FUNCTIONAL GENOMICS OF CARDIOVASCULAR DISEASE RISK

A Thesis
Presented to
The Academic Faculty

by

Jinhee Kim

In Partial Fulfillment
of the Requirements for the Degree
Master in the
School of Biology

Georgia Institute of Technology
August 2013

FUNCTIONAL GENOMICS OF CARDIOVASCULAR DISEASE RISK

Approved by:

Dr. Dr. Greg Gibson, Advisor
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. Melissa Kemp
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: May 17, 2013

To my family and friends

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my advisor Dr. Greg Gibson for his guidance, encouragement and support. He has been a generous supervisor, a heartening teacher and a thoughtful friend to me. He enabled me to successfully complete my thesis. I could not have asked for a better mentor, teacher and guide. I would like to thank my committee members, Dr. Melissa Kemp, Dr. Fredrik Vannberg, and Dr. Soojin Yi for their guidance and feedback on my research. This thesis could never have been completed without their encouragement.

I would like to thank all my lab members, especially great thanks to our lab manager, Dalia Arafat. She has been a good friend of mine all the time. I am also grateful to Linda Kippner for devoting her time for providing valuable comments.

My special thanks to my family. I would not be where I am today without their prayers, love, and supports. The members of New Church of Atlanta, so many to mention, has given me strength and comfort, especially grateful to Namin Jeong, Eunyoung Ahn, Youngjoo Lee, and Seungmin Lee.

Finally, I would like to thank my wonderful husband Mark Hongchul Sohn being by my side throughout this process. You are the answer to my prayer. I am blessed and I thank God for leading me along with your hands.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Understanding variability of heath status is highly likely to be an important component of personalized medicine to predict health status of individuals and to promote personal health. Evidences of Genome Wide Association Study and gene expression study indicating that genetic factors affect the risk susceptibility of individuals have suggested adding genetic factors as a component of health status measurements. In order to validate or to predict health risk status with collected personal data such as clinical measurements or genomic data, it is important to have a well-established profile of diseases.

The primary effort of this work was to find genomic evidence relevant to coronary artery disease. Two major methods of genomic analysis, gene expression profiling and GWAS on gene expression, were performed to dissect transcriptional and genotypic fingerprints of coronary artery disease. Blood-informative transcriptional Axes that can be described by 10 covariating transcripts per each Axis were utilized as a crucial measure of gene expression analysis.

This study of the relationship between gene expression variation and various measurements of coronary artery disease delivered compelling results showing strong association between two transcriptional Axes and incident of myocardial infarction. 244 transcripts closely correlated with death by cardiovascular disease related events were also showing clear association with those two transcriptional Axes. These results suggest potential transcripts for use in risk prediction for the advent of myocardial infarction and cardiac death.

CHAPTER 1

INTRODUCTION

The growing interest in the well-being of modern people has brought on a new era in the health care industry. Researchers accept that we are facing challenges in making the transition from treating illness to promoting health [1, 2]. Understanding variability of heath status is highly likely to be an important component of personalized medicine to predict health status of individuals and to promote personal health. Health status is often assessed by clinical risk measurements including BMI, hypertension, high cholesterol, and diabetes along with several dimensions such as emotional well-being or physical environment [3-5]; however, evidence of Genome Wide Association Study indicating that genetic factors affect the risk susceptibility of individuals have suggested adding genetic factors as a component of health status measurements [6-8]. In order to validate or to predict health risk status with collected personal data such as clinical measurements or genomic data, it is important to have a well-established profile of diseases. Numerous genome-wide association studies (GWAS) that have been performed since 2007 have discovered genetic causes of many kinds of disease including myocardial infarction, rheumatoid arthritis, and diabetes. However, it remains a challenge to develop and standardize complex disease profile networks with various types of data, and to use these to practically enumerate predictive genomic markers, which can be clinically applicable for a specific disease. Interactive and organic health predictive factors utilizing clinical, metabolomic, and genomic data are desired to promote prognostic and therapeutic strategies.

Cardiovascular disease has been studied for decades, and its long-term risk factors are well known [9, 10]. Framingham risk scores is an established method utilized to predict the risk of getting heart disease such as cardiovascular disease, coronary heart

disease, or stroke. Limitations of Framingham risk scores have been observed for with regards to the accuracy that it provides in reliable heart related disease risk estimates [11, 12]. In order to improve the accuracy of predictive risk score for heart disease, numerous studies have been performed, adding more predictive risk factors such as coronary artery calcium score, ankle brachial index or C-reactive protein [13-15]. Despite considerable efforts, the current list of risk factors predictive of subsequent death and occurrence of acute myocardial infarction is still limited.

Preininger et al.[16] evaluated whether clinical functions explain gene expression and observed 9 blood-informative Axes from total gene expression of the healthy cohorts. Those 9 Axes showed trends representing significant biological functions relating to immune status. Variation in quantitative measures along these Axes suggests that individuals can be discriminated with respect to their immune health. This approach will give informative and valuable resources for realization of personalized medicine or treatment [17-19].

Premises in this study were that different health statuses also impact peripheral blood genomic profiles, and these can be discriminated by 9 blood-informative Axes. In addition to testing these hypotheses, how genotype and gene expression are significantly associated, and whether they solely or jointly modulate disease risk factors were evaluated. In order to approach these questions, GWAS on gene expression was performed to evaluate how genetic factors lead to abnormalities of transcript abundance that are associated with cardiac disease and the regulatory polymorphism (eSNP) profiles between patients with cardiovascular disease and adults with no disease were examined to study variance of disease status in the cardiovascular disease cohort from the Emory Cardiology Genebank. The cardiovascular disease cohort from the Emory Cardiology Genebank contains 192 participants in age over 65 years (CDGY1) and an additional replication sample of 163 younger CAD patients (CDGY2). Each phase of cohorts include 43 participants and 50 participants as controls respectively. eSNP profiles

differences were contrasted between CAD patients and the control group and genomic predictors were generated based on blood-informative transcriptional Axes, gene expression and/or regulatory genotypes, for myocardial infarction (MI) and death by cardiac disease. In the course of this study, I asked in what proportions, genetic and environmental effects contribute to divide a population into distinct positions in immune space and how these immune spaces explained by Axes distinguish different statuses of cardiac disease from health.

While profiling the cardiovascular disease cohort of the Emory Cardiology Biobank, strong evidence of genomic variance among different classes of cardiovascular disease was found. This finding will open up a new possibility to identification of peripheral blood gene expression and regulatory genotype profile that is correlated with MI along with comparison profiles of CAD and normal individuals.

1.1 Gene Expression Study

Peripheral blood gene expression is influenced by complex interactions of environmental factors along with genetic factors at baseline [20, 21]. Gene expression profiling has long been studied and is still used as a valuable tool for finding relationship between health and disease or phenotypic traits [22, 23]. Gene expression is potentially more informative than clinical measurements or genotypes itself, since it represents status of molecular mechanisms, such as level of triglyceride, cholesterol, or glucose, that can be attributed to health related traits. It is also influenced by different life-style of individuals [24]. Studies in the Gibson Lab at Georgia Tech have indicated that lifestyle, gender, ethnicity, and geographic location affect gene expression profile [25, 26]. Two previous studies in the Gibson lab, the Morocco study and the Brisbane study, showed gene expression clusters by geographic location between rural and urban populations [25], as well as variation within the urban populations studied [27]. Other authors have shown some evidence that the social condition of different neighborhoods can influence

one metric of physical phenotype, BMI, at significant p-value <0.01 in urban Canada [28] and at p-value <0.03 in the US [29]. Adults with low childhood socioeconomic status (SES) and adolescents in low SES showed significantly elevated immune markers such as IL5, IL6, and IFN-ɣ ($p<0.05$) [30, 31]. Miller et al. also observed the defensive gene expression in respect to resistance to glucocorticoid signaling from subjects with low-SES background [32] . While, these studies provide evidences of how informative gene expression is, gene expression is not only influenced by environmental factors, but also by data sampling criteria, technical effects, and data normalization methods [33, 34]. Since microarray data can be also easily biased by biological factors including age, ratio of genders, and ethnicity, both technical and biological effects were addressed in the study.

Peripheral blood mononuclear cells (PBMC) also include molecular signatures of human disease, and the gene expression of these cells has been successfully utilized in efforts to dissect the complexity of disease on a molecular level. Achiron and Gurevich reviewed peripheral blood gene expression in a model of multiple sclerosis (MS) and concluded that PMBC gene expression can be used as MS disease transcriptional fingerprints and that this will lead to better understanding of the mechanism underlying disease and will enable the evaluation of transcriptional biomarkers for diagnostics and prediction of clinical outcome [35]. Numerous studies have investigated gene expression in the context of heart disease. For example, cardiac markers such as GATA4, MEF2C, Nkx2.5/Csx revealed increased expression of mRNA up to 3.5 fold [36] and 160 genes correlated with the severity of coronary artery disease were found and their partial least square multivariate regression model showed statistically high significance in prediction [37] .

Preininger et al. observed 9 transcriptional Axes from total peripheral blood gene expression in a healthy cohort from the Center for Health Discovery and Well Being

(CHDWB) in Atlanta. Furthermore, CHDWB healthy subjects tend to fall into 8 k-means clusters by 9 modules. 9 modules are representative of T cells, platelets, B cells, cell machinery, TLR signals/interferon, and antiviral (Table 1) [16]. These findings indicate the importance of studying co-regulation factors in each module and of evaluating the results in order to find associations with cardiac disease risks.

Table1. Summary and Functional Interpretation of Blood-Informative Transcriptional Axes. This table is encapsulated by Marcela Preininger et al. [16]

Axis	Number of genes	Keyword
1	866	T-cell signaling
2	237	Platelets
3	99	B-cell signaling
4	982	Post-translational modification
5	1028	Toll-like receptor signaling pathway
6	550	RNA binding
7	169	Viral response
9	571	RNA procession B cell activation
10	242	Signal transduction by phosphorylation

1.2 Genetics of Gene Expression Study

Gene expression shows familiar aggregation and segregation patterns in humans, suggesting a genetic inheritance component of gene expression [38]. In general, despite the effect of many environmental factors such as lifestyle on the levels of gene expression, the trend of genetic inheritance of gene expression is relatively consistent, however study replication rate at the same significant cutoff has not been achieved above 20% [39]. Whether these difficulties of replication are due to power issues or biology is not yet clearly identified [40], but emphasis on trends of association among different phases of studies is remarkably informative and sufficient to show replicability.

GWAS applied gene expression is a powerful tool that has been used to dissect genetic regulation of individual genes both in a healthy population and in a cardiovascular disease cohort [41, 42]. In a recent publication, Rotival et al. found 11 clusters of expression which are influenced by SNPs ($p < 1.15 \times 10^{-9}$) from 1,490 healthy Europeans. The variability of the pattern was explained in range between 1.9% and 24.8% by the lead SNPs and they found 5 patterns that were associated with different cell types. They also replicated 6 associations, which were independent of cell types, with 758 subjects in the Cardiogenic study [43]; however, they only reported associations with SNPs of conserved modules and did not report the association with disease. In this study, the association of eSNPs with cardiovascular disease was proposed as a means to unravel abnormal gene expression patterns at specific genotypes by studying patterns of co-expression and normal genetic regulation. New strategies of profiling genetic association of gene expression in cardiovascular disease and generating a genetic risk score utilizing abnormal genetic regulation will help to make predictions of individual risk of development of abnormal cardiovascular status.

CHAPTER 2

METHODS

The major purpose of this paper is to use microarray gene expression data and whole genome genotype data to generate gene expression profiling and regulatory polymorphism profiling and compare profiles of cardiovascular disease patients with participants with no disease. The gene expression profile of cardiovascular disease will be evaluated with second phase of cohort.

2.1 Sample acquisition and Processing

Sample Acquisition

The cardiovascular disease cohort 1 consists of 192 participants recruited from the Emory Cardiology Genebank, a registry of patients undergoing cardiac catheterization. Out of 192 participants, 17 non Caucasians were excluded to address ethnicity effects on eQTL analysis. Subjects are ranged in age from 41 to 85, with mean of 67, and grouped into four different status, patients with stable CAD defined as >50% stenosis in one or more coronary vessels, angiographically normal, experiencing an acute MI event, and had a history of MI, plus follow-up information for occurrence of future adverse death and MI. These follow-up results were gathered between 2years and 5 years after the day of sample collection. These four classes of status will be designated by CAD, FINE, AcuteMI, and OldCAD, and called CAD status. Each class has approximated equal number of participants. 163 younger CAD patients who are all Caucasians were added to the cohort to replicate study. The second cardiovascular disease cohort includes subjects who are ranged in age from 26 to 83, with mean of 56. Participants are classified according to definition as described with the first cohort. Hereinafter the first cohort is referred as CDGY1(CarDioloGY Phase 1) and CDGY2(CarDioloGY Phase 2). CDGY

cohorts are representative of Atlanta. Biophysical and clinical data of the cohorts are indicated in the Table1.

Sample Processing

All participants of CDGY cohorts donated peripheral blood in Paxgene tubes for gene expression and genotype data over a five year period. CDGY1 and CDGY2 were processed 12 months apart. Gene expression data was generated using hybridization of labeled RNA to Illumina HT-12 bead arrays with probes for all human genes. We consider log base 2 transformed 14,343 probes which are consistently detected across multiple datasets of peripheral blood samples. Genotyping was performed with Illumina OmniExpress SNP chip which has over 730,000 genetic markers. After quality filtering, there were about 610,000 markers retained in the data set. Theses markers are common variants, MAF larger than 0.05.

2.2 Data normalization

Data normalization was a central process for this project. The recent published paper by Qin et al. [34] strongly implied that different normalization methods can crucially affect data. Failure in appropriate normalization can conceal biologically meaningful signals and also generate false positives by technical biased components. In addition, stability and replicability of normalization in other cohorts are also essential. Appropriate data normalization is thus a core issue in gene expression analysis.

Dr. John Storey recently published about new efficient normalization method called Supervised Normalization of Microarrays (SNM). His team showed that SNM method successfully separate biologically meaningful signal from unavoidable technical factors compared with other broadly used normalization methods [44] such as mean or median centered normalization, quantile normalization such as IQR, or rank ordering

normalization . CDGY1 and CDGY2 gene expression data were normalized with SNM method to remove two technical effects, batch and RNA quality, and to adjust for effects of gender, ethnicity, and BMI. Biological variable of interest, four classified CAD status, were statistically adjusted for optimizing biological signals and reducing noise. Supervised normalization of Microarray (SNM) method is implemented in the Bioconductor as the R package.

2.3 Data Analyses

Principal component analysis and Variance component analysis

The first 5 principal components and the magnitude of variance were computed using the Basic Expression Workflow in JMP Genomics. The magnitude of variance is calculated as an average of the proportion of the trait explained by each principal component.

ANOVA and Volcano plots

Analysis of Variance (ANOVA) was performed fitting a model to data from a class variable in order to retain gene sets of possible effects of the selected variable in JMP Genomics. In this study, ANOVA was utilized to query which genes are differentially expressed between MI and nonMI and between cardiac-death and alive. It generates significance as the negative log p-value and fold-change in gene expression between the groups. Volcano plots were drawn with these two numerical data.

Gene expression analysis

The CDGY studies for gene expression profiling applied two mathematical methods, a principal component analysis and a modular analysis. The PCA is a largely utilized method to discompose variables accounting for as much of variability in the data and this method generated first 5 principal components explaining 42% of variation in CDGY1 and 46.3% in CDGY2. The modular analysis is a methodology designed to support system-scale analysis of gene expression study by Chaussabel et al. [45]. They defined 9 Axes of variation which are strongly characterized as different types of immune function. Each Axis is generated by PC1 of the covarying ten defining blood informative transcripts. Each PC1 is explaining approximately 75% of variation in each module. These Axes are characterizing where people are in immune space.

Genetic of gene expression analysis

GWAS applied gene expression was performed to find cis acting regulatory polymorphism (SNPs located within 250kb upstream or downstream of the probe) in CDGY1. Cross correlation analysis in JMP Genomics was used to compute Pearson all pairwise correlations matrix between all 14,343 probes and 610,859 SNPs which are MAF >0.05 set of variables and to test their significance. To clean SNPs which fall within a gene had linkage disequilibrium, a Bioconductor package, Illumina HumanWGDASLv4, was utilized and SNPs within probe sequence were removed from the SNP list.

To address if there are differences in cis acting regulatory effects on gene expression among CAD status, SNP interaction analysis was performed by fitting a model to ask how differently each gene expression is regulated by a SNP in each CAD status.

Cardiovascular disease Risk Analysis

Two-way hierarchical cluster was performed to find signature of fatal risk by cardiovascular disease. Clustering was performed with 244 probes which are retained from ANOVA test with cardiac death index at NLP3(Negative Logarithm of Pvalue). Three deepest clusters were selected to see how many individuals, who died by cardiovascular disease related events, are falling into each cluster. PC1 of 244 probes were calculated and used as one of the suggested genomic risk predictors.

One of major goals of this study is to generate a genomic predictor of MI and fatal risk. Gene expression associated MI risk is driven by local regulatory polymorphism. 16 eSNPs with MAF >0.05 associated with transcripts of cardiac death risk at p-value $<10^{-8}$ were used to generate risk score. The process of generating risk score is as follow: a) Find Axis explaining death by CAD (PC1 of cardiac death associated transcripts), b) ANOVA test in JMP Genomics with the Axis and collect the genes at p-value$<10^{-7}$which is Bonferroni corrected p-value(alpha =0.01), c) Find eSNPs for genes which are selected from ANOVA test, d) Give risk score 2 for homozygous allele for the high risk variant, risk score 1 for heterozygous, and risk score 0 for homozygous allele for the low risk variant, e) Add all 16 risk scores given by each eSNP.

Multi-locus genotypic risk score for blood gene expression was also generated by summing the number of alleles that are associated with expression of elevated expression in each of the Axes. This score was generated using eSNP data from the parallel healthy adult cohort, and then applied to the Cardiology samples.

CHAPTER 3

RESULTS

3.1 Differential Expression associated with Myocardial Infarction

In order to assess whether peripheral blood gene expression associated with cardiovascular disease status, microarray analysis of 14,111 transcripts were performed in two batches of 175 and 163 participants in the Emory Cardiology GeneBank. Whole blood samples were collected over a five year period and those were preserved in PaxGene tubes. RNA was extracted 12 months apart. Anthropomorphic and clinical parameters are indicated in Table 2A and Table 2B. Approximately equal numbers of individuals were presented who were (i) experiencing an incident of myocardial infarction at or within 24 hours prior to sampling, (ii) had a history of myocardial infarction, (iii) were diagnosed with coronary artery disease after enrolling at the clinic, or (iv) had no evident signs of cardiovascular disease. These are referred to as the MI, oldCAD, CAD, and FINE(Controls) samples respectively.

Raw transcript abundance measures from the Illumina BeadStudio were log base 2 transformed and normalized with the open source algorithm Supervised Normalization of Microarrays (SNM). SNM normalization method were performed to remove technical effects of hybridization batch and RNA quality, to adjust for effects of BMI, gender and ethnicity (in phase 1, but all patients in phase 2 were Caucasian), and to estimate the extent of differential expression associated with CAD status. The π_0 estimate for both phases was between 60 and 70%, indicating that this proportion of the transcripts was unaffected by CAD status. This is however suggesting that as many as 30% of 14343 transcripts abundance may be differentiated with respect to CAD status.

Table 2A. Biophysical and Clinical Data of the CDGY Phase1 cohort

CDGY Phase1	ALL Groups (n=175)	FINE (n= 43)	CAD without MI (n=49)	OLD MI (n=46)	Acute MI (n=37)
Age	67 ± 10	71 ± 7	72 ± 10	62 ± 7	61 ± 8
Male (%)	64	44	63	70	81
Systolic BP	142 ± 24	147 ± 21	146 ± 25	139 ± 23	134 ± 23
Diastolic BP	76 ± 11	77 ± 11	75 ± 13	76 ± 11	76 ± 11
Height (m)	1.7 ± 0.1	1.7 ± 0.1	1.7 ± 0.1	1.7 ± 0.1	1.7 ± 0.1
Weight (kg)	87 ± 21	82 ± 23	81 ± 19	90 ± 19	95 ± 19
BMI	29 ± 6	28 ± 7	28 ± 5	31 ± 7	31 ± 6
Diabetes Hx (%)	37	30	24	46	51
Hypertension Hx (%)	81	77	88	83	76
Dyslipidemia Hx (%)	79	72	82	83	78
Smoking Hx (%)	67	51	79	50	57
Gensini Score	36 ± 51	1 ± 3	37 ± 47	56 ± 71	58 ± 63
Statin Use (%)	74	58	84	74	78
Aspirin Use (%)	75	53	84	76	86
ARB ACE Use (%)	57	53	59	57	59
Beta Blockers Use (%)	61	44	57	67	78
Plavix Use (%)	52	14	65	70	57
Dead by Card at sampling	12	1	2	6	3
MI at sampling	37	0	0	0	37
Dead by Card at 5 years	11	0	6	3	2
MI at 5 years	18	4	4	4	6

Table 2B. Biophysical and Clinical Data of the CDGY Phase2 cohort

CDGY Phase2	ALL Groups (n=163)	FINE (n= 50)	CAD without MI (n=42)	OLD MI (n=47)	Acute MI (n=24)
Age	56 ± 10	54 ± 8	55 ± 5	51 ± 8	71 ± 12
Male (%)	66	64	71	57	75
Systolic BP	133 ± 21	134 ± 21	136 ± 22	127 ± 19	137 ± 22
Diastolic BP	75 ± 12	76 ± 12	76 ± 12	74 ± 10	73 ± 12
Height (m)	1.7 ± 0.1	1.7 ± 0.1	1.7 ± 0.1	1.7 ± 0.1	1.8 ± 0.1
Weight (kg)	92 ± 24	97 ± 18	94 ± 27	90 ± 27	86 ±26
BMI	31 ± 8	32 ± 7	31 ± 7	31 ± 9	28 ± 7
Diabetes Hx (%)	34	16	43	40	46
Hypertension Hx (%)	72	68	79	66	79
Dyslipidemia Hx (%)	77	60	88	85	79
Smoking Hx (%)	60	54	69	63	50
Gensini Score	31 ± 51	0.3 ± 1.3	39 ± 49	53 ± 69	33 ± 34
Statin Use (%)	71	56	83	79	67
Aspirin Use (%)	81	64	90	96	71
ARB ACE Use (%)	56	46	62	64	54
Beta Blockers Use (%)	65	54	69	72	67
Plavix Use (%)	49	8	79	64	54
Dead by Card at sampling	2	2	0	0	0
MI at sampling	24	0	0	0	24
Dead by Card at 5 years	6	1	2	2	1
MI at 5 years	0	0	0	0	0

5 principal components of overall gene expression variation which were computed by Variance component analysis in JMP Genomics, which collectively capture approximately 45% of the variance, indicated that CAD status explains 2.8% of these PC in phase1 and 4.2% in phase2, with only PC 3 significantly affected in the same direction in both. Ethnicity and BMI group make stronger contributions to gene expression variation than CAD status. (Figure 1).

The mean values of PC1 to PC5 among the four CAD status samples were compared and indicated that the mean values of PC3 were differentiated in MI samples from the other three CAD status samples in both phases (Figure 2A, B), but the other three status samples are not significantly differentiated from each other. This was confirmed after an SNM data normalization on the z-scores of transcript abundance linearly adjusted for technical and biological covariates. ANOVA was subsequently performed to contrast of each samples pairwise, as well as of MI vs non-MI. Among the CAD and Control groups did not show significant differential gene expression, but a small number of transcripts were expressed differentially in the MI samples in both phases. Figure 3A, B, C are the volcano plots displaying the relationship between fold difference in abundance along the x-axis, and significance as the negative logarithm of the p-value (NLP) on the y-axis. Higher expression in non-MI samples are showed in the right arm. Differentiation in transcript abundance is biased toward up-regulation in the MI samples, and the effect of MI seems to be stronger in the second phase as expected from variance component analysis.

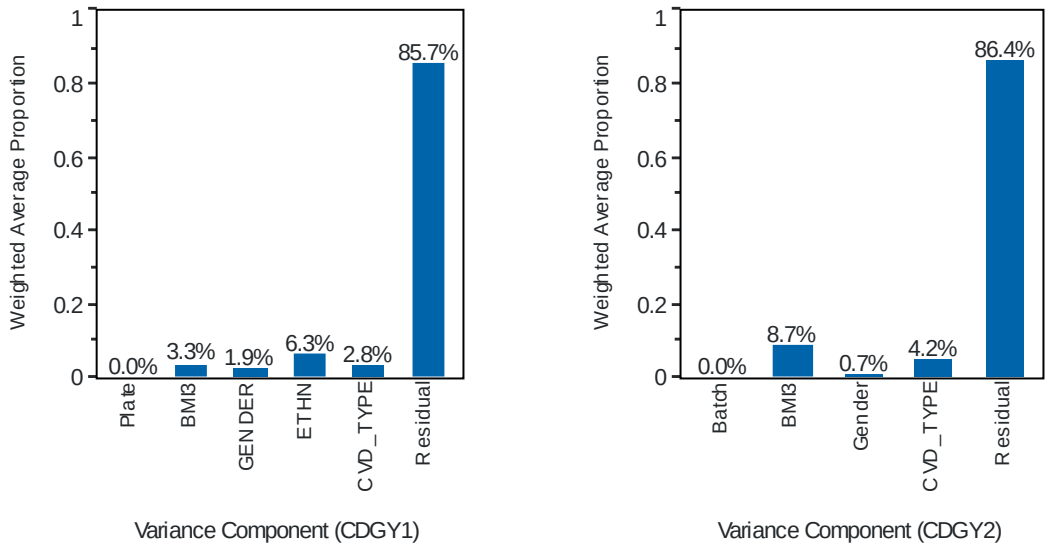

Figure 1. Principal Variance Component analysis in the two phases. In both cases, 85% of the variance among individuals is unattributed, and 2.8% is correlated with CAD status in phase1 and 4.2% in phase2. Ethnicity is only a factor in Phase 1. All individuals in Phase 2 were Caucasian. BMI3 refers to 3 categories of BMI (>30, Obese; >25, Overweight; <25 Normal). Technical effects from Plate and Batch were adjusted and no longer remained.

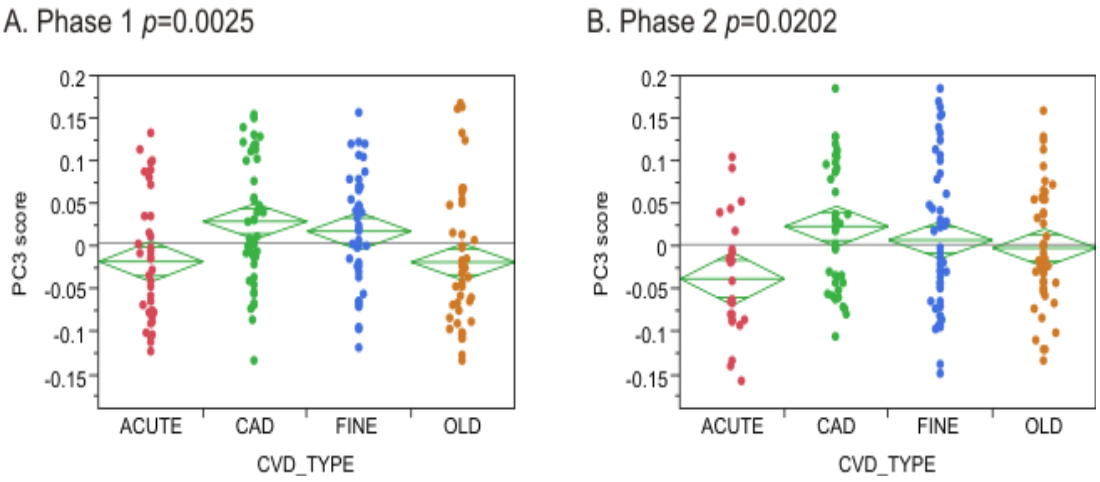

Figure 2 A, B. Differential expression related to class of CAD. Principal component 3 scores by class of CAD in the two phases of gene expression profiling. Replication of differential expression between individuals experiencing an acute MI with meaningful P-values from ANOVA with three degrees of freedom for the CAD class effect, whereas the current CAD, healthy control, and history of CAD (CAD, FINE, OLD respectively are not consistently divergent.

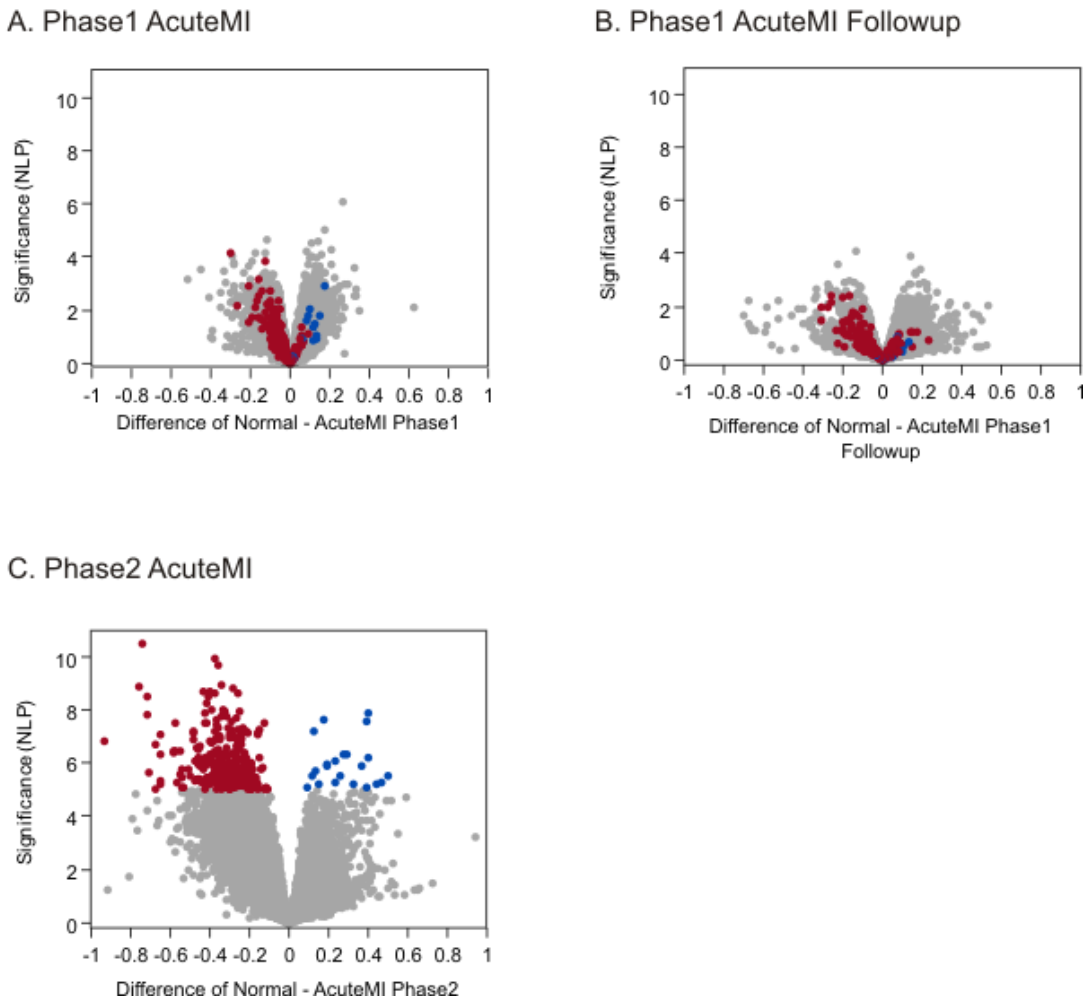

Figure 3 A, B, C. Volcano plots of significance against difference in expression in log2 units of SNM-normalized data. Up-regulated genes in patients with an MI to the left, and down-regulated genes in MI to the right. Transcripts at $p<0.001$ in phase 2 are colored. Almost all of transcripts show same directionality in differential expression in Phase 1 and Phase 1 follow-up, which however showed much less evidence for significant differential expression in the total sample (suspecting possibility of lower technical quality, or unknown covariates biasing against the MI association).

3.2 Axes of Variation implicate altered Neutrophil and Lymphocyte signaling in Myocardial Infarction

Although there is little overlap in the identity of transcripts between phase 1 and phase2 that are significantly associated with MI, the differential expression is consistently in the same direction. Figure 3 is showing this directionality of differential expression of transcripts with MI. The red circles correspond to the significantly up-regulated transcripts in phase 2 (Figure 3B) and are biased toward up-regulation in the MI samples, implying a high false negative rate with a sample of fewer than 50 patients. The blue circles correspond to the down-regulated genes and these are uniformly down-regulated in phase 1. The reciprocal result is seen in phase 2 for significant transcripts of phase1. To explore the nature of the differential expression further, evaluation was performed with the nine Axes of Variation that are highly conserved in peripheral blood gene expression profiles from multiple studies in healthy and disease cohorts. These nine Axes of peripheral blood gene expression variation was recently described and showed in the publication of Preininger et al., [16]

Covariance with ten defining blood information transcripts (BIT) per Axis defines each Axis of Variation. PC1 of the 10 BIT capture the covariance, and these nine defining PCs were fit jointly by multiple linear regression to the full set of 14,343 transcript probes in the analysis. Over 7,000 probes associate with at least one Axis at Bonferroni significance ($p<5.3\times10^{-5}$), and cross-matching of the list of significant associations between the two phases shows on average 80% overlap, ranging from 75% for Axis 4 to 95% for Axis 7. Each transcript associated with the same Axis in both phases was classified as an Axis gene, and we plot the means of the transcript abundance which are differed in the MI samples from the non-MI for each Axis gene by Axis as in Figure 4A,B.

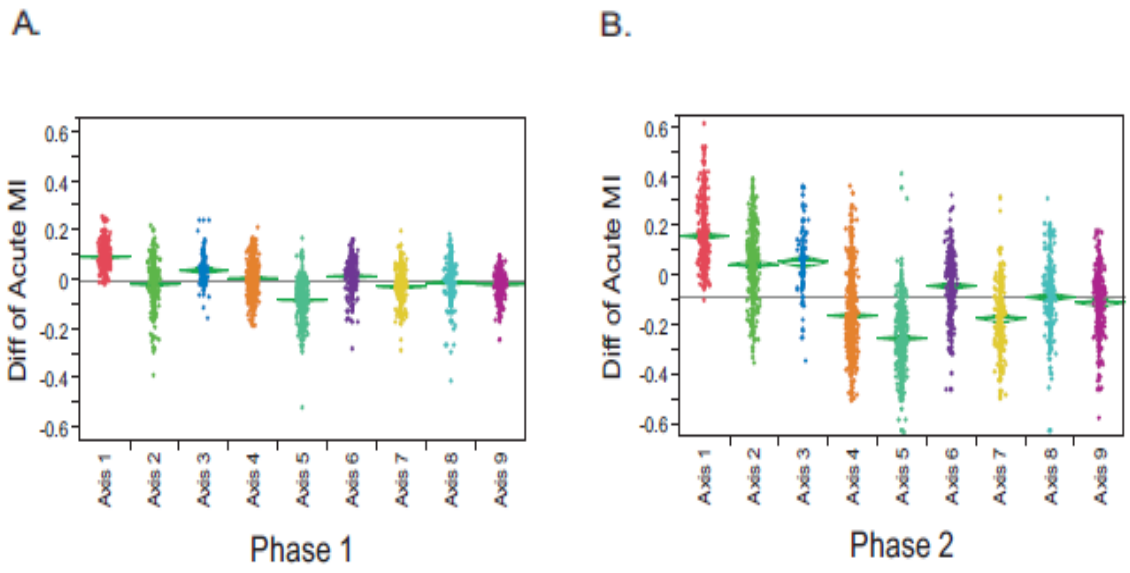

Figure 4 A, B. Replicated association of Axes of Variation with MI. Each dot is represented each gene that is correlated with one of the 9 Axes in both phases. The plots are showing the mean difference between individuals experiencing an MI and nonMI at sampling (higher expression in MI producing negative values). The overall relationship of the Axes with MI status is highly replicated, particularly notably showing down-regulation of Axis 1 and up-regulation of Axis 5. The difference between the studies in Axis 2 is readily explained because this Axis associated with BMI, which is elevated in the phase 2 samples.

Unambiguous down-regulation of genes in Axis1 and up-regulation of genes in Axis5 show the concordant directionality of effects between the two phases. The PC1 scores for these two Axes genes are negatively correlated in this and other studies, because they are partially correlated with lymphocyte (Axis 1) and neutrophil (Axis 5) counts respectively. Cell counts are not available in this cohort, but the data is nevertheless consistent with up-regulation of neutrophil-related gene expression in patients experiencing an MI. The data does not distinguish whether this reflects a predisposition to MI in patients with high neutrophil counts, or an increase in neutrophil counts during MI, both of which have been documented in the literature.

There is an independent evidence for neutrophilia explaining some of the high Axis 5 expression in MI patients. This comes from cross-matching of the Axis 5 genes with genes known to be neutrophil-enriched from published gene expression profiles of distinct blood types. 46 percent of 2,469 genes which is known to be neutrophil-enriched are associated with Axis 5 in this study, and 55 percent of the 2,042 Axis 5 genes are neutrophil-enriched. However, it is possible that differential neutrophil abundance is not simply explains the differential expression, since the overlap is only partial. Similarly, reduced lymphocyte abundance does not solely explain the low Axis 1 scores in MI patients, because T-lymphocyte abundance is strongly associated with only 44 percent of 3,584 Axis 1 genes. Furthermore, gene ontology analysis suggest that genes in Axes 8 and 9 are also involved in aspects of T-cell signaling, while Axis 3 has an excess of genes annotated to B-cell signaling. These results lead to the conclusion that differential gene expression in MI patients is largely attributable to elevation of the ratio of neutrophils to lymphocytes, but also involves differential gene expression within these cell types.

3.3 MI disrupts the effect of genotype on expression of a fraction of genes in peripheral blood

eQTL analysis was performed and found further evidence for differential expression. As described in the method, non-Caucasian samples were excluded and 153 Caucasian samples were remained for eQTL analysis to prevent the possible influence of population stratification between Caucasian, African, and Asian American samples. The genotypes of over 600,000 common polymorphisms with a minor allele frequency <0.05 were obtained and local regulatory effects of theses SNPs on abundance of transcripts was computed utilizing Pearson correlation method in the phase1 samples. Only cis-eSNPs which are located within 250kb were included in the profile. eSNP profiles of associations were filtered subsequently to remove all known common SNPs within the probes in this study and in high linkage disequilibrium ($r^2 > 0.5$) with them.

Local genetic regulation of transcript abundance was detected for 355 probes representing 334 genes at $p<10^{-5}$, with a q-value false discovery rate of just 1%. 87% of these eSNP effects replicated in a parallel healthy adult cohort that we are also studying in Atlanta, implying that they are robust eQTL in adult peripheral blood samples. Interestingly, 5% of all transcripts are Axis 5 genes, but only 0.5% of eSNPs were observed in Axis5 genes, implying significant under-representation of Ais5 genes with eSNPs. On average, each eSNP explains 11% of the abundance of the associated transcript, with a range from 5% to 45%, as observed in other studies.

Linear regression was performed to evaluate whether MI status influences local genetic effects, fitting transcript abundance as a function of genotype, MI status, and the interaction between them. 78 significant interaction effects were observed at $p<0.05$, more than expected by chance, but only a handful at experiment-wide significance. In order to confirm this result, MI samples were removed and eQTL analysis on just the 120 non-MI Caucasian samples was performed again. This resulted in the loss of 95

significant associations as expected due to the reduced power in a smaller sample, but also led to the detection of an additional 45 associations at $p<10^{-5}$, as shown in Figure 5A. . 36 cases show significant interaction effect between genotype and MI status. Each case did not show any significant relationship between genotype and gene expression in the MI samples where a highly significant on exists in the non-MI samples (Figure5B) By contrast, an interaction effect with a modest increase of MI status on the genotypic effect was shown in only 6 out of the 50 cases of reduced significance in the non-MI samples (Figure 5C).

MI reduced the genetic regulation of expression of a subset of genes. This result could be simply explained that the eSNP effect is diluted out in the MI samples where neutrophils are more prevalent if the eSNP effect is predominately observed in lymphocytes or monocytes. To exclude this explanation, an equal number of non-MI whose Axis5 values are high was removed from the phase1 and eSNP profiling was re-performed. As a result, only a handful of cases showed increased significance below the $p<10^{-5,}$ nevertheless significance was reduced for 100 genes due to the loss of power by reducing the size of the study.

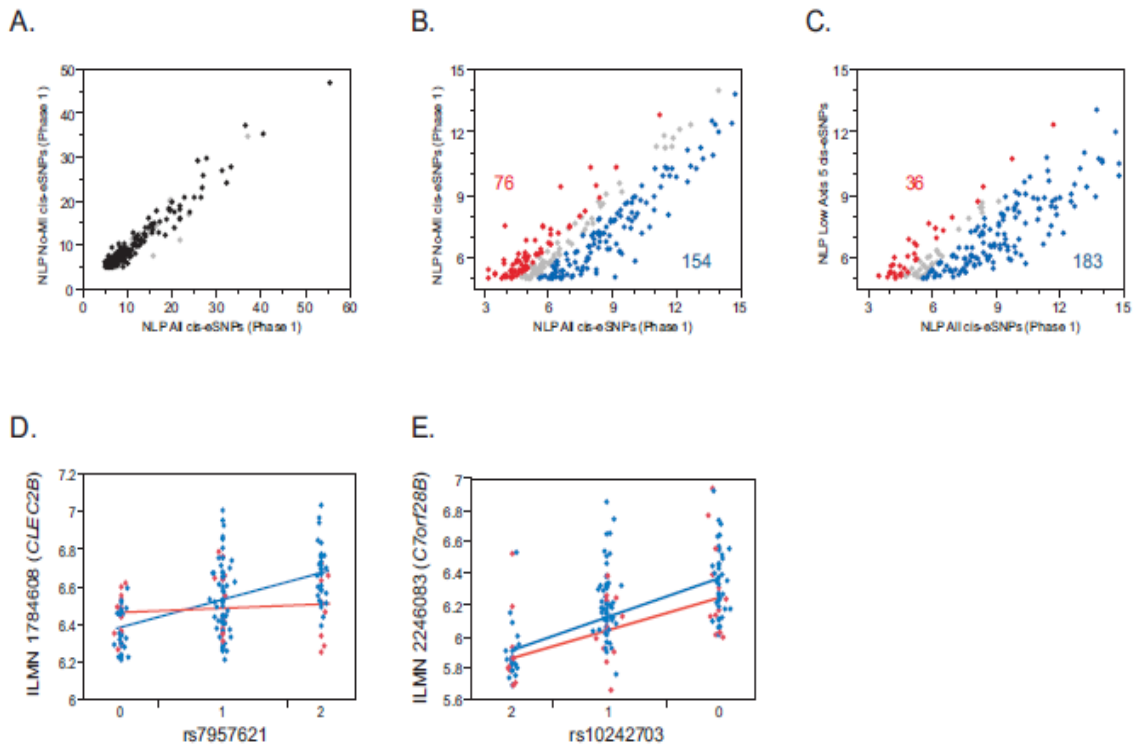

Figure 5 A, B, C, D, E. Expression QTL analyses. (A,B) Plot of significance (negative log P, NLP) of cis eSNPs with a transcript probe located within 250kb of the SNP, contrasting the All Caucasian sample in Phase 1 with the same sample excluding individuals experiencing an MI. (A) Shows the full range of NLP, and (B) eSNPs in the range 5 < NLP < 15. Colored points differ between the full and no-MI samples by at least 0.5 NLP units, red indicating higher significance in the nonMI sample, and blue higher significance in the full sample (ALL). (C) Shows the same analysis as (B) but missing 37 individuals with the highest Axis 5 scores who were not experiencing an MI. Numbers show the number of eSNPs in the red and blue categories across the full range of NLP > 3. (D,E) Representative plots of transcript abundance by genotype, colored with respect to MI status (red experiencing an MI, blue no-MI). Lines show the slope of the two classes, indicating an interaction effect in (D) but no interaction effect in (E).

3.4 Risk of death due to a cardiovascular event may also associate with the Axes of Variation

In the phase 1 sample, 12 individuals had died of a verified heart attack or stroke at sampling and 10 individuals during 5 years of period. Analysis of variance contrasting the MI samples with the remainder revealed 244 transcripts significantly differentially expressed at $p<10^{-3}$. Two-way hierarchical clustering of these transcripts (Figure 6A,B) reveals 3 groups of genes that are clearly co-expressed and distinguish the patients who died subsequent to donation of a sample to the Cardiology GeneBank. The first principal component of the 244 significantly differentiated transcripts is predictive of death-due-to-CVD with an AUC 0.80 (Figure 6C). This result was unambiguously replicated in the phase 2 with 8 patients have subsequently died with an AUC over 0.75 (Figure 6D).

The differential expression associated with death appears to involve the same components of variation as the MI-associated genes. PC1 for the three clusters in Figure 6A are highly correlated with the PC1 scores for Axes 1 and 5, even though the identities of the highly significant genes are distinct from the MI-associated genes (Figure7A,B). If the 5 patients who were experiencing an MI and subsequently died due to a second event are removed from the analyses, the correspondence between 'death-related' gene expression and 'MI-related' gene expression remains since the up-and down-regulated transcripts are still associated with Axes 5 and 1 respectively (Figure 7D). This result implies that peripheral blood gene expression profiles that are associated with MI may also predict likelihood of risk of death due to a cardiovascular event, but we caution that independent replication is essential given the small number of samples.

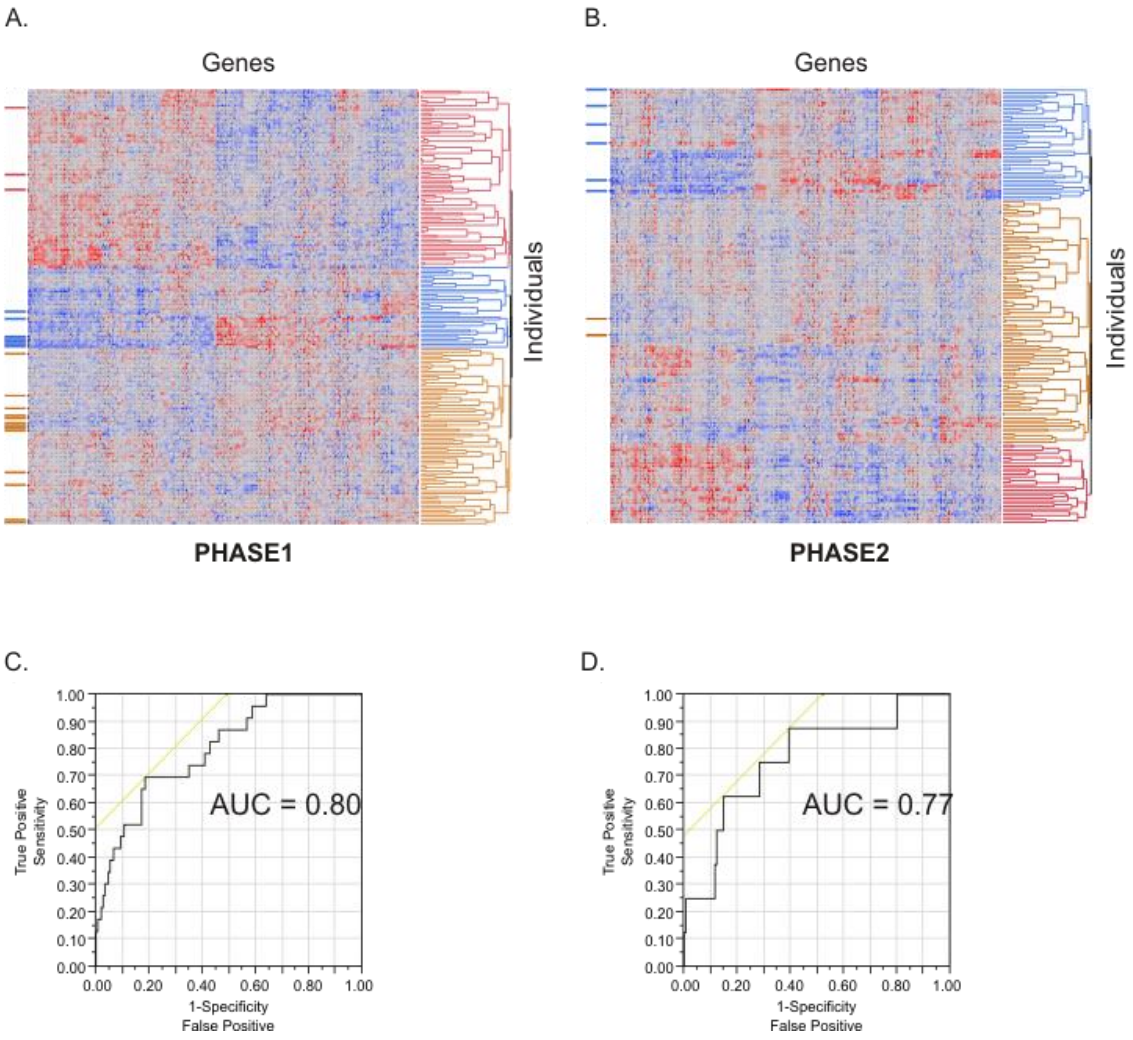

Figure 6 A, B, C, D. Two-way hierarchical plot of normalized transcript abundance (columns) by individual participant (rows), highlighting the three deepest clusters. The left column indicates individuals who have had an MI during the 5 year period since enrolling at the clinic and who have subsequently died of an acute MI or stroke. (A) Note that 7 of the latter are in the 32 person blue cluster in phase1. (B) Two-way hierarchical plot was generated in phase2 with same transcripts in phase1. Three deepest clusters were highlighted. Similarly, blue cluster includes 6 persons out of 8 total individuals who have died by an MI or stroke. Efficient of this model was represented as ROC curve, with AUC 0.8 in phase1 and 0.77 in phase2

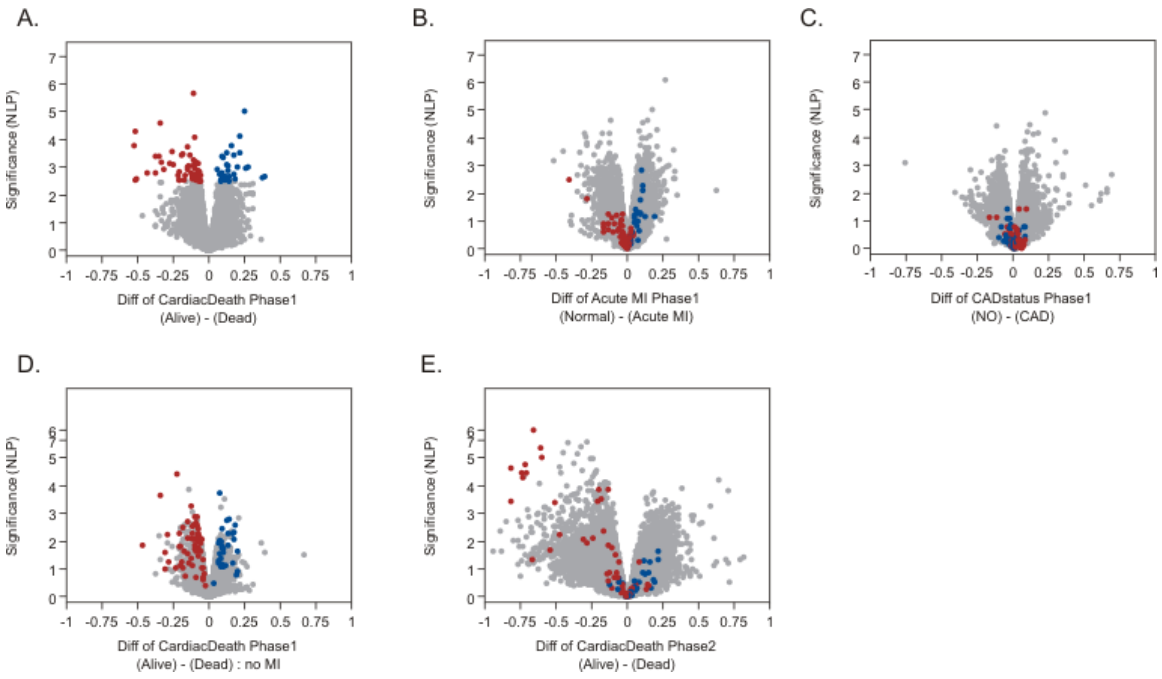

Figure 7 A, B, C, D, E. Volcano plots contrasting differential expression with respect to death due to MI , MI, or CAD status. Colored points are significant in (A) at NLP>3 showing all in the same direction with (B), (D), (E), but there is only a weak, non-significant trend in the CAD versus healthy control contrast in (A). Red circles are up-regulated in the 23 individuals who were confirmed to have died of a cardiac event. Down-regulated transcripts were marked as blue circles. (D) Shows the differential expression in just the 18 participants who died of a cardiovascular event but were not experiencing an MI, and all other non-MI individuals.

3.5 Absence of relationship between gene expression and pharmaceutical usage of measures of CAD

Association of pharmaceutical usage with gene expression in the cohort was also assessed. One-way ANOVA analyses show no significant enrichment for gene expression by prescription status for Statins, Beta Blockers, Plavix, or self-reported Aspirin usage. However, significant association between Axis scores (PC1 of the blood informative transcripts) and Plavix usage or Beta-Blockers was observed in both phases, of which the effects were nevertheless not replicated in both studies (Table 3). Thus, it is most likely that they are due to confounding of the covariance of gene expression in each Axis with unidentified biological factors that may have been common in the drug groups. Alternatively, the factors that were not captured in our analysis such as effects from dosage, prescription interval, or patient compliance are accountable for that drugs influence gene expression.

Similarly, relationship between gene expression and angiographic measures of coronary function was also not detected. For example, Gensini index showed large variability as expected of a cardiology cohort that includes non-CAD individuals and patients with long-term CAD, but was not correlated with any of the Axes of Variation and further had no significant association with individual transcripts. One measure, the count of putative progenitor cells (VEGF+, CD44+) in peripheral blood, showed a notable tendency toward significant association, however were too small fraction of the total cell counts and thus could not be directly associated with total blood gene expression.

Table 3. Association between Axis scores (PC1 of the blood informative transcripts) and drug usage

CDGY Phase1	ARB_ACE	Aspirin	BB	Plavix	Statin
snmAxis1	0.1836	0.3797	0.026	0.1779	0.7524
snmAxis2	0.5009	0.8026	0.1032	0.0224	0.4118
snmAxis3	0.0718	0.7499	0.9961	0.1101	0.8972
snmAxis4	0.8347	0.1061	0.4121	0.0032	0.1836
snmAxis5	0.0346	0.6337	0.0441	0.1246	0.667
snmAxis6	0.2781	0.0678	0.1541	0.1086	0.4888
snmAxis7	0.7535	0.4342	0.495	0.0483	0.1524
snmAxis8	0.7412	0.0215	0.7177	0.6603	0.0173
snmAxis9	0.7349	0.7014	0.0558	0.0282	0.4024
snmAxis10	0.1115	0.2117	0.0255	0.6889	0.3016

CDGY Phase2	ARB_ACE	Aspirin	BB	Plavix	Statin
snmAxis1	0.1034	0.0825	0.3247	0.0361	0.541
snmAxis2	0.158	0.5095	0.1991	0.5845	0.8628
snmAxis3	0.1425	0.6631	0.3242	0.9502	0.7558
snmAxis4	0.7504	0.7459	0.8229	0.6265	0.9286
snmAxis5	0.5094	0.1245	0.5751	0.0431	0.4936
snmAxis6	0.6511	0.9084	0.2751	0.6245	0.5517
snmAxis7	0.7223	0.6653	0.3923	0.6767	0.6182
snmAxis8	0.7105	0.8915	0.3309	0.7893	0.3877
snmAxis9	0.6234	0.5218	0.3063	0.1517	0.9016
snmAxis10	0.6352	0.1897	0.8721	0.0417	0.328

3.6 Genotypic Risk scores

Genome-wide association study has identified 16 significant SNPs that are reproducibly associated with the cardiac death index. Differentiation between the MI samples for the Control was only modest when using a multi-locus genotypic risk score that was generated by summing the number of risk alleles in each person in phase 1. The amount of variance in coronary disease explained was low (Figure 8 A,B), indicating that score generated in this way has low predictive ability in follow up MI. Although it was associated with the expression of a subset of genes in peripheral blood, correlation with the MI-related or any other Axes of Variation were not strong.

An alternative multi-locus genotypic risk score for blood gene expression was generated by summing the number of alleles that are associated with expression of elevated expression in each of the Axes. This score was generated using eSNP data from the parallel healthy adult cohort, and then applied to the Cardiology samples. As expected, the scores are associated with the corresponding Axis scores, since individuals with more cases of elevated expression due to local regulatory polymorphisms will tend to have higher expression of genes in the relevant Axis overall. Compared to individual SNP effects on gene expression, the fraction of the covariance explained was lower. However, genotypic risk scores for Axis 1 and Axis 5 could be regarded as predictors of risk of MI in larger cohorts, since they were capable of differentiating the MI and non-MI samples at nominal levels of significance.

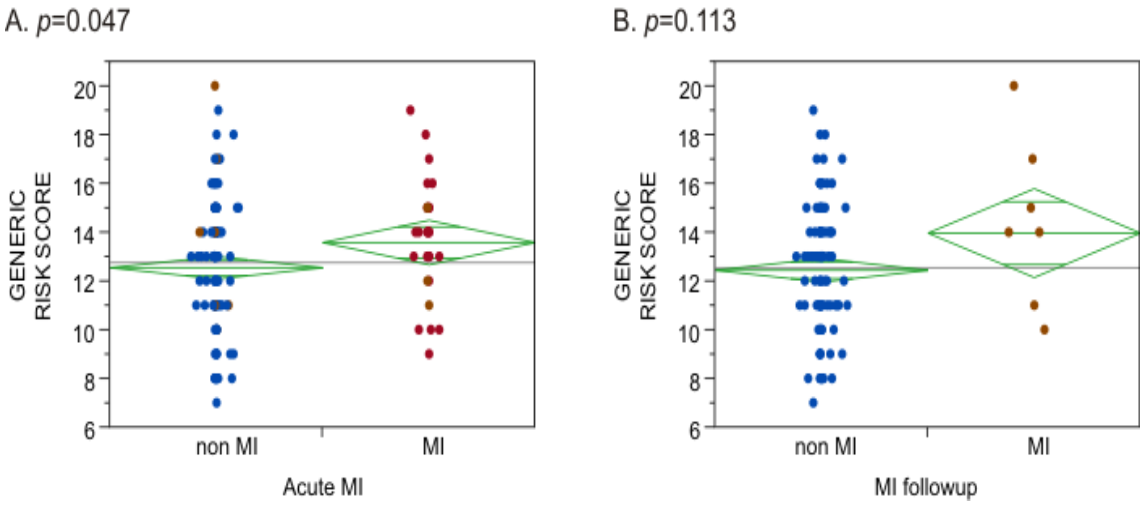

Figure 8 A, B. Generic risk score of MI risk. 244 probes which are retained from ANOVA test with cardiac death index were used for generating genetic risk score. Risk allele is given risk score 2, risk score 1 for heterozygous, and risk score 0 for homozygous allele for the low risk variant. Genetic risk score is ranged between 0 and 32.

CHAPTER 4

DISCUSSION

The primary effort of this work was to find genomic evidence relevant to coronary artery disease. Two major methods of genomic analysis, gene expression profiling and GWAS on gene expression, were performed to dissect transcriptional and genotypic fingerprints of coronary artery disease. Blood-informative transcriptional Axes that can be described by 10 covariating transcripts per each Axis were utilized as a crucial measure of gene expression analysis.

This study of the relationship between gene expression variation and various measurements of coronary artery disease delivered compelling results showing strong association between two transcriptional Axes and incident of myocardial infarction. 244 transcripts (Gene names provided in APPENDIX A) closely correlated with cardiac death were also showing clear association with those two transcriptional Axes. These results suggest potential transcripts for use in risk prediction for the advent of myocardial infarction and cardiac death.

4.1 Correlation with Myocardial Infarction

This study produced evidence of clear signatures of myocardial infarction. First, the results show evidence that two Axes of variation, Axis1 and Axis5, are strongly associated with the MI diagnosis index. In general, about thousands of genes in each Axis from Preininger et al., represent consistently covariating pattern and 60% of total genes are corregulated with one of Axis. Axis1 and Axis5 are in turn enriched in aspects of activities of T-lymphocytes and Neutrophils. Since this cardiology cohort does not include information about cell count, enrichments of T-lymphocytes and Neutrophils activity were evaluated in the healthy Atlanta cohort from CHDWB and with genes

documented in the literature. Up-regulated Axis5 and down-regulated Axis1 were clearly observed with patients experiencing an MI. These signatures were not only revealed in phase1, but also discovered in phase2 of the CDGY cohort. While results from analysis of follow-up incidence of MI did not reach statistical significance, evident trends of high Axis5 and low Axis1 were detected both in phase1 and phase2. These outcomes are also consistent with the clinical data analysis done by Haumer et al., showing increased risk of major adverse cardiovascular events and death in the upper tertile of neutrophil counts and better cardiovascular outcome with lower tertile of lymphocyte counts [46]. In addition, recent re-analyses of the Framingham data by the Peter Wilson's group at Emory University supported our results [47]. These findings lead us to conclude that there is a modification of the ratio of Neutrophils to Lymphocytes in MI patients, or that individuals with a high ratio are at higher risk of MI.

In addition to higher ratio of neutrophils to T-lymphocytes in MI patients, the alteration of genetic regulation on gene expression in patients with myocardial infarction demonstrates loss of the impact of genotypes on gene expression due to MI events. Comparison of cis regulating eSNP profiles for MI patients and non-MI patients in the cohort confirmed our two hypotheses: removal of MI patients increases the statistical significance of subset of eQTL and removal of people with high Axis5 alone (no MI) does not affect the eQTL profiles. These two analyses validated the hypothesis that altered genotypic regulation is not the result of enriched neutrophils circulating in the blood, but due to physiological changes in response to MI.

In conclusion, MI is associated not just with an increase in neutrophil activity, but also with an alteration of gene activity in both neutrophils and lymphocytes. While our results indicate that transcriptional changes are signatures of response to MI events, it is still not clear whether these differences between MI patients and the other groups were triggered due to response to the occurrence of MI or whether they are indicative of

predisposition to MI. With longitudinal data, better clarification of cause and effect would be feasible.

4.2 Predictive Transcripts of Cardiac Death

In this study, clustering with 244 unique transcripts was suggested as a risk predictor of cardiac death. Expression of these transcripts and index of individuals whose deaths were caused by cardiovascular disease related event were strongly associated in phase1 of the CDGY cohort. These signatures of cardiac death were clearly revealed in phase 2 and follow-up cardiac death. Two-way hierarchical clustering of individuals with these transcripts produced 3 clearly defined groups in both phase 1 and phase2 as Figure6 (A,B) and most of patients who died from cardiovascular disease related event fell into a small cluster of 13 individuals showing down-regulated expression of 48% of the 244 suggested transcripts. PC1 of suggested transcripts in phase1 which are correlated with confirmed death by cardiovascular disease were computed and PC1 shows significant association with cardiac death index and Acute MI event index. This is also observed with MI follow-up in phase1 (Figure 9). Not only within phase1, were theses strong association replicated in phase2. This replication study implies that PC1 of our suggested transcripts may contribute to predict MI risk which has not yet been occurred. Validation of these potential markers with larger data set will increase the power of assessment for cardiovascular disease.

PC1 of these genes was highly associated with increased activity in T-lymphocytes and decreased activity in Neutrophils, implying that a higher ratio of Neutrophil activity to T-lymphocytes is suggestive evidence of elevated risk of cardiac death. This outcome is consistent with correlation of Axis1 and Axis5 with myocardial infarction. In addition to this, there was evidence showing that B cell activity was also enriched in the 244 transcripts annotation, indicating that pro-inflammatory responses are closely related with risk of both cardiac death and myocardial infarction.

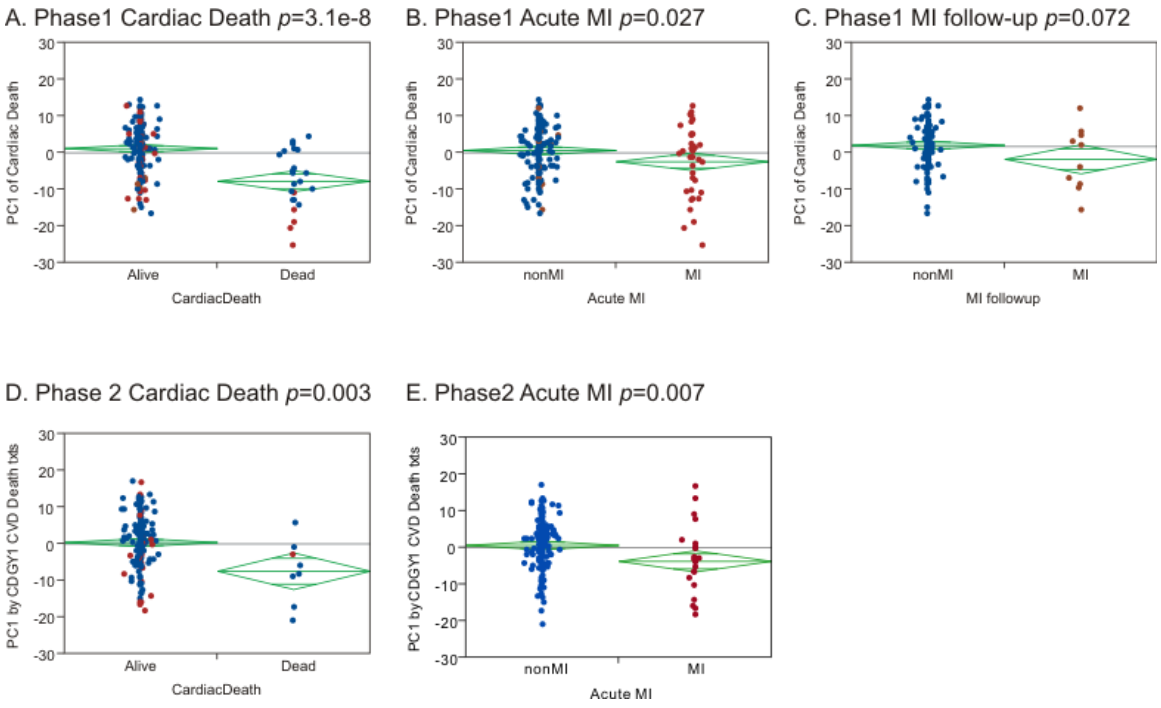

Figure 9 A, B, C, D, E. Association between Principal Component (PC) of cardiac death related transcripts and CAD status. (A) shows PC generated by 244 transcripts retained from ANOVA on cardiac death samples as y-axis. (B) is displaying the significant association of PC with Acute MI samples and (C) with MI follow up samples. The same plots were drawn in phase2 with cardiac death index (D) and Acute MI index (D). 244 transcripts in phase1 were used to generate PC of transcripts in phase2.

4.3 Low Power of Follow-up Analysis and Ambiguities of Calling CAD Types

For the analyses with clinical data, it is crucial to have concrete definitions of criteria for diagnoses. Ambiguous classification of controls or degree of disease affects critical results of analyses and may give rise to loss of power of the study, and even lead to failure of an entire study. For example, participants who are classified as FINE can possibly have underestimated degrees of plaque accumulation, as measured by angiography examination. In this study, such an underestimation could hide signatures of association between the transcriptome and clinical measure of CAD. In order to avoid this, the criteria of classification for calling FINE or different disease types should be carefully addressed.

Unfortunately, the sample sizes of the CDGY cohorts are too small to evaluate whether genotypic differences contribute to the risk of cardiac death and myocardial infarction. The size of follow-up data is also limited in both phases of cohorts. Because of this, the power of replicated study was not reached to optimal power. Since our analyses are inferring that experiencing MI altered genotypic regulation of gene expression relevant to pro-inflammatory response, it is highly desirable to address whether genotypes also affect differentiation of gene expression in respect to coronary artery disease and advent of death to jointly contribute to develop robust genomic markers of risk of adverse events along with transcriptional peculiarity. Validation of predictive genomic markers with a larger number of samples may contribute as a new clinical tool to improve clinical care including preventive care, risk assessment and early diagnosis of cardiovascular disease.

APPENDIX A

POTENTIAL GENOMIC MARKERS FOR MI RISK PREDICTION

PTDSS1	RASGRP2	STIP1	RCAN3	CDK5R1
LDLRAP1	HNRPR	TXNL4A	PDK4	CLIC4
DKFZP761P0423	SIN3A	TPST2	TM7SF2	CD63
CIB1	FBXW11	CRKRS	WDR51B	YWHAE
LPAR5	MAN1C1	SNX6	GREB1	JOSD3
CD79B	PRDM4	TATDN2	GANAB	TSPO
IL7R	MYC	IMMP2L	KIAA0495	BIN2
DDX42	IL7R	OSBPL3	ATP2B4	KIAA1033
ZNF395	PERLD1	IDS	DHFRL1	GLYCTK
RPL18A	PHF15	ASGR1	FMO6P	CD59
OSBPL10	MRPS18B	C1ORF25	RALGPS1	FTHL12
EVL	CD79A	ZUFSP	LOC648164	METRNL
LOC441775	PIK3IP1	SULF2	NLRC4	LOC653778
DENND2D	ETS1	IDUA	WHDC1	LOC440926
CD247	PAFAH1B1	TRAPPC5	ZNF671	TOM1L2
GRAP	SFRS5	XK	PRR11	CYP4A11
PRKCH	PRPF8	STRC	FYB	TRAPPC6B
M-RIP	LOC730316	ZNF544	LOC644330	GINS4
PUM1	GIMAP4	3318796800	C20ORF45	MAP2K2
TAF15	NUB1	UBTD1	ARHGAP30	LGALS3
ST6GAL1	HCP5	TRIM13	MAP2K7	LOC648710
BMS1	CDC2L5	MNAT1	UQCRC2	LOC653604
KRTCAP2	ALDH9A1	DENND4A	MTHFD2L	SIGLEC7
LTA	BTG1	TPCN1	C8ORF53	C4ORF18
ARHGAP17	STK4	KMO	CHPT1	ANXA2P1
SFRS5	TMEM66	SMPDL3A	VCAN	DDX55
LOC728554	ZHX2	RAB1B	EPHB4	CD36
NUMA1	JARID1A	WSB1	MXI1	CCL23
C1ORF77	RPLP2	KIF13B	SMA5	HIST1H2BG
HDAC1	PPP1R11	C21ORF91	LRRC40	CCDC23
SESN1	EIF4H	BRSK1	TMEM183B	MGAT4B
DDX24	NDRG1	PBRM1	NRBP1	ANKRD44
CXXC1	BANP	C6ORF21	AP3D1	C10ORF11
WDR6	RTN3	CYP1B1	COQ6	CRY2
ARHGEF2	PSG9	RAB32	AP2S1	CBR3
ATP1A1	UBE2F	LPCAT1	SEH1L	FTHL2
KIAA1147	STX5	KYNU	C20ORF108	FTHL11
GNG7	PRCC	GOLGA8A	FTHL2	TRAPPC6B
TAF15	LYPLA2	TSTA3	TLR1	C4ORF18
SPOCK2	FLVCR2	LACTB	HS1BP3	NKX3-1
FAM62A	RNH1	KIAA1641	FTHL11	TMEM42
LOC285636	TSSC1	LOC285407	PCTP	LILRA3
FCRLA	C1ORF159	PIK3C2B	S100A12	UBAC1
SPTAN1	LOC391157	LOC402221	HEATR3	FTHL8
FAIM3	FBXO38	DICER1	PBX1	RNASE4
RBM14	TOB1	RASSF4	C11ORF57	FTHL12
ATP8B2	QSOX1	MS4A4A	WSB2	FTHL3
SF3B3	HIAT1	FAHD1	ZNF318	MLH3

Lower principal component of 244 genes is suggested to be used as a genomic predictor of MI risk assessment. Blue colored genes are down-regulated in both phases of death.

REFERENCES

1. Cao, Y., et al., *Cancer research: past, present and future.* Nat Rev Cancer, 2011. **11**(10): p. 749-54.
2. Kovacic, J.C. and V. Fuster, *From treating complex coronary artery disease to promoting cardiovascular health: therapeutic transitions and challenges, 2010-2020.* Clin Pharmacol Ther, 2011. **90**(4): p. 509-18.
3. Krieger, N., *Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology.* Am J Public Health, 1992. **82**(5): p. 703-10.
4. Ardern, C.L., et al., *Discrimination of health risk by combined body mass index and waist circumference.* Obesity Research, 2003. **11**(1): p. 135-142.
5. Mokdad, A.H., et al., *Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001.* Jama-Journal of the American Medical Association, 2003. **289**(1): p. 76-79.
6. Burton, P.R., et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.
7. Tomlinson, I., et al., *A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21.* Nature Genetics, 2007. **39**(8): p. 984-988.
8. Buch, S., et al., *A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease.* Nature Genetics, 2007. **39**(8): p. 995-999.
9. Anderson, K.M., et al., *Cardiovascular-Disease Risk Profiles.* American Heart Journal, 1991. **121**(1): p. 293-298.
10. Taylor, S.H., *Hypertension and Coronary-Artery Disease - a Therapeutic Challenge.* Journal of Cardiovascular Pharmacology, 1991. **18**: p. S39-S44.
11. Brindle, P., et al., *Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study.* BMJ, 2003. **327**(7426): p. 1267.
12. Coleman, R.L., *Framingham, SCORE, and DECODE Risk Equations Do Not Provide Reliable Cardiovascular Risk Estimates in Type 2 Diabetes.* Diabetes Care, 2007. **30**(5): p. 3.
13. Greenland, P., et al., *Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals.* JAMA, 2004. **291**(2): p. 210-5.
14. Fowkes, F.G., et al., *Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis.* JAMA, 2008. **300**(2): p. 197-208.
15. Ridker, P.M., *Clinical application of C-reactive protein for cardiovascular disease detection and prevention.* Circulation, 2003. **107**(3): p. 363-9.
16. Preininger, M., et al., *Blood-informative transcripts define nine common Axes of peripheral blood gene expression.* PLoS Genet, 2013. **9**(3): p. e1003362.
17. Mancinelli, L., M. Cronin, and W. Sadee, *Pharmacogenomics: The promise of personalized medicine.* Aaps Pharmsci, 2000. **2**(1).

18. Storey, *Gene-expression variation within and among human populations (vol 80, pg 502, 2007).* American Journal of Human Genetics, 2007. **80**(6): p. 1194-1194.
19. Whitney, A.R., et al., *Individuality and variation in gene expression patterns in human blood.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(4): p. 1896-1901.
20. Gibson, G., *The environmental contribution to gene expression profiles.* Nature Reviews Genetics, 2008. **9**(8): p. 575-581.
21. Kim, J. and G. Gibson, *Insights from GWAS into the quantitative genetics of transcription in humans.* Genetics Research, 2010. **92**(5-6): p. 361-369.
22. van't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-536.
23. Bomprezzi, R., et al., *Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease.* Human Molecular Genetics, 2003. **12**(17): p. 2191-2199.
24. Booth, F.W., M.V. Chakravarthy, and E.E. Spangenburg, *Exercise and gene expression: physiological regulation of the human genome through physical activity.* Journal of Physiology-London, 2002. **543**(2): p. 399-411.
25. Idaghdour, Y., et al., *Geographical genomics of human leukocyte gene expression variation in southern Morocco.* Nature Genetics, 2010. **42**(1): p. 62-U79.
26. Idaghdour, Y., et al., *A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan amazighs.* Plos Genetics, 2008. **4**(4).
27. Mason, E., et al., *Maternal influences on the transmission of leukocyte gene expression profiles in population samples from Brisbane, Australia.* Plos One, 2010. **5**(12): p. e14479.
28. Ross, N.A., et al., *Body mass index in urban Canada: Neighborhood and metropolitan area effects.* American Journal of Public Health, 2007. **97**(3): p. 500-508.
29. Ludwig, J.e.a., *Neighborhoods, Obesity, and Diabetes — A Randomized Social Experiment.* N Engl J Med, 2011. **365**(16): p. 1509-19.
30. Chen, E., et al., *Socioeconomic status, stress, and immune markers in adolescents with asthma.* Psychosomatic Medicine, 2003. **65**(6): p. 984-92.
31. Sanz-Santos, G., et al., *Gene expression pattern in swine neutrophils after lipopolysaccharide exposure: a time course comparison.* BMC Proc, 2011. **5 Suppl 4**: p. S11.
32. Miller, G.E., et al., *Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling.* Proc Natl Acad Sci U S A, 2009. **106**(34): p. 14716-21.
33. Kerr, M.K. and G.A. Churchill, *Statistical design and the analysis of gene expression microarray data.* Genetical Research, 2001. **77**(2): p. 123-128.
34. Qin, S., et al., *Effect of Normalization on Statistical and Biological Interpretation of Gene Expression Profiles.* Frontiers in Genetics, 2012. **3**.
35. Achiron, A. and M. Gurevich, *Peripheral blood gene expression signature mirrors central nervous system disease: the model of multiple sclerosis.* Autoimmun Rev, 2006. **5**(8): p. 517-22.
36. Wojakowski, W., et al., *Mobilization of CD34/CXCR4+, CD34/CD117+, c-met+ stem cells, and mononuclear cells expressing early cardiac, muscle, and*

endothelial markers into peripheral blood in patients with acute myocardial infarction. Circulation, 2004. **110**(20): p. 3213-20.
37. Sinnaeve, P.R., et al., *Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease.* PLoS ONE, 2009. **4**(9): p. e7037.
38. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-747.
39. Pastinen, T., B. Ge, and T.J. Hudson, *Influence of human genome polymorphism on gene expression.* Human Molecular Genetics, 2006. **15 Spec No 1**: p. R9-16.
40. Liu, Y.J., et al., *Is replication the gold standard for validating genome-wide association findings?* Plos One, 2008. **3**(12): p. e4037.
41. Gibson, G. and B. Weir, *The quantitative genetics of transcription.* Trends in Genetics, 2005. **21**(11): p. 616-623.
42. Hsu, J. and J.D. Smith, *Genome-wide studies of gene expression relevant to coronary artery disease.* Curr Opin Cardiol, 2012. **27**(3): p. 210-3.
43. Rotival, M., et al., *Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans.* PLoS Genet, 2011. **7**(12): p. e1002367.
44. Mecham, B.H., P.S. Nelson, and J.D. Storey, *Supervised normalization of microarrays.* Bioinformatics, 2010. **26**(10): p. 1308-1315.
45. Chaussabel, D., et al., *A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus.* Immunity, 2008. **29**(1): p. 150-64.
46. Haumer, M., et al., *Association of neutrophils and future cardiovascular events in patients with peripheral artery disease.* J Vasc Surg, 2005. **41**(4): p. 610-7.
47. Kaustubh C Dabhadkar, A.J.D., Jawahar L Mehta, Arshed A Quyyumi, Peter W Wilson. *Neutrophil to Lymphocyte Ratio is associated with All-cause, Cardiovascular and Ischemic Heart Disease Mortality: the National Health and Nutrition Examination Survey*. in *American Heart Association Scientific Sessions*. 2012. Los Angeles: Circulation.