A sex-specific evolutionary interaction between *ADCY9* and *CETP*

Isabel Gamache[1,2], Marc-André Legault[1,2,3], Jean-Christophe Grenier[1], Rocio Sanchez[1], Eric Rhéaume[1,2], Samira Asgari[4,5], Amina Barhdadi[1,3], Yassamin Feroz Zada[3], Holly Trochet[1,2], Yang Luo[4,5], Leonid Lecca[6,7], Megan Murray[4], Soumya Raychaudhuri[4,5,8,9,10], Jean-Claude Tardif [1,2], Marie-Pierre Dubé[1,2,3], Julie G. Hussin[1,2*]

Affiliations:

[1] Montreal Heart Institute, Montreal, Québec, Canada

[2] Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada

[3] Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada

[4] Center for Data Sciences, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[5] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02115, USA

[6] Socios En Salud, Lima, Peru

[7] Department of Global Health and Social Medicine, Harvard Medical School

[8] Centre for Genetics and Genomics Versus Arthritis, Manchester Academic Health Science Centre, University of Manchester, Manchester M13 9PL, UK

[9] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

[10] Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[*] E-mail : julie.hussin@umontreal.ca

Abstract

Pharmacogenomic studies have revealed associations between rs1967309 in the adenylyl cyclase type 9 (*ADCY9*) gene and clinical responses to the cholesteryl ester transfer protein (CETP) modulator dalcetrapib, however, the mechanism behind this interaction is still unknown. Here, we characterized selective signals at the locus associated with the pharmacogenomic response in human populations and we show that rs1967309 region exhibits signatures of natural selection in several human populations. Furthermore, we identified a variant in *CETP*, rs158477, which is in long-range linkage disequilibrium with rs1967309 in the Peruvian population. The signal is mainly seen in males, a sex-specific result that is replicated in the LIMAA cohort of over 3,400 Peruvians. We further detected interaction effects of these two SNPs with sex on cardiovascular phenotypes in the UK Biobank, in line with the sex-specific genotype associations found in Peruvians at these loci. Analyses of RNA-seq data further suggest an epistatic interaction on *CETP* expression levels between the two SNPs in multiple tissues. We propose that *ADCY9* and *CETP* coevolved during recent human evolution, which points towards a biological link between dalcetrapib's pharmacogene *ADCY9* and its therapeutic target *CETP*.

Introduction

Coronary artery disease (CAD) is the leading cause of mortality worldwide. It is a complex disease caused by the accumulation of cholesterol-loaded plaques that block blood flow in the coronary arteries. The cholesteryl ester transfer protein (CETP) mediates the exchange of cholesterol esters and triglycerides between high-density lipoproteins (HDL) and lower density lipoproteins (1,2). Dalcetrapib is a CETP modulator that did not reduce cardiovascular event rates in the overall dal-OUTCOMES trial of patients with recent acute coronary syndrome (3). However, pharmacogenomic analyses revealed that genotypes at rs1967309 in the *ADCY9* gene, coding for the ninth isoform of adenylate cyclase, modulated clinical responses to dalcetrapib (4). Individuals who carried the AA genotype at rs1967309 in *ADCY9* had less cardiovascular events, reduced atherosclerosis progression, and enhanced cholesterol efflux from macrophages when treated with dalcetrapib compared to placebo (4,5). In contrast, those with the GG genotype had the opposite effects from dalcetrapib. Furthermore, a protective effect against the formation of atherosclerotic lesions was seen only in the absence of both *Adcy9* and *CETP* in mice (6), suggesting an interaction between the two genes. However, the underlying mechanisms linking *CETP* and *ADCY9*, located 50 Mb apart on chromosome 16, as well as the relevance of the rs1967309 non-coding genetic variant are still unclear.

Identification of selection pressure on a genetic variant can help shed light on its importance. Adaptation to different environments often leads to a rise in frequency of variants, by favoring survival and/or reproduction fitness. An example is the lactase gene (*LCT*) (7–11), where a positively selected intronic variant in *MCM6* leads to an escape from epigenetic inactivation of

LCT and facilitates lactase persistence after weaning (12). Results of genomic studies for phenotypes such as adaptation to high-altitude hypoxia in Tibetans (13), fatty acid metabolism in Inuits (14) or response to pathogens across populations (15) have also been confirmed by functional studies (16–20). Thus, population and regulatory genomics can be leveraged to unveil the effect of genetic mutations at a single non-coding locus and reveal the biological mechanisms of adaptation.

When two or more loci interact during adaptation, a genomic scan will likely be underpowered to pinpoint the genetic determinants. In this study, we took an evolutionary approach on the *ADCY9* and *CETP* candidate genes to specifically study their interaction. As a first step, we used a joint evolutionary analysis to evaluate the potential signatures of selection in these genes, which revealed positive selection pressures acting on *ADCY9*. Genetic associations between the two genes are discovered in Peruvians, a population in which natural selection for high-altitude was previously found on genes related to cardiovascular health (21). In a second step, our analyses of large-scale transcriptomics and available phenome-wide resources bring further evidence of an epistatic interaction between *ADCY9* and *CETP*.

Results

Signatures of selection at rs1967309 in *ADCY9* in human populations

The genetic variant rs1967309 is located in intron 2 of *ADCY9*, in a region of high linkage disequilibrium (LD), and harbors heterogeneous genotype frequencies across human populations from the 1000 Genomes Project (Fig. 1a). Its intronic location makes it difficult to assess its

functional relevance but exploring selective signals around intronic SNPs in human populations can shed light on their importance. In African populations (AFR), the major genotype is AA, which is the homozygous genotype for the ancestral allele, whereas in Europeans (EUR), AA is the minor genotype. The frequency of the AA genotype is slightly higher in Asia (EAS, SAS) and America (AMR) compared to that in Europe, becoming the most frequent genotype in the Peruvian population (PEL). Using the integrated haplotype score (iHS) (22), a statistic that enables the detection of evidence for recent strong positive selection (typically when $|iHS| > 2$), we observed that several SNPs in the LD block around rs1967309 exhibit selective signatures in non-African populations ($|iHS_{SAS}| = 2.66$, $|iHS_{EUR}| = 2.31$), whereas no signal is seen in this LD block in African populations (Fig. 1b, Supplementary Fig. 1, Supplementary text). Our analyses suggest that this locus in *ADCY9* has been the target of recent positive selection in several human populations, with multiple, possibly independent, selective signals detectable around rs1967309. However, recent positive selection as measured by iHS does not seem to explain the notable increase in frequency for the A allele in the PEL population ($f_A=0.77$), compared to the European ($f_A=0.41$), Asian ($f_A=0.44$), and other American populations ($f_A=0.54$ in AMR without PEL).

To test whether the difference between PEL and other AMR allele frequencies at rs1967309 is significant, we used the population branch statistic (PBS) (Methods). This statistic has been developed to locate selection signals by summarizing differentiation between populations using a three-way comparison of allele frequencies between a specific group, a closely related population, and an outgroup (13). It has been shown to increase power to detect incomplete selective sweeps on standing variation. Applying this statistic to investigate rs1967309 allele frequency in PEL, we used Mexicans (MXL) as a closely related group and a Chinese population (CHB) as the outgroup (Methods). Over the entire genome, the CHB branches are greater than PEL and MXL branches

(mean$_{CHB}$=0.020, mean$_{MXL}$=0.008, mean$_{PEL}$=0.009), which reflects the expectation under genetic drift. However, the estimated PEL branch length at rs1967309 (Fig. 1c), which reflects differentiation since the split from the MXL population ($PBS_{PEL,rs1967309}$=0.051, empirical p-value = 0.014), surpasses the CHB branch length ($PBS_{CHB,rs1967309}$=0.049, empirical p-value > 0.05), which reflects differentiation since the split between Asian and American populations, whereas no such effect is seen in MXL ($PBS_{MXL,rs1967309}$=0.026, empirical p-value > 0.05), or for any other AMR populations. Furthermore, the PEL branch lengths at several SNPs in this LD block (Fig. 1c) are in the top 5% of all PEL branch lengths across the whole genome ($PBS_{PEL,95th}$ = 0.031), whereas these increased branch lengths are not observed outside of the LD block (Fig. 1c). These results are robust to the choice of the outgroup and the closely related AMR population (Methods). The increase in frequency of the A allele at rs1967309 is also seen in genotype data from Native American populations (23), with Andeans showing genotype frequencies highly similar to PEL (f_A=0.77, Fig. 1a). The PEL population has a large Andean ancestry (Supplementary Fig. 2, Methods) and almost no African ancestry, strongly suggesting that the increase in AA genotype arose in the Andean population and not from admixture with Africans. The PEL individuals that harbor the AA genotype for rs1967309 do not exhibit a larger genome-wide Andean ancestry than non-AA individuals (p-value=0.30, Mann-Whitney U test). Overall, these results suggest that the ancestral allele A at rs1967309, after dropping in frequency following the out-of-Africa event, has increased in frequency in the Andean population and has been preferentially retained in the Peruvian population's genetic makeup, potentially because of natural selection.

Evidence for co-evolution between *ADCY9* and *CETP* in Peru

The pharmacogenetic link between *ADCY9* and the CETP modulator dalcetrapib raises the question of whether there is a genetic relationship between rs1967309 in *ADCY9* and *CETP*, both located on chromosome 16. Such a relationship can be revealed by analyzing patterns of long-range linkage disequilibrium (LRLD) (24,25), in order to detect whether specific combinations of alleles (or genotypes) at two loci are particularly overrepresented. To do so, we calculated the genotyped-based linkage disequilibrium (r^2) between rs1967309 and each SNP in *CETP* with minor allele frequency (MAF) above 0.05. In the Peruvian population, there are four SNPs, (including 2 in perfect LD in PEL) that exhibit r^2 values with rs1967309 that are in the top 1% of r^2 values (Fig. 2a) computed for all 37,802 pairs of SNPs in *ADCY9* and *CETP* genes with MAF>0.05 (Methods). Despite the r^2 values themselves being low ($r^2_{rs158477}$=0.080, $r^2_{rs158480;rs158617}$=0.089, $r^2_{rs12447620}$=0.090), these values are highly unexpected for these two genes situated 50 Mb apart ($p<0.005$) and thus correspond to a significant LRLD signal. We also computed r^2 between the four identified SNPs' genotypes and all *ADCY9* SNPs with MAF above 0.05 (Fig. 2b). The distribution of r^2 values for the rs158477 *CETP* SNP shows a clear bell-shaped pattern around rs1967309 in *ADCY9*, which strongly suggests the rs1967309-rs158477 genetic association detected is not simply a statistical fluke, while the signal in the region for the other SNPs is less conclusive. The SNP rs158477 in *CETP* is also the only one that has a PEL branch length value higher than the 95th percentile, also higher than the CHB branch length value ($PBS_{PEL,rs158477}$= 0.062, Supplementary Fig. 3a), in line with the observation at rs1967309. Strikingly, this *CETP* SNP's genotype frequency distribution across the 1000G and Native American populations resembles that of rs1967309 in *ADCY9* (Fig. 2c).

Given that the Peruvian population is admixed (26), particular enrichment of genome segments for a specific ancestry, if present, would lead to inflated LRLD between these segments (27–30). However, no significant enrichment is seen at either locus and significant LRLD is also seen in the Andean source population (Supplementary text, Supplementary Fig. 4a,b). Furthermore, we see no enrichment of Andean ancestry in individuals harboring the overrepresented combination of genotypes, AA at rs1967309 + GG at rs158477, compared to other combinations (p-value=0.18, Mann-Whitney U test). These results show that admixture patterns in PEL cannot be solely responsible for the association found between rs1967309 and rs158477. Finally, using a genome-wide null distribution which allows to capture the LRLD distribution expected under the admixture levels present in this sample (Supplementary text), we show that the r^2 value between the two SNPs is higher than expected given their allele frequencies and the physical distance between them (genome-wide empirical p-value=0.01, Fig. 2d). Taken together, these findings strongly suggest that the AA/GG combination is being transmitted to the next generation more often (ie. is likely selectively favored) which reveals a signature of co-evolution between *ADCY9* and *CETP* at these loci.

Still, such a LRLD signal can be due to a small sample size (29). To confirm independently the association between genotypes at rs1967309 of *ADCY9* and rs158477 of *CETP*, we used the LIMAA cohort (31,32), a large cohort of 3,509 Peruvian individuals with genotype information, to replicate our finding. The ancestry distribution, as measured by RFMix (Methods) is similar between the two cohorts (Supplementary Fig. 2), however, the LIMAA cohort population structure shows additional subgroups compared to the 1000G PEL population sample (Supplementary Fig. 5). In this cohort, the pair of SNPs rs1967309-rs158477 is the only pairs identified in PEL who shows evidence for LRLD, with an r^2 value in the top 1% of all pairs of SNPs in *ADCY9* and *CETP*

(Supplementary Fig. 6a,b, 7) (*ADCY9/CETP* empirical p-value=0.003). The r^2 test used above is powerful to detect allelic associations, but the net association measured will be very small if selection acts on a specific genotype combination rather than on alleles. In that scenario, and when power allows it, the genotypic association is better assessed by with a χ^2 distributed test statistic (with four degrees of freedom, χ^2_4) comparing the observed and expected genotype combination counts (25). The test confirmed the association in LIMAA (χ^2_4=82.0, permutation p-value <0.001, genome-wide empirical p-value=0.0003, Supplementary text). The association discovered between rs1967309 and rs158477 is thus generalizable to the Peruvian population and not limited to the 1000G PEL sample.

Sex-specific long-range linkage disequilibrium signal

We explored the effect of sex on the LRLD association found between rs1967309 and rs158477. The allele frequencies at rs1967309 were suggestively different between males and females and sex-stratified PBS analyses suggest that the LD block around rs1967309 is differentiated between sexes in the Peruvians (Supplementary Fig. 3a), although we could not exclude the possibility of random sampling noise (Supplementary text). We further performed sex-stratified LRLD analyses, which revealed that the correlation between rs1967309 and rs158477 is only seen in males (Fig. 3a,b, Supplementary text, Supplementary Fig. 8a,b): the r^2 value rose to 0.348 in males (*ADCY9/CETP* empirical p-value=8.23×10^{-5}, genome-wide empirical p-value<2.85×10^{-4}, N=41) and became non-significant in females (*ADCY9/CETP* empirical p-value=0.78, genome-wide empirical p-value=0.80, N=44). This result cannot be explained by differences of Andean ancestry proportion between males and females (p-value=0.27, Mann-Whitney U test). A permutation analysis that shuffled the sex labels of samples established that the observed difference between

197 the sexes is larger than what we expect by chance (p-value=0.002, Supplementary Fig. 8c,
198 Supplementary text). In the LIMAA cohort, we replicate this sex-specific result (Fig. 3c,d) where
199 the r^2 test is significant in males (*ADCY9*/*CETP* empirical p-value=0.003, N=2,078) but not in
200 females (*ADCY9/CETP* empirical p-value=0.66, N=1,434). The genotypic χ_4^2 test confirms the
201 association between *ADCY9* and *CETP* is present in males (χ_4^2 = 56.6, permutation p-value=0.001,
202 genome-wide empirical p-value=0.002, Supplementary text), revealing an excess of rs1967309-
203 AA + rs158477-GG. This is also the genotype combination driving the LRLD in PEL. In females,
204 the test also shows a weaker but significant effect (χ_4^2 = 37.0, permutation p-value = 0.017,
205 genome-wide empirical p-value = 0.001) driven by an excess of a different genotype combination,
206 rs1967309-AA + rs158477-AA, which is however not replicated in PEL possibly because of lack
207 of power (Supplementary text).
208
209 **Epistatic effects on *CETP* gene expression**
210 LRLD between variants can suggest the existence of gene-gene interactions, especially if they are
211 functional variants (29). In order to be under selection, mutations typically need to modulate a
212 phenotype or an endophenotype, such as gene expression. We have shown previously (6) that
213 CETP and ADCY9 interact in mice to modulate several phenotypes, including atherosclerotic
214 lesion development. To test whether these genes interact in humans, we knocked down (KD)
215 *ADCY9* in hepatocyte HepG2 cells (Methods) and performed RNA sequencing on five KD
216 biological replicates and five control replicates, to evaluate the impact of decreased *ADCY9*
217 expression on the transcriptome. We confirmed the KD was successful as *ADCY9* expression is
218 reduced in the KD replicates (Fig. 4a), which represents a drastic drop in expression compared to
219 the whole transcriptome changes (False Discovery Rate [FDR] = 4.07 x 10^{-14}, Methods). We also

observed that *CETP* expression was increased in *ADCY9-KD* samples compared to controls (Fig. 4a), an increase that is also transcriptome-wide significant (FDR=1.97 x 10^{-7}, ß = 1.257). This increased expression was validated by qPCR, and western blot also showed increased CETP protein product (Supplementary text, Supplementary Fig. 9, Supplementary Table 2). Conversely, knocking down or overexpressing *CETP* did not impact *ADCY9* expression on qPCR (data not shown). These experiments demonstrate an interaction between *ADCY9* and *CETP* at the gene expression level and raised the hypothesis that *ADCY9* potentially modulates the expression of *CETP* through a genetic effect mediated by rs1967309.

To test for potential interaction effects between rs1967309 and *CETP*, we used RNA-seq data from diverse projects in humans: the GEUVADIS project (33), the Genotype-Tissue Expression (GTEx v8) project (34) and CARTaGENE (CaG) (35). We evaluated the effects of the SNPs on expression levels of *ADCY9* and *CETP* by modelling both SNPs as continuous variables (additive model) (Methods). The *CETP* SNP rs158477 was reported as an expression quantitative trait locus (eQTL) in GTEx v7 and, in our models, shows evidence of being a *cis* eQTL of *CETP* in several other tissues (Supplementary text), although not reaching genome-wide significance. To test specifically for an epistatic effect between rs1967309 and rs158477 on *CETP* expression, we included an interaction term in eQTL models (Methods). We note here that we are testing for association for this specific pair of SNPs only, and that effects across tissues are not independent, such that we set our significance threshold at p-value=0.05. This analysis revealed a significant interaction effect (p-value=0.03, ß= -0.22) between the two SNPs on *CETP* expression in GEUVADIS lymphoblastoid cell lines (Fig. 4b, Supplementary Fig. 10a). In rs1967309 AA individuals, copies of the rs158477 A allele increased CETP expression by 0.46 (95% CI 0.26-0.86) on average. In rs1967309 AG individuals, copies of the rs158477 A allele increased CETP expression by 0.24

(95% CI 0.06-0.43) on average and the effect was null in rs1967309 GG individuals (p-value$_{GG}$=0.58). This suggests that the effect of rs158477 on *CETP* expression changes depending on genotypes of rs1967309. The interaction is also significant in several GTEx tissues, most of which are brain tissues, like hippocampus, hypothalamus and substantia nigra, but also in skin, although we note that the significance of the interaction depends on the number of PEER factors included in the model (Supplementary Fig. 11). These factors are needed to correct for unknown biases in the data, but also potentially lead to decreased power to detect interaction effects (36). In CaG whole blood samples, the interaction effect using additive genetic effect at rs1967309 was not significant, similarly to results from GTEx in whole blood samples. However, given the larger size of the dataset, we evaluated a genotypic encoding for the rs1967309 SNP in which the interaction effect is significant (p-value=0.008, Supplementary Fig. 10b) in whole blood, suggesting that rs1967309 could be modulating rs158477 eQTL effect, in this tissue at least, with a genotype-specific effect. We highlight that the sample sizes of current transcriptomic resources do not allow to detect interaction effects at genome-wide significance, however the likelihood of finding interaction effects between our two SNPs on *CETP* expression in three independent datasets is unlikely to happen by chance alone, providing evidence for a functional genetic interaction.

Epistatic effects on phenotypes

The interaction effect of rs1967309 and rs158477 on *CETP* expression in several tissues, found in multiple independent RNA-seq datasets, coupled with the detection of LRLD between these SNPs in the Peruvian population suggest that selection may act jointly on these loci, specifically in Peruvians or Andeans. These populations are well known for their adaptation to life in high

altitude, where the oxygen pressure is lower and where the human body is subjected to hypoxia (37–40). High altitude hypoxia impacts individuals' health in many ways, such as increased ventilation, decreased arterial pressure, and alterations of the energy metabolism in cardiac and skeletal muscle (41,42). To test which phenotype(s) may explain the putative coevolution signal discovered, we investigated the impact of the interaction between rs1967309 and rs158477 on several physiological traits, energy metabolism and cardiovascular outcomes using the UK Biobank and GTEx cohort (Supplementary Table 1). The UK Biobank has electronic medical records and GTEx has cause of death and variables from medical questionnaires (34). The interaction term was found to be nominally significant (p-value<0.05) for forced vital capacity (FVC), forced expiratory volume in 1-second (FEV1) and whole-body water mass, and suggestive (p-value<0.10) for the basal metabolic rate, all driven by the effects in females (Supplementary Fig. 12a). For CAD, the interaction is suggestive (p-value<0.10) and, in this case, driven by males (Supplementary Fig. 12a).

Given this sex-specific result on CAD, the condition targeted by dalcetrapib, we tested the effect of an interaction between sex, rs158477 and rs1967309 (genotypic encoding, see Methods) on binary cardiovascular outcomes including myocardial infarction (MI) and CAD. For CAD, we see a significant three-way interaction effect, meaning that for individuals carrying the AA genotype at rs1967309, the association between rs158477 and CAD is in the opposite direction in males and females. In other words, in rs1967309-AA females, having an extra A allele at rs158477, which is associated with higher CETP expression (Fig. 4b), has a protective effect on CAD. Conversely, in rs1967309-AA males, each A allele at rs158477 increases the probability of having an event (Fig. 5a). Little effect is seen in either sex for AG or GG at rs1967309, although the heterozygotes AG behave differently in females (which further justifies the genotypic encoding of rs1967309). The

beneficial effect of the interaction on CAD thus favors the rs1967309-AA + rs153477-GG males and the rs1967309-AA + rs153477-AA females, two genotype combinations which are respectively enriched in a sex-specific manner in the LIMAA cohort (Supplementary text). Again, observing such a result that concords with the direction of effects in the LRLD sex-specific finding is noteworthy. A significant interaction between the SNPs is also seen in the GTEx cohort (p-value=0.004, Supplementary Fig. 12b,c, Supplementary text), using questionnaire phenotypes reporting on MI, but the small number of individuals precludes formally investigating sex effects. Among the biomarkers studied (Supplementary Table 1), only lipoprotein(a) [Lp(a)] is suggestive in males (p-value=0.08) for an interaction between rs1967309 and rs158477, with the same direction of effect as that for CAD (Fig. 5a,b). Again, given the differences observed between the sexes, we tested the effect of an interaction between sex, rs158477 and rs1967309 (genotypic coding, Methods) on biomarkers, and only Lp(a) was nominally significant in a three-way interaction (p-value=0.049). The pattern is similar to the results for CAD, ie. a change in the effect of rs158477 depending on the genotype of rs1967309 in males, with the effect for AA females in the opposite direction compared to males (Fig. 5b). These concordant results between CAD and Lp(a) support that the putative interaction effect between the loci under study on phenotypes involves sex as a modifier.

Discussion

In this study, we used population genetics, transcriptomics and interaction analyses in biobanks to study the link between *ADCY9* and *CETP*. Our study revealed selective signatures in *ADCY9* and a significant genotypic association between *ADCY9* and *CETP* in two Peruvian cohorts,

specifically between rs1967309 and rs158477, which was also seen in the Native population of the Andes. The interaction between the two SNPs was found to be nominally significant for respiratory and cardiovascular disease outcomes (Fig. 5, Supplementary Fig. 12). Additionally, a nominally significant epistatic interaction was seen on *CETP* expression in many tissues, including the hippocampus and hypothalamus in the brain. Despite brain tissues not displaying the highest *CETP* expression levels, CETP that is synthesized and secreted in the brain could play an important role in the transport and the redistribution of lipids within the central nervous system (43,44) and has been associated with Alzheimer's disease risk (45,46). These findings reinforce the fact that the SNPs are likely functionally interacting, but extrapolating on the specific phenotypes under selection from these results is not straight forward. Identifying the phenotype and environmental pressures that may have caused the selection signal is complicated by the fact that the UK Biobank participants, on which the marginally significant associations have been found, do not live in the same environment as Peruvians. In Andeans from Peru, selection in response to hypoxia at high altitude was proposed to have effects on the cardiovascular system (21). The hippocampus functions are perturbed at high altitude (eg. deterioration of memory (47,48)), whereas the hypothalamus regulates the autonomic nervous system and controls the heart and respiratory rates (49,50), phenotypes which are affected by hypoxia at high altitude (51,52). Furthermore, high altitude-induced hypoxia (53,54) and cardiovascular system disturbances (55,56) have been shown to be associated in several studies (57–61), thus potentially sharing common biological pathways. Therefore, our working hypothesis is that selective pressures on our genes of interest in Peru are linked to the physiological response to high-altitude, which might be the environmental driver of coevolution.

The interaction effect between the *ADCY9* and *CETP* SNPs on both respiratory and cardiovascular phenotypes differs between the sexes, with effects on respiratory phenotypes limited to females (Supplementary Fig. 12a) and cardiovascular disease phenotype associations showing significant three-way sex-by-SNPs effects (Fig. 5). Furthermore, the LRLD signal is present mainly in males (Fig. 3), although the genotype association is also seen in female for a different genotype combination, suggesting the presence of sex-specific selection. This type of selection is very difficult to detect, especially on autosomes, with very few empirical examples found to date in the human genome despite strong theoretical support of their occurrence (62). However, sexual dimorphism in gene expression between males and females on autosomal genes has been linked to evolutionary pressures (63–65), possibly with a contribution of epistasis. As the source of selection, we favor the hypothesis of differential survival over differential reproduction, because the genetic combination between *ADCY9* and *CETP* has high chances to be broken up by recombination at each generation. Even in the case where recombination is suppressed in males between these loci, they would still have equal chances to pass the favored combination to both males and females offspring, which would not explain the sex-specific LRLD signal. We see an enrichment for the rs1967309-AA + rs158477-GG in males and rs1967309-AA + rs158477-AA in females, which are the beneficial combination for CAD in the corresponding sex, possibly pointing to a sexually antagonistic selection pressure, where the fittest genotype combination depends on the sex.

Such two-gene selection signature, where only males show strong LRLD, can happen if a specific genotype combination is beneficial in creating males (through differential sperm cell fitness or in utero survival, for example) or if survival until adulthood is favored with a specific genotype combination compared to other genotypes. To distinguish between the two scenarios, we tested if

the LRLD signal between rs1967309 and rs158477 in the LIMAA cohort depends on age in males but did not detect significant age effects (Supplementary text). Lastly, it may be that the evolutionary pressure is linked to the sex chromosomes (66,67), and a three-way interaction between *ADCY9*, *CETP* and the Y chromosome, for example, remains to be explored.

Despite the fact that we observed the LRLD signal between rs1967309 and rs158477 in two independent Peruvian cohorts, reducing the likelihood that our result is a false positive, one limitation is that the individuals were recruited in the same city (Lima) in both cohorts. However, we show that both populations are heterogeneous with respect to ancestry (Supplementary Fig. 2), suggesting that they likely represent accurately the Peruvian population. As recent admixture and population structure can strongly influence LRLD, we performed several analyses to consider these confounders, in the full cohorts and in the sex-stratified analyses. All analyses were robust to genome-wide and local ancestry patterns, such that our results are unlikely to be explained by these effects alone (extensive details are given in Supplementary text). Unfortunately, we did not have access to expression and phenotypic data from Peruvian individuals, which makes all the links between the selection pressures and the phenotype associations somewhat indirect. Future studies should focus on evaluating the phenotypic impact of the interaction specifically in Peruvians individuals, in cohorts such as the Population Architecture using Genomics and Epidemiology (PAGE) (68), in order to confirm the marginally significant associations found in European cohorts. Indeed, the Peruvian/Andean genomic background could be of importance for the interaction effect observed in this population, which reduces the power of discovery in individuals of unmatched ancestry. Another limitation is the low number of samples per tissue in GTEx and the cell composition heterogeneity per tissue and per sample (69), which can be partially

captured by PEER factors and can modulate the eQTL effects. Therefore, our power to detect tissue-specific interaction effects is reduced in this dataset, making it quite remarkable that we were able to observe multiple nominally significant interaction effects between the loci.

In this study, we discovered a putative epistatic interaction between the pharmacogene *ADCY9* and the drug target gene *CETP*, that appears to be under selection in the Peruvian population. Our approach exemplifies the potential of using evolutionary analyses to help find relationships between pharmacogenes and their drug targets. We characterized the impact of the *ADCY9/CETP* interaction on a range of phenotypes and tissues. Our gene expression results in brain tissues suggest that the interaction could play a role in protection against challenges to the nervous system caused by stress such as hypoxia. In light of the associations between high altitude-induced hypoxia and cardiovascular system changes, the interaction identified in this study could be involved in both systems: for example, ADCY9 and CETP could act in pathways involved in adaptation to high altitude, which could influence cardiovascular risk via their interaction, potentially in a sex-specific manner. Finally, our findings of an evolutionary relationship between *ADCY9* and *CETP* during recent human evolution points towards a biological link between dalcetrapib's pharmacogene *ADCY9* and its therapeutic target *CETP*.

Material and Methods

Population Genetics Datasets

The whole-genome sequencing data from the 1000 Genomes project (1000G) Phase III dataset (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) was filtered to exclude INDELs and

CNVs so that we kept only biallelic SNPs. This database has genomic variants of 2,504 individuals across five ancestral populations: Africans (AFR, *n* = 661), Europeans (EUR, *n* = 503), East Asians (EAS, *n* = 504), South Asians (SAS, *n* = 489) and Americans (AMR, *n* = 347) (70). The replication dataset, LIMAA, has been previously published (31,32) and was accessed through dbGaP [phs002025.v1.p1, dbgap project #26882]. We excluded related individuals as reported previously (31), resulting in a final dataset of 3,509 Peruvians, including 1,433 females and 2,076 males, for analysis. We also identified fine-scale population structure in this cohort and a more homogeneous subsample of 3,243 individuals (1,302 females and 1,941 males) in this cohort was kept for analysis (Supplementary text). The Native American genetic dataset (NAGD) contains 2,351 individuals from Native descendants from the data from a previously published study (23). Individuals were separated by their linguistic families identified by Reich and colleagues (23). NAGD came under the Hg18 coordinates, so a lift over was performed to transfer to the hg19 genome coordinates. Pre-processing details for these datasets are described in Supplementary text.

eQTL Datasets

We used several datasets for which we had both RNA-seq data and genotyping. First, the GEUVADIS dataset (33) for 1000G individuals was used (available at https://www.internationalgenome.org/data-portal/data-collection/geuvadis). A total of 285 non-duplicated European samples (CEU, GBR, FIN, TSI) were kept for analysis. Second, the Genotype-Tissue Expression v8 (GTEx) (34) was accessed through dbGaP (phs000424.v8.p2, dbgap project #19088) and contains gene expression across 54 tissues and 948 donors, genetic and phenotypic information. Phenotype analyses are described in Supplementary text. The cohort contains 67% of males and 33% of females, mainly of European descent (84.6%), aged between 20 and 79 years old. Analyses were done on 699 individuals. Third, we used the data from the

CARTaGENE biobank (35) (CAG project number 406713) which includes 790 RNA-seq whole-blood samples with genotype data, from individuals from Quebec (Canada) aged between 36 and 72 years old. Genotyping and RNA-seq data processing pipelines for these datasets are detailed in Supplementary text. To quantify *ADCY9* gene expression, we removed the isoform transcript ENST00000574721.1 (ADCY9-205 from the Hg38) from the Gene Transfer Format (GTF) file because it is a "retained intron" and accumulates genomic noise (Supplementary text), masking true signals for *ADCY9*. To take into account hidden factors, we calculated PEER factors (71) on the normalized expressions. To detect eQTL effects, we performed a two-sided linear regression on *ADCY9* and *CETP* expressions using R (v.3.6.0) (https://www.r-project.org/) with the formula $lm(p \sim rs1967309 * rs158477 + Covariates)$ for evaluating the interaction effect and $lm(p \sim rs1967309 + rs158477 + Covariates)$ for the main effect of the SNPs. Under the additive model, each SNP is coded by the number of non-reference alleles (G for rs1967309 and A for rs158477), under the genotypic model, dummy coding is used with homozygous reference genotype set as reference. The covariates include the first 5 Principal Components (PCs), age (except for GEUVADIS, information not available), sex, as well as PEER factors. We tested the robustness of our results to the inclusion of different numbers of PEER factors in the models and we report them all for GEUVADIS, CARTaGENE and GTEx (Supplementary Fig. 10, 11). Reported values in the text are for five PEER factors in GEUVADIS and ten PEER factors in CARTaGENE. Covariates specific to each cohort are reported in Supplementary text.

UK biobank processing and selected phenotypes

The UK biobank contains 487,392 genotyped individuals from the UK still enrolled as of August 20th 2020, imputed using the Haplotype Reference Consortium as the main reference panel, and

accessed through project #15357 and UKB project #20168. Additional genetic quality control was done using pyGenClean (v.1.8.3) (72). Variants or individuals with more than 2% missing genotypes were filtered out. Individuals with discrepancies between the self-reported and genetic sex or with aneuploidies were removed from the analysis. We considered only individuals of European ancestry based on principal components (PCs), as it is the largest population in the UK Biobank, and because ancestry can be a confounder of the genetic effect on phenotypes. We used the PCs from UK Biobank to define a region in PC space using individuals identified as "white British ancestry" as a reference population. Using the kinship estimates from the UK Biobank, we randomly removed individuals from kinship pairs where the coefficient was higher than 0.0884 (corresponding to a 3rd degree relationship). The resulting post QC dataset included 413,083 individuals. For the reported phenotypes, the date of baseline visit was between 2006 and 2010. The latest available hospitalization records discharge date was June 30th 2020 and the latest date in the death registries was February 14th 2018. We used algorithmically-defined cardiovascular outcomes based on combinations of operation procedure codes (OPCS) and hospitalization or death record codes (ICD9/ICD10). A description of the tested continuous variables can be found in Supplementary Table 1. We used age at recruitment defined in variable #21022 and sex in variable #31. We ignored self-reported events for cardiovascular outcomes as preliminary analyses suggested they were less precise than hospitalization and death records.

In association models, each SNP analyzed is coded by the number of non-reference alleles, G for rs1967309 and A for rs158477. SNP rs1967309 was also coded as a genotypic variable, to allow for non-additive effects. For continuous traits (Supplementary Table 1) in the UK Biobank, general two-sided linear models (GLM) were performed using SAS software (v.9.4). A GLM model was first performed using the covariates age, sex and PCs 1 to 10. The externally studentized residuals

were used to determine the outliers, which were removed. The normality assumption was confirmed by visual inspection of residuals for the majority of the outcomes, except *birthwt* and *sleep*. For biomarkers and cardiovascular endpoints, regression analyses were done in R (v.3.6.1). Linear regression analyses were conducted on standardized outcomes and logistic regression was used for cardiovascular outcomes. Marginal effects were calculated using margins package in R. In both cases, models were adjusted for age at baseline and top 10 PCs, as well as sex when not stratified. In models assessing two-way (rs1967309 by rs158477) or three-way (rs1967309 by rs158477 by sex) interactions, we used a 2 d.f. likelihood ratio test for the genotypic dummy variables' interaction terms (genotypic model) (Supplementary text).

RNA-sequencing of ADCY9-knocked-down HepG2 cell line

The human liver hepatocellular HepG2 cell line was obtained from ATCC. Cells were cultured in EMEM Minimum essential Medium Eagle's, supplemented with 10% fetal bovine serum (Wisent Inc). 250 000 cells in 2 ml of medium in a six-well plate were transfected using 12.5 pmol of Silencer Select siRNA against human ADCY9 (Ambion cat # 4390826 ID 1039) or Negative Control siRNA (Ambion cat #4390844) with 5 µl of Lipofectamine RNAiMAX reagent (Invitrogen cat #13778) in 500 µl Opti-MEM I reduced serum medium (Invitrogen cat # 31985) for 72h. The experiment was repeated five times at different cell culture passages. Total RNA was extracted from transfected HepG2 cells using RNeasy Plus Mini Kit (Qiagen cat #74136) in accordance with the manufacturer's recommendation. Preparation of sequencing library and sequencing was performed at the McGill University Innovation Center. Briefly, ribosomal RNA was depleted using NEBNext rRNA depletion kit. Sequencing was performed using Illumina NovaSeq 6000 S2 paired end 100 bp sequencing lanes. Basic QC analysis of the 10 samples was

performed by the Canadian Centre for Computational Genomics (C3G). To process the RNA-seq samples, we first performed read trimming and quality clipping using TrimGalore! (73) (https://github.com/FelixKrueger/TrimGalore), we aligned the trimmed reads on the GRCh38 reference genome using STAR (v.2.6.1a) and we ran RSEM (v.1.3.1) on the transcriptome aligned libraries. Prior to normalization with limma and voom, we filtered out genes which had less than 6 reads in more than 5 samples. For ADCY9 and CETP gene-level differential expression analyses, we compared the mean of each group of replicates with a t-test for paired samples. The transcriptome-wide differential expression analysis was done using limma, on all genes having an average of at least 10 reads across samples from a condition. Samples were paired in the experiment design. The multiple testing was taken into account by correcting the p-values with the qvalue R package (v.4.0.0) (74), to obtain transcriptome-wide FDR values.

Natural selection analyses

We used the integrated Haplotype Statistic (iHS) (22) and the population branch statistic (PBS) (70) to look for selective signatures. The iHS values were computed for the each 1000G population. An absolute value of iHS above 2 is considered to be a genome wide significant signal (22). Prior to iHS computation, ancestral alleles were retrieved from 6 primates using the EPO pipeline (version e59) (75) and the filtered 1000 Genomes vcf files were converted to change the reference allele as ancestral allele using bcftools (76) with the fixref plugin. The hapbin program (v.1.3.0) (77) was then used to compute iHS using per population-specific genetic maps computed by Adam Auton on the 1000G OMNI dataset (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rat

es). When the genetic map was not available for a subpopulation, the genetic map from the closest sub-population was selected according to their global FST value computed on the phase 3 dataset. We scanned the *ADCY9* and *CETP* genes using the population branch statistic (PBS), using 1000G sub-populations data. PBS summarizes a three-way comparison of allele frequencies between two closely related populations, and an outgroup. The grouping we focused on was PEL/MXL/CHB, with PEL being the focal population to test if allele frequencies are especially differentiated from those in the other populations. The CHB population was chosen as an outgroup to represent a Eurasian population that share common ancestors in the past with the American populations, after the out-of-Africa event. Using PJL (South Asia) and CEU (Europe) as an outgroup, or CLM as a closely related population (instead of MXL) yield highly similar results. To calculate F_{ST} for each pair of population in our tree, we used vcftools (78) by subpopulation. We calculated normalized PBS values as in (21), which adjusts values for positions with large branches in all populations, for the whole genome. We use this distribution to define an empirical threshold for significance based on the 95th percentile of all PBS values genome-wide for each of the three populations.

Long-range linkage disequilibrium

Long-range linkage disequilibrium (LRLD) was calculated using the function geno-r2 of vcftools (v.0.1.17) which uses the genotype frequencies. LRLD was evaluated in all subpopulations from 1000 Genomes Project Phase III, in LIMAA and NAGD, for all biallelic SNPs in *ADCY9* (chr16:4,012,650-4,166,186 in Hg19 genome reference) and *CETP* (chr16:56,995,835-57,017,756 in Hg19 genome reference). We analyzed loci from the phased VCF files that had a MAF of at least 5% and a missing genotype of at most 10%, in order to retain a maximum of SNPs in NAGD which has higher missing rates than the others. We extracted the 99th percentile of all pairs of

comparisons between *ADCY9* and *CETP* genes to use as a threshold for empirical significance and we refer to these as *ADCY9/CETP* empirical p-values (Supplementary text). In LIMAA, we also evaluated the genotypic association using a χ^2 test with four degrees of freedom (χ^2_4) using a permutation test, as reported in (25) (Supplementary text).

Furthermore, for both cohorts, we created a distribution of LRLD values for random pairs of SNPs across the genome to obtain a genome-wide null distribution of LRLD to evaluate how unusual the genotypic association between our candidate SNPs (rs1967309-rs158477) is while taking into account the cohort-specific background genomic noise/admixture and allele frequencies. We extracted 3,513 pairs of SNPs that match rs1967309 and rs158477 in terms of MAF, physical distance (in base pairs) and genetic distance (in centiMorgans (cM), based on the PEL genetic map) between them in both cohorts (Supplementary text), and report genome-wide empirical p-values based on this distribution. For the analyses of LRLD between *ADCY9* and *CETP* stratified by sex, we considered the same set of SNP pairs that we used for the full cohorts, but separated the dataset by sex before calculating the LRLD values. To evaluate how likely the differences observed in LRLD between sex are, we also performed permutations of the sex labels across individuals to create a null distribution of sex specific effects (Supplementary text).

Local ancestry inference

To evaluate local ancestry in the PEL subpopulation and in the LIMAA cohort, we constructed a reference panel using the phased haplotypes from 1000 Genomes (YRI, CEU, CHB) and the phased haplotypes of NAGD (Northern American, Central American and Andean) (Supplementary text). We kept overlapping positions between all datasets, and when SNPs had the exact same genetic position, we kept the SNP with the highest variance in allele frequencies across

563 all reference populations (Supplementary text). We ran RFMix (v.2.03) (79) (with the option
564 'reanalyze-reference' and for 25 iterations) on all phased chromosomes. We estimated the whole
565 genome average proportion of each ancestry using a weighted mean of the chromosome specific
566 proportions given by RFMix based on the chromosome size in cM. For comparing the overall
567 Andean enrichment inferred by RFMix between rs1967309/rs158477 genotype categories, we
568 used a two-sided Wilcoxon-t-test. To evaluate the Andean local ancestry enrichment specifically
569 at *ADCY9* and *CETP*, we computed the genome-wide 95th percentile for proportion of Andean
570 attribution for all intervals given by RFMix.

571 **Supplemental Information description**

572 Supplementary figures 1-14

573 Supplementary tables 1-2

574 Supplementary methods and texts

575

576 **Acknowledgements**

577 We thank all members of the Hussin lab for their constructive comments and feedback throughout
578 this project. This work was completed thanks to computational resources provided by Compute
579 Canada clusters Graham and Beluga. This work was funded by the *Institut de Valorisation des*
580 *Données* (IVADO) and the Montreal Heart Institute (MHI) Foundation. J.G.H. is a *Fonds de la*
581 *Recherche en Santé* (FRQS) Junior 1 fellow. I.G. receives a PhD scholarship from the MHI
582 Foundation and M.A.L. holds a PhD scholarship from Canadian Institutes of Health Research.
583 M.P.D. holds the Canada Research Chair in Precision Medicine Data Analysis. J.C.T. holds the
584 Canada Research Chair in Personalized Medicine and the Université de Montréal endowed
585 research chair in atherosclerosis.

586

587 **Author Contributions**

588 Conceptualization: I.G., M.P.D and J.G.H.; Data curation: I.G., M.A.L., J.C.G.; Statistical and
589 bioinformatic analyses: I.G., M.A.L., J.C.G., H.T, S.A, A.B. and Y.F.Z.; Data acquisition: J.C.G.,
590 J.G.H, Y.L., L.L., M.M. and S.R.; Wet lab experimentation: R.S. and E.R.; Writing – Original
591 draft: I.G. and J.G.H.; Writing – Review & editing: I.G., M.A.L., J.C.G., R.S., E.R., S.A., H.T.,
592 Y.L., S.R., J.C.T., M.P.D. and J.G.H.; Supervision: J.C.T., M.P.D. and J.G.H.; Funding
593 acquisition: J.C.T., M.P.D., J.G.H.

Competing Interests statement

J.G.H. has received speaker honoraria from Dalcor and District 3 Innovation Centre. J.C.T. reports grants from Government of Quebec, National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health (NIH), the MHI Foundation, from Bill and Melinda Gates Foundation, Amarin, Esperion, Ionis, Servier, RegenXBio; personal fees from Astra Zeneca, Sanofi, Servier; and personal fees and minor equity interest from Dalcor. M.P.D. and J.C.T. have a patent Methods for Treating or Preventing Cardiovascular Disorders and Lowering Risk of Cardiovascular Events issued to Dalcor, no royalties received, a patent Genetic Markers for Predicting Responsiveness to Therapy with HDL-Raising or HDL Mimicking Agent issued to Dalcor, no royalties received, and a patent Methods for using low dose colchicine after myocardial infarction with royalties paid to Invention assigned to the Montreal Heart Institute. M.P.D. reports personal fees and other from Dalcor and personal fees from GlaxoSmithKline, other from AstraZeneca, Pfizer, Servier, Sanofi. The remaining authors have nothing to disclose.

References

1. Lagrost L. Regulation of cholesteryl ester transfer protein (CETP) activity: review of in vitro and in vivo studies. Biochim Biophys Acta. 1994 Dec 8;1215(3):209–36.
2. Shinkai H. Cholesteryl ester transfer-protein modulator and inhibitors and their potential for the treatment of cardiovascular diseases. Vasc Health Risk Manag. 2012;8:323–31.
3. Schwartz GG, Olsson AG, Abt M, Ballantyne CM, Barter PJ, Brumm J, et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. N Engl J Med. 2012 Nov 29;367(22):2089–99.
4. Tardif J-C, Rhéaume E, Lemieux Perreault L-P, Grégoire JC, Feroz Zada Y, Asselin G, et al. Pharmacogenomic determinants of the cardiovascular effects of dalcetrapib. Circ Cardiovasc Genet. 2015 Apr;8(2):372–82.
5. Tardif J-C, Rhainds D, Brodeur M, Feroz Zada Y, Fouodjio R, Provost S, et al. Genotype-dependent effects of dalcetrapib on cholesterol efflux and inflammation. Circ Cardiovasc Genet. 2016 Aug;9(4):340–8.
6. Rautureau Y, Deschambault V, Higgins M-È, Rivas D, Mecteau M, Geoffroy P, et al. ADCY9 (Adenylate Cyclase Type 9) inactivation protects from atherosclerosis only in the absence of CETP (Cholesteryl Ester Transfer Protein). Circulation. 2018 Oct 16;138(16):1677–92.
7. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004 Jun;74(6):1111–20.
8. Enattah NS, Trudeau A, Pimenoff V, Maiuri L, Auricchio S, Greco L, et al. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. Am J Hum Genet. 2007 Sep;81(3):615–25.
9. Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, et al. Genome flux and stasis in a five millennium transect of European prehistory. Nature Communications. 2014 Oct 21;5(1):5257.
10. Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG. The origins of lactase persistence in Europe. PLOS Computational Biology. 2009 août;5(8):e1000491.
11. Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, et al. The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet. 2003 Jul;67(Pt 4):298–311.
12. Labrie V, Buske OJ, Oh E, Jeremian R, Ptak C, Gasiūnas G, et al. Lactase non-persistence is directed by DNA variation-dependent epigenetic aging. Nat Struct Mol Biol. 2016 Jun;23(6):566–73.
13. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of fifty human exomes reveals adaptation to high altitude. Science. 2010 Jul 2;329(5987):75–8.
14. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. 2015 Sep 18;349(6254):1343.
15. Hollenbach JA, Thomson G, Cao K, Fernandez-Vina M, Erlich HA, Bugawan TL, et al. HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. Human Immunology. 2001 Apr 1;62(4):378–90.
16. Li P, Zhao J, Kothapalli KSD, Li X, Li H, Han Y, et al. A regulatory insertion-deletion polymorphism in the FADS gene cluster influences PUFA and lipid profiles among Chinese

652 adults: a population-based study. The American Journal of Clinical Nutrition. 2018 Jun
653 1;107(6):867–75.
654 17. Reynolds LM, Dutta R, Seeds MC, Lake KN, Hallmark B, Mathias RA, et al. FADS
655 genetic and metabolomic analyses identify the Δ5 desaturase (FADS1) step as a critical control
656 point in the formation of biologically important lipids. Scientific Reports. 2020 Sep
657 28;10(1):15873.
658 18. Tashi T, Scott Reading N, Wuren T, Zhang X, Moore LG, Hu H, et al. Gain-of-function
659 EGLN1 prolyl hydroxylase (PHD2 D4E:C127S) in combination with EPAS1 (HIF-2α)
660 polymorphism lowers hemoglobin concentration in Tibetan highlanders. J Mol Med. 2017 Jun
661 1;95(6):665–70.
662 19. Meyer D, C. Aguiar VR, Bitarello BD, C. Brandt DY, Nunes K. A genomic perspective
663 on HLA evolution. Immunogenetics. 2018 Jan 1;70(1):5–27.
664 20. Blais M-E, Zhang Y, Rostron T, Griffin H, Taylor S, Xu KY, et al. High frequency of
665 HIV mutations associated with HLA-C suggests enhanced HLA-C-restricted CTL selective
666 pressure associated with an AIDS-protective polymorphism. J Immunol. 2012 May
667 1;188(9):4663–70.
668 21. Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, et al. Natural selection
669 on genes related to cardiovascular health in high-altitude adapted Andeans. Am J Hum Genet.
670 2017 Nov 2;101(5):752–67.
671 22. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the
672 human genome. PLOS Biology. 2006 Mar 7;4(3):e72.
673 23. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing
674 Native American population history. Nature. 2012 Aug;488(7411):370–4.
675 24. Lewontin RC, Kojima K. The evolutionary dynamics of complex polymorphisms.
676 Evolution. 1960;14(4):458–72.
677 25. Rohlfs RV, Swanson WJ, Weir BS. Detecting coevolution through allelic association
678 between physically unlinked loci. The American Journal of Human Genetics. 2010 May
679 14;86(5):674–85.
680 26. Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, et al. Evolutionary
681 genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc Natl Acad Sci U
682 S A. 2018 Jul 10;115(28):E6526–35.
683 27. Li W-H, Nei M. Stable linkage disequilibrium without epistasis in subdivided
684 populations. Theoretical Population Biology. 1974 Oct 1;6(2):173–83.
685 28. Nei M, Li WH. Linkage disequilibrium in subdivided populations. Genetics. 1973
686 Sep;75(1):213–9.
687 29. Park L. Population-specific long-range linkage disequilibrium in the human genome and
688 its influence on identifying common disease variants. Scientific Reports. 2019 Aug 6;9(1):11380.
689 30. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping
690 the medical future. Nat Rev Genet. 2008 Jun;9(6):477–85.
691 31. Asgari S, Luo Y, Akbari A, Belbin GM, Li X, Harris DN, et al. A positively selected
692 FBN1 missense variant reduces height in Peruvian individuals. Nature. 2020 Jun;582(7811):234–
693 9.
694 32. Luo Y, Suliman S, Asgari S, Amariuta T, Baglaenko Y, Martínez-Bonet M, et al. Early
695 progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. Nat
696 Commun. 2019 Aug 21;10(1):3765.
697 33. Lappalainen T, Sammeth M, Friedländer MR, ‘t Hoen PAC, Monlong J, Rivas MA, et al.

698 Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013
699 Sep;501(7468):506–11.
700 34. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.
701 35. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet J-P, Knoppers B, et al. Cohort
702 profile of the CARTaGENE study: Quebec's population-based biobank for public health and
703 personalized genomics. Int J Epidemiol. 2013 Oct;42(5):1285–99.
704 36. Brynedal B, Choi J, Raj T, Bjornson R, Stranger BE, Neale BM, et al. Large-scale trans-
705 eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. Am J
706 Hum Genet. 2017 Apr 6;100(4):581–91.
707 37. Beall CM. Two routes to functional adaptation: Tibetan and Andean high-altitude
708 natives. Proc Natl Acad Sci U S A. 2007 May 15;104(Suppl 1):8655–60.
709 38. Brutsaert TD, Parra EJ, Shriver MD, Gamboa A, Rivera-Ch M, León-Velarde F.
710 Ancestry explains the blunted ventilatory response to sustained hypoxia and lower exercise
711 ventilation of Quechua altitude natives. Am J Physiol Regul Integr Comp Physiol. 2005
712 Jul;289(1):R225-234.
713 39. Julian CG, Moore LG. Human genetic adaptation to high altitude: evidence from the
714 Andes. Genes (Basel). 2019 février;10(2):150.
715 40. Moore LG. Human genetic adaptation to high altitudes: current status and future
716 prospects. Quat Int. 2017 Dec 15;461:4–13.
717 41. Milledge JS, West JB, Schoene RB. High altitude medicine and physiology [Internet]. 4th
718 ed. CRC Press; 2007. Available from: https://books.google.ca/books?id=lj59BgAAQBAJ
719 42. Murray AJ. Energy metabolism and the high-altitude environment. Experimental
720 Physiology. 2016;101(1):23–7.
721 43. Albers JJ, Tollefson JH, Wolfbauer G, Albright RE. Cholesteryl ester transfer protein in
722 human brain. Int J Clin Lab Res. 1992;21(3):264–6.
723 44. Yamada T, Kawata M, Arai H, Fukasawa M, Inoue K, Sato T. Astroglial localization of
724 cholesteryl ester transfer protein in normal and Alzheimer's disease brain tissues. Acta
725 Neuropathol. 1995 Dec 1;90(6):633–6.
726 45. Murphy EA, Roddey JC, McEvoy LK, Holland D, Hagler DJ, Dale AM, et al. CETP
727 polymorphisms associate with brain structure, atrophy rate, and Alzheimer's disease risk in an
728 APOE-dependent manner. Brain Imaging Behav. 2012 Mar;6(1):16–26.
729 46. Oestereich F, Yousefpour N, Yang E, Ribeiro-da-Silva A, Chaurand P, Munter LM. The
730 cholesteryl ester transfer protein (CETP) raises cholesterol levels in the brain and affects
731 presenilin-mediated gene regulation. bioRxiv. 2020 Nov 24;2020.11.24.395186.
732 47. Lieberman P, Morey A, Hochstadt J, Larson M, Mather S. Mount Everest: a space
733 analogue for speech monitoring of cognitive deficits and stress. Aviat Space Environ Med. 2005
734 Jun;76(6 Suppl):B198-207.
735 48. Shukitt-Hale B, Stillman MJ, Welch DI, Levy A, Devine JA, Lieberman HR. Hypobaric
736 hypoxia impairs spatial memory in an elevation-dependent fashion. Behavioral and Neural
737 Biology. 1994 Nov 1;62(3):244–52.
738 49. Horiuchi J, McDowall LM, Dampney RAL. Vasomotor and respiratory responses evoked
739 from the dorsolateral periaqueductal grey are mediated by the dorsomedial hypothalamus. The
740 Journal of Physiology. 2009;587(21):5149–62.
741 50. Rahmouni K. Cardiovascular regulation by the arcuate nucleus of the hypothalamus:
742 neurocircuitry and signaling systems. Hypertension. 2016 Jun;67(6):1064–71.
743 51. Bärtsch Peter, Gibbs J. Simon R. Effect of altitude on the heart and the lungs. Circulation.

2007 Nov 6;116(19):2191–202.
52. Hainsworth R, Drinkhill MJ, Rivera-Chira M. The autonomic nervous system at high altitude. Clin Auton Res. 2007 Feb;17(1):13–9.
53. Bigham AW, Lee FS. Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes Dev. 2014 Oct 15;28(20):2189–204.
54. Moore LG. Measuring high-altitude adaptation. J Appl Physiol (1985). 2017 Nov 1;123(5):1371–85.
55. Abe H, Semba H, Takeda N. The roles of hypoxia signaling in the pathogenesis of cardiovascular diseases. J Atheroscler Thromb. 2017 Sep 1;24(9):884–94.
56. Lee JW, Ko J, Ju C, Eltzschig HK. Hypoxia signaling in human diseases and therapeutic targets. Exp Mol Med. 2019 Jun 20;51(6):1–13.
57. Faeh D, Gutzwiller F, Bopp M, Swiss National Cohort Study Group. Lower mortality from coronary heart disease and stroke at higher altitudes in Switzerland. Circulation. 2009 Aug 11;120(6):495–501.
58. Naeije R. Physiological adaptation of the cardiovascular system to high altitude. Progress in Cardiovascular Diseases. 2010 May 1;52(6):456–66.
59. Ostadal B, Kolar F. Cardiac adaptation to chronic high-altitude hypoxia: Beneficial and adverse effects. Respiratory Physiology & Neurobiology. 2007 Sep 30;158(2):224–36.
60. Riley CJ, Gavin M. Physiological changes to the cardiovascular system at high altitude and its effects on cardiovascular disease. High Altitude Medicine & Biology. 2017 Mar 15;18(2):102–13.
61. Savla JJ, Levine BD, Sadek HA. The effect of hypoxia on cardiovascular disease: friend or foe? High Alt Med Biol. 2018 Jun;19(2):124–30.
62. Morrow EH, Connallon T. Implications of sex-specific selection for the genetic basis of disease. Evol Appl. 2013 Dec;6(8):1208–17.
63. Connallon T, Clark AG. Sex linkage, sex-specific selection, and the role of recombination in the evolution of sexually dimorphic gene expression. Evolution. 2010 Dec;64(12):3417–42.
64. Parsch J, Ellegren H. The evolutionary causes and consequences of sex-biased gene expression. Nature Reviews Genetics. 2013 Feb;14(2):83–7.
65. Williams TM, Carroll SB. Genetic and molecular insights into the development and evolution of sexual dimorphism. Nature Reviews Genetics. 2009 Nov;10(11):797–804.
66. Cox RM, Cox CL, McGlothlin JW, Card DC, Andrew AL, Castoe TA. Hormonally mediated increases in sex-biased gene expression accompany the breakdown of between-sex genetic correlations in a sexually dimorphic lizard. The American Naturalist. 2017 Jan 25;189(3):315–32.
67. McGlothlin JW, Cox RM, Brodie ED III. Sex-Specific Selection and the Evolution of Between-Sex Genetic Covariance. Journal of Heredity. 2019 Jul 1;110(4):422–32.
68. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019 Jun;570(7762):514–8.
69. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204–13.
70. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.
71. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of

790 expression residuals (PEER) to obtain increased power and interpretability of gene expression
791 analyses. Nat Protoc. 2012 Feb 16;7(3):500–7.
792 72. Lemieux Perreault L-P, Provost S, Legault M-A, Barhdadi A, Dubé M-P. pyGenClean:
793 efficient tool for genetic data clean up before association testing. Bioinformatics. 2013 Jul
794 1;29(13):1704–5.
795 73. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
796 EMBnet.journal. 2011 May 2;17(1):10–2.
797 74. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical
798 Society: Series B (Statistical Methodology). 2002;64(3):479–98.
799 75. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl
800 comparative genomics resources. Database (Oxford) [Internet]. 2016 Feb 20 [cited 2021 Feb
801 23];2016. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4761110/
802 76. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence
803 alignment/map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
804 77. Maclean CA, Chue Hong NP, Prendergast JGD. hapbin: an efficient program for
805 performing haplotype-based scans for positive selection in large genomic datasets. Molecular
806 Biology and Evolution. 2015 Nov 1;32(11):3027–9.
807 78. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
808 call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.
809 79. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling
810 approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013 Aug
811 8;93(2):278–88.
812

Main figures

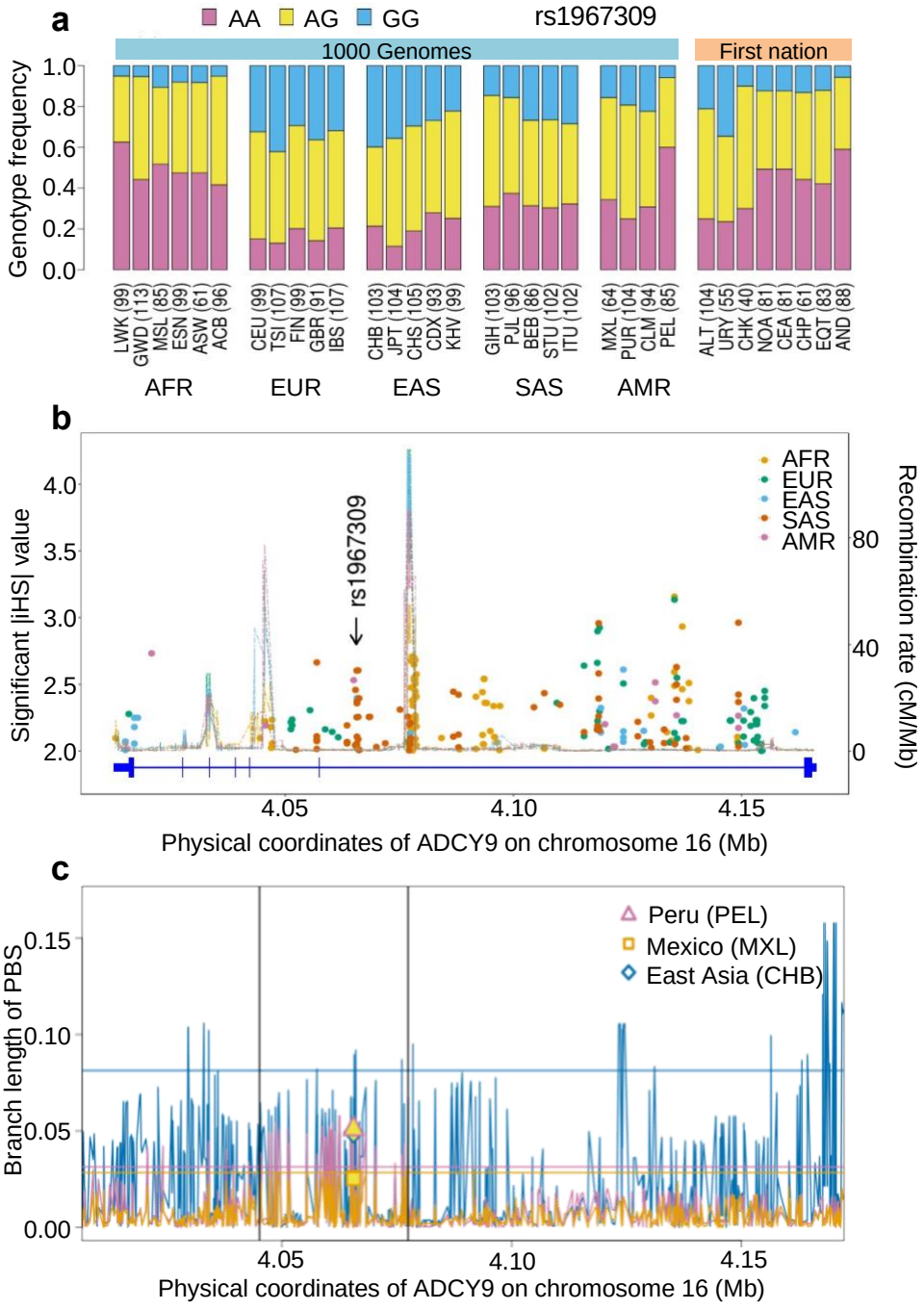

Fig. 1. Natural selection signature at rs1967309 in ADCY9. (A) Genotype frequency distribution of rs1967309 in populations from the 1000 Genomes (1000G) Project and in Native Americans. (B) Significant iHS values (absolute values above 2) for 1000G continental

818 populations and recombination rates from 1000G population-specific genetic maps, in the *ADCY9*
819 gene. (C) PBS values in the *ADCY9*, comparing the CHB (outgroup), MXL and PEL. Horizontal
820 lines represent the 95th percentile PBS value genome-wide for each population. Vertical black lines
821 represent the highest recombination rates around rs1967309 from 1000G population-specific
822 genetic maps. Abbreviations: Altaic from Mongolia and Russia: ALT; Uralic Yukaghir from
823 Russia: URY; Chukchi Kamchatkan from Russia: CHK; Northern American from Canada,
824 Guatemala and Mexico: NOA; Central American from Costal Rica and Mexico: CEA; Chibchan
825 Paezan from Argentina, Bolivia, Colombia, Costa Rica and Mexico: CHP; Equatorial Tucanoan
826 from Argentina, Brazil, Colombia, Gualana and Paraguay: EQT; Andean from Bolivia, Chile,
827 Colombia and Peru: AND. For 1000G populations, abbreviations can be found here
828 https://www.internationalgenome.org/category/population/.

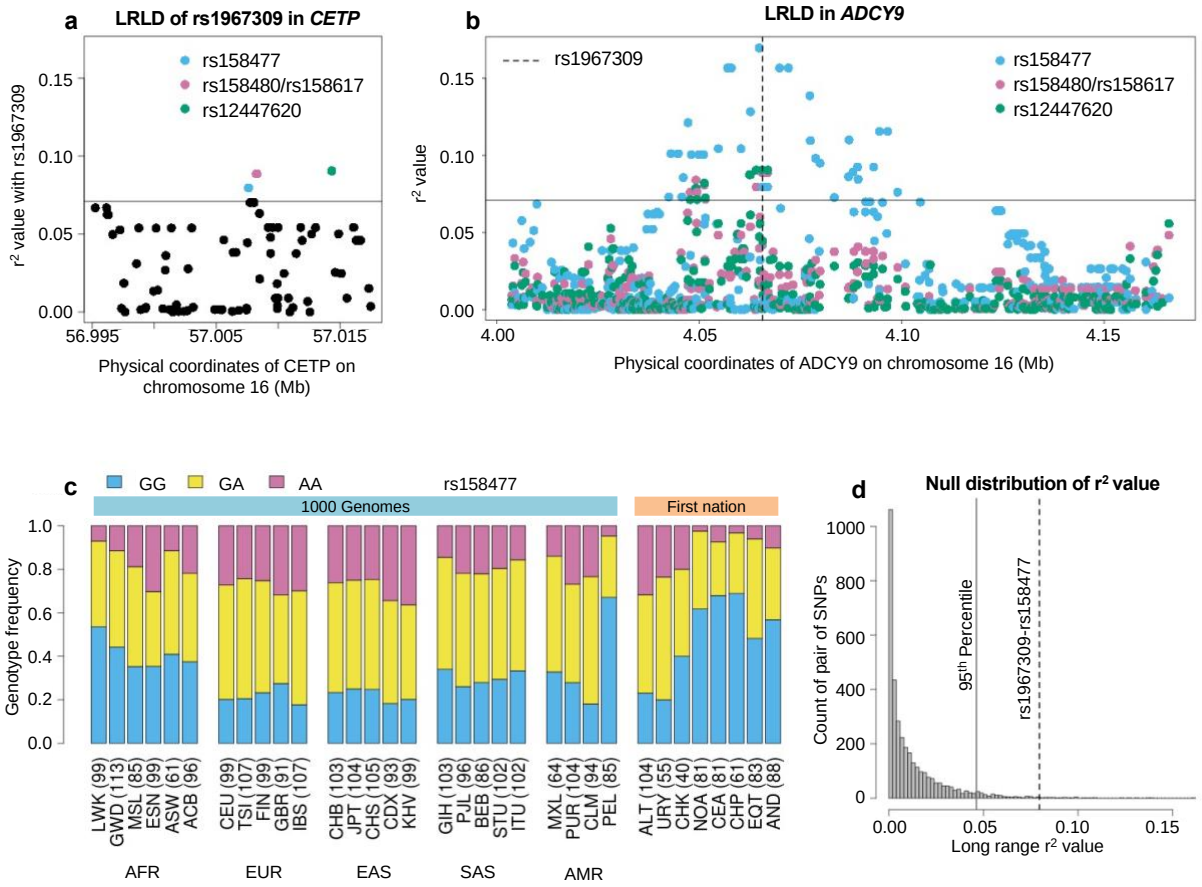

Fig. 2. Long-range linkage disequilibrium between rs1967309 and rs158477 in Peruvians from Lima, Peru. (A) Genotype correlation (r^2) between rs1967309 and all SNPs with MAF>5% in *CETP*, for the PEL population. (B) Genotype correlation between the 3 loci identified in (a) to be in the 99th percentile and all SNPs with MAF>5% in *ADCY9*. The dotted line indicates the position of rs1967309. The horizontal lines in (a,b) represents the threshold for the 99th percentile of all comparisons of SNPs (MAF>5%) between *ADCY9* and *CETP*. (C) Genotype frequency distribution of rs158477 in 1000G and Native American populations. (D) Genomic distribution of r^2 values from 3,513 pairs of SNPs separated by between 50-60 Mb and 61±10 cM away across all Peruvian chromosomes from the PEL sample, compared to the rs1967309-rs158477 r^2 value (dotted grey line) (genome-wide empirical p-value=0.01). The vertical black line shows the

840 threshold for the 95th percentile threshold of all pairs. Abbreviations: Altaic from Mongolia and
841 Russia: ALT; Uralic Yukaghir from Russia: URY; Chukchi Kamchatkan from Russia: CHK;
842 Northern American from Canada, Guatemala and Mexico: NOA; Central American from Costal
843 Rica and Mexico: CEA; Chibchan Paezan from Argentina, Bolivia, Colombia, Costa Rica and
844 Mexico: CHP; Equatorial Tucanoan from Argentina, Brazil, Colombia, Gualana and Paraguay:
845 EQT; Andean from Bolivia, Chile, Colombia and Peru: AND. For 1000G populations,
846 abbreviations can be found here https://www.internationalgenome.org/category/population/
847

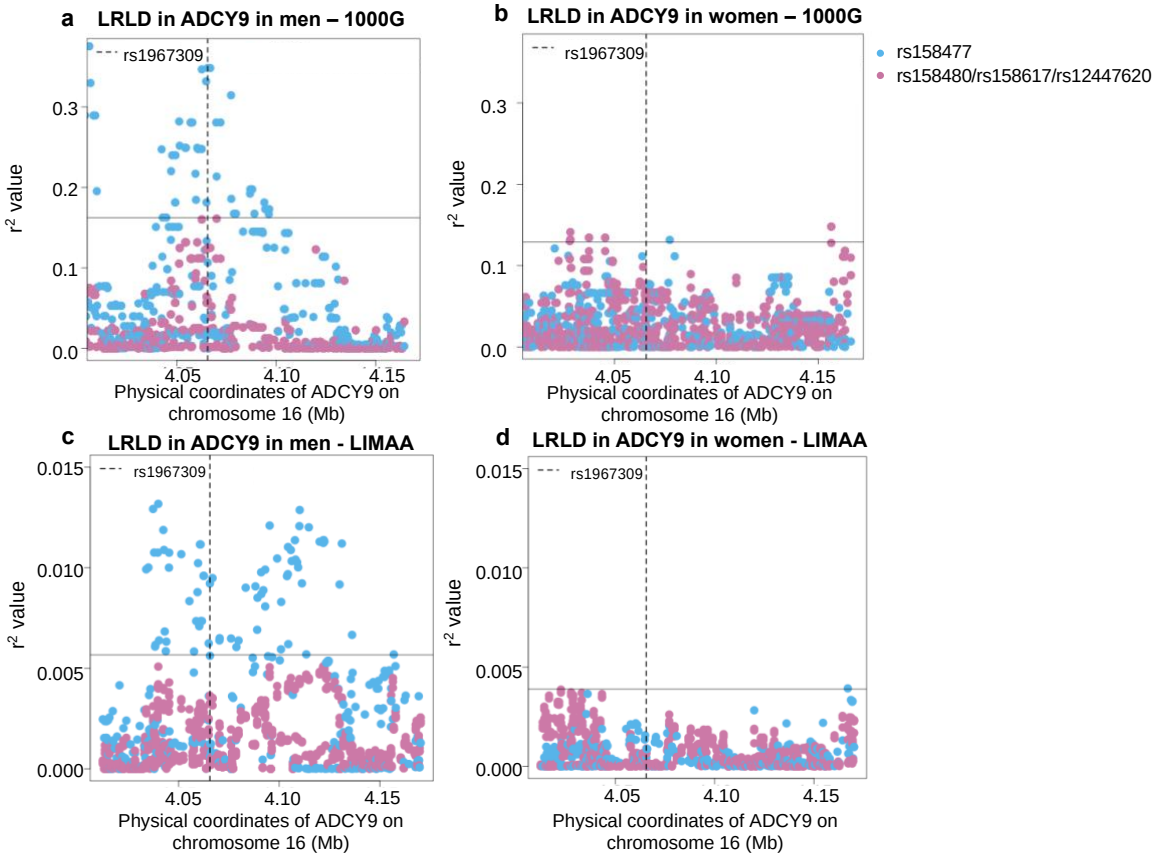

Fig. 3. Sex-specific long-range linkage disequilibrium. Genotype correlation between the loci identified in CETP in Figure 2A and all SNPs with MAF>5% in *ADCY9* for the PEL population (A,B) and LIMAA cohort (C,D) in males (A,C) and in females (B,D). SNPs rs158480, rs158617 and rs12447620, are in near-perfect LD in these subsets. The horizontal line shows the threshold for the 99th percentile of all comparisons of SNPs (MAF>5%) between *ADCY9* and *CETP*. The dotted line represents the position of rs1967309. Blue dots represent the rs158477 SNPs and pink represents the other three SNPs identified in Figure 2A.

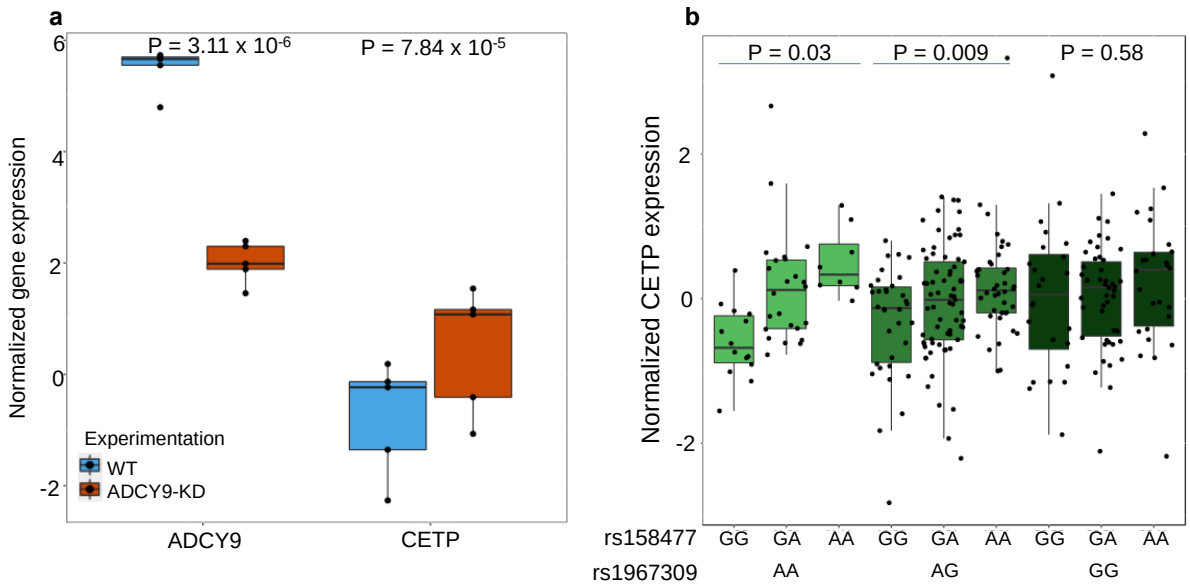

Fig. 4. Effect of *ADCY9* on *CETP* expression. (A) Normalized expression of *ADCY9* or *CETP* genes depending on wild type (WT) and *ADCY9-KD* in HepG2 cells from RNA sequencing on five biological replicates in each group. P-values were obtained from a two-sided Wilcoxon paired test. (B) *CETP* expression depending on the combination of rs1967309 and rs158477 genotypes in GEUVADIS for 285 individuals of European descent according to principal component analysis. P-values reported were obtained from a two-sided linear regression model after stratification by rs1967309 genotypes.

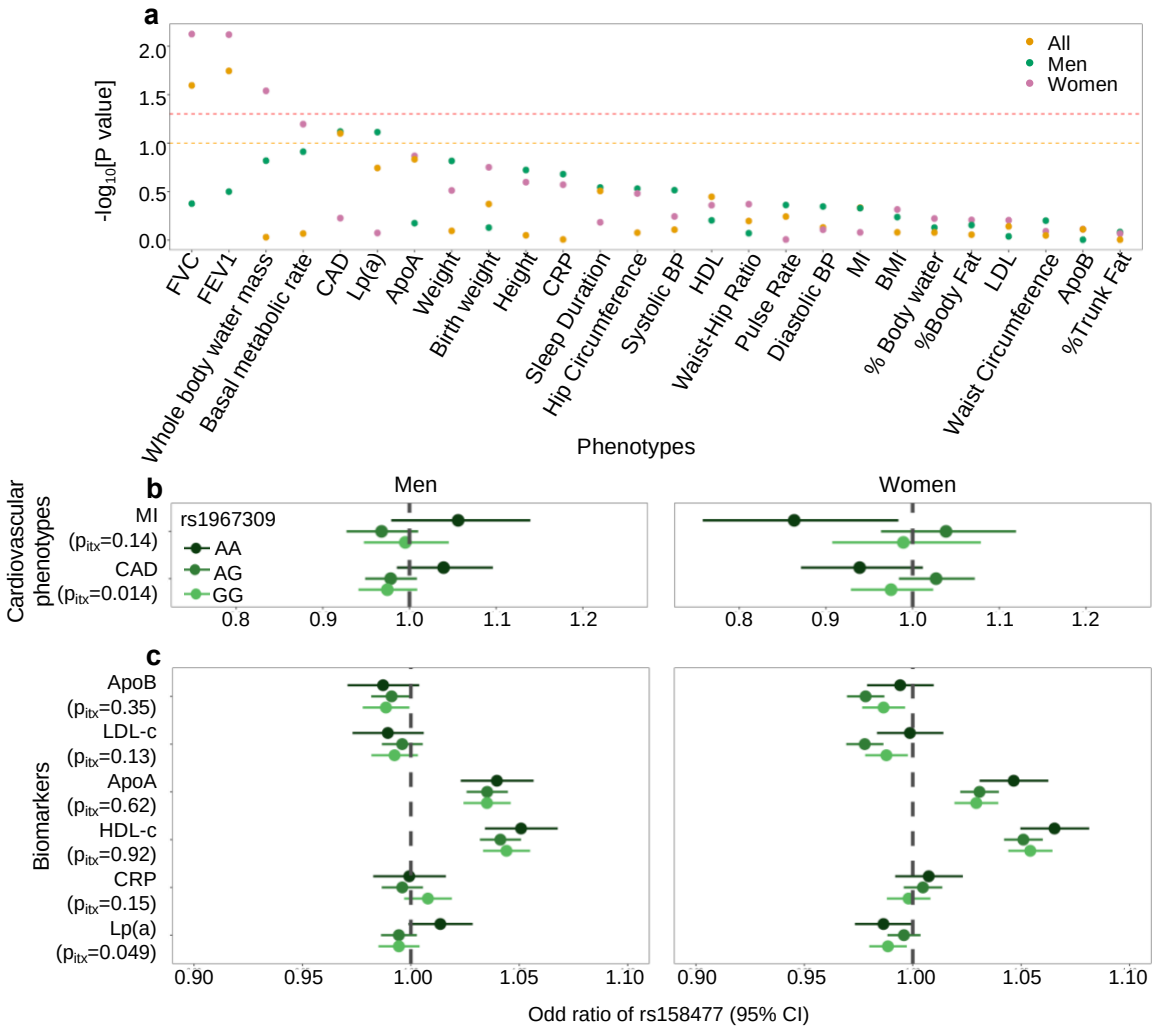

Fig. 5. Epistatic association of rs1967309 and rs158477 on phenotypes in the UK biobank. Effect of the rs158477 SNP on cardiovascular phenotypes (A) and biomarkers (B) depending on the genotype of rs1967309 (genotypic encoding) in males and females. The p-values p_{itx} reported are from a likelihood ratio test comparing models with and without the three-way interaction term between the SNPs and sex. See Supplementary Table 1 for the list of abbreviations.