# Identification of variants in PCSK9 affecting LDL-C levels

## Introduction

The *PCSK9* gene product mediates breakdown of liver LDL receptors, so that blocking PCSK9 action increases the levels of LDL receptors and the rate of removal of LDL from the bloodstream. Ongoing phase III clinical trials are assessing whether drugs targeting PCSK9, that have been shown to efficaciously reduce LDL cholesterol (LDL-C) concentrations, also result in a reduction in risk of cardiovascular events. We seek to complement these trials by using genetic variants in *PCSK9* that affect levels of LDL-C, mimicking the action of these drugs, in Mendelian randomisation analyses in the China Kadoorie Biobank (CKB) prospective study.

Some *PCSK9* variants identified in Europeans as associated with LDL cholesterol (LDL-C) are absent from or present at low frequency in Chinese.. Given that existing SNPs identified in Europeans may tag the causal variant on the background of different linkage disequilibrium, and rare/functional variants that are monomorphic in Europeans may be polymorphic in Asian individuals, we seek to identify additional *PCSK9* variants associated with LDL-C levels in Chinese populations.

## Study Population

Genome-wide genotyping data are available for ~33k CKB participants. These subjects were selected in two phases. Phase 1 excluded prevalent self-reported cardiovascular disease, cancer and/or statin use, and comprised cases and controls ascertained (although their status may have been revised subsequently) according to the first-reported incident event of:

* haemorrhagic stroke (ICH, 5,019)
* ischaemic stroke (IS, 5,663)
* subarachnoid haemorrhage (SAH, 455)
* myocardial infarction or fatal ischaemic heart disease (MI/IHD, 1,761)
* controls with no incident cardiovascular disease, matched to ICH cases for age/gender/RC (10,038)

There were in addition some subjects who were genotyped as a result of linkage errors that have subsequently been corrected. These can be treated as unselected controls.

Phase 2 included 5,081 subjects who attended the second resurvey. These subjects had no prior exclusions or other ascertainment, except that some individuals had already been selected for Phase 1 and, therefore, were absent from Phase 2. (GWAS data are also available for 5,376 COPD exacerbation cases not attending resurvey 2, but these will not be included in this analysis because there is no LDL-C data for these subjects.) For consistency with phase 1, subjects in phase 2 with prevalent self-reported cardiovascular disease, cancer and/or statin use will be excluded from the analysis.

Subjects will be excluded from analysis according to standard GWAS QC (elevated heterozygosity, call rate, gender mismatch etc.). For the initial analysis, there will be no exclusions on the basis of relatedness.

## LDL-C measurements

The majority of study subjects with GWAS data also have measurements of LDL-C from one or both of two separate sources: (i) “**indirect**” calculation according to the Friedewald equation, based on measurements of total cholesterol, HDL-C and triglycerides from whole blood during the second resurvey (measured using a portable handheld lipidometer); (ii) “**direct**” measurements (in a biochemistry laboratory using conventional instruments), from serum samples derived from blood taken at baseline, first resurvey and/or second resurvey.

Where subjects have more than one LDL-C measurement, one dataset for that subject will be used. For Phase 1 subjects, this will be selected in the following order of priority: direct baseline (preferred); direct first resurvey; direct second resurvey; indirect. For Phase 2 and “unselected” subjects, indirect measures will be used. All covariates (e.g. age) will be taken from the same survey as the LDL-C data. 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.

Table 1 shows the number of subjects in each ascertainment category and the corresponding source(s) of LDL-C data, to date (before any exclusions for prior incident CVD events). Bold text/shading shows the data to be used in the analysis. The 5,070 indirect measurements will be analysed as a single group. They are primarily from control or unselected subjects and the 81 additional datasets from ascertained cases will not cause any appreciable bias in this analysis – in fact, they will reverse the depletion of cases bringing this group closer to the composition of an unselected population sample.

## SNP data

Previous analyses (Willer, et al., 2013) have identified a broad recombination interval (including the *PCSK9* structural gene) within which many SNPs show association with LDL-C at genome wide significance, but there are also additional SNPs lying outside this region (see Figure 1).

Based on this, we will test associations for all SNPs that pass initial QC (clustering parameters, call rate, plate and batch effects) in the range chr1:55.45-56.0Mb (hg19). Before QC this comprises 157 SNPs (assayed using 171 probesets). No SNPs will be excluded on the basis of low MAF since we hope to find strong instruments for MR analysis and these could potentially include rare variants. To avoid unnecessarily stringent QC, other QC criteria (e.g. Hardy-Weinberg equilibrium, manual review of clustering) will take place following the analysis, for only those SNPs showing significant LDL-C associations.

## Primary association analysis

For all analyses, subjects will initially be stratified according to (i) assay type (indirect/direct measurement); and (ii) according to ascertainment (ICH, IS, SAH, MI/IHD, controls). Each individual analysis will fit a per-allele additive effect model using plink v1.90, with covariates as follows:

* region code (as 9 dummy binary variables)
* sex
* age at time of measurement
* leading principal components

As recent studies have shown that fasting time has minimal impact on LDL-C levels, fasting time will not be included as a covariate.

Outputs for each SNP from each analysis will include:

* Sample size
* Effect allele
* Allele effect size and standard error (in mmol/l or SDs)
* P-value
* Between strata heterogeneity (Cochran’s Q), where relevant

The results overall will be visually represented using “LocusZoom” style plots.

### Analysis of untransformed LDL-C

Association analyses will be conducted within each stratum. Results for analyses of “direct” LDL-C measurements will be combined in an inverse-variance-weighted fixed-effects meta-analysis, using METAL.

From analysis of 1,119 resurvey 2 attendees for whom LDL-C data were available from both direct and indirect methods, it is clear that there is good correlation (R2=0.52) between the two methods. The regression line has a slope near to unity, but with a non-zero intercept. Therefore, it can be hypothesised that analyses of data acquired using the two methods should give effect sizes that are similar.

Thus, the results of the “direct” meta-analysis and of the “indirect” analysis will be combined in a further fixed-effects inverse-variance-weighted meta-analysis to give effect sizes in mmol/l. The validity of this approach will be checked by assessing heterogeneity (Cochran’s Q) between the two sets of results.

### Analysis of standardised LDL-C

The association results from each individual stratum will be transformed to provide effect sizes scaled to the SD for LDL-C in that stratum. Fixed-effects inverse-variance-weighted meta-analysis across all strata (both direct and indirect) will then yield overall effect size estimates in SD units.

### Analysis of transformed LDL-C

For each stratum, LDL-C data will be regressed against region, sex and age, and the residuals will be rank inverse-normal transformed (RINT) to give a standard normal distribution. These sets of RINT data will be concatenated into a single data set, which will be used for a single association analysis including leading principal components (only) as covariates. This will yield overall effect size estimates in SD units.

### Sensitivity analyses

The impact of the following modifications to each analysis will be assessed:

* including age2 as covariate
* including genotyping batch id as covariate
* including fasting time as covariate
* including RC, sex, age as covariates for RINT analysis
* within “direct” LDL-C measurements, using an unstratified analysis with strata as covariates
* exclusion of related individuals

## Interpretation

It is expected that the different analytical approaches will yield similar results. If this is not the case, the results of sensitivity analyses will be used to determine which approach is to be relied upon.

In light of the prior knowledge that the PCSK9 locus is associated with LDL-C levels, in assessing significance of association we will apply a 5% FDR threshold (Benjamini–Hochberg).

The SNP with the lowest P-value will not necessarily be the best instrument (e.g. 2 SNPs with similar P-value but the one with the larger P-value having a large effect size). Where there is any doubt as to the lead SNP, patterns of LD between them and their positions relative to *PCSK9* and other genes in the region will be used to try to determine the better choice.

The primary aim of the proposed analysis is to find the best instrument for future Mendelian randomisation experiments and as such correction for inflation is not a vital prerequisite for identifying the lead SNP(s) to be used as instruments. However, if there is a question as to whether any SNPs exhibit significant association, we will repeat the analysis on a pruned genome-wide subset of SNPs and estimate inflation using the LD score approach (Bulik-Sullivan, et al., 2015).

## Identification of additional independent effects

Following identification of a lead SNP showing significant association with LDL-C, we will run repeat analyses conditioned on the lead variant. If a significant independent effect is identified, the new lead SNP will be added as a further covariate and process will be repeated, until no further independent effects are identified.

## Additional analyses

### European/Chinese comparison

To investigate whether our identified lead variant(s) for associations with LDL-C levels in Chinese are consistent with what has previously been identified in Europeans, we will meta-analyse with published results and test for heterogeneity. In the absence of significant heterogeneity, we will consider using the reported effect sizes from Europeans (e.g. if they are more accurate estimates).

To further elucidate any differences in LD matrices between the European and Chinese individuals, comparison of European and Chinese LD in the *PCSK9* region will be performed using varLD.

### Joint model

We will assess the association with LDL-C levels of a joint model incorporating all independently-associated SNPs. This will be conducted using gene scores derived from (a) unweighted allele count; (b) internal-weights derived from a (jackknifed) joint model of all associated SNPs; (c) external weights derived from European analyses.

### Imputation analysis

In the event that no significantly-associated individual SNPs are identified, we will carry out local imputation into the 1000 genomes EAS reference dataset supplemented (if possible) with the CONVERGE dataset, using SHAPEIT2 and IMPUTE2 or their successors. Association of imputed genotypes will be conducted as above, but using SNPTEST2.