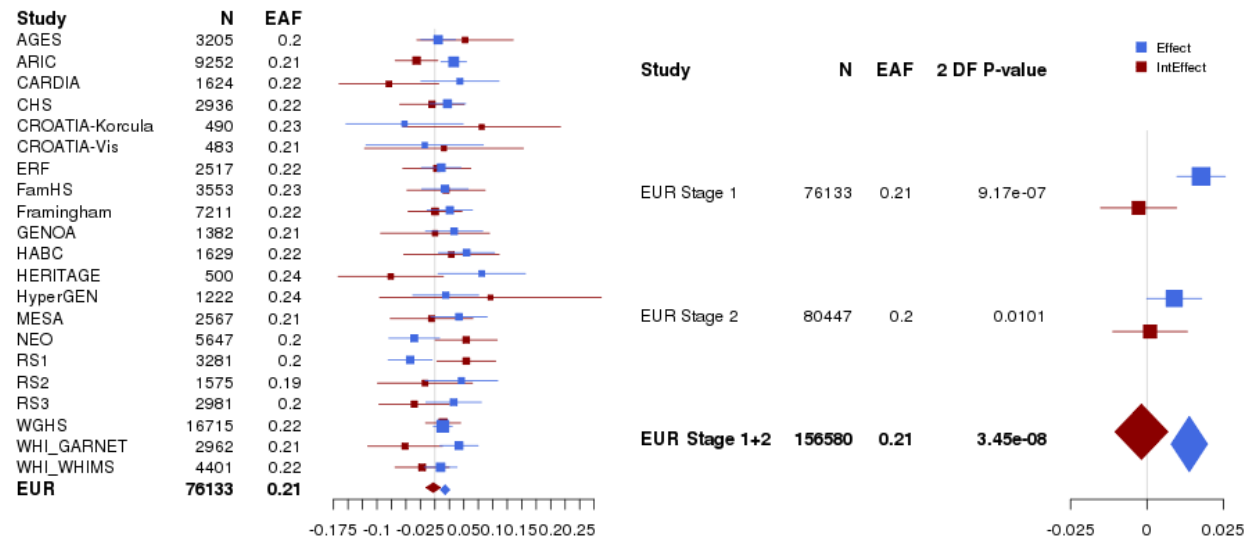
4g. Variant rs56076449

rs56076449, 5:132442190 ~ META TG REGDRINK



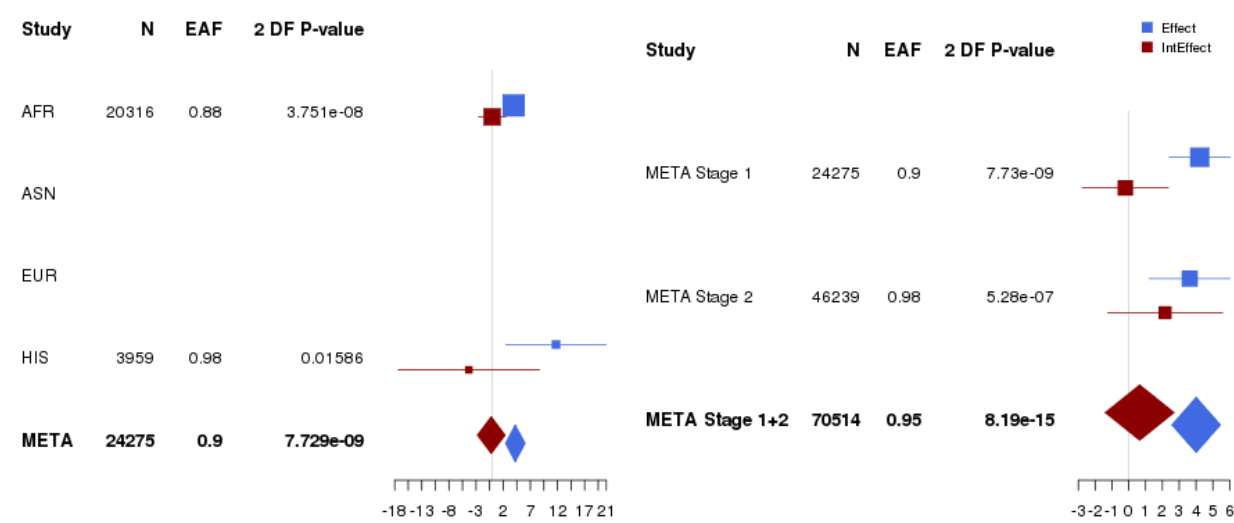
4h. Variant rs2963472

rs2963472, 5:157999022 ~ EUR TG REGDRINK



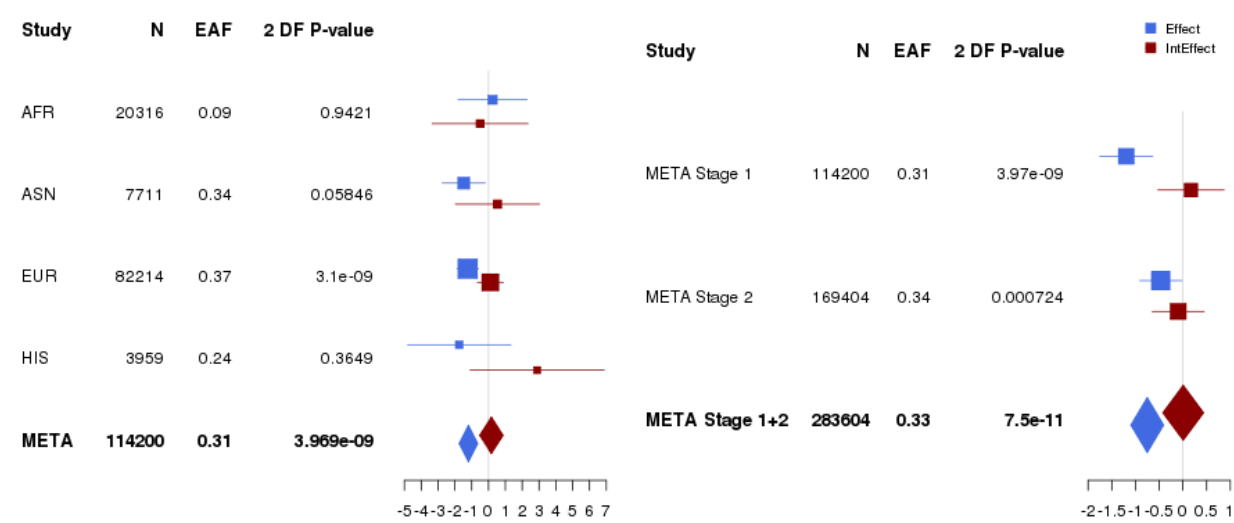
4i. rs73729083

rs73729083, 7:137559799 ~ META LDL-C CURDRINK



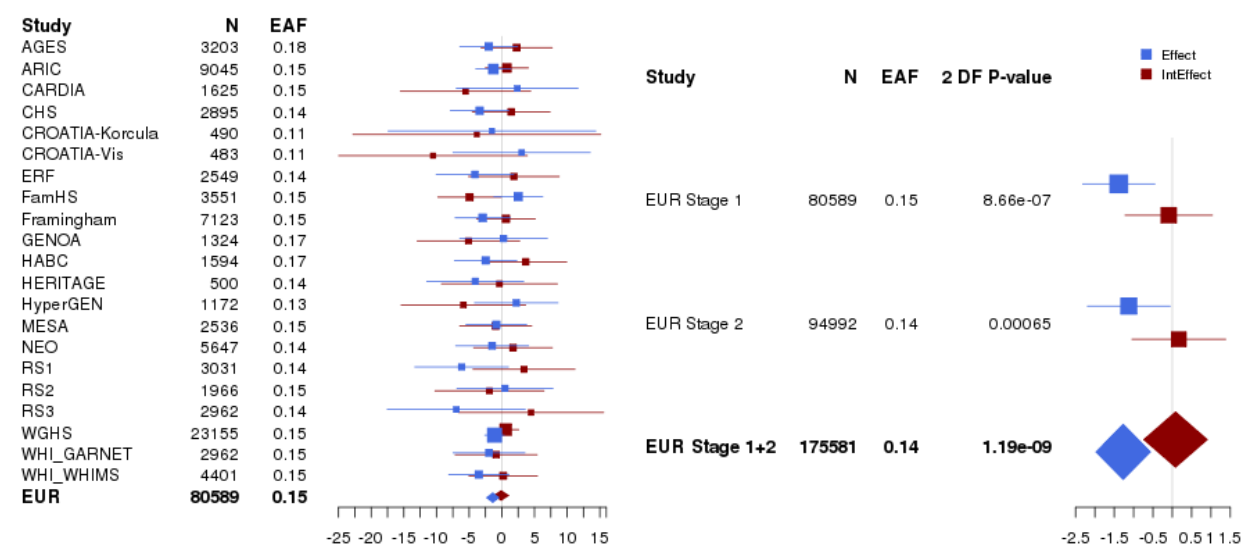
4j. Variant rs2911971

rs2911971, 8:6607634 ~ META LDL-C CURDRINK



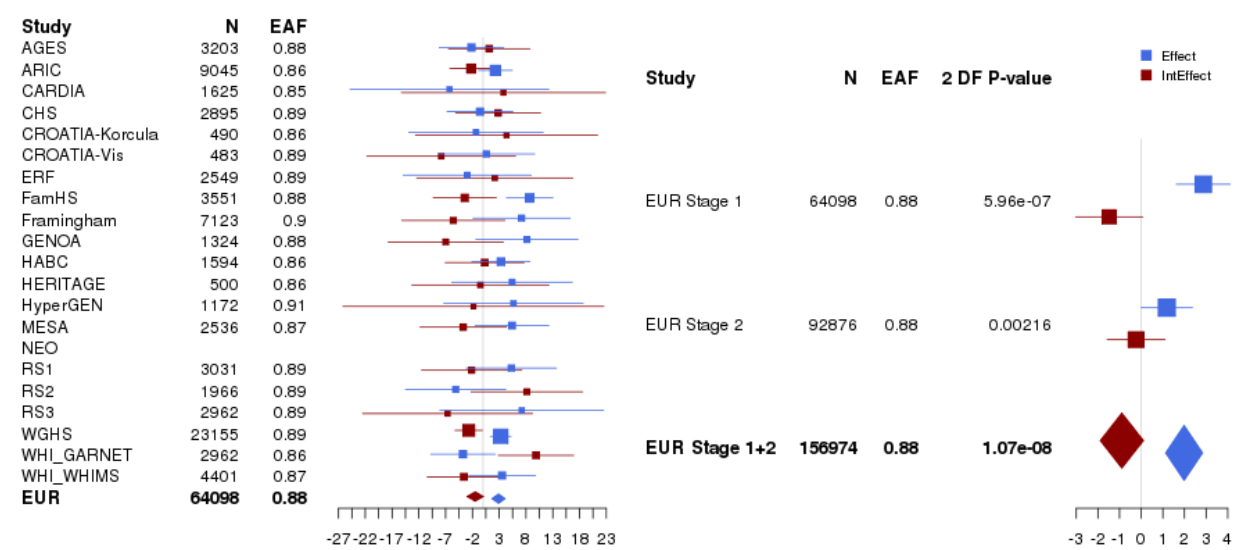
4k. Variant rs7035578

rs7035578, 9:78745177 ~ EUR LDL-C CURDRINK



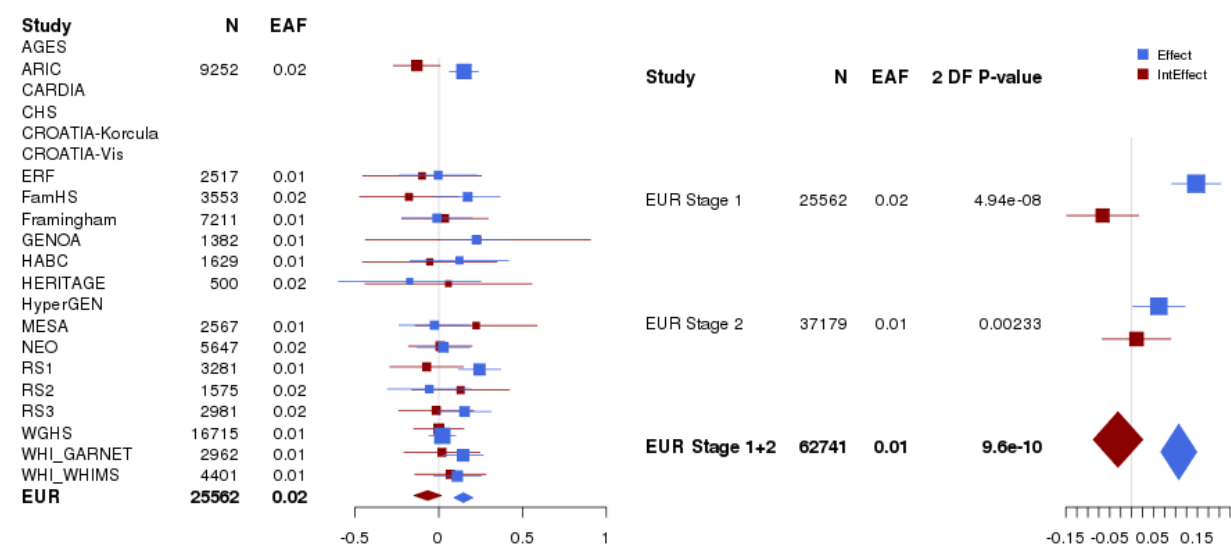
rs13284665

rs13284665, 9:131513370 ~ EUR LDL-C CURDRINK



4m. Variant rs41274050

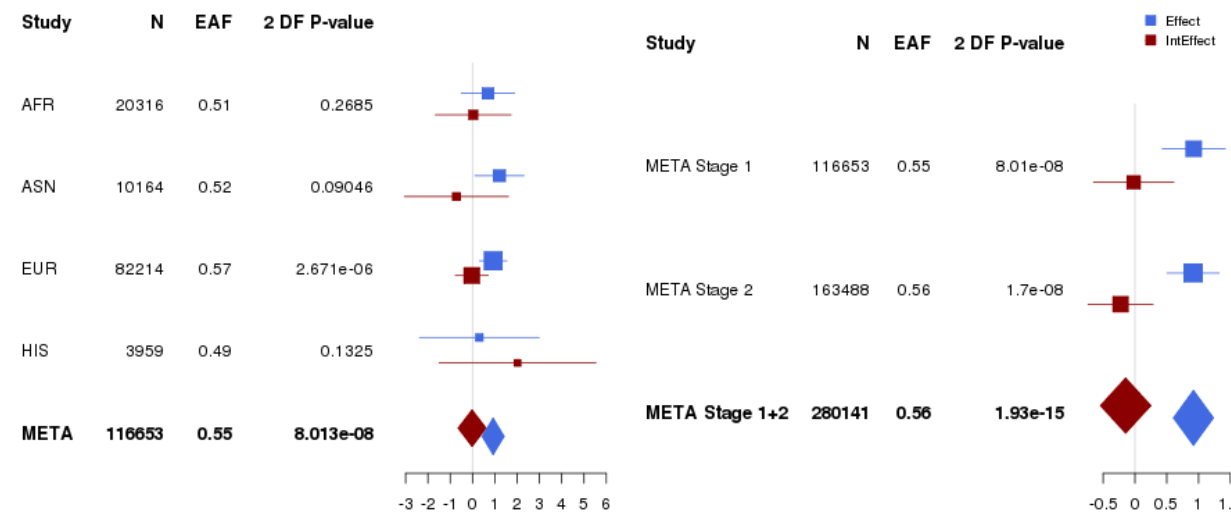
rs41274050, 10:52573772 ~ EUR TG REGDRINK



4n. Variant

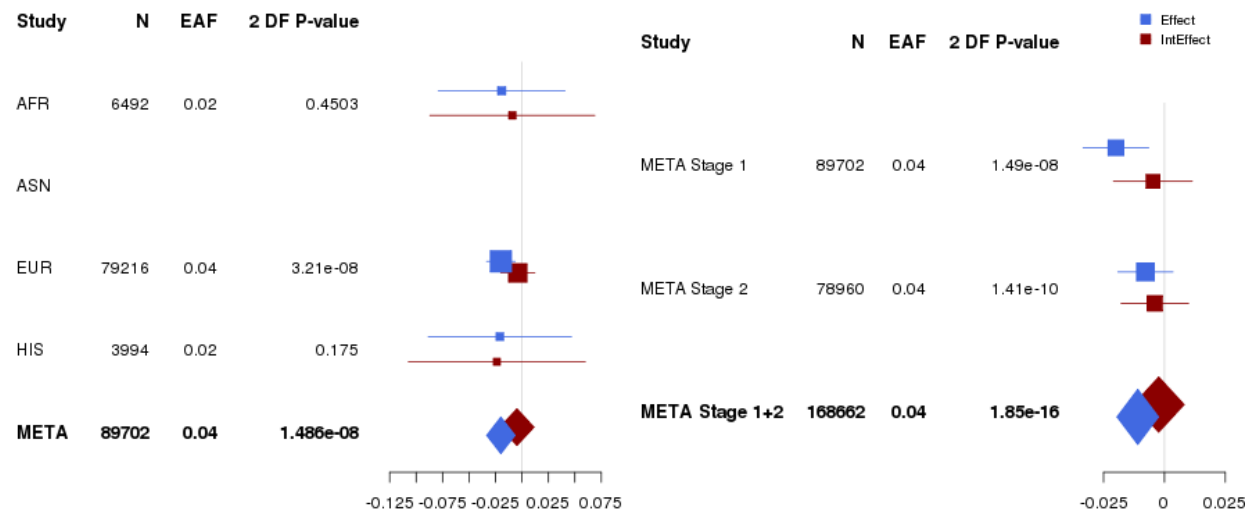
rs7904973

rs7904973, 10:124693587 ~ META LDL-C CURDRINK



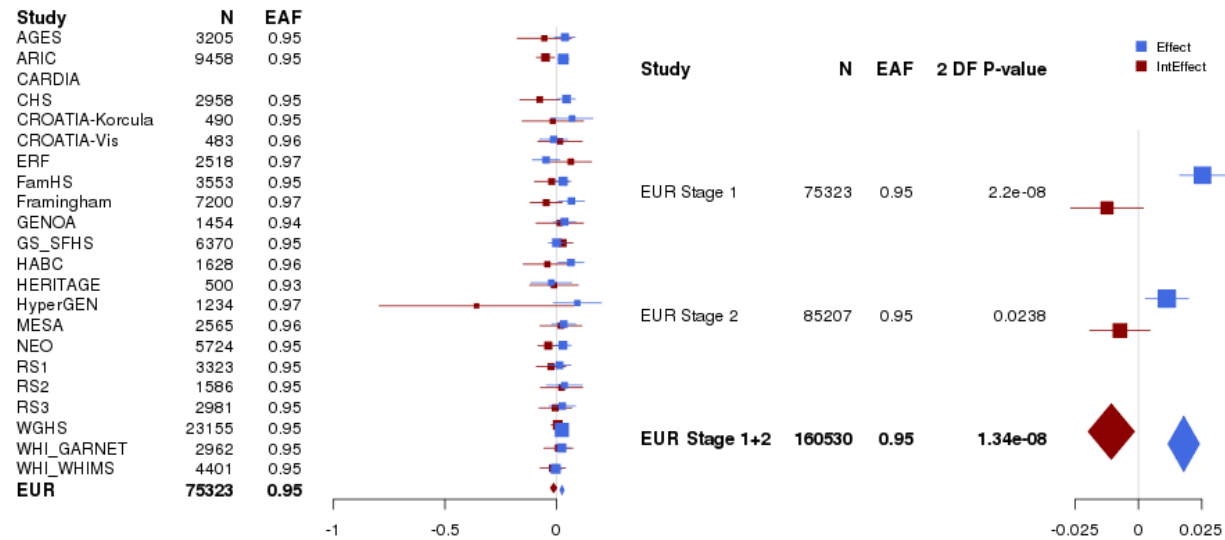
4o. Variant rs190528931

rs190528931, 11:63911273 ~ META HDL-C CURDRINK



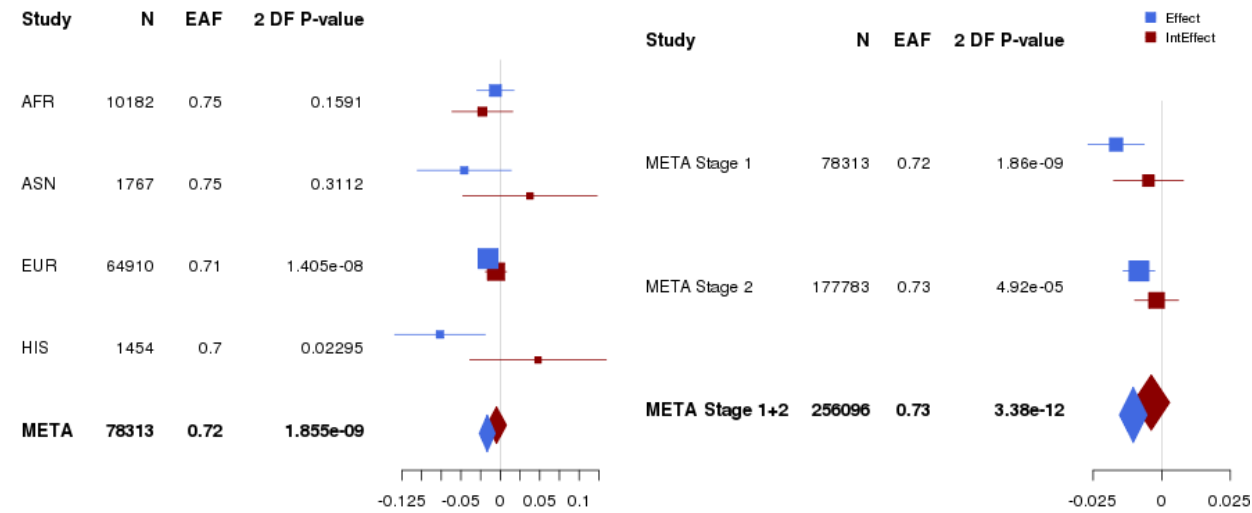
4p. Variant rs4898521

rs4898521, 12:49755162 ~ EUR HDL-C REGDRINK



4q. Variant rs7140110

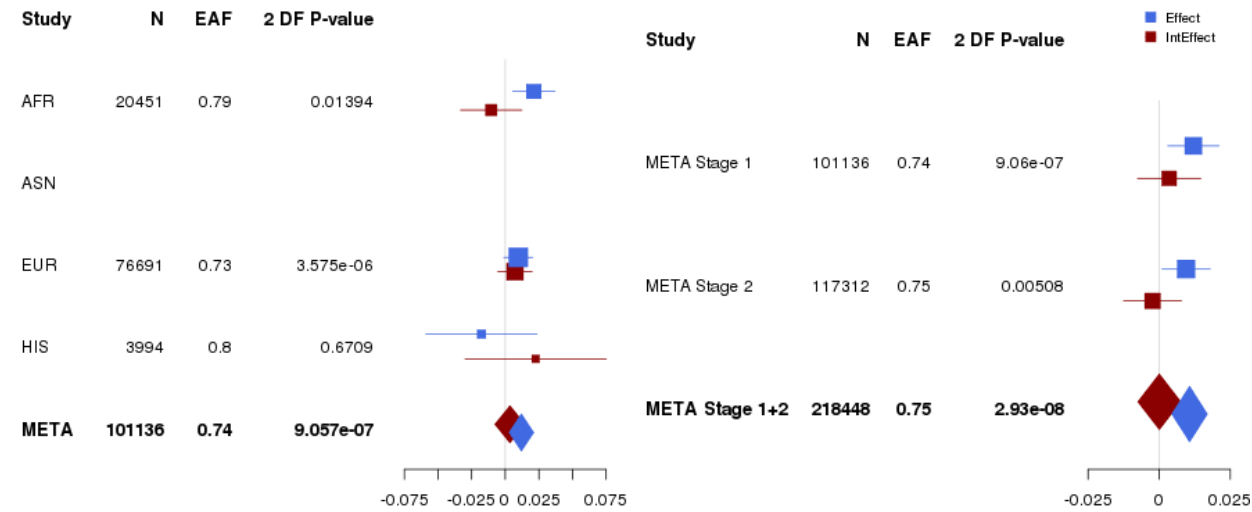
rs7140110, 13:114544024 ~ META TG CURDRINK



4r. Variant

rs6063050

rs6063050, 20:45604240 ~ META TG CURDRINK





Web Figure 5. Tissue, cell type, and physiological system expression of genes at identified loci associated with A) high-density lipoprotein cholesterol, B) low-density lipoprotein cholesterol, and C) triglycerides.

