RESEARCH ARTICLE

Clinical Utility of a Coronary Heart Disease Risk Prediction Gene Score in UK Healthy Middle Aged Men and in the Pakistani Population

Katherine E. Beaney[1], Jackie A. Cooper[1], Saleem Ullah Shahid[2], Waqas Ahmed[3], Raheel Qamar[4], Fotios Drenos[1,5], Martin A. Crockard[6], Steve E. Humphries[1]*

1 Centre for Cardiovascular Genetics, British Heart Foundation Laboratories, Institute of Cardiovascular Science, University College London, University Street, London, United Kingdom, 2 Department of Microbiology and Molecular Genetics, New Campus, University of the Punjab, Lahore, Pakistan, 3 Department of Microbiology, University of Haripur, Haripur, Pakistan, 4 COMSATS Institute of Information Technology, Park Road, Chak Shahzad, Islamabad, Pakistan, 5 Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, 6 Molecular Diagnostics Group, Randox Laboratories Ltd, Crumlin, United Kingdom

* steve.humphries@ucl.ac.uk

Abstract

Background

Numerous risk prediction algorithms based on conventional risk factors for Coronary Heart Disease (CHD) are available but provide only modest discrimination. The inclusion of genetic information may improve clinical utility.

Methods

We tested the use of two gene scores (GS) in the prospective second Northwick Park Heart Study (NPHSII) of 2775 healthy UK men (284 cases), and Pakistani case-control studies from Islamabad/Rawalpindi (321 cases/228 controls) and Lahore (414 cases/219 controls). The 19-SNP GS included SNPs in loci identified by GWAS and candidate gene studies, while the 13-SNP GS only included SNPs in loci identified by the CARDIoGRAMplusC4D consortium.

Results

In NPHSII, the mean of both gene scores was higher in those who went on to develop CHD over 13.5 years of follow-up (19-SNP p=0.01, 13-SNP p=7x10^{-3}). In combination with the Framingham algorithm the GSs appeared to show improvement in discrimination (increase in area under the ROC curve, 19-SNP p=0.48, 13-SNP p=0.82) and risk classification (net reclassification improvement (NRI), 19-SNP p=0.28, 13-SNP p=0.42) compared to the Framingham algorithm alone, but these were not statistically significant. When considering only individuals who moved up a risk category with inclusion of the GS, the improvement in

risk classification was statistically significant (19-SNP p=0.01, 13-SNP p=0.04). In the Pakistani samples, risk allele frequencies were significantly lower compared to NPHSII for 13/19 SNPs. In the Islamabad study, the mean gene score was higher in cases than controls only for the 13-SNP GS (2.24 v 2.34, p=0.04). There was no association with CHD and either score in the Lahore study.

Conclusion

The performance of both GSs showed potential clinical utility in European men but much less utility in subjects from Pakistan, suggesting that a different set of risk loci or SNPs may be required for risk prediction in the South Asian population.

Introduction

Despite being largely preventable, coronary heart disease (CHD) remains the most common cause of the death worldwide [1, 2]. There is a growing CHD burden in the developing world [3], particularly in South Asian countries such as Pakistan where the prevalence of CHD in urban Karachi has approximately doubled since 1970[4]. This increase is most likely due to the adoption of a more "Western" lifestyle combined with a greater susceptibility to metabolic syndrome[5]. Many CHD risk factors, such as smoking status and hypertension, can be modified through lifestyle changes or drug therapy[6], which can greatly reduce CHD morbidity and mortality[7, 8]. Given that the atherosclerotic process can begin many decades before clinical manifestation, this provides an opportunity for preventative measures to be employed to avoid its escalation[9].

In order to target lifestyle and therapeutic intervention appropriately, those most at risk of developing CHD must be identified. A number of conventional risk factor (CRF) scores to determine 10-year risk CHD risk have been developed. These include the Framingham risk score [10], SCORE [11] and QRISK2 [12]. Individuals are then classified on the basis of the CRF score. Until recently the cut-off for the high risk category (those who qualify for statin treatment for primary prevention of CHD) was set at 20% ten year risk of CHD[6]. However, lower cut-offs have been proposed in both the USA [13] and the UK, with the 2014 National Institute for Health and Care Excellence (NICE) guidelines setting the threshold at 10%[14]. To date, such risk scores have provided only modest discrimination. Most events occur in individuals classified as being at intermediate risk [1, 15]. By improving risk prediction tools, preventative measures can be targeted more appropriately.

It is has been estimated that heritability of CHD is 40–50% [16] and more than forty CHD associated loci have been identified using both traditional candidate gene studies and genome-wide association studies (GWAS)[17, 18]. The most recent publication from the CARDIoGRAMplusC4D consortium [18] increased the number of robustly associated loci to 46. The inclusion of genetic information is therefore a good candidate to improve CHD risk prediction. As each variant is associated with a modest effect size, the addition of a single CHD risk variant to a CRF score does not improve risk stratification [19–21]. However, combining even a small number of variants into a "gene score" has been shown to improve classification or discrimination of individuals [22, 23].

The objective of this study was to determine whether the inclusion of a CHD risk gene score has clinical utility in the European population. We sought to do this by investigating whether addition of a gene score improves discrimination and prediction over and above the use of the

Framingham CRF score alone in a prospective study of middle-aged men from the UK. We then looked to determine if the same genetic variants can be used in individuals from the Indian Subcontinent. Two gene scores, one with 19 SNPs taken from both GWAS and candidate gene studies and one with 13 SNPs where only SNPs in loci identified by the CARDIo-GRAMplusC4D consortium were included, were tested.

Methods

NPHSII

NPHSII is a prospective CHD study of approximately 3000 men as described previously [2]. Briefly, middle-aged men (50–64) were recruited from 9 general practices in the UK. Anyone with a history of CHD or diabetes was excluded. There was a median of 13.5 years follow-up. CHD was defined as acute myocardial infarction (MI), silent MI or undergoing coronary surgery. All subjects gave written informed consent and the study had ethical approval from the national research ethics service (NRES) Committee London-Central.

Pakistani cohorts

Two case-control studies from Pakistan were collected. One group was collected from the Rawalpindi Institute of Cardiology, Pakistan. All cases had had an MI as defined by a positive test for troponin T, ST segment changes on electrocardiogram and typical chest pain radiating in the chest that was not relieved at rest. This study will hereafter be referred to as the "Islamabad" study. The study had approval from the Institutional Review Board and Ethics Committee of Shifa College of Medicine, Shifa International Hospital, Islamabad and all subjects gave written informed consent. The second group were collected from hospitals in Lahore, Pakistan. All cases had CHD as defined by echocardiogram or angiography. This study will hereafter be referred to as the "Lahore" study. All subjects gave written informed consent and the study had ethical approval from the institutional ethical committee, University of the Punjab, Lahore. In both studies controls were taken from the general population.

Gene Scores

The SNPs included in the gene scores are presented in Table 1, along with the source publication(s). Each SNP was weighted by its effect size as reported in their respective publications and the gene scores were calculated by multiplying the number of risk alleles by the natural log of the odds ratio for each SNP and adding the products together. It was assumed that all SNPs were acting in an additive manner apart from rs1799983, which was treated in a recessive manner. A 13 SNP score was also considered using only SNPs in loci identified by the most recent publication from the CARDIoGRAMplusC4D consortium [18].

Statistical analysis

Analysis for NPHSII was conducted using STATA v13.1[24]. Analysis for the Islamabad study and genotype frequency comparisons between all studies were conducted using R v3.0.3 [25]. Analysis for the Lahore study was conducted using SPSS version 22.0[26]. Variables were compared between the CHD and no-CHD groups using two-sided t-tests for continuous variables and χ^2 tests for categorical variables. Hardy Weinberg equilibrium was assessed using χ^2 tests. The Framingham 10-year risk score was calculated using the equations described in [10]. The relationship between quintile of gene score and CHD was investigated using logistic regression.

Table 1. SNPs included in the gene scores.

Gene/Locus	SNP	SNP Location	Risk Allele	Odds Ratio	Reference
MIA3*	rs17465637	Intergenic	C	1.14	Kathiresan et al. [36]
9p21*	rs10757274	Intergenic	G	1.29	Kathiresan et al. [36]
DAB2IP	rs7025486	Intergenic	A	1.16	Harrison, Cooper [37]
CXCL12*	rs1746048	Intergenic	C	1.17	Kathiresan et al. [36], Samani, Erdmann [38]
ACE	rs4341	Intergenic	G	1.22	Casas, Cooper [17]
NOS3	rs1799983	E289D	T	1.31	Casas, Cooper [17]
APOA5*	rs662799	Promoter Variant	G	1.19	Sarwar, Sandhu [39]
SMAD3	rs17228212	Intergenic	C	1.21	Samani, Erdmann [38]
APOB*	rs1042031	E4181K	A	1.73	Casas, Cooper [17]
CETP	rs708272	Intronic	C	1.28	Casas, Cooper [17]
LPA*	rs3798220	I1891M	C	1.92	Clarke, Peden [40]
LPA*	rs10455872	Intergenic	G	1.70	Clarke, Peden [40]
MRAS*	rs9818870	Intergenic	T	1.15	Erdmann, Grosshennig [41]
LPL*	rs328	S447X	C	1.25	Casas, Cooper [17]
LPL	rs1801177	D9N	A	1.33	Sagoo, Tatt [42]
SORT1*[+]	rs646776[+]	Intergenic	A	1.19	Kathiresan et al. [36]
PCSK9*	rs11591147	R46L	G	1.43	Benn, Nordestgaard [43]
APOE*	rs429358	C112R	C	1.06	Bennet, Di Angelantonio [44]
APOE*	rs7412	C158R	T[++]	0.80	Bennet, Di Angelantonio [44]

SNPs marked with an asterisk (*) are included in both the 19 and 13 SNP gene score.

[+]rs599839 was genotyped instead of rs646776, r^2 = 0.95 in Europeans

[++]For rs7412, the protective SNP is included in the gene score

doi:10.1371/journal.pone.0130754.t001

Results

Results from NPHSII

The baseline characteristics for the participants of NPHSII are presented in Table 2. As expected, the men who went on to develop CHD were older, had higher BMI, higher blood

Table 2. Baseline Characteristics of NPHSII participants.

Trait	NPHSIINo CHD(n = 2491)	NPHSIICHD(n = 284)	p-value
Age (years)	56.0 (3.4)	56.6 (3.5)	7×10^{-3}
Male (%)	100	100	-
Smoking (%)	27	37.0	5×10^{-4}
BMI (kg/m^2)	26.4 (3.5)	26.9 (3.4)	0.04
Systolic Blood Pressure (mmHg)	137.9 (19.1)	142.5 (19.4)	2×10^{-4}
Diastolic Blood Pressure (mmHg)	84.3 (11.2)	86.7 (11.4)	9×10^{-4}
Cholesterol (mmol/l)	5.70 (1.01)	6.06 (1.02)	$<1 \times 10^{-4}$
Triglyceride (mmol/l)	1.77 (0.93)	2.05 (1.06)	$<1 \times 10^{-4}$
Framingham 10 year CHD risk	0.10 (0.07–0.15)	0.14 (0.09–0.20)	9×10^{-4}
19 SNP Gene Score	3.30 (0.57)	3.42 (0.55)	0.01
13 SNP Gene Score	2.43 (0.48)	2.53 (0.45)	7×10^{-3}

All variable are present as the mean plus standard deviation, apart from the Framingham 10 year CHD risk score where the mean and interquartile range are given.

doi:10.1371/journal.pone.0130754.t002

Table 3. Reclassification of NPHSII participants with the addition of the gene scores to the Framingham conventional risk factor score.

Gene Score	CHD Status	Framingham 10 year CHD risk	Reclassification: Increased Risk	Reclassification: Decreased Risk	Proportion Increased Risk
19 SNP	No CHD	<10%	103	132	22.9%
19 SNP	CHD	≥ 10%	11	9	45.8%
			Overall NRI: 4.5% (-3.7%-12.7%) p = 0.28		Increased risk p = 0.01
13 SNP	No CHD	<10%	115	145	20.0%
13 SNP	CHD	≥ 10%	13	12	34.2%
			Overall NRI: 3.0% (-4.3%-10.3%) p = 0.42		Increased risk p = 0.04

NRI = net reclassification index.

doi:10.1371/journal.pone.0130754.t003

pressure, higher cholesterol, higher triglycerides and a higher proportion were smokers, at baseline. Moreover, those who went on to develop CHD had a statistically significantly higher 10 year CHD risk as calculated using the Framingham score.

The genotype frequency of each SNP in NPHSII is shown in S1 Table. All SNPs except rs1042031, in *APOB*, were in Hardy-Weinberg equilibrium. A comparison of risk allele frequency in those who did and those who did not go on to develop CHD is presented in S4 Table. There was a statistically significant difference between the two groups for rs10757274 at the 9p21 locus and rs1746048 (close to gene *CXCL12*), with the risk allele frequency being higher in the CHD group for both SNPs.

The genotype information was combined using a weighted genetic risk score. For the 19 SNP GS, full data for Framingham score plus gene score was available for 1164 individuals, 114 of whom developed CHD. For the 13 SNP GS full data for Framingham score plus gene score was available for 1437 individuals, 143 of whom developed CHD. Both gene scores were higher in those who developed CHD during follow-up (19 SNP GS p = 0.01, 13 SNP GS p = 7×10^{-3}).

We examined the ability of the two gene scores to reclassify those who did and those who did not go on to develop CHD. As shown in Table 3, although more participants appeared to be reclassified correctly with the inclusion of both the 19 SNP and 13 SNP GS compared to the Framingham score alone, this increase was not statistically significant in either case. When only those who were reclassified into the high risk category (set at 10% ten-year risk) were considered (those with a strong genetic predisposition to CHD with intermediate and low CRF scores), a statistically significant number of participants were reclassified correctly (Table 3). This was also the case when the cut-off for high risk was set at 20% (shown in S5 Table).

The predicative ability of using the Framingham CRF score was compared to the Framingham score plus the genetic risk score. There was no statistically significant increase in the area under the ROC curve for either the 19 SNP GS (p = 0.48) or the 13 SNP GS (p = 0.82) (S1 Fig).

As shown in Fig 1 and S6 Table, there was a statistically significant trend for higher quintiles of gene score to have higher odds ratio for CHD compared to the bottom quintile for both the 19 SNP GS (p = 8×10^{-3}) and the 13 SNP GS (p = 0.01) after adjustment for age.

Results from Pakistani studies

We then attempted to transpose these scores in two case-control samples from Pakistan. The available conventional risk factor information for both studies is presented in Table 4. Data was missing for up to 60% of participants for some variables in the Islamabad study. As expected, in both studies the case group was older and had a higher proportion with diabetes, hypertension and smokers. There was no difference in the proportion of males present in each group. Surprisingly, in the Islamabad study, mean LDL cholesterol was not statistically
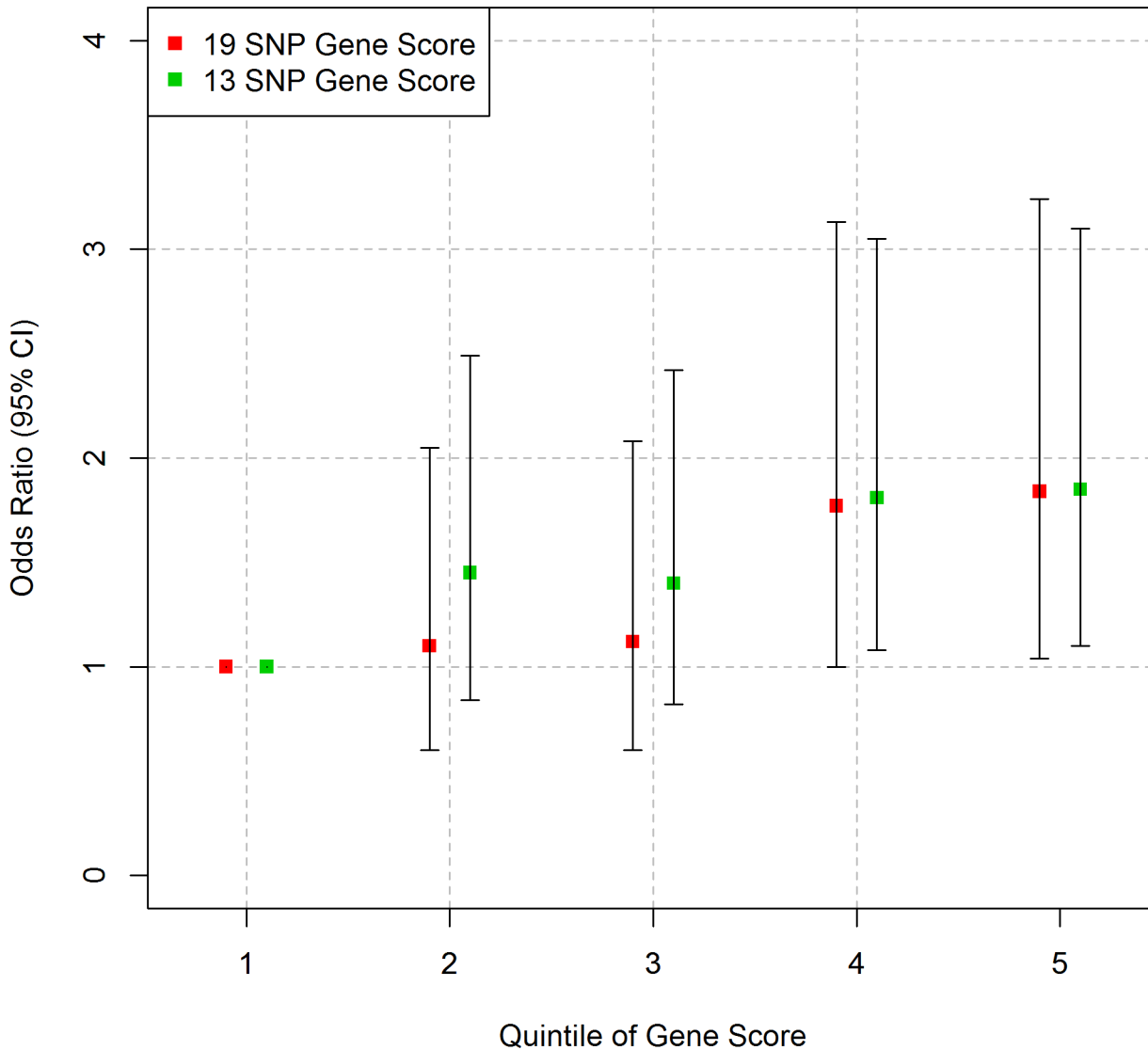
Fig 1. Association between gene score and CHD in NPHSII. Logistic regression (age adjusted) was performed for each group. Error bars represent 95% confidence intervals.

doi:10.1371/journal.pone.0130754.g001

significantly different between cases and controls. This can be attributed to the treatment of those post-MI with lipid lowering therapies. BMI, total cholesterol, triglycerides, family history and HDL cholesterol data were available for some of the participants of the Islamabad study. Only for HDL cholesterol was the difference between the two groups statistically significant, being lower in the cases.

The 19 SNPs were genotyped in both Pakistani sample sets and the results are presented in S1 Table. For the Islamabad study, five SNPs were not in Hardy-Weinberg equilibrium—*MIA3* rs17465637, *CXCL12* rs1746048, *MRAS* rs9818870, *LPL* rs1801177 and *SMAD3* rs17228212-with an excess of homozygotes. Genotypes were confirmed by sequencing. In the Lahore study, only one SNP–*LPA* rs10455872- was not in Hardy Weinberg equilibrium. The risk allele frequencies did not differ between the two Pakistani groups as shown in Table 5. The data from the control groups was combined and compared to that from the group in NPHSII.

Table 4. Characteristics of the Pakistani sample sets.

	Islamabad			Lahore		
Trait	ControlsN = 228	CasesN = 321	p-value	ControlsN = 219	CasesN = 404	p-value
Age (mean)	38(11.83)	53(11.80)	$<2.2 \times 10^{-16}$	56 (10.50)	59 (12.60)	2×10^{-3}
Gender (% Female)	34%	31%	0.52	46%	41%	0.27
BMI (kg/m^2)	24.2(3.95)	24.3(4.08)	0.85	-	-	-
TC (mmol/l)	4.52(1.38)	4.71(3.77)	0.56	-	-	-
LDL-C (mmol/l)	2.66(0.78)	2.55(0.99)	0.33	2.19 (0.44)	2.74 (0.75)	6.50×10^{-22}
HDL-C (mmol/l)	1.35(1.01)	0.96(0.22)	1.01×10^{-4}	-	-	-
TG (mmol /l)*	1.44 (0.91)	1.58 (1.06)	0.22	-	-	-
Diabetes (%)	1	32	6.6×10^{-9}	14	65	5.10×10^{-34}
Hypertension (%)	15	46	4.9×10^{-7}	16	62	9.00×10^{-28}
Family History (%)	42	42	0.96	-	-	-
Smoking (%)	25	46	1.25×10^{-3}	11	30	4.30×10^{-8}
19 SNP Gene Score	2.89 (0.50)	2.94(0.50)	0.35	2.92(0.53)	2.93(0.50)	0.75
13 SNP Gene Score	2.24(0.42)	2.34(0.42)	0.04	2.21(0.39)	2.21(0.39)	0.95

Categorical variables were compared using a χ^2 test while continuous variables were compared using Welch's t-tests.

* Log transformed data. Geometric mean and approximate SD are given.

doi:10.1371/journal.pone.0130754.t004

Table 5. Comparison of risk allele frequencies between the control groups from the Pakistani studies and between the combined total of the Pakistani control groups and NPHSII.

Gene/Locus	SNP	RAF Islamabad Controls	RAF Lahore Controls	p-value	RAF NPSHII(95% CI)	p-value
MIA3	rs17465367	0.64(0.59–0.69)	0.63(0.58–0.67)	0.68	0.71(0.69–0.72)	3.42×10^{-5}
9p21	rs10757274	0.44(0.40–0.49)	0.46(0.42–0.51)	0.60	0.48(0.47–0.50)	3.16×10^{-3}
DAB2IP	rs7025486	0.31 (0.26–0.35)	0.32(0.27–0.36)	0.84	0.26(0.17–0.24)	4.70×10^{-4}
CXCL12	rs1746048	0.65 (0.61–0.70)	0.64(0.59–0.68)	0.65	0.86(0.85–0.87)	$<2.20 \times 10^{-16}$
SMAD3	rs17228212	0.19 (0.15–0.22)	0.18(0.14–0.21)	0.76	0.31(0.30–0.32)	4.88×10^{-14}
MRAS	rs9818870	0.10 (0.07–0.13)	0.09(0.06–0.12)	0.75	0.16(0.15–0.17)	2.61×10^{-6}
SORT1	rs599839	0.72 (0.68–0.77)	0.74(0.70–0.79)	0.55	0.78(0.77–0.79)	2.03×10^{-3}
ACE	rs4341	0.41 (0.36–0.45)	0.47(0.43–0.52)	0.05	0.52(0.50–0.53)	3.09×10^{-5}
NOS3	rs1799983	0.16 (0.13–0.20)	0.18(0.15–0.22)	0.51	0.33(0.32–0.35)	$<2.20 \times 10^{-16}$
APOA5	rs662799	0.15 (0.12–0.18)	0.17(0.13–0.20)	0.60	0.06(0.05–0.07)	$<2.20 \times 10^{-16}$
APOB	rs1042031	0.15 (0.12–0.19)	0.13(0.09–0.16)	0.24	0.18(0.17–0.19)	9.75×10^{-3}
CETP	rs708272	0.55 (0.50–0.60)	0.56(0.51–0.61)	0.83	0.56(0.55–0.58)	0.62
LPA	rs3789220	0.01 (0–0.02)	0.003(0–0.01)	0.81	0.02(0.01–0.02)	0.02
LPA	rs10455872	0.01 (0–0.03)	0.01(0.00–0.03)	1	0.07(0.07–0.08)	1.06×10^{-10}
PCSK9	rs11591147	1.00	0.995 (0.00–0.01)	0.50	0.99 (0.99–0.99)	0.07
APOE	rs429358	0.09 (0.06–0.11)	0.11 (0.08–0.14)	0.52	0.17 (0.16–0.18)	1.60×10^{-9}
APOE	rs7412	0.96 (0.94–0.98)	0.96 (0.94–0.98)	1	0.91 (0.90–0.92)	2.69×10^{-6}
LPL	rs328	0.92 (0.89–0.94)	0.91 (0.89–0.94)	0.98	0.90 (0.89–0.91)	0.12
LPL	rs1801177	0.01 (0–0.02)	0	0.14	0.01 (0.01–0.02)	0.04

Comparisons were performed using tests of proportion. RAF = Risk Allele Frequency, CI = Confidence Interval.

doi:10.1371/journal.pone.0130754.t005

The risk allele frequency was lower in the Pakistani group for 13 SNPs and higher for three SNPs compared to NPHSII.

The 19 and 13 SNP GSs were calculated for both Pakistani studies and the results are shown in Table 4. For the 19 SNP GS, full genotyping was available for 304 samples (119 controls/175 cases) in the Islamabad study and for 438 samples (130 controls/308 cases) in the Lahore study. For the 13 SNP GS, full genotyping was available for 317 samples (123 controls/194 cases) in the Islamabad study and for 490 samples (145 controls/345 cases) in the Lahore study. In the Islamabad sample, a statistically significant increase in the mean gene score was observed between cases and controls for the 13 SNP GS (p = 0.04), but not for the 19 SNP GS (p = 0.35). For the Lahore sample, the mean gene score was not statistically significantly different between cases and controls for either score (19 SNP GS p = 0.75, 13 SNP GS p = 0.95). The 19 and 13 SNP GSs were found to be *lower* in the controls compared to those in NPHSII who did not go on to develop CHD in both the Islamabad (19 SNP GS p = 2.92×10^{-14}, 13 SNP GS p = 5.62×10^{-6}) and the Lahore cohorts (19 SNP GS p = 7.78×10^{-13}, 13 SNP GS p = 1.49×10^{-9}).

While the higher quintiles of the 13 SNP GS gene score appear to have a greater odds ratio for outcome compared to the bottom quintile in the Islamabad sample, this was not statistically significant. No trend between quintile of gene score and outcome was observed for the 19 SNP GS in the Islamabad sample or for either GS in the Lahore sample (Fig 2 and S7 Table).

Discussion

The aim of including genetic information in CHD risk prediction is to identify those who have a high overall risk but who would be classed as being at low or intermediate risk using conventional risk factor scores. In this study, we found that a 19 SNP GS, which uses CHD risk weighting from published meta-analyses, showed improved risk classification over and above the Framingham score alone in NPHSII when those men who went up a risk category (for example moving into the high risk category) were considered. The exclusion of six SNPs in non-CARDIoGRAMplusC4D loci had a modest detrimental effect on the utility of the gene score in the UK men, reducing the overall NRI non-significantly from 4.5% to 3.0% with a similar non-significant decrease in the number of men correctly reclassified into the high risk group. Thus the 19 SNP GS, appears to have the better utility. By contrast, in two case-control studies from Pakistan the gene scores did not perform as well. While in the subjects from Islamabad the 13 SNP GS showed a significant case-control effect, neither score showed utility in the samples from Lahore.

There are a number of potential reasons why the results from the European group were not confirmed in Pakistanis. Firstly both Pakistani studies comprise only a few hundred samples and thus sample size and statistical power is an issue. We estimated that should the same effect found in NPHSII be present in the Pakistani group, to have 80% power to detect it (at the 5% significance threshold), approximately 340 cases and approximately 340 controls would be required for both gene scores. This is well above the number of individuals for whom complete genotyping was available in both Pakistani studies. Secondly, "cases" were defined differently in the two Pakistani studies. There was very little difference between the mean gene score in cases and controls in the Lahore sample which used a much broader definition of CHD, compared to the Islamabad group which used an MI phenotype (more like the "hard" endpoints used in the prospective NPHSII). A further limitation of the study is that using retrospectively collected cases restricts the case group to those who have survived which may introduce a survivor bias to the sample.

A further possibility is SNP selection. While eight of the SNPs are non-synonymous (and thus are likely to be the functional), and the *APOA5* promoter variant has been shown to be directly functional [27] many are not in protein coding regions. It is possible that such SNPs
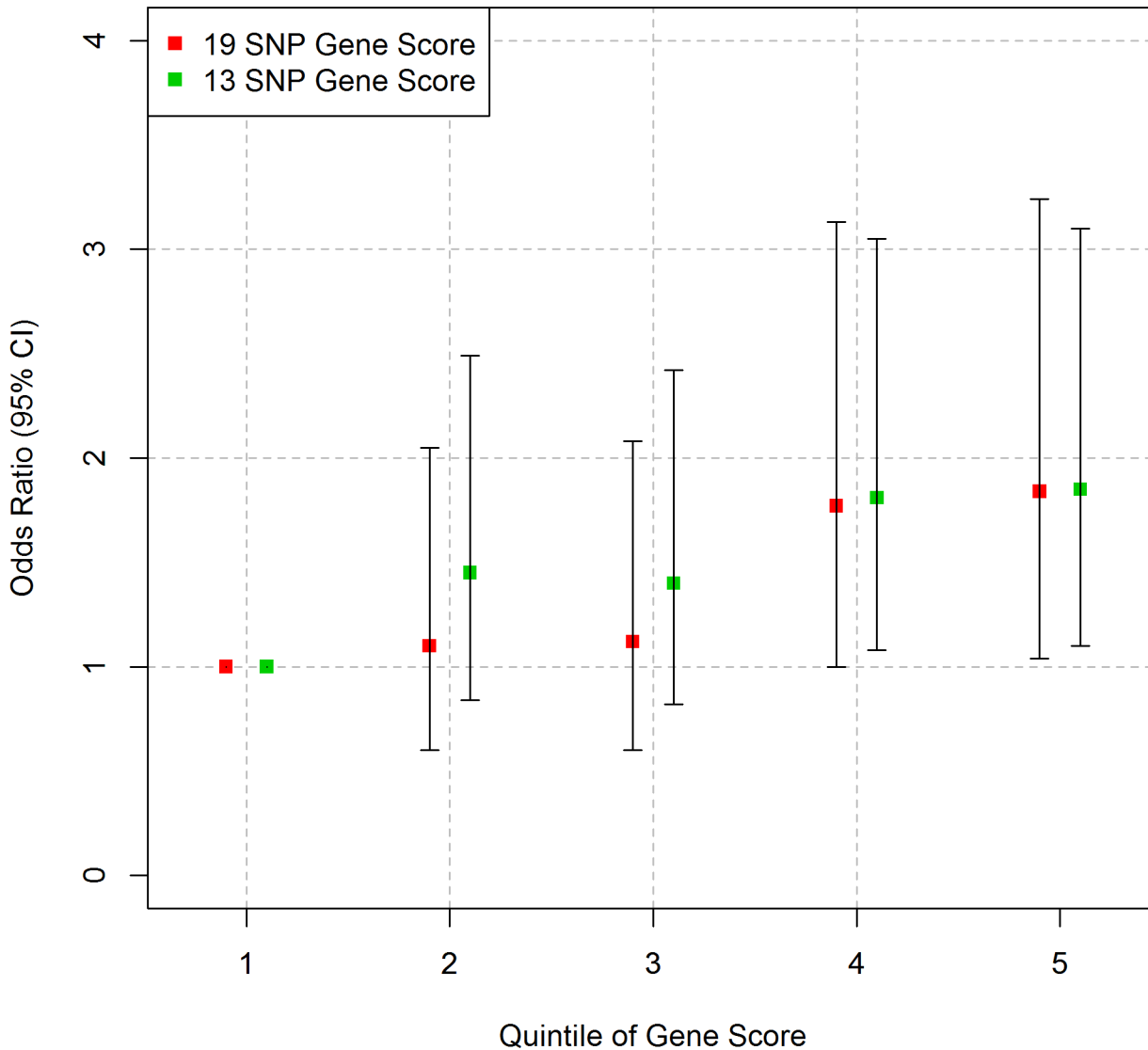
Fig 2. Association between gene score and outcome in the Pakistani samples. Logistic regression was performed for each group. (A) Islamabad study, outcome is MI, and (B) Lahore study, outcome is CHD. Error bars represent 95% confidence intervals.

merely "tag" the functional SNP located in the same linkage disequilibrium (LD) block. Given the differing LD patterns between ethnicities, it is possible that some of the SNPs used here are in strong LD with the functional SNP in Europeans but not in other ethnic groups. For example, the SNPs rs599839 and rs646776 are often used to tag each other at the *CELSR2-PSRC1--SORT1* locus, with rs599839 being known to be functional [28]. Data from the HapMap project [29] shows that the LD between the two SNPs is $r^2 = 0.95$ in the Northern European group and $r^2 = 0.76$ in the Japanese group but the two SNPs are not in LD in the Yoruba group (from Ibadan, Nigeria). However, a study of approximately 2000 Indians collected from Mumbai and Bangalore found the two SNPs to be in strong LD ($r^2 = 0.98$) in this group. Furthermore, a number of SNPs at the 9p21 locus were genotyped in the PROMIS study of ethnic Pakistanis living in the UK and the LD pattern was found to be similar to Europeans at this locus [30] and a comparison of the genetic architecture of the *LPA* gene in individuals of

different ethnicities (including South Asians) living in Canada found similar haplotype blocks to be present in the groups studied [31]. Therefore, for these loci we conclude that it is appropriate to use the SNP identified in studies with European participants. To what extent this can be extrapolated to the other SNPs and loci included is not clear, and for example, a study of *CETP* polymorphisms in the Punjabi population from Northern India found the LD pattern to differ from that observed in Europeans[32] but overall information regarding LD patterns in South Asians is limited. Moreover, given that the risk allele frequency is statistically significantly lower for 13 of the 19 SNPs in the Pakistani group compared to NPHSII, it is unsurprising that the gene scores are also statistically significantly lower. This shows that even if many of the SNPs are functional, then gene scores based on these SNPs will perform less well in the Pakistani group. In order to optimise the use of the gene scores for use in those of South Asian ethnicity, the functional SNP should be identified (bearing in mind that the functional SNP may differ between ethnic groups) in all cases to maximise the performance of a gene score based on this set of loci, and it may be necessary to design a South-Asian specific gene score.

We can only speculate whether the modest performance of the gene scores observed in Pakistani subjects here will be seen in samples from other parts of the Indian Subcontinent, as there is considerable genetic diversity within the region and since score performance is in part dependent on the frequency of the risk alleles. To date, allele frequency information across the Indian Subcontinent is limited for the SNPs used. A small number of Punjabi subjects (n = 96) from Lahore were genotyped as part of the 1000 genomes project (phase 3) [33]. No difference in allele frequencies was observed for any of the SNPs in this data set when compared with the control groups presented here. The risk allele frequency of the 9p21 SNPs genotyped in the PROMIS study were statistically significantly higher than observed here [30]. While a similar risk allele frequency to that found here was observed for rs662799 (in the *APOA5* promoter) in a Punjabi Pakistani case-control study, the risk allele frequency observed for rs599839 (at the *SORT1* locus) in that study was much lower than observed here [34]. Another factor which may influence how representative the results presented here are, is that five of the 19 SNPs genotyped in the Islamabad case-control study were not in Hardy-Weinberg equilibrium. In each case this was due to an excess of homozygotes compared to what would be expected. This is likely caused by the presence of population sub-structure[35]. Clearly much more data is required from larger samples representative of different regions and groups from the Indian Subcontinent to determine the frequencies of these SNPs.

Overall this work demonstrated that the use of a 19 SNP gene score has the potential to improve risk stratification in European individuals over and above classical CHD risk factors. Reducing the number of SNPs in the score to only the 13 with GWAS-proven effects on CHD risk had a modest detrimental effect on the utility. We do not have definitive evidence of the utility of either score in subjects from Pakistan and larger samples are required to determine this.

Supporting Information

S1 Fig. ROC curves for CHD prediction in NPHSII.
(TIF)

S1 Table. Risk allele frequency for each SNP in NPHSII and the Pakistani sample sets.
HWE = Hardy Weinberg Equilibrium, CI = Confidence Interval.
(DOCX)

S2 Table. Comparison of risk allele frequency between Lahore controls and NPHSII. Comparisons were performed using proportion tests. CI = Confidence Interval.
(DOCX)

S3 Table. Comparison of risk allele frequency between Islamabad controls and NPHSII.
Comparisons were performed using tests of proportion. CI = Confidence Interval.
(DOCX)

S4 Table. Comparison of risk allele frequencies in those who did and did not develop CHD during follow-up of NPHSII. Comparisons were performed using proportion tests.
CI = Confidence Interval.
(DOCX)

S5 Table. Reclassification of NPHSII participants with the addition of the gene scores to the Framingham conventional risk factor score (20% cut-off for high risk). NRI = Net Reclassification Index.
(DOCX)

S6 Table. Odds ratio for CHD in NPHSII by quintile of gene score, compared to the lowest quintile. Logistic regression, age adjusted, was performed for each group. OR = Odds Ratio, CI = Confidence Interval.
(DOCX)

S7 Table. Odds ratio for outcome (MI for Islamabad, CHD for Lahore) by quintile of gene score, compared to the lowest quintile. Logistic regression, adjusted for age and sex, was performed for each group. OR = Odds Ratio, CI = Confidence Interval. For the Islamabad study full data was available for 258 participants for the 19 SNP GS (106 controls/152 cases) and 268 participants for the 13 SNP GS (110 controls/158 cases). For the Lahore study full data was available for 438 participants for the 19 SNP GS (130 controls/308 cases) and 490 participants for the 13 SNP GS (145 controls/345 cases).
(DOCX)

Author Contributions

Conceived and designed the experiments: SEH. Performed the experiments: KEB SUS. Analyzed the data: KEB JAC SUS. Contributed reagents/materials/analysis tools: WA RQ FD. Wrote the paper: KEB SEH. Discussed study in role as industrial supervisor to KEB: MAC.

References

1. Cooper JA, Miller GJ, Humphries SE. A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. Atherosclerosis. 2005; 181(1):93–100. doi: 10.1016/j.atherosclerosis.2004.12.026 WOS:000230215300012. PMID: 15939059

2. Miller GJ, Bauer KA, Barzegar S, Foley AJ, Mitchell JP, Cooper JA, et al. The effects of quality and timing of venepuncture on markers of blood coagulation in healthy middle-aged men. Thromb Haemost. 1995; 73(1):82–6. Epub 1995/01/01. PMID: 7740501.

3. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. Curr Probl Cardiol. 2010; 35(2):72–115. Epub 2010/01/30. doi: 10.1016/j.cpcardiol.2009.10.002 PMID: 20109979; PubMed Central PMCID: PMC2864143.

4. Aziz K, Faruqui A, Patel N, Jaffery H. Prevalence and Awareness of Cardiovascular Disease Including Lifestyles in a Lower Middle Class Urban Community in an Asian Country Pakistan Heart Journal. 2008; 41(3–4):11–20.

5. Jafar TH, Qadri Z, Chaturvedi N. Coronary artery disease epidemic in Pakistan: more electrocardiographic evidence of ischaemia in women than in men. Heart. 2008; 94(4):408–13. Epub 2007/07/25. doi: 10.1136/hrt.2007.120774 PMID: 17646192; PubMed Central PMCID: PMC2565583.

6. Wood D, Wray R, Poulter N, Williams B, Kirby M, Patel V, et al. JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. Heart. 2005;91. doi: 10.1136/hrt.2005.079988 WOS:000234060500001.

7. Ford ES, Ajani UA, Croft JB, Critchley JA, Labarthe DR, Kottke TE, et al. Explaining the decrease in U. S. deaths from coronary disease, 1980–2000. New Engl J Med. 2007; 356(23):2388–98. Epub 2007/06/08. doi: 10.1056/NEJMsa053935 PMID: 17554120.

8. Bakhru A, Erlinger TP. Smoking cessation and cardiovascular disease risk factors: results from the Third National Health and Nutrition Examination Survey. PLoS medicine. 2005; 2(6):e160. Epub 2005/06/25. doi: 10.1371/journal.pmed.0020160 PMID: 15974805; PubMed Central PMCID: PMC1160573.

9. Falk E, Nakano M, Bentzon JF, Finn AV, Virmani R. Update on acute coronary syndromes: the pathologists' view. Eur Heart J. 2013; 34(10):719–28. Epub 2012/12/18. doi: 10.1093/eurheartj/ehs411 PMID: 23242196.

10. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97(18):1837–47. Epub 1998/05/29. PMID: 9603539.

11. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003; 24(11):987–1003. doi: 10.1016/s0195-668x(03)00114-3 WOS:000183577800002. PMID: 12788299

12. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008; 336 (7659):1475–82. Epub 2008/06/25. doi: 10.1136/bmj.39609.449676.25 PMID: 18573856; PubMed Central PMCID: PMC2440904.

13. Goff DC Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB Sr., Gibbons R, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2013. Epub 2013/11/14. doi: 10.1161/01.cir.0000437741.48606.98 PMID: 24222018.

14. National Institute for Health and Care Excellence: Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. 2014.

15. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ. 2010; 340:c2442. Epub 2010/05/15. doi: 10.1136/bmj.c2442 PMID: 20466793; PubMed Central PMCID: PMC2869403.

16. Peden JF, Farrall M. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. Hum Mol Genet. 2011; 20(R2):R198–205. Epub 2011/08/31. doi: 10.1093/hmg/ddr384 PMID: 21875899; PubMed Central PMCID: PMC3179381.

17. Casas JP, Cooper J, Miller GJ, Hingorani AD, Humphries SE. Investigating the genetic determinants of cardiovascular disease using candidate genes and meta-analysis of association studies. Ann Hum Genet. 2006; 70(Pt 2):145–69. Epub 2006/04/22. doi: 10.1111/j.1469-1809.2005.00241.x PMID: 16626327.

18. 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. doi: 10.1038/ng.2480 WOS:000312838800009. PMID: 23202125

19. Talmud PJ, Cooper JA, Palmen J, Lovering R, Drenos F, Hingorani AD, et al. Chromosome 9p21.3 coronary heart disease locus genotype and prospective risk of CHD in healthy middle-aged men. Clin Chem. 2008; 54(3):467–74. doi: 10.1373/clinchem.2007.095489 WOS:000253570400005. PMID: 18250146

20. Brautbar A, Ballantyne CM, Lawson K, Nambi V, Chambless L, Folsom AR, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk in Communities study. Circulation Cardiovascular genetics. 2009; 2(3):279–85. Epub 2009/12/25. doi: 10.1161/CIRCGENETICS.108.817338 PMID: 20031596; PubMed Central PMCID: PMC2771929.

21. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med. 2009; 150(2):65–72. Epub 2009/01/21. PMID: 19153409; PubMed Central PMCID: PMC2629586.

22. Davies RW, Dandona S, Stewart AF, Chen L, Ellis SG, Tang WH, et al. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. Circulation Cardiovascular genetics. 2010; 3(5):468–74. Epub 2010/08/24. doi: 10.1161/CIRCGENETICS.110.946269 PMID: 20729558; PubMed Central PMCID: PMC3035486.

23. Hughes MF, Saarela O, Stritzke J, Kee F, Silander K, Klopp N, et al. Genetic Markers Enhance Coronary Risk Prediction in Men: The MORGAM Prospective Cohorts. PLoS One. 2012; 7(7). doi: 10.1371/journal.pone.0040922. WOS:000306806600029.

24. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2013.

25. R Core Team. R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org/. 2014.

26. IBMCorp. IBM SPSS Statistics for Windows, Version 22.0. 22.0 ed. Armonk, NY: IBM 2013.

27. Talmud PJ, Palmen J, Putt W, Lins L, Humphries SE. Determination of the functionality of common APOA5 polymorphisms. J Biol Chem. 2005; 280(31):28215–20. Epub 2005/06/09. doi: 10.1074/jbc. M502144200 PMID: 15941721.

28. Linsel-Nitschke P, Heeren J, Aherrahrou Z, Bruse P, Gieger C, Illig T, et al. Genetic variation at chromosome 1p13.3 affects sortilin mRNA expression, cellular LDL-uptake and serum LDL levels which translates to the risk of coronary artery disease. Atherosclerosis. 2010; 208(1):183–9. doi: 10.1016/j. atherosclerosis.2009.06.034 WOS:000273567300031. PMID: 19660754

29. The International HapMap Project. Nature. 2003; 426(6968):789–96. Epub 2003/12/20. doi: 10.1038/ nature02168 PMID: 14685227.

30. Saleheen D, Alexander M, Rasheed A, Wormser D, Soranzo N, Hammond N, et al. Association of the 9p21.3 Locus With Risk of First-Ever Myocardial Infarction in Pakistanis Case-Control Study in South Asia and Updated Meta-Analysis of Europeans. Arterioscl Throm Vas. 2010; 30(7):1467–U498. doi: 10. 1161/atvbaha.109.197210 WOS:000278856600026.

31. Lanktree MB, Anand SS, Yusuf S, Hegele RA. Comprehensive analysis of genomic variation in the LPA locus and its relationship to plasma lipoprotein(a) in South Asians, Chinese, and European Caucasians. Circulation Cardiovascular genetics. 2010; 3(1):39–46. Epub 2010/02/18. doi: 10.1161/ CIRCGENETICS.109.907642 PMID: 20160194.

32. Schierer A, Been LF, Ralhan S, Wander GS, Aston CE, Sanghera DK. Genetic variation in cholesterol ester transfer protein, serum CETP activity, and coronary artery disease risk in Asian Indian diabetic cohort. Pharmacogenet Genomics. 2012; 22(2):95–104. Epub 2011/12/07. doi: 10.1097/FPC. 0b013e32834dc9ef PMID: 22143414; PubMed Central PMCID: PMC3269125.

33. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. Epub 2012/11/07. doi: 10.1038/nature11632 PMID: 23128226; PubMed Central PMCID: PMC3498066.

34. Iqbal R, Jahan N, Sun Y, Xue H. Genetic association of lipid metabolism related SNPs with myocardial infarction in the Pakistani population. Mol Biol Rep. 2014; 41(3):1545–52. Epub 2014/01/10. doi: 10. 1007/s11033-013-3000-x PMID: 24402875.

35. Hedrick PW. Genetics of Populations. 4th ed. Sudbury, Mass.; London: Jones and Bartlett Publishers, Inc; 2011.

36. Kathiresan S, Altschuler D, Anand S, Ardissino D, Asselta R, Ball SG, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009; 41(3):334–41. doi: 10.1038/ng.327 WOS:000263640200023. PMID: 19198609

37. Harrison SC, Cooper JA, Li K, Talmud PJ, Sofat R, Stephens JW, et al. Association of a sequence variant in DAB2IP with coronary heart disease. Eur Heart J. 2012; 33(7):881–8. doi: 10.1093/eurheartj/ ehr075 WOS:000302497800020. PMID: 21444365

38. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. New Engl J Med. 2007; 357(5):443–53. doi: 10.1056/ NEMJoa072366 WOS:000248436300005. PMID: 17634449

39. Sarwar N, Sandhu MS, Ricketts SL, Butterworth AS, Di Angelantonio E, Boekholdt SM, et al. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. Lancet. 2010; 375 (9726):1634–9. Epub 2010/05/11. doi: 10.1016/S0140-6736(10)60545-4 PMID: 20452521; PubMed Central PMCID: PMC2867029.

40. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, et al. Genetic Variants Associated with Lp(a) Lipoprotein Level and Coronary Disease. New Engl J Med. 2009; 361(26):2518–28. doi: 10. 1056/NEJMoa0902604 WOS:000273003700007. PMID: 20032323

41. Erdmann J, Grosshennig A, Braund PS, Konig IR, Hengstenberg C, Hall AS, et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet. 2009; 41(3):280–2. doi: 10.1038/ ng.307 WOS:000263640200006. PMID: 19198612

42. Sagoo GS, Tatt I, Salanti G, Butterworth AS, Sarwar N, van Maarle M, et al. Seven lipoprotein lipase gene polymorphisms, lipid fractions, and coronary disease: a HuGE association review and meta-analysis. Am J Epidemiol. 2008; 168(11):1233–46. Epub 2008/10/17. doi: 10.1093/aje/kwn235 PMID: 18922999.

43. Benn M, Nordestgaard BG, Grande P, Schnohr P, Tybjaerg-Hansen A. PCSK9 R46L, low-density lipoprotein cholesterol levels, and risk of ischemic heart disease: 3 independent studies and meta-analyses. J Am Coll Cardiol. 2010; 55(25):2833–42. Epub 2010/06/29. doi: 10.1016/j.jacc.2010.02.044 PMID: 20579540.

44. Bennet AM, Di Angelantonio E, Ye Z, Wensley F, Dahlin A, Ahlbom A, et al. Association of apolipoprotein E genotypes with lipid levels and coronary risk. JAMA. 2007; 298(11):1300–11. Epub 2007/09/20. doi: 10.1001/jama.298.11.1300 PMID: 17878422.