PLOS | GENETICS

RESEARCH ARTICLE

Genetic loci associated with coronary artery disease harbor evidence of selection and antagonistic pleiotropy

Sean G. Byars[1,2]*, Qin Qin Huang[1,2,3], Lesley-Ann Gray[1,2], Andrew Bakshi[1], Samuli Ripatti[4,5,6], Gad Abraham[1,2,3], Stephen C. Stearns[7], Michael Inouye[1,2,3]*

1 Centre for Systems Genomics, School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia, **2** Department of Pathology, The University of Melbourne, Parkville, Victoria, Australia, **3** Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia, **4** Institute of Molecular Medicine Finland, University of Helsinki, Helsinki, Finland, **5** Department of Public Health, University of Helsinki, Helsinki, Finland, **6** Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, **7** Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, United States of America

* sean.byars@unimelb.edu.au (SGB); minouye@baker.edu.au (MI)

Citation: Byars SG, Huang QQ, Gray L-A, Bakshi A, Ripatti S, Abraham G, et al. (2017) Genetic loci associated with coronary artery disease harbor evidence of selection and antagonistic pleiotropy. PLoS Genet 13(6): e1006328. https://doi.org/10.1371/journal.pgen.1006328

Editor: Sarah Pendergrass, Geisinger Health System, UNITED STATES

Received: August 31, 2016

Accepted: May 2, 2017

Published: June 22, 2017

Data Availability Statement: The HapMap3 genotype and gene expression data are respectively available at hapmap.org and http://www.ebi.ac.uk/arrayexpress/. The 1000 genomes data is available at http://www.1000genomes.org/data. The coronary artery disease genome-wide risk meta-analysis data is available at http://www.cardiogramplusc4d.org. The Framingham Heart Study phenotype and genotype data is available at https://www.ncbi.nlm.nih.gov/gap.

Abstract

Traditional genome-wide scans for positive selection have mainly uncovered selective sweeps associated with monogenic traits. While selection on quantitative traits is much more common, very few signals have been detected because of their polygenic nature. We searched for positive selection signals underlying coronary artery disease (CAD) in worldwide populations, using novel approaches to quantify relationships between polygenic selection signals and CAD genetic risk. We identified new candidate adaptive loci that appear to have been directly modified by disease pressures given their significant associations with CAD genetic risk. These candidates were all uniquely and consistently associated with many different male and female reproductive traits suggesting selection may have also targeted these because of their direct effects on fitness. We found that CAD loci are significantly enriched for lifetime reproductive success relative to the rest of the human genome, with evidence that the relationship between CAD and lifetime reproductive success is antagonistic. This supports the presence of antagonistic-pleiotropic tradeoffs on CAD loci and provides a novel explanation for the maintenance and high prevalence of CAD in modern humans. Lastly, we found that positive selection more often targeted CAD gene regulatory variants using HapMap3 lymphoblastoid cell lines, which further highlights the unique biological significance of candidate adaptive loci underlying CAD. Our study provides a novel approach for detecting selection on polygenic traits and evidence that modern human genomes have evolved in response to CAD-induced selection pressures and other early-life traits sharing pleiotropic links with CAD.

Author summary

How genetic variation contributes to disease is complex, especially for those such as coronary artery disease (CAD) that develop over the lifetime of individuals. One of the fundamental

Funding: This study was supported by the National Health and Medical Research Council (NHMRC) of Australia (grant no. 1062227) and the National Heart Foundation of Australia. MI was supported by a Career Development Fellowship co-funded by the NHMRC and the National Heart Foundation of Australia (no. 1061435). GA was supported by an NHMRC Peter Doherty Early Career Fellowship (no.1090462). SR was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (Grant No 213506 and 129680), Academy of Finland (Grant No 251217 and 285380), the Finnish Foundation for Cardiovascular Research, the Sigrid Juselius Foundation, Biocentrum Helsinki and the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No 201413 (ENGAGE) and 261433 (BioSHaRE-EU), and Horizon 2020 Research and Innovation Programme under grant agreement No 692145 (ePerMed). The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

questions about CAD––whose progression begins in young adults with arterial plaque accumulation leading to life-threatening outcomes later in life––is why natural selection has not removed or reduced this costly disease. It is the leading cause of death worldwide and has been present in human populations for thousands of years, implying considerable pressures that natural selection should have operated on. Our study provides new evidence that genes underlying CAD have recently been modified by natural selection and that these same genes uniquely and extensively contribute to human reproduction, which suggests that natural selection may have maintained genetic variation contributing to CAD because of its beneficial effects on fitness. This study provides novel evidence that CAD has been maintained in modern humans as a by-product of the fitness advantages those genes provide early in human lifecycles.

Introduction

It is well established that modern human traits are a product of past evolutionary forces that have shaped heritable variation, but we are far from a good understanding of whether recent natural selection has acted on these and how this process has left its imprint across the genome. While many genome-wide multi-population scans have searched for signatures of positive selection [1–9], these studies have detected relatively few adaptive candidates for common traits or diseases [10–12]. This suggests that classic 'selective sweeps' have been relatively rare in recent human history [13–16] and that current tools may not be appropriate for detecting and validating smaller shifts in adaptive variation, thus limiting our understanding of how natural selection acts on common diseases and traits [12]. Research in this area is also important as the combination of positive selection and significant GWAS signals at the same locus supports the existence of functional variation for disease. Here, we aimed to comprehensively identify selection signals for coronary artery disease (CAD) loci with methods designed to detect recent signals of positive selection. We compared selection signals in 12 worldwide populations (HapMap3) with CAD genetic risk (CARDIoGRAMplusC4D) to help understand how selection acts on disease variation at the genetic level and prioritize genes most likely modified in relation to CAD. We examined the association between selection signals and gene expression to further test whether adaptive candidates are functionally important for CAD in terms of gene regulation. Lastly, we tested if CAD genes are associated with reproductive fitness to try to understand why this common disease persists in modern humans.

Classic population genetics theory describes positive selection with the selective-sweep (or hard-sweep) model, in which a strongly advantageous mutation increases rapidly in frequency (often to fixation) resulting in reduced heterozygosity of nearby neutral polymorphisms due to genetic hitch-hiking [17, 18] and a longer haplotype with higher frequency. Many methods have been developed to detect these signatures [19, 20], including traditional tests that detect differentiation in allele frequencies among populations (i.e. Wright's fixation index, Fst [21]) and more recently developed within population tests for extended haplotype homozygosity (i.e. integrated haplotype score, iHS [9]). Some of the most convincing examples of human adaptive evolution have been uncovered for traits influenced by single loci with large effects. For example, the lactase persistence (*LCT*) and Duffy-null (*DARC*) mutations affecting expression of key proteins in milk digestion [10] and malarial resistance [22] both display hallmarks of selective sweeps. Other loci that are not clearly monogenic but also show selective sweeps are associated with high-altitude tolerance (*EPAS1* [23]) and skin pigmentation (*SLC24A5* and

KITLG [24]). These studies show that rapid selective sweeps mainly occurred for new mutations with large effects on phenotypes.

Motivated by these initial successes and the increasing availability of global population data genotyped on higher resolution arrays (i.e. HapMap Project, 1000 Genomes Project), many recent genome-wide scans for candidate adaptive loci have recently been performed [11]. These suggest that selection may have operated on a variety of biological processes [10] in ways that differ among populations (i.e. local adaptation) [25], been prevalent in genetic variation linked to metabolic processes [26], and may often target intergenic regions and gene regulatory variants rather than protein-coding regions [12]. Often only the larger signals underlying monogenic (or near-monogenic) traits are typically considered for follow-up because of losses in the statistical power needed to quantify significance for smaller candidate adaptive signals after correcting for genome-wide multiple testing [20]. The adaptive status of many smaller candidate signals also remains uncertain due to inconsistencies in results between studies that utilized the same data [14], and it is inherently more difficult to functionally validate candidate adaptive signals underlying complex polygenic traits compared to monogenic traits where only one or a few variants may have been under selection due to their influence on fitness [27, 28].

In contrast to population genetics, research in quantitative genetics has shown that rapid adaptation can often occur on complex traits that are highly polygenic [29, 30]. Under the 'infinitesimal (polygenic) model', such traits are likely to respond quickly to changing selective pressures through smaller allele frequency shifts in many polymorphisms already present in the population [13, 31]. Selection on such variation is generally less likely to push it towards fixation due to genetic correlations, thus producing smaller changes in surrounding heterozygosity over time that are harder to detect with most current population genetic methods [14, 28, 32]. Note that polygenic and classic sweep models are not mutually exclusive [13, 33], for alleles with small- and large-effects may both underlie a polygenic trait, which suggests that there will be some variation in the degree to which candidate alleles are modified after selective events. Because most common diseases are highly polygenic, we need to improve how we detect and classify the adaptive signatures underlying these traits.

Recent studies investigating genomic selection on polygenic traits have taken two approaches. The first scans for significant selection signals for a subset of large effect SNPs that have previously been identified as genome-wide significant. For example, Ding and Kullo [34] found significant population differentiation (Fst) for 8 of 158 index SNPs underlying 36 cardiovascular disease phenotypes, and Raj et al. [35] observed elevated positive selection scores (Fst, iHS) for 37 of 416 index susceptibility SNPs underlying 10 inflammatory-diseases. The second approach tests if aggregated shifts in genome-wide significant allele frequencies are associated with phenotypic differences by population, latitudinal, or environmental gradients, which might indicate local adaptation. For example, Castro and Feldman [36] used 1300 index SNPs underlying many polygenic traits and found elevated adaptive signals (Fst and iHS) above background variation, and Turchin et al. [37] demonstrated moderately higher frequency of 139 height-increasing alleles in a Northern (taller) compared to Southern (shorter) European populations. These approaches all assume that genome-wide significant variants are the most probable selection targets, but many if not most such variants are tags for the causal variants, which may be at lower frequencies. This suggests an approach more sensitive for detecting subtle signals of polygenic selection is needed.

We chose CAD as a model for examining polygenic selection signals for complex disease because it has (and continues to) impose considerable disease burden (and possible selection pressure) in humans [38], its underlying genetic architecture has been extensively studied [39, 40] and many of its risk factors (cholesterol, blood pressure) have been under recent natural

selection [41] related to potential pleiotropic effects or tradeoffs with CAD. Antagonistic pleiotropy describes gene effects on multiple linked traits where selection on one may cause negative fitness effects (i.e. reproduction, survival, and disease) in the other due to their antagonistic genetic association [42]. Two common misconceptions are that CAD only occurs in older people and is a disease that has mainly afflicted modern humans. If either were true, selection might not have had either the opportunity or sufficient time to affect genetic variation associated with CAD. However, CAD begins to manifest during reproductive ages [43, 44] and disease origins can be detected even in adolescence through degree of atherosclerosis [44, 45] and myocardial infarction events [46]. CAD is also a product of many heritable risk factors (cholesterol, weight, blood pressure) whose variation is expressed during the reproductive period, when CAD could drive selection directly or indirectly. Furthermore, CAD has impacted human populations since at least the ancient Middle Kingdom period, with atherosclerosis detectable in Egyptian mummies [47]. This suggests that there has been enough time for evolutionary responses to CAD to have occurred, genomic signatures from which may be detectable in modern humans.

By combining several 1000 Genomes-imputed datasets including HapMap3 and Finnish SNP data, a large genetic meta-analysis of CAD, HapMap3 gene expression data and lifetime fitness data from the Framingham Heart Study, we sought to address the reason(s) why CAD exists in humans by answering the following questions: 1) Has selection recently operated on CAD loci? 2) How do selection signals underlying CAD loci vary among populations and are they enriched for gene regulatory effects? 3) Do candidate adaptive signatures overlap directly with CAD genetic risk and is this useful for highlighting disease-linked selection signals? 4) Do CAD-linked selection signals display functional effects and evidence of antagonistic pleiotropy, in that they are also linked to biological processes or traits influencing reproduction?

Results

To test for selection signals for variants directly linked with CAD, we utilized SNP summary statistics from 56 genome-wide significant CAD loci in Nikpay et al. [40], the most recent and largest CAD case-control GWAS meta-analysis to date, to identify 76 candidate genes for CAD (see Methods). Nikpay used 60,801 CAD cases and 123,504 controls from a mix of individuals of mainly European (77%), south (13% India and Pakistan) and east (6% China and Korea) Asian, Hispanic and African American (~4%) descent with genetic variation imputed to a high-density using the 1000 Genomes reference panel. By investigating all SNPs in CAD genes, we aimed to improve detection of smaller polygenic selection signals for the range of functional genic variants and short-range intergenic regulatory variants that would be missed with approaches that only consider genome-wide significant SNPs.

Signals of positive selection within coronary artery disease loci

We utilised the integrated Haplotype Score (iHS) to estimate positive selection for each SNP underlying CAD genes within each population separately. Because iHS is typically used to detect candidate adaptive SNPs where the selected alleles may not have reached fixation [9], this estimate is well suited for detecting recent signals of selection as opposed to other measures [20]. iHS is also better suited for detecting selection acting on standing variation in polygenic traits [20, 48].

Candidate selection signals were found for many of the 76 CAD genes within each of the 12 worldwide populations (11 HapMap3 populations and Finns; Fig 1A for top 40 based on their association with CAD log odds genetic risk, S1 Fig for all 76). These were defined as 'peaks' of

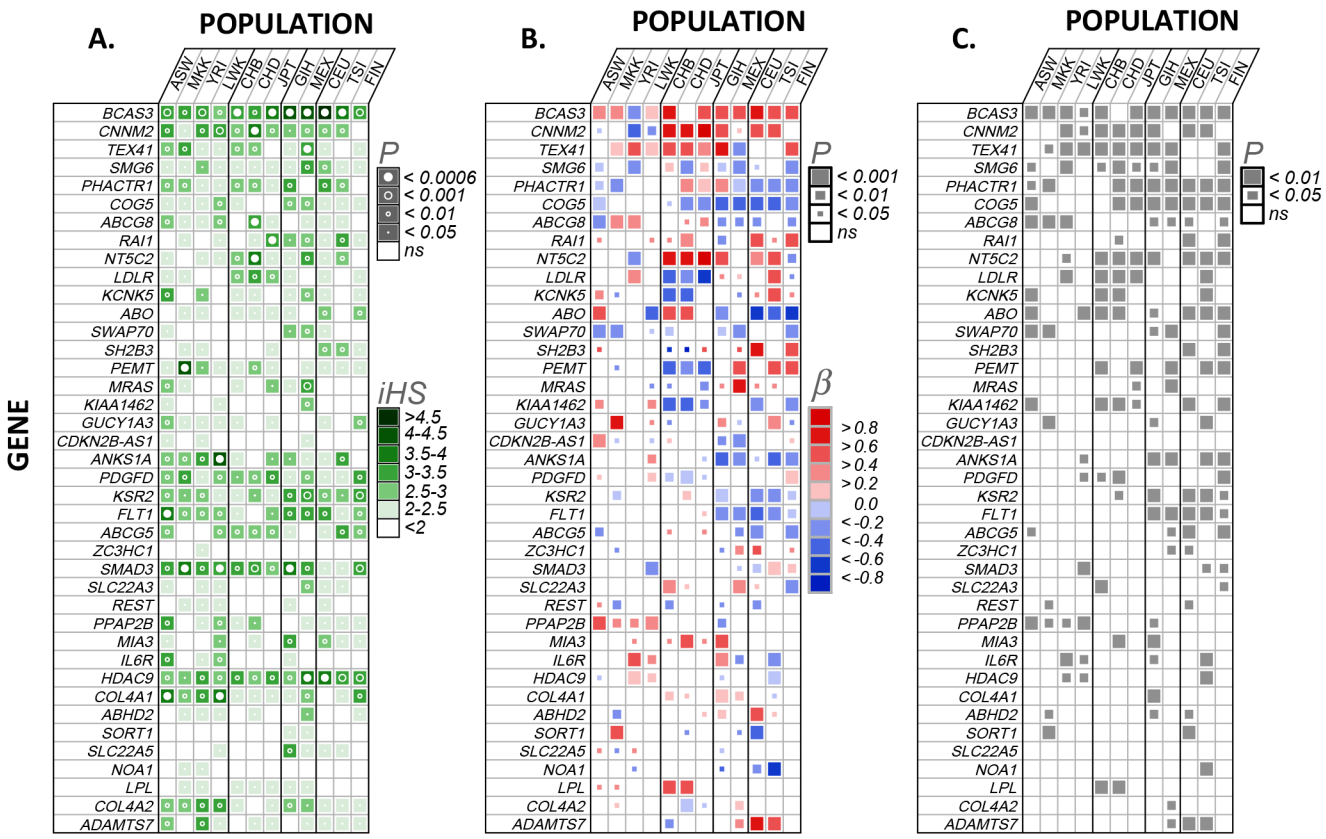

Fig 1. Association of coronary artery disease (CAD) genetic risk and positive signatures of selection in 12 worldwide populations. The 40 of 76 CAD genes investigated are shown that have at least four significant selection-risk associations in Panel B across all 12 populations. **Panel A.** Magnitude and significance of largest positive selection signal (integrated haplotype score, iHS) within each gene-population combination. P values (circles within squares) were obtained from 10000 permutations. Bonferroni corrected p value limit also shown ($\alpha = 0.05/76 = 0.000657$) with closed circles. **Panel B.** Null hypothesis: no association between CAD genetic risk and positive selection, tested using mixed effects model with SNP estimates of CAD log odds genetic risk and iHS while accounting for gene LD structure as a random effect (first eigenvector from LD matrix per gene). Scaled regression coefficients were obtained directly from regressions, each p value from 10000 permutations. **Panel C.** Null hypothesis: association between genetic risk and positive selection for SNPs within CAD genes no different than non-CAD associated genes. Permuted p values were estimated by comparing each p value in Panel B against 100 nominal p values obtained by randomly choosing (without replacement) 100 non-CAD associated genes of similar size across the genome and using the same mixed effects model setup as described above. **Populations**. Grouped by ancestry, African (ASW, African ancestry in Southwest USA; MKK, Maasai in Kinyawa, Kenya; YRI, Yoruba from Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya), East-Asian (CHB, Han Chinese subjects from Beijing; CHD, Chinese in Metropolitan Denver, Colorado; JPT, Japanese subjects from Tokyo), European (CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; TSI, Tuscans in Italy; FIN, Finnish in Finland), GIH (Gujarati Indians in Houston, TX, USA), MEX (Mexican ancestry in Los Angeles, CA, USA).

https://doi.org/10.1371/journal.pgen.1006328.g001

significantly elevated iHS scores across SNPs within each gene-population combination, with the apex approximating the likely positional target of positive selection.

The results for the largest iHS score per gene and population (Fig 1A) show that most candidate selection signals were relatively small, but a few larger signals were detected. For example, out of the 912 gene-by-population combinations (S1 Fig), 354 (38%) contained weak-moderate candidate selection signals (significant iHS between 2–3), 84 (9%) contained moderate-strong signals (significant iHS between 3–4), and 6 (0.6%) had very strong signals (significant iHS > 4). The 6 largest candidate signals were found in the following gene-population combinations: *BCAS3* in GIH (iHS = 4.45), MEX (iHS = 4.23) and CEU (iHS = 4.86), *PEMT* in MKK (iHS = 4.24), *ANKS1A* in LWK (iHS = 4.03), and *CXCL12* in JPT (iHS = 4.10), with all iHS p values <0.0001. Six genes (*BCAS3*, *SMG6*, *PDGFD*, *KSR2*, *SMAD3*, *HDAC9*) exhibited

Table 1. Leading multiple candidate selection signals in *PHACTR1* SNPs.

rs2015764	rs4142300	rs8180558			rs4715043	rs6924689	
12788283 bp	12825772 bp	12919989 bp			12987641 bp	13025819 bp	
MEX, 2.08*	GIH, 3.73***	ASW, 2.43**	LWK, 2.75**	YRI, 2.17*	CHB, 2.26*	CHD, 2.89**	JPT, 3.00**
rs4273688—rs11760186					rs9349549		
13192799–13196011 bp, intron 7					13277029 bp, intron 11		
ASW, 2.38**	LWK, 2.00*	MKK, 2.95**	CHD, 2.05*	GIH, 2.13*	MKK, 2.91**	CEU, 2.71**	TSI, 2.96**

Values include SNP rsID, build 37 base pair chromosomal position, population, absolute integrated Haplotype Score |iHS| and permuted p values (*p<0.05; **p<0.01; ***p<0.001).

Populations: ASW (African ancestry in Southwest USA), MKK (Maasai in Kinyawa, Kenya), YRI (Yoruba from Ibadan, Nigeria), LWK (Luhya in Webuye, Kenya), CHB (Han Chinese subjects from Beijing), CHD (Chinese in Metropolitan Denver, Colorado), JPT (Japanese subjects from Tokyo), CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection), TSI (Tuscans in Italy), GIH (Gujarati Indians in Houston, TX, USA), MEX (Mexican ancestry in Los Angeles, CA, USA).

https://doi.org/10.1371/journal.pgen.1006328.t001

candidate selection signals consistently within all populations (Fig 1A), and many genes also contained consistent selection signals for all populations within similar ancestral groups (e.g. African, European etc, Fig 1A).

Within CAD genes, multiple candidate selection signals were sometimes present (particularly within larger genes, within separate linkage disequilibrium (LD)-blocks); these varied between and sometimes within a population. For example, eleven (of the twelve) populations had candidate selection signals in *PHACTR1* introns 4, 7 or 11 (Table 1; see also S2 Fig, comparing cross-population selection signals in *PHACTR1*) that were in separate LD-blocks (see S2 Fig, LD plots). For eight populations, there was a broad and relatively weak set of candidate selection signals in intron 4 (the largest *PHACTR1* intron, ~300kb in length). Intron 4 is also the location of the published CAD index SNP (rs12526453) for *PHACTR1*. Other interesting candidate selection signals present in other CAD genes (S1 Fig) are not discussed here. Such patterns suggest that candidate selection signals are sometimes complex and often do not correspond to the SNPs with largest effect on disease.

Relationship between CAD genetic risk and selection across populations

For each CAD gene within each population, we used a mixed effects linear model to regress SNP-based estimates of CAD log odds genetic risk (ln(OR), obtained from cardiogramplusc4d.org) against iHS selection scores (see Methods). We accounted for LD structure by including the first eigenvector from an LD matrix of correlations (r^2) between SNPs within each gene as a random effect.

For a subset of CAD loci, we found significant quantitative associations between disease risk and selection signals and for each of these the direction of this association was often consistent between populations (Fig 1B). Furthermore, when compared to a null distribution of genes selected randomly from the genome, the strength of the CAD log odds versus selection signal at most loci was statistically significant (Fig 1C). Fig 1B shows 40 genes ranked based on those with the most consistent number of significant associations across the 12 populations, with those that showed fewer than four significant associations excluded. Positive and negative associations indicate elevated selection signals present in regions with higher or lower CAD log odds genetic risk, respectively.

In the comparison across populations, directionality of significant selection-risk associations tended to be most consistent for populations within the same ancestral group (Fig 1B). For example, in *PHACTR1*, negative associations were present within all European populations

(CEU, TSI, FIN), and in *NT5C2* strong positive associations were present in all East Asian populations (CHB, CHD, JPT). Other negative associations that were consistent across all populations within an ancestry group included five genes in Europeans (*COG5*, *ABO*, *ANKS1A*, *KSR2*, *FLT1*) and four genes (*LDLR*, *PEMT*, *KIAA1462*, *PDGFD*) in East Asians.

Additional consistent positive associations included four genes (*CNNM2*, *TEX41*, *NT5C2*, *MIA3*) in East Asians, three (*BCAS3*, *RAI1*, *KCNK5*) in Europeans, and one (*PPAP2B*) in Africans. In comparison to other ancestral groups, African populations showed fewer significant selection-risk associations (27.9% of all 76-gene x 12-population combinations) than Asians (31.5%) or Europeans (32.8%). Some associations were consistent in all but one population (e.g. *CNNM2*, *ABCG8* in Europeans; *BCAS3*, *KCNK5* in Asians; *CNNM2*, *TEX41* in Africans) or unique to one population within an ancestral group (e.g. *TEX41* in FIN, *COG5* in ASW).

Below we focus on *BCAS3* (Fig 2) and *PHACTR1* (Fig 3), two of the strongest selection-risk associations which, when adjusting for LD (see Methods), displayed varying directionality between at least two populations.

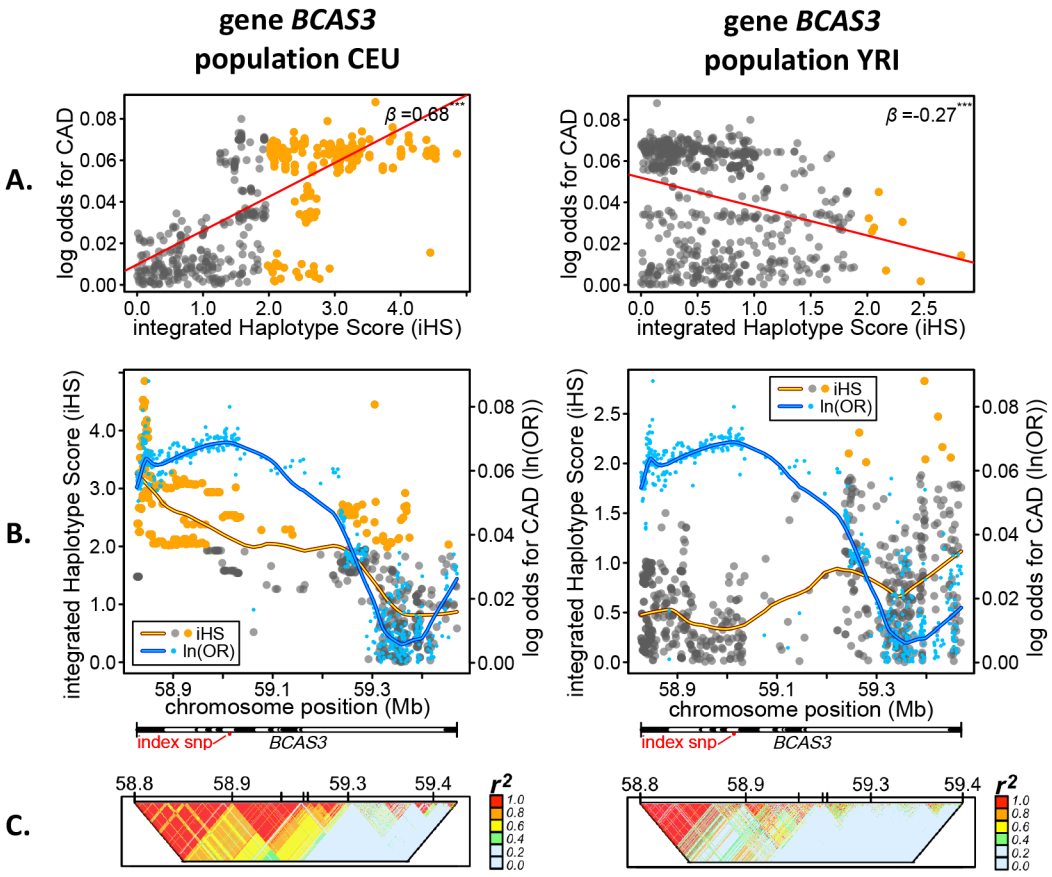

Fig 2. Quantitative links between coronary artery disease risk and selection signals in *BCAS3*. A. Correlation between selection signals (iHS) and coronary artery disease (CAD) log odds genetic risk (log odds, ln (OR)), both represented as absolute values. Red line/upper right value, *β* from mixed effects regression. **B.** Base pair positional comparison of selection signals and CAD genetic risk across *BCAS3*. Blue points, CAD log odds values; grey-orange or non-significant-significant points, iHS scores. Horizontal bar shows *BCAS3* gene (and intron) span and location of lead index SNP. Blue/orange lines are smoothed lines estimated with loess function in R. **C.** LD plots, r^2. Populations: CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; YRI, Yoruba from Ibadan, Nigeria.

https://doi.org/10.1371/journal.pgen.1006328.g002

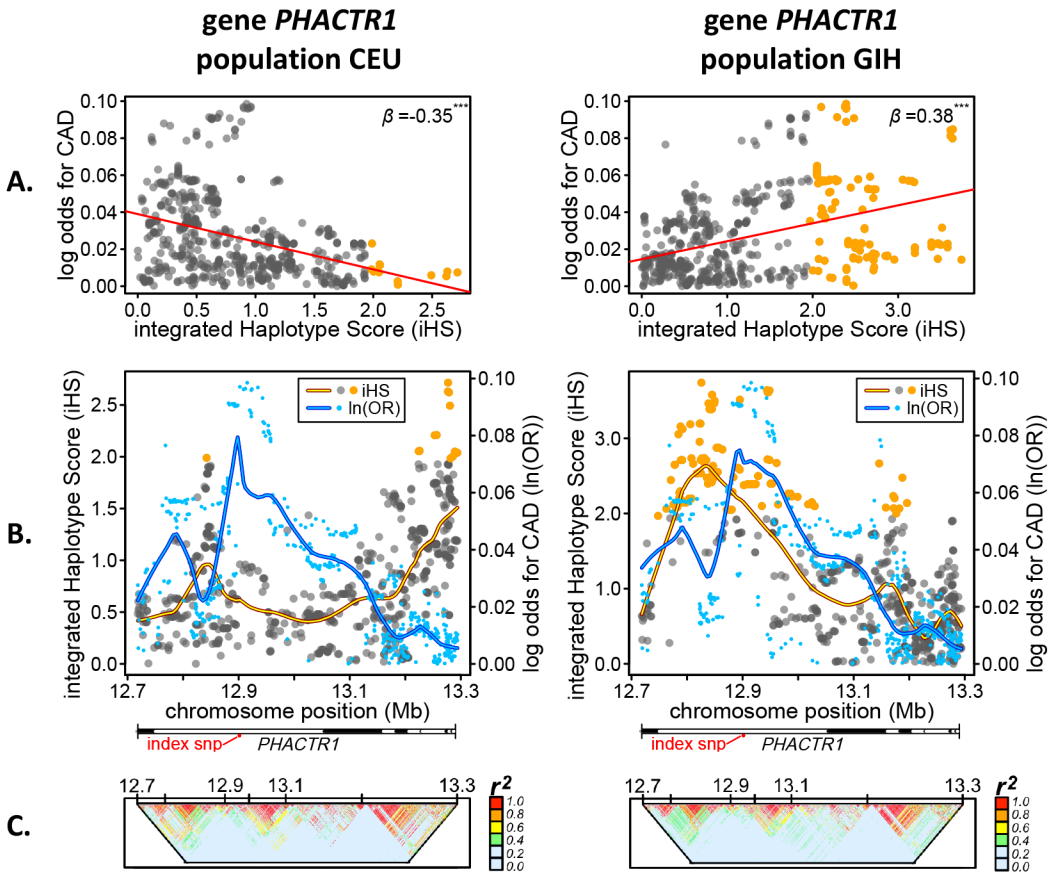

Fig 3. Quantitative links between coronary artery disease risk and selection signals in *PHACTR1*. A. Correlation between selection signals (iHS) and coronary artery disease (CAD) log odds genetic risk (ln(OR)), both represented as absolute values. Red line/upper right value, *β* from mixed effects regression. **B.** Base pair positional comparison of selection signals and CAD genetic risk across *PHACTR1*. Blue points, CAD log odds values; grey-orange or non-significant-significant points, iHS scores. Horizontal bar shows *PHACTR1* gene (and intron) spans and location of index SNP if present. **C.** LD plots, r^2. Populations: CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; GIH, Gujarati Indians in Houston, TX, USA.

https://doi.org/10.1371/journal.pgen.1006328.g003

Genetic risk of CAD vs positive selection in *BCAS3*

The genetic risks of CAD for variants in *BCAS3* were positively correlated with an extremely large candidate adaptive signal in all European and two of three East Asian populations (Fig 1B). For example in CEU, the largest iHS score was 4.85 and highly significant, and was elevated across most of *BCAS3* (Fig 2B CEU, spanning introns 1–18 and various LD-blocks, Fig 2C), which matched the approximate trends in CAD log odds giving rise to a highly significant positive correlation (Fig 2A CEU). In contrast, in YRI there was no detectable selection signal close to the index SNP (Fig 2B YRI), but weak-moderate signals were present towards the end of *BCAS3* (Fig 2B YRI, introns 18–19, smaller LD-blocks Fig 2C), which also corresponded with lower CAD log odds (Fig 2B, YRI) thus giving rise to a significant negative correlation in Fig 2A.

Genetic risk of CAD vs positive selection in *PHACTR1*

For all European populations, *PHACTR1* (see CEU example, Fig 3A) selection peaks were typically located within regions of consistently lower CAD log odds (Fig 3B). This contrasted with

most other non-European populations where the highest candidate selection peaks were located within regions with elevated CAD log odds (including the index CAD SNP rs12526453, intron 4). The largest selection peak in GIH (Fig 3B) overlapped the CAD log odds peak in *PHACTR1* giving rise to the strong positive association seen in Fig 3A. The two distinctive selection peaks in both CEU and GIH were separated by different LD-blocks (Fig 3C), suggesting that these may have developed independently within *PHACTR1*. Interestingly, the negative association found for the MKK population was due to the location of the selection peaks more closely matching those of the European populations in intron 11 (S2 Fig).

Enrichment of gene regulatory variants under selection at CAD loci

To establish whether variants with evidence of selection in CAD genes also showed evidence of function, we performed an eQTL scan in 8 HapMap3 populations with matched LCL gene expression. We compared all SNPs in each CAD locus against expression for each focal gene within each population.

We found that SNPs with significant integrated Haplotype Scores (iHS) were often also involved in gene regulation, compared to SNPs with non-significant selection scores (Fig 4, Kolmogorov-Smirnov test p value <0.001). To assess which biological pathways were enriched for the highest-ranked genes according to Fig 1B, i.e. those where selection scores were most closely associated with CAD log odds genetic risk, we included the top 10 genes into the Enrichr analysis tool [49] and found that these genes are especially enriched in pathways related to metabolism, focal adhesion and transport of glucose and other sugars. More interestingly, we found connections to reproductive phenotypes in the associations of these genes with pathways, ontologies, cell types and transcription factors. For example, we found links to ovarian steroidogenesis and genes expressed in specific cell types and tissues including the ovary, endometrium and uterus (see S1 Table for Enrichr outputs).

Enrichment of CAD loci for lifetime reproductive outcomes and antagonistic effects

To test whether CAD genes are directly associated with human lifetime reproductive success (LRS or total number of children born across reproductive lifetimes), a prerequisite for responses to selection, we examined their association with LRS for women in the Framingham Heart Study (FHS). Out of the 76 CAD genes (representing 20,254 SNPs in total; a minimum, average and maximum of 18, 266 and 2121 SNPs tested per gene, respectively), 51 genes contained SNPs that were significantly nominally associated with LRS ($p<0.05$), 30 genes contained SNPs associated at $p<0.01$ and 12 genes contained SNPs associated at $p<0.001$, based on both nominal p values from FaST-LMM and permuted p values (see S2 Table). For example, the most significant associations per gene included rs56152906 in *PPAP2B* (p = 5.23E-06, permuted $p<0.0001$), rs7896502 in *LIPA* (p = 0.0002, permuted p = 0.0001) and rs2479409 in *PCSK9* (p = 0.0003, permuted p = 0.0001) including a further 9 (*COL4A2*, *FLT1*, *HDAC9*, *KSR2*, *LPA*, *MIA3*, *PDGFD*, *PLG*, *SMAD3*) genes with significant LRS associations at permuted $p<0.001$. The two previous studies that have investigated genome-wide SNP associations with LRS found associations with similar levels of evidence to our study. For example, the leading SNP in Kosova et al. [50] for completed family size was rs10966811 with p = 5.57E-06. The top two leading SNPs in Aschebrook-Kilfoy et al. [51] for LRS were rs10009124 (p = 7.65E-08) and rs1105228 (p = 2.16E-06).

When we considered these associations using fastBAT that combines SNP associations within a gene (accounting for LD-redundancy) into single gene-level p value, similar results were obtained with 8 genes significantly associated with LRS (e.g. *PPAP2B*, p = 0.0004, permuted p = 0.001, *SMAD3*, p = 0.0061, permuted p = 0.007, *MIA3*, p = 0.008, see S2 Table).

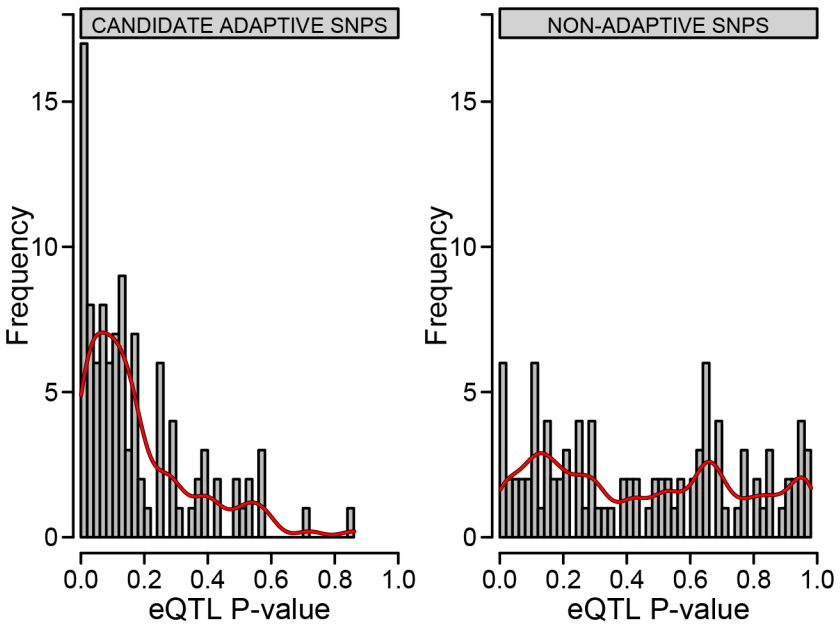

Fig 4. Comparing positive selection with gene regulation. Summary distribution of permuted eQTL p values for SNPs with (left) or without (right) a significant selection signal. SNPs with a significant selection signal (iHS) were chosen by taking the largest significant positive selection signal (if one was present) within each gene-population combination. The same number of SNPs without a significant selection signal were also randomly drawn across all gene-population combinations for comparison. These SNPs were used in an eQTL analysis where they were regressed (including gender as a covariate) against their associated gene probe's expression.

https://doi.org/10.1371/journal.pgen.1006328.g004

To test the null hypothesis that CAD variation is no more significantly associated with LRS than is variation in the rest of the genome, we used a permutation approach. We sampled 20,254 non-CAD related SNPs (matched within MAF bins to the CAD SNPs) randomly (without replacement) across the genome 100 times. The permuted p value was based on the number of times each random sample of 20,254 non-CAD SNPs shared significantly more associations with LRS than did the 20,254 CAD SNPs. The total sample of randomly selected SNPs (n = 2,025,400) was also compared against the 20,254 CAD SNPs with a Kolmogorov-Smirnov (K-S) test. We found that CAD genetic variation was significantly (p = 9.49E-08 and p = 1.90E-07 based on one- and two-sided K-S tests, respectively; permuted $p<0.01$) more enriched for LRS compared to the rest of the genome (see S2 Table for other fitness-related traits), providing strong evidence in the FHS for shared fitness effects at CAD loci. This was also the case when we tested this at the gene-level using fastBAT results (permuted p = 0.026, S2 Table).

To test whether effects between CAD loci and LRS were antagonistic, we cross-referenced the genome-wide significant index SNPs for CAD from Nikpay [40] with significant SNPs for LRS from the FaST-LMM analysis. Of the 56 CAD index SNPs in Nikpay [40], 53 were genotyped or imputed in the FHS to a high confidence. In FHS, six of those SNPs (11.3%) were significantly associated with LRS (FaST-LMM $p < 0.05$), with 5 out of those 6 antagonistic, i.e. the allele that increases LRS also increases risk for CAD (see S3 Table). For example, in *FLT1*, rs9319428-A significantly increases both LRS (ß = 0.041, p = 0.0143) and CAD risk (ß = 0.039, p = 7.13E-05), and similarly, rs2048327-C in *LPA* significantly increases both LRS (ß = 0.041, p = 0.00894) and CAD risk (ß = 0.057, p = 2.46E-09). This suggests that antagonistic effects occur in some loci, but the power to detect and define this for smaller effect variants on LRS is limited in the FHS (e.g. see S3 Fig for power estimates). Compared to the CARDIoGRAMplusC4D study [40]

where the 56 genome-wide significant CAD index SNPs were obtained using a meta-sample of ~184,000 individuals, SNP effects on LRS were based on 1,579 women from the FHS. Given that power to detect small effects (i.e. |ß|<~0.3–0.4 or OR <~1.2–1.3) in these studies is poor when *n* is small (i.e. ~1000 individuals [52]) suggests that larger samples of women and men with completed reproduction are needed to test for antagonistic effects comprehensively to avoid false negatives.

We further tested whether SNPs are associated with both LRS and CAD due to potential confounding effects rather than antagonistic pleiotropy, i.e. confounding effects would occur if CAD SNPs influence LRS, which in turn cause significant changes in CAD risk due to physiological, hormonal or social changes related to childbearing/rearing. We tested the association between CAD SNPs and CAD in FHS females, stratified by LRS (see S3 Fig for full analysis). We found no significant effect of LRS modifying SNP effects on CAD (see S3 Fig), which supports the antagonistic pleiotropy hypothesis, however we caution that larger, better powered studies may show some level of attenuation.

Extending this investigation to understand why CAD genes are significantly enriched for LRS, i.e. what possible underlying reproductive processes are contributing, we performed an extensive systematic literature search on the 40 top-ranked genes in Fig 1 and a random set of 20 non-CAD genes. While gene set enrichment had been performed (above) suggesting some connections to reproductive phenotypes, such tools cannot capture the full range of possible effects on multiple fitness traits, some that are themselves rarely tested in other mammalian (non-human) species due to ethical limitations. We found evidence for direct links between CAD genes and fitness (S4 and S5 Tables) including genes associated with reproductive (*PPAP2B*, [53]) or twinning (*SMAD3*, [54]) capacity and number of offspring produced (e.g. *KIAA1462*, [55], *SLC22A5*, [56]). *PHACTR1*, *LPL*, *SMAD3*, *ABO* and *SLC22A5* may contribute to reproductive timing (menarche, menopause) in women [57–59] and animals [60]. Expression of *PHACTR1* [61], *KCNK5* [62], *MRAS* and *ADAMST7* [63] appear to regulate lactation capacity. Some gene deficiencies also cause pregnancy loss (e.g. *LDLR*, [64], *COL4A2*, [65]). Evidence for other pleiotropic links related to fitness included 25 genes that shared links with traits expressed during pregnancy (S4 and S5 Tables), i.e. variation that can negatively influence the health and survival outcomes of both the fetus and mother [66]. For example, a variant of *CDKN2B-AS1* significantly contributes to risk of fetal growth restriction [67], both *FLT1* [68] and *LPL* [69] are significantly differentially expressed in placental tissues from pregnancies with intrauterine growth restriction (IUGR), and preeclampsia and *LDLR*-deficient mice had litters with significant IUGR [70]. A further 29 and 19 genes were linked to traits that can directly influence female and male fertility, respectively (13 influence both) (S4 and S5 Tables). For example, *BCAS3* and *PHACTR1* are highly expressed during human embryogenesis [71, 72], *SWAP70* is intensely expressed at the site of implantation [73], and *PHACTR1* may play a role in receptivity to implantation [74]. For *ABCG8* and *KSR2*, animal models provide further support as gene expression deficiency can cause infertility in females (*ABCG8*, [75]) and males (*KSR2*, [76]).

Pleiotropic connections were also apparent in the classification of specific disorders or from studies investigating single-gene effects. For example, women with polycystic ovarian syndrome (PCOS) have higher rates of infertility due to ovulation failure and modified cardiovascular disease risk factors (i.e. diabetes, obesity, hypertension [77]). While reduced fecundity associated with PCOS might suggest it would not fit the model of antagonistic pleiotropy, some hypothesize that it is an ancient disorder and may have provided a rearing advantage in ancestral food-limited environments [78]. A number of CAD genes in this study (e.g. *PHACTR1*, *LPL*, *PDGFD*, *IL6R*, *CNNM2*) are found differentially expressed in PCOS women [79–83], suggesting possible links between perturbed embryogenesis and angiogenesis. In males, this can be demonstrated

with a mutation in *SLC22A5* that causes both cardiomyopathy and male infertility due to altered ability to break down lipids [84, 85]. More generally, many recent studies link altered cholesterol homeostasis with fertility, which is most apparent in patients suffering from hyperlipidemia or metabolic syndrome [86, 87].

For the random set of non-CAD genes that were approximately the same size as the top 20 genes in Fig 1, we were only able to find three (out of 20) with at least one potential link with fitness (S6 Table) using the same systematic literature search further demonstrating the relative abundance of CAD loci effects on fitness earlier in life.

Discussion

This study identified many candidate adaptive signals suggesting that selection on CAD loci is much more widespread than previously appreciated (also see S1 Discussion). It has previously been suggested [12] and demonstrated [88] that selection on gene expression levels has been an important element of human adaptation in general. We confirm this result for CAD associated loci. Positive selection signals within CAD loci were more likely than random SNPs to be associated with gene expression levels in *cis* (Fig 4).

We found evidence that some of these signals may be a result of selection pressures induced directly by CAD itself. This finding is important for highlighting genes that may have been modified directly by selection on disease phenotypes and also for our general understanding of how quickly human genomes can respond to selection induced by changing environments. Subsequent fitness and biological process analyses and a thorough literature review demonstrated that CAD loci are enriched for lifetime reproductive success in women and also linked to other male and female reproductive phenotypes, which suggests both their potential to respond to natural selection and their possible role via antagonistic pleiotropy in the reproductive tradeoffs that would help to explain why CAD is common in modern humans. While the connection between cardiovascular disease and lifetime parity is not novel (e.g. see [89, 90]), it is not known whether this connection is due to hormonal, physiological, social or selective processes. The current study provides the first evidence for a selective and antagonistic mechanism.

Coronary artery disease-induced changes to human genomes

One of our most interesting findings was the significant association between selection signals and CAD log odds genetic risk. This approach of integrating genome scans of positive selection with genome-wide genotype-phenotype data has been promoted previously as a tool to uncover biologically meaningful selection signals of recent human adaptation [12, 88] but has rarely been applied. Among the exceptions, Jarvis et al. [91] found a cluster of selection and association signals coinciding on chromosome 3 that included genes *DOCK3* and *CISH*, which are known to affect height in Europeans.

For highly-ranked genes (according to the number of significant associations present within the 12 populations) in Fig 1B such as *BCAS3*, *CNNM2*, *TEX41*, *SMG6* and *PHACTR1*, the consistent overlap between selection and genetic risk of CAD suggests that many of these may have been modified by CAD-linked selective pressures. If so, then two conditions must have been met. Firstly, CAD was present for long enough to be involved in these genetic alterations, an evolutionary process which generally takes thousands of years. Indeed, precursors of CAD (i.e. atherosclerosis) are detectable in very early civilizations [47]. Secondly, the effects of CAD were directly or indirectly expressed during the reproductive period and trait variation was under natural selection due to its effects on reproductive success.

It is only possible for natural selection to directly act on CAD if those outcomes modify individual fitness relative to others in the same population. As outlined in the introduction, this is possible as CAD outcomes (i.e. myocardial infarction) do occur in young adults. However, early-life CAD outcomes are relatively rare, suggesting selection is more likely to operate indirectly on CAD via its risk factors (or other pleiotropically linked traits, discussed below), which provides a more likely explanation for the close associations we found between positive selection and genetic risk. Supporting this, phenotypic selection has been found operating on CAD risk factors [41], suggesting that these selection pressures are still present in modern humans.

Some genes had large signals of selection but showed weak or no consistent overlap with CAD genetic risk. For example *HDAC9* (Histone Deacetylase 9) shows extensive evidence for having undergone recent selection within most populations, especially those of European or Mexican decent, but little or no overlap with CAD risk was evident in most populations. This suggests positive selection has operated on this gene due to its effects on a trait unrelated to CAD, which may not be surprising given *HDAC9*'s broad biological roles (as a transcriptional regulator, cell-cycle progression) and association with other very different phenotypes including ulcerative colitis [92] and psychiatric disorders [93]. This further demonstrates that this approach is useful for separating candidate selection signals important for the disease or phenotype of interest from those that aren't.

Pleiotropic effects that establish the genetic foundations of tradeoffs

We found direct evidence in the Framingham Heart Study for shared fitness effects at CAD loci, which were specifically significantly enriched for their effects on female lifetime reproductive success relative to the rest of the genome. This novel finding shows a connection between direct fitness and later disease expressed through CAD loci. An extensive literature review supported this conclusion. All 40 CAD genes from Fig 1 shared at least one (often more) connection with fitness (S4 and S5 Tables). Some appear to directly influence fitness (offspring number, age at menarche, menopause, survival), while many were associated with early-life reproductive traits that are likely to correlate with fitness, including variation in ability to fertilize/conceive or fetal growth, development and survival. This suggests further pleiotropic links between CAD and early-life fitness-related traits. Directly testing for antagonistic effects between fitness and CAD, we found evidence at specific loci for the leading CAD index SNPs, where the allele that significantly increased LRS also significantly increased CAD. We further found no evidence that this link between CAD and LRS was due to confounding (e.g. physiological, hormonal) effects of LRS on CAD risk. While this is promising, the Framingham study is limited in its power to detect small fitness and CAD effects; better powered studies may yet be needed to definitively establish antagonistic effects at all loci. Fitness traits collected on genotyped populations are currently rare, but this is likely to change as more biobank-scale studies come online.

To facilitate interpretation of selection occurring on early-life traits or CAD phenotypic risk factors that share pleiotropic connections and possible evolutionary tradeoffs with coronary artery disease, we present a conceptual figure (Fig 5). These pleiotropic effects are important because many of them affect traits expressed early in life, some extremely early in life. Any allele that increases reproductive performance enough early in life to more than compensate for a loss of associated fitness late in life will be selected [42]. Such a mechanism has been recently suggested to help explain the maintenance of polymorphic disease alleles in modern human populations [94]. While such tradeoffs have been previously tested for in humans using genotypes, LRS and lifespan (e.g. [95]), there is not yet much evidence that such a mechanism influences human disease. A 2017 study by Rodríguez et al. [96] that used an indirect measure of fitness

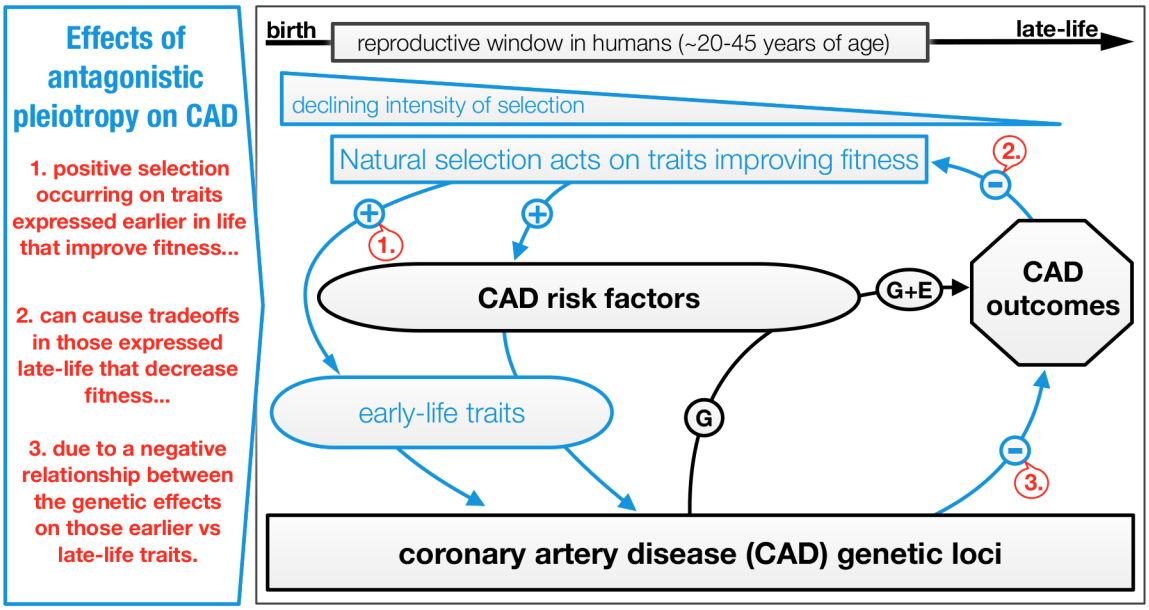

Fig 5. Conceptual figure of potential evolutionary tradeoffs between coronary artery disease (CAD) burden and other phenotypes as a consequence of antagonistic pleiotropy (AP) [42]. As a simple example, AP describes gene effect on two traits (pleiotropy) that oppositely (antagonistic) affect individual fitness at different ages. Selection on that gene conferring a fitness advantage and disadvantage at different ages depends on the size and timing of the effects. An advantage during the ages with the highest probability of reproduction (between~20–45 years of age in humans) would increase fitness (lifetime reproductive success) more than a similarly sized disadvantage at later ages would decrease it. This concept is part of the well-known evolutionary theory of ageing, which describes tradeoffs in energy invested into growth, reproduction and survival [97]. In the figure above, intense natural selection occurring on CAD loci as a result of fitness advantages (+ signs, red text callout box 1.) conferred by genetically correlated risk factors ('CAD risk factors' box) or early-life traits ('early-life traits' box) trades off with the deleterious effects of these genes on fitness (i.e. CAD burden) later in life (- sign, red text callout box 2.) where the intensity of selection is weak. This occurs because of the negative relationship between genetic effects on early vs late-life traits (- sign, red text callout box 3.), which could help explain the high prevalence and maintenance of CAD in modern human populations. Over shorter timescales, lifetime probability of CAD is modified by a combination of genetic and environmental risk factors (e.g. [98]). There is evidence that such antagonistic effects have operated on CAD loci given: significant associations between CAD genetic risk and selection found (Figs 1 and 2); CAD genes are significantly enriched for lifetime reproductive success (S2 Table) and may also effect other early-life traits known to modify fitness (S4 Table); suggestive evidence was found for an antagonistic relationship between CAD and LRS (S3 Table); phenotypic selection has been found operating on CAD phenotypic risk factors [41].

https://doi.org/10.1371/journal.pgen.1006328.g005

provides support for antagonistic pleiotropy acting on general early health and later life disease. Our study demonstrates that CAD genes are significantly and directly enriched for fitness with evidence that some of the leading CAD effect SNPs share an antagonistic relationship with fitness through significant positive effects on LRS. This provides support for such a mechanism influencing CAD and may help to explain our vulnerability to this disease.

Study limitations

There are also some limitations to our approach. We utilized CAD genetic risk estimated from a meta-analysis based on predominantly European (77%) with smaller contributions from south/east Asian (19%), Hispanic and African American (~4%) ancestry [40]. Genetic risk variation for CAD might be different in the un-represented (i.e. Mexican) or less-represented (i.e. African) populations in this meta-analysis. If that were the case, it would reduce the usefulness of comparing selection and risk estimates in those populations. We also saw fewer significant selection-risk associations in the African populations (Fig 1B), however this may be due to selection signals in the African populations being less obvious than those in East Asian and

European populations, perhaps due to lesser linkage disequilibrium, as is consistent with results from previous studies [99]. Calculating disease risk and selection variation from populations within the same ancestral group might help resolve this, however it only represents a potential shortcoming for our cross-population analyses and not observations of antagonistic pleiotropy.

Summary

In this study, we found evidence that natural selection has recently operated on CAD associated variation. By comparing positive selection variation with genetic risk variation at known loci underlying CAD, we were able to identify and prioritize genes that have been the most likely targets of selection related to this disease across diverse human populations. That selection signals and the direction of selection-risk relationships varied among some populations suggests that CAD-driven selection has operated differently in these populations and thus that these populations might respond differently to similar heart disease prevention strategies. The pleiotropic effects that genes associated with CAD have on traits associated with reproduction that are expressed early in life strongly suggests some of the evolutionary reasons for the existence of human vulnerability to CAD.

Methods

Defining loci linked to coronary artery disease

We started with the 56 lead index SNPs from Supplementary Table 5 in Nikpay et al. [40] corresponding to 56 CAD loci. When the index SNP was genic, all SNPs within that gene were extracted (using NCBI's dbSNP) including directly adjacent intergenic SNPs ±5000bp from untranslated regions (UTR) in LD r^2>0.7 (with any respective genic SNP). When the index SNP was intergenic, that SNP and other directly adjacent SNPs ±5000bp and in LD>0.7 (with the index SNP) were extracted and combined with SNPs from the respective linked gene listed in Nikpay including SNPs ±5000bp from UTR regions in LD r^2>0.7 with that gene. This resulted in SNP lists for 56 genes. To further explore other genes not directly connected with lead index SNPs, but that were within the CAD loci identified by the two most recent CARDIoGRAMplusC4D studies–including Deloukas et al. [39] (i.e. 46 loci and 61 genes listed in their Tables 1–2) and Nikpay et al. [40] (i.e. 10 loci and 15 genes listed in their Table 1)—we extracted SNPs within each of those genes (plus SNPs ±5000bp from UTR regions in LD r^2>0.7 with that gene). This resulted in SNP lists for a further 20 genes, bringing the total number of candidate genes for CAD to 76.

The per-SNP log odds (ln(OR)) values for the 76 genes were obtained for the additive model from Nikpay et al. [40] available at http://www.cardiogramplusc4d.org/downloads and used in the analysis described below.

Preparation of HapMap3 samples

Genotype data (1,457,897 SNPs, 1,478 individuals) were downloaded for 11 HapMap Phase 3 (release 3) populations (http://www.hapmap.org [100]) including: Yoruba from Ibadan, Nigeria (YRI), Maasai in Kinyawa, Kenya (MKK), Luhya in Webuye, Kenya (LWK), African ancestry in Southwest USA (ASW), Utah residents with ancestry from northern and western Europe from the CEPH collection (CEU), Tuscans in Italy (TSI), Japanese from Tokyo (JPT), Han Chinese from Beijing (CHB), Chinese in Metropolitan Denver, Colorado (CHD), Gujarati Indians in Houston, TX, USA (GIH), and Mexican ancestry in Los Angeles, CA, USA (MEX). We also included another HapMap3 population, the Finnish in Finland (FIN) sample

(ftp://ftp.fimm.fi/pub/FIN_HAPMAP3 [101]). These data had already been pre-filtered, i.e. SNPs were excluded that were monomorphic, call rate < 95%, MAF<0.01, Hardy-Weinberg equilibrium $p<1x10^{-6}$.

Before phasing and imputation, we performed a divergent ancestry check with flashpca [102] to check accuracy of population assignments, converted SNP data from build 36 to 37 with UCSC LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver), checked strand alignment in Plink v1.9 [103] to ensure all genotypes were reported on the forward strand, and kept only autosomal SNPs. To speed up imputation, data were first pre-phased with Shapeit v2 [104] using the duoHMM option that combines pedigree information to improve phasing and default values for window size (2Mb), per-SNP conditioning sates (100), effective population size (n = 15000) and genetic maps from the 1000 Genomes Phase 3 b37 reference panel (ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

Phased data were imputed in 5 Mb chunks across each chromosome with Impute v2 [105]. We then removed any multiallelic SNPs (insertions, deletions etc) from the imputed data and excluded SNPs with call rate < 95%, HWE $p<1x10^{-6}$ and MAF<1%. The final dataset was then phased with Shapeit v2, and alleles were converted to ancestral and derived states using python script. Ancestral allele states came from 1000 Genomes Project FASTA files and derived 6-primate (human, gorilla, orangutan, chimp, macaque, marmoset) Enredo-Pecan-Ortheus alignment [106] from the Ensembl Compara 59 database [107].

Estimating signatures of recent selection

Integrated Haplotype Score (iHS): Using the package rehh [108] in R version 3.1.3, per SNP iHS scores were calculated within each population (after excluding non-founders) using methods described previously [9]. iHS could not be calculated for SNPs without an ancestral state, or whose population minor allele frequency is <5%, or for some SNPs that are close to chromosome ends or large regions without SNPs [9]. Rehh was also used to standardize (mean 0, variance 1) iHS values empirically to the distribution of available genome-wide SNPs with similar derived allele frequencies. For analyses in the main text, we considered a SNP to have a candidate selection signal if it had an absolute iHS score > 2, a permuted p value <0.05, and was within a 'cluster' of SNPs that also had elevated iHS scores. Although permuting p values is computationally more intensive, it provides more flexibility to detect smaller selection signals that may be incorrectly classified with the more stringent Bonferroni correction that is often applied to these estimates. For the analyses described below, even though we only used iHS estimates for the SNPs defined in the CAD genes (and additional SNPs for permutation purposes), we calculated per-SNP iHS scores genome-wide (rather than locally, i.e. within 1MB regions around focal SNPs), for this provides more accurate estimates because final adjustments are made relative to other genome-wide SNPs of similar sized derived allele frequency classes. P values for iHS scores were permuted based on comparison of nominal p values against 10000 randomly selected estimates from within the same derived allele frequency classes.

Comparing CAD genetic risk and quantitative selection signals

We first tested the null hypothesis that there is no association between CAD genetic risk and signals of positive selection for CAD genes. For each gene within each population, we used a mixed effects linear model to regress SNP-based estimates of CAD log odds (ln(OR)) genetic risk against selection scores (iHS) resulting in 912 separate regressions. To account for LD structure (and potential confounding of highly correlated SNPs) within each gene, we also included the first eigenvector derived from an LD matrix of correlations (r^2) between SNPs

within each gene as a random effect. We chose to model LD structure with mixed-effects models rather than LD-prune because for many genes, the SNP samples would have been too small for regression analyses. Also, it would be very difficult to properly capture both selection and the CAD log odds peaks needed to compare these variables. We did however investigate alternative models to validate our approach (i.e. running the same models without the LD structure variable; using smaller multiple LD-pruned subsets of SNPs per gene) with consistent results suggesting our approach was largely robust to LD effects and likelihood of false positive associations. We accounted for multiple testing by permuting p values for each regression based on comparing each nominal p value against 10000 permuted p values derived from shuffling iHS scores.

Genes were then ranked based on the number of significant associations summed across the 12 populations. The 40 genes with at least four or more significant associations are shown in Fig 1B. To illustrate the positional architecture of these selection-risk associations, plots for selected highly-ranked genes are shown in Figs 1 and 2. By demonstrating how CAD genetic risk peaks and valleys correspond to variation in the magnitude of selection scores (iHS), this allowed visual assessment of potential modifications made to the phenotype-genotype map by selective pressures imposed directly or indirectly by CAD. It also helped us localize selection peaks within genes and compare them between populations. Similar peaks suggested similar selection and different peaks suggested local adaptation. This way of presenting the results also allowed us to detect the smaller adaptive shifts in allele frequencies typically expected to underlie selection on polygenic traits.

We then tested a second null hypothesis: that the selection-risk associations using the CAD genes are not unique compared to non-CAD associated loci. For each of the 76 CAD genes, we randomly (without replacement) chose 100 genes of similar length across the genome and performed the same mixed effects regression procedure described above for each gene by population combination using both CAD log odds values from Nikpay et al. [40], iHS scores estimated from the SNP data, and the first LD eigenvector from SNPs within a gene. Permuted p values were derived by comparing the nominal p value for each CAD gene against the 100 null distribution p values from the non-CAD associated genes. Results are shown in Fig 1C.

Identifying functional targets of selection

To examine whether candidate adaptive signals within each gene corresponded to a gene's regulatory variation, we regressed SNPs within focal genes and gender against that gene's probe expression levels, which had previously been quantified in lymphoblastoid cell lines from circulating peripheral blood using Illumina's Human-6 v2 Expression BeadChip for eight of the 12 populations [109]. Given gene expression in peripheral blood is known to be an important marker for cardiovascular disease, we therefore might expect this cell type a good candidate to search for association between selection signals and regulatory variants important for these genes. The raw gene microarray expression data had previously been normalized on a log2 scale using quantile normalization for replicates of a single individual then median normalization for each population [109]. P values for each SNP-probe association were permuted using 10000 permutations by randomly shuffling gene probes expression. P values were then extracted for the most significant iHS score for each gene-population combination and compared to the same number of p values randomly drawn from different LD blocks underlying SNPs with non-significant iHS scores across each gene-population combination. A Kolmogorov-Smirnov test was used to compare the distribution of p values from each. To examine what biological processes were associated with the top ranked genes from Fig 1, we uploaded the top 10 genes into Enrichr (http://amp.pharm.mssm.edu/Enrichr/) to define associated

pathways (i.e. KEGG 2016, kegg.jp/kegg), ontologies (MGI Mammalian phenotypes, informatics.jax.org), cell types (Cancer cell line Encyclopedia, broadinstitute.org/ccle) and transcription factors (ChEA 2015, amp.pharm.mssm.edu/lib/chea.jsp).

Testing for fitness effects of CAD loci

We tested whether CAD SNPs were directly associated with human fitness. For a trait to evolve, this is one of the main prerequisites, but it also helps demonstrate whether alleles that influence disease also influence reproduction, which in the case of CAD suggests there may be antagonistic trade-offs between early versus late life.

We used the Framingham Heart Study dataset because it has completed reproductive outcomes (lifetime reproductive success (LRS) or number of children ever born), genotypes, pedigree data, cardiovascular outcomes and demographic and socioeconomic data. LRS was derived from clinical questionnaires and further validated with pedigree data. We did not include other datasets for validation here, as it is extremely hard to find others that include all these variables. There were 1,579 women from the Original and Offspring cohorts who had genotypes and all phenotypes available also after excluding non-founders. FHS 500k Affymetrix genotypes were 1000-Genomes imputed using the same pipeline described above bringing the total number of SNPs available (at MAF>1%) to 7,486,901. LRS was adjusted to deal with secular demographic change where data was broken into six groups based on year women were born and divided by the mean reproductive success of women in that group (same as described in [41]).

We examined the association of all SNPs available in the FHS for the 76 CAD genes (20,254 SNPs) with LRS using linear mixed models implemented in FaST-LMM [110] that account for potential confounding effects of genetic similarity by including a *k*-spectral decomposition variable derived from the realized relationship matrix (RRM) of an LD-pruned subset of SNPs. SNPs used for RRM were not in LD with CAD SNPs to avoid proximal contamination. Factors that may affect LRS—education, smoking status, whether the person was born in the US, and estrogen usage (hormone therapy or contraceptive use)—were included as covariates. Permutations with 10,000 iterations were also run for each SNP in order to validate nominal p values obtained directly from FaST-LMM, where permuted p values were based on the number of times nominal p values for 10,000 randomly chosen SNPs (within a similar MAF bin) were greater than the target SNP nominal p value. Bonferroni and FDR adjustment was also applied to p values based on 20,254 tests.

To test the null hypothesis that CAD SNPs are collectively no more enriched for fitness compared to non-CAD SNPs, we randomly sampled without replacement 20,254 non-CAD SNPs (matched within MAF bins to the CAD-SNP sample) 100 times. The permuted p value was based on the number of times (out of 100) that the number of significant p values in the random sample exceeded that for the CAD SNP sample. We also compared the distribution of p values between all randomly chosen SNPs (2,025,400) from this analysis to the 20,254 CAD SNPs with a Kolmogorov-Smirnov test. One- and two-sided tests were run. The one-sided test specifically tested whether the distribution of p values for CAD SNPs was stochastically larger compared to non-CAD SNPs.

We then used fastBAT [111] to test whether CAD is enriched for fitness at the gene-level. fastBAT combines SNP-based summary-level data from GWAS and LD reference data to give locus-based estimates of association. We also ran 100 permutations for each of the 76 genes in order to estimate a permuted p value for each locus. Permuted p values were based on the number of times p values for the 100 randomly chosen similarly-sized genes were greater than that for each CAD gene. We also tested whether CAD genes were collectively more enriched

for fitness at the gene-level, relative to non-CAD genes. We randomly chose 76 non-CAD genes of similar size with 300 permutations (sampling without replacement) and asked how many times the number of p values for the non-CAD genes exceeded that of the CAD genes.

Other traits that may influence fitness were also tested (using the same analysis/permutation tests as above) for FHS women including age at first and last birth, interbirth interval, menarche and menopause. Menarche and menopause were derived from questionnaires, while birth timing/spacing were estimated from pedigrees. We did not consider reproductive outcomes for men as that data was only available from pedigrees, which is less reliable than clinical records. Age at first and last birth and interbirth interval were also adjusted for secular demographic changes (same as above).

Alternative simplified models for LRS, AFB and ALB were run in FaST-LMM where fitness measures were not adjusted for temporal effects and no covariates were included. This boosted sample sizes (S2 Table) due to avoiding missing values associated with covariates, however results were largely comparable (S2 Table) suggesting no power gain from unadjusted models.

Testing for antagonistic effects between fitness and CAD

We only tested for antagonistic effects for LRS as that is the most direct measure of fitness and was the only fitness trait where CAD SNPs were significantly and consistently enriched across (un)adjusted models (S2 Table). To assess whether antagonistic effects were present between LRS and CAD, genome-wide significant CAD index SNPs were taken from Nikpay et al. [40] and cross-referenced with significant LRS SNPs from the FaST-LMM regression results. In an extended analysis (results not shown), we also included any SNP in high-LD ($r^2 >= 0.8$) and proximal (±1MB) to the index SNP to boost the number of SNPs available for comparison: results were virtually identical for the significance of LRS and CAD effects and the consistency of antagonistic effects. An antagonistic effect was defined as an allele that significantly increased LRS and significantly increased CAD risk. We also tested whether CAD SNPs were associated with both LRS and CAD due to other confounding (rather than pleiotropic) effects (see S3 Fig for methods and findings).

Supporting information

S1 Fig. Association of coronary artery disease (CAD) risk and genomic signatures of selection in 12 worldwide populations. All 76 genes are shown ranked according to Fig 1B. Boxes show magnitude and significance of largest positive selection signal (integrated haplotype score, iHS) within each gene-population combination. P values (circles within squares) were obtained from 10000 permutations. Bonferroni corrected p value limit also shown ($\alpha = 0.05/76 = 0.000657$) with closed circles. **Populations**. Grouped by common ancestry, African (ASW, African ancestry in Southwest USA; MKK, Maasai in Kinyawa, Kenya; YRI, Yoruba from Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya), East- Asian (CHB, Han Chinese subjects from Beijing; CHD, Chinese in Metropolitan Denver, Colorado; JPT, Japanese subjects from Tokyo), European (CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; TSI, Tuscans in Italy; FIN, Finnish in Finland), GIH (Gujarati Indians in Houston, TX, USA), MEX (Mexican ancestry in Los Angeles, CA, USA).
(PDF)

S2 Fig. Comparing cross-population candidate selection signals in *PHACTR1*. Per-SNP integrated Haplotype Scores (iHS) plotted by chromosome position within *PHACTR1* (including LD plots below each) for 12 worldwide populations. Permuted p value significance for each score coded by color (grey, non-significant; orange, $p < 0.05$). Red dashed line indicates

position of index SNP for *PHACTR1*. Grey columns in background represent intron spans. Populations are clustered by common ancestry, African (ASW, African ancestry in Southwest USA; MKK, Maasai in Kinyawa, Kenya; YRI, Yoruba from Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya), East-Asian (CHB, Han Chinese subjects from Beijing; CHD, Chinese in Metropolitan Denver, Colorado; JPT, Japanese subjects from Tokyo), European (CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; TSI, Tuscans in Italy; FIN, Finnish in Finland), GIH (Gujarati Indians in Houston, TX, USA), MEX (Mexican ancestry in Los Angeles, CA, USA).
(PDF)

S3 Fig. Testing whether CAD SNPs are associated with both LRS and CAD due to pleiotropy or confounding effects. Confounding effects would occur if coronary artery disease (CAD) SNPs modestly affected lifetime reproductive success (LRS), which in turn caused significant changes in CAD risk due to physiological, hormonal or social changes related to childbearing/rearing [1, 2]. If this was the case, we would expect that the effect of CAD SNPs on CAD should diminish when adjusting for LRS; in the case of pleiotropy, it would not. Grey dots represent regression coefficients (β) for index SNPs on CAD outcomes with (model 2) or without (model 1) stratifying for LRS. β's are exponeniated coefficients from Cox proportional hazard models that were also adjusted for other potentially confounding effects on CAD (see methods below).
(PDF)

S1 Table. Selected Enrichr analysis outputs for top 10-ranked CAD genes with highest genetic risk-selection associations from Fig 1B. Enrichr outputs includes KEGG 2016 Pathways (http://www.kegg.jp/kegg/download/), MGI Mammalian Phenotype Level 3 (http://www.informatics.jax.org/), Cancer Cell Line Encyclopaedia (http://portals.broadinstitute.org/ccle/data/browseData), and ChEA 2015 (http://amp.pharm.mssm.edu/lib/cheadownload.jsp).
(PDF)

S2 Table. Testing association of CAD SNPs with human fitness in the Framingham Heart Study women. First three columns give number of individuals available and used in analyses. Four FaST-LMM columns provide summary of leading results including leading and highest ranked SNP(s) (and associated genes). Three fastBAT columns provide leading gene(s). Final four columns provide statistics for testing the Null hypothesis that CAD variation is no more enriched for fitness compared to non-CAD variation found genome-wide.
(PDF)

S3 Table. Testing for antagonistic pleiotropy for SNPs with significant effects on lifetime reproductive success (LRS) and coronary artery disease (CAD). *Left table*: provides statistics for CAD index SNPs derived directly from the CARDIoGRAMplusC4D 1000 Genomes-based GWAS meta-analysis (see [1] or http://www.cardiogramplusc4d.org/data-downloads/ for further details of data and variables). *Right table*: provides corresponding FaST-LMM regression statistics of these SNPs on LRS based on Framingham Heart Study women (first six columns), rows correspond to SNPs in the left table. Last four columns test for antagonistic effects between CAD and LRS with the last two providing these tests only when the LRS beta was significant. This shows that when SNPs with significant effects on both LRS and CAD are considered, most (5 of 6, or 83%) were antagonistic, i.e. the allele that increases LRS also increases CAD risk.
(PDF)

S4 Table. Pleiotropic links between coronary artery disease (CAD) and early- life fitness-related traits due to shared genetic loci. The table below provides extensive support (143 studies) that antagonistic pleiotropy is likely to be present for CAD genes due to their consistent connections with fitness-related traits expressed early in life. See Fig 5 for discussion and conceptual overview of these potential effects. Fitness-related traits include fertility potential, reproductive outcomes, pregnancy outcomes, fetal growth and survival, i.e. affecting the ability of an organism to reproduce and transfer genes to the next generation. The first 3 columns give CAD gene rank (no.; based on rank of 40 genes from Fig 1B), name and full name. Columns 4–8 provide key details of each study where CAD genes also contribute to traits that influence fitness, including what species that was demonstrated in, what biological process or fitness effects that gene is impacting, what fitness class that effect is likely to impact (e.g. dysfunctional spermatogenesis or embryogenesis will affect male and female fertility, ability to conceive), what the observed genetic effect or mechanism that gene was associated with.
(PDF)

S5 Table. Summary of types of pleiotropic connections between coronary artery disease (CAD) and fitness-related traits. Counts are based on S1 Table, 'fitness class' column. Most fitness-related traits were related to female potential fertility (29 of 40 genes had these effects) and pregnancy outcomes (25 of 40 genes had these effects). Some genes had broad or specific effects on fitness-related traits. For example, number of fitness classes affected ranged from 6 for *ABO* (had fitness effects across all classes) to 1, for example *CNNM2* (evidence for fitness effects in pregnancy outcomes class).
(PDF)

S6 Table. Pleiotropic links between randomly chosen genes and early-life fitness-related traits. Fitness-related traits include fertility potential, reproductive outcomes, pregnancy outcomes, fetal growth and survival, i.e. affecting the ability of an organism to reproduce and transfer genes to the next generation. The first column gives coronary artery disease (CAD) gene (first 20 of 40 CAD genes from Fig 1B/S1 Table). Columns 2–3 give name (abbreviated, full) of randomly chosen genes matched for approximate length for each CAD gene. Columns 4–8 provide key details of each study where random genes also contribute to traits that influence fitness, including what species that was demonstrated in, what biological process or fitness effects that gene is impacting, what fitness class that effect is likely to impact (e.g. dysfunctional spermatogenesis or embryogenesis will affect male and female fertility, ability to conceive), what the observed genetic effect or mechanism that gene was associated with.
(PDF)

S1 Discussion. Widespread candidate signals of positive selection on CAD loci. Extended discussion on candidate adaptive signals found on coronary artery disease (CAD) loci in relation to the polygenic model of selection and previous studies examining genomic selection on broader cardiovascular disease loci.
(PDF)

Acknowledgments

We are grateful to the CARDIoGRAMplusC4D consortium for making their large-scale genetic data available. A list of members of the consortium and the contributing studies is available at www.cardiogramplusc4d.org. We further thank reviewers for their valuable input on earlier drafts of this manuscript.

Author Contributions

Conceptualization: SGB MI.

Formal analysis: SGB QQH AB.

Funding acquisition: MI.

Methodology: SGB MI.

Supervision: MI.

Visualization: SGB.

Writing – original draft: SGB MI.

Writing – review & editing: SGB QQH LG AB SR GA SCS MI.

References

1. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome research. 2002; 12(12):1805–14. https://doi.org/10.1101/gr.631202 PMID: 12466284
2. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005; 437(7062):1153–7. https://doi.org/10.1038/nature04240 PMID: 16237444
3. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, et al. Genomic regions exhibiting positive selection identified from dense genotype data. Genome research. 2005; 15 (11):1553–65. https://doi.org/10.1101/gr.4326505 PMID: 16251465
4. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome research. 2006; 16(8):980–9. https://doi.org/10.1101/gr.5157306 PMID: 16825663
5. Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet. 2007; 71:354–69. https://doi.org/10.1111/j.1469-1809.2006.00341.x PMID: 17233754
6. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419(6909):832–7. https://doi.org/10.1038/nature01140 PMID: 12397357
7. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Human genomics. 2004; 1(4):274–86. https://doi.org/10.1186/1479-7364-1-4-274 PMID: 15588487
8. Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS biology. 2007; 5(7):1587–602.
9. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology. 2006; 4(3):e72. https://doi.org/10.1371/journal.pbio.0040072 PMID: 16494531
10. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying Recent Adaptations in Large-Scale Genomic Data. Cell. 2013; 152(4):703–13. https://doi.org/10.1016/j.cell.2013.01.035 PMID: 23415221
11. Haasl RJ, Payseur BA. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Molecular ecology. 2015.
12. Scheinfeldt LB, Tishkoff SA. Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genet. 2013; 14(10):692–702. https://doi.org/10.1038/nrg3604 PMID: 24052086
13. Pritchard JK, Di Rienzo A. Adaptation—not by sweeps alone. Nat Rev Genet. 2010; 11(10):665–7. https://doi.org/10.1038/nrg2880 PMID: 20838407
14. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current biology: CB. 2010; 20(4):R208–15. https://doi.org/10.1016/j.cub.2009.11.055 PMID: 20178769

15. Fu W, Akey JM. Selection and adaptation in the human genome. Annual review of genomics and human genetics. 2013; 14:467–89. https://doi.org/10.1146/annurev-genom-091212-153509 PMID: 23834317

16. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic Selective Sweeps Were Rare in Recent Human Evolution. Science. 2011; 331(6019):920–4. https://doi.org/10.1126/science.1198878 PMID: 21330547

17. Kaplan NL, Hudson RR, Langley CH. The Hitchhiking Effect Revisited. Genetics. 1989; 123(4):887–99. PMID: 2612899

18. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 2007; 89(5–6):391–403. https://doi.org/10.1017/S0016672308009579 PMID: 18976527

19. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2010; 365(1537):185–205. https://doi.org/10.1098/rstb.2009.0219 PMID: 20008396

20. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006; 312(5780):1614–20. https://doi.org/10.1126/science.1124309 PMID: 16778047

21. Wright S. Genetical structure of populations. Nature. 1950; 166(4215):247–9. PMID: 15439261

22. Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet. 2002; 70(2):369–83. https://doi.org/10.1086/338628 PMID: 11753822

23. Beall CM, Cavalleri GL, Deng LB, Elston RC, Gao Y, Knight J, et al. Natural selection on EPAS1 (HIF2 alpha) associated with low hemoglobin concentration in Tibetan highlanders. P Natl Acad Sci USA. 2010; 107(25):11459–64.

24. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science. 2005; 310(5755):1782–6. https://doi.org/10.1126/science.1116238 PMID: 16357253

25. Fraser HB. Gene expression drives local adaptation in humans. Genome research. 2013; 23(7):1089–96. https://doi.org/10.1101/gr.152710.112 PMID: 23539138

26. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome research. 2009; 19(5):711–22. https://doi.org/10.1101/gr.086652.108 PMID: 19411596

27. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007; 449(7164):913–U12. https://doi.org/10.1038/nature06250 PMID: 17943131

28. Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome research. 2006; 16(6):702–12. https://doi.org/10.1101/gr.5105206 PMID: 16687733

29. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Harlow, England; New York: Prentice Hall; 1996. xv, 464 p. p.

30. Grant PR, Grant BR. Predicting Microevolutionary Responses to Directional Selection on Heritable Variation. Evolution. 1995; 49(2):241–51.

31. Hermisson J, Pennings PS. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics. 2005; 169(4):2335–52. https://doi.org/10.1534/genetics.104.036947 PMID: 15716498

32. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013; 28(11):659–69. https://doi.org/10.1016/j.tree.2013.08.003 PMID: 24075201

33. Chevin LM, Hospital F. Selective Sweep at a Quantitative Trait Locus in the Presence of Background Genetic Variation. Genetics. 2008; 180(3):1645–60. https://doi.org/10.1534/genetics.108.093351 PMID: 18832353

34. Ding KY, Kullo IJ. Geographic differences in allele frequencies of susceptibility SNPs for cardiovascular disease. Bmc Med Genet. 2011; 12.

35. Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. Common Risk Alleles for Inflammatory Diseases Are Targets of Recent Positive Selection. Am J Hum Genet. 2013; 92(4):517–29. https://doi.org/10.1016/j.ajhg.2013.03.001 PMID: 23522783

36. Casto AM, Feldman MW. Genome-Wide Association Study SNPs in the Human Genome Diversity Project Populations: Does Selection Affect Unlinked SNPs with Shared Trait Associations? Plos Genet. 2011; 7(1).

37. Turchin MC, Chiang CWK, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet. 2012; 44(9):1015–+. https://doi.org/10.1038/ng.2368 PMID: 22902787

38. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Blaha MJ, et al. Heart disease and stroke statistics—2014 update: a report from the American Heart Association. Circulation. 2014; 129(3):e28–e292. https://doi.org/10.1161/01.cir.0000441139.02102.80 PMID: 24352519

39. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013; 45(1):25–U52. https://doi.org/10.1038/ng.2480 PMID: 23202125

40. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015; 47 (10):1121–+. https://doi.org/10.1038/ng.3396 PMID: 26343387

41. Byars SG, Ewbank D, Govindaraju DR, Stearns SC. Colloquium papers: Natural selection in a contemporary human population. Proc Natl Acad Sci U S A. 2010; 107 Suppl 1:1787–92.

42. Williams GC. Pleiotropy, Natural Selection, and the Evolution of Senescence. Evolution. 1957; 11 (4):398–411.

43. Jalowiec DA, Hill JA. Myocardial infarction in the young and in women. Cardiovascular clinics. 1989; 20(1):197–206. PMID: 2653633

44. Rubin JB, Borden WB. Coronary Heart Disease in Young Adults. Curr Atheroscler Rep. 2012; 14 (2):140–9. https://doi.org/10.1007/s11883-012-0226-3 PMID: 22249950

45. Tuzcu EM, Kapadia SR, Tutar E, Ziada KM, Hobbs RE, McCarthy PM, et al. High prevalence of coronary atherosclerosis in asymptomatic teenagers and young adults: evidence from intravascular ultrasound. Circulation. 2001; 103(22):2705–10. PMID: 11390341

46. Morillas P, Bertomeu V, Pabon P, Ancillo P, Bermejo J, Fernandez C, et al. Characteristics and outcome of acute myocardial infarction in young patients—The PRIAMHO II study. Cardiology. 2007; 107 (4):217–25. https://doi.org/10.1159/000095421 PMID: 16953107

47. Allam AH, Thompson RC, Wann LS, Miyamoto MI, Nur El-Din Ael H, El-Maksoud GA, et al. Atherosclerosis in ancient Egyptian mummies: the Horus study. JACC Cardiovascular imaging. 2011; 4 (4):315–27. https://doi.org/10.1016/j.jcmg.2011.02.002 PMID: 21466986

48. Wollstein A, Stephan W. Inferring positive selection in humans from genomic data. Investigative genetics. 2015; 6:5. https://doi.org/10.1186/s13323-015-0023-1 PMID: 25834723

49. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016; 44(W1):W90–7. https://doi.org/10.1093/nar/gkw377 PMID: 27141961

50. Kosova G, Scott NM, Niederberger C, Prins GS, Ober C. Genome-wide Association Study Identifies Candidate Genes for Male Fertility Traits in Humans. Am J Hum Genet. 2012; 90(6):950–61. https://doi.org/10.1016/j.ajhg.2012.04.016 PMID: 22633400

51. Aschebrook-Kilfoy B, Argos M, Pierce BL, Tong L, Jasmine F, Roy S, et al. Genome-Wide Association Study of Parity in Bangladeshi Women. PloS one. 2015; 10(3).

52. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics & informatics. 2012; 10(2):117–22.

53. Pokharel K, Peippo J, Andersson G, Li M, Kantanen J. Transcriptome profiling of Finnsheep ovaries during out-of-season breeding period. Agricultural and Food Science. 2015; 24:1–9.

54. Mbarek H, Steinberg S, Nyholt DR, Gordon SD, Miller MB, McRae AF, et al. Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. Am J Hum Genet. 2016; 98(5):898–908. https://doi.org/10.1016/j.ajhg.2016.03.008 PMID: 27132594

55. Huang C, Lin Y, Ding S, Lo L, Wang P, Lin E, et al. Efficient SNP Discovery by Combining Microarray and Lab-on-a-Chip Data for Animal Breeding and Selection. Microarrays. 2015; 4(4):570–95. https://doi.org/10.3390/microarrays4040570 PMID: 27600241

56. Mote BE, Koehler KJ, Mabry JW, Stalder KJ, Rothschild MF. Identification of genetic markers for productive life in commercial sows. J Anim Sci. 2009; 87(7):2187–95. https://doi.org/10.2527/jas.2008-1017 PMID: 19359509

57. Balgir RS. Menarcheal age in relation to ABO blood group phenotypes and haemoglobin-E genotypes. J Assoc Physicians India. 1993; 41(4):210–1. PMID: 8270560

58. Pyun JA, Kim S, Cho NH, Koh I, Lee JY, Shin C, et al. Genome-wide association studies and epistasis analyses of candidate genes related to age at menarche and age at natural menopause in a Korean population. Menopause. 2014; 21(5):522–9. https://doi.org/10.1097/GME.0b013e3182a433f7 PMID: 24045676

59. Spencer KL, Malinowski J, Carty CL, Franceschini N, Fernandez-Rhodes L, Young A, et al. Genetic Variation and Reproductive Timing: African American Women from the Population Architecture Using Genomics and Epidemiology (PAGE) Study. PloS one. 2013; 8(2).

60. Rempel LA, Nonneman DJ, Wise TH, Erkens T, Peelman LJ, Rohrer GA. Association analyses of candidate single nucleotide polymorphisms on reproductive traits in swine. J Anim Sci. 2010; 88(1):1–15. https://doi.org/10.2527/jas.2009-1985 PMID: 19749016

61. Patel OV, Casey T, Dover H, Plaut K. Homeorhetic adaptation to lactation: comparative transcriptome analysis of mammary, liver, and adipose tissue during the transition from pregnancy to lactation in rats. Funct Integr Genomic. 2011; 11(1):193–202.

62. Wang M, Moisa S, Khan MJ, Wang J, Bu D, Loor JJ. MicroRNA expression patterns in the bovine mammary gland are affected by stage of lactation. J Dairy Sci. 2012; 95(11):6529–35. https://doi.org/10.3168/jds.2012-5748 PMID: 22959945

63. Colodro-Conde L, Zhu G, Power RA, Henders A, Heath AC, Madden PA, et al. A twin study of breast-feeding with a preliminary genome-wide association scan. Twin Res Hum Genet. 2015; 18(1):61–72. https://doi.org/10.1017/thg.2014.74 PMID: 25475840

64. McLean MP, Zhao Z, Ness GC. Reduced hepatic LDL-receptor, 3-hydroxy-3-methylglutaryl coenzyme A reductase and sterol carrier protein-2 expression is associated with pregnancy loss in the diabetic rat. Endocrine. 1995; 3(10):695–703. https://doi.org/10.1007/BF03000200 PMID: 21153157

65. Kuo DS, Labelle-Dumais C, Gould DB. COL4A1 and COL4A2 mutations and disease: insights into pathogenic mechanisms and potential therapeutic targets. Hum Mol Genet. 2012; 21(R1):R97–110. https://doi.org/10.1093/hmg/dds346 PMID: 22914737

66. Lin S, Leonard D, Co MA, Mukhopadhyay D, Giri B, Perger L, et al. Pre-eclampsia has an adverse impact on maternal and fetal health. Transl Res. 2015; 165(4):449–63. https://doi.org/10.1016/j.trsl.2014.10.006 PMID: 25468481

67. Sayed AAA. Molecular genetic studies in pregnancies affected by preeclampsia and intrauterine growth restriction: University of Nottingham; 2011.

68. Fritz RB. Trophoblast Retrieval And Isolation From e Cervix (tric) For Non-Invasive Prenatal Genetic Diagnosis And Prediction Of Abnormal Pregnancy Outcome: Wayne State University; 2015.

69. Tabano S, Alvino G, Antonazzo P, Grati FR, Miozzo M, Cetin I. Placental LPL gene expression is increased in severe intrauterine growth-restricted pregnancies. Pediatr Res. 2006; 59(2):250–3. https://doi.org/10.1203/01.pdr.0000199441.62045.a1 PMID: 16439587

70. Bhasin KK, van Nas A, Martin LJ, Davis RC, Devaskar SU, Lusis AJ. Maternal low-protein diet or hypercholesterolemia reduces circulating essential amino acids and leads to intrauterine growth restriction. Diabetes. 2009; 58(3):559–66. https://doi.org/10.2337/db07-1530 PMID: 19073773

71. Kakourou G, Jaroudi S, Tulay P, Heath C, Serhal P, Harper JC, et al. Investigation of gene expression profiles before and after embryonic genome activation and assessment of functional pathways at the human metaphase II oocyte and blastocyst stage. Fertil Steril. 2013; 99(3):803–+. https://doi.org/10.1016/j.fertnstert.2012.10.036 PMID: 23148922

72. Siva K, Venu P, Mahadevan A, Shankar SK, Inamdar MS. Human BCAS3 Expression in Embryonic Stem Cells and Vascular Precursors Suggests a Role in Human Embryogenesis and Tumor Angiogenesis. PloS one. 2007; 2(11).

73. Liu J, Fu YY, Sun XY, Li FX, Li YX, Wang YL. Expression of SWAP-70 in the uterus and feto-maternal interface during embryonic implantation and pregnancy in the rhesus monkey (Macaca mulatta). Histochem Cell Biol. 2006; 126(6):695–704. https://doi.org/10.1007/s00418-006-0206-1 PMID: 16786323

74. Zhou L, Li R, Wang R, Huang HX, Zhong K. Local injury to the endometrium in controlled ovarian hyperstimulation cycles improves implantation rates. Fertil Steril. 2008; 89(5):1166–76. https://doi.org/10.1016/j.fertnstert.2007.05.064 PMID: 17681303

75. Solca C, Tint GS, Patel SB. Dietary xenosterols lead to infertility and loss of abdominal adipose tissue in sterolin-deficient mice. J Lipid Res. 2013; 54(2):397–409. https://doi.org/10.1194/jlr.M031476 PMID: 23180829

76. Moretti E, Collodel G, Mazzi L, Russo I, Giurisato E. Ultrastructural study of spermatogenesis in KSR2 deficient mice. Transgenic Res. 2015; 24(4):741–51. https://doi.org/10.1007/s11248-015-9886-4 PMID: 26055731

77. Dokras A. Cardiovascular disease risk in women with PCOS. Steroids. 2013; 78(8):773–6. https://doi.org/10.1016/j.steroids.2013.04.009 PMID: 23624351

78. Azziz R, Dumesic DA, Goodarzi MO. Polycystic ovary syndrome: an ancient disorder? Fertil Steril. 2011; 95(5):1544–8. Epub 2010/10/29. https://doi.org/10.1016/j.fertnstert.2010.09.032 PMID: 20979996

79. Kenigsberg S, Bentov Y, Chalifa-Caspi V, Potashnik G, Ofir R, Birk OS. Gene expression microarray profiles of cumulus cells in lean and overweight-obese polycystic ovary syndrome patients. Molecular human reproduction. 2009; 15(2):89–103. https://doi.org/10.1093/molehr/gan082 PMID: 19141487

80. Manneras-Holm L, Benrick A, Stener-Victorin E. Gene expression in subcutaneous adipose tissue differs in women with polycystic ovary syndrome and controls matched pair-wise for age, body weight, and body mass index. Adipocyte. 2014; 3(3):190–6. https://doi.org/10.4161/adip.28731 PMID: 25068085

81. Salilew-Wondim D, Wang Q, Tesfaye D, Schellander K, Hoelker M, Hossain MM, et al. Polycystic ovarian syndrome is accompanied by repression of gene signatures associated with biosynthesis and metabolism of steroids, cholesterol and lipids. J Ovarian Res. 2015; 8.

82. Scotti L, Parborell F, Irusta G, De Zuniga I, Bisioli C, Pettorossi H, et al. Platelet-derived growth factor BB and DD and angiopoietin1 are altered in follicular fluid from polycystic ovary syndrome patients. Mol Reprod Dev. 2014; 81(8):748–56. https://doi.org/10.1002/mrd.22343 PMID: 24889290

83. Yan L, Wang A, Chen L, Shang W, Li M, Zhao Y. Expression of apoptosis-related genes in the endometrium of polycystic ovary syndrome patients during the window of implantation. Gene. 2012; 506 (2):350–4. https://doi.org/10.1016/j.gene.2012.06.037 PMID: 22789864

84. Kilic M, Ozgul RK, Coskun T, Yucel D, Karaca M, Sivri HS, et al. Identification of Mutations and Evaluation of Cardiomyopathy in Turkish Patients with Primary Carnitine Deficiency. Jimd Rep. 2012; 3:17–23. https://doi.org/10.1007/8904_2011_36 PMID: 23430869

85. Tamai I. Pharmacological and pathophysiological roles of carnitine/organic cation transporters (OCTNs: SLC22A4, SLC22A5 and Slc22a21). Biopharm Drug Dispos. 2013; 34(1):29–44. https://doi.org/10.1002/bdd.1816 PMID: 22952014

86. Maqdasy S, Baptissart M, Vega A, Baron S, Lobaccaro JMA, Volle DH. Cholesterol and male fertility: What about orphans and adopted? Mol Cell Endocrinol. 2013; 368(1–2):30–46. https://doi.org/10.1016/j.mce.2012.06.011 PMID: 22766106

87. Schisterman EF, Mumford SL, Chen Z, Browne RW, Boyd Barr D, Kim S, et al. Lipid concentrations and semen quality: the LIFE study. Andrology. 2014; 2(3):408–15. https://doi.org/10.1111/j.2047-2927.2014.00198.x PMID: 24596332

88. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. Gene Expression Levels Are a Target of Recent Natural Selection in the Human Genome. Mol Biol Evol. 2009; 26(3):649–58. https://doi.org/10.1093/molbev/msn289 PMID: 19091723

89. Barrett-Connor E, Bush TL. Estrogen and coronary heart disease in women. Jama. 1991; 265 (14):1861–7. Epub 1991/04/10. PMID: 2005736

90. Beral V. Long term effects of childbearing on health. Journal of epidemiology and community health. 1985; 39(4):343–6. Epub 1985/12/01. PMID: 4086966

91. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, et al. Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. Plos Genet. 2012; 8(4):299–313.

92. Haritunians T, Taylor KD, Targan SR, Dubinsky M, Ippoliti A, Kwon S, et al. Genetic Predictors of Medically Refractory Ulcerative Colitis. Inflamm Bowel Dis. 2010; 16(11):1830–40. https://doi.org/10.1002/ibd.21293 PMID: 20848476

93. Lang B, Alrahbeni TM, Clair DS, Blackwood DH, International Schizophrenia C, McCaig CD, et al. HDAC9 is implicated in schizophrenia and expressed specifically in post-mitotic neurons but not in adult neural stem cells. American journal of stem cells. 2012; 1(1):31–41. PMID: 23671795

94. Carter AJ, Nguyen AQ. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. Bmc Med Genet. 2011; 12:160. https://doi.org/10.1186/1471-2350-12-160 PMID: 22151998

95. Wang X, Byars SG, Stearns SC. Genetic links between post-reproductive lifespan and family size in Framingham. Evol Med Public Health. 2013; 2013(1):241–53. https://doi.org/10.1093/emph/eot013 PMID: 24481203

96. Rodríguez JA, Marigorta UM, Hughes DA, Spataro N, Bosch E, Navarro A. Antagonistic pleiotropy and mutation accumulation influence human senescence and disease. Nature Ecology & Evolution. 2017; 1:1–5.

97. Stearns SC. The evolution of life histories. Oxford; New York: Oxford University Press; 1992. xii, 249 p. p.

98. Abraham G, Havulinna AS, Bhalala OG, Byars SG, de Livera AM, Yetukuri L, et al. Genomic prediction of coronary heart disease. European Heart Journal. 2016; 37(43):3267–78. https://doi.org/10.1093/eurheartj/ehw450 PMID: 27655226

99. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited evidence for classic selective sweeps in African populations. Genetics. 2012; 192(3):1049–64. https://doi.org/10.1534/genetics.112.144071 PMID: 22960214

100. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449(7164):851–61. https://doi.org/10.1038/nature06258 PMID: 17943122

101. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L, et al. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome research. 2010; 20(10):1344–51. https://doi.org/10.1101/gr.106534.110 PMID: 20810666

102. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. PloS one. 2014; 9(4):e93766. https://doi.org/10.1371/journal.pone.0093766 PMID: 24718290

103. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4:7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

104. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. Plos Genet. 2014; 10(4):e1004234. https://doi.org/10.1371/journal.pgen.1004234 PMID: 24743097

105. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. Plos Genet. 2009; 5(6):e1000529. https://doi.org/10.1371/journal.pgen.1000529 PMID: 19543373

106. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome research. 2008; 18(11):1829–43. https://doi.org/10.1101/gr.076521.108 PMID: 18849525

107. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012; 40(Database issue):D84–90. https://doi.org/10.1093/nar/gkr991 PMID: 22086963

108. Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012; 28(8):1176–7. https://doi.org/10.1093/bioinformatics/bts115 PMID: 22402612

109. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of Cis Regulatory Variation in Diverse Human Populations. Plos Genet. 2012; 8(4):272–84.

110. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nature methods. 2011; 8(10):833–5. https://doi.org/10.1038/nmeth.1681 PMID: 21892150

111. Bakshi A, Zhu ZH, Vinkhuyzen AAE, Hill WD, Mcrae AF, Visscher PM, et al. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. Sci Rep-Uk. 2016;6.