# Allelic associations in 91 proteins

Authors to be added here

## Contact information

For general questions about SCALLOP, please contact Anders Malarstig (anders.malarstig@ki.se). For technical issues about TRYGGVE, please contact Lasse Folkersen (lasfol@cbs.dtu.dk).

For questions regarding SCALLOP/INF, please contact Jing Hua Zhao (jhz22@medschl.cam.ac.uk) and James Peters (jp549@medschl.cam.ac.uk).

***Last updated 14/8/2019***

## Abstract

The fundamental role of proteins in biological processes can hardly be overstated. Although heavily featured in research on diseases and therapeutic targets, the introduction of large-scale genomic data in these is relatively recent and shown great promise into the underlying biology in relation to disease and treatment. Here we report a concerted effort on genomewide association studies of proteins from discovery cohorts within the SCALLOP consortium involving 15,335 individuals as well as replication cohorts, all using 1000 Genomes imputed genotypes and 91 proteins from Olink/Inflammation panels. Through genomewide meta-analysis we have identified 146 loci associated with protein levels. Regional associations involving these loci, colocolisation as well as Mendelian randomization analysis into the causal variants, links with gene expression levels and with disease and other outcomes are reported. While findings related to proteins such as IL6 and OPG were seen in the literature, a lot of them are novel; our work will contribute to the wealth of the literature and warrant future efforts.

## Introduction

The fundamental role of proteins in biological processes can hardly be overstated. Although they widely featured in diseases and therapeutic investigations, the addition of large-scale genomic data is relatively recent as has been reported by Sun et al. (2018) from the INTERVAL study. Here we assembled data from protein-wide genomic studies (pGWAS) in the SCAndinavian coLLaboration for Olink plasma Protein genetics (SCALLOP) consortium, including a study of 966 individuals with sequence data all with Olink/inflammation proteins consisting of multiplex biomarker panels of 92 assays for proteins related to specific biological processes for protein quantitative trait loci (pQTLs).

The panel was designed to study inflammation but also part of other panels in the Olink platforms. This is in line with the notion that inflammation is vital in immune response such as removal of harmful stimuli and initiation of the healing process, but that growing evidence indicates it is indispensable for a wide range of pathological processes and/or disorders which include inflammatory bowel disease, asthma and dermatological conditions, multiple cardiovascular and neurological diseases, as well as cancer.

We conducted comprehensive *in silico* experiments leading to identification of pQTLs, and followed by their functional annotation including cis-/trans- effects, pleiotropic effects, colocalisation with gene expression as well as assessment of the causal role on a range of disease outcomes such as CHD through Mendelian randomization (MR). We have made our implementation widely available. Our approaches and findings should contribute to a greater understanding of the genomics of inflammation-related proteins in these areas.

## Results

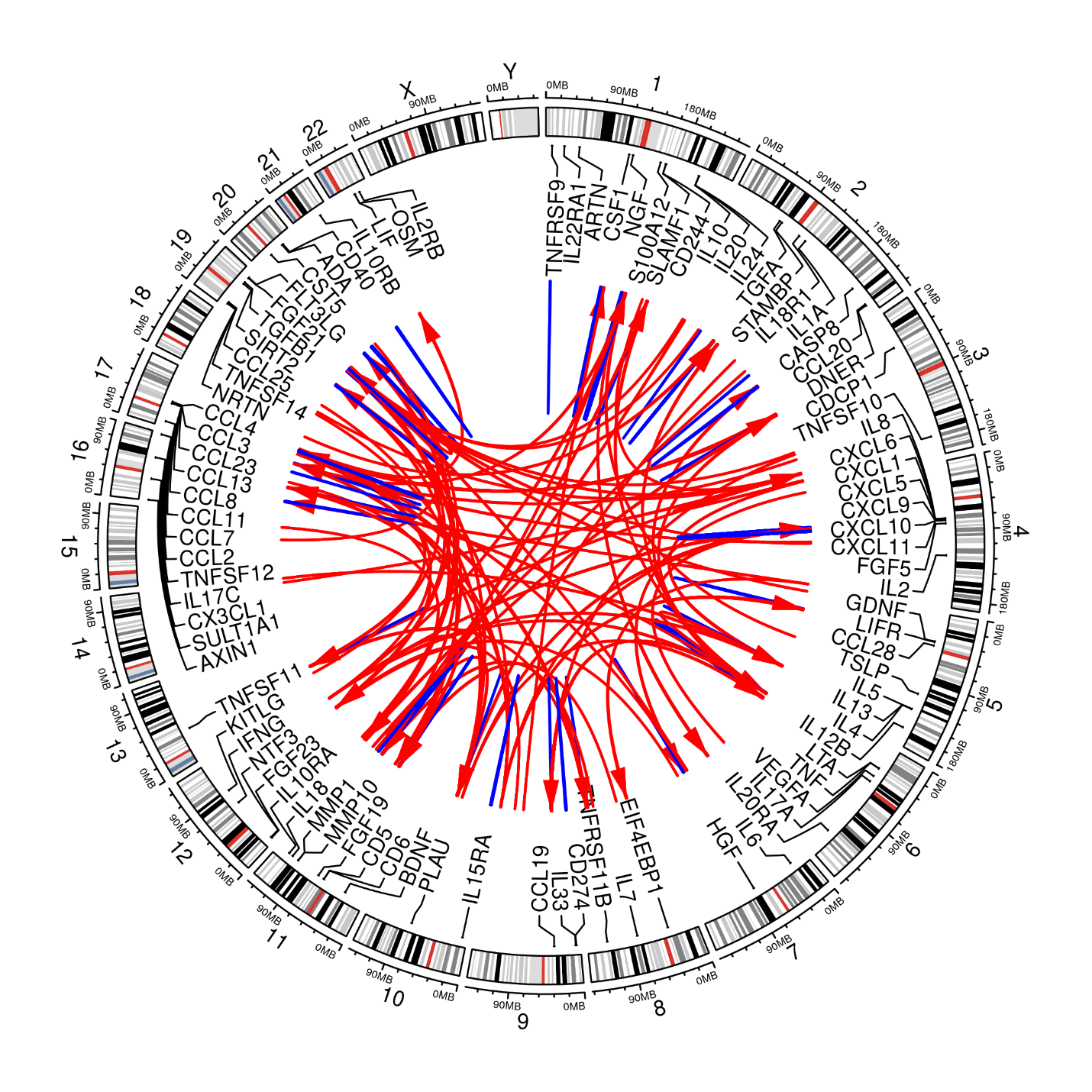
**Independent signals**

Based on a distance approach, a total of 146 independent signals were identified. These signals were based on 69 (20 only cis, 13 trans only and 36 both) proteins. The signals are further classified into 55 cis and 91 trans, signals respectively. GCTA cojo-slct analysis on these regions led to 278 primary and 65 secondary signals totaling 343. (Table 1 and Figure 1).

**Table 1. Classification of sentinel signals**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cis** | **Trans** | **Total** |
| Primary | 174 | 104 | 278 |
| Secondary | 47 | 18 | 65 |
| Total | 221 | 122 | 343 |

**Figure 1. cis/trans signals**. Signals are shown according to chromosomal positions with blue lines indicate cis and red lines trans.

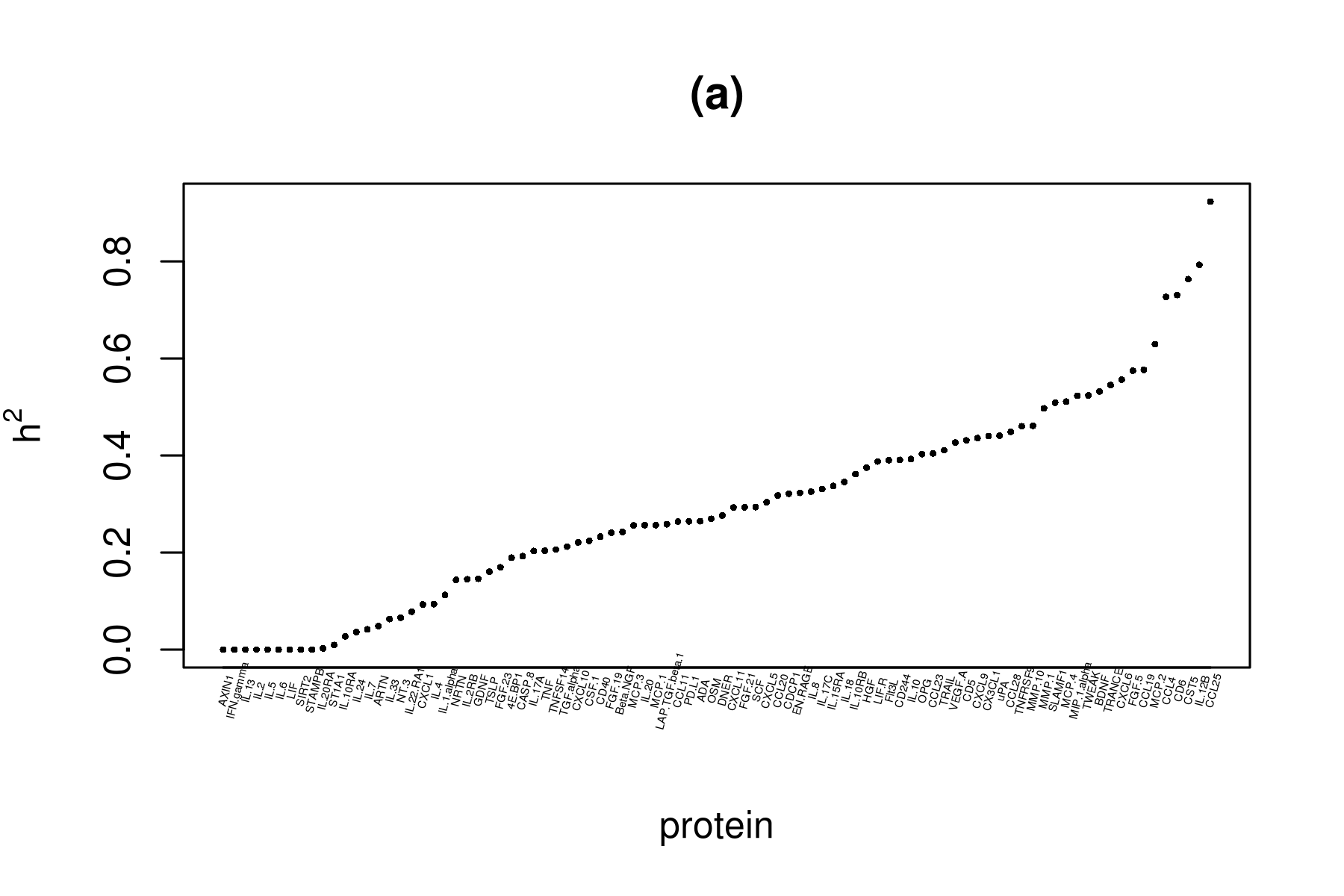


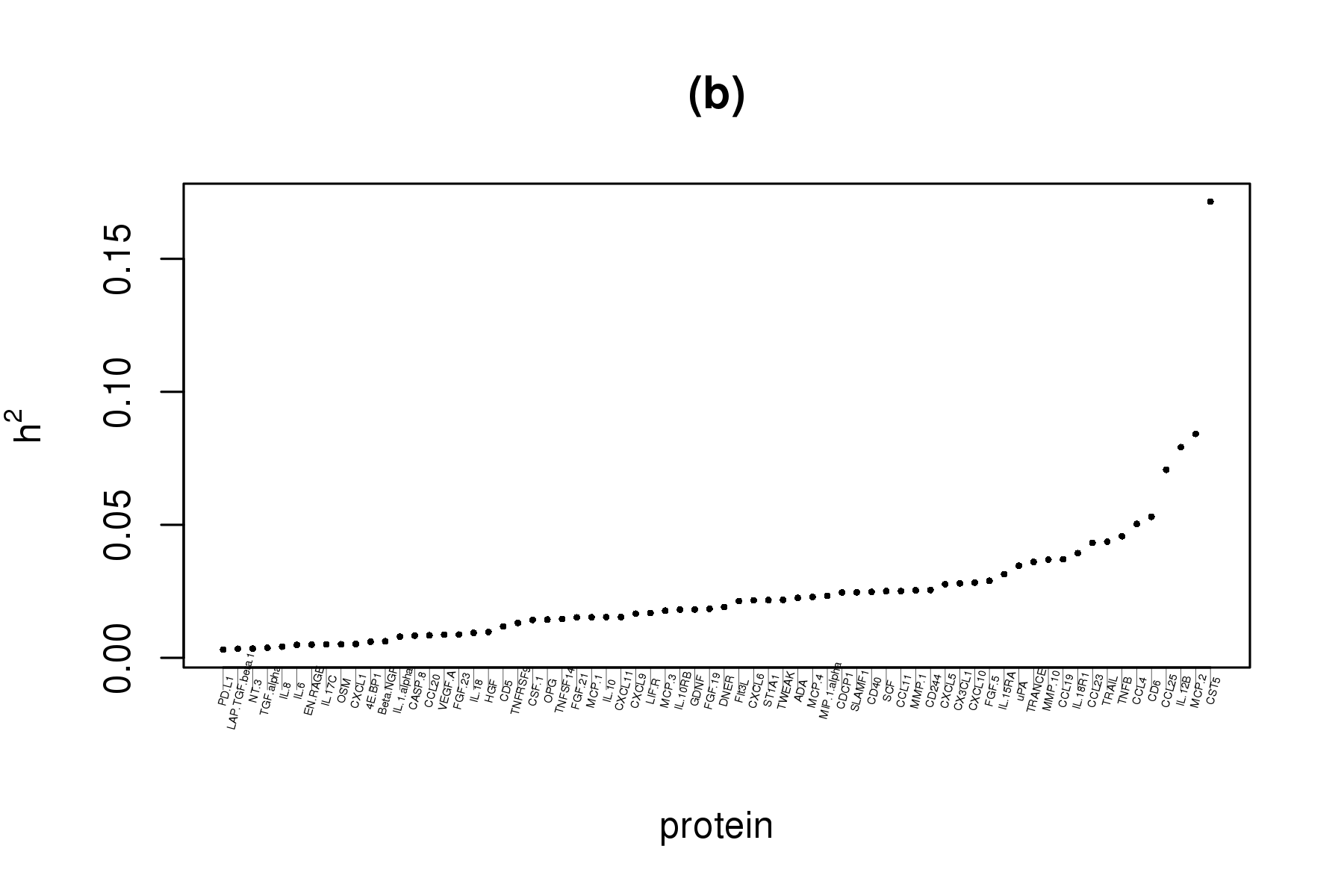
**Genomic heritability and variance explained**

Shown in Figure 2 are the genomic heritabilities obtained from the INTERVAL study. The estimates from INTERVAL study vary from almost zero for AXIN1, etc. to 0.92 for CCL25.

Variance explained by the sentinels from the meta-analysis varies from 0.003 for PD.L1 to 0.172 for CST5. The estimates were based on mixed model with adjustment for sex, age and PC1-PC20.

**Figure 2. Variance explained by SNPs**. (a). **genomewide** SNPs. (b). **sentinel** SNPs.



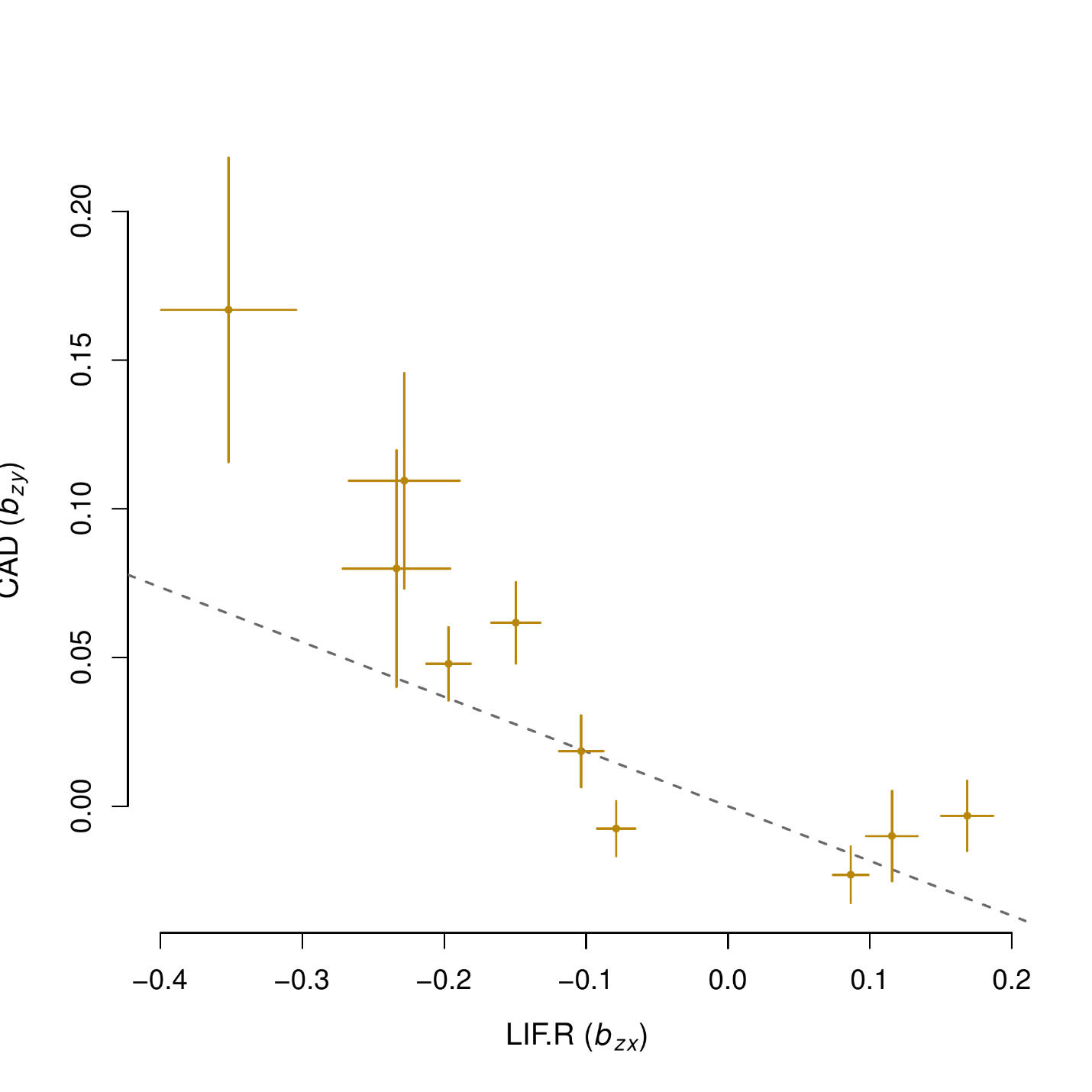


The list of variants is shown in Tables xx. The Q-Q, Manhattan, LocusZoom, and forest plots are shown in Supplementary Figures xx. The cis/trans classification is shown in Supplementary Table x.

**MR**

The generalised summary-data-based MR through GCTA is shown in Figure 3.

**Figure 3**. Effect size plot for LIF.R according to GCTA.



**Inflammation score**

## Discussions

By assembling the largest sample size so far, we were able to identify and validate protein-specific genetic associations in the OLINK/INF panel, followed by characterization with respect to their cis/trans effects, pleiotropic roles, and as instruments for causal inference through Mendelian randomization as well as biological pathways. By analogy to polygenic score and protein score (Ganz et al. 2016), an inflammation score are also be built.

We made our code available from the web; some of which were made generic through R package gap.

## Methods

## Olink Proximity Extension Assay (PEA) technology

Multiplex immunoassays that measure 92 proteins across 96 samples simultaneously using only one microliter of serum, plasma, etc.

A pair of oligonucleotide-labeled antibodies (“probes”) are allowed to pair-wise bind to the target protein present in the sample in a homogeneous assay, with no need for washing. When the two probes are in close proximity, a new PCR target sequence is formed by a proximity-dependent DNA polymerization event. The resulting sequence is subsequently detected and quantified using standard real-time PCR.

Un-truncated values of protein abundance was used.

### The INTERVAL study

The INTERVAL study was designed (Moore et al. 2014).

### The SCALLOP/INF1 consortium

The contributing cohorts in this study are listed in Supplementary Table studies. The raw measurements, including those beyond lower limit of detection, were subject to a rank-based inverse normal transformation. Quality control on cohort level involves SNP/sample call rates, gender mismatch, abnormal inbreeding coefficient, failed cryptic relatedness test, ancestry outlier, heterozygosity and Hardy-Weinberg equilibrium test. 1000 genomes imputation, build 37 (hg19) positions. At the time of analysis, BDNF was dropped from the panel so will not be analysed.

### Association analysis

Multiple linear regression was conducted on protein data and genotypes as well as sex, age, principal components and other cohort specific covariates under an additive genetic model. As individual level data from KORA study rather than GWAS summary statistics were available, the protein normalization and association testing were done centrally.

Software which account for genotype uncertainty, such as SNPTEST were used but due to the relatively large number of proteins, results based on PLINK were also accepted, both amended with outputs from qctool –snp-stats.

## Meta-analysis

**Quality controls**. Prior to the meta-analysis, extensive effort were paid on quality control of the GWAS summary statistics with available information such as MAF, HWE, and imputation score. To facilitate this, cohort-level Q-Q and Manhattan plots were generated with R package qqman and QCGWAS. Meta-analysis were performed using the inverse-variance weighted analysis of regression betas and standard errors, as implemented in the software METAL, version 28.8.2018 and the results were additionally visualized with regional association plots from LocusZoom 1.4. Variants with frequencies below 0.01 and/or above 0.99 were excluded from the meta-analysis. To address the issue that largest studies will drive the signals and the heterogeneity, those achieving genomewide significant will be from least three studies with combined sample size greater than 3,500 plus heterogeneity *I2*<30% or in the case of *I2*>=30% all studies should be nominally significant at 0.05 with consistent direction of effects. Forest plots were generated using customised functions in R to facilitate the meta-analysis.

**Identification of independent signals.** A distance-based approach was used and reframed as an algorithm here. It takes as input signals multiple correlated variants in particular region(s) which reach genomewide significance and output three types of sentinels in a region-based manner. For a given protein, the algorithm proceeds as follows:

**Algorithm *sentinels***

**Step 1**. for a particular chromosomal region, the width of the region is calculated according to the start and end chromosomal positions and if it is smaller than the flanking distance, the variant with the smallest P value is taken as sentinel (I) otherwise goes to **step 2**.

**Step 2**. The variant at **step 1** is only a candidate and a flanking region is generated. If such a region contains no variant the candidate is recorded as sentinel (II) and a new iteration starts from the variant next to the candidate.

**Step 3**. When the flanking is possible at **step 2** but the P value is still larger than the candidate at **step 2**, the candidate is again recorded as sentinel (III) but next iteration starts from the variant just after the variant at the end position; otherwise the variant is updated as a new candidate where the next iteration starts.

As in earlier report, the HLA locus is counted as single variant. Note type II results at **step 2** would be seen when a chromosome contains two trans signals. The function *sentinels* is part of the R/gap package at GitHub. We have used Bonferroni correction (5 x 10-10) for genomewide association and +/- 1MB flanking regions. In the comparison we used a version of –log10(p) which is based on effect size and its standard error of the association statistic which accommodates very small P value with high precision.

The algorithm is similar in spirit to swiss (<https://github.com/statgen/swiss>), i.e., swiss --dist-clump --clump-dist 1000000 but simpler to set up. We have conducted *in silico* experiments compared to PLINK clumping and GCTA joint/conditional analysis with respect to reference panel, sample size, inclusion of indels, and LD. The reference panel involves 1000Genomes release 3 data as well as UK10K+1000Genomes imputed data from the INTERVAL study. As expected the results were in favour of large reference panel, larger sample size, the inclusion of indels, and more signals from clumping than GCTA. Our experiment with approximately independent LD-blocks (Berisa et al 2016) found association peaks were separated into two neighbouring blocks.

**Finemapping**.Locus-specific analysis was done around each sentinel variant by conditional analysis through GCTA and INTERVAL-based imputed genotypes as the reference panels, and contrasted with software finemap. To make it feasible to cross-compare with additional software JAM, we also obtained independent variants for each region linking with a sentinel, such that LD r2 < 0.8 ensuring the sentinel being included. We used credible for these variants as a quantitative measure of being causal, through customized function (similar to the gwas-credible-sets JavaScript repository) in R for GCTA and built-in implementations in JAM and finemap.

**Annotation**. For signals as identified above, their cis/trans classifications were obtained using customised bash and R functions. PhenoScanner was used for variant annotation, and for signals with effects on multiple proteins their association was also assessed with R package hyprcoloc.

**Replication and contrast with previously reported signals**. Results from PhenoScanner highlights replication of cis/trans signals for OPG, which reported earlier (Kwan et al. 2014). Replication was also done through independent studies.

**Genomic heritability and variance explained**. INTERVAL individual level data were used for heritability estimation. Genetic relationship matrix was built on SNPs pruned at r2=0.8 and MAF=0.01, all using PLINK. Genomic heritability was then estimated with GCTA with adjustment for age, sex and PC1-PC20. Variants explained for each protein were approximated with where *T* is the total number of sentinel variants, the chi-squared statistic and the associate sample size, respectively (Giri, et al. 2019).

**Mendelian randomization**. **Generalised Summary-data-based Mendelian Randomisation (GMSR) as implemented in GCTA was used.**

**Pathway analysis**.Attempt was also made for cis signals.

**Transcriptomewide association analysis**. This mirrors work by Mancuso et al. (2017) and work on mQTL by McRae et al. (2018) and Yengo, et al. (2018).

## Acknowledgements and funding information

We wish to thank participants from the SCALLOP studies and collaboration from colleagues for making this work possible.

**CROATIA-Korcula/Split/Vis:** MRC Human Genetics Unit core funding, the Ministry of Science, Education and Sport in the Republic of Croatia (216-1080315-0302) and the Croatian Science Foundation (grant 8875) (funding to I. Rudan, C. Hayward, S.M. Kerr, O. Polasek, V. Vitart, and J. Marten).

**INTERVAL:** The INTERVAL study: NHSBT (11-01-GEN) and the NIHR-BTRU in Donor Health and Genomics (NIHR BTRU- 2014-10024) at the University of Cambridge in partnership with NHSBT. This study was partially funded by Merck. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health of England, or NHSBT. The Cardiovascular Epidemiology Unit at the University of Cambridge: UK MRC (G0800270), BHF (SP/09/002), UK NIHR Cambridge Biomedical Research Centre, ERC (268834), and European Commission Framework Programme 7 (HEALTH-F2-2012-279233).

**KORA F4 and KORA S3:** The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ and by the Competence Network Asthma and COPD (ASCONET), network COSYCONET (subproject 2, BMBF FKZ 01GI0882) funded by the German Federal Ministry of Education and Research (BMBF).

**NSPHS:** Swedish Medical Research Council (K2007-66X-20270-01-3, 2011-5252, 2012-2884 and 2011-2354), the Foundation for Strategic Research (SSF). NSPHS as part of European Special Populations Research Network (EUROSPAN) was also supported by the European Commission FP6 STRP (01947, LSHG-CT-2006-01947).

**ORCADES:** Supported by the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), the Royal Society, the MRC Human Genetics Unit, Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). ORCADES DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh.

## References

Folkersen L, et al. (2017). Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genetics 13(4), doi.org/10.1371/journal.pgen.1006706.

Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, Sterling DG, Williams SA (2016). Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* 315(23):2532-41. doi: 10.1001/jama.2016.5951.

Giri A, et al. (2019). Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* 51:51-62.

Kwan JSH, et al. (2014). Meta-analysis of genome-wide association studies identiﬁes two loci associated with circulating osteoprotegerin levels. Hum Mol Genet 23(24): 6684–669.

Mancuso N, et al. (2017). Integrating gene expression with summary association statistics to identify susceptibility genes for 30 complex traits. Am J Hum Genet 100:473-487

McRae AF, et al. (2018). Identification of 55,000 Replicated DNA Methylation QTL. *Sci Rep.* 8(1):17605. doi: 10.1038/s41598-018-35871-w.

Niewczas MA, et al. (2019). A signature of circulating inflammatory proteins and development of end-stage renal disease in diabetes. *Nat Med. https://doi.org/10.1038/s41591-019-0415-5*

Sun BB, et al. (2018). Genomic atlas of the human plasma proteome. Nature 558: 73–79.

Yengo L, et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum Mol Genet*. 27(20):3641-3649. doi: 10.1093/hmg/ddy271.

## Supplementary information

## URLs

Additional information about this investigation is available from GitHub, <https://github.com/jinghuazhao/INF>.

OLINK, <https://www.olink.com/data-you-can-trust/technology/>, SCALLOP, <https://www.olink.com/scallop/>; METAL <https://github.com/statgen/METAL>; PLINK, <http://zzz.bwh.harvard.edu/plink>; KING, <http://people.virginia.edu/~wc9c/KING/>; LDetect-data (for approximately independent LD blocks), <https://bitbucket.org/nygcresearch/ldetect-data>; FUSION, <http://gusevlab.org/projects/fusion/>, LocusZoom, <https://github.com/statgen/locuszoom-standalone>; R. <https://cran.r-project.org>; Uniprot, [https://www.uniprot.org](https://www.uniprot.org/):

## Cohort information

**Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably (BioFINDER)**. The study is based in Sweden and affiliated to the Clinical Memory Research Unit and The Biomedical centre, both at Lund University. Patients are consecutively included from the Memory and Neurology clinics at Skåne University Hospital as well as the Memory Clinic at Ängelholm’s Hospital. More than 1600 patients with mild cognitive symptoms, dementia and parkinsonian symptoms as well as cognitively healthy elderly have so far been enrolled in the study. The subjects undergo repeated examinations of advanced MRI (including fMRI, DTI, DKI, ASL and MPRAGE), CSF and plasma analysis, amyloid and tau PET, detailed clinical assessments and neuropsychological examinations. Skin biopsies are also collected and the fibroblasts are reprogrammed to iN and iPS cells.

**Estonian Genome Center at the University of Tartu (EGCUT)**. The cohort size is currently 51,535 gene donors (≥18 years of age), which closely reflects the age, sex and geographical distribution of the Estonian population. Estonians represent 83%, Russians 14%, and other nationalities 3% of all participants. All subjects have been recruited randomly by general practitioners (GP) and physicians in hospitals.  The participants are individuals who have joined the Estonian biobank after hearing about it during promotion events, media, friends, etc. or visiting GP offices or hospitals for other reasons.

**INTERVAL.** The INTERVAL study is a prospective cohort study of approximately 50,000 participants of mostly European ancestry, nested within a pragmatic randomized trial of blood donors. Between 2012 and 2014, blood donors 18 years and older were consented and recruited from 25 NHSBT (National Health Service Blood and Transplant) static donor centers across England. Participants are predominantly healthy individuals since people with major disease (myocardial infarction, stroke, cancer etc.) are ineligible for donation, as are those who report being unwell or having had recent illness or infection. Participants completed online questionnaires containing basic lifestyle and health-related information, including self-reported height and weight, ethnicity, current smoking status, alcohol consumption, doctor-diagnosed anemia, use of medications (hormone replacement therapy, iron supplements) and menopausal status. The INTERVAL study was approved by the Cambridge (East) Research Ethics Committee and UK Biobank was approved by the North West Multi-center Research Ethics Committee (MREC). Informed consent was obtained from all participants.

See Moore, et al. (2014), Astle, et al. (2016), Sun, et al. (2018).

**Cooperative Health Research in the Region of Augsburg (KORA).** It is a series of independent population based studies from the general population living in the region of Augsburg, Southern Germany. The **KORA S3** study including 4,856 individuals was conducted in 1994/95. Spirometry was measured during a follow up in 1997/98 for all participants younger than 60 years who did not smoke or use inhalers one hour before the test. **KORA F4** including 3,080 individuals was conducted from 2006-2008 as a follow-up study to KORA S4 (1999-2001). Genotypes were available on Affymetrix Axiom chips on 3788 individuals. The imputation was done through IMPUTE 2.3.2 with 1000Genomes phase 3 reference panel. A total of 1,070 individuals with both genotypes and protein data were used in association analysis via SNPTEST 2.5.2, where the per-sample missing proportion was obtained from qctool 2.0.1, and covariates were sex, age, and five PCs from GCTA 1.91.7beta. IBD information was obtained via KING 2.1.6. The total number of variants were 81,651,446.

**MadCam**. The MadCam trial samples were all from baseline but the patients included were moderate ulcerative cholitis. Several details on the trial are available from [Vermeire S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Vermeire%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28527704), et al. (2017).

**Northern Sweden Population Health Study** (NSPHS) represents a cross-sectional study conducted in the communities of Karesuando (samples gathered in 2006) and Soppero (2009) in the subarctic region of the County of Norrbotten, Sweden. Spirometry was performed in sitting position without noseclips using a MicroMedicalSpida 5 spirometer (http://www. medisave.co.uk). Three consecutive 28 lung function measurements per participant were done and the maximum value per measured lung function parameter was used for further analysis. Relatedness was taken into account by applying the "polygenic" linear mixed effects model. Genome-wide association analysis was performed using a score test, a family-based association test27 which uses the residuals and the variance-covariance matrix from the polygenic model and the SNP fixed effect coded under an additive model.

**Orkney Complex Disease Study** (ORCADES) is an ongoing family-based, cross-sectional study in the isolated Scottish archipelago of Orkney. Spirometry was performed in the sitting position without nose clips, using a Spida handheld spirometer. Measurements were repeated once and the better reading was used for analysis.

Imputation panel was HRC.

**RECOMBINE**. COMBINE is a five year project which started June 15 2008. COMBINE was initially funded by six Swedish Research organizations, Vinnova, Vårdalstiftelsen, Reumatikerförbundet, Invest in Sweden Agency, KK-stiftelsen, Stiftelsen för strategisk forskning. The overall objective of COMBINE is to use unique Swedish advantages to improve understanding of why inflammatory diseases develop, what are the most essential goals for patients to achieve, and to develop and implement novelprevention and therapy for these diseases. COMBINE has enabled a novel participation from patients and patient organizations in design and interpretation of research. It has also - in collaboration with other initiatives - been active in workingwith clinical care for implementation of science in clinical practice. Finally, COMBINE has enabled a novel way for collaboration between translational andclinical science and pharmaceutical/biotech industry.

# Stabilization of Atherosclerotic Plaque by Initiation of Darapladib Therapy Trial (STABILITY). Briefly, STABILITY was a randomized, double‐blind, controlled trial that enrolled patients with a history of CHD, including previous myocardial infarction (MI), previous percutaneous coronary intervention or coronary artery bypass grafting, or multivessel coronary disease confirmed by angiography, and on statin therapy unless contraindicated or not tolerated. In addition, at least one of the following risk factors was required for enrollment: age ≥60 years, diabetes mellitus requiring pharmacotherapy, moderate renal impairment, smoking ≥5 cigarettes per day at study entry or within the past 3 months, polyvascular arterial disease, poorly controlled hypertension, or high‐density lipoprotein cholesterol <40 mg/dL. Patients were excluded if they had liver disease, severe renal dysfunction, history of nephrectomy or kidney transplantation, heart failure with New York Heart Association class III or IV, or severe asthma or if they had a percutaneous coronary intervention, coronary artery bypass grafting, or a major surgical procedure planned. Study participants were randomized to receive either a 160‐mg oral dose of darapladib daily or placebo. The median duration of follow‐up was 3.7 years (25th–75th percentiles: 3.5–3.8 years). The study was approved by the institutional review committee in each participating country, and all patients provided written informed consent.

**STANLEY**. The study consists of lah1 and swe6 subcohorts.

**VIS**. CROATIA-Vis is a family-based, cross-sectional study in the isolated island of Vis, Croatia that included 1,056 examinees aged 18-93. It is a genetic epidemiology study that aims to discover genetic factors that influence traits (e.g. height) or the risk of common complex diseases. The cohort is very well characterised with detailed phenotyping and genotyping information available.

Imputation panel was HRC.

## References

[Vermeire S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Vermeire%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28527704), et al. (2017). Anti-MAdCAM antibody (PF-00547659) for ulcerative colitis (TURANDOT): a phase 2, randomised, double-blind, placebo-controlled trial. *Lancet* 390(10090):135-144. doi: 10.1016/S0140-6736(17)30930-3.

White HD, et al. (2014). Darapladib for preventing ischemic events in stable coronary heart disease. [*N Engl J Med.*](https://www.ncbi.nlm.nih.gov/pubmed/24678955?dopt=Abstract) 370(18):1702-11. doi: 10.1056/NEJMoa1315878.

## Supplementary tables

Studies. Study information

## Q-Q plots

## Manhattan plots

## Regional association plots

## Forest plots

## Pathway analysis

## PheWAS results

## EWAS results