New genes predisposing to coronary artery disease detected by genome-wide gene-based association analysis

I. Zorkoltseva[1], N. Belonogova[1], A. Shadrina[1], A. Kirichenko[1], Y. Tsepilov[1,2], T. Axenovich[1,2]

[1] Institute of Cytology and Genetics, Novosibirsk, Russia

[2] Novosibirsk State University, Novosibirsk, Russia

Background: To date genome-wide association analyses (GWAS) have identified more than 300 single nucleotide polymorphisms at 163 genetic loci associated with CAD. However, there is no full understanding about the causal genes for CAD and the mechanisms of their action.

Aim: We aimed to perform a post GWAS analysis to identify genes whose polymorphism may influence the risk of CAD and enhance our understanding of the genetic architecture of CAD.

Materials: GWAS summary statistics computed using UK Biobank White British individuals from the Lee Lab (https://www.leelabsg.org/) and the Gene ATLAS (http://geneatlas.roslin.ed.ac.uk/)

Phenotypes: Ischaemic Heart Disease (IHD) (PheCode 411) (N = 408 458)

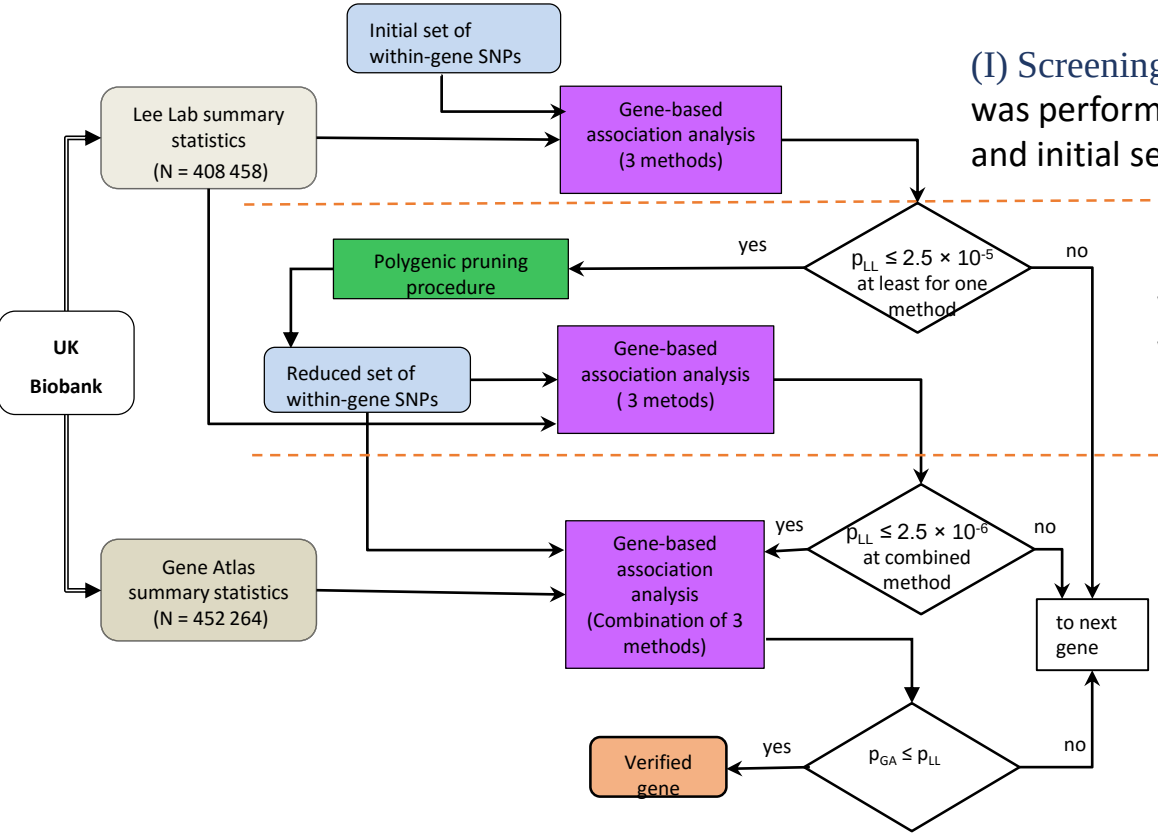
Coronary atherosclerosis (CA) (PheCode 411.4) (N = 397 126)

Methods: Three-stage gene-based association analysis was performed using:

- four sets of genetic variants within a gene differing in their protein coding properties
- three methods, SKAT-O, PCA, and ACAT-V, implemented in sumFREGAT package

A ‘polygene pruning’ procedure was used to eliminate the influence of strong GWAS signals outside the gene

The pipeline of three-stage gene-based association analysis for each phenotype and set of SNPs within the gene

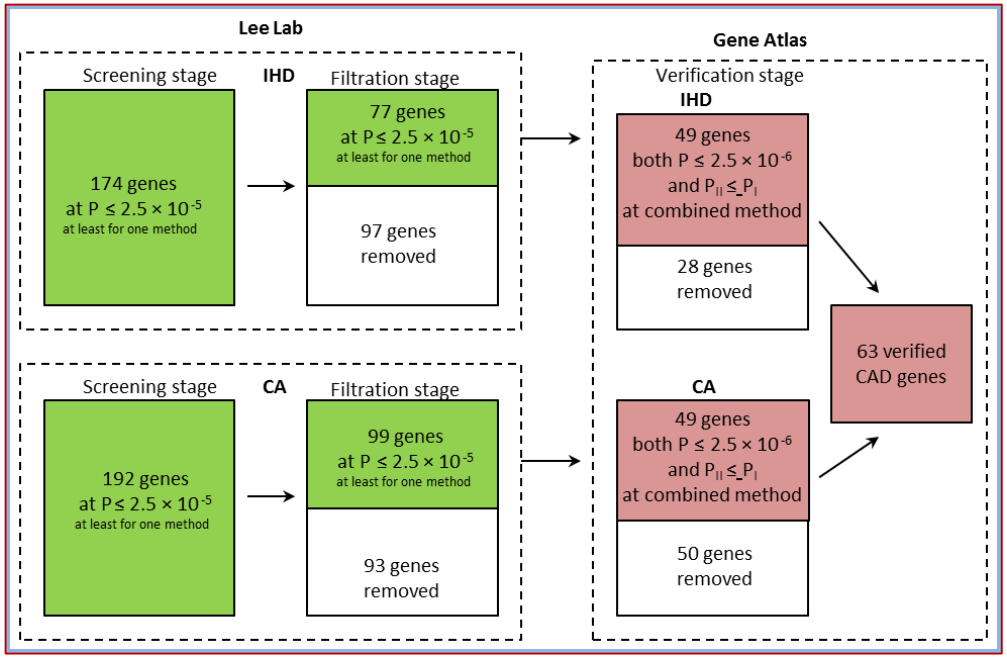


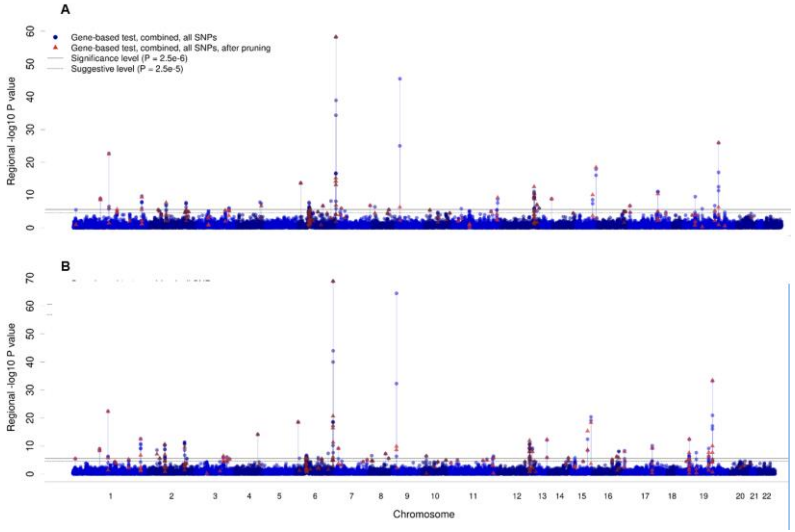
(I) Screening: a gene-based association analysis was performed using Lee Lab summary statistics and initial sets of SNPs.

(II) Filtration: a '*polygene pruning' procedure* was applied. Gene-based association analysis was repeated using the SNPs remaining after pruning.

(III) Verification: a gene-based association analysis was performed using Gene ATLAS summary statistics and the SNPs remaining after pruning.
Loci for which *p-values* in the extended Gene ATLAS dataset were lower than the corresponding *p-values* in the Lee Lab dataset were considered confirmed.

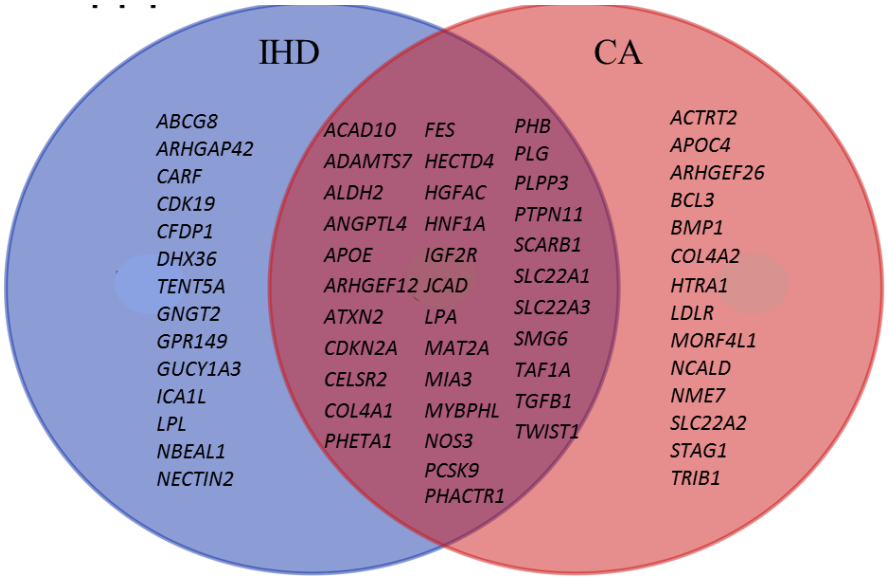
Results of different stages of gene-based association analysis





Manhattan plots showing the association of genes with IHD (A) and CA (B) in a genome-wide gene-based association study before (blue) and after (red) polygene pruning procedure. About half of genes were removed after pruning.

63 confirmed CAD-associated genes



- 26 genes did not contain any SNPs reaching the genome-wide significance threshold of $5.0 \cdot \times 10^{-8}$ in GWAS.
- The proportion of genes being closest to the top GWAS signals was 0.52.
- 17 genes were identified using protein-coding SNPs sets. Ten of them included 25 potentially pathogenic variants.
- 44 genes were identified using a set of non-coding within-gene variants. For 10 out of them we found evidence of pleiotropy with gene expression.
- 3 genes, CDK19, NCALD, and ARHGEF12, are new.

Conclusion: Using the UK Biobank GWAS summary statistics in gene-based analysis, we found 63 genes significantly associated with CAD due to their within-gene polymorphisms. Many of these genes are well known. Some known CAD genes (such as *FURIN* and *SORT1*) did not show the gene-based association because their variants had low GWAS signals or gene-based association was inflated by the strong GWAS signal outside the gene. For several known CAD genes, we demonstrated that their effects can be explained not only or not at all by their own variants but by the variants within the neighboring genes controlling their expression. Using a number of bioinformatics techniques, we suggested potential mechanisms underlying gene-CAD associations. Three genes, *CDK19, NCALD,* and *ARHGEF12,* were not previously associated with CAD. The role of these genes should be clarified in further studies.

Acknowledgements: This work was supported by the Federal Agency of Scientific Organizations via the Institute of Cytology and Genetics, SB RAS (project 0259-2021-0009/AAAA-A17-117092070032-4) and the Russian Foundation for Basic Research (project 20-04-00464).