RESEARCH NOTE

Genome-based exome-sequencing analysis identifies *GYG1, DIS3L, DDRGK1* genes associated with myocardial infarction in Korean

JI-YOUNG LEE[1,2†],SANGHOONMOON[1†], YUN KYOUNG KIM[1], SANG-HAK LEE[3], BOK-SOO LEE[4],MIN-YOUNG PARK[5],JEONGEUY PARK[4*], YANGSOO JANG[3*],AND BOK-GHEE HAN[1*]

[1]*Center for Genome Science, Korea National Institute of Health, KCDC, Chungcheongbuk-do 363-951, Korea*

[2]*Cardiovascular Research Institute and Cardiovascular Genome Center, Yonsei University Health System, Seoul 120-752, Korea*

[3]*Cardiology Division, Department of Internal Medicine, Severance Cardiovascular Hospital, Yonsei University College of Medicine, Seoul120-752, Korea*

[4]*Division of Cardiology,Samsung Medical Center, Seoul 135-710, Korea.*

[5]*DNA link, Seoul 121-850, Korea.*

Running title: Exome sequencing and myocardial infarction

Key words: *Combined multivariate and collapsing method,myocardial infarction, exome-sequencing*

†These authors contributed equally to this work

*To whom correspondence should be addressed

Introduction

Myocardial infarction (MI)is a complex disease caused by genetic and environmentalfactors(Kessler *et al.* 2013; Yusuf *et al.* 2004).In the pastdecade, genome-wide

association studies (GWAS) have been applied to identify genetic susceptibility factors for MI andmore than 46 risk lociwhich are associated with coronary artery disease and MI have been identified (Anderson *et al.* 2007; CARDIoGRAMplusC4D Consortium *et al.* 2013; Kessler *et al.* 2013; Peden and Farrall 2011; Schunkert *et al.* 2010). However, the variants identified by GWASexplain only a small fraction of the heritable component of MI risk(CARDIoGRAMplusC4D Consortium *et al.* 2013),mobilizing interest in rare, low-frequency variants that might contribute to the risk of this complex disease (Altshuler *et al.* 2008; Manolio *et al.* 2009). We hypothesized that genetic influences were typically larger in young onset of MI. To identify the geneticvariants that confer susceptibility to MI, we screened the susceptibility loci for MI using exome sequencing and validated candidatevariants in replication sets.

Materials and methods

Subjects

In this study, we used a two-stage approach, which first employed exome sequencing in a series of individuals and a large-scale follow-up genotyping of identified candidate variants in larger case-controls. The first phase of screening included 167MIpatients and 100 control subjects. The diagnosis of MI was based on typical chest pain over 430 min, characteristic electrocardiographic patterns of acute myocardialinfarction, and elevated creatine kinase-MB and troponin I levels. The patients with a familial hypercholesterolemia, known vasculitides, end-stage renal disease and congenital heart disease were excluded from this study.Young onset of MI was defined as age of first diagnosis less than 45 in male, 55 in female, respectively(Malmberg *et al.* 1994; Shah *et al.* 2016).Controls were eligible if they did not have diagnosed subclinical disease assessing anthropometric and biochemical examinations including

ECG.Several studies have been described in the same subjects in detail (CARDIoGRAMplusC4D Consortium *et al.* 2013; Cho *et al.* 2009; Lee *et al.* 2013).The mean (±SD) ages were 42.4 (±7.9) years for thepatients and 49.3(±6.1) years for the control subjects.

At the replication stage, we selected252MIpatients and 1,000 control individuals currently participating in the KoGES to confirm the results of thescreening stage.The252 patientsnot included in discovery stage, and the control individuals are from a large urban cohort. Demographic characteristics, including age, sex, ethnicity,disease history, lifestyle, andmedication information, wereobtained from a detailed questionnaire.Genomic DNAwas extracted from peripheral blood leukocyte pelletsusing a DNA extraction kit (QIAGEN, Valencia, CA, USA).The study protocol was approved by the institutional review boards at the KoreaNational Institute of Health and at each collaborating institute. Informedconsent was obtained from all participants.

Target exon capture and Exome Sequencing

Genomic DNA was extracted from peripheral blood using a QIAamp DNA Blood Mini kit(Qiagen, Hilden, Germany). Exome sequencing was performed by the Agilent SureSelect Human All Exon 50Mb Capture kit and 100-bp paired-end sequencing onto Illumina HiSeq 2000, according to the manufacturer's protocol. Aligning to the reference genome was conducted by the Burrow-Wheeler Aligner (v0.6.1). In addition, removal of duplicate PCR read and recalibrating of quality scores were performed by Picard (v1.6.7) (http://picard.sourceforge.net/), and Genome Analysis Toolkit (v2.3.6), respectively (McKenna *et al.* 2010).Variant calls were retained with a minimum of 20x coverage using GATK genotyper(v2.3.6).An average of 6.93 billion bases of high-quality sequence was generated per individual. Though the targeted capture method enables a high rate of enrichment for the targeted

regions, non-specific capture also occurred. A 52% of aligned reads mapped to the targeted regions of the genome.

Statistical analysis

The association test for annotated variants used Fisher's exact tests to compare the frequencies of variants identified from the sequence data between patients and control subjects. Existing association methods for common variants may not workfor rare variants(Bansal *et al.* 2010). We thereforedeployed the combined multivariate and collapsing (CMC) method(Li and Leal 2008) and C-alpha methods (Neale *et al.* 2011). These methods, provide robust results when dealing with low frequencies and rare variants, whichmight predispose an individual to a given risk (Pan and Shen 2011). In addition, we applied a permutation, to avoid the inflation of type I error rates. All statistics was applied using the PLINK /SEQ (v0.10) (https://atgu.mgh.harvard.edu/plinkseq) and R packages (v 3.2.1) (http://cran.r-project.org).

Results

The demographic and baseline clinical characteristicsareshown in Table 1. Age, body mass index, and HDL-cholesterol differed significantly between MI patients and control individuals. This confirms that the control individuals we selected were very healthy and therefore, any confounding risk factors could be ruled out.

We found 3,420 SNPs with p values of <0.05 byanalyzing differences in allele distribution between the patients and the control subjects. In the gene-based analysis, we selected $p<1\times10^{-5}$(CMC or C-alpha methods) as the cut-off for highly suggestive loci. After ruling out the outliers (number of rare variants (less than 4) or with unknown function), we selected 44 variants to

replicate the association with MI. Using combined analysis of the common and rare variants, we confirmed that 3 gene (*GYG1*,*DIS3L andDDRGK1*) wereassociated with MIat the discovery stage (p<0.05). The results of the statistical analysis of the variants are shown inTable 2. In the analysis at the replication stage, variants in*CYP4A22* did not show a significant association with MI risk (p>0.05).

Discussion

This study is among the first to explore the presence and effect of rare and common variation in early onset MI. We evaluated specific genes chosen on the basis of CMC methods and confirmed that 3 genes(*GYG1*,*DIS3L andDDRGK1*)were associated with MI at the discovery and replication stages.

GYG1, the strongest gene, encodes a member of the glycogenin family. Glycogenin is a glycosyltranferase that catalyzes the formation of a short glucose polymer from uridine diphospate glucose in an authglucosylation reaction. Sequencing of *GYG1*,the glycogenin-1 gene, revealed a nonsense mutation in 1 allele and a missensemutation; (Thr83Met), in the other. The missense mutation resulted in the inactivationof the autoglucosylation of glycogenin-1, which is necessary for priming glycogensynthesis in muscle(Moslemi *et al.* 2010). Furthermore, mutations in genes involved in glycogen metabolism cause glycogenstorage cardiomyopathies(Arad *et al.* 2005)

Notably, *DDRGK1, encodingDDRGK* domain containing 1, is a protein-coding gene associated with thrombocytopenia and hepatitis, which explains MI-related platelet pathogenesis (Ochi *et al.* 2010; Tanaka *et al.* 2011). Previous exome sequencing identified rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction at an early age (Do *et al.* 2015). In addition, recent

consortium study showed that carriers of novel mutations in *ANGPTL4* were associated with protection from myocardial infarction (Myocardial Infarction *et al.* 2016). However, most studies in exome sequencing were conducted among populations of European ancestry suggesting lack of replication in other populations. The size of our study sample limited our ability to detect previously identified variants among European ancestries. Hence, efforts to identify additional variation underlying MI will require much larger study samples. While we include relatively small number of subjects compared to other studies, this is the first evidence of three novel genes (*GYG1, DIS3L,* DDRGK1) detected using gene-based exome sequencing in an Asian population. Due to a lack of additional functional studies, the mechanisms by these genes influence MI pathogenesis remain unknown.

For the road ahead, further research will be required to determine the functional association of these genes with MI risk, and these associations will need to be confirmed in other ethnic populations.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all participants and investigators of the Korea GenomeEpidemiology Study (KoGES). This work was supported by grants from the Korea Centers for Disease Control and Prevention (4845-301), an intramural grant from the Korea National Institute of Health (2012-N73003-00)

References

Altshuler, D., Daly, M. J. and Lander, E. S. 2008 Genetic mapping in human disease. *Science* **322**, 881-888.

Anderson, J. L., Carlquist, J. F., Horne, B. D. and Hopkins, P. N. 2007 Progress in unravelin g the genetics of coronary artery disease and myocardial infarction. *Curr. Atheroscler. Re p.***9**, 179-186.

Arad, M., Maron, B. J., Gorham, J. M., Johnson, W. H., Jr., Saul, J. P., Perez-Atayde, A. R. *et al.* 2005 Glycogen storage diseases presenting as hypertrophic cardiomyopathy. *N. En gl. J. Med.***352**, 362-372.

Bansal, V., Libiger, O., Torkamani, A. and Schork, N. J. 2010 Statistical analysis strategies fo r association studies involving rare variants. *Nat. Rev. Genet.***11**, 773-785.

CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., As simes, T. L. *et al.* 2013 Large-scale association analysis identifies new risk loci for coro nary artery disease. *Nat. Genet.***45**, 25-33.

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H. J. *et al.* 2009 A large-scal e genome-wide association study of Asian populations uncovers genetic factors influencin g eight quantitative traits. *Nat. Genet.***41**, 527-534.

Do, R., Stitziel, N. O., Won, H. H., Jorgensen, A. B., Duga, S., Angelica Merlini, P. *et al.* 2 015 Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myoc ardial infarction. *Nature***518**, 102-106.

Kessler, T., Erdmann, J. and Schunkert, H. 2013 Genetics of coronary artery disease and myo cardial infarction--2013. *Curr. Cardiol. Rep.***15**, 368.

Lee, J. Y., Lee, B. S., Shin, D. J., Woo Park, K., Shin, Y. A., Joong Kim, K. *et al.* 2013 A genome-wide association study of a coronary artery disease risk variant. *Journal of hum an genetics***58**, 120-126.

Li, B. and Leal, S. M. 2008 Methods for detecting associations with rare variants for commo n diseases: application to analysis of sequence data. *Am. J. Hum. Genet.***83**, 311-321.

Malmberg, K., Bavenholm, P. and Hamsten, A. 1994 Clinical and biochemical factors associat ed with prognosis after myocardial infarction at a young age. *Journal of the American C ollege of Cardiology***24**, 592-599.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J. *et al.* 2009 Finding the missing heritability of complex diseases. *Nature***461**, 747-753.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. *et al.* 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation D

NA sequencing data. *Genome research***20**, 1297-1303.

Moslemi, A. R., Lindberg, C., Nilsson, J., Tajsharghi, H., Andersson, B. and Oldfors, A. 201 0 Glycogenin-1 deficiency and inactivated priming of glycogen synthesis. *N. Engl. J. Me d.***362**, 1203-1210.

Myocardial Infarction, G., Investigators, C. A. E. C., Stitziel, N. O., Stirrups, K. E., Masca, N. G., Erdmann, J. *et al.* 2016 Coding Variation in ANGPTL4, LPL, and SVEP1 and th e Risk of Coronary Disease. *The New England journal of medicine***374**, 1134-1144.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M. *et al* . 2011 Testing for an unusual distribution of rare variants. *PLoS Genet.***7**, e1001322.

Ochi, H., Maekawa, T., Abe, H., Hayashida, Y., Nakano, R., Kubo, M. *et al.* 2010 ITPA pol ymorphism affects ribavirin-induced anemia and outcomes of therapy--a genome-wide stud y of Japanese HCV virus patients. *Gastroenterology***139**, 1190-1197.

Pan, W. and Shen, X. 2011 Adaptive tests for association analysis of rare variants. *Genet. Ep idemiol.***35**, 381-388.

Peden, J. F. and Farrall, M. 2011 Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum. Mol. Genet.***20**, R198-205.

Schunkert, H., Erdmann, J. and Samani, N. J. 2010 Genetics of myocardial infarction: a progr ess report. *Eur. Heart J.***31**, 918-925.

Shah, N., Kelly, A. M., Cox, N., Wong, C. and Soon, K. 2016 Myocardial Infarction in the "Young": Risk Factors, Presentation, Management and Prognosis. *Heart, lung & circulatio n***25**, 955-960.

Tanaka, Y., Kurosaki, M., Nishida, N., Sugiyama, M., Matsuura, K., Sakamoto, N. *et al.* 2011 Genome-wide association study identified ITPA/DDRGK1 variants reflecting thrombocyto penia in pegylated interferon and ribavirin therapy for chronic hepatitis C. *Hum. Mol. Ge net.***20**, 3507-3516.

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F. *et al.* 2004 Effect of po tentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet***364**, 937-952.

Received 4 October 2016, in revised form 1 January 2017; accepted 17 February 20017
Unedited version published on 22 February 2017

Table 1. Baseline clinical attributes of the study subjects

	Case (N=167)	Control (N=100)	*p*-value
Age, years	42.4 (7.9)	49.3 (6.1)	<.0001
Female (%)	21.6	60	<.0001
Hypertension (%)	97.6	0	<.0001
BMI (kg/m2)	25.4 (4.2)	21.4 (1.2)	<.0001
Total cholesterol (mg/dl)	190.4 (56.6)	182.7 (30.5)	<.0001
HDL-cholesterol (mg/dl)	42.3 (15.0)	48.3 (11.1)	<.0001
LDL-cholesterol (mg/dl)	116 (43.4)	110.2 (27.1)	<.0001
Triglycerides (mg/dl)	177.0 (184.8)	120.7 (50.9)	<.0001

Data are presented as mean ± SD or (%).

Table 2. Gene-based statistical significance analyses at the discovery and replication stages

						Discovery				Replication			
Gene	**CHR**	**BP**	**REF**	**ALT**	**SNP[1)]**	**No. of SNP**	**No. of rare SNP†**	***p*-value[2)]**	***p*-value[3)]**	**No. of SNP***	**No. of rare SNP†**	***permutated p*-value[2)]**	***permutated p*-value[3)]**
CYP4A22	1	47603172	C	G	47603172*	20	13	8.84×10^{-3}	2.18×10^{-8}	11	6	0.163671	0.170592
		47603188	C	T	rs76011927								
		47603297	C	G	47603297								
		47603312	C	T	47603312								
		47606476	C	T	47606476								
		47609492	A	G	rs191687330*								
		47609531	C	A	47609531								
		47610065	A	C	rs149718343*								
		47610100	C	T	47610100*								
		47610221	G	A	47610221*								
		47610274	C	T	47610274*								
		47610275	G	A	47610275								
		47610293	C	T	rs968322*								
		47610590	A	C	rs2405593*								
		47611765	G	A	rs150794228*								
		47611833	G	A	47611833*								
		47614376	C	A	47614376								
		47614379	C	A	rs185005287								
		47614434	C	T	rs4926600								
		47614485	A	G	rs72684327*								
GYG1	3	148711957	C	T	148711957*	8	6	2.88×10^{-3}	1.15×10^{-6}	5	3	$< 5 \times 10^{-6}$	$< 5 \times 10^{-6}$
		148712028	T	C	148712028*								
		148714060	G	A	rs143942810								
		148714146	C	T	rs149479866*								
		148714238	G	T	148714238								
		148727127	G	A	148727127								
		148727133	G	A	rs4938*								

		148744731	T	C	148744731*								
DIS3L	15	66604032	C	T	rs28616181	11	5	3.16×10^{-4}	4.52×10^{-5}	7	1	$< 5 \times 10^{-6}$	0.001483
		66610984	G	A	rs141774650*								
		66612965	T	C	rs17851970*								
		66618143	T	C	66618143								
		66618308	G	A	rs141173452								
		66618342	A	G	rs3803412								
		66618345	A	C	66618345*								
		66618663	G	A	rs117749912*								
		66621347	C	T	66621347*								
		66625161	A	G	rs11071885*								
		66625470	G	A	rs3759785*								
DDRGK1	20	3175957	C	T	3175957	5	4	5.09×10^{-5}	1.43×10^{-5}	2	1	0.46987	0.004
		3175991	C	T	3175991*								
		3176009	T	C	rs2295552*								
		3180666	C	A	rs35327491								
		3180717	C	T	3180717								

1) Novel SNPs were notified by base position (GRCh37.p10), 2) P-value were calculated by Combined Multivariate and Collapsing Method (CMC)

3) P-value were calculated by C-alpha Score Test (C-ALPHA), †minor allele frequency<1%, *Successfully replicated SNP

CHR; chromosome, BP; Base position, REF; Reference SNP , ALT; Alternative SNP, CYP4A22; Cytochrome P450, Family 4, Subfamily A, Polypeptide 22, GYG1; Glycogenin 1, DIS3L; DIS3 like exosome 3'-5' exoribonuclease, DDRGK1; DDRGK domain containing 1