Streptococcus species abundance in the gut is linked to subclinical coronary atherosclerosis in 8,973 participants from the SCAPIS cohort

Sergi Sayols-Baixeras[1,2*], Koen F Dekkers[1*], Ulf Hammar[1], Gabriel Baldanzi[1], Yi-Ting Lin[3,4], Shafqat Ahmad[1,5], Diem Nguyen[1], Georgios Varotsis[1], Sara Pita[6,7], Nynne Nielsen[6], Aron C. Eklund[6], Jacob B Holm[6], H Bjørn Nielsen[6], Louise Brunkwall[8], Filip Ottosson[8,9], Christoph Nowak[3], Daniel Jönsson[8,10,11], Dan Ericson[12], Björn Klinge[11,13], Peter M Nilsson[8,14], Andrei Malinovschi[15], Lars Lind[16], Göran Bergström[17,18], Johan Sundström[19,20], Johan Ärnlöv[3,21], Gunnar Engström[8], J. Gustav Smith[22,23,24], Marju Orho-Melander[8#] and Tove Fall[1#]

[*]Shared first authorship

[#]Shared senior authorship

[1]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, EpiHubben, MTC-huset, Uppsala, Sweden

[2]CIBER Cardiovascular diseases (CIBERCV), Instituto de Salud Carlos III, Madrid, Spain.

[3]Division of Family Medicine and Primary Care, Department of Neurobiology, Care Science and Society, Karolinska Institutet, Huddinge, Sweden

[4]Department of Family Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Taiwan

[5] Preventive Medicine Division, Harvard Medical School, Brigham and Women's Hospital, Boston, MA, United States

[6]Clinical Microbiomics A/S, Copenhagen, Denmark

[7]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark

[8]Department of Clinical Sciences in Malmö, Lund University, Malmö, Sweden

[9]Section for Clinical Mass Spectrometry, Danish Center for Neonatal Screening, Department of Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark

[10]Public Dental Service of Skåne, Lund, Sweden.

[11]Department of Periodontology, Faculty of Odontology, Malmö University, Malmö, Sweden.

[12]Deparmtent of Cariology, Faculty of Odontology, Malmö University, Malmo, Sweden.

[13]Department of Dental Medicine, Karolinska Institutet, Solna, Sweden.

[14]Department of Internal Medicine, Skånes University Hospital, Malmö, Sweden

[15]Department of Medical Sciences, Clinical Physiology, Uppsala University, Uppsala, Sweden.

[16]Department of Medical Sciences, Clinical Epidemiology, Uppsala University, Uppsala Science Park, Uppsala, Sweden

[17]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[18]Department of Clinical Physiology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden

[19]Department of Medical Sciences, Clinical Epidemiology, Uppsala University, Uppsala, Sweden

[20]The George Institute for Global Health, University of New South Wales, Sydney, Australia

[21]School of Health and Social Studies, Dalarna University, Falun, Sweden

[22]The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University and the Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden

[23]Department of Cardiology, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden

[24]Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Lund University, Lund, Sweden

WORD COUNT: 4,472

Abstract

Coronary atherosclerosis is the main pathophysiological mechanism underlying myocardial infarction. The gut microbiota has been implicated in cardiometabolic disease but its relationship with subclinical coronary atherosclerosis is unknown. We identified 73 gut metagenomics species associated with coronary artery calcium score (CACS) in 8,973 SCAPIS participants without previous cardiovascular disease. *Streptococcus* associations were overrepresented and were validated in an independent case-control study together with eight non-*Streptococcus* spp. We further found enrichment for bacterial genes linked to amino acid and carbohydrate degradation functions. Gut *Streptococcus* spp. were associated with circulating biomarkers of inflammation and infection response, bile acids, androgenic steroids and sphingomyelins, and were associated with their homologous species in the oral cavity, which were in turn associated with oral pathologies. This study provides robust evidence of the association of *Streptococcus* spp., with subclinical atherosclerosis and markers of systemic inflammation and infection, calling for studies re-investigating the infectious hypothesis in atherosclerosis pathogenesis.

Atherosclerotic cardiovascular diseases (ACVD), such as ischemic heart disease and ischemic stroke, are the major causes of death and disability [1]. The formation of atherosclerotic plaques is a silent, complex, and progressive process characterized by accumulation of lipids, fibrous elements, calcium minerals, and inflammatory molecules in the subendothelial space[2]. While the underlying mechanisms of atherosclerosis remain incompletely understood, it has been proposed that the gut microbiota composition could contribute to accelerated atherosclerotic development by transmission of bacteria into circulation, resulting in either subsequent direct infection of the atherosclerotic plaque or systemic inflammation associated with infection at other sites[2]. For instance, experimental studies suggested that oral challenge with *Streptococcus* spp. accelerates atherosclerotic plaque growth and macrophage invasion[3]. Alternatively, certain bacteria could affect the atherosclerosis process by modulating the host metabolism or interaction with dietary components to produce both beneficial and harmful molecules[2]. Gut microbiota composition has already been linked to cardiovascular risk factors, such as obesity[4], insulin resistance[5] and type 2 diabetes[6], although the causal relationships remain unclear. A number of case-control studies of symptomatic coronary atherosclerosis with up to 1,241 participants have pointed to differences in abundance of more than 500 gut species[7–15]. However, the findings are often not reproducible and are prone to bias; comparison groups are often non-comparable in terms of medical treatment and lifestyle factors, and there is risk of reverse causation. The importance of employing large cohorts in microbiome research of earlier phases of the atherosclerosis process are thus compelling[7,13,14,16,17].

Here, we identified associations between gut microbiota composition, in particular *Streptococcus* spp., and asymptomatic coronary atherosclerosis, determined by computed tomography-derived coronary artery calcium score (CACS), in a large cohort of middle-aged Swedes from the Swedish CArdioPulmonary bioImage Study (SCAPIS)[18] and validated in a geographically distinct case-control study of symptomatic atherosclerotic disease. Furthermore, we identified associations of gut *Streptococcus* spp. with circulating inflammatory and infection biomarkers. These species were further associated with the corresponding oral *Streptococcus* spp. that in turn were associated with worse oral health. In conclusion, we identify a subset of gut microbiota species with enrichment for *Streptococcus*

that are robustly associated with CACS, laying the foundation for future studies on causal relationships and investigations of the plausibility of these species as potential intervention targets to reduce cardiovascular risk.

Results

Large Swedish cohort profiled with deep shotgun metagenomics of fecal samples and detailed coronary atherosclerosis imaging. In the current study, we took advantage of SCAPIS, a unique resource for epidemiological studies combining a large sample size with extensive and in-depth phenotypic information, including cutting-edge molecular techniques and direct imaging of the cardiovascular disease (CVD) processes[18]. We selected 8,973 individuals recruited at the Malmö and Uppsala sites, aged 50–64 years, with no history of symptomatic CVD. A description of the main sociodemographic and clinical characteristics of the study population is presented in **Table 1**, including the prevalence of calcified coronary plaques, measured as CACS and categorized as absent (CACS=0), mild (1–100), moderate (101–400), and extensive (>400) calcification of coronary arteries. The prevalence of CACS>0 in the study sample was 40.3%, comparable to the whole SCAPIS population[19], which in turn showed a high agreement with contrast-enhanced coronary computed tomography angiography (CCTA), an alternate measurement of atherosclerosis[19]. Only 5.5% of the participants with CACS=0 were classified as having any coronary atherosclerosis.

Gut microbiota composition and richness are associated with asymptomatic atherosclerosis and attenuated by adjustment for lifestyle factors, diet and medication. To test the primary hypothesis that gut microbiota composition is associated with CACS at the asymptomatic disease stage, we first tested whether the alpha and beta diversity measures were associated with CACS. Alpha diversity, a measure of overall species richness and evenness within each sample, was inversely associated with CACS (β=-0.16 (95% confidence interval (CI) = -0.26, -0.07, p-value=$8.6*10^{-4}$)) using linear regression adjusting for age, sex, country of birth, center site and extraction plate (basic model). However, after further adjustment for smoking, physical activity, fiber and total energy intake, and self-reported medication for dyslipidemia, hypertension and/or diabetes (full model), no association

was observed (β=-0.04 (95% CI=-0.15, 0.06, p-value=0.38)). These covariates were included because our hypothetical causal diagram indicated them as potential confounders for the microbiota-CACS association (**Extended Data Fig. 1**). Further, in permutational multivariate analysis of variance (PERMANOVA), we found that beta diversity, an indicator of the overall similarity among samples, differed across the CACS categories in both models, although the fully adjusted model was attenuated (r^2_{basic}=0.0007; p-value$_{basic}$=1*10^{-04} vs r^2_{full}=0.0004; p-value$_{full}$=0.036). Pairwise comparisons revealed that the overall microbiome composition in participants with CACS=0 was different compared to other categories, with increasing distances with higher CACS classes (**Figure 1 and Supplementary Table 1**). These findings suggest that gut microbiome composition and richness are associated with asymptomatic atherosclerosis, but that differences are partly due to differences in lifestyle factors, diet, and medication across groups.

Specific species are robustly associated with subclinical atherosclerosis, especially species belonging to the *Streptococcus* genus. Extensive simulation studies prior to the analysis supported the use of linear regression modelling for high power, interpretability, and low frequency of false positive findings (**Extended Data Fig. 2**). We found the relative abundance of 73 out of 1,985 tested species associated with CACS at a false discovery rate (FDR) of 5% using the basic model further adjusted for Shannon diversity index. For 60 of these, the relative abundance was positively associated with CACS and for 13 negatively associated (**Fig. 2a, b, d, Supplementary Tables 2 and 3**). Gene-set enrichment analysis (GSEA) for genera revealed an overrepresentation of the *Streptococcus* genus in the associations with a positive effect estimate (**Fig. 2a, c, and Supplementary Table 4**). In the full model (further adjusted for Shannon diversity index), 63 species remained associated with CACS in 8,155 individuals with complete data on all covariates (**Fig. 2a, d, and Supplementary Table 3**). *Streptococcus anginosus*, *Streptococcus oralis* subsp. *oralis*, *Anaerotignum lactatifermentans*, *Escherichia coli*, and *Eubacteriales* sp. (internal identifier HG3A.1354) were most strongly species associated with CACS based on p-value, and all of them were positively associated. To assess whether the difference in the number of associated species between basic and full model was due to smaller sample size or confounding by the included covariates in the full model, we re-analyzed the 73 species

associations with the basic model restricted to the 8,155 individuals with complete data that were included in the full model. This analysis showed highly correlated (Pearson correlation=0.99) but slightly attenuated estimates compared to the full model, supporting the difference in the number of associated species between basic and full model to be mainly due to lower power and to some extent confounding (**Supplementary Table 3 and Extended Data Fig. 3**). A comparison of the main clinical characteristics of carriers and non-carriers of the most strongly associated species *S. anginosus* and *S. oralis* subsp. *oralis* (**Supplementary Table 5**) revealed that these *Streptococcus* spp. were more prevalent in women. Carriers had on average higher triglycerides, blood pressure, body mass index (BMI), high-sensitivity C-reactive protein (hsCRP), leukocytosis, neutrophilia, prevalence of diabetes, ulcerative and Crohn's disease, and medication use compared to the non-carriers. Sensitivity analyses performed using partial Spearman correlation showed loss-of-significance for 32 species. However, extensive sensitivity analyses did not find any evidence for this difference resulting from single influential observations nor non-linear effects, and differences were thereby likely to be due to lower power compared to linear model (**Supplementary Information, Supplementary Table 6 and 7**). Even if not directly indicated as confounders, we further assessed the effect of adjustment for BMI and proton-pump inhibitors (PPI) medication, as well as excluding specific groups of individuals such as those with Crohn's disease or those with antibiotic treatment the past 12 months from the analyses. . We also tested the effect of not adjusting the basic and full models for Shannon diversity index to observe if some of these factors could explain the identified associations of the species and CACS (extensively explained in **Supplementary Information, Extended Data Fig. 4, Supplementary Table 6 and 7**). Statistical significance was lost for a few species in some of the analyses; however, effect estimates remained largely similar. Overall, these findings indicate that the associations identified in the main analysis were robust and were not in general affected by statistical model misspecification, PPI or antibiotic use, or gastrointestinal disease. Moreover, our findings show that the identified species are in general associated with CACS independent of established cardiovascular risk factors.

Sex modifies the association between certain species and subclinical atherosclerosis. We re-assessed the 63 species associated with CACS in the full model stratifying for sex, and tested whether the effect estimates in the female strata were different to the male strata (**Extended Data Fig. 5, Supplementary Tables 3 and 8**). *Streptococcus agalactiae*, *Rothia mucilaginosa*, and two *Eubacteriales* spp. (HG3A.0242 and HG3A.0854) showed different effect estimates across sex (Cochran's Q-test p-value<0.05) where the two first mentioned species were associated in females only and the two latter in males only. These findings indicate that sex-specific associations are present between gut microbiota and subclinical coronary atherosclerosis.

Few CACS-related species are also associated with carotid atherosclerosis. Next, using ordinal regression analysis with basic model adjustment and a FDR of 5%, we searched for associations of CACS-associated species with carotid atherosclerosis in 8,955 SCAPIS participants with available information of carotid atherosclerosis measured with ultrasound, in addition to deep shotgun metagenomics and CACS (**Supplementary Table 9**). Subjects were categorized as "no identified atherosclerosis" (n=3,821), "atherosclerosis in one carotid artery" (n=2,779), and "atherosclerosis in both arteries" (n=2,355). Carotid atherosclerosis was detectable in 57% of the SCAPIS participants which is higher than in previous large-cohort studies: 34% in the Atherosclerosis Risk in Communities study (ARIC)[20], 47% in the Multi-Ethnic Study of Atherosclerosis (MESA)[21], 44% in the Malmö Diet and Cancer study (MDC)[22], and 45% in Risk Evaluation For INfarct Estimates (REFINE)-Reykjavik study[23]. Twenty-one of the CACS-related species were associated with carotid atherosclerosis in the basic model including *Streptococcus mutans*. However, *Blautia obeum*, *Clostridium phoceensis*, *Oscillibacter* sp. (HG3A.0243), *Intestinimonas* sp. (HG3A.1018) and three *Eubacteriales* spp. (HG3A.0242, HG3A.0511, HG3A.1158) remained significant in the ordinal regression with the full model adjustment (**Supplementary Table 9**). These results may suggest that gut microbiota associations differ between the two vascular beds; however, it might also depend on differences in the measurement methods.

Fifteen CACS-related species, including seven *Streptococcus* spp., are validated using a case-control study of symptomatic atherosclerotic cardiovascular patients. We next attempted to

replicate findings using shotgun metagenomics data from a case-control study[13] of 210 patients with symptomatic ACVD and 163 controls from a Chinese population representing a later stage of the atherosclerotic process, wider age range (32–107 years), different genetic background and geographical area, and also likely different dietary habits and medication usage, which are all factors that can modulate gut microbiota composition. Of the 73 species associated with CACS in the basic model in the current study, we could identify 64 species using the same gene sequence signatures as in the SCAPIS study. Fifteen of these 64 species, including seven *Streptococcus* spp. were associated with ACVD with consistent effect direction as in SCAPIS using logistic regression models adjusted for age, sex and Shannon diversity index at an FDR of 5%, where disease status was the dependent variable and the rank-base inverse transformation of the relative abundance was set as the independent variable of interest (**Extended Data Fig. 6 and Supplementary Table 10**).

Trehalose and fructose degradation genes are enriched in CACS-related species, especially in *Streptococcus* genus. Next, to identify possible bacterial functions that might be involved in the coronary subclinical atherosclerosis and shared between the associated bacteria, we used GSEA to identify functional gut metabolic modules (GMM)[24] enriched for the CACS-associated species. The GSEA were based on ranked p-values from the basic association model with CACS stratified by the direction of the regression coefficient. The analysis revealed four enriched functional GMM in the bacterial species positively associated with CACS**,** and one enrichment in species negatively associated with CACS (**Extended Data Fig. 7a, b** and **Supplementary Table 11** and **12**). The positive enriched modules were involved in amino acid degradation (threonine degradation I and II) and carbohydrate degradation (trehalose degradation and fructose degradation), while the negative module was involved in acetogenesis. To test if these functions were enriched due to the overrepresentation of *Streptococcus* spp. associated with CACS, we performed GSEA applying leave-one (taxon)-out analysis (**Extended Data Fig. 7c**, **Supplementary Table 13** and **14**). This analysis revealed that the trehalose and fructose degradation functions were strongly attenuated when we removed *Streptococcus* spp., and to a lesser extent the threonine degradation II, supporting that *Streptococcus* spp. are important contributors for the association of these functions with CACS.

Streptococcus spp. have a high ability to catalyze different types of carbohydrates including most common dietary sugars (fructose, glucose and sucrose) forming compounds involved in virulence processes including biofilm buildup[25]. Trehalose degradation has been linked to virulence, and its role as a stress protectant has been linked to protection from desiccation freezing, starvation, and osmotic stress[26,27]. For threonine degradation I, *Roseburia* was the important genus, for threonine degradation II, *Bifidobacterium,* and for homoacetogenesis, *Romboutsia.* Together, these results show that CACS-associated species share, at least partially, some specific bacterial functions related to amino acid and carbohydrate degradation, and that *Streptococcus* spp. were the most important contributor to the enrichment for trehalose degradation genes, a function previously linked to bacterial virulence.

Plasma bile acids, androgenic steroids, and sphingomyelin-associations with CACS-associated *Streptococcus* spp. are overrepresented. Next, we identified 2,377 associations between seven CACS-associated *Streptococcus* spp., and 873 of the 1,412 plasma metabolites detected in 7,252 SCAPIS participants using partial Spearman correlation adjusted at 5% FDR (**Extended Data Fig. 8** and **Supplementary Table 15)**. Among the positive associations, primary bile acid metabolism was enriched in four *Streptococcus* spp. (*S. gordonii, S. parasanguinis, S. salivarius*, and *S. oralis*. subsp. *oralis*), while secondary bile acid metabolism, acetylated peptides, and analgesics and anesthetics drugs sub-pathway were enriched in *S. anginosus*; and plasmalogen in *S. oralis* subsp. *oralis* (**Fig. 3a** and **Supplementary Table 16**). At individual metabolite level, all 11 primary bile acid metabolites were positively associated with all the *Streptococcus* spp. with adjusted p-values<0.05, with the exception of chenodeoxycholic acid sulfate. We further observed that metabolites from analgesic and anesthetic medications were positively associated with all the *Streptococcus* spp., particularly those metabolites derived from paracetamol. Furthermore, omeprazole was strongly associated with the *Streptococcus* spp., with the exception of *S. agalactiae*. In contrast, androgenic steroids metabolites and sphingomyelins showed enrichment among the metabolites negatively correlated with *S. gordonii, S. parasanguinis, S. salivarius,* and *S. mutans; and S. mutans* and *S. anginosus* respectively (**Figure 3a** and **Supplementary Table 17)**. Notably, all androgenic steroids and 26 of the 29 sphingomyelins metabolites were negatively associated with CACS-related *Streptococcus* spp. with adjusted p-

values<0.05. For androgenic steroids, sex-differences in the effect estimates were tested. Eleven metabolites reported different effect estimates across sex (Cochran's Q-test p-value<0.05). Collectively, these results show the tight relationship between CACS-associated *Streptococcus* spp. and endogenous and exogenous metabolites. Next, we additionally adjusted the full model assessing the relationship between *Streptococcus* spp. and CACS for the *Streptococcus*-associated metabolites (with <30% missing data) involved in sub-pathways related to primary and secondary bile acids metabolism, acetylated peptides, plasmalogen, androgenic steroids, sphingomyelins, analgesic and anesthetic drugs, and/or partially characterized molecules to assess the potential mediation by these metabolites (**Extended Data Fig. 9** and **Supplementary Table 18**). The inclusion of these metabolites in the models did not attenuate the association between *Streptococcus* spp. and CACS compared with the full model restricted to complete data (n=5,683), indicating that the association is not mediated by these metabolites (**Extended data Fig. 9**).

***Streptococcus* spp. are associated with markers of systemic inflammation and infection.** To test whether *Streptococcus* spp. were associated with markers of systemic inflammation and infection, respectively, we evaluated the associations between these species and high-sensitivity plasma C-reactive protein (hsCRP) (n=7,248), and counts of leukocytes (n=7,237) and neutrophils (n=7,235). Five out of seven CACS-associated *Streptococcus spp.* (*S. anginosus, S. parasanguinis, S. oralis* subsp. *oralis, S. gordonii* and *S. salivarius*) were positively associated with hsCRP and leukocyte counts. The same five *Streptococcus* spp. and *S. mutans* were positively associated with neutrophil counts (**Fig. 3b** and **Supplementary Table 19**). These models were adjusted for age, sex, country of birth, smoking, physical activity, fiber and total energy intake, and self-reported medication for dyslipidemia, hypertension and/or diabetes, Shannon diversity index, center site and extraction plate. The inclusion of BMI and PPI into the models attenuated the estimates but all five species remained associated with hsCRP, while only *S. parasanguinis* and *S. salivarius* remained associated with leukocytes and neutrophils counts. *S. oralis* subsp. *oralis* was also associated with neutrophils counts (**Supplementary Table 20**).

Overall, these findings suggest that these *Streptococcus* spp. are related to markers of inflammation and infection, which have been described as plausible mechanisms in gut microbiota effects on the atherosclerosis process[2].

Gut *Streptococcus* spp. are correlated with oral *Streptococcus* spp. and the latter are associated with oral health. *Streptococcus* spp. are commonly localized in the oral cavity and are linked to worse oral health status. We investigated the correlation between the abundance of CACS-related *Streptococcus* spp. in the fecal and saliva samples from 343 participants in the Malmö Offspring Dental Study (MODS) with an age range between 23 and 71 years who underwent a thorough dental examination within 4 to 12 months after fecal sampling in the Malmö Offspring Study (MOS)[28]. The gene signature in the saliva samples mapped to *S. anginosus*, *S. parasanguinis*, *S. gordonii*, *S. mutans*, *S. oralis*, and two different *S. salivarius* (Ho1B.0002 and Ho1B.0234) in the catalogue. Unfortunately, we could not identify *S. agalactiae* in the salvia samples. We applied partial Spearman correlations with cluster-robust standard errors accounting for relatedness between the CACS-related *Streptococcus* spp. in the gut and the corresponding species in the oral cavity adjusting for age, sex, country of birth, and extraction plate of the fecal samples. Five *Streptococcus* spp. (*S. anginosus*, *S. parasanguinis*, *S. gordonii*, *S. mutans* and *S. salivarius*) from the fecal samples were positively associated with their homologous species in the oral cavity (rho=0.15–0.30) (**Supplementary Table 21**). To investigate whether the oral *Streptococcus* spp. corresponding to CACS-associated gut bacteria species were associated with oral health, we fitted a series of ordinal regressions with cluster-robust standard errors with three different outcomes (filled surfaces, surfaces with caries (initial and manifest), or gingival inflammation in 637 participants from MODS. These models were adjusted for age, sex, smoking, education, oral hygiene, activity realized the hour before the dental examination including eating, brushing teeth or/and smoking, and Shannon diversity index (**Extended data Fig. 10**). *S. anginosus* was associated with all three outcomes at 5% FDR. *S. mutans* was associated with gingival inflammation and caries. *S. parasanguinis* and *S. salivarius* were associated with filled surfaces while *S. gordonii* was associated with gingival inflammation (**Fig. 3c** and **Supplementary Table 22**). Further adjustments for BMI, PPI and antibiotic treatment provided similar estimated

effects, indicating that these associations are independent of these factors (**Supplementary Table 23**). These findings showed a correlation between oral and gut microbiota, which may explain a migration of these species from the oral cavity to the gut. Furthermore, we observed four *Streptococcus* spp. associated with oral health, which may contribute to mechanisms underpinning the association between oral health and atherosclerosis process.

Discussion

Gut bacteria have been proposed to affect atherosclerosis progression and development through infections local or distal to the atherosclerotic plaque, or through production of metabolites affecting the atherosclerotic process[2,29]. The association of gut microbiota with coronary atherosclerosis has previously only been studied in symptomatic patients, who are often under treatment, resulting in high risk of bias. To address the previous biased sampling, we leveraged the large population-based SCAPIS cohort with detailed image-based measurements of coronary artery atherosclerosis, and deep characterization of the gut microbiome using shotgun metagenomics. We identified 73 species associated with CACS, with an enrichment of *Streptococcus* spp. and gene functions involved in amino acid and carbohydrate degradation. Our findings further supported that gut *Streptococcus* spp. were independently associated with endogenous and exogenous plasma metabolites, with inflammatory and infection markers, and with their bacterial homologues in the oral cavity, which were associated with worse oral health. Together, these findings provide robust evidence of the association of *Streptococcus* spp., with subclinical atherosclerosis and markers of systemic inflammation and infection, calling for studies to re-investigate the role of bacteria in atherosclerosis pathogenesis.

The *Streptococcus* genus was clearly enriched in the associations between species and CACS, in line with observations in earlier ACVD case-control studies[7,13,14]. Specifically, we observed *S. anginosus*, *S. oralis* subsp. *oralis*, *S. parasanguinis*, *S. gordonii*, *S. salivarius*, *S. mutans*, and *S. agalactiae* associated with increased CACS. These species commonly colonize the oropharyngeal cavity and the digestive tract[30] and all belong to the viridans streptococci group (VGS), except for *S. agalactiae* that

329 is a β-hemolytic non-VGS. An overview of the current knowledge of the involvement of these specific
330 bacteria in cardiovascular pathologies is provided in **Supplementary Table 24**. VGS has the ability to
331 infect the valves and the coronary vessels accounting for 20% of infective endocarditis cases, and they
332 have been isolated from human coronary atherosclerotic plaque samples from coronary artery disease
333 patients[31–33]. Some studies on animal models support a causal link between *Streptococcus* species and
334 the atherosclerotic process[3,34,35]. For instance, atherosclerosis-prone mice orally challenged with *S.*
335 *sanginius* had pro-inflammatory responses in the aorta, and accelerated atherosclerosis[3]. In the present
336 study, we observed strong positive associations between abundance of CACS-associated
337 *Streptococcus* species in the gut and hsCRP, leukocytosis and neutrophilia, which could have been
338 triggered by low-grade bacteremia. Among patients with bacteremia, VGS is a common cause[36,37].
339 VGS species are early colonizers, and may contribute to or initiate biofilm formations. Biofilms are
340 syntrophic beneficial poly-microbial communities that facilitate bacterial survival in aerobic
341 environments such as the atherosclerotic plaque[38] and biofilms have been observed in atherosclerotic
342 lesions associated with the fibrous cap[39], however causal relationships are not clear. Both biofilms and
343 VSG can induce persistent inflammation, attract monocytes into the endothelial space, and contribute
344 to platelet aggregation, all of which are the requisites for promoting the atherosclerosis
345 development[40,41]. Both local and distal infections require translocation of bacterial species to the
346 bloodstream. In the oral cavity, VGS form biofilms on the tooth surface and they can enter to the
347 bloodstream following mucosal barrier injuries, for example daily dental care activities, invasive
348 dental procedures, and oral pathologies, which have been associated with increased risk of
349 atherosclerosis, myocardial infarction, and stroke[42–48]. Our results indicated associations between the
350 oral VGS and worse oral health, which can potentially be the entry point to the bloodstream.

351 A potential modifier of *Streptococcus* spp. abundance would be antimicrobial treatment. However,
352 multiple clinical trials have demonstrated inefficacy of anti-infective therapies in mitigating
353 atherosclerotic cardiovascular events[49]. The CLARICOR study[50], in which patients with stable
354 coronary artery disease were treated with clarithromycin, an antibiotic that can be used to treat
355 *Streptococcus* spp. infections, reported increased mortality in the treatment arm. One possible

356 explanation for this lack of efficacy even though *Streptococcus* spp. would be causally related to
357 atherosclerosis, could be the formation of *Streptococcus*-associated biofilms that increase the bacterial
358 resistance to antibiotic treatment up to 1,000 times[51], or alternatively, antibiotic treatment could lead to
359 a recolonization of more pathogenic bacteria. Furthermore, the treatment window could exist much
360 earlier in the atherosclerosis process. In the current study, the association between the *Streptococcus*
361 spp. and calcification in the coronary arteries remained after exclusion of those participants treated
362 with antibiotics the past year.

363 Additionally, gut microbiota composition could affect atherosclerosis development through effecting
364 the host metabolism. We identified 1,412 associations between CACS-associated *Streptococcus*
365 species in the gut and plasma metabolites, with an overrepresentation of associations with plasma bile
366 acids, androgenic steroids, and sphingomyelins. Gut microbiota species are essential to transform
367 primary bile acids to secondary bile acids that enter circulation and interact with host bile acid
368 receptors. Knock-out mice of these receptors were shown to be protected against atherosclerosis
369 development or progression[52]. Elevated plasma bile acids have also previously been associated with
370 increased risk of coronary plaques in asymptomatic individuals[53]. However, other studies have
371 reported a null or a protective role in symptomatic disease[54,55].

372 *Streptococcus* spp. were also inversely associated with androgenic steroids and sphingomyelins.
373 Androgenic steroids, which have a similar structure to bile acids, are recycled through enterohepatic
374 circulation, which is partially regulated by gut microbiota[56]. Low levels of dehydroepiandrosterone
375 (DHEA) have previously reported to associate with an increased risk of death caused by CVD in
376 elderly men[57]. The role of sphingomyelins in the cardiovascular outcomes remains controversial. Some
377 studies reported higher levels of sphingomyelin in coronary heart disease (CHD) patients compared to
378 controls and these elevated levels were associated with earlier subclinical atherosclerosis[58–61].
379 However, other studies suggest that long-chain saturated sphingomyelins may protect from CHD
380 incidence[62], while follow-up studies indicated no association between plasma sphingomyelins and
381 incidence of CHD[63]. Collectively, these observations show that *Streptococcus* spp. are correlated with
382 many metabolites that were suggested in previous studies to be related to CVD.

There are some limitations in the current study. First, few participants presented high levels of subclinical atherosclerosis, and thus limiting statistical power, which was however counteracted by the size of the cohort analyzed that represents a population at least seven times larger than those analyzed previously. Second, microbial composition can vary extensively throughout the gastrointestinal tract and quantification of the microbial communities in fecal samples represents the microbial population at the distal colon, but does not comprise other sites[64] such as the small intestine, meaning that our study design might not allow discovery of species that are underrepresented in fecal samples. Third, we were unable to validate 64% of our findings in the discovery in the validation attempt. Here a more similar study of larger size would have been useful. As the validation was performed in a small case-control study of symptomatic disease, we could not determine whether the lack of replication for some of our findings was caused by lack of true association, lack of power, differences in study design, and/or due to cross-country differences. Fourth, our study does not take into consideration the different interactions among bacterial species such as synergistic effects in the relationship with coronary atherosclerosis. Finally, the cross-sectional study design prevented causal inferences. Different causal inference methods could be used in future studies to disentangle the underlying relationships and determine whether the identified species and suggested mediators are causally related to atherosclerosis development.

In conclusion, by combining data from a large population-based cohort study and highly accurate bioimaging to evaluate subclinical coronary atherosclerosis, we identified seven *Streptococcus* spp. associated with CACS, biomarkers of inflammation, and with their oral counterparts. These *Streptococcus* spp. may affect the atherosclerosis plaque development by direct infection or alteration of host metabolism. Future studies investigating the causal relationship in these associations will show whether these species can be used as potential biomarkers or treatment targets.

Online Methods

Study design and participants. SCAPIS was used as the primary data source. SCAPIS is a national Swedish general population study including 30,000 subjects aged 50–64 years, at six study sites,

409 focusing on phenotypes relevant to CVD, chronic obstructive pulmonary disease, and related
410 metabolic disorders[18]. Participants from the Uppsala (n=4,541) and Malmö (n=4,432) centers were
411 included in the present study after excluding 846 participants with prevalent CVD (self-reported
412 myocardial infarction, angina, atrial fibrillation, heart valve disease, previous bypass surgery or
413 percutaneous coronary intervention, revascularization of other arterial vessel, and stroke) before the
414 baseline visit or missing information on country of birth or CACS. All participants gave written
415 informed consent before participation. This investigation followed the principles expressed in the
416 Declaration of Helsinki and was approved by the Swedish Ethical Review Authority
417 (Etikprövningsmyndigheten Dnr 2010-228-31M, Dnr. 2018/315).

418 The results were validated in a published case-control study[13] of 214 cases and 171 controls of Han
419 Chinese origin. The phenotypic data for the validation study are publicly available through the
420 *curatedMetagenomicData* R package[65]. In this study, patients and controls aged 40–107 were recruited
421 at the Medical Research Center of Guangdong General Hospital. Cases showed clinical manifestations
422 of stable angina, unstable angina, or acute myocardial infarction. The diagnosis was confirmed by
423 coronary angiography, and individuals with ≥50% stenosis in a single or multiple vessels were
424 included in the study. At the time of medical examination, controls were free of any clinical symptom
425 of ACVD including peripheral artery disease (coronary artery disease or myocardial infarction),
426 cardiomyopathy, renal failure, peripheral neuropathy, systemic disease, and/or stroke[13].

427 We investigated the association between *Streptococcus* spp. located in the gut and in the oral cavity in
428 the MODS, a sub-study of MOS. MOS was performed in 2013–2021 including 5,259 adults (18–71
429 years old), which consisted of children and grandchildren from participants examined at the baseline
430 (1992–1996) of the Malmö Diet Cancer Study Cardiovascular Arm, and it aimed to identify gene-
431 environment interactions of major diseases. The attendance rate of MOS was 47.9% and details of the
432 study can be found in Brunkwall et al.[66]. Participants attending MOS 2014–2018 were eligible to
433 participate in MODS (n=2,643) after the second visit in MOS. In total 831 individuals were recruited
434 in MODS. The participants underwent a thorough dental examination including clinical examination,
435 panoramic and bite-wing radiography. The Malmö Offspring Dental Study (MODS) was approved by

436 the Regional Ethics committee (Regionala etikprövningsnämden, REPN) in Lund (Dnr 2013/761),
437 which is part of the Malmö Offspring Study (MOS) with ethical approval from Regional Ethics
438 committee (REPN) in Lund (Dnr 2012/594).

439 **CACS determination.** The total calcium score was measured in SCAPIS participants by summing the
440 CACS in the left main coronary artery, anterior descending artery, circumflex artery, and right
441 coronary artery. The specific methods for this measurement are available in Bergström et al.[18].

442 **Carotid plaque assessment.** The numbers of plaques in the left and right carotid arteries were
443 determined in SCAPIS participants from two-dimensional grey-scale ultrasound images obtained
444 using a standardized protocol with a Siemens Acuson S2000 ultrasound scanner equipped with a 9L4
445 linear transducer (both from Siemens, Germany) following the Mannheim consensus[67]. Carotid plaque
446 presence was categorized in three categories. None: absence of identified plaques in both vessels;
447 unilateral: identified plaques in one of the two vessels; and bilateral: identified plaques in both
448 vessels[18].

449 **Inflammation and infection markers.** Clinical chemistry, including hsCRP, was carried out on
450 venous blood from SCAPIS participants. A blood cell count, including white blood cell differential,
451 was also performed.

452 **Oral health phenotypes.** Five trained dentists performed the dental examination in MODS
453 participants. Caries was detected using standard clinical criteria aided by mirror, probe (Hu-Friedy
454 EXD57), and bite-wing radiographs. Cavitated lesions that extend into the dentin were recorded as
455 manifest caries and a primary lesion not reaching the stage of manifest as initial. Initial and manifest
456 lesions were summed for a combined variable of surfaces with caries. Filled surfaces included both
457 fillings and crowns. Both caries and fillings were recorded on all teeth, counting five surfaces.
458 Gingival inflammation was recorded as percentage of bleeding on probing excluding wisdom teeth
459 and counting on six surfaces per tooth using a Hu-Friedy PCPUNC157 probe.

460 **Other phenotypes.** In SCAPIS the sociodemographic, lifestyle, health, and cardiovascular risk factor
461 information were collected using validated and standardized questionnaires[18]. Self-reported

cardiovascular, Crohn's, and ulcerative diseases, as well as self-reported medication for high blood pressure, dyslipidemia, and diabetes were categorized as binary variables. The recruitment center was also categorized as a binary variable (Malmö or Uppsala). Self-reported smoking was categorized as never, former, and current smoker. The participants were grouped into sedentary, moderate exercise, moderate but regular exercise, and regular exercise and training in leisure time categories according to self-reported physical activity. The participants' country of birth was categorized as Scandinavia, Europe, Asia, and other. The Scandinavia group included participants who were born in Sweden, Denmark, Norway, or Finland. BMI was determined by dividing the weight (measured in kg) by the square of the height (measured in meters). The participants who received PPI drugs were classified into a binary category, using plasma metabolomics information. The participants who had measurable omeprazole and/or pantoprazole levels in the plasma over the detection limit were classified as PPI drug users. Total energy intake and fiber intake were derived from the food frequency questionnaire. Both variables were natural log-transformed. Participants who reported values of ln(total energy intake) over or below the geometric mean of ln(total energy intake) ± 3 standard deviations in the population were excluded. Linkage with the drug prescription register was performed for antibiotic drug use (Anatomical Therapeutical Chemical code J01). Those participants who received a prescription for these drugs during the year preceding their attendance of the baseline visit were classified as participants with antibiotic drug treatment.

In the validation case-control study[13] , the sociodemographic information was collected using questionnaires. No information on Crohn's and ulcerative diseases, cardiovascular medication, and physical activity was available. Smoking exposure was available only for 8.8% of the participants and none of them were under antibiotic treatment.

In the MOS population, the sociodemographic, lifestyle, and medication treatment were collected using validated and standardized questionnaires. Country of birth was categorized in two levels (Sweden or other) and self-reported information of the PPI was used. BMI was measured dividing the weight (measured in kg) by the square of the height (measured in meters). Participants belonging to the same family were registered in a variable called "family id". For the participants included in

MODS, we obtained information using standardized questionnaires filled out during the dental examination. Smoking was categorized as never, former and current smoker. The participants were grouped in primary education, secondary education or university degree according to the level of education acquired. Self-reported information of the antibiotic usage was categorized as binary variable depending on if the participants received antibiotic treatment the last three months. The activity performed by the participant during the previous hour before attending the dental examination included three variables categorized as binary depending on whether the participant did or did not the following actions: eat, smoke, and/or brush their teeth. The oral hygiene was assessed using the Löe plaque index and the mean degree of plaque per tooth surface was calculated[68]. Therefore, we summed the number of surfaces with plaques identified during the dental exploration divided by the number of teeth multiplied by six surfaces.

Metagenomics. *General considerations.* For SCAPIS-Malmö and MOS samples, the whole analytical process was performed together in the project "lungut", from DNA extraction to relative abundance calculation for each identified species, at Clinical Microbiomics A/S (Copenhagen, Denmark), following standardized methodology. The samples from lungut and SCAPIS-Uppsala were randomized on box-level (16 samples per box) and all the samples from these three studies were processed together during the years 2019 and 2020. MODS saliva samples were also carried out in the same company following the same pipeline during the year 2020, but it was not processed together with the three other studies. For the validation study[13], the analytical pipeline for the metagenomics analysis from the FASTQ files to the relative abundance calculation for each identified species was also performed in Clinical Microbiomics A/S (Copenhagen, Denmark).

Handling and analyses of SCAPIS, MOS and MODS. Clinical Microbiomics A/S (Copenhagen, Denmark) processed three different projects that were: lungut (fecal samples from SCAPIS-Malmö and MOS samples), SCAPIS-Uppsala (fecal samples) and MODS (saliva samples). For lungut and SCAPIS-Uppsala, DNA was extracted from fecal samples using NucleoSpin® 96 Soil (Macherey-Nagel, Germany) from the same batch (Lot: 1903/001) to limit the technical bias. For MODS, DNA was extracted from 250 µL saliva using the same tools. At least one negative control (no sample

material) was included per batch of sample during the extraction process. One positive control (mock) was included per batch during the whole laboratory process for all the projects, including DNA sequencing. DNA extraction quality was evaluated using agarose gel electrophoresis and the quantity was determined by Qubit 2.0 fluorometer for the three projects. The genomic DNA was randomly sheared into fragments of approximately 350 bp. The fragmented DNA was used for library construction using NEBNext® Ultra Library Prep Kit for Illumina (New England Biolabs, MA, USA). The sample index pairs were unique for each sample per run. The prepared DNA libraries were purified using AMPure XP kit, and evaluated using Agilent 2100 Bioanalyzer to determine fragment size distribution. Before sequencing, the concentration of the final libraries were determined using quantitative real-time PCR. The libraries were sequenced using an Illumina Novaseq 6000 instrument using 2×150 bp paired-end reads. The sequencing process generated on average 26.3 million read pairs per sample in lungut, 25.3 million read pairs in SCAPIS-Uppsala and 26.3 million read pairs in MODS. Reads with >10% ambiguous bases, or >50% bases with Phred score (Qscore) <5 were removed. On average, 97.9% of the sequenced bases had a Qscore >20 in lungut, 97.8% in SCAPIS-Uppsala and 97.4% in MODS. Reads that mapped the human reference genome GRCh38 using Bowtie 2 v.02.3.4.1[69] (selecting default settings) were removed from FASTQ files. The remaining reads, classified as high-quality non-host reads (NQNH), were mapped to the gene catalog using BWA mem v.0.7.16a. The reads were considered mapped if the following criteria were met: an alignment of ≥100 bases, ≥95% identity in this alignment, mapping quality (MAPQ) ≥20, and ≤10 bases failing to align with the gene sequence at either end. Reads meeting previous criteria except the MAPQ threshold were considered multi-mapped. A gene count table was created with the number of mapped read pairs for each gene. Two specific gene catalogues were built. The first catalogue was built for the fecal samples including 6,813 samples from lungut, 4,876 from SCAPIS-Uppsala, 9,428 from Pasolli et al.[70], and 3,486 publicly available genome assemblies for isolated microbial strains, selected for their relevance or potential relevance to the human gut or because they are used in commercially available mock microbial communities. The second catalogue was built for the saliva samples including 706 MODS samples, 1,305 oral samples compiled from 21 publicly available data sets, and 1,326 publicly available genome assemblies from isolated microbial strains, selected for their

544 relevance or potential relevance to the human mouth, and 81 genome assemblies corresponding to
545 commercially available mock microbial communities. The HQNH reads from the samples were
546 assembled using MEGAHIT (v.1.1.1)[71,72] into contigs of ≥500 bp. The contigs from lungut, SCAPIS-
547 Uppsala, Pasolli et al.[70], and genome assemblies were combined, and genes were predicted using
548 Prodigal Gene Prediction Software (v.2.6.3, metagenomics/anonymous mode; ;
549 https://github.com/hyattpd/Prodigal). The contigs from MODS were combined with genome
550 assemblies and genes were predicted using the same software. Genes and partial genes with a length
551 <102 bp were removed, resulting in a set of 2.95×10^9 genes in the human gut catalogue and $1.89\text{x}10^8$
552 genes in the human oral catalogue. For the human gut catalogue, the gene sequences were clustered
553 using MMseqs2 (Release 11)[73] ("pre-clustered" at 98% identity over 95% coverage of the longer
554 sequence (>3 kbp), followed by 93% identity over 70% coverage of the shorter sequence (≤ 3 kbp)).
555 For each cluster, a representative sequence was chosen based on the following criteria: first prioritize
556 sequences derived from metagenome assembly (lungut, SCAPIS-Uppsala and Pasolli[70]) over those
557 derived from isolated strain (genomes); then prioritize sequences representing the largest (cardinality)
558 pre-cluster; then prioritize the longest sequence. The two sets of representative sequences were then
559 re-clustered using the same criteria. The resulting sets of short and long cluster representatives were
560 combined as follows: 1) All short cluster representatives were compared to all long cluster
561 representatives and all alignments at 93% identity over 70% coverage of the shorter sequence were
562 identified. 2) All genes that did not have an alignment were retained. 3) For genes that did have an
563 alignment, the short gene but not the long gene were retained. The resulting set of 33.5 million
564 sequences was then filtered to retain only sequences that represent a cluster with ≥ 1 reference-derived
565 sequence and/or ≥ 5 metagenome-derived sequences, or must have been specifically selected for its
566 relevance, e.g. as a pathogen or as a component of a mock community to build a non-redundant human
567 gut gene catalog (version "HG3A") of 14,147,921 microbial genes. For the human oral catalogue, the
568 gene sequences were clustered using MMseqs2 (Release 11)[73] ("pre-clustered" at 93% identity over
569 70% coverage of the longer sequence, followed by 93% identity over 70% coverage of the shorter
570 sequence). We searched for alignments between long gene cluster representatives and short gene
571 cluster representatives at 93 % identity over 70 % coverage of the shorter sequence; for all matches we

572 removed the long gene and retained the short gene. The merged long and short genes were filtered to
573 remove sequences with tetramer entropy below 4, resulting in a non-redundant human oral gene
574 catalog (version "Ho01") of 8,554,253 microbial genes.

575 Metagenomic species (MGS) core gene sets were defined as bins of co-abundant genes identified
576 using gene abundances from the correspondent non-redundant gene catalog across the cohorts that
577 passed the quality assessment according to Nielsen et al.[74]. Species abundance was estimated
578 according to the signature gene set, which was assigned using 100 genes with the highest correlation
579 to the median core gene abundance for each species. A table of species counts taking into account the
580 total gene counts for the signature gene per species was created. A metagenomics species was
581 considered detected if the read pairs were mapped to least three of the 100 signature genes. Species
582 that did not fulfill this criterion were set to 0, resulting in a 99.6% of specificity, according to the
583 internal benchmarks. The species count table was normalized for effective gene length (accounting for
584 the read length). The relative abundance of each species was estimated normalizing it to the sum
585 (100%). All analyses were performed at the species level.

586 For beta diversity analyses and the comparison of carriers and non-carriers between the two center
587 sites, downsized MGS relative abundance data was used. The estimation of downsized MGS was
588 performed by random sampling without replacement from the gene count table corresponding to the
589 signature genes. Both lungut and SCAPIS-Uppsala were downsized to 210,430 reads. One sample
590 from SCAPIS-Uppsala was discarded due to it presented only 1,473 reads mapped to the signature
591 genes and downsizing all the samples to 1,473 reds would result in a huge loss of precision.

592 The taxonomical information was annotated after comparing all the genes on the two catalogues with
593 NCBI RefSeq database[75] for archaea, bacterial, fungal, protozoa, and viral genomes, using BLAST
594 algorithms. Human gut catalogue was compared with NCBI RefSeq downloaded on 02 May 2021 and
595 the human oral catalogue was compared with the version downloaded on 27 January 2020. To
596 annotate at the various taxonomic ranks, we required different levels of identity (95%, 95%, 85%,
597 75%, 65%, 55%, 50% and 45% for subspecies, species, genus, family, order, class, phylum, and
598 superkingdom, respectively) and a minimum of 80 % sequence coverage. If >75% of the MGS genes

mapped to a single species the MGS was annotated to this species. For genus, family, order, class and phylum the thresholds were set to 60%, 50%, 40%, 30% and 25% respectively. Furthermore, at genus and species level the MGS was not annotated to this level if >10% of the genes mapped an alternative species or genus.

The functional annotation was performed by comparing each gene in the catalog to the EggNOG (v. 5.0)[76] orthologous groups database (http://eggnogdb.embl.de/) using EggNOG-mapper software (v. 2.0.1)[77]. This comparison provided annotation to the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) database (https://www.genome.jp/kegg/). The functional potential profile was determined with GMM[24], which includes 103 metabolic pathways that represent a cellular enzymatic process. MGS were assigned to a GMM if they contained at least two-thirds of the KOs required for the functionality of the module. If the module consisted of three or fewer steps, the MGS must contain all the steps. If the module contains alternative paths, the MGS only have to contain one of the paths.

For lungut, SCAPIS-Uppsala and MODS, no detectable levels of DNA were observed for negative controls, while detectable levels of DNA were observed for mock samples. The mock samples showed a coefficient of variation estimated by the Shannon diversity index of 3.30% in lungut and 3.05% in SCAPIS-Uppsala. The coefficient of variation for 158 pairs of biological replicates randomly introduced in the analysis in SCAPIS-Uppsala center (where Clinical Microbiomics was blind to this information) was 1.49%.

Analyses of the validation study. All the FASTQ files from the validation study were directly downloaded from the European Nucleotide Archive (ENA) under the project code “PRJEB21528”. The bioinformatics processing of reads, mapping to the catalogue, MGS count table generation and MGS relative abundance calculation steps were performed following the same algorithm used in the SCAPIS and MOS samples, with the exception of a minimum read-to-gene alignment length >90 bp to accommodate the 2X100 paired-end sequencing in the validation study.

Metabolomics data. Fasting plasma samples were stored at –80 °C until they were processed by Metabolon Inc (Durham, NC, USA), as described by Evans et al.[78]. The order of the SCAPIS samples were randomized and they were analyzed together with quality control standards including pure water, solvents used for metabolite extraction and a pool of human samples maintained by Metabolon Inc (Durham, NC, USA). Proteins from these samples were removed by precipitating them adding methanol and applying vigorous shanking using Glem Mills GenoGrinder 200 and centrifugation. The metabolite identification was carried out under different settings using Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer (MS) interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The settings were a reverse phase (RP)/UPLC-MS/MS method with positive-ion mode electrospray ionization (ESI), a RP/UPLC-MS/MS with negative ion mode ESI, and a Hydrophilic interaction (HILIC)/UPLC-MS/MS with negative ion mode. Then, Metabolon's hardware and software were used to extract the raw data, identify the peaks, and process the specific quality controls. The peak measurement areas for each metabolite were divided by the median peak area of samples in that batch (n=144). The compounds were identified by comparison to Metabolon library, which contains over 3,300 commercially purified standards and recurrent unknown entities. Metabolite quantification was performed according to the area-under-the-curve quantification of the corresponding peaks. For each metabolite, if the metabolite measurement failed to reach the detection threshold were imputed from the minimum observed value for that metabolite. Each metabolite was assigned to a superpathway, which includes broad metabolic pathway terms, and a subpathway, which includes narrow metabolic pathway terms, during the annotation process. Drug metabolites were categorized as binary variables and the remaining metabolites were natural log plus one (ln+1) transformed.

Statistical analysis. *Simulation to determine the main statistical model.* The statistical method for the models in which CACS was the outcome was selected from simulation data. The simulated dataset was built by shuffling randomly the first delivered data (n=438) to simulate the null hypothesis. We ran 12 models using different transformations in CACS and in microbial species: Linear model with

ln+1 transformation on CACS and microbial species; linear model with bootstrapping standard errors with ln+1 transformation on CACS and microbial species; linear model with bootstrapping based on the residuals with ln+1 transformation on CACS and microbial species; linear model using robust standard errors with ln+1 transformation on CACS and microbial species; linear model with ln+1 on CACS and a center log ratio transformation on the species; negative binomial model with bootstrapping standard errors with ln+1 transformation on the microbial species; negative binomial model with ln+1 transformation on microbial species; negative binomial model with center log transformation on the microbial species; hurdle negative model in which we applied a logistic regression in the zero part of the hurdle negative model and a negative binomial model in the count part of the hurdle negative binomial model with the species ln+1 transformed; spearman correlation; ordinal regression with a ln+1 transformation on the microbial species; and a two-step model, in which logistic regression was first applied, followed by a negative binomial for CACS>0 with a ln+1 transformed species. The final model was selected according to the performance in the simulation based on the inflation factor and prioritizing methods that allow investigating the associations, in accordance with the hypothetical causal diagram created using DAGitty (www.dagitty.net; **Extended Data Fig. 1**) (regression over correlation), and easy interpretation. The linear model using ln+1 transformation in CACS and in the gut microbiota species performed well and it was one of the easiest regression methods for result interpretation.

Association between gut microbiota diversity and atherosclerosis. Alpha and beta diversity were estimated using the R package *vegan*[79]. Alpha diversity was assessed using Shannon diversity index and beta diversity using Bray Curtis dissimilarity. For alpha diversity, we fitted linear regression model using CACS as the outcome and adjusting the model for age, sex, country of birth, and technical variables including center site and metagenomics extraction plate within each center site (we included an indicator variable for center site and interaction terms between metagenomics extraction plate and center site, but no main effect for metagenomics extraction plate (since this would be redundant)). We further adjusted this model for smoking, physical activity, dietary indicators, and self-reported medication for dyslipidemia, hypertension and/or diabetes. For beta diversity, we performed

PERMANOVA analyses with 9,999 permutations and assessing the marginal effects of the terms with beta-diversity as the outcome in the model. The independent variable was the four categories of CACS and the models were adjusted for the same two set of covariates used in the alpha diversity analyses. Pairwise comparisons were carried out using PERMANOVA models. Multiple testing was adjusted for using 5% Benjamini-Hochberg FDR.

Association between gut microbiota species and coronary atherosclerosis. In the discovery cohort, a series of linear multivariable regressions using ln+1 of CACS as the dependent variable and ln+1 of the 1,985 microbial species as independent variable. We fitted a model for each species separately. Two sets of covariates were selected according to the assumptions of the causal framework (**Extended Data Fig. 1**). The basic model was adjusted for age, sex, country of birth, and technical variables including center site, metagenomics extraction plate within each center site and Shannon diversity index. The full model was additionally adjusted for smoking, physical activity, ln(total energy intake), ln(fiber intake), and medication (self-reported medication prescribed for dyslipidemia, high blood pressure, and/or diabetes). Multiple testing was adjusted for using 5% Benjamini-Hochberg FDR. The analyses were performed jointly as well as sex-stratified. To test if the estimate effects were different in the sex-stratified analyses the Cochran's Q-test was used.

We used GSEA from the *fgsea* R package[80] to determine whether species associated with CACS were enriched for certain genera. This statistical method was applied on the ranked p-values for positive and negative regression coefficients separately of the associations between species and the phenotype of interest in the basic model. The enrichment p-values were controlled using 5% Benjamini–Hochberg FDR.

Sensitivity analysis. Analyses for species significantly associated with CACS were repeated without adjusting for Shannon diversity index. The analyses were also repeated including BMI and PPI drugs as additional covariates in the full model. The full model was also fitted additionally adjusting for traditional cardiovascular risk factors (BMI, diabetes, systolic blood pressure, diastolic blood pressure, total cholesterol, cholesterol in low-density lipoproteins, and cholesterol in high-density lipoproteins) to test if the effect of the gut microbiota species on coronary atherosclerosis was independent of these

705 traditional cardiovascular risk factors. We also performed sensitivity analyses excluding data for
706 participants with Crohn's disease, ulcerative disease and for participants who underwent antibiotic
707 treatment during the year preceding their attending the baseline visit. The associations from the main
708 analysis (basic and full model using linear regression) were tested using partial Spearman correlations
709 to ensure consistency using different statistical methods. In case of a discrepancy between the main
710 model (linear regression) and the sensitivity model (Spearman correlation) a study of influential
711 observations that may drive the linear association was performed by calculating unscaled dfbeta
712 values. If the most influential observation had higher dfbeta values compared to the regression
713 coefficient and the values were in the same direction as the regression coefficient, the finding from the
714 linear regression was deemed unreliable. The residuals from the association between the species
715 considered as reliable and CACS using linear regression were plotted against the exposure to identify
716 any trend between the two variables.

717 *Association between gut microbiota species and carotid atherosclerosis.* To assess the association
718 between CACS-related species and carotid atherosclerosis, a series of ordinal multivariable
719 regressions were performed using the presence of carotid plaque as the dependent variable for carotid
720 atherosclerosis and the microbiota species significantly associated with CACS as the independent
721 variable including only one species in each model. The relative abundance of these species were ln+1
722 transformed. The models were adjusted using the covariates in the basic and full model. Multiple
723 testing was adjusted using 5% Benjamini-Hochberg FDR.

724 *Validation in a case-control study with symptomatic atherosclerotic cardiovascular patients.*
725 Multivariable logistic regressions were fitted with disease status as the binary outcome. The exposure
726 variables were the species with adjusted p-value <0.05 from the basic model used in the discovery
727 cohort that were available in the validation dataset. We did a rank-based inverse normal transformation
728 of the relative abundance of these species for the analysis. The models were adjusted for age, sex and
729 Shannon diversity index. Those associations with a p-value adjusted for Benjamini–Hochberg <0.05
730 and showing an effect in the same direction as we observed in the discovery cohort were considered as
731 validated.

Functional analysis based on gut metabolic modules. GSEA was used to assess if species belonging to a specific functional GMM were enriched in their association with CACS compared to species in the other functional models. The analysis was performed on ranked p-values for positive and negative regression coefficients separately, and controlled using 5% Benjamini–Hochberg FDR. To determine if these results were driven by an enrichment of species belonging to a same genus we carried out a repeated GSEA but applying a leave-one-(taxon)-out in the analysis, which consist in removing the species belonging to one genus each time.

Associations between Streptococcus spp. and plasma metabolites. Partial Spearman correlations were fitted using the R package *ppcor*[81] to assess the correlations between the significant CACS-associated *Streptococcus* spp. with plasma metabolites controlling for multiple testing using 5% Benjamini–Hochberg FDR. The models were adjusted using the same covariates as previously defined in the full model. GSEA was also performed on ranked p-values for positive and negative correlation coefficients separately to evaluate if the streptococci associated with plasma metabolites were enriched for certain metabolite subpathway. The enriched p-values were controlled using 5% Benjamini-Hochberg FDR. To investigate the possible mediation effect of these metabolites in the association between *Streptococcus* spp. and CACS we additionally adjusted the full model for metabolites involved in enriched subpathways with <30% missing values and adjusted p-value <0.05.

Associations between Streptococcus spp. and systemic inflammatory and infection biomarkers. A series of linear multivariable regressions were fitted to assess the association between CACS-related *Strepptococcus* spp. (independent variable) and hsCRP and counts of neutrophils and leukocytes (outcomes) controlling for multiple testing using Benjamini-Hochberg at 5% FDR level. The relative abundance of these species were ln+1 transformed, while the outcome variables were natural log transformed. The models were adjusted for the same covariates as the full model. These models further adjusted for BMI and PPI were also performed.

Associations between gut and oral Streptococcus spp. and between oral Streptococcus spp. and oral health phenotypes. To investigate the association between CACS-related *Streptococcus* spp. in the gut with their homologue in the oral cavity, a series of partial Spearman correlation with cluster-robust

standard errors (family id as a cluster) were fitted adjusted for age, sex, country of birth and gut metagenomics extraction plate. The models were not adjusted for oral metagenomics extraction plate because previous quality controls showed no significant effect from the plate on the oral microbiome. The oral *Streptococcus* spp. associated with an adjusted p-value <0.05 with their homologue in the gut were then associated with three oral health phenotypes consisting of filling surfaces, caries and gingival inflammation using ordinal regressions. The oral health phenotypes were the dependent variable in the regressions and the ln+1 transformation of the relative abundance of the *Streptococcus* spp. the independent variable. These ordinal regression with cluster-robust standard errors (family id as a cluster) were adjusted for age, sex, smoking, education, oral hygiene, activity realized the hour before attending to the dental examination, and oral Shannon diversity index. The models were additionally adjusted for BMI, PPI and antibiotic treatment. A 5% Benjamini–Hochberg FDR was applied to denote statistical significance in the analyses.

Data availability.

The availability of individual data are limited due to the sensitive nature of the data and it can only be used with previous ethical approval. Therefore, the metagenomics sequences has been anonymized and they are available in ENA under accession code “PRJEB51353”. A subset of anonymized metabolomics data (n=125) is available in MetaboLights under the study identifier “MTBLS407”. However, all the data used in this work is available from the authors upon reasonable request and with previous written permission from the Swedish Ethical Review Authority and the SCAPIS Data Access Board.

Code availability.

The source code used to generate the results for the analysis is available at https://github.com/MolEpicUU/GUTSY-CACS.

Acknowledgements

The main financial support for the study was from the European Research Council (ERC-2018-STG-801965 (TF); ERC-CoG-2014-649021 (MO-M) and ERC-STG-2015-679242 (JGS)). Further funding

785 was also provided from the Swedish Research Council (VR 2019-01471 (TF), 2018-02784 (MO-M),
786 2018-02837 (MO-M), 2019-01015 (JÄ), 2020-00243 (JÄ), 2019-01236 (GE), 2017-02554 (JGS); and
787 EXODIAB 2009-1039 (MO-M)); the Swedish Heart-Lung Foundation [Hjärt-Lungfonden 20190505
788 (TF), 20200711 (MO-M), 20180343 (JÄ), 2019-0526 (JGS), 20200173 (GE)]; the A.L.F.
789 governmental grant (2018-0148 (MO-M)); the Novo Nordisk Foundation (NNF20OC0063886 (MO-
790 M)); the Swedish Diabetes Foundation (DIA 2018-375 (MO-M)); the Swedish Foundation for
791 Strategic Research (LUDC-IRC 15-0067 (MO-M)); Formas (2020-00989 (SA)); Göran Gustafsson
792 foundantion [2016 (TF)]; and Axel and Signe Lagerman's foundation (TF),

793 The main funding body of The Swedish CArdioPulmonary bioImage Study (SCAPIS) is the Swedish
794 Heart-Lung Foundation. The study is also funded by the Knut and Alice Wallenberg Foundation; the
795 Swedish Research Council and VINNOVA (Sweden's Innovation agency); the University of
796 Gothenburg and Sahlgrenska University Hospital; Karolinska Institutet and Stockholm County
797 Council; Linköping University and University Hospital; Lund University and Skåne University
798 Hospital; Umeå University and University Hospital; and Uppsala University and University Hospital.

799 The main funding body of Malmö Offspring Dental Study (MODS) was Oral Health Related Research
800 by Region Skåne (OFRS 422361, OFRS 512951, OFRS 567711, OFRS 655561, OFRS 752071, OFRS
801 853031, OFRS 931171 and OFRS968144). The Malmö Offspring Study (MOS) has been funded by
802 the Research Council of Sweden (VR 521-2013-2756 (PMN)), the Swedish Heart and Lung
803 Foundation [Hjärt-Lungfonden 20150427 (PMN)], and by funds ("ALF") obtained from the local
804 Region Skåne County Council (PMN). In addition, funding has been obtained from Ernhold
805 Lundströms Stiftelse (LB).

806 The computations and data handling were enabled by resources provided by the Swedish National
807 Infrastructure for Computing (SNIC) at Uppsala University (SNIC CENTRE) partially funded by the
808 Swedish Research Council through grant agreement no. 2018-05973. The computations were
809 performed using resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced
810 Computational Science (UPPMAX) under Project SNIC sens2019512.

Author contributions

SA, DJ, GE, JGS, JÄ, MO-M and TF obtained the financial support for the study. SS-B, KFD, UH, JÄ, MO-M and TF planned and designed the study. LB, AM, LL, GöB, JS, JÄ, GE, JGS, JS, MO-M and TF collected the SCAPIS data. LB, PMN, GE and MO-M collected the MOS data. DJ, BK and DE planned and collected MODS data. SP, NN, ACE, JBH, HBN preprocessed samples and data and carried out the metagenomics bioinformatics analyses. SS-B prepared the R scripts and carried out all the associations. SS-B, KFD, DN, DJ, JÄ, MO-M and TF wrote the manuscript. SS-B, GaB and GV created the figures. All authors contributed with the critical interpretation of the results and the manuscript.

Competing interest statement

The authors declare the following competing interests: S.P, N.N., A.E., J.B.H. and H.B.N. are employees for the company Clinical Microbiomics A/S, where the sample processing has been performed from DNA extraction to the estimations of relative abundance for the metagenomics species. J. Ä. has received lecture fees from Novartis and AstraZeneca, and served on advisory boards for AstraZeneca and Boerhinger Ingelheim, all unrelated to the present paper. PMN has received lecture fees from Novartis, Novo Nordisk, Amgen and Boerhinger Ingelheim unrelated to the present paper. The other authors declare no competing interests.

References

1. Roth, G. A. *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology* **76**, 2982–3021 (2020).

2. Jonsson, A. L. & Bäckhed, F. Role of gut microbiota in atherosclerosis. *Nature Reviews Cardiology* **14**, 79–87 (2017).

3. Hashizume-Takizawa, T. *et al.* Oral challenge with Streptococcus sanguinis induces aortic inflammation and accelerates atherosclerosis in spontaneously hyperlipidemic mice. *Biochemical and Biophysical Research Communications* **520**, 507–513 (2019).

4. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444**, 1022-1023 (2006).

5. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).

6. Wang, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).

7. Liu, S., Zhao, W., Liu, X. & Cheng, L. Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-cohort microbial signatures and potential therapeutic target. *FASEB Journal* **34**, 14166–141681 (2020).

8. Aryal, S., Alimadadi, A., Manandhar, I., Joe, B. & Cheng, X. Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. *Hypertension* **76**, 1555–1562 (2020).

9. Liu, F. *et al.* Alterations of Gut Microbiome in Tibetan Patients With Coronary Heart Disease. *Frontiers in Cellular and Infection Microbiology* **10**, 373 (2020).

10. Zheng, Y. Y. *et al.* Gut Microbiome-Based Diagnostic Model to Predict Coronary Artery Disease. *Journal of Agricultural and Food Chemistry* **68**, 3548–57 (2020).

11. Toya, T. *et al.* Coronary artery disease is associated with an altered gut microbiome composition. *PLoS ONE* **15**, e0227147 (2020).

12. Liu, Z. *et al.* The intestinal microbiota associated with cardiac valve calcification differs from that of coronary artery disease. *Atherosclerosis* **284**, 121–8 (2019).

13. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications* **8**, 845 (2017).

14. Liu, H. *et al.* Alterations in the gut microbiome and metabolism with coronary artery disease severity. *Microbiome* **7**, 68 (2019).

15. Fromentin, S. *et al.* Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nature Medicine 2022* **37**, 1–12 (2022).

16. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Communications* **3**, 1245–1253 (2012).

17. Sattar, N. & Preiss, D. Reverse Causality in Cardiovascular Epidemiological Research: More Common Than Imagined? *Circulation* **135**, 2369–2372 (2017).

18. Bergström, G. *et al.* The Swedish CArdioPulmonary BioImage Study: Objectives and design. *Journal of Internal Medicine* **278**, 645–659 (2015).

19. Bergström, G. *et al.* Prevalence of Subclinical Coronary Artery Atherosclerosis in the General Population. *Circulation* **144**, 916–929 (2021).

20. Li, R. *et al.* B-Mode-Detected Carotid Artery Plaque in a General Population. Atherosclerosis Risk in Communitites (ARIC) Study Investigators. *Stroke* **25**, 2377-2383 (1994).

21. Gepner, A. D. *et al.* Comparison of Coronary Artery Calcium Presence, Carotid Plaque Presence, and Carotid Intima-Media Thickness for Cardiovascular Disease Prediction in the Multi-Ethnic Study of Atherosclerosis. *Circulation: Cardiovascular Imaging* **8**, e002262 (2015).

22. Persson, M. *et al.* Soluble urokinase plasminogen activator receptor: a risk factor for carotid plaque, stroke, and coronary artery disease. *Stroke* **45**, 18–23 (2014).

878 23. Sturlaugsdottir, R. *et al.* Prevalence and determinants of carotid plaque in the cross-sectional
879 REFINE-Reykjavik study. *BMJ Open* **6**, e012457 (2016).

880 24. Vieira-Silva, S. *et al.* Species-function relationships shape ecological properties of the human
881 gut microbiome. *Nature Microbiology* **1**, 16088 (2016).

882 25. Lemos, J. A. *et al.* The Biology of Streptococcus mutans. *Microbiology Spectrum* **7** (2019).

883 26. Iturriaga, G., Suárez, R. & Nova-Franco, B. Trehalose Metabolism: From Osmoprotection to
884 Signaling. *International Journal of Molecular Sciences* **10**, 3793–3810 (2009).

885 27. Lè Ne Tournu, H., Fiori, A. & Dijck, P. van. Relevance of Trehalose in Pathogenicity: Some
886 General Rules, Yet Many Exceptions. *PLoS Pathogens* **9**, e1003447 (2013).

887 28. Jönsson, D. *et al.* Periodontal disease is associated with carotid plaque area: the Malmö
888 Offspring Dental Study (MODS). *Journal of Internal Medicine* **287**, 301–309 (2020).

889 29. Ma, J. & Li, H. The Role of Gut Microbiota in Atherosclerosis and Hypertension. *Frontiers in*
890 *Pharmacology* **9**, 1082 (2018).

891 30. Doern, C. D. & Burnham, C.-A. D. It's Not Easy Being Green: the Viridans Group Streptococci,
892 with a Focus on Pediatric Clinical Manifestations. *Journal of Clinical Microbiology* **48**, 3829
893 (2010).

894 31. Ott, S. J. *et al.* Detection of diverse bacterial signatures in atherosclerotic lesions of patients
895 with coronary heart disease. *Circulation* **113**, 929–937 (2006).

896 32. Koren, O. *et al.* Human oral, gut, and plaque microbiota in patients with atherosclerosis.
897 *Proceedings of the National Academy of Sciences* **108**, 4592–4598 (2011).

898 33. Viehman, J. A. *et al.* Chapter 3 - Microbiology of endocarditis. in Infective Endocarditis 43-59
899 (ed. Kilic A, 2022).

900 34. Kesavalu, L. *et al.* Increased atherogenesis during Streptococcus mutans infection in ApoE-null
901 mice. *Journal of Dental Research* **91**, 255–260 (2012).

902 35. Brandsma, E. *et al.* A Proinflammatory Gut Microbiota Increases Systemic Inflammation and
903 Accelerates Atherosclerosis. *Circulation Research* **124**, 94–100 (2019).

904 36. Razonable, R. R. *et al.* Bacteremia due to viridans group streptococci with diminished
905 susceptibility to levofloxacin among neutropenic patients receiving levofloxacin prophylaxis.
906 *Clinical Infectious Diseases* **34**, 1469–1474 (2002).

907 37. Chia, J. S. *et al.* Induction of Cytokines by Glucosyltransferases of Streptococcus mutans.
908 *Clinical and Diagnostic Laboratory Immunology* **9**, 892-897 (2002).

909 38. Chhibber-Goel, J. *et al.* Linkages between oral commensal bacteria and atherosclerotic
910 plaques in coronary artery disease patients. *npj Biofilms and Microbiomes* **2**, 7 (2016).

911 39. Lanter, B. B. & Davies, D. G. Propionibacterium acnes recovered from atherosclerotic human
912 carotid arteries undergoes biofilm dispersion and releases lipolytic and proteolytic enzymes in
913 response to norepinephrine challenge In vitro. *Infection and Immunity* **83**, 3960–3971 (2015)

914 40. Nagata, E., de Toledo, A. & Oho, T. Invasion of human aortic endothelial cells by oral viridans
915 group streptococci and induction of inflammatory cytokine production. *Molecular Oral*
916 *Microbiology* **26**, 78–88 (2011).

41. Snow, D. E. *et al.* The presence of biofilm structures in atherosclerotic plaques of arteries from legs amputated as a complication of diabetic foot ulcers. *Journal Wound Care* **25**, S16–S22 (2016).

42. Leishman, S. J., Do, H. L. & Ford, P. J. Cardiovascular disease and the role of oral bacteria. *Journal of Oral Microbiology* **2** (2010).

43. Nakano, K. & Ooshima, T. Common knowledge regarding prevention of infective endocarditis among general dentists in Japan. *Journal of Cardiology* **57**, 123–130 (2011).

44. Yumoto, H. *et al.* The Pathogenic Factors from Oral Streptococci for Systemic Diseases. *International Journal of Molecular Sciences* **20**, 4571 (2019).

45. Nomura, R. *et al.* Contribution of Severe Dental Caries Induced by Streptococcus mutans to the Pathogenicity of Infective Endocarditis. *Infection and Immunity* **88**, e00897 (2020).

46. Kim, K. *et al.* Severity of dental caries and risk of coronary heart disease in middle-aged men and women: a population-based cohort study of Korean adults, 2002–2013. *Scientific Reports* **9**, 1–7 (2019).

47. Kilian, M. Streptococcus and enterococcus: Pharyngitis; scarlet fever; skin and soft tissue infections; streptococcal toxic shock syndrome; pneumonia; meningitis; urinary tract infections; rheumatic fever; post-streptococcal glomerulonephritis. in *Medical Microbiology: Eighteenth Edition* 183–198 (Elsevier Inc., 2012).

48. Rafferty, B. *et al.* Impact of monocytic cells on recovery of uncultivable bacteria from atherosclerotic lesions. *Journal of Internal Medicine* **270**, 273–280 (2011).

49. Sethi, N., Safi, S., Korang, S. & Hróbjartsson, A. Antibiotics for secondary prevention of coronary heart disease. *Cochrane Database Systemic Reviews* **2021**, CD003610 (2021).

50. Jespersen, C. M. Randomised placebo controlled multicentre trial to assess short term clarithromycin for patients with stable coronary heart disease: CLARICOR trial. *British Medical Journal* **332**, 22–24 (2006).

51. Sharma, D. & Khan, A. U. Antibiotics versus biofilm: an emerging battleground in microbial communities. Antimicrobial Resistance and Infection Control **8**, 76 (2019).

52. Brown, J. M. & Hazen, S. L. Microbial modulation of cardiovascular disease. *Nature Reviews Microbiology* **16**, 171–181 (2018).

53. Zhang, B. C. *et al.* Increased serum bile acid level is associated with high-risk coronary artery plaques in an asymptomatic population detected by coronary computed tomography angiography. *Journal of Thoracic Disease* **11**, 5063–5070 (2019).

54. Steiner, C. *et al.* Bile Acid Metabolites in Serum: Intraindividual Variation and Associations with Coronary Heart Disease, Metabolic Syndrome and Diabetes Mellitus. *PLOS ONE* **6**, e25006 (2011).

55. Charach, G. *et al.* The association of bile acid excretion and atherosclerotic coronary artery disease. *Therapeutic Advances in Gastroenterology* **4**, 95 (2011).

56. Cross, T. W. L., Kasahara, K. & Rey, F. E. Sexual dimorphism of cardiometabolic dysfunction: Gut microbiome in the play? *Molecular Metabolism* **15**, 70-81 (2018).

956 57. Ohlsson, C., Vandenput, L. & Tivesten. DHEA and mortality: What is the nature of the
957 association? *The Journal of Steroid Biochemistry and Molecular Biology* **145**, 248–253 (2015).

958 58. Poss, A. M. *et al.* Machine learning reveals serum sphingolipids as cholesterol-independent
959 biomarkers of coronary artery disease. *The Journal of Clinical Investigation* **130**, 1363–1376
960 (2020).

961 59. Ganna, A. *et al.* Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident
962 Coronary Heart Disease. *PLOS Genetics* **10**, e1004801 (2014).

963 60. Ottosson, F. *et al.* A plasma lipid signature predicts incident coronary artery disease.
964 *International Journal of Cardiology* **331**, 249–254 (2021).

965 61. Nelson, J. C., Jiang, X. C., Tabas, I., Tall, A. & Shea, S. Plasma sphingomyelin and subclinical
966 atherosclerosis: findings from the multi-ethnic study of atherosclerosis. *American Journal of*
967 *Epidemiology* **163**, 903–912 (2006).

968 62. Sigruener, A. *et al.* Glycerophospholipid and Sphingolipid Species and Mortality: The
969 Ludwigshafen Risk and Cardiovascular Health (LURIC) Study. *PLOS ONE* **9**, e85724 (2014).

970 63. Yu, Z., Peng, Q. & Huang, Y. Potential therapeutic targets for atherosclerosis in sphingolipid
971 metabolism. *Clinical Science* **133**, 763 (2019).

972 64. S Leite, G. G. *et al.* Mapping the Segmental Microbiomes in the Human Small Bowel in
973 Comparison with Stool: A REIMAGINE Study. *Digestive Diseases and Sciences* **65**, 2595–2604
974 (2020).

975 65. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature*
976 *Methods* **14**, 1023–24 (2017).

977 66. Brunkwall, L. *et al.* The Malmö Offspring Study (MOS): design, methods and first results.
978 *European Journal of Epidemiology* **36**, 103–116 (2021).

979 67. Touboul, P. J. *et al.* Mannheim Carotid Intima-Media Thickness and Plaque Consensus (2004–
980 2006–2011): An Update on Behalf of the Advisory Board of the 3rd and 4th Watching the Risk
981 Symposium 13th and 15th European Stroke Conferences, Mannheim, Germany, 2004, and
982 Brussels, Belgium, 2006. *Cerebrovascular Disease* **34**, 290 (2012).

983 68. Löe, H. The Gingival Index, the Plaque Index and the Retention Index Systems. The *Journal of*
984 *Periodontology* **38**, 610–616 (1967).

985 69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
986 *2012 9:4* **9**, 357–359 (2012).

987 70. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000
988 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20
989 (2019).

990 71. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced
991 methodologies and community practices. *Methods* **102**, 3–11 (2016).

992 72. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node
993 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
994 *Bioinformatics* **31**, 1674–1676 (2015).

995 73. Steinegger M. & Söding J. Clustering huge protein sequence sets in linear time. *Nature*
996 *Communications* **9**, 2542 (2018).

997 74. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex
998 metagenomic samples without using reference genomes. *Nature Biotechnology* **32**, 822–828
999 (2014).

1000 75. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic
1001 expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).

1002 76. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated
1003 orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**,
1004 D309–D314 (2019).

1005 77. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-
1006 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the
1007 Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).

1008 78. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted
1009 ultrahigh performance liquid chromatography/ electrospray ionization tandem mass
1010 spectrometry platform for the identification and relative quantification of the small-molecule
1011 complement of biological systems. *Analytical Chemistry* **81**, 6656–6667 (2009).

1012 79. Oksanen, J. *et al.* vegan: Community Ecology Package. R package version 2.5-7.
1013 https://CRAN.R-project.org/package=vegan. (2020).

1014 80. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* (2021).

1015 81. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients.
1016 *Communications for Statistical Applications and Methods* **22**, 665–674 (2015).

1017

1018

1019 **Table 1** Descriptive characteristics of participants in the discovery cohort (SCAPIS study)

	Malmö **n=4,541**	**Uppsala** **n=4,432**
CACS[†‡]	0.00 [0.00; 27.0]	0.00 [0.00; 15.0]
Clinical CACS, n (%)[‡]:		
0	2,574 (56.7)	2,781 (62.7)
1–100	1,379 (30.4)	1,187 (26.8)
101–400	398 (8.76)	319 (7.20)
>400	190 (4.18)	145 (3.27)
Carotid arteries with identified plaques, n (%):		
None	1,862 (41.1)	1,959 (44.2)
Unilateral	1,388 (30.7)	1,391 (31.4)
Bilateral	1,275 (28.2)	1,080 (24.4)
Age (years)*	57.3 (4.28)	57.6 (4.39)
Sex, female, n (%)	2,480 (54.6)	2,336 (52.7)
Country of birth, n (%):		
Scandinavia	3,575 (78.7)	4,001 (90.3)
Asia	237 (5.22)	170 (3.84)
Rest of Europe	629 (13.9)	165 (3.72)
Other	100 (2.20)	96 (2.17)
Triglycerides, mmol/L[†]	1.10 [0.80; 1.50]	1.09 [0.83; 1.53]
LDL cholesterol, mg/L*[‡]	3.64 (0.94)	3.57 (0.92)
HDL cholesterol, mmol/L[†‡]	1.60 [1.30; 2.00]	1.40 [1.20; 1.70]
Total cholesterol*	5.48 (1.00)	5.73 (1.05)
SBP, mmHG*[‡]	122 (16.4)	125 (15.9)
DBP, mmHG*[‡]	75 (9.65)	77 (9.79)

BMI, kg/m^{2‡}	27.2 (4.58)	27.0 (4.36)
HsCRP, mmol/L†‡	1.20 [0.60; 2.40]	1.20 [0.59; 2.30]
Neutrophil counts*10^{9}/L†	3.00 [2.40; 3.80]	2.80 [2.30; 3.50]
Leukocyte counts*10^{9}/L†	5.60 [4.70; 6.70]	5.30 [4.60; 6.30]
Total energy intake, Kcal/day†	1,578 [1,220; 2,066]	1,612 [1,268; 2,048]
Fiber intake, g/day†	11.3 [8.58; 14.2]	11.5 [9.01; 14.1]
Diabetes, n (%)	198 (4.55)	171 (4.07)
Crohn's and ulcerative disease, n (%):	43 (1.00)	52 (1.24)
Medication for dyslipidemia, n (%)	292 (6.71)	274 (6.53)
Medication for high blood pressure, n (%)	834 (19.2)	764 (18.2)
Medication for diabetes, n (%)	165 (3.80)	140 (3.34)
Proton pump inhibitor, n (%)	159 (4.51)	122 (2.78)
Antibiotic treatment (J01 class), n (%)	931 (20.5)	819 (18.5)
Smoking, n (%):		
Never smoker	1,948 (44.0)	2,469 (58.6)
Former smoker	1,681 (38.0)	1,351 (32.1)
Current smoker	795 (18.0)	394 (9.35)
Physical activity in leisure time, n (%):		
Sedentary	594 (13.8)	425 (10.2)
Moderate exercise	2,107 (48.9)	1,873 (45.1)
Moderate but regular exercise	1,114 (25.8)	1,357 (32.7)
Regular exercise and training	498 (11.5)	496 (11.9)

1020

1021 * Mean (standard deviation).

1022 † Median [interquartile range].

1023 ‡ CACS, coronary artery calcium score; LDL, low-density lipoprotein; HDL, high-density lipoprotein;

SBP, systolic blood pressure; DBP, diastolic blood pressure; BMI, body mass index; hsCRP, high-sensitivity C-reactive protein.

Fig. 1 Beta diversity across the coronary artery calcium score (CACS) categories. Representation of the two principal coordinates (PCo) based on Bray-Curtis dissimilarity between individuals colored according to clinical categories of CACS, a measure of asymptomatic atherosclerosis. The centroid of the two first coordinates is shown as a triangle. Absent: CACS=0; Mild: 1≤CACS<101; Moderate: 101≤CACS<401; and Extensive: CACS>400.

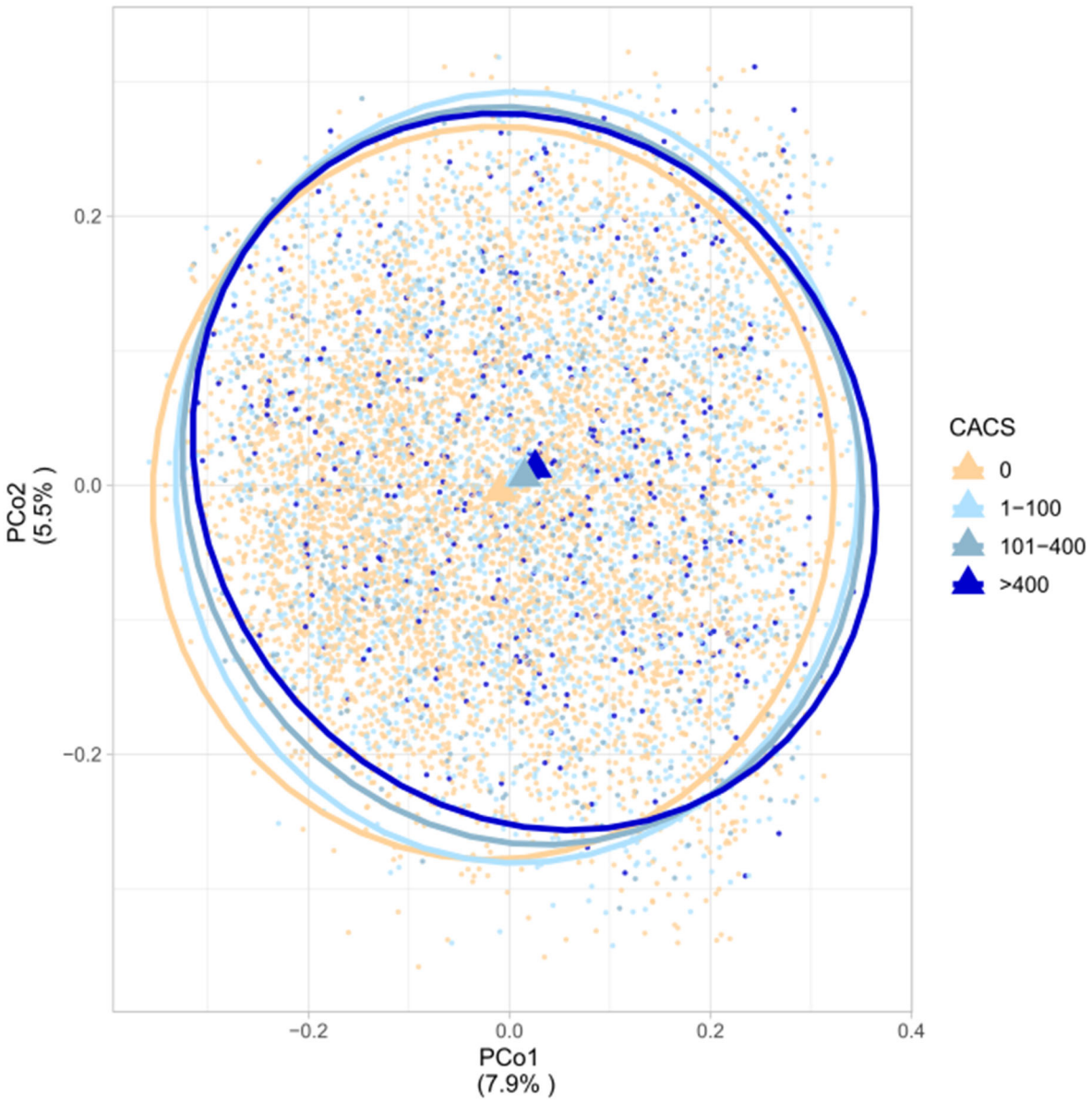


Fig. 2 Metagenomic species associated with coronary artery calcium score (CACS) in the basic model and the full model showed an overrepresentation of *Streptococcus* spp. a, Cladogram of species associated with CACS with a p-value <0.05 (unadjusted for multiple testing) in the basic model. Blue shade indicates the genera overrepresented (FDR<5%) in the enrichment analysis using gene-set enrichment analysis. Only the metagenomics species identified at least to species level were highlighted. Red circles indicate the streptococcal species associated with CACS in the full model (FDR<5%). Turquoise circles indicate the non-streptococcal species associated with CACS in the full model (FDR<5%). **b**, Volcano plot representing all the associations between species and CACS in basic model. Red dots indicate the significant streptococcal species associated with CACS (FDR<5%), turquoise dots indicate significant non-streptococcal species associated with CACS (FDR<5%), and grey dots indicate the associations between non-significant species and CACS (FDR≥5%). **c,** Dot plot showing the results of the enrichment analysis using gene-set enrichment analysis at genus level using the ranked p-values with positive regression coefficients of the associations between species and CACS in the basic model. Modules with less than 15 elements were removed from the plot to improve the visualization. *Streptococcus* was the only enriched genus (FDR<5%) and it is indicated with a blue dot. **d**, Forest plot of the *Streptococcus* spp. associated with CACS in the basic mode and full model with FDR<5%. The dots represent the estimate of the association between the *Streptococcus* spp. and the CACS, and the bar represents the 95% confidence intervals. The estimates and the confidence intervals were ln(x+10) -1 transformed to improve the figure visualization. Orange color represents sex-combined analysis, purple color represents results for the female population, and blue color represents results for the male population. The * indicates the existence of a sex-effect modification in the association between the *Streptococcus spp.* and the CACS (unadjusted for multiple testing).

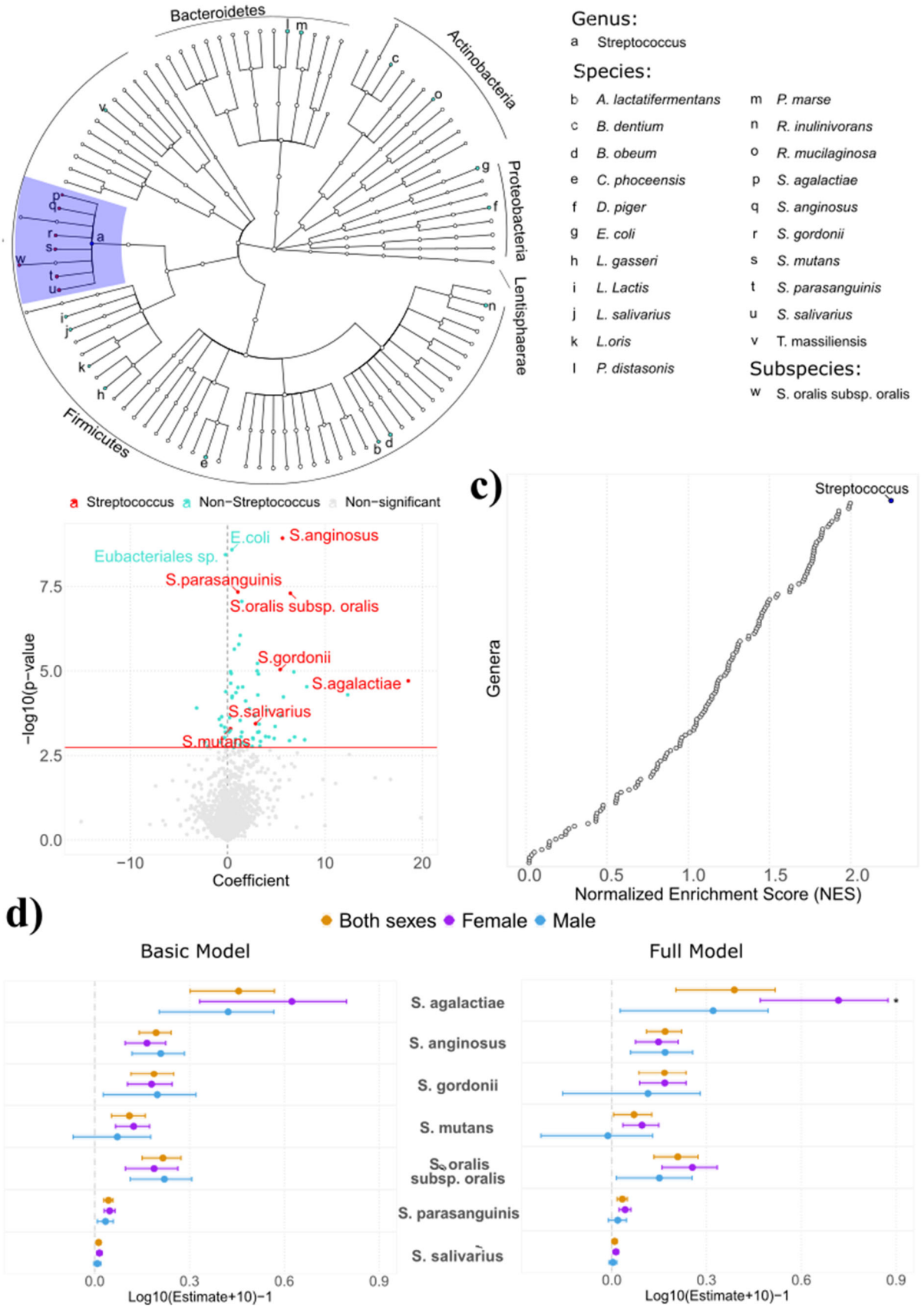
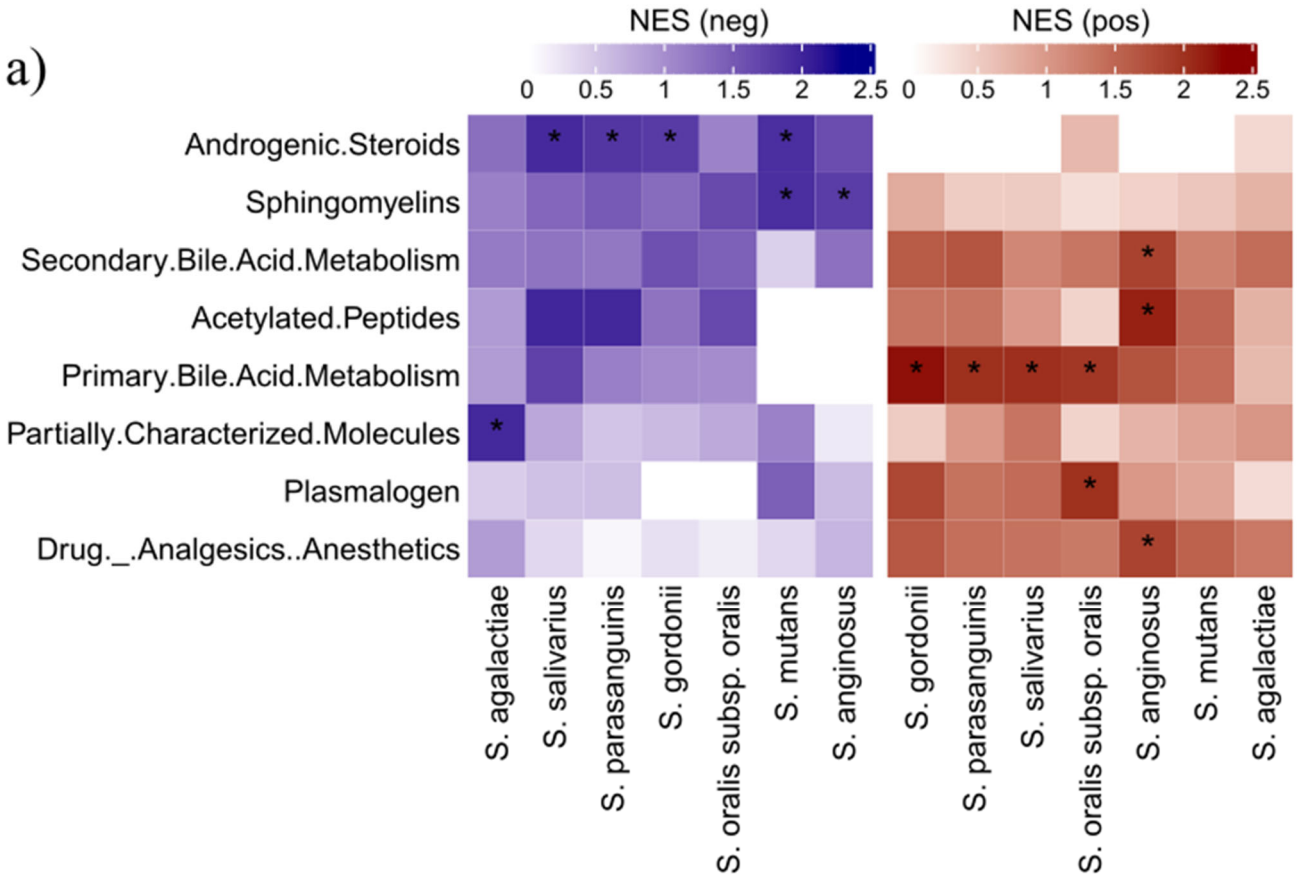
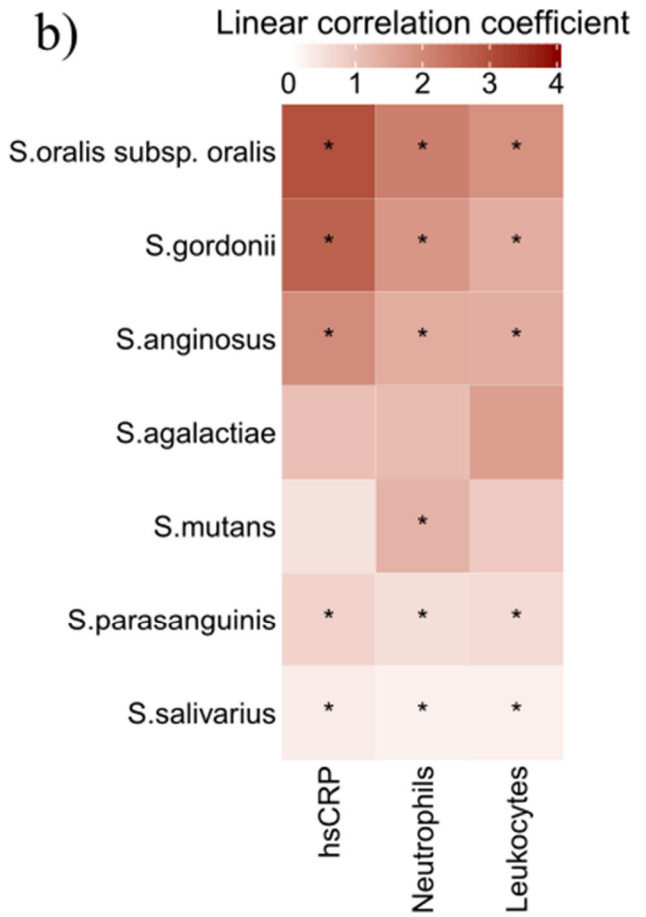
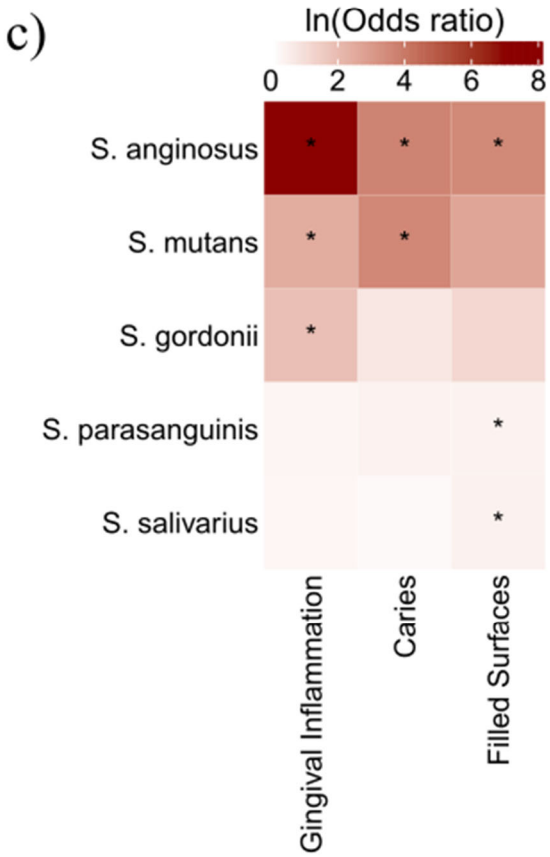


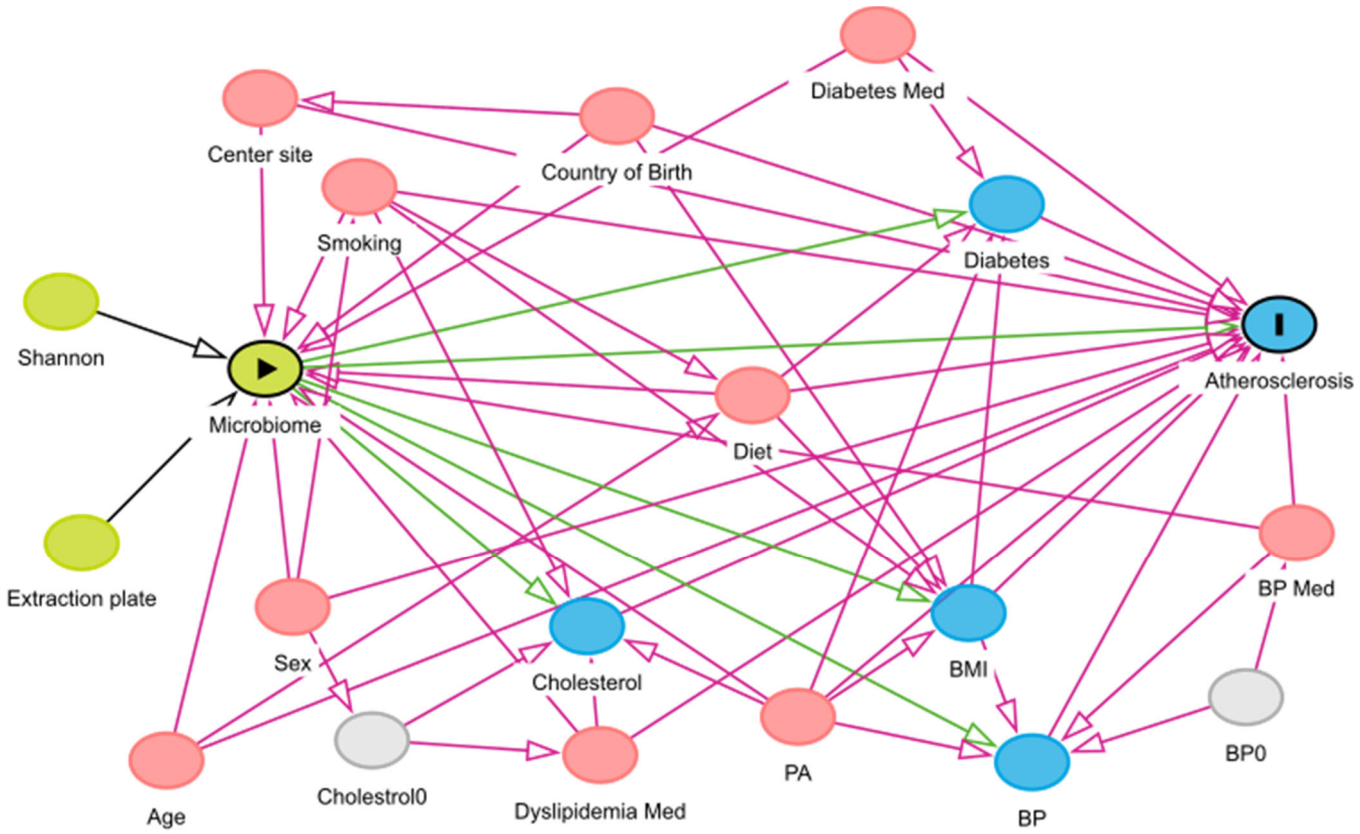
Fig. 3 Heatmaps of associations between CACS-related gut *Streptococcus* spp. and plasma metabolites sub-pathways, inflammatory and infection markers, and oral *Streptococcus spp.* related to oral health phenotypes. **a,** Heatmap showing the enrichment of metabolite sub-pathways based in full adjusted associations between CACS-related gut *Streptococcus* spp. and plasma metabolites. Blue represents the normalized enrichment score (NES) on the enrichment analysis for the negative associations between the *Streptococcus* spp. and plasma metabolites and red represents the NES on the enrichment analysis based on the positive associations. Significant enrichments at 5% FDR are displayed with an asterisk (*). **b,** Heatmap showing linear associations between CACS-related gut *Streptococcus* spp. and three inflammatory markers (hsCRP: high-sensitivity C-reactive protein, neutrophils and leukocytes) adjusted for age, sex, country of birth, center site, metagenomics extraction plate within the center site, Shannon diversity index, smoking, physical activity, fiber and total energy intake, and self-reported medication for dyslipidemia, hypertension and/or diabetes. The three inflammatory markers were scaled to mean of 0 and standard deviation of 1. The heatmap is colored based on the magnitude of the linear correlation coefficient. Significant associations at 5% FDR are displayed with an asterisk (*). **c)** Heatmap showing the associations between oral CACS-related gut *Streptococcus* spp. and three oral health phenotypes (Filled surface, Caries and Gingival inflammation) adjusted for age, sex, smoking, education, oral hygiene, activity realized the hour before attending to the dental examination, and Shannon diversity index. The heatmap is colored based on the natural logarithm of the odds ratio (ln(odds ratio)). Significant associations at 5% FDR are displayed with an asterisk (*).



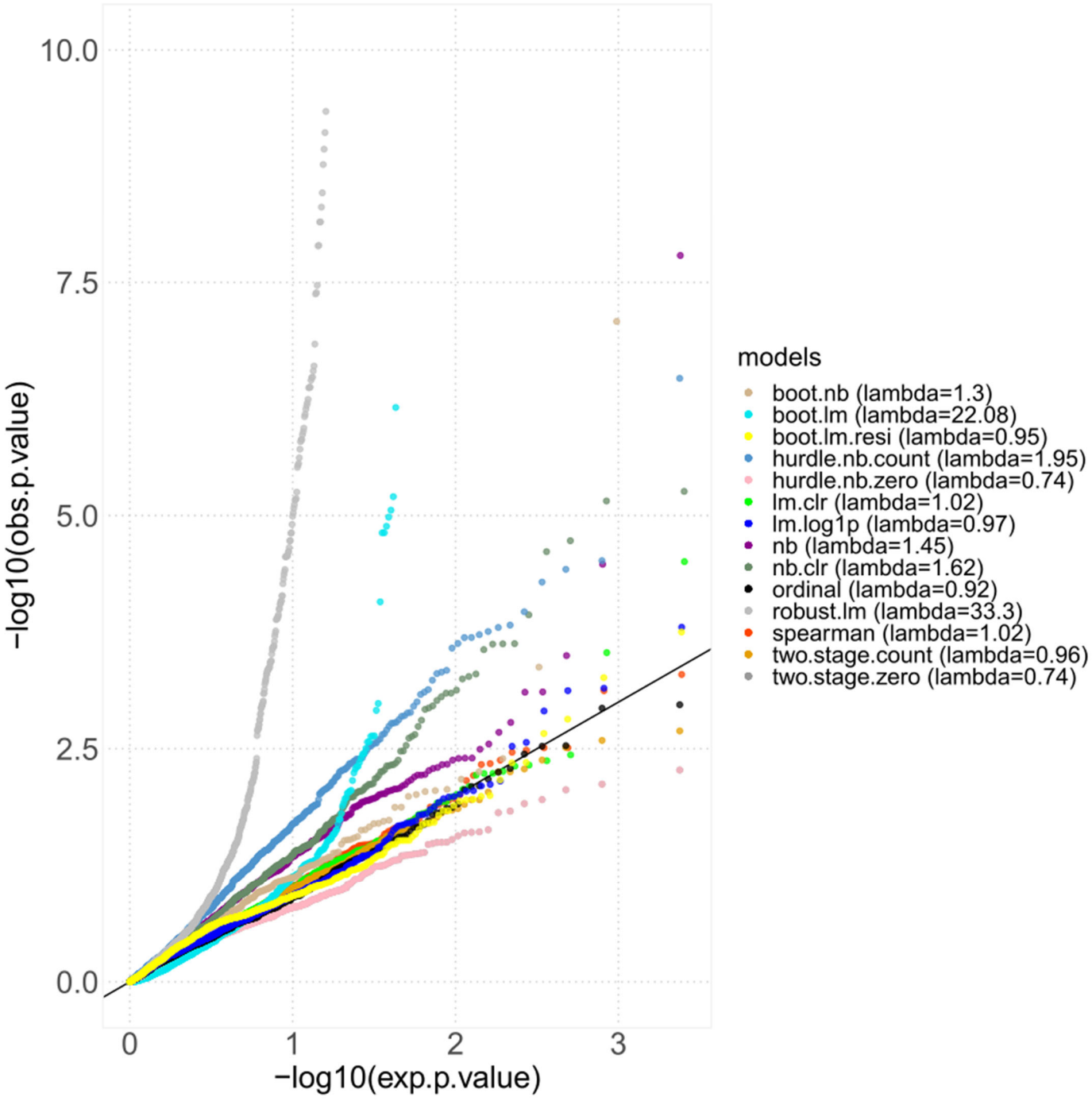




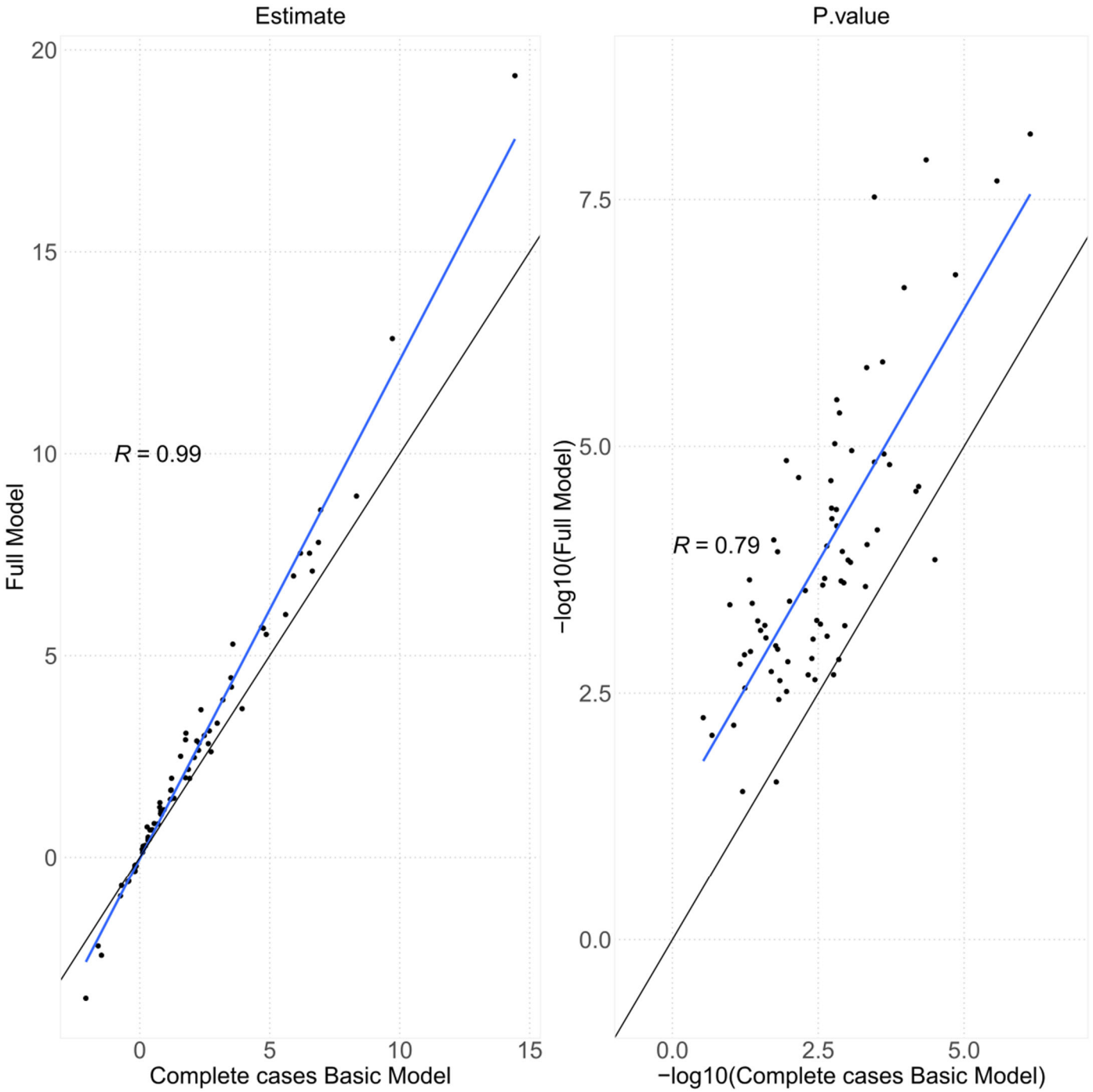
Extended Data Fig. 1 Directed acyclic graph of the assumed framework in the associations between the gut microbiome and atherosclerosis. The graph was generated in DAGitty v3.0 assuming gut microbiota species are causally associated with atherosclerosis. A directed edge (or "arrow") from one node to another represents a direct effect between these two nodes. Red circles represent confounder variables, blue circles mediator variables, green circles technical source of variation, and grey circles unobserved variables. BP, blood pressure; Dyslipidemia Med, medication for dyslipidemia; Diabetes Med, medication for diabetes; BP med, medication for blood pressure; BMI, body mass index; PA, physical activity.



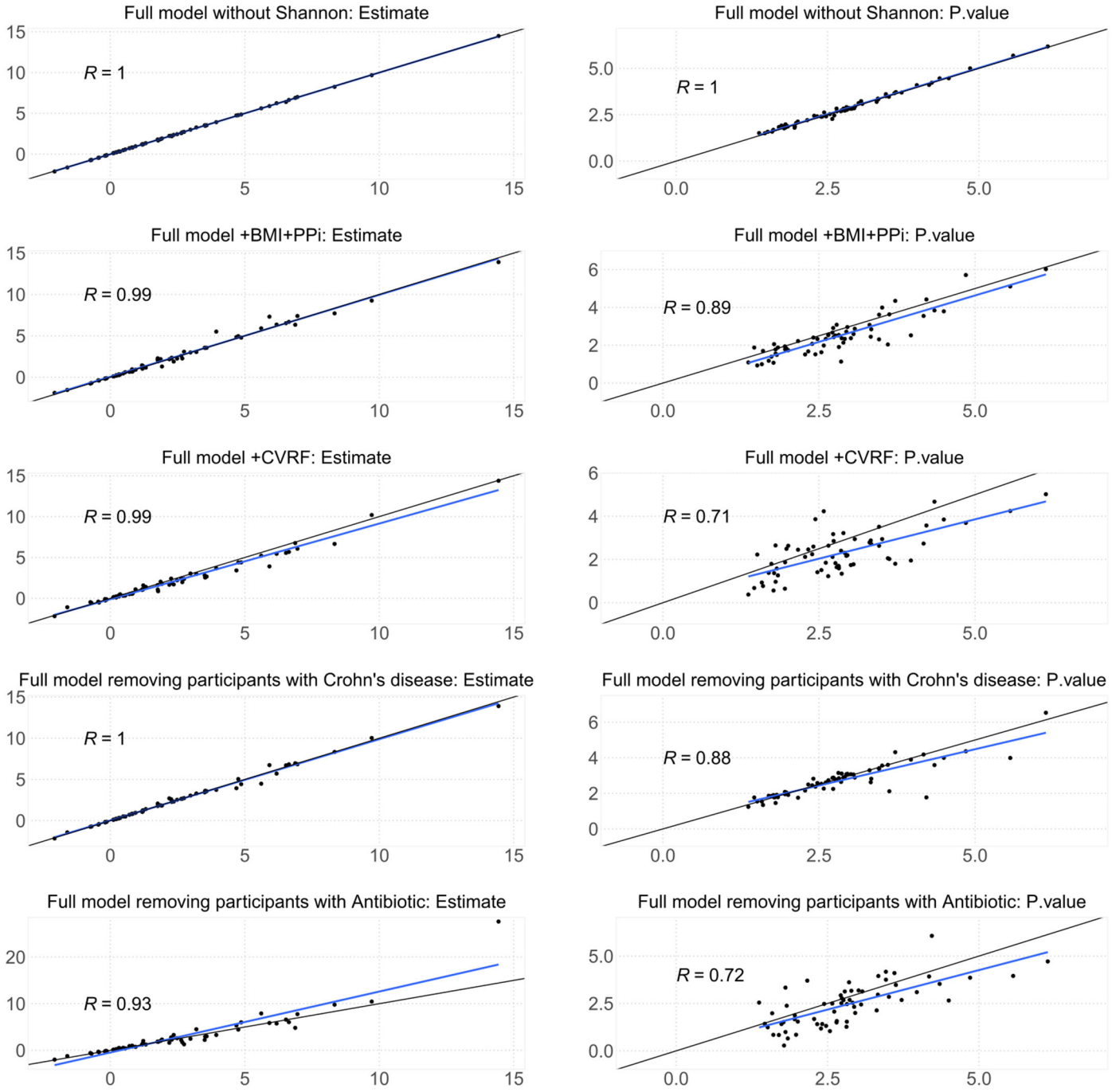
Extended Data Fig. 2 Quantile-quantile plot (qq-plot) of p-values from simulation studies performed prior to the analysis of the relationship between species and coronary artery calcium score (CACS). The first data delivered from Clinical Microbiomics A/S (n=438) was shuffled randomly to create a simulation dataset with maintained variable distribution. The y axis has been truncated at 10 units to improve the visualization of the qq-plots. Then, 12 models were tested to compare their performance in the dataset and are depicted in the qq-plot. Boot.lm, linear model with bootstrapping standard errors; boot.lm.resi, linear model with bootstrapping based on the residuals; boot.nb; negative binomial model with bootstrapping standard errors; hurdle.nb.count, the count part of the hurdle negative binomial model; hurdle.nb.zero, the zero part of the hurdle negative binomial model; lm.clr, linear model transforming the species data using center log ratio transformation; lm.log1p , linear model transforming the species data using natural logarithm of the relative abundance plus one; nb, negative binomial model; nb.clr, negative binomial transforming the species data using center log ratio transformation; ordinal, ordinal regression model, robust.lm, linear model using robust standard errors; spearman, partial spearman correlation; two.stage.count, linear model on the counts different to zero after a first step using natural logistic regression; and two.stage.zero, logistic regression applied before a second step using linear regression on the counts different to zero.



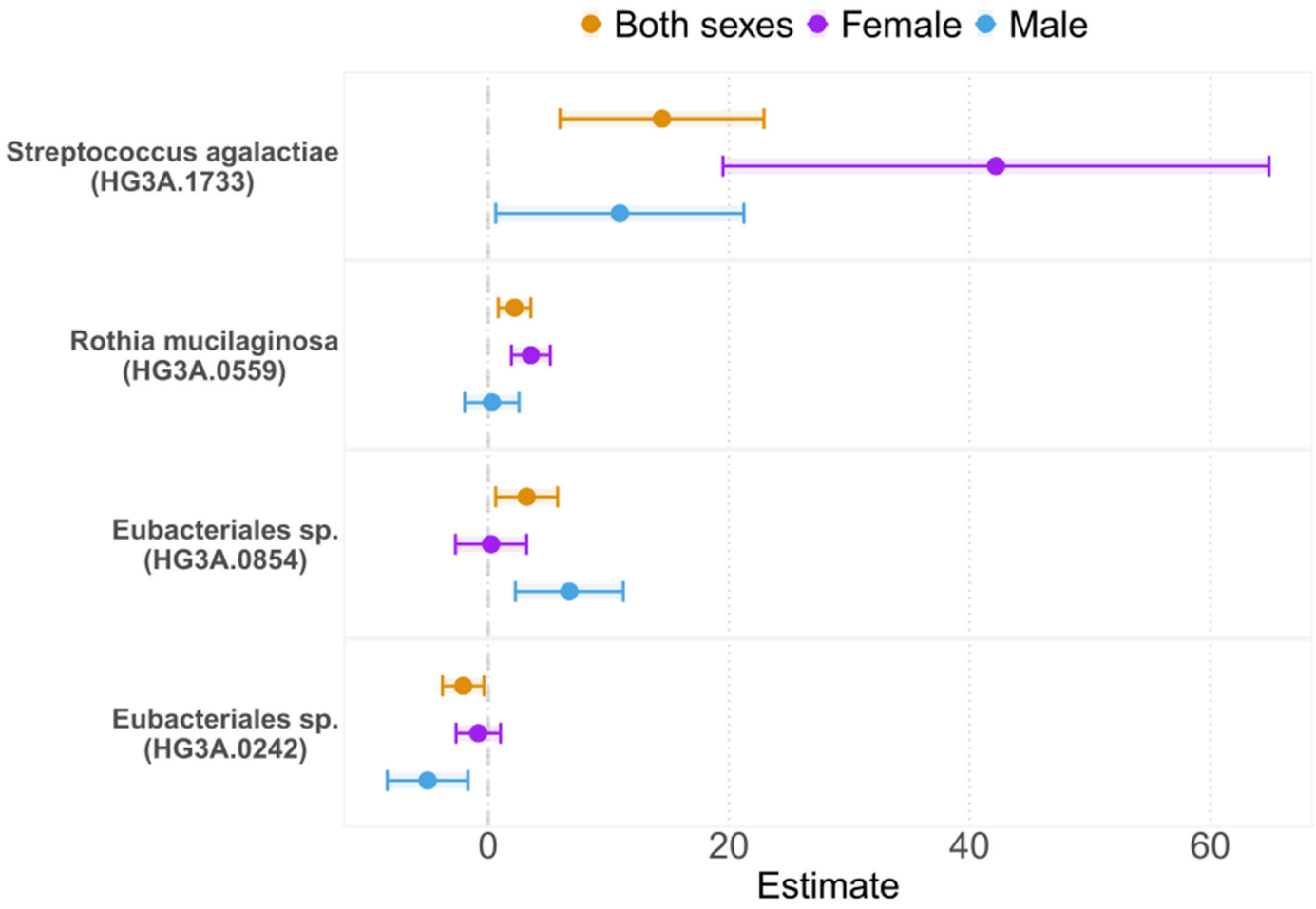
Extended Data Fig. 3 Scatterplot showing the correlation between the estimates and the p-values of the basic model and the full model in 8,155 participants with complete data on all covariates. Blue line indicates the linear regression showing a slightly lower slope compared to the black line, which represents the regression with slope =1. The Pearson correlation (R) between the main model and the sensitivity model is displayed in the figure.



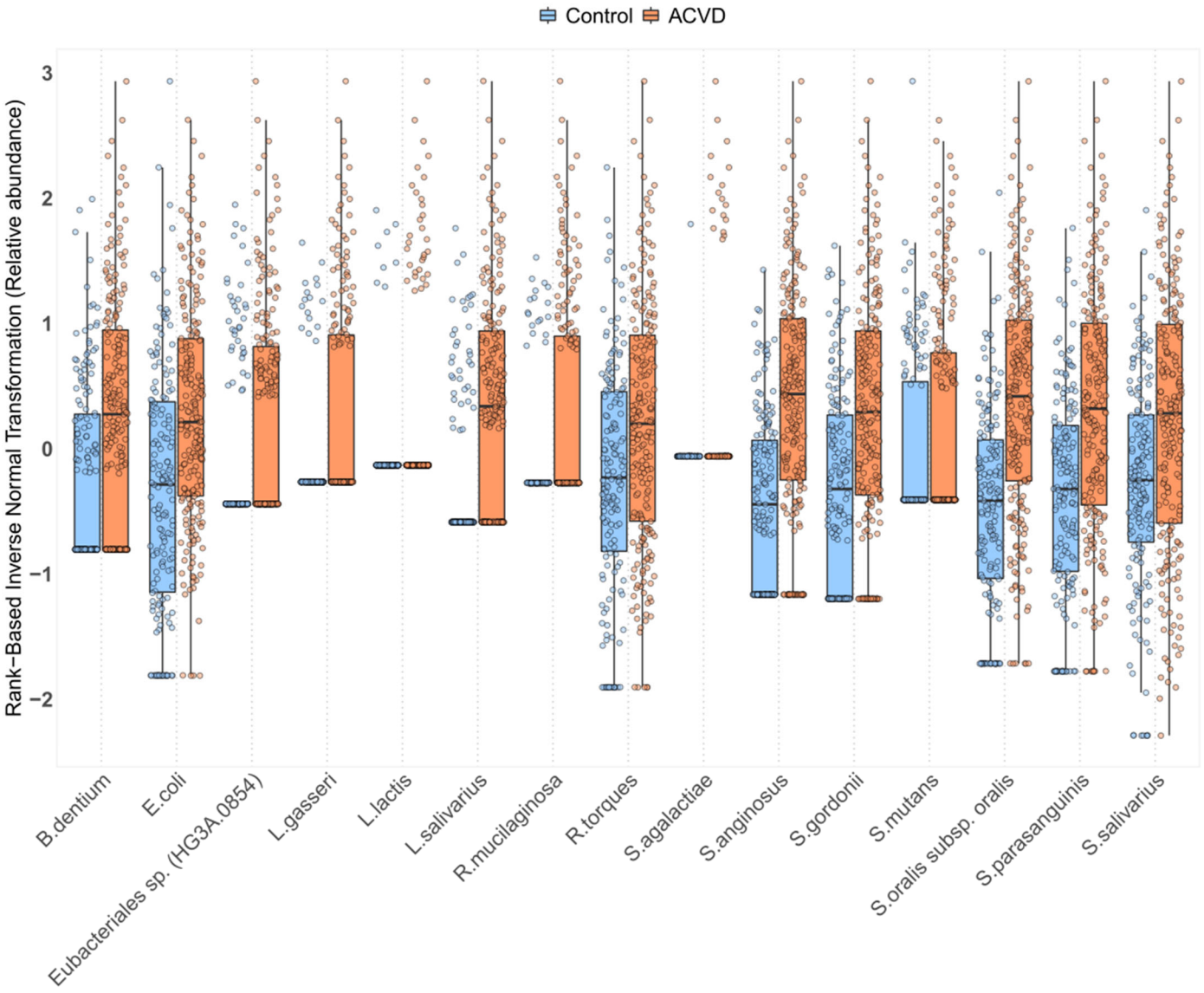
Extended Data Fig. 4 Scatterplot showing the correlation between the main analysis using basic and full models and sensitivity models excluding participants with Crohn's disease, antibiotic drug users, without adjusting for Shannon diversity index, and after adjusting the full model for body mass index (BMI), participants receiving proton-pump inhibitors (PPI) and/or established cardiovascular risk factors (CVRF). Blue lines indicate the linear regression and the black lines represent the regression with slope =1. The Pearson correlation (R) between the main model and the sensitivity model is displayed in the figure.



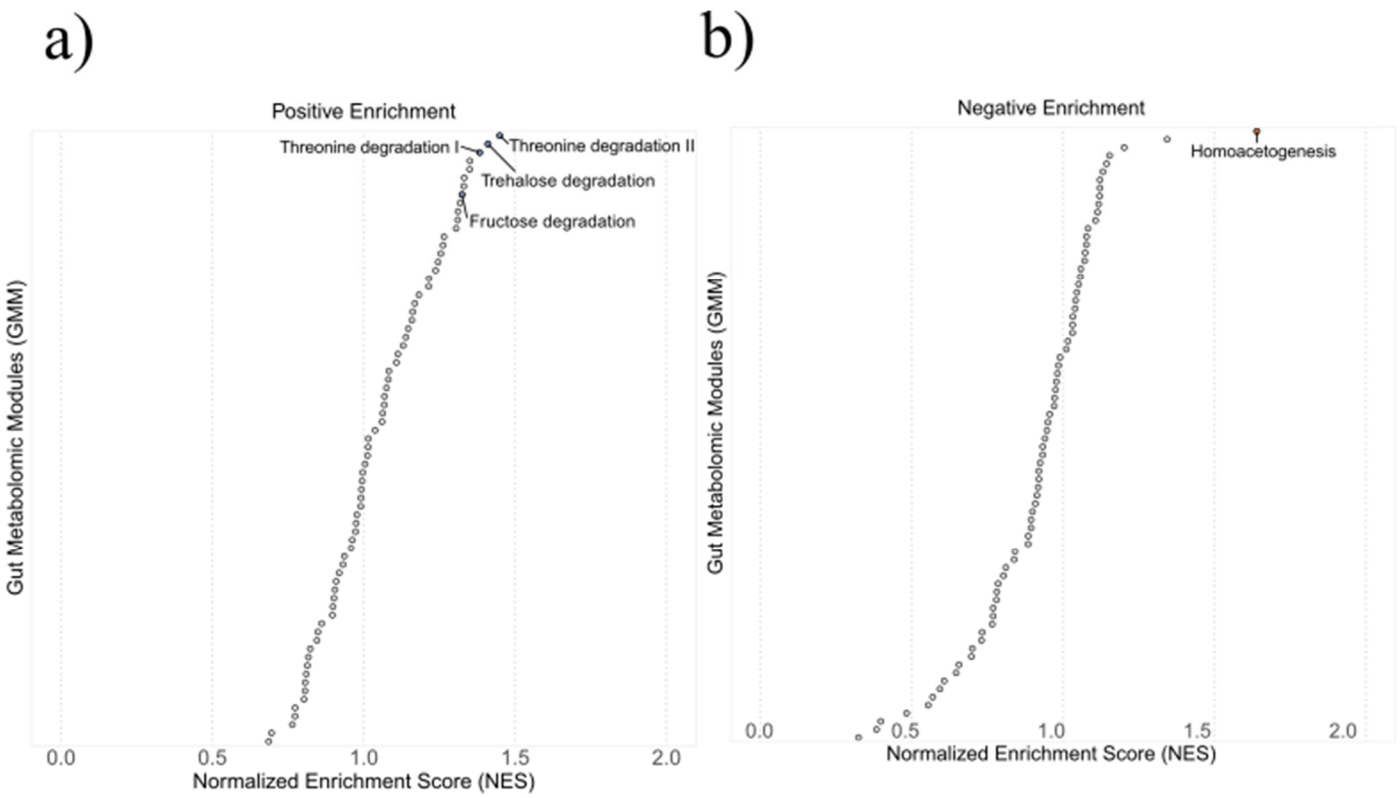
Extended Data Fig. 5 Forest plot of the species associated with CACS showing sex-effect modification with p-value <0.05 (unadjusted for multiple testing) in the full model. The dot represents the estimate of the association between the species and the CACS, and the bar represents the 95% confidence interval. Orange color represents sex-combined analysis, purple color represents results for the female population, and blue color represents results for the male population.

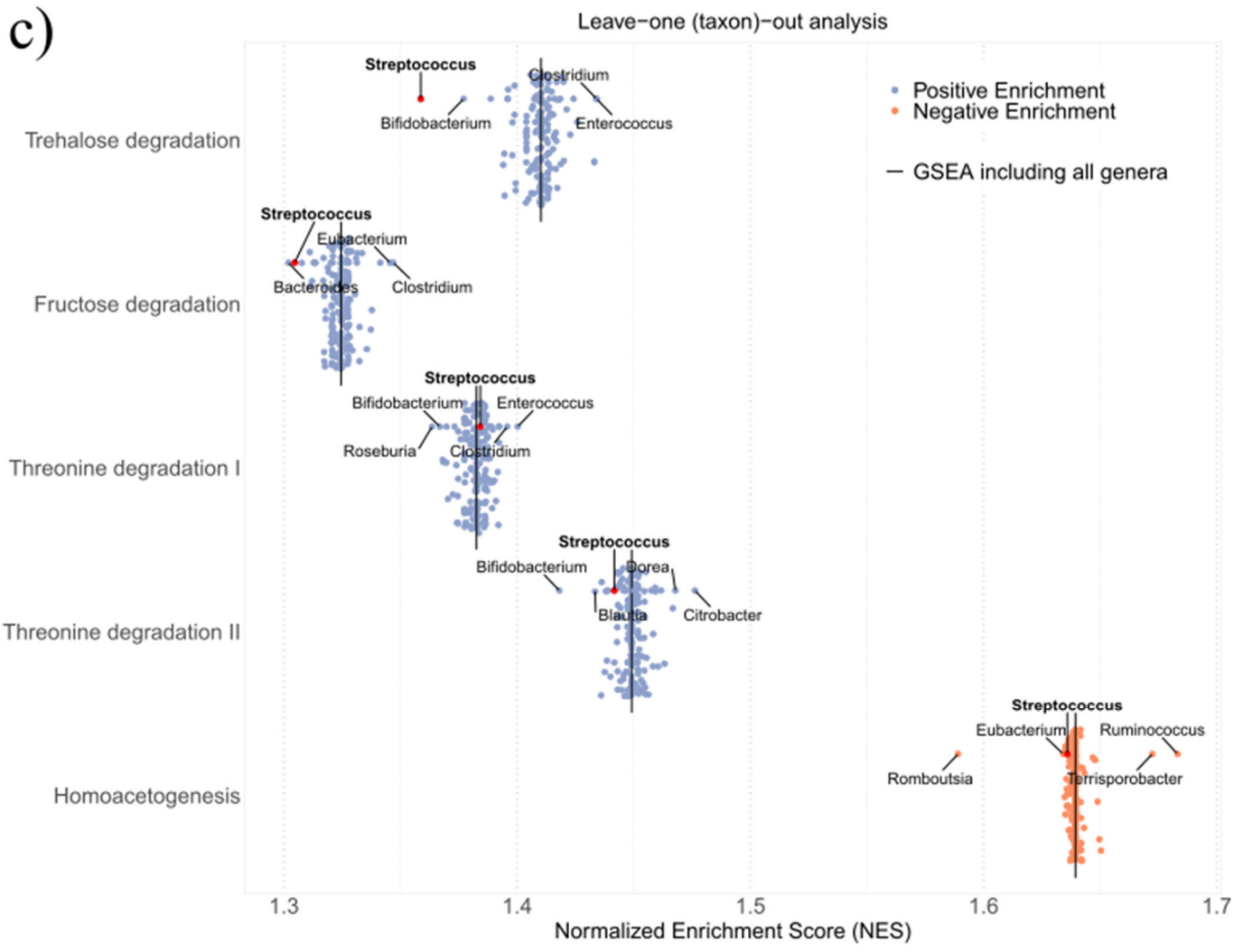


Extended data Fig. 6 The distribution of fifteen species associated with increased coronary artery calcium score in the SCAPIS study over 210 cases and 163 controls of symptomatic atherosclerotic cardiovascular disease in the case-control study by Jie Z et al. The species were identified using the same gene signature as in SCAPIS and the resulting relative abundance were rank-based inverse normal transformed.

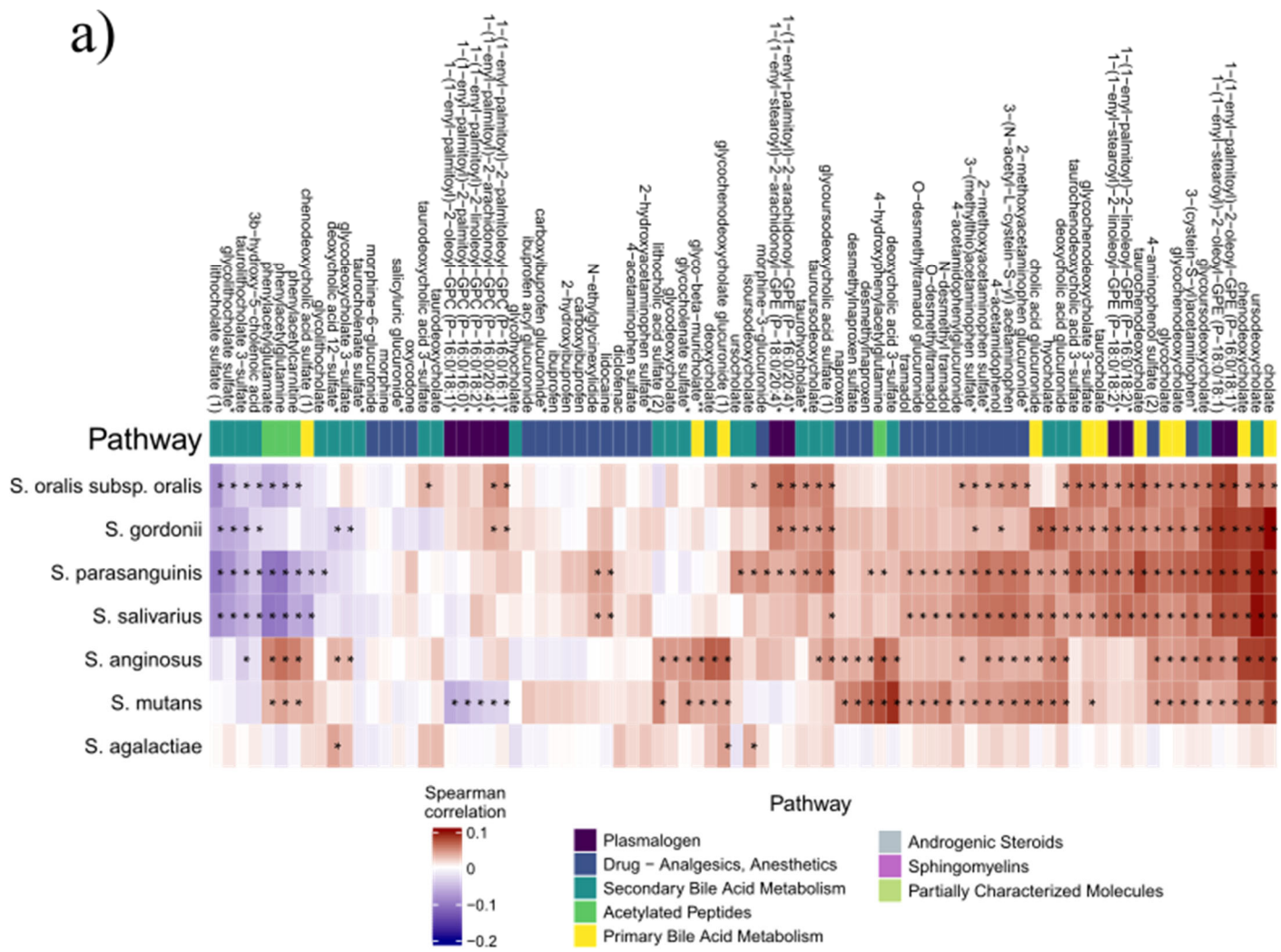


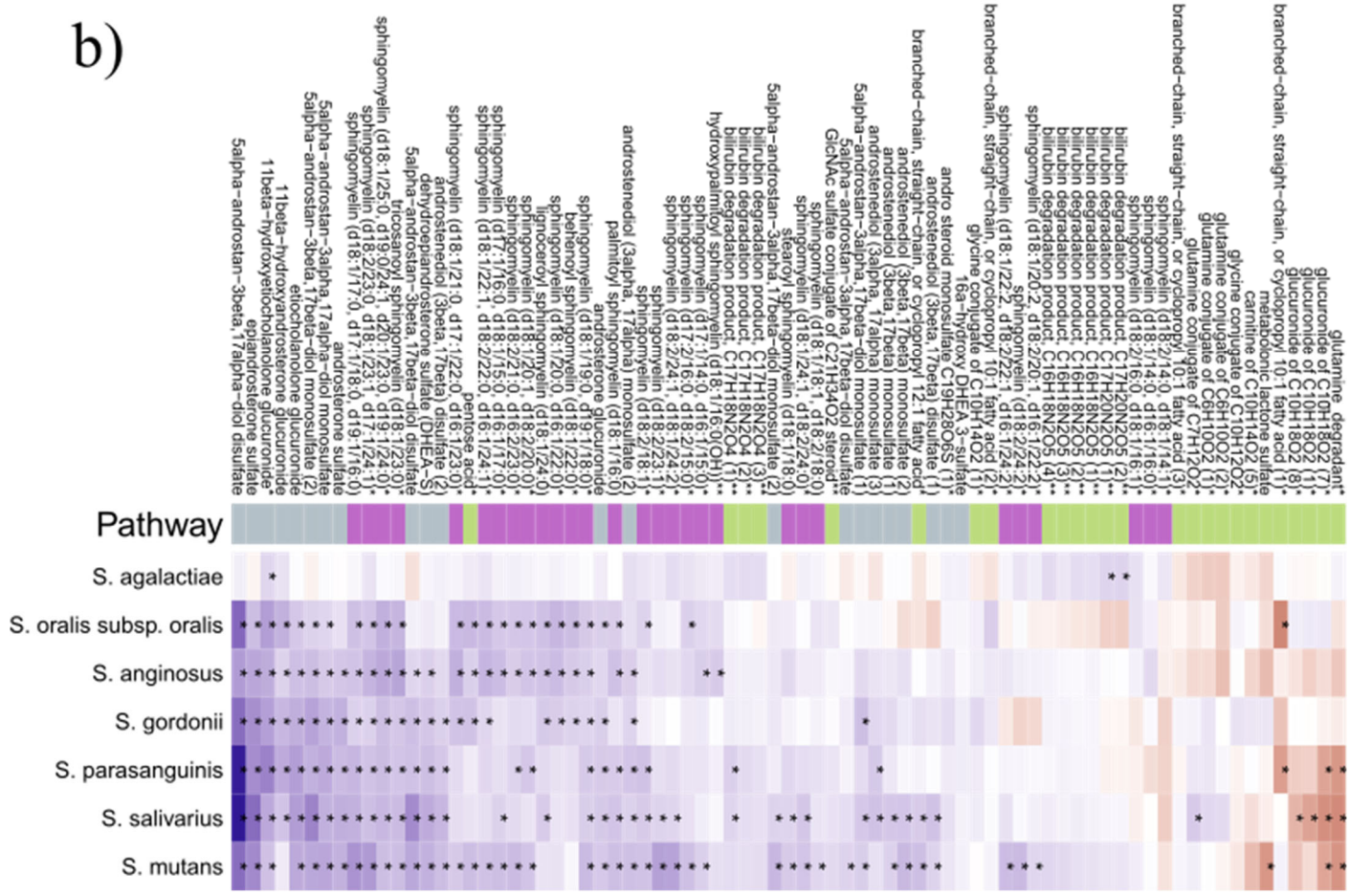
Extended data Fig. 7 Functional gut metabolic modules (GMM) enriched on the positive associations between species and CACS in the basic model. a, Dot plot showing the results of the enrichment analysis using gene-set enrichment analysis (GSEA) to identify functional GMM using the ranked p-values of the positive associations between species and CACS in the basic model. Blue dots indicate the enriched GMM. **b,** Dot plot showing the results of the enrichment analysis using gene-set enrichment analysis (GSEA) to identify functional GMM using the ranked p-values of the negative associations between species and CACS in the basic model. Orange dots indicate the enriched GMM. **c,** GSEA applying leave-one (taxon)-out analysis removing one genus at a time for each enriched GMM. Red dots indicate the analysis removing *Streptococcus* genus and the black lines indicate the enrichment analysis without removing any genus. The genera indicated in the figure are the two removed genus that causes lower normalized enrichment score (NES) or the two removed genus that causes higher normalized enrichment score (NES).



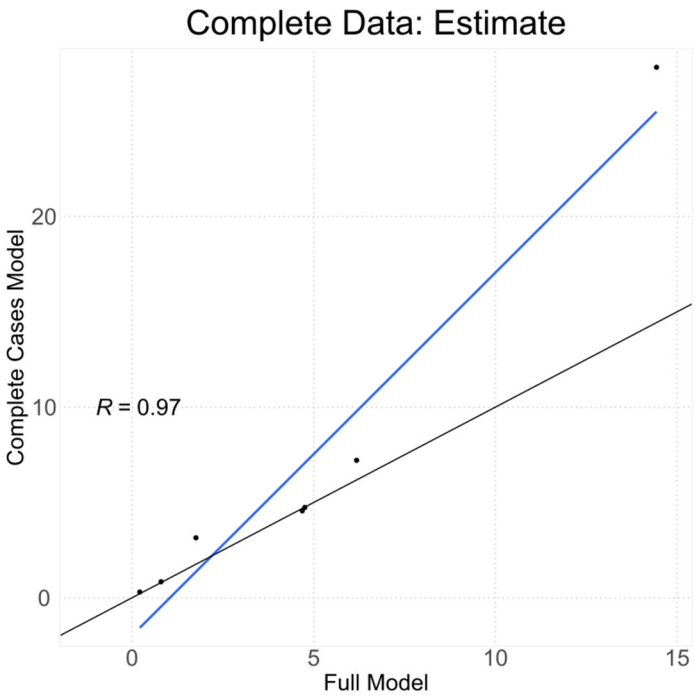


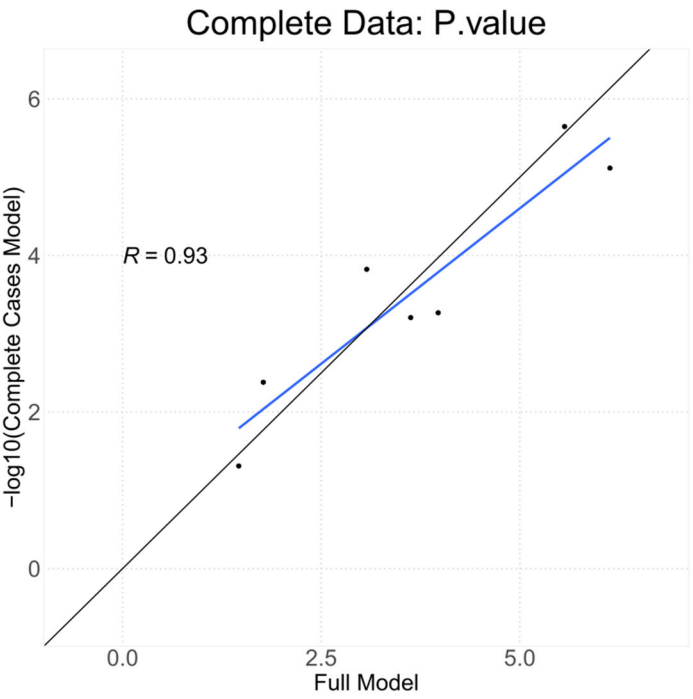
Extended data Fig. 8 Correlations between CACS-associated species and plasma metabolites of enriched metabolic sub-pathways. a, Heatmap showing the partial Spearman correlations between CACS-associated *Streptococcus* spp. and plasma metabolites involved in the significantly enriched metabolic sub-pathway based on the ranked p-values of positive associations adjusted for the covariates used in the full model (model adjusted for age, sex, country of birth, smoking, physical activity, total energy intake, fiber intake, self-reported medication prescribed for dyslipidemia, high blood pressure, and/or diabetes, and technical variables including center site, metagenomics extraction plate and Shannon diversity index). The asterisk (*) indicates the associations significant at 5% FDR. **b,** Heatmap showing the partial Spearman correlations between CACS-associated *Streptococcus* spp. and plasma metabolites involved in the significantly enriched metabolic sub-pathway based on the ranked p-values of negative associations adjusted for the covariates used in the full model. The asterisk (*) indicates the associations significant at 5% FDR.

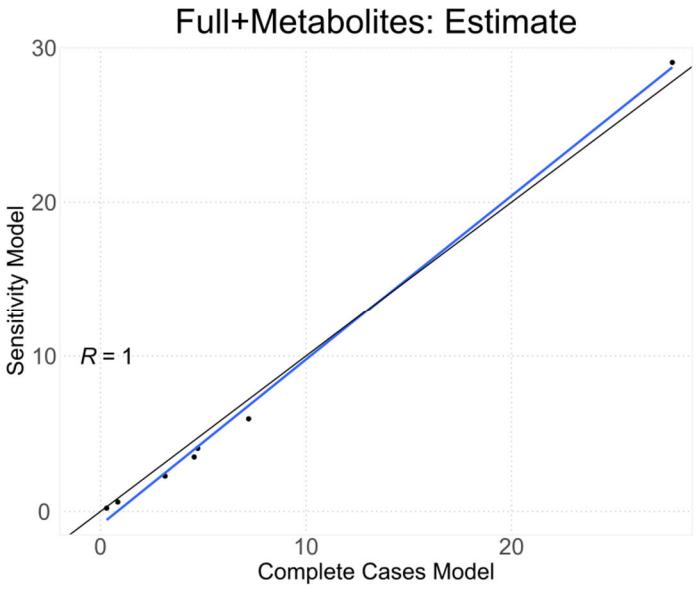


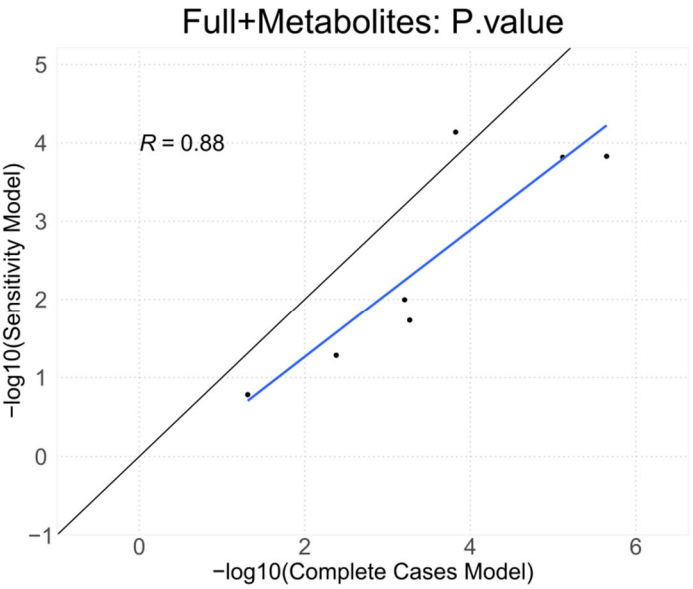


Extended data Fig. 9 Scatterplot showing the correlation between the estimates and the p-values of the associations between *Streptococcus* spp. and CACS in the full model (n=5,683) with and without adjustment for metabolites with <30% missing data involved in sub-pathways related to primary and secondary bile acids metabolism, acetylated peptides, plasmalogen, androgenic steroids, sphingomyelins, analgesic and anesthetic drugs, and/or partially characterized molecules; and the same analysis but restricting the analysis to those participants with complete data. Blue lines indicate the linear regression and the black lines represent the regression with slope =1. The Pearson correlation (R) between the main model and the sensitivity model is displayed in the figure.









Extended data Fig. 10 Directed acyclic graph of the hypothetical causal framework in the association between the oral *Streptococcus* spp. and three phenotypes of oral health (filled surface, caries and gingival inflammation). The graph was generated in DAGitty v3.0 assuming causal association between *Streptococcus* spp. and oral health. A directed edge (or "arrow") from one node to another represents a direct effect between these two nodes. Red circles represent confounder variables, blue circles mediator variables, and green circles technical source of variation. Dyslipidemia Med, medication for dyslipidemia; and Last Hour Act, activity realized the hour before attending to the dental examination.