# Finding PCSK9 variants affecting LDL-C level in Chinese population

1. Introduction

LDL-C (Low-Density Lipoprotein Cholesterol) level in bloodstream is a significant health factor. It has been well established that high LDL-C leads to increased cardiovascular risk. Given the measurement of LDL-C concentration, and the genotype data from the CKB project, we can perform association studies in Chinese cohorts to identify relevant genes, and the corresponding genetic variants.

PCSK9 (proprotein convertase subtilisin/kexin type 9) is one of the relevant genes. The convertase binds to LDLR (Low-Density Lipoprotein Receptor) and causes its degradation. While LDLR removes LDL-C from circulation on the surface of liver, inactivation of the protein induces higher LDL-C levels and increased risk of cardiovascular events. PCSK9 was intensively studied in Caucasian populations. Drugs blocking the protein have been developed to reduce LDL-C level in blood.

Some Caucasian PCSK9 variants identified as associated with LDL-C are absent from or present at low frequencies in Chinese. We seek to identify additional PCSK9 variants associated with LDL-C levels in Chinese populations.

1. Data

The genotype data set is the CKB genotyping phase1-2, QC stage 3. Only autosome SNPs (Single Nucleotide Polymorphisms) were used, leaving 636670 variations and 32205 subjects in the set.

Two types of LDL-C levels were used in this study. They are the indirect measurements estimated from the levels of total cholesterol using the Friedewald equation, and the direct measurements from a biochemistry laboratory using conventional instruments.

The direct LDL-C values were from the file LDL-c\_biochem\_data.xlsx in the directory K:\kadoorie\Groups\Genetics\PROJECTS\PCSK9.

The indirect LDL-C values and subjects ages were obtained from the CKB snapshot 10.

The subject gender information was extracted from the genotype data set.

Subject stratification was based on the file GWAS\_SNPdata\_samples.xlsx in the directory K:\kadoorie\Groups\Genetics\Data Archive\Project Sample Lists\Lists\

1. PCA (Principal Component Analysis)

PCA components are often employed as covariates in association studies to account for population structure. A LD (Linkage Disequilibrium)-free subset of 120201 SNPs was selected for the purpose. Before PCA we used plink (version 1.90) to estimate pairwise IBD (identity-by-descent) distances between all pairs of subjects. Pairs of subjects with IBD values higher than 0.05 were considered as related.

We used an iterative approach to remove the relatives. Given the network of related individuals, the degrees of connections were calculated. We rejected the most connected node (subject), before updated the degrees. The process was iterated till no subjects were related to each other.

6990 subjects were discharged. PCA was performed using the rest and projected onto the discharged ones later. plink and eigensoft produced very similar results. We used the components calculated by plink.

1. Stratification

Subjects were grouped into 6 strata according to their ascertainment and the type of LDL-C measures available. For Phase 1 subjects, LDL-C values were selected in the following order of priority: direct baseline (preferred); direct first resurvey; direct second resurvey; indirect. For LDL-C data not from baseline, subjects will be excluded if they have had an incident CVD event prior to the date on which the sample/measurement was taken.

|  |  |  |  |
| --- | --- | --- | --- |
| **stratum** | **Ascertainment** | **size** | **LDL-C type** |
| stratum\_1 | ICH | 4762 | direct |
| stratum\_2 | IS | 5210 | direct |
| stratum\_3 | SAH | 167 | direct |
| stratum\_4 | MI/IHD | 1265 | direct |
| stratum\_5 | Control | 6687 | direct |
| stratum\_6 | unselected/resurvey 2 | 4177 | indirect |

1. Linear regression

We employed plink for genome-wide linear regression. Genetic variants were regressed against raw LDL-C levels in each stratum. Covariates including the region code, sex, age and the 10 leading principal components were utilized. For the smallest stratum (stratum 3), we used only sex, age and the 2 leading components.

|  |
| --- |
| Stratum 1, raw LDL-C |
| Stratum 2, raw LDL-C |
| Stratum 3, raw LDL-C |

|  |
| --- |
| Stratum 4, raw LDL-C |
| Stratum 5, raw LDL-C |
| Stratum 6, raw LDL-C (indirect) |

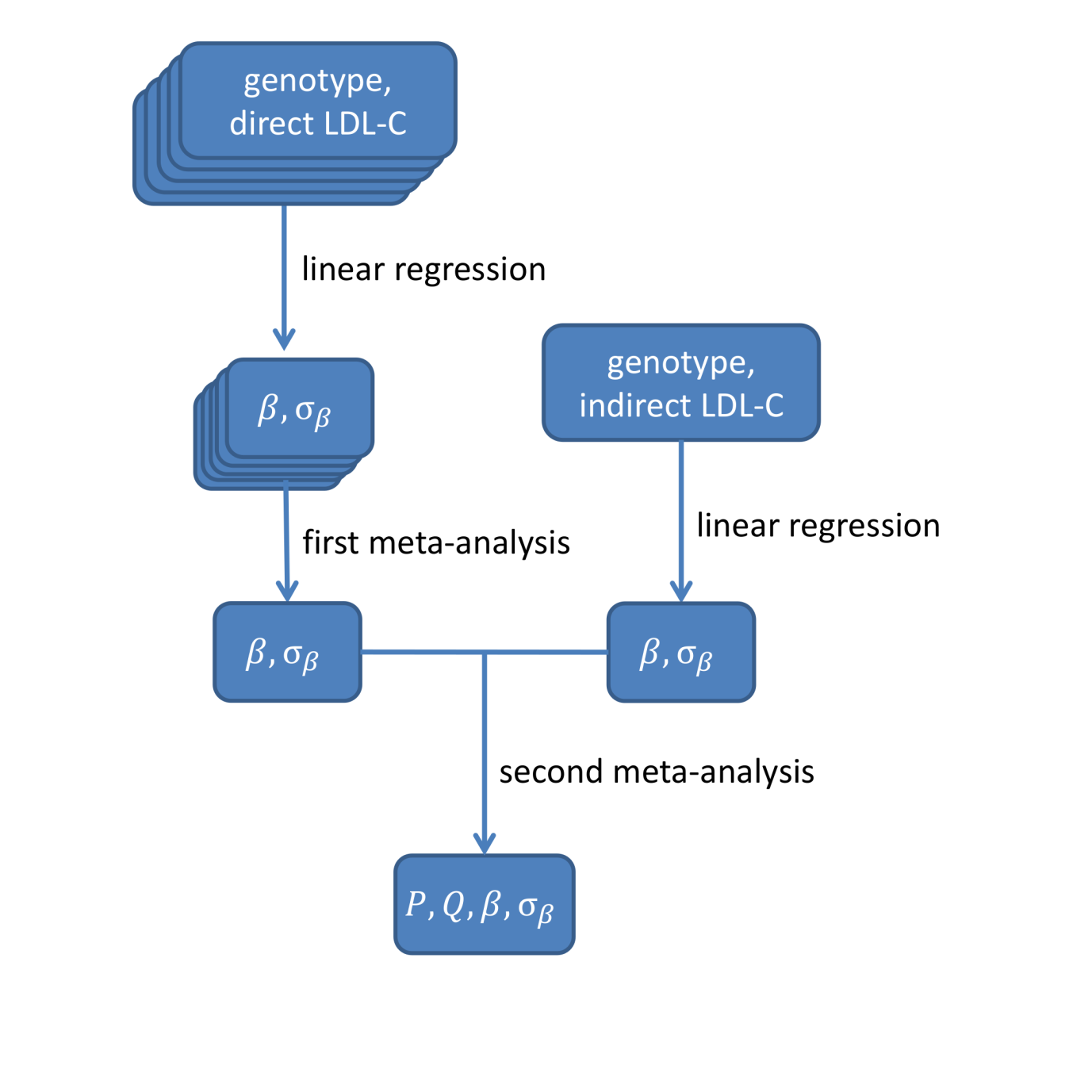
1. meta-analysis

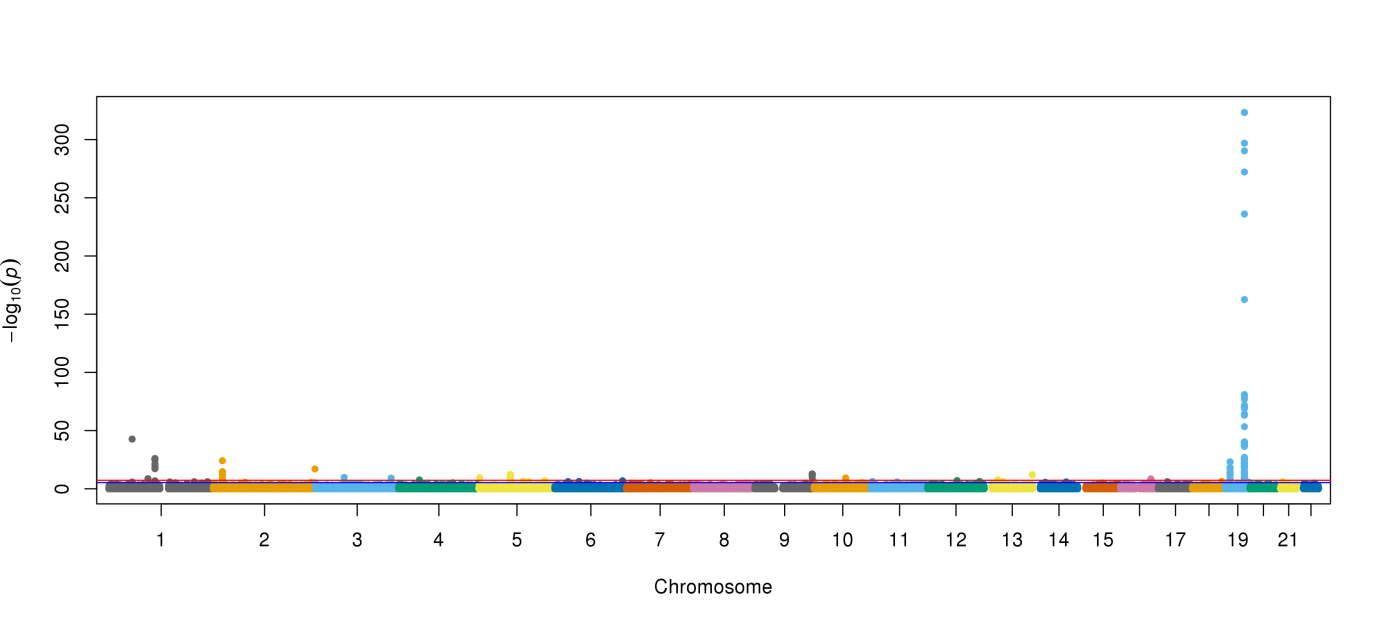
Genome-wide meta-analysis was performed using Metal.

Because of the heterogeneity between the direct and indirect LDL-C values, two meta-analyses were executed.

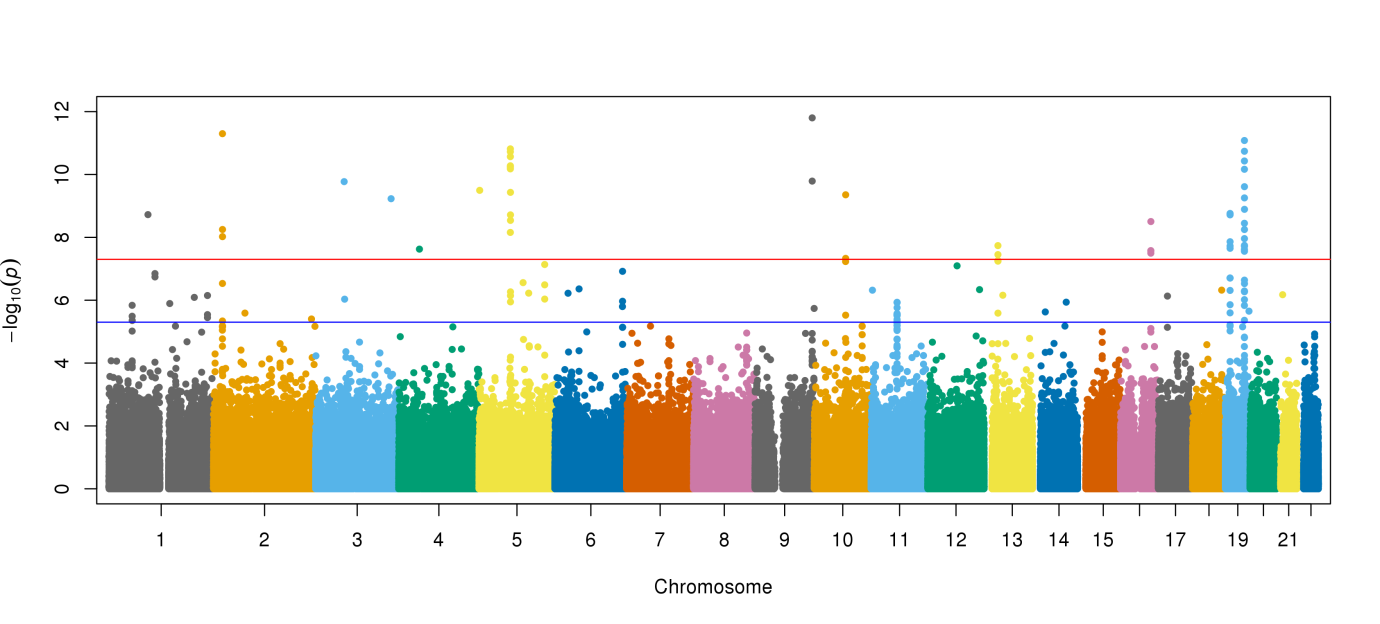
First we combined the beta and the standard error of beta estimated in the linear regression of the first five strata, using the inverse-variance-weighted fixed-effects scheme. The direct LDL-C meta-analysis output was then combined with the linear regression results of stratum 6 applying the same scheme.

The validity of this approach was checked by assessing heterogeneity (Cochran’s Q) between the two sets of results.





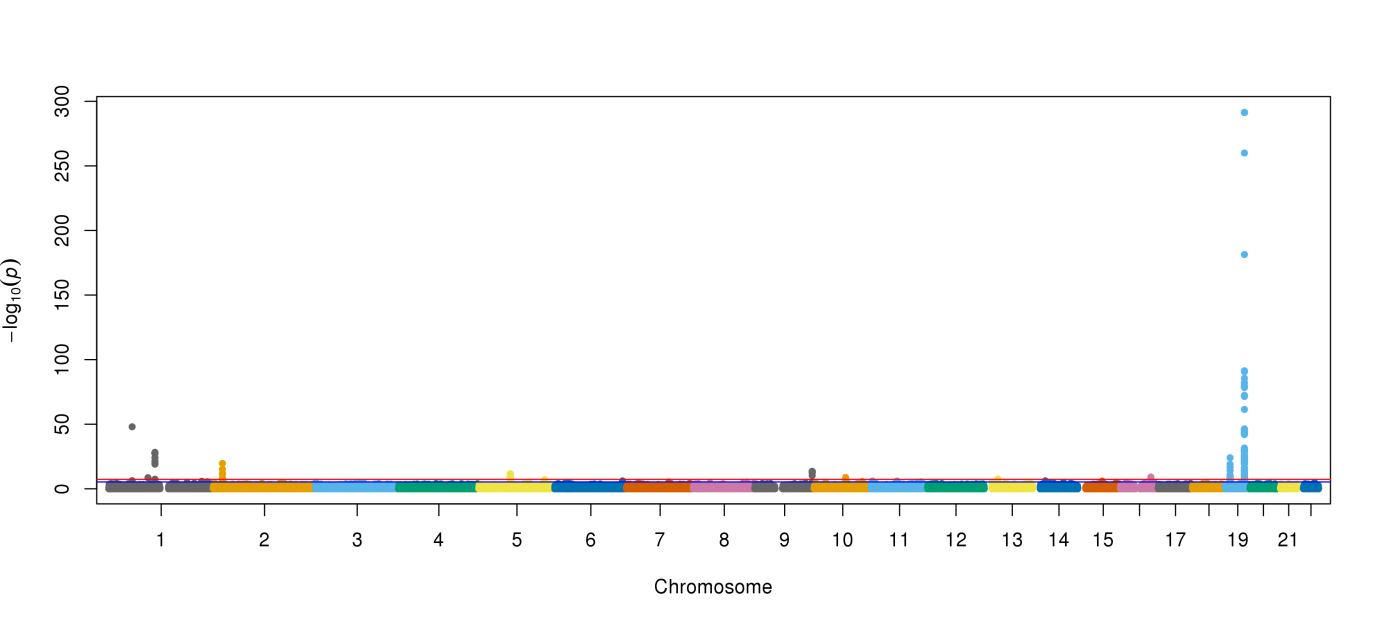
The group of the most highly-associated SNPs were located around the well-known gene APOE (Apolioprotein E).The association between this gene and LDL-C levels has been observed in many studies.



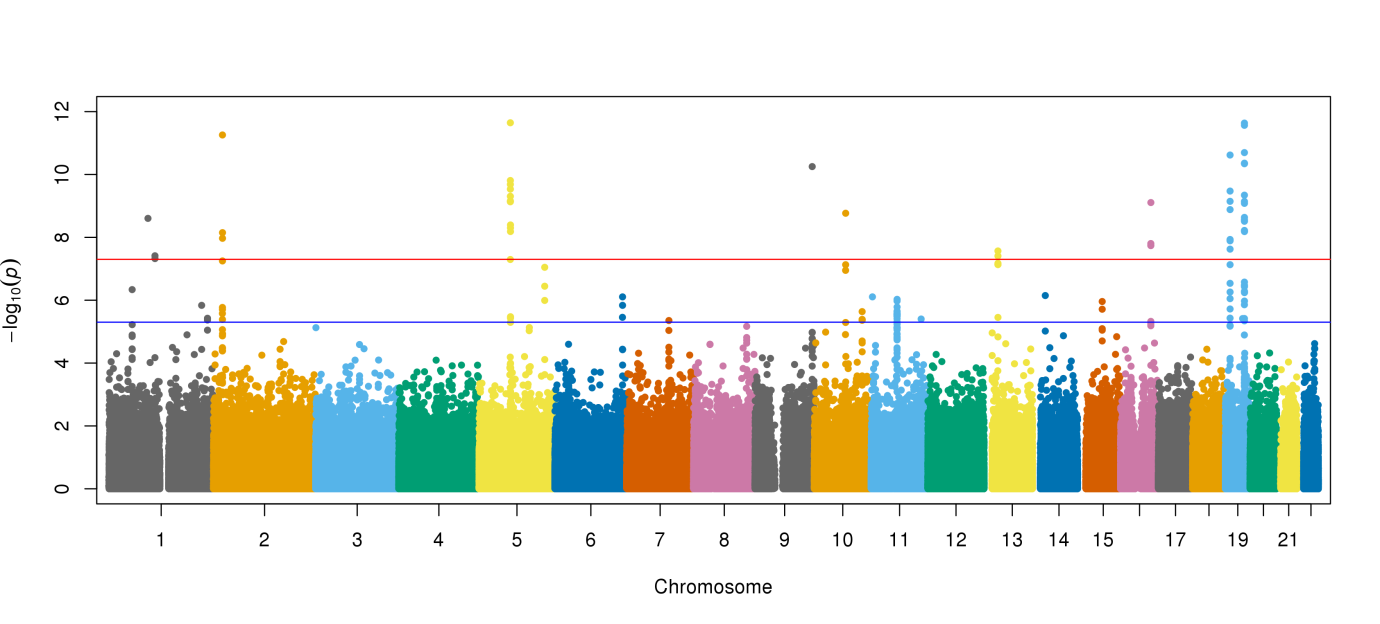
The lower part of this Manhattan plot shows other significant hits across the genome.

1. RINT (Rank Inverse-Normal Transformed)

In each stratum, the raw LDL-C values were regressed against region, sex and age. We rank inverse-normal transformed the residuals to obtained phenotype sets following the standard normal distribution. The sets of RINT values were concatenated into a single set. In the linear regression of these RINT LDL-C values, the 10 leading principal components were used as covariates.



Again, the APOE SNPs are highly significant. (The top 4 SNPs have P-values of 0 according to plink! We had to replace the value with the minimal positive number in the double precision float format to generate this plot.)



The lower part shows a very similar set of SNPs being significantly associated.

1. Conditional analysis

There are 160 variants in the PCSK9 region (chr1:55.4-56.0Mb, hg19). To find the independently-associated SNPs we iteratively performed the linear regression and the meta-analysis.

After each iteration, the SNP with the lowest P value found in the previous meta-analysis (or linear regression for RINT LDL-C) was to be added to the list of SNP hits. The association study would then be re-performed conditioned on these SNPs. We stopped when no additional SNP hits with P value less than 10-4 can be found or conditioned upon.

|  |
| --- |
| The original meta-analysis result. The vertical lines are 10-4 and 0.05 accordingly. |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **SNP** | **CHR** | **BP** | **MAF** | **BETA** | **SE\_BETA** | **Q** | **DIR** | **P** | | AX-83389438 | 1 | 55509585 | 1.36E-02 | -0.386 | 2.79E-02 | 0.9749 | -- | 1.92E-43 | | AX-39912161 | 1 | 55513061 | 0.1947 | -0.0391 | 8.10E-03 | 0.7648 | -- | 1.45E-06 | | AX-31642001 | 1 | 55517883 | 2.87E-01 | -0.0331 | 7.10E-03 | 0.5008 | -- | 3.23E-06 | | AX-11576926 | 1 | 55630151 | 0.1165 | 0.0465 | 1.01E-02 | 0.09788 | ++ | 4.42E-06 | | AX-31642169 | 1 | 55521242 | 0.2477 | -0.0331 | 7.50E-03 | 0.334 | -- | 9.60E-06 | | AX-31641677 | 1 | 55509939 | 0.06613 | 0.0508 | 0.0129 | 0.5738 | ++ | 8.51E-05 | | AX-11541856 | 1 | 55529187 | 0.06007 | -0.0525 | 0.0135 | 0.9953 | -- | 9.53E-05 | |

|  |
| --- |
| The second round of meta-analysis, conditioned on AX-83389438 |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **SNP** | **CHR** | **BP** | **MAF** | **BETA** | **SE\_BETA** | **Q** | **DIR** | **P** | | AX-31642001 | 1 | 55517883 | 0.2873 | -0.0408 | 7.10E-03 | 0.5588 | -- | 9.94E-09 | | AX-39912161 | 1 | 55513061 | 0.1947 | -0.0457 | 8.00E-03 | 0.7961 | -- | 1.24E-08 | | AX-31642169 | 1 | 55521242 | 0.2477 | -0.0401 | 7.50E-03 | 0.3943 | -- | 7.90E-08 | | AX-11576926 | 1 | 55630151 | 0.1165 | 0.0521 | 1.00E-02 | 0.1211 | ++ | 2.10E-07 | | AX-51209582 | 1 | 55498949 | 0.3514 | 0.03 | 6.70E-03 | 0.7234 | ++ | 8.52E-06 | | AX-31641243 | 1 | 55498982 | 0.1261 | -0.0427 | 9.80E-03 | 0.6374 | -- | 1.24E-05 | | AX-31641677 | 1 | 55509939 | 0.06613 | 0.0557 | 1.29E-02 | 0.5771 | ++ | 1.68E-05 | | AX-39912159 | 1 | 55512995 | 0.07861 | -0.051 | 1.20E-02 | 0.8625 | -- | 1.97E-05 | |

|  |
| --- |
| The third round of meta-analysis, conditioned on AX-83389438 and AX-31642001 |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **SNP** | **CHR** | **BP** | **MAF** | **BETA** | **SE\_BETA** | **Q** | **DIR** | **P** | | AX-11541856 | 1 | 55529187 | 0.06007 | -0.066 | 1.37E-02 | 0.8062 | -- | 1.51E-06 | | AX-39912161 | 1 | 55513061 | 0.1947 | -0.0329 | 0.0089 | 0.964 | -- | 0.0002102 | |

|  |
| --- |
| The fourth round of meta-analysis, conditioned on AX-83389438, AX-31642001 and 11541856 |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **SNP** | **CHR** | **BP** | **MAF** | **BETA** | **SE\_BETA** | **Q** | **DIR** | **P** | | AX-31642169 | 1 | 55521242 | 0.2477 | -0.043 | 0.0116 | 0.4119 | -- | 0.000213 | |

|  |
| --- |
| The original RINT linear regression result. The vertical line is 10-4. |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **CHR** | **SNP** | **BP** | **A1** | **NMISS** | **BETA** | **SE** | **STAT** | **P** | | 1 | AX-83389438 | 55509585 | T | 22265 | -0.6048 | 0.04112 | -14.71 | 9.68E-49 | | 1 | AX-39912161 | 55513061 | T | 22236 | -0.06014 | 0.01192 | -5.044 | 4.60E-07 | | 1 | AX-11576926 | 55630151 | C | 22191 | -0.06745 | 0.0149 | -4.528 | 5.98E-06 | | 1 | AX-31642001 | 55517883 | C | 22198 | -0.04585 | 0.01051 | -4.365 | 1.28E-05 | | 1 | AX-11541856 | 55529187 | G | 22226 | 0.08612 | 0.01984 | 4.341 | 1.42E-05 | | 1 | AX-31641677 | 55509939 | C | 22212 | -0.0789 | 0.01907 | -4.136 | 3.54E-05 | | 1 | AX-31642169 | 55521242 | T | 22213 | -0.04547 | 0.01102 | -4.126 | 3.71E-05 | | 1 | AX-31641243 | 55498982 | A | 22180 | -0.05734 | 0.01439 | -3.986 | 6.73E-05 | | 1 | AX-39912159 | 55512995 | A | 22229 | -0.06977 | 0.01764 | -3.956 | 7.65E-05 | |

|  |
| --- |
| The second round RINT linear regression, conditioned on AX-83389438 |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **CHR** | **SNP** | **BP** | **A1** | **NMISS** | **BETA** | **SE** | **STAT** | **P** | | 1 | AX-39912161 | 55513061 | T | 22233 | -0.07046 | 0.01188 | -5.928 | 3.10E-09 | | 1 | AX-31642001 | 55517883 | C | 22195 | -0.0581 | 0.01049 | -5.542 | 3.03E-08 | | 1 | AX-31642169 | 55521242 | T | 22210 | -0.05674 | 0.01099 | -5.162 | 2.47E-07 | | 1 | AX-11576926 | 55630151 | C | 22188 | -0.07637 | 0.01484 | -5.147 | 2.67E-07 | | 1 | AX-31641243 | 55498982 | A | 22177 | -0.06568 | 0.01433 | -4.584 | 4.59E-06 | | 1 | AX-31641677 | 55509939 | C | 22209 | -0.0865 | 0.01899 | -4.555 | 5.26E-06 | | 1 | AX-51209582 | 55498949 | C | 22176 | -0.04496 | 0.009939 | -4.524 | 6.10E-06 | | 1 | AX-39912159 | 55512995 | A | 22226 | -0.07815 | 0.01756 | -4.45 | 8.61E-06 | | 1 | AX-11629248 | 5.57E+07 | G | 22249 | -0.04597 | 0.01162 | -3.954 | 7.71E-05 | |

|  |
| --- |
| The third round RINT linear regression, conditioned on AX-83389438 and AX-39912161 |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **CHR** | **SNP** | **BP** | **A1** | **NMISS** | **BETA** | **SE** | **STAT** | **P** | | 1 | AX-11576926 | 5.56E+07 | C | 22156 | -0.0596 | 0.01521 | -3.918 | 8.95E-05 | | 1 | AX-11541856 | 55529187 | G | 22191 | 0.07542 | 0.01974 | 3.82 | 0.0001338 | |

|  |
| --- |
| The fourth round of RINT regression, conditioned on AX-83389438, AX-39912161 and AX-11576926 |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **CHR** | **SNP** | **BP** | **A1** | **NMISS** | **BETA** | **SE** | **STAT** | **P** | | 1 | AX-11541856 | 55529187 | G | 22115 | 0.06655 | 0.01986 | 3.351 | 0.0008058 | |

SNPs selected in the two association studies

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SNP | RSID | BP | BETA | SE\_BETA | META\_1\_Q | META\_2\_Q | DIR |
| AX-83389438 | rs151193009 | 55509585 | -0.3860 | 0.0279 | 0.938 | 0.975 | ----- -- |
| AX-39912161 | rs10888897 | 55513061 | -0.0391 | 0.0081 | 0.779 | 0.765 | ----- -- |
| AX-31642001 | rs624612 | 55517883 | -0.0331 | 0.0071 | 0.296 | 0.501 | ----- -- |
| AX-11576926 | rs6663252 | 55630151 | 0.0465 | 0.0101 | 0.671 | 0.098 | +++++ ++ |
| AX-31642169 | rs471705 | 55521242 | -0.0331 | 0.0075 | 0.886 | 0.334 | ----- -- |
| AX-31641677 | rs4275490 | 55509939 | 0.0508 | 0.0129 | 0.864 | 0.574 | ++-++ ++ |
| AX-11541856 | rs505151 | 55529187 | -0.0525 | 0.0135 | 0.487 | 0.995 | ----- -- |
| AX-31641243 |  | 55498982 | -0.0374 | 0.0098 | 0.154 | 0.620 | ----- -- |
| AX-39912159 | rs2495478 | 55512995 | -0.0455 | 0.0120 | 0.886 | 0.846 | --+-- -- |
| AX-51209582 |  | 55498949 | 0.0227 | 0.0067 | 0.853 | 0.722 | +++++ ++ |
| AX-31642257 |  | 55522425 | -0.0415 | 0.0127 | 0.381 | 0.637 | --+-- -- |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNP | MAF | Function | META\_P | RINT\_P | META\_COND\_P | RINT\_COND\_P |
| AX-83389438 | 0.0136 | R93->C | 1.92E-43 | 9.68E-49 | 1.92E-43 | 9.68E-49 |
| AX-39912161 | 0.1947 | intronic | 1.45E-06 | 4.60E-07 |  | 3.10E-09 |
| AX-31642001 | 0.2873 | intronic | 3.23E-06 | 1.28E-05 | 9.94E-09 |  |
| AX-11576926 | 0.1165 | intronic | 4.42E-06 | 5.98E-06 |  | 8.95E-05 |
| AX-31642169 | 0.2477 | intronic | 9.60E-06 | 3.71E-05 | 2.13E-04 |  |
| AX-31641677 | 0.0661 | intronic | 8.51E-05 | 3.54E-05 |  |  |
| AX-11541856 | 0.0601 | E670->G | 9.53E-05 | 1.42E-05 | 1.51E-06 |  |
| AX-31641243 | 0.1261 |  | 1.25E-04 | 6.73E-05 |  |  |
| AX-39912159 | 0.0786 | intronic | 1.41E-04 | 7.65E-05 |  |  |
| AX-51209582 | 0.3514 |  | 7.57E-04 | 7.66E-04 |  |  |
| AX-31642257 | 0.0690 |  | 1.09E-03 | 5.54E-04 |  |  |

1. Backwards feature selection

The selected SNPs were filtered using the stepAIC function in the MASS library of R. Dosages of the 6 SNPs were calculated. We used the backward stepwise selection on the linear models of RINT LDL-C values with SNP dosages. 5 SNPs were in the last model. The only one ejected was AX-11576926.

1. Some side notes

The leading SNP hit AX-83389438 (rs151193009) was found only in the CHB, CHS, JPT and KHV cohorts in the 1000 Genomes Project (phase 3) data set, with the MAF ranges between 1-2%.

Two SNPs in this region are also available in the SNP 384 panel. (The tables to be updated)

AX-11150762 (rs11206510)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| META | CHR | BP | MAF | a1 | a2 | meta\_beta | meta\_se | meta\_q | meta\_dir | meta\_p |
| raw\_dir\_ldl\_meta\_0 | 1 | 55496039 | 0.0564 | T | C | 0.0308 | 0.0151 | 0.4555 | +++-+ | 0.04063 |
| raw\_dir\_ldl\_meta\_1 | 1 | 55496039 | 0.0564 | T | C | 0.0351 | 0.015 | 0.4839 | +++-+ | 0.01933 |
| raw\_dir\_ldl\_meta\_2 | 1 | 55496039 | 0.0564 | T | C | 0.0287 | 0.0151 | 0.4496 | +++-+ | 0.05688 |
| raw\_dir\_ldl\_meta\_3 | 1 | 55496039 | 0.0564 | T | C | 0.0245 | 0.0151 | 0.4445 | +++-+ | 0.1048 |
| Rint\_dir\_ldl\_meta\_0 | 1 | 55496039 | 0.0564 | T | C | 0.0444 | 0.0228 | 0.5506 | +++-+ | 0.05113 |
| rint\_dir \_ldl\_meta\_1 | 1 | 55496039 | 0.0564 | T | C | 0.0514 | 0.0227 | 0.5647 | +++-+ | 0.02332 |
| rint\_dir \_ldl\_meta\_2 | 1 | 55496039 | 0.0564 | T | C | 0.0132 | 0.0238 | 0.5117 | +-+-+ | 0.579 |
| rint\_dir \_ldl\_meta\_3 | 1 | 55496039 | 0.0564 | T | C | 0.0166 | 0.0238 | 0.5042 | +-+-+ | 0.4857 |

AX-39911995 (rs2479409)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| META | CHR | BP | MAF | a1 | a2 | meta\_beta | meta\_se | meta\_q | meta\_dir | meta\_p |
| raw\_dir\_ldl\_meta\_0 | 1 | 55504650 | 0.3181 | A | G | -0.0073 | 0.0075 | 0.5959 | ---+- | 0.3301 |
| raw\_dir\_ldl\_meta\_1 | 1 | 55504650 | 0.3181 | A | G | -0.0156 | 0.0075 | 0.5641 | ---+- | 0.0378 |
| raw\_dir\_ldl\_meta\_2 | 1 | 55504650 | 0.3181 | A | G | -0.0124 | 0.0076 | 0.5271 | ---+- | 0.1011 |
| raw\_dir\_ldl\_meta\_3 | 1 | 55504650 | 0.3181 | A | G | -0.0148 | 0.0076 | 0.4978 | ---+- | 0.05061 |
| rint\_dir \_ldl\_meta\_0 | 1 | 55504650 | 0.3181 | A | G | -0.0117 | 0.0114 | 0.7847 | ---+- | 0.3039 |
| rint\_dir \_ldl\_meta\_1 | 1 | 55504650 | 0.3181 | A | G | -0.0249 | 0.0113 | 0.7438 | ---+- | 0.02813 |
| rint\_dir \_ldl\_meta\_2 | 1 | 55504650 | 0.3181 | A | G | -0.0096 | 0.0117 | 0.7401 | ---+- | 0.4134 |
| rint\_dir \_ldl\_meta\_3 | 1 | 55504650 | 0.3181 | A | G | -0.0118 | 0.0117 | 0.7811 | --++- | 0.3156 |