DNA METHYLATION AND GENE EXPRESSION OF P2RY12 IN A MULTI-ETHNIC COHORT

BY

MICHAEL WESLEY CAMMARATA, MD

A Thesis Submitted to the Graduate Faculty of

WAKE FOREST UNIVERSITY GRADUATE SCHOOL OF ARTS AND SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In Clinical and Population Translational Sciences

August 2015

Winston-Salem, North Carolina

Approved By

David Herrington, M.D., MHS., Advisor

Timothy Howard, PhD., Chairperson

Yongmei Lui, M.D., PhD.

Fang-Chi Hsu, PhD.
ACKNOWLEDGEMENTS

David Herrington, M.D., MHS.

Yongmei Lui, M.D., PhD.

Fang-Chi Hsu, PhD.

Timothy Howard, PhD.

Georgia Saylor

Elsayed Z. Soliman, M.D., M.Sc., M.S.

Gregory L. Burke, M.D.

The MESA Epigenomics & Transcriptomics Study was funded by NHLBI grant R01HL101250 to Wake Forest University Health Sciences.
# TABLE OF CONTENTS

I. **LIST OF ILLUSTRATIONS AND TABLES** ............................................................... iv  
II. **LIST OF ABBREVIATIONS** ................................................................................. vi  
III. **ABSTRACT** ..................................................................................................... vii  
IV. **CHAPTER I – BACKGROUND** ......................................................................... 1  
V. **CHAPTER II – METHYLATION AND EXPRESSION OF P2Y12 IN A**  
   **MULTI-ETHNIC COHORT** .................................................................................. 24  
VI. **CHAPTER III – ADDITIONAL TOPICS AND ANALYSES** .............................. 50  
VII. **CURRICULUM VITAE** .................................................................................... 68
LIST OF ILLUSTRATIONS AND TABLES

Chapter I

• Figure 1 - The Metabolic Pathway of Oral Anti-Platelet Agents .................. 3

Chapter II

• Table I - Participant Baseline Characteristics and study variables .......... 32
• Table II - Relationship between methylation of potential cis-acting CpG loci with P2RY12 expression, significant loci from individual locus modeling .................. 33
• Figure 1 - -Log P value vs. Beta coefficient significant loci individual locus modeling .......................................................... 33
• Table III - Relationship between methylation of potential cis-acting CpG loci with P2RY12 expression, results from backward selection multivariable linear regression model .................. 34
• Table IV - Beta values of backward selected methylation loci by race .......... 34
• Figure 2 - Gene map with significant CpG loci from backward selection modeling .......................................................... 37
• Figure 3 - Gene transcript expression of P2RY12 by race ....................... 37
• Supplemental Table I - Analysis of baseline characteristics vs. P2RY12 gene transcript concentration .................. 42
• Supplemental Table II - Relationship between methylation of potential cis-acting CpG loci with P2RY12 expression, loci from individual locus modeling .................. 42
Chapter III

- Table I: Significant variables in model after different selection techniques utilized ................................................................. 54

- Figure I: Type I (A) and II (B) error rates plotted against number of predictor values across different methods for correcting multiple comparisons ................................................................. 63

- Figure II: Type I and II error rates compared against number of predictor variables while using (A) false discovery rate, (B) Bonferroni correction, (C) modified Bonferroni correction, and (D) nominal p-value of 0.05 ................................................................. 63

- Table II: Alpha and beta values by number of total and significant predictors and method of correction for multiple comparisons .......... 64

- Table III: Primary analysis repeated using four different adjustments for multiple comparisons................................................................. 65
LIST OF ABBREVIATIONS

ACS - Acute Coronary Syndrome
ADP - Adenosine Diphosphate
ANOVA - Analysis of Variance
CAD - Coronary Artery Disease
CPTP - Cyclopentyl-triazolo-pyrimidines
CYP - Cytochrome P450
DNA - Deoxyribonucleic Acid
FDR - False Discovery Rate
HTPR - High on-treatment Platelet Reactivity
MESA - Multi-Ethnic Study on Atherosclerosis
MLR - Multiple Linear Regression
NSTEMI - Non-ST Elevation Myocardial Infarction
P2RY12 - P2Y12 ADP Receptor Gene
PCI - Percutaneous Coronary Intervention
PLA - Platelet Light Aggregometry
PRU - Platelet Reactivity Unit
RIN - RNA Integrity Score
RNA - Ribonucleic Acid
SNP - Single Nucleotide Polymorphism
STEMI - ST Elevation Myocardial Infarction
UA - Unstable Angina
ABSTRACT

Background: P2Y12 receptor is the target for thienopyridine mediated platelet inhibition. Epigenetic factors, including DNA methylation could influence gene expression and subsequent in vitro and clinical response to thienopyridine therapy. Our aim was to determine the relationships between DNA methylation at cis acting CpG sites and P2RY12 gene expression.

Methods: 1264 MESA participants (590 White, 272 Black, 402 Hispanic) underwent methylomic and transcriptional profiling of monocytes using Illumina HumanMethylation450 BeadChip and HumanHT-12 v4 Expression BeadChip respectively. Data were adjusted for chip, chip position, age, gender, race, MESA site and contamination with non-monocytes. Single locus regression was performed between the 58 methylation loci within 500KB of the P2RY12 gene and gene expression levels. Multiple linear regression with backwards elimination procedure was used to identify a subset of the 58 CpG sites independently contributing to prediction of gene expression.

Results: Of the 59 loci, 11 were nominally significantly associated with gene expression in univariate analyses. Multiple linear regression identified 12 loci significantly associated with gene expression at the p=0.05 level with an overall model $r^2$ of 0.12. The 3 most significant CpG loci were significant in both models and would survive correction for multiple comparisons using a Bonferroni correction. In race/ethnicity stratified analyses, different loci were significant in each race/ethnic group, though all models had $r^2$ values of 0.08-0.15.

Conclusion: Variation in degree of methylation at numerous CpG sites in the vicinity of the P2RY12 gene were associated with P2Ry12 mRNA expression. Association
between individual CpG loci and expression varied by race/ethnic group. Further studies are needed to determine whether methylation at these sites significantly influences platelet function or clinical responses to thienopyridines.
CHAPTER I - BACKGROUND

Coronary artery disease (CAD) is one of the most common medical ailments in the United States, affecting approximately 6% of the population.\textsuperscript{1} Though the prevalence of disease has decreased over the last 5 years, incidence rates of many CAD risk factors are on the rise.\textsuperscript{1} The use of percutaneous coronary intervention (PCI) is standard therapy for some patients with stable CAD and for all patients with acute coronary syndromes (ACS) including unstable angina (UA), non-ST elevation myocardial infarction (NSTEMI), and ST-elevation myocardial infarction (STEMI).\textsuperscript{2} In the US alone, there are approximately 600,000 stents placed annually.\textsuperscript{2} While generally viewed as a safe practice, there are complications to such procedures. Stent thrombosis, which occurs when a thrombus forms inside of a placed stent, is one of the most devastating complications.\textsuperscript{3} Rates of stent thrombosis have improved since the advent of dual antiplatelet therapy from 5% or higher to ~1-2% in the 12 months following stent placement. However, stent thrombosis still has similar complications and prognosis to acute STEMI, including potential for development of congestive heart failure, papillary muscle rupture, ventricular wall rupture, and death.\textsuperscript{4,5} Stent thrombosis has cost the United States over $100 million per year for direct hospital-based costs alone (2012 dollars).\textsuperscript{3} The highest risk of stent thrombosis occurs within the first 30 days after placement of a stent and decreases over time with an overall rate of ~3.3 percent over the four years following the procedure.\textsuperscript{6}

Combination therapy including a thienopyridine medication, such as clopidogrel, and aspirin is the current standard of care for patients undergoing PCI. These therapies are essential to prevent thrombotic events following admission for acute coronary syndrome.\textsuperscript{7} In the era since drug eluting stents became widely used, along with the use of dual antiplatelet therapy, the incidence of stent thrombosis has decreased from >5%
to ~1-2% at one year. Unfortunately, despite receiving the recommended doses of aspirin and clopidogrel, some patients will still experience stent thrombosis. Thienopyridine resistance syndromes are one important factor contributing to these potentially life-threatening adverse events.

Racial disparities in rates of stent thrombosis have also been recognized. In a meta-analysis comparing rates of stent thrombosis in the African American population to Caucasian populations, it was found that the rates of stent thrombosis are 3.67% versus 1.25%, respectively. This is despite medication adherence rates in the African American study population equal to or greater than those seen in the Caucasian population. This study was controversial given the retrospective design, as well as difference in risk factors between the two populations. One further retrospective study, however, confirmed these results. These studies showed that race/ethnicity could play an important role in the epidemiology of stent thrombosis and/or clopidogrel resistance syndrome. These studies notwithstanding, African Americans have been historically underrepresented in clinical trials assessing the efficacy of thienopyridine medications. Thus, the fund of knowledge about the effects of thienopyridine and the prevalence of thienopyridine resistance in AA is less well developed that in Caucasians.

**METABOLISM OF ORAL ANTIPLATELET MEDICATIONS**

The thienopyridine metabolism pathway is complex and involves many organ systems and proteins encoded by multiple genes (Figure 1). Understanding of this pathway helps guide targets for research into resistance syndromes.

Clopidogrel has the most complex pathway of the oral antiplatelet medications and involves the most intermediate steps, including the initial absorption step which is
mediated by the ABCB1 transport enzyme in the intestinal epithelial cell. There it undergoes conversion to inactive metabolite (~85%) by esterases. After absorption, the drug then passes through the liver and undergoes metabolism by multiple cytochrome P450 enzymes (primarily CYP2C19), in

**Figure 1 – The Metabolic Pathway of Oral Anti-Platelet Agents.**
a 2-step process, leading to a small fraction being converted to active metabolite while most of the drug is inactivated.\textsuperscript{10-12} Cytochrome P450 (CYP) 2C19 is the most studied of these enzymes and has been implicated in clopidogrel resistance syndromes.

Prasugrel (a more novel thienopyridine) bypasses the intestinal absorption protein and is rapidly metabolized into a thiolactone metabolite. However, it must still undergo conversion in the liver to an active metabolite primarily through the CYP3A4 and CYP2B6 enzymes. Multiple cytochrome p450 enzymes have activity in this conversion step and thus prasugrel is less dependent on the 2C19 enzyme for its conversion into an active metabolite and therefore has more predictable activity than does clopidogrel.\textsuperscript{12,13}

The most novel oral antiplatelet medication currently available (ticagrelor) is itself an active metabolite and requires no further enzymatic changes. This non-thienopyridine medication is the first in a new chemical class of cyclopentyl-triazolo-pyrimidines (CPTP).\textsuperscript{14} Importantly, the mechanism of action for all of the oral antiplatelet medications intersect at a final common site, the P2\textsubscript{Y\textsubscript{12}} ADP Receptor, encoded on the P2RY12 gene, where they antagonize the binding of ADP to the receptor, thus inhibiting platelet activation and thrombosis.\textsuperscript{11}

**THIENOPYRIDINE RESISTANCE**

Previous attempts to classify individual patients as high vs. low risk for stent thrombosis have produced mixed results. The various mechanisms investigated to-date include mutations of the genes encoding for the three main proteins involved in the metabolism of clopidogrel (ABCB1, CYP2C19, P2RY12) and use of point of care platelet function assays such as VerifyNow.
CYP2C19

Proton pump inhibitors were, at one point, believed to competitively inhibit the CYP2C19 enzyme as well.\textsuperscript{15} Because Omeprazole is metabolized using CYP2C19 enzyme it was believed that occupying the enzyme would allow more clopidogrel substrate to be deactivated by other enzymes involved in catabolism of the medication. Further investigation has revealed that omeprazole does, in fact, decrease clopidogrel effect in platelet function assays leading to high on-treatment platelet reactivity (HTPR).\textsuperscript{16} Clinical trials have been widely varied in their results and large meta-analysis while showing minimal increased risk, demonstrated the wide variation in outcomes of studies measuring the effect of omeprazole on the action of clopidogrel.\textsuperscript{17}

Further, a “smoker's paradox” exists in the efficacy of clopidigrel therapy whereby clopidogrel therapy seems to be more effective in participants who smoke than their non-smoking counterparts. Gurbel et.al. described this phenomenon and hypothesized that smoking had an activating effect on the CYP2C19 enzyme\textsuperscript{18} which was further supported by the PARODOX clinical trial where it observed that participant who smoked had higher platelet inhibition than those who did not smoke.\textsuperscript{19}

The initial genetic studies aimed at predicting clopidogrel resistance investigated CYP2C19 enzyme loss of function mutations. The CYP2C19*2 variant was the most common of four loss of function alleles identified and became the most commonly studied.\textsuperscript{20-22} This variant allele contains a loss of function mutation (681G-A) which renders the enzyme inactive. Patients homozygous for the variant allele undergoing PCI had an increased risk of subsequent cardiovascular events (HR 1.98).\textsuperscript{23} Specifically, there was an increased risk of stent thrombosis associated with homozygosity of CYP2C19*2 (1.5% vs 0.4%) in a study with low number of events.\textsuperscript{24} While genetic
mutation in this gene was initially promising, it failed to translate to a clinically viable means of determining clopidogrel resistance. Further studies indicated that though they did not depend on the CYP2C19 enzyme for their activation, the more novel anti-platelet agents had resistance syndromes as well. The PLATO trial demonstrated that ticagrelor still had a significant degree of stent thrombosis at 1 year, though less than that of clopidogrel (1.3% vs 1.9%). However, despite the decreased rate of stent thrombosis, novel antiplatelet agents have shown an increase in minor and major bleeding risk (16.1% vs 14.6%). Further studies investigating the effect of increased dose of clopidogrel on high risk genotypes revealed that it may be possible to overcome the effect in patients with heterozygosity for CYP2C19*2, but not in homozygotes. In summary, though initially promising, SNP variations in the CYP2C19 enzyme have been mixed in their ability to predict clinical events and stent thrombosis. Further, though increasing dose of medication based on genotype can increase platelet inhibition measured by in-vitro tests of platelet aggregation, this has not translated into a clinically applicable way of preventing stent thrombosis.

**ABCB1**

A key protein involved in thienopyridine absorption in the intestine is the efflux pump P-glycoprotein which is encoded by the ABCB1 gene. Investigation into this protein as a culprit for thienopyridine resistance were initially promising. One study found that carriers of two variant alleles 3435C→T had a higher rate of cardiovascular events (death from any cause, nonfatal stroke, or myocardial infarction) at 1 year than those with the ABCB1 CC phenotype. Individuals homozygous for 3435C→T mutation were shown to have increased risk for cardiovascular events after stenting. They had decreased platelet inhibition after loading dose of clopidogrel, and these findings were independent of the effect of the CYP2C19 mutations. Differing data were later
presented showing possible protective association on the carrier status of this mutation. The GIFT study provided further evidence for lack of association between ABCB1 status and clopidogrel pharmacodynamics. Therefore, while absorption of clopidogrel through the ABCB1 pathway was initially thought to be an important step in the metabolism and action of clopidogrel, this has not translated clinically into prediction or prevention of adverse outcomes.

**P2RY12**

The final common point in the pathways of all the oral anti-platelet agents is the P2Y12 G-couple receptor for ADP, encoded by the P2RY12 gene, and which primarily resides on the platelet cell surface. As few studies have been performed to evaluate the effect of genetic variants of the P2RY12 gene and clinical outcomes, study of this protein was the next logical step. Several rare genetic syndromes involving P2RY12 have been reported to cause a mild bleeding diathesis. One report of deficiency of P2Y12 was published and described a lifelong bleeding disorder characterized by the inability of ADP to fully reverse platelet function. Other agents (thrombin or collagen) were able to fully reverse the dysfunction in-vitro. Others patients have been identified with genetic mutations of the receptor who also have bleeding disorders, similarly irreversible by ADP administration in-vitro. Results of studies investigating genetic mutations of the P2RY12 locus and clinical outcomes have been mixed. In the FAST-MI registry, none of the selected polymorphisms were found to have an association with risk for combined outcome of death from any cause, nonfatal myocardial infarction, or stroke. However, in further studies an association was found between several P2RY12 SNPs and drug resistance when measured by the VerifyNow point of care platelet function testing assay. While results of clinical studies are mixed using the P2RY12 enzyme as a predictor, analysis of this enzyme continues to be a promising area of research given the
history of bleeding diatheses associated with both variations in level and in structure/function of the receptor.

**VERIFYNOW**

Inhibition of platelet function in response to thienopyridine therapy has long been thought to mediate clopidogrel resistance. However, the gold standard of platelet function, platelet light aggregometry (PLA), is a time consuming and technically difficult test to perform. The VerifyNow Point of care platelet function assay (the most widely used point of care assay) was developed and has been shown to be strongly correlated with platelet light aggregometry for both peak and late inhibition ($r=0.75$ and $0.73$ respectively). The VERITAS study found large inter-individual variations in the level of platelet inhibition using VerifyNow assay and it was suspected that measuring platelet function after initiation of clopidogrel would help predict individuals with resistance. Further investigation showed that while there was some predictive ability of the point of care testing, it only carried a sensitivity of 78% and specificity of 68% using a platelet reactivity units (PRU) level of $>234$. However, this study only had only two stent thrombosis events. While there was some evidence of predictive ability, further studies found that no events were avoided by doubling drug dose or by reducing PRU value below the 235 threshold. Therefore, point of care platelet function assays have been a reliable measure of platelet reactivity especially when compared to the gold standard PLA. Unfortunately, while platelet function as a whole has had some success in the prediction of thienopyridine resistance syndromes, it has not yielded clinically meaningful results in the prevention of the disease. While platelet function may be a clinical indicator of risk, it has not been shown to adequately predict clinical resistance to thienopyridine or other P2RY12 targeting therapies.
EPIGENETICS IN HUMAN DISEASE

While genetic variation has been proven important in human disease, it has failed to explain much of the variation in most complex disease states, including thienopyridine resistance syndromes. Due to the heterogeneity of results in the prediction of stent thrombosis or adverse outcomes using platelet function and genotyping, novel techniques may be better able to aid in prediction of resistance syndromes. An emerging area of research involves the analysis of epigenetic modifications. Epigenetic modification including methylation of the genome at specific sites has great importance in the regulation of gene expression. DNA methylation occurs on the cytosine residue of some CpG dinucleotides throughout the genome. It is generally observed that methylation of the promoter region of a gene down regulates expression and methylation within a gene is believed to increase the expression of that gene.

Two main techniques have been developed to determine which residues are methylated and to what extent. The first of these techniques is through the DNA sequencing method known as pyrosequencing. Pyrosequencing relies on the detection of release of pyrophosphate during the incorporation of DNA base rather than due to the termination of transcription as in traditional sequencing modalities. This is accomplished through the enzymatic conversion of pyrophosphate into a light signal. When DNA is treated with bisulfite, cytosine residues are converted into uridine residues, unless they are methylated. Then DNA amplification is performed using PCR. By comparing the degree to which one locus is converted, the ratio of DNA strands that are methylated at that locus are can be determined. Pyrosequencing has the advantage of being able to be performed in any area of the genome so long as a primer is made to sequence, and it is fairly inexpensive to perform per test. However, it suffers in that it can analyze DNA methylation in a region limited to 100 base pairs.
Newer technology has made available a hybridization array chip-based method of evaluating DNA methylation. These chips have the advantage of being able to look at CpG loci throughout the entire genome in one reaction. The current version of a chip produced by the Illumina Corporation is able to determine methylation levels of >485,000 sites. Bead based chips such as this use microscopic beads which bind a specific CpG site and are read through fluorescence determined by a map provided with the chip. These technologies correlate well with both pyrosequencing reactions and with older bead chip platforms, are able to be processed rapidly, and can very useful in methylome-wide studies. However, they may be prohibitively expensive on small scale studies.

A protein must be transcribed and translated in order for it to have any effect on cell biology. Novel methods have been developed to measure the degree to which a gene is transcribed. Three methods have been widely accepted to determine the degree to which RNA is produced for a specific gene. RNA can be quantitatively measured using northern blot technique. RNA must first be separated and purified, then can be separated with the use of an agarose gel and then radioactively labeled using a complementary probe to the RNA of interest. Autoradiography can then be performed. This is a less widely used technique recently due to the time consuming nature of the work, as well as the use of dangerous regents used in the method. RNA may also be quantified using a real-time quantitative polymerase chain reaction (RT-qPCR) in the “TaqMan” assay. In this method, an RNA transcript will be reverse transcribed by Taq polymerase. In subsequent steps, a DNA based primer is used in a PCR reaction to amplify the cDNA transcript and by use of florescent tag level then fluorescence can be measured using a standard competitive transcript for optimal quantification of transcript.
Modern gene transcription chips have been designed for broader quantification of the gene transcript levels of thousands of genes simultaneously. Similar to the DNA methylation chip, it uses microbeads attached to transcript specific probes and uses computer analysis to quantify the amount of RNA transcript present. This method, while relatively inexpensive per gene, quantifies thousands of gene transcript levels simultaneously. For small studies, it may be prohibitively expensive at this time, though it is minimally labor intensive and allows for correlation in genome wide association studies and/or methylome wide association studies.\(^{43}\)

The use of these novel techniques have already proven useful in cancer biology. Cancer lines have been noted to have a large variation in the degree of methylation between different types of cancer and when compared with non-cancerous cells.\(^{37,44}\) Addition hyper-methylation has been noted at CpG islands associated with tumor suppressor genes and those involved in the process of apoptosis.\(^{37}\) Epigenetic profiles of some cancer types have also been associated with poor prognosis for response to therapeutic agents. Increased risk for disease progression/metastasis has also been reported.\(^{45}\) It has also been hypothesized that by changing DNA methylation profiles targeted to specific types of cancers, better outcomes may be achieved.\(^{46}\) Novel techniques have recently been developed (though in the early stages of testing) based on their ability to effect the DNA methylation profile of certain genes. Two medications, Azacitidine and decitabine, are DNA methyltransferase inhibitors. Another medication, entinostat, is a Multiple Histone Deacetylase inhibitor. While only phase I clinical trial data is currently available, these medications appear promising even on this small scale level.\(^{47}\) While still in early stages of discovery, epigenetic modification in the form of DNA methylation has already become a promising area of investigation in cancer genetics and is beginning to be utilized in different fields as well, including with some
common neurological disorders. Hypermethylation of the NEP gene has shown some association with Alzheimer’s dementia. Conversely, hypomethylation of the PADI2 gene has been associated with multiple sclerosis.

Though most of the epigenetic associations, prognostics, and therapeutics currently recognized have been within the field of cancer medicine, other areas have been investigating this as well, including cardiovascular medicine. Initial investigation in the PDAY cohort study found no association between global levels of DNA methylation and subclinical atherosclerosis. However, this study was in very young individuals and viewed the genomes level of methylation as a whole rather than individual genes. More recent investigation has implicated DNA methylation of the FOXP3 gene as a possible mechanism in the progression of coronary artery disease. Methylation of two genes (LY75 and ADORA2A) has also been associated with dilated cardiomyopathy. These genes were subsequently shown in a knockout model to have a functional role in dilated cardiomyopathy in a zebra fish model. Data are currently scarce regarding the prognostic value of epigenetic modifications in cardiovascular disease or in possible therapeutics related to these novel risk factors.

Little has been documented to-date on the role of epigenetic modification or gene expression in thrombosis/coagulation. Unfortunately, one of the key cellular players in thrombosis (platelets) does not contain genomic DNA, and therefore epigenetic studies have been particularly difficult to perform. Since epigenetic changes tend to be cell-specific, the ideal cell type to address the effects of epigenetic changes on platelet reactivity in response to therapy would be the platelets themselves, or more precisely the nucleated precursor to platelets – megakaryocytes. It has been hypothesized, however, that surrogate cell lines could be used, specifically those of hematopoietic lineage closely related to platelets. Monocytes, for example, have long been known to
be involved in the process of thrombosis and are known to be a source of many of the coagulation factors. Monocytes also bind to platelets and form highly thrombogenic platelet-monocyte complexes. These cells have also more broadly been implicated in the pathogenesis of common cardiovascular diseases including coronary artery disease (subclinical atherosclerosis through myocardial infarction) and risk factors including diabetes and dyslipidemias. A suspected role in venous thromboembolism has also been described. Monocytes have proven to be a very reliable and easily accessible source of DNA for studies on epigenetic modifications and their role in human disease states.
SPECIFIC AIMS:

Currently, the effects of epigenetic modifications on expression of genes relevant for thienopyridine action remain uncertain. Monocytes are a much more easily accessible cell type than platelets and are also known to express the P2RY12 gene and participate in thrombosis and hemostasis. The MESA Epigenetic Ancillary study has collected genome-wide methylation data and gene expression data in purified monocytes from ~1260 participants. Therefore, the overall goal for this proposal is to determine if epigenetic modifications in the proximity of the P2RY12 gene alter its expression. If confirmed, this finding would justify more extensive in-vitro and clinical evaluations of the effects of these epigenetic modifications on response to thienopyridine therapy.

Specific Aim 1: To evaluate the association between methylation of individual CpG loci within 500 Kb of the P2RY12 gene and expression of the P2Y12 gene product in monocytes.

Null-Hypothesis 1: Individual DNA methylation levels at CpG sites within 500 KB of the P2RY12 gene will have no association with expression of the gene in peripheral monocytes.

Specific Aim 2: To determine what proportion of the variability in P2RY12 gene expression is accounted for by methylation of all relevant cis-acting DNA methylation sites.

Null-Hypothesis 2: No association exists between the combined effects of DNA methylation

Primary outcome: Gene transcript expression of P2RY12 in peripheral monocytes
**Secondary Aim 1:** To determine whether race/ethnic group is associated with the degree of expression of the P2RY12 gene.

**Null-Hypothesis:** Race/ethnic group does not have an association with expression of the P2RY12 gene.

**Secondary Aim 2:** To evaluate whether race/ethnic group acts as an effect modifier of the association between DNA Methylation levels in their association with gene expression of the P2RY12 gene.

**Null-Hypothesis:** Race/ethnic group does not have an effect modification of the association between DNA methylation within 500 KB of the P2RY12 gene and expression of that gene.
REFERENCES


CHAPTER II – METHYLATION AND EXPRESSION OF P2RY12 IN A MULTI-ETHNIC COHORT

INTRODUCTION

Combination therapy with a thienopyridine medication, such as clopidogrel and aspirin is the current standard of care for patients undergoing percutaneous coronary intervention following admission for acute coronary syndrome.\textsuperscript{1,2} Despite receiving recommended doses of aspirin and clopidogrel, some patients still experience in-stent thrombosis. One factor potentially contributing to these adverse events is clopidogrel resistance possibly due to genetic variants that alter drug activation or modify the structure or abundance of its target receptor.\textsuperscript{3-5} A common single nucleotide polymorphism (SNP) in the CYP2C19 gene was the first and most widely recognized genetic factor thought to affect platelet response to clopidogrel.\textsuperscript{6} More recently, SNPs in the P2Y12 ADP receptor gene (P2RY12), the primary target for the active metabolite of thienopyridine, have also been evaluated. Although there is some evidence that these variants may alter platelet response to thienopyridines in-vitro \textsuperscript{7,8}, to-date their effect on clinical treatment response is minimal or not yet adequately evaluated.

Recent advances in molecular techniques now make it possible to also determine if epigenetic factors such as variations in methylation of CpG sites may also alter risk for clinical phenotypes or response to therapies. Several examples of epigenetic changes (most commonly DNA methylation) are known to alter abundance or translation of factors altering drug metabolism. One of the most relevant examples, an inverse correlation between the degree of methylation in the promoter region and level of protein expressed in the transporter protein \textit{ABCB1}. In addition, multiple examples of inter-individual and inter-tissue variability in cytochrome levels have been described and at least partially explained by variable DNA methylation.\textsuperscript{9-11} To-date the effects of
Epigenetic modifications on expression of genes relevant for thienopyridine action have yet to be described.

Epigenetic patterns tend to be cell specific, so the ideal cell type to address the effects of epigenetic changes on P2RY12 would be platelets, or more precisely their nucleated precursor [megakaryocytes]. However, monocytes are a much more easily accessible cell type that are also known to express the P2RY12 gene and participate in thrombosis and hemostasis. The goal of this investigation is to determine if association exists between DNA methylation of potentially cis-acting CpG sites in monocytes and expression of the cognate P2RY12 mRNA transcript. If verified, such a finding could justify additional clinical studies to determine if methylation of the relevant CpG sites are associated with adverse outcomes, or could be used to adjust dosing in patients being treated with thienopyridines.

METHODS

STUDY POPULATION

The Multi-Ethnic Study of Atherosclerosis (MESA) was designed to investigate the prevalence, correlates, and progression of subclinical cardiovascular disease in a population cohort of 6,814 participants. Since its inception in 2000, five clinic visits collected extensive clinical, socio-demographic, lifestyle and behavior, laboratory, nutrition, and medication data. The present analysis uses data from purified monocyte samples from the April 2010-February 2012 MESA examination (exam 5) of 1264 randomly selected MESA participants (55-94 years old, Caucasian (47%), African American (21%), Hispanic (32%), and female (51%)) from four MESA field centers (Baltimore, MD; Forsyth County, NC; New York, NY; and St. Paul, MN). The study
protocol was approved by the Institutional Review Board at each site. All participants underwent the informed consent process.

**SAMPLE PREPARATION**

*Purification of Monocytes*

Centralized training of technicians, standardized protocols, and extensive quality control (QC) measures were implemented for collection, on-site processing, and shipment of MESA specimens and routine calibration of equipment. Blood was initially collected in sodium heparin-containing Vacutainer CPTTM cell separation tubes (Becton Dickinson, Rutherford, NJ) to separate peripheral blood mononuclear cells from other elements within 2 hours from blood draw. Subsequently, monocytes and T cells were isolated with anti-CD14 and anti-CD4 monoclonal antibody coated magnetic beads, respectively, using AutoMACs automated magnetic separation unit (Miltenyi Biotec, Bergisch Gladbach, Germany). Based on flow cytometry analysis of 18 specimens, monocyte samples were consistently > 90% pure.

*DNA/RNA extraction*

DNA and RNA QC metrics included optical density measurements, using a NanoDrop spectrophotometer and evaluation of the integrity of 18s and 28s ribosomal RNA. The rRNA was found to be of high quality by analysis of electropherogram with two distinct bands. Additional RNA QC testing was performed using the Agilent 2100 Bioanalyzer with RNA 6000 Nano chips (Agilent Techonology, Inc., Santa Clara, CA, USA) following manufacturer’s instructions. RNA with RIN (RNA integrity) scores ≥9.0 (on a scale of 1-10) has previously been used to denote a high quality sample with minimal degradation.\(^{16}\) This was used as the cutoff for global expression microarrays. The median RIN for our 1264 samples was 9.9.
Gene expression quantification

Illumina Human HT-12 v4 Expression BeadChip and Illumina Bead Array Reader were used to perform the Genome wide expression analysis, following the Illumina expression protocol. The Illumina TotalPrep-96 RNA Amplification Kit (Ambion/Applied Biosystems, Darmstadt, Germany) was used for reverse transcription, and amplification with 500 ng of input total RNA (at 11 ml). An amount of 700 ng of biotinylated cRNA was hybridized to a BeadChip at 58°C for 16–17 h. To avoid potential biases due to batch, chip and position effects, a stratified random sampling technique was used to assign individual samples (including 24 common control samples) to specific BeadChips (12 samples/chip) and chip position.

Quantification of DNA Methylation

The Illumina HumanMethylation450 BeadChip and HiScan reader were used to perform the epigenome-wide methylation analysis. The EZ-96 DNA MethylationTM Kit (Zymo Research, Orange, CA, USA) was used for bisulfite conversation with 1 µg of input DNA (at 45°C). An amount of 4 µl of bisulfite-converted DNA were used for DNA methylation assays, following the Illumina Infinium HD Methylation Protocol. This consisted of a wholegenome amplification step followed by enzymatic end-point fragmentation, precipitation and resuspension. The resuspended samples were hybridized on HumanMethylation 450 BeadChips at 48°C for 16 h. The individual samples were assigned to the Bead-Chips and to chip position, using the same sampling scheme as that for the expression BeadChips.

Pre-processing and QC of Microarray Data

Data pre-processing and QC analyses were performed in R (http://www.r-project.org/) using Bioconductor (http://www.bioconductor.org/) packages. For
expression data, data corrected for local background were obtained from Illumina’s proprietary software GenomeStudio. QC analyses and bead-type summarization (average bead signal for each type after outlier removal) were performed using the beadarray package. Detection P-values were computed using the negative controls on the array. The neqc function of the limma package was used to perform a normal–exponential convolution model analysis to estimate non-negative signal, quantile normalization using all probes (gene and control, detected and not detected) and samples, addition of a recommended (small) offset, log2 transformation and elimination of control probe data from the normalized expression matrix. Multidimensional scaling plots showed that the five common control samples were highly clustered together and identified three outlier samples, which were excluded subsequently.

The Illumina HumanHT-12 v4 Expression BeadChip included 48K transcripts. Statistical analyses were limited to probes retained after applying the following QC elimination criteria: probes for the X or the Y chromosome, non-detectable expression in ≥90% of MESA samples using a detection P-value cut-off of 0.0001, existence of any known SNPs, overlap with a repetitive element or region, low variance across the samples (≤10th percentile) or putative and/or not well-characterized genes, i.e. gene names starting with KIAA, FLJ, HS, Cxorf, MGC or LOC. We included 8370 autosomal gene transcripts for analysis.

Bead-level methylation data were summarized in GenomeStudio. Because a two-channel system and both Infinium I and II assays were used, normalization was performed in several steps using the lumi package. We first adjusted for color bias using ‘smooth quantile normalization’. Next, the data were background-adjusted by subtracting the median intensity value of the negative control probes. Lastly, data were normalized across all samples by standard quantile normalization applied to the bead-
type intensities and combined across Infinium I and II assays and both colors. QC measures included checks for sex and race/ethnicity mismatches, and outlier identification by multidimensional scaling plots. The final methylation value for each methylation probe was computed as the M-value, essentially the log ratio of the methylated to the unmethylated intensity. The M-value is well suited for high-level analyses and can be transformed into the beta-value, an estimate of the percent methylation of an individual CpG site that ranges from 0 to 1 \( M = \text{logit}(\text{beta-value}) \).

The Illumina HumanMethylation450 BeadChip included probes for 485K CpG sites. Of these 485K CpG sites, 416,507 passed the QC elimination criteria including: probes for the X or the Y chromosome, ‘detected’ methylation levels in 90% of MESA samples using a detection P-value cut-off of 0.05, existence of any SNPs within 10 base pairs of the targeted CpG site, location outside of the 1 MB intervals on both sides of each gene or overlap with a repetitive element or region.

Pre-processing with global normalization removed large position and chip effects across all probes; however, probe-specific chip effects were found for some CpG sites and gene transcripts, whereas probe-specific position effects existed for some CpG sites but were ignorable for all gene transcripts. These probe specific effects were included as covariates in all subsequent analyses.

**CpG Locus Selection**

Methylation site identification was performed using the UCSC Genome Browser (http://genome.ucsc.edu). A search was performed for the P2RY12 gene. A field containing ± 500 KBP was obtained. The ENCODE DNA methylation database was overlaid on the field. All CpG loci within ±500 KBP from the start codon of the P2RY12 gene (chromosome 3: 151,102,544) which were available on the Illumina
HumanMethylation 450 BeadChip were including in the analysis. A total of 59 methylation loci fit this description. Functional analysis of the significant loci was performed by using data from the ENCODE project accessed through the UCSC Genome Browser. Overlaying a map for the histone H3 lysine 4 mono-methylation (H3K4me1) modification allowed for identification of CpG loci in regulatory regions whereas overlaying of the histone H3 lysine 4 tri-methylation (H3K4me3) modification allowed for identification of loci located in or near promoter regions of other genes. No loci were located within the promoter region of the gene, though; one locus was very near the promoter.

STATISTICAL ANALYSES

Baseline Characteristics/Covariates

Normal distribution of all continuous variables (M-value of CpG loci, log_2 transformed P2RY12 gene transcript levels, and age) was confirmed and all baseline characteristics were compared across race/ethnic groups utilizing ANOVA procedure for continuous variables and chi-square test for categorical values (Table I). Potential confounding variables were identified though individual modeling. All significant variable were adjusted for in further modeling. (Supplemental Table I)

Association Analysis

In an exploratory analysis, to identify potential cis-acting CpG loci individually associated with the cognate gene’s expression, M-values of the 59 identified CpG loci were individually regressed as a function of P2RY12 gene expression while adjusting for significant baseline covariates (age, gender, racial/ethnic group, site at MESA exam 5,
methylation chip/well, enrichment scores). Of the 59 CpG loci, 11 were associated with gene expression at the p<0.05 level (Table II/Figure 1).

A model was then created to determine which loci would best predict P2RY12 gene expression levels in a multi-variable analysis. The 59 previously identified CpG loci were included in this multiple regression analysis. While adjusting for the significant baseline covariates, a backwards selection procedure was utilized. An significant alpha level of <0.05 determined which individual loci would be retained in the model. Adjusted $R^2$ value was reported for each of the significant loci in this model (Table III).

Racial/Ethnic Group Analyses

Initially, ANOVA analysis identified racial/ethnic differences in the levels of P2RY12 expression (Table I). To further investigate this difference and any potential effect modification of race on individual CpG locus methylation, interaction analysis was used. All 59 CpG loci as well as an interaction term for each variable were entered into a multiple linear regression model, adjusting for baseline covariates (which the exception of race/ethnic group). There was a significant interaction between two of these loci and a trend towards significant in three additional loci. Stratified analysis was then performed to determine which individual loci were significantly associated with P2RY12 gene expression in each racial/ethnic group. While adjusting for baseline covariates (except race/ethnic group), a stratified multiple linear regression analysis was preformed again using a backward selection procedure with a significant alpha level of <0.05 to be retained in the model. The results of this analysis were reported in table IV.

All analyses were performed using SAS statistical software version 9.3 (SAS inc., Cary, NC).
Table I. Participant Baseline Characteristics and study variables

<table>
<thead>
<tr>
<th></th>
<th>All Racial/Ethnic groups n=1264</th>
<th>Caucasian n=590 (46.7%)</th>
<th>African American n=272 (21.5%)</th>
<th>Hispanic n=402 (31.8%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yrs</td>
<td>60.2 (±9.5)</td>
<td>60.8 (±9.6)</td>
<td>60.5 (±9.1)</td>
<td>58.9 (±9.4)</td>
<td>0.007</td>
</tr>
<tr>
<td>Gender, Female %</td>
<td>650 (51.4%)</td>
<td>285 (48.3%)</td>
<td>163 (59.9%)</td>
<td>202 (50.3%)</td>
<td>0.005</td>
</tr>
<tr>
<td>Hypertension (95% CI)</td>
<td>493 (39.1%)</td>
<td>264 (44.8%)</td>
<td>64 (23.5%)</td>
<td>165 (41.2%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>255 (20.2%)</td>
<td>73 (12.4%)</td>
<td>86 (31.6%)</td>
<td>96 (23.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Aspirin use</td>
<td>589 (46.6%)</td>
<td>311 (52.7%)</td>
<td>99 (36.4%)</td>
<td>179 (44.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Tobacco use</td>
<td>108 (8.5%)</td>
<td>47 (8.0%)</td>
<td>34 (12.5%)</td>
<td>27 (6.7%)</td>
<td>0.02</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>568 (44.9%)</td>
<td>345 (58.5%)</td>
<td>102 (37.5%)</td>
<td>121 (30.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Log₂ P2RY12 Expression</td>
<td>7.05 (±0.58)</td>
<td>7.14 (±0.56)</td>
<td>6.88 (±0.59)</td>
<td>7.03 (±0.58)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*p-value was calculated using chi-square test for categorical variables and ANOVA for continuous variables. Continuous variables are reported as mean ± standard deviation; categorical variables reported as count (%).
Table II: Relationship between methylation of potential cis-acting CpG loci with P2RY12 expression, significant loci from individual locus modeling.

<table>
<thead>
<tr>
<th>Methylation Site</th>
<th>β (SE)*</th>
<th>Semi-partial $R^2$ (95% CI)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg05094216</td>
<td>0.189 (0.04)</td>
<td>0.019 (0.007, 0.036)</td>
<td>4.5 x10^{-7}</td>
</tr>
<tr>
<td>cg24630764</td>
<td>-0.271 (0.06)</td>
<td>0.017 (0.006, 0.034)</td>
<td>4.2 x10^{-6}</td>
</tr>
<tr>
<td>cg07736012</td>
<td>0.234 (0.06)</td>
<td>0.010 (0.002, 0.024)</td>
<td>0.0003</td>
</tr>
<tr>
<td>cg00472710</td>
<td>0.098 (0.03)</td>
<td>0.009 (0.001,0.022)</td>
<td>0.0006</td>
</tr>
<tr>
<td>cg16780454</td>
<td>0.085 (0.03)</td>
<td>0.008 (0.001,0.020)</td>
<td>0.0013</td>
</tr>
<tr>
<td>cg21766308</td>
<td>0.085 (0.03)</td>
<td>0.007 (0.001,0.019)</td>
<td>0.0032</td>
</tr>
<tr>
<td>cg01754219</td>
<td>0.180 (0.07)</td>
<td>0.005 (0.000, 0.016)</td>
<td>0.0113</td>
</tr>
<tr>
<td>cg09432154</td>
<td>-0.152 (0.06)</td>
<td>0.004 (0, 0.015)</td>
<td>0.0185</td>
</tr>
<tr>
<td>cg09410045</td>
<td>0.132 (0.06)</td>
<td>0.004 (0, 0.014)</td>
<td>0.0222</td>
</tr>
<tr>
<td>cg12775479</td>
<td>0.160 (0.07)</td>
<td>0.004 (0, 0.013)</td>
<td>0.0295</td>
</tr>
<tr>
<td>cg15001056</td>
<td>-0.125 (0.06)</td>
<td>0.003 (0, 0.012)</td>
<td>0.0425</td>
</tr>
</tbody>
</table>

Loci in bold would are most significantly associated, are significant in both models and would survive a correction for multiple comparisons.

*β value represents the predicted change in expression for each unit increase of methylation at that site.

Figure 1: -Log P value vs. Beta coefficient significant loci individual locus modeling.
Table III: Relationship between methylation of potential cis-acting CpG loci with P2RY12 expression, results from backward selection multivariable linear regression model.

<table>
<thead>
<tr>
<th>Methylation Site</th>
<th>( \beta(SE)* )</th>
<th>Semi-partial ( R^2 ) (95% CI)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg05094216</td>
<td>0.232 (0.04)</td>
<td>0.024 (0.010, 0.043)</td>
<td>8.1 x10^{-9}</td>
</tr>
<tr>
<td>cg24630764</td>
<td>-0.290 (0.06)</td>
<td>0.017 (0.006, 0.033)</td>
<td>1.2 x10^{-6}</td>
</tr>
<tr>
<td>cg07736012</td>
<td>0.248 (0.06)</td>
<td>0.011 (0.002, 0.025)</td>
<td>9.4 x10^{-5}</td>
</tr>
<tr>
<td>cg11919271</td>
<td>-0.141 (0.05)</td>
<td>0.006 (0.001, 0.017)</td>
<td>0.004</td>
</tr>
<tr>
<td>cg09410045</td>
<td>0.168 (0.06)</td>
<td>0.005 (0, 0.016)</td>
<td>0.006</td>
</tr>
<tr>
<td>cg01754219</td>
<td>0.185 (0.07)</td>
<td>0.005 (0, 0.016)</td>
<td>0.008</td>
</tr>
<tr>
<td>cg11046673</td>
<td>-0.097 (0.04)</td>
<td>0.005 (0, 0.016)</td>
<td>0.007</td>
</tr>
<tr>
<td>cg00091349</td>
<td>0.122 (0.05)</td>
<td>0.005 (0, 0.015)</td>
<td>0.010</td>
</tr>
<tr>
<td>cg00472710</td>
<td>0.073 (0.03)</td>
<td>0.005 (0, 0.015)</td>
<td>0.009</td>
</tr>
<tr>
<td>cg14191476</td>
<td>0.090 (0.04)</td>
<td>0.003 (0, 0.012)</td>
<td>0.043</td>
</tr>
<tr>
<td>cg17582100</td>
<td>-0.121 (0.06)</td>
<td>0.003 (0, 0.012)</td>
<td>0.043</td>
</tr>
<tr>
<td>cg09432154</td>
<td>-0.136 (0.06)</td>
<td>0.003 (0, 0.012)</td>
<td>0.045</td>
</tr>
</tbody>
</table>

\( r^2 = 0.1 \)

Loci in bold would are most significantly associated, are significant in both models and would survive a correction for multiple comparisons

*\( \beta \) value represents the predicted change in expression for each unit increase of methylation at that site.

Table IV: Beta values of backward selected loci by race

<table>
<thead>
<tr>
<th>White</th>
<th>AA</th>
<th>Hispanic</th>
<th>P-value interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg00137750</td>
<td>0.32334</td>
<td>-0.2635</td>
<td>0.63</td>
</tr>
<tr>
<td>cg01330290</td>
<td>0.29379</td>
<td>0.20281</td>
<td>0.25</td>
</tr>
<tr>
<td>cg01754219</td>
<td>0.30215</td>
<td>0.006 (0, 0.016)</td>
<td>0.004</td>
</tr>
<tr>
<td>cg03379681</td>
<td>0.1181 (0.06)</td>
<td>0.005 (0, 0.015)</td>
<td>0.008</td>
</tr>
<tr>
<td>cg00472710</td>
<td>0.073 (0.03)</td>
<td>0.005 (0, 0.015)</td>
<td>0.009</td>
</tr>
<tr>
<td>cg14191476</td>
<td>0.090 (0.04)</td>
<td>0.003 (0, 0.012)</td>
<td>0.043</td>
</tr>
<tr>
<td>cg17582100</td>
<td>-0.121 (0.06)</td>
<td>0.003 (0, 0.012)</td>
<td>0.043</td>
</tr>
<tr>
<td>cg09432154</td>
<td>-0.136 (0.06)</td>
<td>0.003 (0, 0.012)</td>
<td>0.045</td>
</tr>
</tbody>
</table>

*Beta values significant at the \( p<0.05 \) level reported, after correction for multiple comparisons, no interaction terms remain significant.
RESULTS

The baseline characteristics of the different racial ethnic groups are described in (Table I). There were significant unadjusted differences across race/ethnic groups with respect to gender distribution, presence of hypertension, diabetes, aspirin use, tobacco use and alcohol use. After Bonferroni correction of the p-values of all baseline covariates, the only covariates which remained significant were: age, gender, racial/ethnic group, site at exam 5, enrichment scores and methylation chip, cell line enrichment scores (b-cells, t-cells, NK-cells and neutrophils), methylation chip and chip well, MESA exam 5 site (Table II). These covariates were used to adjust subsequent analyses.

In our exploratory analysis we were able to identify 11/59 significant cis-acting CpG loci (Table II - full table of results including non-associated CpG loci can be found in the Supplemental Table II). Of these loci, three were located in known DNA regulatory regions. The remainder of the loci were located in intronic regions or regions outside of the gene which were not previously identified as being DNA regulatory regions. Figure 1 illustrates the association between significant loci identified and their regression coefficient (beta values).

Our primary analysis identified 12/59 significant cis-acting CpG loci in a multivariable model remaining significant after adjustment for the significant covariates (Table III). Four of these loci were located in known DNA regulatory regions, three were located in intronic regions, and five were located in regions of DNA not previously identified as regulatory regions (Figure 2). These 12 loci were able to explain a total of 12% of the variation in P2RY12 expression as measured by an R² statistic (Table III). Seven loci were significant in both the initial exploratory analysis and in the primary multiple regression analysis. The three most strongly associated loci (cg24630764,
cg05094216, and cg07736012) were strongly significant in both models and were able to account for 5.2% of the overall variability of P2RY12 expression. Of these loci, 2 were located in intronic regions of the gene and the third was located within a DNA regulatory region upstream of the gene start codon. An additional CpG site (cg11919271), located within the promoter region of the gene, was found to become significant in the multivariable model, however, this site has a smaller association, beta coefficient and level of significance than the other three loci.

**Racial/Ethnic Groups**

In the initial evaluation, we identified racial/ethnic difference between expression of P2RY12 in our ANOVA analysis. The analysis was strongly significant (p= 6.54x10^{-9}). Pairwise comparison found significant differences between all of the groups (after Bonferroni correction) with Caucasians, having the highest levels of gene transcript concentration and African Americans having the lowest (Figure 3).

Due to the strongly significant difference in expression between the groups, an interaction analysis was pursued. Our analysis identified two loci with interaction terms which were significant (p<0.05) and an additional three loci with a trend towards significance (0.05< p >0.10). We found in our stratified analysis that overall 17/59 different loci were significantly associated with gene expression levels in at least one of the racial/ethnic groups race (10 for Caucasians, 4 for African Americans and 6 for Hispanics). In each individual racial ethnic group, different CpG loci were found to be significant. The Hispanic group shared one locus with the Caucasian group and two loci with the African American group, however, the Caucasian and African American groups shared no significant loci.
Figure 2: Gene map with significant CpG loci from backward selection modeling

Figure 3: Gene transcript expression of P2RY12 by race
DISCUSSION

We have shown an association between DNA methylation around our target gene (P2RY12) and expression of its RNA transcript. In addition, a significant degree of variation exists in our study population in terms of expression of P2RY12. This variability in expression can be at least partially explained by variations in the methylome of our population. Even a small degree of variation in RNA expression could result in a large variation of the concentration of receptor and an even larger difference in target cell inhibition due to the exponential increase at each step of the molecular cascade. This may help explain some of the variability in response to thienopyridine therapy in the general population. Variability in expression may also help to explain some adverse outcome related to side effects of these medications (specifically bleeding).

Though rates of thienopyridine failure have decreased over the course of past decade and novel agents have been approved by the FDA, there remains a significant rate of failure. In one recent trial, more than 2% of patients may still have failure as evidenced by myocardial infarction or stroke, even with 30 months of antiplatelet therapy.\textsuperscript{22} Prior investigation into major risk factors for thienopyridine resistance, including trials of genetic variation in the metabolism of thienopyridines, has explained little of the variability in response to thienopyridine therapy. Even use of novel agents less dependent on metabolism for their action has proven to have a significant rate of failure.\textsuperscript{5, 23-25}

There exists substantial data on genome wide methylation and global RNA transcript expression patterns.\textsuperscript{10} Some data from the PDAY cohort found association between global levels of DNA methylation and subclinical atherosclerosis in young individuals.\textsuperscript{26} However, data remain scarce regarding the prognostic value of epigenetic
modifications and cardiovascular disease. There have been few studies investigating single gene epigenetic modification or the relationship between cardiovascular disease and epigenetic modification (specifically DNA methylation). The limited evidence that does exist, however, supports an effect on human disease, especially in dilated cardiomyopathies and coronary artery disease progression.\textsuperscript{27, 28} These studies did not evaluate the effect of the modification on gene transcription, however. Most of the previous work looking at thienopyridine responsiveness has been focused on SNPs, particularly related to the ABCB1 and CYP2C19 genes with the most promising data coming from the latter. To date no data has been published evaluating the effect of epigenetics on this area of disease.

One recently published paper looked at two CpG loci in the promoter region of the P2RY12 gene and did not find a significant association with platelet reactivity after loading and initiation of maintenance dose of clopidogrel, except in certain sub-populations.\textsuperscript{29} Our data may help to support this finding. Of the two CpG loci we included in this study, one did not reach significance in its association with gene expression in either model, while the other reached significance only when included in the multi-locus modeling. Our findings suggest that, with relation to P2RY12 mRNA transcript expression, CpG loci located outside of the known promoter region of the gene have a higher association than those within the promoter region.

\textit{Racial/Ethnic Groups}

The differences in expression between the racial/ethnic groups are particularly interesting. African Americans have been known to have higher thienopyridine failure rate than their Caucasian cohorts. Absolute differences in failure rates have been
quantified as up to 2.5% higher in the African American population despite having equal or greater rates of adherence to medication therapy. Our data do suggest a potential racial/ethnic interaction with expression of P2RY12, with significance or a trend towards significance in 5 different cis-acting CpG loci. In addition to this, when adjusted gene expression levels were compared between the different racial/ethnic groups there was a significant difference between all three racial/ethnic groups (Figure 3). Interestingly, there also appeared to be different methylation loci which were associated with gene expression in the different racial/ethnic groups. No loci were shared between the African American group and the Caucasian group. Hispanics had significant loci what were shared with African Americans and with Caucasians, however, they also had loci which were specific to themselves as well (Table IV). It unclear what effect differences in gene expression have on the concentration of receptor on the cell surface or the efficacy of thienopyridines. Our understanding of the underlying mechanisms of these racial/ethnic differences remains incomplete and it is not clear if this could account for the previously observed differences in clinical outcomes.

LIMITATIONS

Several key limitations are noted in this study. First is our use of monocytes as a surrogate cell type for the more desired platelet (or their nucleated precursor – megakaryocytes). Epigenetic profiles tend to be relatively cell line specific, however, it has been hypothesized that cell lines closely related to platelets may be an adequate surrogate and be more easily studied than megakaryocytes. Second, our ability to draw conclusion about clinical utility of this data is limited by our endpoint. There are no data at this point to support a direct association between gene expression of P2RY12
and clinical endpoints of myocardial infarction, stent thrombosis or stroke. Future studies aimed specifically at platelets and/or prospectively evaluated clinical endpoints are needed to fully evaluate the utility of monocyte DNA methylation and RNA transcript expression in the prediction of thienopyridine resistance.

Another key limitation of this study is its cross sectional nature. We cannot draw conclusions regarding the effect DNA methylation may have on gene expression or vice versa. Further prospective data would be required to draw such conclusions.

**CONCLUSIONS**

Our data demonstrate that an association exists between the degree of methylation of CpG sites located within ±500 KBP of the P2RY12 gene and RNA transcript expression of that gene. Additionally there appears to be racial/ethnic differences in which CpG loci and to what degree they are associated with expression. Given the cross sectional nature of this data, conclusions cannot be drawn regarding causal relationship between DNA methylation and gene expression in this instance, however, these results warrant further prospective investigation.
**Supplemental Table I:** Analysis of baseline characteristics vs. P2RY12 gene transcript concentration.

<table>
<thead>
<tr>
<th></th>
<th>β (std. error)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.007 (0.002)</td>
<td>2 x10^{-4}***</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.11 (0.033)</td>
<td>0.001***</td>
</tr>
<tr>
<td>Race/Ethnic Group</td>
<td>--</td>
<td>2.4 x10^{-6}***</td>
</tr>
<tr>
<td>Site at Exam 5</td>
<td>--</td>
<td>0.001***</td>
</tr>
<tr>
<td>Enrichment scores*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B lymphocytes</td>
<td>-0.15 (0.222)</td>
<td>0.510</td>
</tr>
<tr>
<td>T lymphocytes</td>
<td>-0.39 (0.146)</td>
<td>0.007</td>
</tr>
<tr>
<td>NK cells</td>
<td>1.33 (0.229)</td>
<td>6.3 x10^{-9}***</td>
</tr>
<tr>
<td>Neutrophils</td>
<td>1.39 (0.372)</td>
<td>2 x10^{-4}***</td>
</tr>
<tr>
<td>Methylation Chip*</td>
<td>--</td>
<td>0.0013 ***</td>
</tr>
<tr>
<td>Chip Well*</td>
<td>--</td>
<td>0.613</td>
</tr>
<tr>
<td>Tobacco use**</td>
<td>-0.139 (0.065)</td>
<td>0.032</td>
</tr>
<tr>
<td>Aspirin use**</td>
<td>-0.025 (0.033)</td>
<td>0.42</td>
</tr>
<tr>
<td>Diabetes**</td>
<td>0.072 (0.040)</td>
<td>0.07</td>
</tr>
<tr>
<td>Alcohol use**</td>
<td>0.032 (0.034)</td>
<td>0.35</td>
</tr>
<tr>
<td>Hypertension**</td>
<td>0.096 (0.034)</td>
<td>0.005</td>
</tr>
</tbody>
</table>

--β values for categorical values not reported (chi-square test performed)

*The covariates age, gender, race/ethnic group, site, enrichment scores, chip, and chip well were univariate analyses

**Adjusted based on age, gender, race/ethnic group, site, enrichment scores, chip, and well.

***Remain significant after Bonferroni correction

**Supplemental Table II:** Relationship between methylation of potential cis-acting CpG loci with P2RY12 expression, loci from individual locus modeling.

<table>
<thead>
<tr>
<th>Methylation Site</th>
<th>β (SE)*</th>
<th>Semi-partial $r^2$ (95% CI)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg05094216</td>
<td>0.189 (0.04)</td>
<td>0.019 (0.007, 0.036)</td>
<td>4.5 x10^{-7}</td>
</tr>
<tr>
<td>cg24630764</td>
<td>-0.271 (0.06)</td>
<td>0.017 (0.006, 0.034)</td>
<td>4.2 x10^{-6}</td>
</tr>
<tr>
<td>cg07736012</td>
<td>0.234 (0.06)</td>
<td>0.010 (0.002, 0.024)</td>
<td>0.0003</td>
</tr>
<tr>
<td>cg00472710</td>
<td>0.098 (0.03)</td>
<td>0.009 (0.001, 0.022)</td>
<td>0.0006</td>
</tr>
<tr>
<td>Cg00472710</td>
<td>0.099 (0.03)</td>
<td>0.0093 (0.002, 0.023)</td>
<td>0.0006</td>
</tr>
<tr>
<td>cg16780454</td>
<td>0.085 (0.03)</td>
<td>0.008 (0.001, 0.020)</td>
<td>0.0013</td>
</tr>
<tr>
<td>cg21766308</td>
<td>0.085 (0.03)</td>
<td>0.007 (0.001, 0.019)</td>
<td>0.0032</td>
</tr>
<tr>
<td>cg01754219</td>
<td>0.180 (0.07)</td>
<td>0.005 (0.000, 0.016)</td>
<td>0.0113</td>
</tr>
<tr>
<td>Gene</td>
<td>Value</td>
<td>Standard Error</td>
<td>P Value</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>cg09432154</td>
<td>-0.152 (0.06)</td>
<td>0.004 (0.015)</td>
<td>0.0185</td>
</tr>
<tr>
<td>cg09410045</td>
<td>0.132 (0.06)</td>
<td>0.004 (0.014)</td>
<td>0.0222</td>
</tr>
<tr>
<td>cg12775479</td>
<td>0.160 (0.07)</td>
<td>0.004 (0.013)</td>
<td>0.0295</td>
</tr>
<tr>
<td>cg15001056</td>
<td>-0.125 (0.06)</td>
<td>0.003 (0.012)</td>
<td>0.0425</td>
</tr>
<tr>
<td>Cg15814508</td>
<td>0.115 (0.07)</td>
<td>0.0023 (0.011)</td>
<td>0.0901</td>
</tr>
<tr>
<td>Cg17582100</td>
<td>-0.100 (0.06)</td>
<td>0.0021 (0.1)</td>
<td>0.1067</td>
</tr>
<tr>
<td>Cg1919271</td>
<td>-0.069 (0.04)</td>
<td>0.0020 (0.010)</td>
<td>0.1112</td>
</tr>
<tr>
<td>Cg18509239</td>
<td>0.129 (0.08)</td>
<td>0.0019 (0.010)</td>
<td>0.1237</td>
</tr>
<tr>
<td>Cg2179358</td>
<td>0.098 (0.07)</td>
<td>0.002 (0.009)</td>
<td>0.1633</td>
</tr>
<tr>
<td>Cg00219763</td>
<td>0.136 (0.10)</td>
<td>0.0015 (0.009)</td>
<td>0.1715</td>
</tr>
<tr>
<td>Cg00091349</td>
<td>0.059 (0.04)</td>
<td>0.0014 (0.009)</td>
<td>0.1810</td>
</tr>
<tr>
<td>Cg18143869</td>
<td>-0.073 (0.05)</td>
<td>0.0014 (0.009)</td>
<td>0.1839</td>
</tr>
<tr>
<td>Cg07548668</td>
<td>0.100 (0.08)</td>
<td>0.0013 (0.008)</td>
<td>0.2007</td>
</tr>
<tr>
<td>Cg01876514</td>
<td>0.080 (0.06)</td>
<td>0.0013 (0.008)</td>
<td>0.2054</td>
</tr>
<tr>
<td>Cg03379681</td>
<td>0.081 (0.07)</td>
<td>0.0012 (0.008)</td>
<td>0.2257</td>
</tr>
<tr>
<td>Cg12956884</td>
<td>0.094 (0.08)</td>
<td>0.0003 (0.008)</td>
<td>0.2304</td>
</tr>
<tr>
<td>Cg06512350</td>
<td>-0.089 (0.08)</td>
<td>0.0011 (0.008)</td>
<td>0.2343</td>
</tr>
<tr>
<td>Cg03604388</td>
<td>0.066 (0.06)</td>
<td>0.0011 (0.008)</td>
<td>0.2426</td>
</tr>
<tr>
<td>Cg11046673</td>
<td>-0.040 (0.03)</td>
<td>0.0010 (0.008)</td>
<td>0.2521</td>
</tr>
<tr>
<td>Cg06838302</td>
<td>-0.075 (0.07)</td>
<td>0.0001 (0.007)</td>
<td>0.2944</td>
</tr>
<tr>
<td>Cg25499537</td>
<td>0.048 (0.05)</td>
<td>0.0008 (0.007)</td>
<td>0.3074</td>
</tr>
<tr>
<td>Cg12754495</td>
<td>0.044 (0.05)</td>
<td>0.0008 (0.007)</td>
<td>0.3258</td>
</tr>
<tr>
<td>Cg11485946</td>
<td>0.079 (0.08)</td>
<td>0.0007 (0.007)</td>
<td>0.3442</td>
</tr>
<tr>
<td>Cg12842552</td>
<td>0.079 (0.09)</td>
<td>0.0006 (0.006)</td>
<td>0.3867</td>
</tr>
<tr>
<td>Cg07375150</td>
<td>0.048 (0.06)</td>
<td>0.0006 (0.006)</td>
<td>0.3885</td>
</tr>
<tr>
<td>Cg23351724</td>
<td>-0.071 (0.08)</td>
<td>0.0006 (0.006)</td>
<td>0.3889</td>
</tr>
<tr>
<td>Cg10552106</td>
<td>0.075 (0.09)</td>
<td>0.0006 (0.006)</td>
<td>0.3961</td>
</tr>
<tr>
<td>Cg19926250</td>
<td>0.069 (0.08)</td>
<td>0.0006 (0.006)</td>
<td>0.4020</td>
</tr>
<tr>
<td>Cg17795586</td>
<td>-0.077 (0.09)</td>
<td>0.0005 (0.006)</td>
<td>0.4094</td>
</tr>
<tr>
<td>Cg06855506</td>
<td>-0.048 (0.06)</td>
<td>0.0005 (0.006)</td>
<td>0.4113</td>
</tr>
<tr>
<td>Cg00290355</td>
<td>0.049 (0.06)</td>
<td>&lt;0.0001 (0.006)</td>
<td>0.4203</td>
</tr>
<tr>
<td>Cg26182497</td>
<td>0.047 (0.06)</td>
<td>0.0004 (0.006)</td>
<td>0.4540</td>
</tr>
<tr>
<td>Cg16034546</td>
<td>0.066 (0.09)</td>
<td>0.0004 (0.006)</td>
<td>0.4595</td>
</tr>
<tr>
<td>Cg00137750</td>
<td>-0.049 (0.08)</td>
<td>0.0003 (0.005)</td>
<td>0.5171</td>
</tr>
<tr>
<td>Cg12207930</td>
<td>-0.029 (0.05)</td>
<td>0.0003 (0.005)</td>
<td>0.5348</td>
</tr>
<tr>
<td>Cg14191476</td>
<td>0.023 (0.04)</td>
<td>0.0002 (0.005)</td>
<td>0.5973</td>
</tr>
<tr>
<td>Cg03906843</td>
<td>0.028 (0.05)</td>
<td>0.0002 (0.005)</td>
<td>0.6031</td>
</tr>
<tr>
<td>Cg24474182</td>
<td>0.022 (0.05)</td>
<td>0.0002 (0.005)</td>
<td>0.6332</td>
</tr>
<tr>
<td>Cg05684005</td>
<td>-0.026 (0.06)</td>
<td>0.0002 (0.005)</td>
<td>0.6401</td>
</tr>
<tr>
<td>Cg05657760</td>
<td>-0.025 (0.05)</td>
<td>0.0002 (0.005)</td>
<td>0.6460</td>
</tr>
<tr>
<td>Cg15295441</td>
<td>-0.015 (0.04)</td>
<td>0.0001 (0.004)</td>
<td>0.6824</td>
</tr>
<tr>
<td>Cg22843488</td>
<td>0.009 (0.02)</td>
<td>&lt;0.0001 (0.004)</td>
<td>0.6860</td>
</tr>
<tr>
<td>Sample ID</td>
<td>β Value (SE)</td>
<td>p Value (CI)</td>
<td>Expression (P)</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------</td>
<td>--------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Cg09571369</td>
<td>0.024 (0.70)</td>
<td>&lt;0.0001 (0, 0.004)</td>
<td>0.6995</td>
</tr>
<tr>
<td>Cg01114224</td>
<td>0.023 (0.06)</td>
<td>&lt;0.0001 (0, 0.004)</td>
<td>0.7105</td>
</tr>
<tr>
<td>Cg04422741</td>
<td>0.018 (0.05)</td>
<td>0.0001 (0, 0.004)</td>
<td>0.7130</td>
</tr>
<tr>
<td>Cg19227710</td>
<td>0.019 (0.06)</td>
<td>0.0001 (0, 0.004)</td>
<td>0.7395</td>
</tr>
<tr>
<td>Cg15625324</td>
<td>-0.015 (0.06)</td>
<td>&lt;0.0001 (0, 0.003)</td>
<td>0.8085</td>
</tr>
<tr>
<td>Cg13229867</td>
<td>0.005 (0.04)</td>
<td>&lt;0.0001 (0, 0.002)</td>
<td>0.8867</td>
</tr>
<tr>
<td>Cg07269138</td>
<td>0.008 (0.08)</td>
<td>&lt;0.0001 (0, 0.002)</td>
<td>0.9201</td>
</tr>
<tr>
<td>Cg16899444</td>
<td>-0.006 (0.10)</td>
<td>&lt;0.0001 (0, 0.001)</td>
<td>0.9473</td>
</tr>
<tr>
<td>Cg01330290</td>
<td>-0.009 (0.06)</td>
<td>&lt;0.0001 (0,0)</td>
<td>0.9763</td>
</tr>
</tbody>
</table>

*β value represents the predicted change in expression for each unit increase of methylation at that site.
REFERENCES


MULTIPLE LINEAR REGRESSION MODELING

Multiple Linear Regression (MLR) analysis is a popular statistical procedure for investigating relationships in the ecological sciences. A response (or outcome) variable and one or more predictor variables are selected and compared in order to describe the relationship between variables. Regression coefficients, standard errors, and statistical significance can be obtained for each predictor variable in the MLR analysis. Non-significant variables can then be removed or retained in the model based on their significance and known historical contribution to the outcome variable. This method of statistical analysis can be helpful in several situations, specifically when confounding variables need to be controlled for in an analysis, when exploring possible prognostic variables with little or no prior information of which variables are important, or when developing an prognostic model for prediction of the outcome variable. Non-significant factors can be removed from the model and the model performed again to obtain updated regression coefficients.¹

One problem with the approach of Multiple Linear Regression modeling comes when there are a large number of potential explanatory variables being analyzed in the model. At larger numbers, it is expected that one or more of the variables would be significant (depending on the level of significance selected). In these situations, there is no completely satisfactory way of searching for the best prediction model without incurring an inflated type I error rate. Unfortunately, many of the corrective techniques used to limit the type I error rate increase the type II error rate to an unacceptable level.
as well. Exploratory analysis can performed initially between each predictor variable and the outcome variable in a univariate fashion to identify potential variables to include in the MLR. When this is performed, however, a level of significance of p<0.2 is recommended due to unpredictable interactions between variables. This method can increase the type II error rate by not including variables that may be significantly associated with the outcome once they are adjusted for other variables in the model.

This is the basis for selection procedures. These selection procedures have become widely used in the identification of a subset of potential predictors for a biological trait. Whenever able, an “all possible regressions procedure” is a preferred method of determining which predictor variables fit the data the best. In this procedure, all of the possible models that can be formed are modeled. This precludes its use in modeling with large numbers of predictor variables. \(^{1}\) For k variables, a total of \((2^k-1)\) models will be run. As the number of predictor variables increases, the computation time and power required quickly become too large for practical application of this selection model. In the P2RY12 dataset, with 58 predictor variables, this would equate to over 288 quadrillion different models being analyzed. Obviously, this would take an unattainable amount of time and computer resources to obtain. In fact, SAS will not support running this type of model with more than 10 variables.

Other selection procedures have been developed as means of obtaining a “best model,” in situations where the number of predictor variables preclude the preferred “all possible” method of model selection. First, we will discuss one of the simplest methods of model selection – the maximum model method. In this method, the maximum model is entered and the p-value cutoff is used to eliminate variables which do not meet
significance in the model. All of the variables which meet the minimum p value are then entered into a model of their own and re-analyzed. In theory, this allows all of the predictor variables to be analyzed during the analysis, thereby minimizing the possibility of type II error, as can become problematic with some of the other selection methods.

A similar selection method to the maximum model method is the backward selection technique. In the backward selection method, all variables are entered into an initial regression model. In a stepwise fashion, the variable furthest from meeting the minimum selection criteria (highest p value, lowest $R^2$ value, lowest F value, etc.) in its association with the outcome variable is removed from the analysis and all of the remaining variables are again entered into the model. The user must pre-specify the selection criteria for leaving the multiple linear regression model.

A forward selection procedure starts by regressing all of the predictor variables in a univariate fashion against the outcome variable. Then, in a stepwise fashion, the most highly associated variable is entered into the model. At each step there is regression of each individual predictor variable to again select the variable with the strongest association, which will then be added into the model at that step. When there are no further variables significantly associated with the outcome variable, the procedure stops. The user must pre-specify selection criteria to enter the regression model. The stepwise selection procedure is a variant of the forward selection method, however, at the end of each addition the model is re-analyzed and any variable which is no longer meets the criteria for significance is removed from the model. This method is intended to reduce the number of non-significant variables which may be left in the final model. The user will have to pre-specify both selection criteria to enter and be removed from the model.
A forward selection procedure has a modest effect on lowering of the type I error rate and a modest effect on increasing the type II error rate.

Due to the aforementioned reasons, some have advocated using both forward selection and backward elimination procedures, with or without a true stepwise selection procedure, to find at least some degree of agreement between the models to verify validity. In 2009, it was suggested that while this method of correction may be useful, it is rare that the forward and backward selection models differ by a significant degree. As the number of predictor variables increased, the rate of having an incorrect model increases. It was also suggested that as the same size increases (ie sample sizes of 1000 and 5000), while the models contained some “noise” variables, they were at least inclusive of all of the truly significant predictor variables.

ANALYZING P2RY12 DATA

The P2RY12 data in our primary analysis was presented as a backward selection model. We initially added all 58 of our predictor variables into the model and used a p-value of 0.05 to remain in the model. Due to the debate on which selection procedure may be most accurate and whether the models would agree, the primary analysis was repeated in four separate models using the different selection methods. An inclusion criterion of 0.05 was used for both the forward and backward selection models and the stepwise selection model used an inclusion criterion of 0.1 and exclusion criteria of 0.05. Additionally, a full model was performed using all 58 variables and identifying all predictor variables significant at the 0.05 level (Table I). The forward and stepwise selection procedures had the same results and thus these were combined into forward selection. There was a reasonable level of agreement between the selection models
and the full model. The full model contained eight total significant variables, while the forward selection contained all of these plus one additional “noise” variable. The backward selection contained all except for one of the predictor variables, but contained an additional five “noise variables.” The analyses all had the three most significant (and most strongly correlated) variables in common. These results agree with the prior published data suggesting reasonable agreement between models, however, there was additional agreement and inclusion of the significant predictor variables from a full model.

**Table I:** Significant variables in model after different selection techniques

<table>
<thead>
<tr>
<th>CpG Locus</th>
<th>p-value</th>
<th>CpG Locus</th>
<th>p-value</th>
<th>CpG Locus</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg05094216</td>
<td>8.9E-07</td>
<td>cg05094216</td>
<td>8.07E-09</td>
<td>cg05094216</td>
<td>1.3E-07</td>
</tr>
<tr>
<td>cg24630764</td>
<td>5.79E-08</td>
<td>cg24630764</td>
<td>1.2E-06</td>
<td>cg24630764</td>
<td>5.2E-06</td>
</tr>
<tr>
<td>cg07736012</td>
<td>0.00057</td>
<td>cg07736012</td>
<td>9.4E-05</td>
<td>cg07736012</td>
<td>0.00011</td>
</tr>
<tr>
<td>cg11919271</td>
<td>0.04523</td>
<td>cg11919271</td>
<td>0.00355</td>
<td>cg11919271</td>
<td>0.00589</td>
</tr>
<tr>
<td>cg09410045</td>
<td>0.01355</td>
<td>cg09410045</td>
<td>0.00573</td>
<td>cg09410045</td>
<td>0.01239</td>
</tr>
<tr>
<td>cg01754219</td>
<td>0.01052</td>
<td>cg01754219</td>
<td></td>
<td></td>
<td>0.01304</td>
</tr>
<tr>
<td>cg00091349</td>
<td>0.00421</td>
<td>cg00091349</td>
<td>0.00992</td>
<td>cg00091349</td>
<td>0.02565</td>
</tr>
<tr>
<td>cg11046673</td>
<td>0.01555</td>
<td>cg11046673</td>
<td>0.00712</td>
<td>cg11046673</td>
<td>0.03101</td>
</tr>
<tr>
<td>cg00472710</td>
<td>0.00405</td>
<td>cg00472710</td>
<td>0.00924</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>cg14191476</td>
<td>0.04277</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>cg17582100</td>
<td>0.04383</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>cg09432154</td>
<td>0.04534</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>cg01754219</td>
<td>0.00755</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CORRECTION FOR MULTIPLE COMPARISONS

Unfortunately, any selection method which is decided upon will ultimately have an overly optimistic model given inflated type I error rate associated with multiple analyses. This is a problem with any multiple regression model even without a selection procedure, as there are multiple comparisons involved in any multiple linear regression model. Because of this fact there has been much debate in the literature regarding the validity of results from a multiple linear regression model, specifically in models using selection procedures, as there is rarely a correction applied to the selection p-value. As the number of predictor variables increases, the type I error level increase based on the formula: $\alpha_{cor} = 1 - (1 - \alpha)^k$ where $\alpha$ is the selected level of significance, $k$ represents the number of predictor variables and $\alpha_{cor}$ is the corrected type I error level. In the case of 58 predictor variables, there would be an increased type I error rate of 0.94 (giving a 94% chance of finding at least one significant variable by chance alone). To adjust for this, multiple methods have been proposed.

First, the simplest and most conservative method for correcting an inclusion p-value is to perform a Bonferroni correction based on the total number of predictor variables in a maximum model ($\alpha_{cor} = \alpha / k$). This method is straightforward to apply to a simple MLR model, however, it does not come without its own problems. This method may be overly conservative and does not take into account the multiple steps used in a selection procedure which exponentially increases the number of total analyses performed during the analysis. Mundfrom et. al. showed that when using this method of correcting for type I error inflation that there is significant overcorrection, especially in cases where there are substantial number of non-significant predictor variables. While
the type I error does increase in a predictable fashion with the increased number of total variables used, the total type I error increases less rapidly as the proportion of non-significant variables increases. If all variables were deemed significant, then the Bonferroni corrected values are very close to the nominal p-value selected.\textsuperscript{4} Using this simulated data and extrapolating out to a larger number of variables with a significant portion of non-associated variables, the Bonferroni corrected alpha level would quickly become exceedingly low and would drastically increase the type II error rate. This is likely the less acceptable of the two types of errors, especially when attempting to create a prediction model. When applied to a stepwise model, unfortunately, there happen to be many more total analyses. At each step of a forward selection model, each predictor variable not currently in the model is tested. In addition, it is very difficult to ascertain how many steps total it will take to complete the model in advance.

Correcting for each analysis at each step of a selection procedure is likely to be too overly conservative. One property of conditional probability can be very helpful concerning selection procedures. Assuming that A is an event of very high probability, meaning P(A) is close to one (ie. a small p value or a high F statistic), and an event B has a probability which is at least of modest effect, P(B|A) is approximately the same as P(B) if not slightly smaller. Leading to the relationship:

\[
P(B|A) = \frac{P(AB)}{P(A)} \approx P(B)
\]

Therefore, the probability of including variable B in a model at each step of a selection procedure is approximately equal to the probability of including B in a model without having A in the model already. This would not hold true for variables with smaller
regression coefficients or larger P values. Therefore, we can assume that for the
variables with the strongest associations added in the initial steps of the procedure, that
their p-values would not need to be adjusted for the number of steps taken (but should
still be corrected for total number of variables in the model) and that the variables added
in the later steps of the procedure would be of little consequence to the overall
procedure given small effect sizes to being with.\(^5\)

**SELECTION OF AN APPROPRIATE ALPHA LEVEL**

To this point, the approaches used to make a prediction model and the problems
associated with correction in each of the various selection types has been discussed.
However, determining a sufficient level of correction to use on MLR models to limit both
type I and type II error to a reasonable level may be a much more important discussion.
As was discussed in the prior section, stepwise selection procedures may increase the
type I error level slightly more than using a maximum model due to the increased
number of analyses. However, since these analyses are related and the overall effect of
including a non-significant value (type I error) is minimal, it can be reasonably assumed
that the same alpha level of significance could be used in any of the stepwise selection
procedures as we would use in a full MLR model.

It is important to bear in mind that there are two different types of error which
need to be minimized. It is very important for prediction models to have all truly
associate variables included in the model (avoiding type II error and increasing the
power of the analysis), while not including variables with true H\(_o\) (avoiding type I error) is
important to a slightly lesser degree. Since the two types of error are inversely related
(ie. as \(\alpha\) increased, \(\beta\) decreases) there is generally a tradeoff between improving
sensitivity for a predictor variable and the specificity of the variables included in a model. Several adjustment methods for individual variable p-values have been suggested.

**SINGLE STEP METHODS OF CORRECTION**

Single step methods of correction imply that at one step the p-value for inclusion in a model is corrected based on the number of predictor variables entered into model. Several different methods for calculating the correction have been proposed, however, for ease of explanation, the Bonferroni method (which, as previously stated, is the simplest formula and most conservative of these approaches) will be discussed. Using single step methods for correcting the desired p-value for multiple corrections was the topic of Mundfrom et al in 2006. They showed that there is significant over-correction by performing a straightforward Bonferroni correction to MLR modeling. The observed alpha levels of their analysis did show that the Bonferroni method likely overcorrected for type I error rates when there were multiple significant variables included in the models which were performed. However, they did not comment on the combined type II error rate. While this was true, they were able to hold the rate of type I error down to a minimum by applying this correction method.

In order to adequately perform and adjust a stepwise selected procedure, one would need the ability to select one p-value prior to running the model. Unfortunately, Mundfrom et al. only extrapolated data for up to eight predictor variables. This, unfortunately, would not be adequate for determining appropriate alpha levels for models containing more variables than that. In theory, there would be a drastic reduction in type I error and increase in type II error with increasing variables based on their data.
FALSE DISCOVERY RATE

The False discovery rate (FDR) was first proposed in 1995 as an alternative method to the traditional single step adjustment methods (Bonferroni, Tukey, etc.) This correction has been applied, especially in genetic and epigenetic testing, when very large numbers of individual predictor variables (sometimes hundreds of thousands) are tested in succession to identify a number of possible predictor variables. As would be expected in large databases (such as those used in genomic research), the use of a Bonferroni type of correction would lead to incredibly small p-value to declare significance. For instance, if 500,000 single variables (loci of either genetic or epigenetic origin) were tested for significance, they would have to reach a p-value of $10^{-7}$ to reach significance at the 0.05 level. This greatly increases the type II error rate, and as most of these studies are trying to identify new (perhaps borderline in significance) variables, the type II error is the more concerning of the two types of error. The FDR gives a less conservative level of significance to the analysis in order reduce the type II error. They also discussed the family wide error rate, which is the probability of making at least one false discovery (type I error) in all of the analyses. Traditional correction methods for multiple comparisons aim to keep the FWER less than a pre-designated level.

Methodology behind the false discovery rate is more complicated than the more straight forward single step type correction. The procedure was first described in 1995 by Benjamini and Hochberg. They proposed that while it is clear that correction for multiple comparisons are useful and needed, they are often not applicable to certain types of medical and basic science research, substantially reduce the power of studies, and overcorrect family wide error rate in tests which may be related (i.e., those of
different effects of the same treatment or data which may have some mild correlation between predictor variables). While the Bonferroni correction focuses mainly on reducing type I error rate while ignoring the type II error rate, they proposed a correction procedure to help minimize both type I and type II error simultaneously. The FDR formula is based on testing of \( m \) null hypotheses. Of these \( m \) hypotheses, \( m_0 \) are truly null and \( R \) is the number of hypotheses rejected (deemed to be significant). Multiple further variables were proposed based on this basic framework. The variable \( m \) is the only known variable in advance of testing.

<table>
<thead>
<tr>
<th></th>
<th>Declared non-significant</th>
<th>Declared significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True null hypotheses</td>
<td>U</td>
<td>V</td>
<td>( m_0 )</td>
</tr>
<tr>
<td>Non-true null hypotheses</td>
<td>T</td>
<td>S</td>
<td>( m - m_0 )</td>
</tr>
<tr>
<td></td>
<td>m - R</td>
<td>R</td>
<td>( m )</td>
</tr>
</tbody>
</table>

The proportion of false discoveries (type I error), is defined by the proportion of rejected null hypothesis which were erroneously rejected:

\[
Q = \frac{V}{V + S} = \frac{V}{R}
\]

And the FDR is defined as

\[
FDR = Q_e = E[Q] = E\left[\frac{V}{V + S}\right] = E\left[\frac{V}{R}\right] \text{ where } \left[\frac{V}{R}\right] = 0 \text{ when } R = 0
\]

60
This essentially states that for x number of expected discoveries, we are willing to tolerate (FDR) proportion of those being false. If all null hypotheses are true, the FDR is equivalent to the FWER. If the \( m_0 < m \), then the FDR will be smaller than or equal to the FWER. In order to apply an FDR correction to a problem one must first have to consider testing \( H_1, H_2, \ldots, H_m \). An assumption is made that the corresponding \( P_1 \leq P_2 \leq \ldots \leq P_m \). In this case, each subsequent analysis would be tested at a p value corresponding to:

\[
P_i \leq \frac{i}{m} q \quad \text{where } q \text{ is the overall FDR level which is pre-specified.}
\]

This would be a very difficult measure to use in a stepwise procedure as the inclusion vs exclusion p-value would have to be adjusted at each step in the analysis to correct for the FDR, though this could potentially be very helpful in a full model MLR approach.

**TYPE I AND TYPE II ERROR RATES WITH CORRECTION METHODS**

In this analysis, the data from Mundorf et al.\(^4\) were expanded upon. Simulated data were created and type I and type II error rates were tested with three different methods for controlling the error. The simulated data were created using JMP version 11.2. First 30,000 observations were simulated for one outcome variable and 105 different predictor variables (45 of which were significantly associated at the 0.1 level with the outcome variable). These were then grouped into populations of 300. A MLR model was performed and 45 of 105 variables were found to be significant while all others failed to reach a nominal p-value. MLR modeling was performed and actual type I and type II error rates were calculated over a total of 100 iterations. Type I error was defined as any iteration which had a non-significant variable being associated at the
alpha level chosen. Type II error was defined as any variable which was significant in the overall model being left out of iteration. MLR was performed with models containing a total of 2, 4, 8, 16, 32 and 60 predictor variables, and for models with 0, ¼ and ½ of the variables being significantly associated in each category (except for models with 2 predictor variables which only had 0 and ½ groups). Each of the MLR were tested at an unadjusted $\alpha$-level of 0.05, a Bonferroni adjusted $\alpha$-level of 0.05/k, where k is the number of predictors in the model, a modified version of the Bonferroni correction using 0.05/n, where n is the number of known significant variables in the model (and when n=0 true bonferroni correction was used), as well as using the FDR(Figure I & II). Results of models with ¼ of the variables being significant are displayed in Figure I and II, with full data displayed in Table II. As the number of predictor variables increases, so do both the type I and type II error rates. When Bonferroni method is used to correct p-values in the model, the type II error rate becomes extraordinarily high even when only eight predictor variables are entered into the model. The opposite can be said for using a nominal p-value of 0.05. Using a nominal p-value of 0.05, the type II error rate increases at a more modest rate, however, the type I error rate very rapidly becomes highly elevated and reaches a level of almost one by the time 60 variables are tested. The modified version of the Bonferroni correction and the FDR seem to take a middle ground and does not retain low type I or type II error levels well. In the prior paper, with larger numbers the alpha level for the Bonferroni correction was trending downward as the predictor variables increased, however, in our model they increased slightly, but remained fairly low overall. None of these correction techniques is particularly adequate given either highly inflated type I or type II error rates. Furthermore, in the case of
Figure I: (A) Type I error rate plotted against number of predictor values across different methods for correcting multiple comparisons. (B) Type II error rate plotted against number of predictor values across different methods for correcting multiple comparisons.

Figure II: alpha (type I) and beta (type II) error rates compared against number of predictor variables while using (A) false discovery rate, (B) Bonferroni correction, (C) modified Bonferroni correction, and (D) nominal p-value of 0.05.
Table II. Alpha and beta values by number of total and significant predictors and method of correction for multiple comparisons.

<table>
<thead>
<tr>
<th></th>
<th>FDR</th>
<th>Bonferroni</th>
<th>Mod Bonferroni</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>type I</td>
<td>type II</td>
<td>type I</td>
<td>type II</td>
</tr>
<tr>
<td>2</td>
<td>zero</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>half</td>
<td>0.04</td>
<td>0.2</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>zero</td>
<td>0.04</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>quarter</td>
<td>0.03</td>
<td>0.2</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>half</td>
<td>0.05</td>
<td>0.34</td>
<td>0.04</td>
</tr>
<tr>
<td>8</td>
<td>zero</td>
<td>0.06</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>quarter</td>
<td>0.1</td>
<td>0.43</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>half</td>
<td>0.09</td>
<td>0.57</td>
<td>0.03</td>
</tr>
<tr>
<td>16</td>
<td>zero</td>
<td>0.03</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>quarter</td>
<td>0.14</td>
<td>0.55</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>half</td>
<td>0.24</td>
<td>0.43</td>
<td>0.02</td>
</tr>
<tr>
<td>32</td>
<td>zero</td>
<td>0.07</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>quarter</td>
<td>0.31</td>
<td>0.69</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>half</td>
<td>0.4</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>60</td>
<td>zero</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>quarter</td>
<td>0.56</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>half</td>
<td>0.74</td>
<td>0.56</td>
<td>0.08</td>
</tr>
</tbody>
</table>
creating a prediction model the type II error rate tends to be the one that should be minimized.

**MULTIPLE COMPARISONS AND P2RY12 DATA**

In order to show differences between models when using alternative methods of correcting for multiple comparisons, the P2RY12 data were used in a further analysis. Using the backward selection p-values from the prior analysis, four separate methods of correction were applied. Results are displayed in Table III. Interestingly, FDR and Bonferroni correction for multiple comparisons came up with the same models with three significant variables. These were the same three predictor variables which were most strongly associated with P2RY12 expression levels in the previous analysis. This was interesting given the vastly different observed alpha levels seen when comparing the different correction methods in the prior analysis. The modified Bonferroni correction method revealed a five variable model and the nominal p-value was obviously the full presented model from the previous chapter.

**Table III.** Primary analysis repeated using four different adjustments for multiple comparisons

<table>
<thead>
<tr>
<th>CpG Locus</th>
<th>p-value</th>
<th>FDR</th>
<th>B</th>
<th>MB</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg05094216</td>
<td>8.07E-09</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg24630764</td>
<td>1.2E-06</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg07736012</td>
<td>9.4E-05</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg11919271</td>
<td>0.00355</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg09410045</td>
<td>0.00573</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg11046673</td>
<td>0.00712</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg01754219</td>
<td>0.00755</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg00472710</td>
<td>0.00924</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg00091349</td>
<td>0.00992</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg14191476</td>
<td>0.04277</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg17582100</td>
<td>0.04383</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg09432154</td>
<td>0.04534</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*FDR – False discovery rate, B – Bonferroni, MB – Modified Bonferroni, N – nominal p-value 0.05
CONCLUSION

When it comes to evaluation of MLR models, there has been no consensus statement on the reporting of results. As has been shown, there are multiple different methods for correcting for the problem of multiple comparisons in the models. Depending on the perspective of the reader, different approaches could be taken and be equally appropriate. If creation of a prediction model is desired, it may be warranted to use a nominal p-value or FDR method of correction for adjusting results in order to retain as many truly significant variables in the model. On the other hand, if the reader is more interested in minimizing “noise” variables, a more stringent Bonferroni correction would be more appropriate. The recommendation of this author would be for reporting of raw p-values with the data and allowing the reader the ability to adjust based on their particular point of view. Alternatively, reporting models with the most conservative and most liberal selection criteria could also be an appropriate method of reporting though this may limit the ability of the reader to make their own conclusions. In addition, when performing MLR, the method of selection should again be based on intent of model. In the creation of a prediction model, inclusion of the most predictor variables may improve prediction, however, the models obtained from the forward and backward selection were very similar to the full model method of selection. Whenever feasible, an all possible regressions model should be employed. However, in models where large numbers of predictor variables make this impossible, all of the selection procedures produced similar results.
REFERENCES


2. Pohlmann JT. Controlling the Type I Error Rate in Stepwise Regression Analysis. 1979.


CURRICULUM VITAE

EDUCATION

- **Master of Science, Clinical and Population Translational Science** - in progress  
  Wake Forest University, Department of Public Health  
  Winston-Salem, NC  
  2012-2015

- **Doctor of Medicine**  
  University of Florida, College of Medicine  
  Gainesville, FL  
  2004-2008

- **Bachelor of Science in Biology, Cum Laude**  
  University of Miami, College of Arts and Sciences  
  Coral Gables, FL  
  2000-2004

TRAINING

- **Interventional Cardiology Fellowship - Future**  
  Section on Cardiovascular Medicine  
  Wake Forest University – Baptist Hospital  
  2016-2017

- **Cardiovascular Diseases Fellowship – Research Track**  
  Section on Cardiovascular Medicine  
  Wake Forest University – Baptist Hospital  
  2012-2016

- **Internal Medicine Internship/Residency**  
  Department of Internal Medicine  
  University of North Carolina Hospitals  
  2008-2011

WORK EXPERIENCE

- **Hospitalist**  
  Rex Hospital  
  Raleigh, NC  
  2011-2012
- **Teaching Assistant**  
  University of Florida, College of Medicine  
  Gainesville, FL  
  2008

- **Anatomy Assistant**  
  University of Florida, College of Medicine  
  Gainesville, FL  
  2005

### Professional Activities

- **Clinical and Population Translational Science Program Representative**  
  Graduate Student Association, Wake Forest University, Winston-Salem, NC.  
  2012-2013

- **Epidemiology & Prevention of Cardiovascular Disease Seminar Attendee**  
  American Heart Association, Tahoe, CA  
  July 2012

### Honors/Awards

- **Internal Medicine Resident Research Award**, UNC Hospitals, Chapel Hill, NC  
  2010

- **Chapman Scholar**, University of Florida, College of Medicine, Gainesville, FL  
  2005-2008

### Professional Organizations/Certifications

- American College of Cardiology – Member in training  
  2012-2015

- American Board of Internal Medicine – Board Certification  
  2011-2015

- North Carolina Medical Board Licensure  
  2011-2015

- American College of Physicians – Member  
  2010-2012

- American Medical Association – Member  
  2010-2012
PUBLICATI0NS/PRESENTATIONS


- Soliman EZ, Cammarata M, Li Y. Explaining the inconsistent associations of PR interval with mortality: The role of P-duration contribution to the length of PR interval. Heart Rhythm. 2014 Jan;11(1):93-8


- Cammarata MW, Rossi JS, Dharmavaram J, Walko CM, Muldrew KL, Week KE, Gabriel DA, Stouffer GA. Heterozygosity for *2 but not *17 CYP2C19 Allele Effects Platelet Reactivity Measured by VerifyNow. ACC Scientific Sessions April 2011 (abstract)