Online-only Data Supplement: Materials and Methods

Dissecting the Roles of microRNAs in Coronary Heart Disease via Integrative Genomics Analyses

Tianxiao Huan[1,2], Jian Rong [3], Kahraman Tanriverdi[4], Qingying Meng[5], Anindya Bhattacharya[5], David D. McManus[4], Roby Joehanes[1,2,6,7,8], Themistocles L. Assimes[9], Ruth McPherson[10], Nilesh J. Samani[11,12], Jeanette Erdmann[13, 14], Heribert Schunkert[15,16], Coronary ARteryDIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) Consortium, Paul Courchesne[1,2], Peter J. Munson[8], Andrew D. Johnson[1,17], Christopher J. O'Donnell[1,17], Bin Zhang[18,19], Martin G. Larson[3], Jane E Freedman[4]*, Daniel Levy[1,2]* and Xia Yang[5]*

[1]The National Heart, Lung, and Blood Institute's Framingham Heart Study, 73 Mt. Wayte Avenue, Framingham, MA, USA; [2]The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, MD, USA; [3]Department of Mathematics and Statistics, Boston University, Boston, MA, USA; [4]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA; [5]Department of Integrative Biology and Physiology, University of California, Los Angeles, CA, USA; [6]Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health, Bethesda, MD, USA; [7]Harvard Medical School, Boston, MA, USA; [8]Hebrew SeniorLife, Boston, MA, USA; [9]Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, USA; [10]Departments of Medicine and Biochemistry, University of Ottawa, Ottawa, Canada; [11]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK; [12]NIHR Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, UK; [13]Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, Lübeck, Germany; [14]DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany, [15]Deutsches Herzzentrum München, Technische Universität München, 80636 München, Germany; [16]DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany; [17]Cardiovascular Epidemiology and Human Genomics Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD, USA; [18]Department of Genetics and Genomic Sciences, [19]Institute of Genomics and Multiscale Biology, New York, NY;

Running title: *Huan et al.; Integrative analysis of miRNAs in CHD*

Correspondence should be addressed to:

Jane E. Freedman, MD
University of Massachusetts Medical School
AS7-1051, 368 Plantation St.
Worcester, MA 01605-4319 USA
Email: jane.freedman@umassmed.edu
Phone: (508) 856-6961

Daniel Levy, MD
Framingham Heart Study
Center for Population Studies
National Heart, Lung, and Blood Institute
73 Mt. Wayte Avenue, Suite 2
Framingham, MA 01702
Email: Levyd@nih.gov
Phone: 508-935-3458
Fax: 508-872-2678

Xia Yang, PhD
Department of Integrative Biology and Physiology
University of California, Los Angeles
Los Angeles, CA 90095
Email: xyang123@ucla.edu
Phone: 310-206-1812
Fax: 310-206-9184

Materials and Methods

CHD Case Control Study

We designed a CHD case-control study with CHD cases and controls matched based on age and statin use from Framingham Heart Study (FHS) participants. Peripheral whole blood samples were collected and RNA was isolated for mRNA and miRNA profiling as described previously[1, 2].

In two recent reports[1, 2], we explored differential mRNA expression and derived gene networks associated with CHD in 188 case-control pairs. Among the 188 pairs, 186 pairs had successful miRNA profiling and these 186 pairs were used to investigate miRNA-mRNA associations in subsequent analyses. Of these, 176 cases and 185 controls had genome-wide genotype data[3] and data from these individuals were used in the SNP-miRNA association or expression quantitative trait loci (eQTL) analyses. This study was approved under Boston University Medical Center's Institutional Review Board protocol H-27984. Informed consent was obtained from each participant.

miRNA expression profiling

We used quantitative polymerase chain reaction (qRT-PCR) based TaqMan miRNA assays to conduct miRNA expression profiling of 754 miRNAs. TaqMan miRNA assays have been previously demonstrated with sufficient specificity, reproducibility and sensitivity [4-9]. Fasting peripheral whole blood samples (2.5ml) were collected in PAXgene Blood RNA™ tubes (Qiagen, Valencia, CA) and frozen at –80^{o}C during FHS offspring cohort examination 8 (2005-2008). Total RNA was isolated from the frozen PAXgene Blood RNA tubes (Asuragen, Inc. Austin, TX) and a 2100 Bioanalyzer Instrument (Agilent, Santa Clara, CA) was used for RNA quality assessment. Isolated RNA samples were converted to complementary DNA (cDNA) using TaqMan miRNA Reverse Transcription Kit and MegaPlex Human RT Primer Pool Av2.1 and Pool Bv3.0. (Life Technologies, Foster City, CA) using TaqMan assays in 2 panels for a total of 754 miRNAs attempted. The cDNA samples were pre-amplified using TaqMan PreAmp Master Mix and PreAmp Primers, Human Pool A v2.1 and Pool B v3.0 (Life Technologies, Foster City, CA). We routinely ran blank spaces and, as in prior experiments showing excellent reproducibility using the Biomark dynamic array platform in conjunction with multiplexed reverse transcriptase reactions for miRNA profiling, we did not encounter cross-contamination with this platform. The BioMark system is quantitative for low abundance miRNAs [9].

Measurements were completed using quantitative RT-PCR and, as such, threshold cycle (Ct) values were used to evaluate the miRNA expression by counting the number of amplification cycles required for the fluorescent signal to exceed the background level. Lower Ct values indicated higher expression levels of miRNA. As PCR cycle over 30 is commonly considered over the linear range which may cause saturation of PCR product of normal expression levels of miRNA and make it not accurate for quantitative comparison, miRNAs with Ct a value greater than 30 were considered to show no expression in our experiment. All data used herein are available online in dbGaP (http://www.ncbi.nlm.nih.gov/gap, accession number phs000007).

miRNA expression data normalization

We considered different models for miRNA data normalization through variance analysis of general linear models with miRNA raw Ct as response variable. First, in Model 1, we ran general linear model on miRNA raw Ct values using mean Ct values of the top 50 ubiquitously expressed miRNAs as independent variable (**Supplementary Figure 3A**). Then we ran Model 2, by adjusting for 4 technical variables, namely isolation batch, RNA concentration, RNA quality (defined as RNA integrity number

[RIN]), and 260/280 ratio (defined as the ratio of the absorbance at 260 and 280nm; measured using a spectrophotometer). **Supplementary Figure 3A** and **3B** show the distribution of the R-squares for every miRNA from Model 1 and Model 2, respectively. Our results indicated that adjusting for technical variables (Model 2) gave rise to more stable model R-squares (indicating the variance proportion of measured miRNA expression values that explained by tested variables): ~80% miRNAs have R-squares between 0.2-0.6 in Model 2, giving a reasonable amount of adjustment for most miRNAs, whereas Model 1 showed a dichotomized normalization pattern with 25% and 50% miRNAs having R-squares <0.052 and 0.6-0.8, respectively. Therefore, we used Model 2 for miRNA data normalization and kept residuals for further analysis.

Imputing Cell Counts

The cell count proportions of whole blood were only measured in 2,138 FHS individuals in the FHS third generation cohorts, but not in the samples used in this study. We found that cell counts could be accurately imputed using a Partial Least Squares regression method[10]. The estimated cell count proportion values imputed were highly consistent with the measured values in the FHS third generation cohorts with cross-validated estimates of prediction accuracy (r^2) for white blood cell, red blood cell, platelet, lymphocyte percent, monocyte percent, eosinophil percent and basophil percent being 0.61, 0.41, 0.25, 0.83, 0.81, 0.89, and 0.25, respectively. We therefore estimated the cell count proportions in the 372 samples in the current study using the same method.

Identification of differentially expressed miRNA signatures in CHD

Of the 754 miRNAs assayed, 271 were found to be expressed (Ct<30) in more than 50 samples and were used for further analysis. For these 271 miRNAs, we used relative miRNA expression (Cr) values defined as Cr=30-Ct. Differentially expressed individual miRNA signatures between CHD cases and controls were identified by a conditional logistic regression model by conditioning on matched pairs and accounting for diabetes and technical covariates, using the *clogit()* R package (http://cran.r-project.org/web/packages/survival/). False discovery rate (FDR) was calculated using the Benjamini-Hochberg (BH) method[11].

Identification of SNPs associated with miRNA expression

Approximately 550,000 SNPs were genotyped using the Affymetrix 500K mapping array and Affymetrix 50K gene-focused MIP array. Quality control procedures and genotype imputation to a set of approximately 2.5 million HapMap SNPs were described in detail previously[3].

Linear regression was used to determine the association between the age-, sex-, CHD status-, and technical covariate-adjusted miRNA expression values and the imputed SNP genotypes, yielding results for approximately 271 miRNAs x 2.5 million SNPs. SNPs were filtered at minor allele frequency (MAF) <0.05 based on the genotype data. Associated SNP-miRNA pairs that reside within 1Mb of each other (referred to as *cis*-miR-eSNPs) were identified. The significant threshold was chosen at BH[11] corrected FDR<0.05 (corresponding p=8.7e-6). As there were missing values for miRNA, we further filtered out miR-eSNPs with MAF <0.05 in samples with both miRNA and genotype data available.

miRNA and mRNA correlation analysis

The Affymetrix Human Exon 1.0 ST array was used to quantify mRNA expression levels of about 18,000 transcripts genome-wide in peripheral whole blood from CHD cases and controls, as described previously[1, 2]. miRNA-mRNA associations were assessed by Pearson correlation of miRNA and mRNA expression residuals, each adjusted for age-, sex-, and technical covariates. Technical covariates included isolation batch, RNA concentration, RNA quality (defined as RNA integrity number [RIN]), and 260/280

ratio (defined as the ratio of the absorbance at 260 and 280nm; measured using a spectrophotometer). As both mRNA and miRNA data were pre-adjusted for covariates, the correlations reported in the current study are partial correlations. The BH method[11] was used to calculate FDR values and the threshold for significance was set to FDR<0.05. The miRNA-mRNA correlation analysis was conducted separately in cases, controls, and in all samples adjusting for CHD status.

The co-expression patterns in cases and controls were replicated using separate sets of CHD cases (n=63) and controls (n=1000) in the FHS, whose miRNA and mRNA expression were measured using the same platforms and protocols as for the CHD case-control study.

Identification of differential miRNA-mRNA co-expression pairs between CHD cases and controls

The CHD-related differential miRNA-mRNA co-expression pairs were defined as the miRNA-mRNA pairs co-expressed in either CHD cases (defined as case-specific pairs) or in controls (control-specific pairs) but not in both at FDR <0.05 and that showed an interaction in relation to CHD status at FDR <0.2 as described below.

To capture potential miRNA-mRNA interactions that were associated with CHD status, we used the following logistic regression model:

$$\mathrm{CHD}_{\mathrm{status}} = \mu + \beta_1\,(\mathrm{miRNA}_{\mathrm{resid}}) + \beta_2(\mathrm{mRNA}_{\mathrm{resid}}) + \beta_3\,(\mathrm{miRNA}_{\mathrm{resid}} \times \mathrm{mRNA}_{\mathrm{resid}}) + \varepsilon$$

where $\mathrm{miRNA}_{\mathrm{resid}}$ and $\mathrm{mRNA}_{\mathrm{resid}}$ denote the residuals of each miRNA and mRNA after adjustment for age, sex and technical covariates; μ was the overall trait mean; β_1 and β_2 were the regression coefficients of $\mathrm{miRNA}_{\mathrm{resid}}$ and $\mathrm{mRNA}_{\mathrm{resid}}$, respectively; β_3 was the regression coefficients of the $\mathrm{miRNA}_{\mathrm{resid}}$ and $\mathrm{mRNA}_{\mathrm{resid}}$ interaction; ε was the residual error. The p-value of β_3 determines whether the miRNA and mRNA interaction is associated with CHD status. An FDR<0.2 for the β_3 p-values for the miRNA-mRNA interaction pairs was considered significant.

Testing the putatively causal relationship between CHD-related gene sets and CHD

We used SNP set enrichment analysis [12] to test if any of the CHD gene sets are putatively causal for CHD by showing enrichment with low p value CHD SNPs from the CARDIoGRAM GWAS[13]. For the CHD miRNA differential expression signatures, we linked the miRNAs with CHD GWAS results via miR-eSNPs. For the CHD case-specific (or control-specific) miRNA-mRNA co-expression pairs, we linked the findings to CHD GWAS results using both miR-eSNPs and mRNA eSNPs[14-17]. We used Fisher's exact test and the Kolmogorov-Smirnov (KS) test to evaluate if these CHD gene sets were enriched for either miR-eSNPs or mRNA eSNPs, displaying low p-value CHD associations in the CHD GWAS[13]. If a gene set was significant at p<0.017 (Bonferroni correction for three gene sets) by either Fisher's exact or the KS tests, and p<0.05 by the other test, we considered the gene set to be putatively causal for CHD.

Additional blood miR-eSNPs

Our group currently is conducting a miRNA eSNP project from whole blood of 5,329 individuals. Linear mixed regression model was used to determine the association between the age-, sex-, technical covariate- and familiar relatedness- adjusted miRNA expression values and the imputed SNP genotypes from the 1000-genome resource, yielding results for approximately 271 miRNAs x 10 million SNPs. This ongoing project (unpublished) generated 5269 *cis*-miR-eSNPs and 270 *trans*-miR-eSNPs at FDR<0.1, with all of the *cis*-miR-eSNPs from the CHD case-control study replicated, supporting the reliability of the *cis*-miR-eSNPs. The *cis*-miR-eSNPs from this ongoing larger miRNA study were used to complement the miR-

eSNPs identified in the CHD case-control study to validate the results of the SNP set enrichment analysis (see details in "Testing the causal relationship between CHD-related gene sets and CHD") of the CHD-related miRNA sets.

Comparison of miRNA-mRNA co-expression pairs with miRNA target databases

We used known or predicted miRNA-mRNA target pairs collected from six publicly available databases [18-23]. Among these six databases, miRTarBase [18], TarBase [19], miRecords [20] and miR2Disease [21] annotated experimentally validated miRNA-mRNA pairs, whereas miRecords [20], miRDB [22] and MicroRNA.org [23] reported computationally predicted miRNA-mRNA pairs by utilizing more than 10 miRNA target predicting algorithms. The predicted miRNA-mRNA target pairs based on these databases were compared with the miRNA-mRNA co-expression pairs identified from the current study for overlaps.

Pathway and gene ontology enrichment analysis

Each CHD gene set identified in this study was classified using Gene Ontology (GO) [24] databases to identify biological processes potentially involved in CHD. Fisher's exact test was used to calculate enrichment p values and results were corrected by the 825 unique GO biological process terms, yielding a p value significance threshold of 0.05/825=6.0e-5.

References

1. Huan T, Zhang B, Wang Z, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013;33:1427-1434
2. Joehanes R, Ying S, Huan T, et al. Gene expression signatures of coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013;33:1418-1426
3. Levy D, Ehret GB, Rice K, et al. Genome-wide association study of blood pressure and hypertension. *Nat Genet*. 2009;41:677-687
4. Chen C, Tan R, Wong L, Fekete R, Halsey J. Quantitation of microrna s by real-time rt-qpcr. *Methods Mol Biol*. 2011;687:113-134
5. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ. Real-time quantification of microrna s by stem-loop rt-pcr. *Nucleic Acids Res*. 2005;33:e179
6. Schmittgen TD, Lee EJ, Jiang J, Sarkar A, Yang L, Elton TS, Chen C. Real-time pcr quantification of precursor and mature microrna. *Methods*. 2008;44:31-38
7. Mestdagh P, Feys T, Bernard N, Guenther S, Chen C, Speleman F, Vandesompele J. High-throughput stem-loop rt-qpcr mirna expression profiling using minute amounts of input rna. *Nucleic Acids Res*. 2008;36:e143
8. Jensen SG, Lamy P, Rasmussen MH, Ostenfeld MS, Dyrskjot L, Orntoft TF, Andersen CL. Evaluation of two commercial global mirna expression profiling platforms for detection of less abundant mirnas. *BMC genomics*. 2011;12:435
9. Jang J, Simon V, Feddersen R, Rakhshan F, Schultz D, Zschunke M, Lingle W, Kolbert C, Jen J. Quantitative mirna expression analysis using fluidigm microfluidics dynamic arrays. *BMC genomics*. 2011;12:144
10. Boardman AE, Hui BS, Wold H. The partial least squares-fix point method of estimating interdependent systems with latent variables. *Communications in statistics-theory and methods*. 1981;10:613-639
11. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995:289-300
12. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *American journal of human genetics*. 2010;86:581-591
13. Schunkert H, Konig IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*. 2011;43:333-338
14. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;452:423-428
15. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*. 2008;6:e107
16. Yang X, Zhang B, Molony C, et al. Systematic genetic and genomic analysis of cytochrome p450 enzyme activities in human liver. *Genome research*. 2010;20:1020-1036
17. Huan T, Zhang B, Wang Z, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013
18. Hsu SD, Lin FM, Wu WY, et al. Mirtarbase: A database curates experimentally validated microrna-target interactions. *Nucleic Acids Res*. 2011;39:D163-169
19. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG. Tarbase 6.0: Capturing the exponential growth of mirna targets with experimental support. *Nucleic Acids Res*. 2012;40:D222-229
20. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. Mirecords: An integrated resource for microrna-target interactions. *Nucleic Acids Res*. 2009;37:D105-110

21. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. Mir2disease: A manually curated database for microrna deregulation in human disease. *Nucleic Acids Res*. 2009;37:D98-104
22. Chen Y, Zhu J, Lum PY, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008;452:429-435
23. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microrna.Org resource: Targets and expression. *Nucleic Acids Res*. 2008;36:D149-153
24. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000;25:25-29