STRUCTURAL GENETIC VARIATION AND DYSLIPIDEMIA AMONG MEN IN THE MULTICENTER AIDS COHORT STUDY

by

Rebecca Bosko Marino

B.S., Slippery Rock University, 2003

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Rebecca Bosko Marino

It was defended on

August 15, 2014

and approved by

Dissertation Advisor: Jeremy J Martinson DPhil
Assistant Professor,
Infectious Diseases and Microbiology
Graduate School of Public Health
University of Pittsburgh

Committee Members:
Lawrence Kingsley DrPH
Professor
Infectious Diseases and Microbiology
Graduate School of Public Health
University of Pittsburgh

Charles Rinaldo PhD
Chair and Professor
Infectious Diseases and Microbiology
Graduate School of Public Health
University of Pittsburgh

Robert Ferrell PhD
Professor
Human Genetics
Graduate School of Public Health
University of Pittsburgh

Copyright © by Rebecca Bosko Marino

2014

Jeremy J Martinson DPhil

STRUCTURAL GENETIC VARIATION AND DYSLIPIDEMIA AMONG MEN IN THE MULTICENTER AIDS COHORT STUDY

Rebecca Bosko Marino, PhD

University of Pittsburgh, 2014

ABSTRACT

While highly active antiretroviral therapy has resulted in slowing the rate of progression to AIDS among individuals infected with human immunodeficiency virus, it has also resulted in detrimental metabolic lipid changes. As this dyslipidemia is not observed for all individuals receiving antiviral therapy, genetic factors likely influence the increased susceptibility for some. We designed this study to investigate the role of human gene copy number variation (CNV) in therapy associated dyslipidemia and risk assessment as well as investigating mRNA expression levels to identify new genetic variants associated with this lipid dysfunction.

A custom multiplex ligation dependent probe amplification assay was developed to analyze CNV of reverse cholesterol transport pathway (RCT) genes within individuals (n=320) enrolled in the Multicenter AIDS Cohort Study (MACS). The resulting analysis demonstrated that CNV was present in extremely low levels within these genes as the only loss identified and verified was observed for *CETP* in one individual. To further identify lipid metabolism associated genes, blood-derived RNA from 437 MACS participants was analyzed using the Illumina Human HT-12 microarray. Significant transcripts were present only for variation in HDL-C and Triglyceride levels with 4 differentially expressed transcripts (*HDC*, *CPA3*, *GATA2* & *SLC45A3*) repeatedly identified. Finally, to determine if CNV can alter the functionality of single nucleotide

polymorphism (SNP) genotyping, we analyzed SNPs in regions with/without CNV by Fluorescence Polarization, TaqMan SNP genotyping assays and Sanger Sequencing. SNPs in regions of no CNV were observed to have 3 distinct genotype groups but in the presence of CNV, this distinction was lost resulting in a continuous spread of allele values.

These results show that CNV is not a major factor in the development of antiviral therapy-associated dyslipidemia. Other genetic variants, such as *HDC*, may explain some of the variability. Furthermore, when CNV is present it hinders the ability to SNP genotype when using the standard assumption of three genotype groups. As antiretroviral therapy is becoming more available for the over 35 million living with HIV-1, identification of factors leading to antiviral-associated dyslipidemia is important for Public Health. Here, we have identified genes that could serve as markers for lipid level changes helping physicians custom tailor therapy and care for these individuals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The work within this document is the original, soon to be published, independent work by Rebecca B Marino.

The study documented here was conceived after an epic battle over the "graduate early because you know it all" project. It promised design freedom, complexity and the possibility of a great new find... Yet gave many moons of troubleshooting, swapping experiments, and questioning my sanity. Further still periods of elation during moments of success when the lab gods decided to grant favor upon my humble bench. While the findings at first seemed small, they weren't at all. And now at the end, I find that there is more I want to comprehend of these lipid-associated genes. However, this torch will be passed to a future student who may glance at my final notebook to find a list of lab life lessons tucked away in the back outlining the pitfalls that can set a PhD back.

I would like to thank my advisor, Dr. Jeremy Martinson, for mentoring me as a lab technician and giving me the push to pursue a PhD. I aspire to be as knowledgeable as him and understand that this involves **READING** the literature. I would like to thank my doctoral committee; Dr. Lawrence Kingsley, Charles Rinaldo, and Dr. Robert Ferrell, who helped to guide the work in the right direction when I was too close to see what I needed. I would also like to thank the staff and faculty of the Department of Infectious Diseases and Microbiology for providing the environment and knowledge necessary for me to complete my doctoral degree. Not to mention, the office gals and Joe for the random chit chat and confidence boosting support when I'd stop by for office supplies or order issues.

I would like to thank my parents Christine Bosko and Joseph Bosko (Michelle Bosko), my grandparents; George and Erma Anthony, Joseph and Anna Bosko, my Aunt and Uncles; Aunt Lynnie and Delvy McElwain, George and Lynda Anthony, my siblings; Joseph and Megan Bosko and friends; Monica Jo Tomaszewski, Matt Nicholaou, LaToya Strong, Sara Chadwick, Laura Wasil, Stella Berendam, and Anwesha Sanyal.

You provided love, support, advice and at times financial support during this process that made it all possible. Thank you to those who provided the extra inspiration needed to push this along. Meg, your constant admiration, made me want to strive to make you proud of your big sis while Delvy's constant "So when are you graduating?" made me just want to be done.

I'm finally done!

Most of all, I would like to thank my husband D. Joshua Marino. Your unwavering support, even in the midst of my most irritable moments, enabled me to keep going when the science wasn't being friendly. I only hope to be able to provide the same support as you tackle grad school.

Without all of you, I would have never made it this far or accomplished so much. Thank you!

1.0 INTRODUCTION

1.1 HUMAN IMMUNODEFICIENCY VIRUS

Human immunodeficiency virus (HIV) is a single-stranded, positive-sense, enveloped RNA lentivirus of the Retroviridae family. This family of viruses reverse transcribes their RNA genome to double-stranded DNA using an RNA-dependent DNA polymerase, reverse transcriptase (RT), packaged in the mature virion. Once transcribed, the 10kb viral genome is then integrated into the host genome of the infected cell using the virally provided integrase enzyme at which point HIV is completely dependent on the host cell for transcription and translation of its genome during replication.

1.1.1.1 Types

There are two types of HIV; HIV-1 is a highly infectious global pandemic[1] while HIV-2 is endemic mainly in West Africa due to reduced pathogenicity and poor transmission rates[2] . Classification of HIV-1 subgroups is based upon genetic sequence differences. There are 4 HIV-1 lineages, of which group M represents the predominant (~90% of HIV-1 infections) while groups N, O and P have extremely low prevalence[3]. Within group M, there are subtypes A-H that vary in prevalence based on geographical location across the world. Phylogenetic analysis of HIV-1, HIV-2 and various lineages of SIV (chimpanzee, gorilla, sooty mangabey) have

illustrated that HIV-1 and HIV-2 lineages are products of multiple cross-species transmission events[3]. HIV-1 groups M and N are most closely related to the chimpanzee SIV (SIVcpz) and are likely to originate from SIVcpz while group P appears to be of gorilla SIV (SIVgor) origin[3-6]. The outlier group O appears does not distinctly resemble any of the SIV lineages analyzed and appeared to be of either SIVcpz or SIVgor origin[3]. HIV-2 groups A-H (only A-B have spread to humans) on the other hand are of sooty mangabey origin[2, 3, 7, 8].

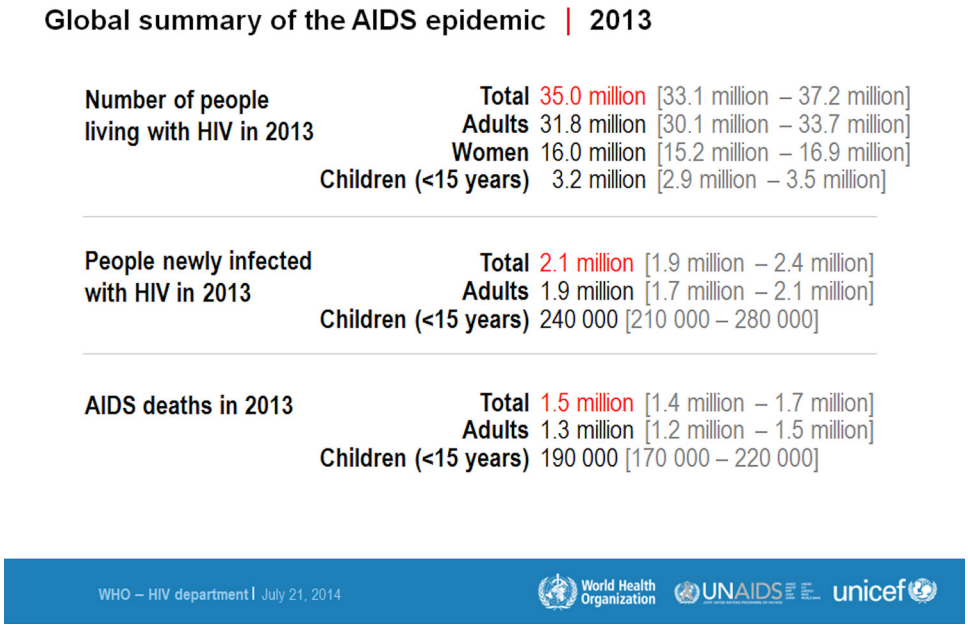


Figure 1.1: AIDS Epidemic Statistics

1.1.1.2 Global Burden

The burden of HIV infection is evident in the 2013 statistics released by the World Health Organization (WHO) that illustrated an estimated 35 million people worldwide were living with HIV-1 and that 2.1 million people were newly infected that year. Those that had AIDS related deaths were 1.5 million, a lower number than those infected indicating that the number of those living with HIV will continue to rise. The majority of individuals for each of these stats were

adults as illustrated in the WHO table shown in Figure 1.1. But roughly 9% of those living with HIV were children under the age of 15. The regions (Africa and South East Asia) with the highest prevalence of HIV-1 infection also have the highest rates of AIDS associated death as illustrated in the 2012 WHO world maps in Figure 1.2.

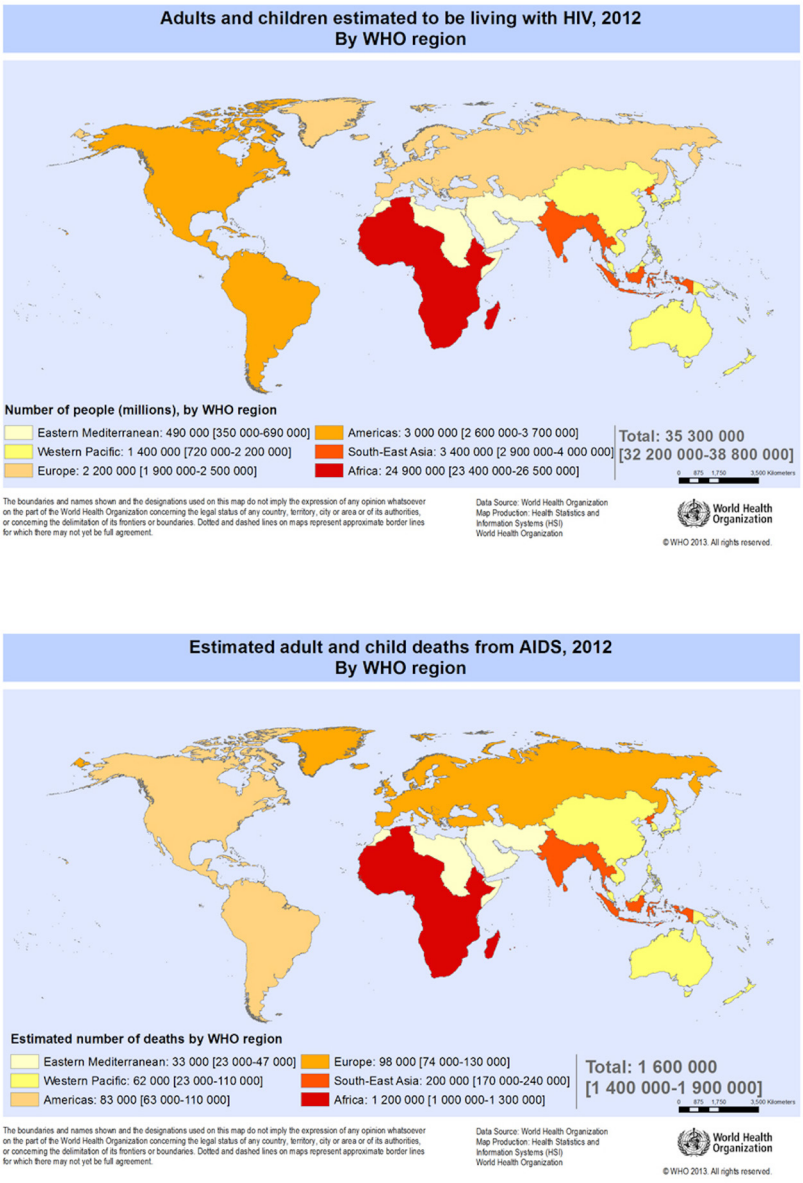


Figure 1.2: World Maps of HIV Burden and AIDS Deaths

1.1.1.3 HIV-1 Life Cycle

HIV-1 is transmitted through direct contact of blood or bodily fluids. The primary modes of transmission are through unprotected sex, interaction with infected blood (accidental or intentional needle prick) and mother to child. Initial infection occurs with the CCR5-trophic strain of the virus that predominantly infects macrophages. The HIV infects the target cell through interaction of the viral envelope protein gp160 (comprised of gp120 dimer & gp41 heterodimer) with the cell surface. Initially gp120 binds with the CD4 surface protein and then binds to one of two chemokine coreceptors (CCR5 or CXCR4) based on the tropism of the infecting virus. Binding of the coreceptor triggers gp41 mediated fusion entry. Following fusion, the viral genome is uncoated and released into the cytoplasm where the positive-sense single stranded RNA virus is reverse transcribed using the viral reverse transcriptase into double stranded DNA. The provirus is then migrated into the nucleus where the viral genome is integrated into the host cell genome using viral integrase. Once integrated, the viral genome is transcribed by the host cellular machinery with the aid of viral accessory protein Tat. Some of the resulting transcripts are then cleaved into mRNA, while others are left as whole transcripts (full length RNA genome for new viruses) before moving out to the cytoplasm. The viral mRNA is then translated into viral proteins using host machinery in the cytoplasm (Gag-Pol precursor polyprotein: capsid, matrix, integrase and RT) or endoplasmic reticulum (Envelope polyprotein precursor). Viral proteins derived from the Gag-Pol precursor polyprotein form a complex that migrates to the host cellular membrane for viral assembly while Envelope proteins are transported to the Golgi apparatus for glycosylation and cleavage into mature Env proteins before being transported to the cellular membrane. The viral proteins and genome assemble at

the membrane and then bud from the cell. The now free virus uses its viral protease, packaged during assembly, to cleave Gag-Pol generating a mature virus capable of infecting other cells.

1.2 HIGHLY ACTIVE ANTI-RETROVIRAL THERAPY

Prior to the advent of anti-retroviral therapy, the only outcome for individuals diagnosed with HIV was progression to AIDS and succumbing to opportunistic infections. The remaining lifespan after infection without treatment is generally short.

Antiretroviral drugs are divided into 5 classes based on how they interfere with the life cycle of HIV-1. These include entry inhibitors, fusion inhibitors, reverse transcriptase inhibitors, integrase inhibitors and protease inhibitors.

Entry inhibitors – Prevent binding of the virus to surface receptors.

Fusion inhibitors – Interfere with virus' ability to fuse to the cell membrane.

Reverse transcriptase inhibitors

Nucleoside Reverse Transcriptase Inhibitors (NRTI) – compete with cellular deoxynucleotides during reverse transcription. Incorporation results in truncated sequence.

Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) – Bind to site on reverse transcriptase changing its confirmation and inhibiting its function

Integrase inhibitors – Prevent integration of viral dsDNA transcript from integrating with the host genome.

Protease inhibitors – Inhibit viral protease from cleaving Gag-Pol during maturation.

As reverse transcriptase is highly error prone, the viral genome is mutated regularly resulting in escape mutants that are no longer affected by antivirals in the way that they were initially. In some cases the drug is completely ineffective against the new mutant virus. Because of this, drug therapy for HIV infection combines a cocktail of antiviral drugs that is adjusted per patient basis to avoid drug resistance. This combination therapy that began in 1996 is called highly active antiretroviral therapy (HAART) and consists of at least two or more classes of antivirals. Use of HAART has been effective of reducing viral loads in patients to undetectable as well as elevating CD4+ T-cell counts.

1.3 HAART-ASSOCIATED DYSLIPIDEMIA

1.3.1 Infection Alone

HIV-1 infection alone has been shown to result in metabolic changes. Before the advent of HAART, as infected individuals progressed to AIDS they were observed to have decreases in high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), total plasma cholesterol, apolipoprotein-A-1 and apolipoprotein-B-100, as well as increases in triglycerides, free fatty acids and very low density lipoprotein (vLDL)[9]. This combination of hypertriglyceridemia (increase in vLDL and triglycerides) and decrease in HDL-C has been shown to increase susceptibility to atherosclerosis[10]. A possible genetic component to metabolic changes is suggested by Grunfeld's study observing that only 53% of patients experienced hypertriglyceridemia. Additionally, during autopsies of children who died of AIDS prior to HAART, damage to the vascular endothelium was observed. As these children lack the

traditional risk factors (smoking, high blood pressure, obesity, etc) for coronary heart disease, HIV-1 infection related changes are implicated in their abnormal development of endothelial damage[11, 12].

1.3.2 HAART-Associated

Following development of HAART, particularly use of protease inhibitors, individuals experience changes in lipid levels observed as increases of triglycerides, cholesterol, vLDL and LDL along with increased incidence of hypertriglyceridemia. Of those receiving therapy in a study by Carr et al, 74% of patients experienced triglyceride and cholesterol increases. As HDL levels remain decreased in these individuals, this lipid triad of decreased HDL and increased triglycerides and LDL is deemed atherogenic dyslipidemia due to the increased risk of atherosclerosis known to be associated with this combination of lipids. In addition to serum lipid alterations, some experience redistribution of body fat termed lipodystrophy, a combination of lipoatrophy and lipohypertrophy. Lipoatrophy is observed as fat loss from regions of the body (arms, legs, buttocks and face) while lipohypertrophy is excessive fat accumulation in the dorso-cervical region as well as hepatic, cardiac, intra-thoracic and subcutaneous regions. The prevalence of lipodystrophy ranges from 11 to 88% in several studies and is a consequence of a lack of any standard set of criteria defining lipodystrophy[13-15]. Despite this, the abnormal fat distributions in conjunction with metabolic changes in HDL and triglycerides along with insulin resistance (observed to increase in those on HAART as well) are some of the criteria that define metabolic syndrome, which is a combination of abnormalities known to increase risk of CVD in the general population. Traditional risk factors for CVD in the general public include

hypertension, smoking, diabetes, obesity, family history of CVD, total cholesterol, LDL-C, HDL-C, lipoprotein (a), triglycerides.

Furthermore, multiple studies have reported increased CVD risk and myocardial infarction (MI) incidence in those on therapy exhibiting characteristics of dyslipidemia. Janizewski et al evaluated the hypertriglyceridemic waist phenotype, a screening tool that had reliably identified visceral adiposity[16] and elevated CVD risk in the general population[17], for its ability to identify CVD risk among those infected with HIV. In doing so they not only showed that the screening tool was effective, but also found that individuals with the highest triglyceride levels and waist circumference had the highest levels of adipose tissue, metabolic syndrome, type 2 diabetes and highest Framingham risk score for CVD risk[18]. They also observed that increased visceral fat accumulation and peripheral fat atrophy as seen in decreased subcutaneous leg fat are associated with elevated triglycerides and decreased HDL. Furthermore, individuals without lipodystrophy were observed to have reduced CVD risk factors, except for HDL levels, when compared with individuals with fat redistribution[19].

Within the data collection on adverse events of anti-HIV drugs (DAD) study group, a steady increase in incidence of MI was observed as exposure to antiretroviral therapy increased before plateauing off at the 4yr mark[20]. They also observed an increased MI incidence, partially explained by dyslipidemia, for individuals exposed to protease inhibitors for more than 6 years when compared to those with no exposure[21]. No association of MI was observed for non-nucleoside reverse-transcriptase inhibitors. Other factors identified to have an association with

increased incidence of MI were higher total serum cholesterol, triglycerides and the presence of diabetes[20].

Other studies investigated HDL particle size identifying that HAART naïve individuals had a subpopulation of HDL particles similar to those with coronary heart disease and significantly different from healthy controls[22]. Furthermore, various forms of HDL particles (large and small) were shown to be significantly associated with CVD and non-fatal coronary heart disease[23].

As the impact of HIV infection on CVD risk is unknown in those who are aging while infected with HIV, it is imperative to understand the mechanisms behind HAART-associated dyslipidemia. Because the occurrence of dyslipidemia, while having high prevalence, is not seen in the entire HIV-1 infected population, it suggests that other factors such as genetic variation play a role.

1.4 GENETIC VARIATIONS ASSOCIATED WITH DYSLIPIDEMIA

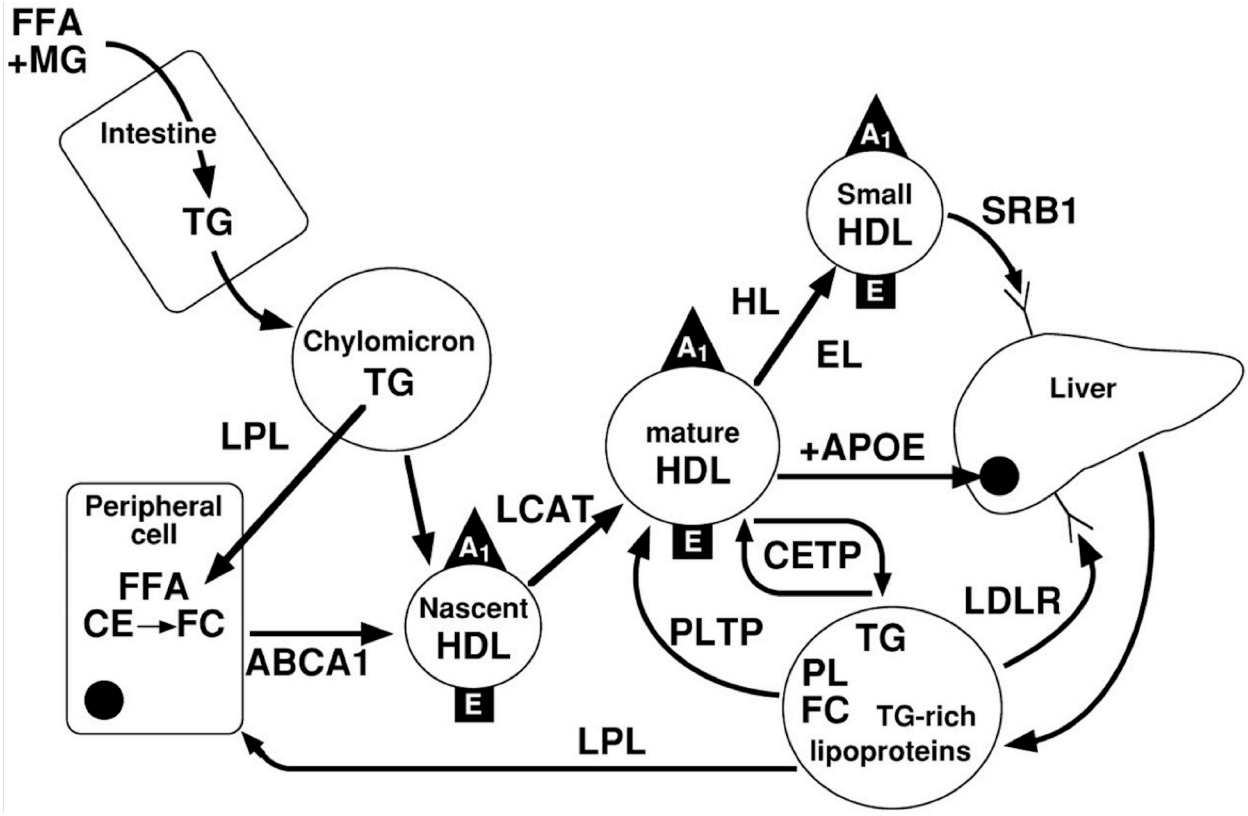
The impact of genetic variations on lipid levels within the blood is a major topic of current research. In the broad sense of genetics, gender and ethnicity have both been shown to have an association with serum lipids. Women generally have lower total cholesterol and LDL as well as higher HDL when compared to men[24]. As observed in our recent study, when investigating ethnicity those of European ancestry have the more atherogenic phenotype (higher LDL and lower HDL) while those of African European or Asian ancestry are observed to have an

atheroprotective phenotype (lower LDL and high HDL)[25]. Previous studies had similar findings[26-28]. These results would suggest that women of African descent generally have the best serum lipid profiles while Caucasian males have the worst.

Another well-studied genetic variation in the role of lipid levels is that of single nucleotide polymorphisms (SNPs) which are individual based changes within the genome that can result in one of 3 genotypes at the site of the SNP. Initially studies involved identifying candidate genes and performing small (<50 SNPs) scale investigations of the variations within. With the advent of array-based SNP genotyping platforms, the number of SNPs that could be studied at one time quickly rose exponentially from those smaller scale investigations to the current assays that can detect over a million SNPs. And thus, the genome wide association study (GWAS) was born. This type of study is an approach using the array-based platforms to examine an extremely large number of SNPs at once in order to find new associations between common disease states and SNPs on the array by analyzing samples from individuals with the condition of interest and healthy controls.

Various GWAS studies have identified polymorphisms associated with cardiovascular disease risk. Of the polymorphisms identified, many of these are involved with lipid and glucose metabolism with numerous genes falling in the reverse cholesterol transport pathway. This pathway depicted in Figure 1.3 transfers cholesterol from the periphery to the liver for degradation or recycling and is key for keeping cholesterol levels balanced in the body.

Variation within genes of and those that act upon members of the reverse cholesterol transport pathway have been shown to have a direct impact on lipid levels as well as expression levels of some of the RCT genes. Studies by Morabia and Knoblauch shown that SNPs within the RCT genes are directly associated with lipid levels[29, 30].



"Graphical representation of the reverse cholesterol transport (RCT) metabolic pathway, from which we selected 11 genes: ABCA1, APOA1, APOE, CETP, EL, HL, LCAT, LDLR, LPL, PLTP and SR-BI. Dietary free fatty acids and monoglycerides form triglycerides (TG) in the intestine, which are transported by chylomicrons. TG are also transported as very low-density lipoproteins (VLDL) formed in the liver. LPL releases free fatty acids from chylomicrons and VLDL in peripheral tissues (heart, muscle, adipose). The TG-depleted 'chylomicron remnants' take on APOE and serve, in the macrophages, as a basis for the construction of APOA1 and APOE-containing HDL particles. These nascent HDL particles interact with peripheral cells and acquire cholesterol and phospholipids through a transport process facilitated by ABCA1. Nascent HDL evolves into mature HDL in part via the PLTP-mediated transfer of phospholipids and free cholesterol from TG-rich lipoproteins to HDL, and via the esterification of free cholesterol within the HDL particle by the LCAT enzyme. These cholesteryl esters (CE) form the core of the mature HDL, which can be further enriched with APOE prior to their uptake as particles in the liver. CEs can also be selectively transferred, in exchange for TG, to TG-rich lipoproteins through the action of CETP. These TG-rich lipoproteins can then undergo hepatic endocytosis via the action of LDL-receptors. HL and EL hydrolyze HDL-TG and phospholipids, thereby reducing the size of HDL and stimulating the SR-BI-mediated selective hepatic uptake of CE."

Morabia A et al. Hum. Mol. Genet. 2003;12:2733-2743

Figure 1.3: Reverse Cholesterol Transport Pathway

In particular, Knoblauch's study revealed that SNP haplotypes within these genes explained a substantial portion of the genetic variability for LDL (67%) and HDL (58%) levels. Furthermore, mutations within the LDL-receptor (*LDLR*) have been identified to be the underlying cause of familial hypercholesterolemia (FH), a rare autosomal dominant hypercholesterolemia disorder [31, 32]. Over 1000 mutations have been reported within this gene, of which the deleterious forms result in non-functional truncated versions of LDLR that are no longer capable of binding LDL-C for transport and degradation in the liver[33]. The resulting consequence of decreased levels of functional LDLR is hypercholesterolemia, an increase of LDL-C within the plasma, which is a risk factor for atherosclerosis and cardiovascular disease. Another set of mutations that can have a dramatic impact on the RCT pathway and lipid levels are those of proprotein convertase subtilisin/kexin type 9 (*PCSK9*). The product of *PCSK9* binds to the cell surface LDL receptor molecules, targeting them for removal from the surface and subsequent degradation. Individuals who exhibit gain of function mutations are observed to have increased LDL-C levels while those with loss of function mutations have decreased levels, along with reduced risk of coronary heart disease[34-38]. Mutations within *PCSK9* have also been shown to modify LDLR expression levels.

An additional form of variation that is becoming a popular candidate for study is copy number variation (CNV). This type of variation involves a duplication, deletion or inversion of a section of DNA that can range in size from 50bp to entire sections of a chromosome. When the variation contains an entire gene including the promoter and 3' UTR, activation of transcription for the original gene can result in transcription of all of its copies provided they have the same promoter sequence of the original copy. The end result is an increase of protein product directly

proportional to the number of gene copies if no other stringent post-translation regulation exists to reduce the amount of mRNA produced. This exact phenomenon has been illustrated for *CCL3L1* and *DEFB4*, where increased copy number resulted in higher mRNA levels and increased protein levels to a point of saturation after a certain copy number has been reached, at which point the amount of protein produced reaches a plateau[39-42]. Such a variation in any of the RCT genes could disrupt the delicate balance of the reverse cholesterol transport influencing serum lipid levels (similar to that seen for *PCSK9* SNPs). Unfortunately, little is available in the literature regarding CNV in genes associated with lipid metabolism, except for rare variants in a few genes (*LDLR*, *LPL*, *ABCA1* and *LIPC*)[43, 44]. It should be noted that while *LDLR* does have an extensive amount of literature pertaining to CNV, those variants are found as insertion and deletion events within the gene rather than encompassing the whole *LDLR* gene, and are primarily seen in familial hypercholesterolemia patients[43-49]. Therefore, they are unlikely to have an impact on RCT in the general population. Furthermore, the Database of Genomic Variants (DGV), a collection of structural variation data among healthy individuals, has rare CNV documented for the RCT genes, but this primarily consists of insertions and deletions rather than whole gene duplications[50].

1.5 WHOLE-GENOME EXPRESSION

Advancing technology in the area of whole genome sequencing has made large scale genetic analysis more accessible to the scientific community. The use of this sequencing data combined with whole-transcriptome profiles can provide a more complete picture of the associations drawn during a study. However, transcriptome studies require the use of tissues specific to the analysis

at hand as gene expression varies in a tissue-specific manner. Collection of such material can be difficult if the tissue is not easily accessible, volunteers are not willing to donate or the procedure itself is too risky to undertake. Consequently, studies have sought out a surrogate for tissue specific RNA and used whole blood (PBMCs) derived RNA. Studies using circulating blood have been able to identify biomarkers and predictors of disease outcome for cardiovascular disease. And, gene expression profiles in leukocytes have identified lipid level associations with genes in lipid metabolism and the inflammatory response. These findings indicate that the use of blood derived RNA for a transcriptome analysis on CVD and atherosclerosis could produce viable targets for therapy and risk assessment.

For this reason, we designed a study to analyze CNV within the reverse cholesterol transport pathway in an effort to identify if such variation could alter expression levels of RCT genes thereby having an effect on serum lipid levels and potentially CVD risk. Additionally, we aimed to determine if whole blood derived RNA was sufficient to identify differences in expression in genes, particularly among those with extremely altered lipid levels. We also aimed to illustrate how varying copies of a gene could influence SNP genotyping, as nucleotide variation within these genes have already been shown to have associations to serum lipids. If CNV is present in these genes that hinders accurate SNP genotyping then use of those SNPs to assess disease risk would be inaccurate.

2.0 STUDY PREMISE

2.1 HYPOTHESIS

Copy Number Variation (CNV) in genes associated with lipid and glucose metabolism can alter levels of gene products, thereby disrupting the functionality of metabolic pathways and leading to HAART related dyslipidemia. This CNV may also impair SNP-based assays for disease association in these genes.

2.1.1 Specific Aim 1: Copy Number Variation in Reverse Cholesterol Transport Genes

Quantify Copy Number Variation in lipid metabolism genes and determine its association with dyslipidemia.

2.1.2 Specific Aim 2: Transcriptome Analysis

Identify gene transcripts that are differentially expressed between the atherogenic and atheroprotective phenotypes.

2.1.3 Specific Aim 3: Single Nucleotide Polymorphisms and CNV

Determine the effect of Copy Number Variation on the functionality of Single Nucleotide Polymorphism (SNP) genotyping.

3.0 MATERIALS AND METHODS

3.1 SAMPLES

3.1.1 MACS Sample Description

Experimental samples were obtained from the Multicenter AIDS Cohort Study (MACS). The MACS is a four center (Baltimore, MD; Chicago, IL; Pittsburgh, PA; and Los Angeles, CA) ongoing prospective study, founded in 1984, of the natural and treated histories of HIV-1 infection in homosexual and bisexual men. Participants attend clinics bi-annually for a physical exam and sample collection, and complete extensive questionnaires about their medical history, behavior changes, and overall quality of life.

3.1.1.1 Cross-Section from 2005

Blood samples were drawn from MACS participants in 2005, their Peripheral blood mononuclear cells (PBMC) were spun down into pellets and subsequently frozen at each center. We received pellets from 1,945 individuals (HIV seropositive n=955 & seronegative n=950) that were matched**.** DNA was extracted from these samples using the Qiagen kit, as described on page 19.

3.1.1.2 ARRA Cross-Section

As part of the American Recovery and Reinvestment Act, funding was granted to perform transcriptome analysis using a large sample set to identify genes involved in dyslipidemia that may or may have not been previously identified. Samples for this part of the study were collected from the Pitt Men's Study, the Pittsburgh site of the MACS. The 437 subjects that attended the clinic between August 2010 and July 2011 had blood drawn into a PAXgene tube for whole transcriptome analysis. Three participants had two tubes collected at the initial time of blood draw while 50 individuals had their blood sample collected during two separate visits and 3 individuals had blood collected during three separate visits. These additional samples served as quality control samples for the transcriptome assay.

3.1.1.3 Laboratory Control Samples

Control samples used for this study consisted pre-existing laboratory controls (n=5) and samples from the Coriell repository with known amounts of *DEFB103* gene CNV (n=4). The 5 pre-existing controls were initially collected as blood samples in 2004 from individuals at the Graduate School of Public Health. DNA was extracted using the Phenol Chloroform method. Samples from the Coriell Institute cell repository (NA07048, NA10846, NA10861, and NS12911) were obtained as extracted DNA.

3.1.2 Nucleic Acid Extraction

3.1.2.1 DNA

Phenol chloroform

DNA was extracted from control blood samples using a phenol chloroform extraction protocol. Blood (3-10mL) is transferred into 50mL Falcon tubes and filled with sterile distilled water to hemolyse the red blood cells. Tubes are placed on ice for 5-10 minutes then centrifuged at 3000rpm for 10min to pellet the white blood cells. The supernatant is discarded before the pellet is resuspended in 25mL of the Triton/Sucrose Buffer to lyse the white cell membrane while leaving the nuclear membrane intact. Once again the tubes are set on ice for 10min then spun at 3000rpm for 10min to pellet the white cell nuclei. The supernatant is discarded and the pellet of cell nuclei is now rinsed in 5mL of PBS before being spun at 3000rpm for 5min. The supernatant is discarded again and the pellet is resuspended in 10mL of a lysis buffer containing Proteinase "K", then incubated at 50-60 °C for 1 hour with occasional shaking. Once the incubation period has ended, 10mL of the phenol/chloroform/water mix (70:20:10 v/v) is added and shaken to emulsify before centrifuging at 3000rpm for 10min. The upper aqueous layer is transferred to a fresh 50mL Falcon tube and 10mL of chloroform is added, shaken and spun at 3000rpm for 10min. The upper aqueous layer is again transferred to a new Falcon tube and 1mL of 4M sodium acetate is added then mixed gently. In order to precipitate the DNA, 25mL of 100% ethanol is added to each tube and mixed. This tube is held at 4°C and spun at 3000rpm for 10min. The supernatant is discarded and the pellet is then washed in 10mL of 70% ethanol and spun again as above. The tubes are then inverted and allowed to air dry for 30-60 minutes before

resuspending the DNA in 1mL of 1X TE buffer. DNA concentration is either determined by Quant-iT PicoGreen dsDNA Assay Kit [Life Technologies, Carlsbad CA] described in Section 3.1.3 or A260 spectrophotometry. These samples were subsequently stored at 4°C.

Qiagen QIAmp

When samples were received as PBMC pellets, the Qiagen QIAamp DNA Blood Mini Kit [Qiagen Inc., Valencia, CA], utilizing the Blood or Body Fluid Spin Protocol, was used to extract DNA following the manufacturer's protocol. DNA yield from this procedure generally ranges from 3-12ug in an elution volume of 400uL 1X TE. Resulting genomic DNA was quantified by Quant-iT PicoGreen dsDNA Assay Kit [Life Technologies, Carlsbad, CA] and stored at -20 °C.

3.1.2.2 RNA

Blood samples from MACS participants were collected in PAXgene tubes (PreAnalytix/Qiagen, Valencia, CA) to stabilize intracellular RNA prior to processing. Tubes were inverted 8-10 times upon collection then stored upright for a minimum of 2 hours and a maximum of 72 hours to allow full stabilization of RNA before freezing at -20°C. Frozen samples were brought to room temperature prior to RNA extraction and purification with the Qiagen PAXgene Blood RNA Kit IVD that is specifically designed for extracting intracellular RNA from blood stabilized in the PAXgene Blood RNA tubes. Following extraction, samples were placed back in -20°C before being taken to the University of Pittsburgh Genomics and Proteomics Core Laboratories (GPCL) for quantification and further sample processing. Samples are currently being stored at -80°C.

3.1.3 Nucleic Acid Quantification

Quantification of double stranded DNA was performed using the Quant-iT PicoGreen dsDNA Assay Kit supplied by Life Technologies (Carlsbad, CA). This method of quantification is dependent on the ability of the PicoGreen dsDNA quantification reagent to fluoresce once it has bound DNA and a standard curve for a comparative quantification. A standard curve is generated by performing a serial dilution of the Lambda bacteriophage DNA (100ug/mL in TE) supplied in the kit. The Lambda DNA is first diluted to a concentration of 1000ng/mL by adding 1911uL of 1X TE to 39uL of the stock Lambda DNA in a 2.0 mL microcentrifuge tube. The tube is mixed and 1300uL are transferred into a new microcentrifuge tube containing 650uL of 1X TE. This process is repeated until 15 separate dilution tubes are generated. To avoid issues with retention differences between pipette tips, the same tip is used to perform the serial transfer between tubes. A 16^{th} tube containing only 1X TE is also filled as a DNA blank. From these dilutions, a 100uL volume is aliquotted vertically within a Greiner U-bottom clear 96-well plate with triplicate samples running horizontally such that the first 8 dilutions fill columns 1-3 while the following 8 fill columns 4-6.

To prepare the sample DNA for quantification, 99uL of 1X TE is added to each well of a new Greiner U-bottom plate and 1uL of sample DNA is added. As with the standard curve, these samples were also analyzed in triplicate. The sample DNA was also mixed using a 200uL pipette containing a filter tip to ensure even consistency. Previously, freezing of DNA samples was observed to clump the DNA at the bottom of stock plate wells resulting in inaccurate concentrations that ranged the extremes depending on the well height from which the sample was taken.

Once the standard curve and samples were sorted out, the thawed PicoGreen dsDNA quantification reagent is diluted by adding 172.5uL of the reagent to 34.328mL of 1X TE in a 50mL Falcon tube. The tube is gently inverted to evenly distribute the reagent before 100uL was added to each well of both the standard curve and sample plates. The plates are then sealed with AlumaSeal II [Excel Scientific/AF-100/non-sterile] and wrapped in aluminum foil to protect from light. They were allowed to incubate for a minimum of 5min at room temperature before being spun down at 1000rpm and read spectrophotometrically at the GPCL facility located on the University of Pittsburgh campus. From the standard curve a best-fit linear line was determined and from the line's equation sample DNA was quantified.

3.2 MULTIPLEX LIGATION-DEPENDENT PROBE AMPLIFICATION

3.2.1 Sample Selection

Samples (n=366) were identified on the basis of serum lipid measures obtained in their 2005 visit. We used the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) [51] to identify lipid levels associated with higher risk of heart disease (HDL-C <40 mg/dL and/or LDL-C > 130 mg/dL) or with lower risk (HDL-C > 60 mg/dL and/or LDL-C < 100 mg/dL). DNA was extracted from frozen PBMC pellets using the Qiagen QIAamp DNA Blood Mini Kit, following the Blood and Body Fluid Spin Protocol. DNAs were stored at -20°C.

We also obtained 4 DNA control samples (NA07048, NA10846, NA10861, and NS12911) from the Coriell Institute cell repository. Three of these individuals served as reference samples, as the CNV of their Defensin B103A gene *DEFB103A* was already known[52-54]. We also used 5 laboratory reference DNA samples as additional controls.

3.2.2 MLPA Probes

3.2.2.1 Reverse Cholesterol Transport Pathway

As MRC-Holland does not supply probes for all of the genes within the Reverse Cholesterol Transport pathway, we opted to use their P300 Human DNA Reference-2 probemix and augmented this with a custom probe set. Our custom RCT probes were designed according to the MRC-Holland Synthetic Probe Design Manual criteria (v10-update 04-02-2009) using the human genome 18 reference assembly for gene sequences and the SUNY Stony Brook MAPD browser (http://bioinform.arcan.stonybrook.edu/mlpa2/cgi-bin/mlpa.cgi) with the following conditions: (Hyb Temp = 60°C; Min Tm = 70°C; [Na+] = 0.35M; [Mg2+] = 0; Min delta G = 0 kcal/mol; Protocol = Electrophoresis: stuffer (JS98_v1)). [55, 56] The top ranked probe pairs (LPO & RPO) were selected and UCSC Blat searches were performed to identify oligos and complete probes that exclusively bound only the gene of interest. We designed one probe for each of our RCT genes under the assumption that if CNV encompassing a whole gene were present then a single probe would initially detect this variation, and this could be further investigated. Sequences can be found in Appendix Section: MLPA Probe Oligos on page 132.

3.2.2.2 Positive Control Copy Number Assay

To ensure our MLPA conditions were able to detect a range of CNV when a gold standard CNV reference sample is not available, we genotyped the Defensin B103A gene *DEFB103A* and the chemokine receptor gene *CCR5* in a subset of our experimental samples, and in samples from the Coriell repository that have known genotypes for these genes (NA07048 – *DEFB103A* 4 copies/*CCR5* wt/Δ32, NA10846 – *DEFB103A* 5 copies/*CCR5* wt/wt, NA10861 – *DEFB103A* 3 copies/*CCR5* wt/wt). The MRC-Holland P139 Defensin probemix set was selected as a positive control to type the defensin cluster based on its previous use to type CNV in these samples [3-5]. In addition to the predesigned defensin MLPA kit, we took our original P300/custom RCT probe mix and exchanged 3 of the lipid probes (*APOC2*, *APOA1*, and *APOE*) for a *DEFB103A* probe and another two custom probes designed to detect the full and Δ32 deletion forms of CCR5[6]. The *DEFB103A* probe used was a shortened form of the 04389-L03745 probe from the MRC-Holland P139 probemix, as the original probe length would have conflicted with another reference probe in the P300. The *CCR5* probes consisted of one left hand probe oligo that stopped at the site of the Δ32 deletion and two right hand oligos that were specific to either the wild type sequence or the sequence directly following the Δ32 deletion. With these probes, a wt/wt homozygote has two copies of the wt probe target and zero copies of the Δ32 target, a Δ32/Δ32 homozygote has zero copies of the wt target and two copies of the Δ32 target, and a wt/Δ32 heterozygote has one copy of each probe target.

All probes were synthesized through Integrated DNA Technologies (Coralville, IA) as DNA oligos. Additional 5' phosphorylation of the right-hand probe oligo was performed in our laboratory for probes that were ordered lacking this modification. This was accomplished using

T4 Polynucleotide Kinase (New England Biolabs, Ipswich, MA). Standard conditions for a non-radioactive phosphorylation call for up to 300pmol of 5' termini in a 50uL reaction volume that contains 1X of the Kinase buffer (supplied), 1mM of ATP (NEB Adenosine 5' Triphosphate – P0756S) and 10 units of T4 Polynucleotide Kinase. We followed this protocol with the exception that we used between 13-30pmol of each RPO oligo, as this was the amount present in 10uL of our rehydrated IDT RPOs. The reactions were ran for 1hr at 37°C in 8-strip tubes followed by a 20 min 65°C enzyme deactivation. Appendix Section A.1 on page 132 lists all the probe oligos and specifics about each.

3.2.3 MLPA Procedure

3.2.3.1 MLPA Hybridization and Amplification

On the first day, 5uL of ~15ng/uL DNA was added to a well within a 96 well low profile PCR plate for each sample. The plate was sealed with a polymer PCR plate sealing mat and place in an Applied Biosystems Incorporated GeneAmp PCR System 9700 thermocycler (Life Technologies, Carlsbad CA). The samples were denatured at 98°C for 5 minutes and cooled to 25°C. Once at room temperature, 3uL of the vortexed hybridization master mix was added to each sample using an 8-channel pipette containing filter tips. To ensure proper mixing of the probes with the sample, pipette mixing was used when the master mix was added and the samples were then spun down in the centrifuge to remove air bubbles. The plate was once again placed into the thermocycler, this time with a compression pad on top of the polymer sealing mat to ensure a tight seal. The samples were then incubated for 1 minute at 95°C and then incubated at 60°C for 18-19 hours overnight.

On the second day, the samples were cooled to and held at 54°C in the cycler while 32uL of Ligase-65 mix (3uL Buffer A, 3uL Buffer B, 25uL sterile distilled H_2O, & 1uL Ligase-65 enzyme) was added to each well then mixed gently in with the pipette. To ensure as little evaporation as possible, eight-strip caps were used to cover each column of the plate when the compression pad and sealing mat were removed. The cycler program was continued at 54°C for 15 minutes followed by an enzyme heat inactivation at 98°C for 5minutes. The samples were then cooled down and held at 15°C. At this point, 20uL of the ligation product was transferred to a new plate for the one-tube PCR protocol (MDP-v001, update 17-06-2011). At room temperature, 5uL of the polymerase master mix containing a FAM labeled PCR primer was added to each sample and mixed with a pipette. The plate was then spun and placed in a cycler preheated to 60°C where the ligation products were then amplified using the cycling program listed below. Following amplification, 5uL of the PCR product was added to a plate containing 15uL of sterile distilled water in order to dilute the brightness of the product before the plate was sealed with AlumaSeal II [Excel Scientific/AF-100/non-sterile] and wrapped in aluminum foil.

3.2.3.2 MLPA Fragment Separation Conditions

Fragment separation was performed at the University of Pittsburgh Genomics and Proteomics Core Laboratories, using an ABI 3730xl Genome Analyzer (Life Technologies, Carlsbad, CA) with the following conditions: 1.6 kVolts injection voltage; 25 kVolt run voltage; 50cm capillary; 10 sec injection time; POP7 column; LIZ labeled GS-500 size marker standard.

3.2.4 MLPA Analysis

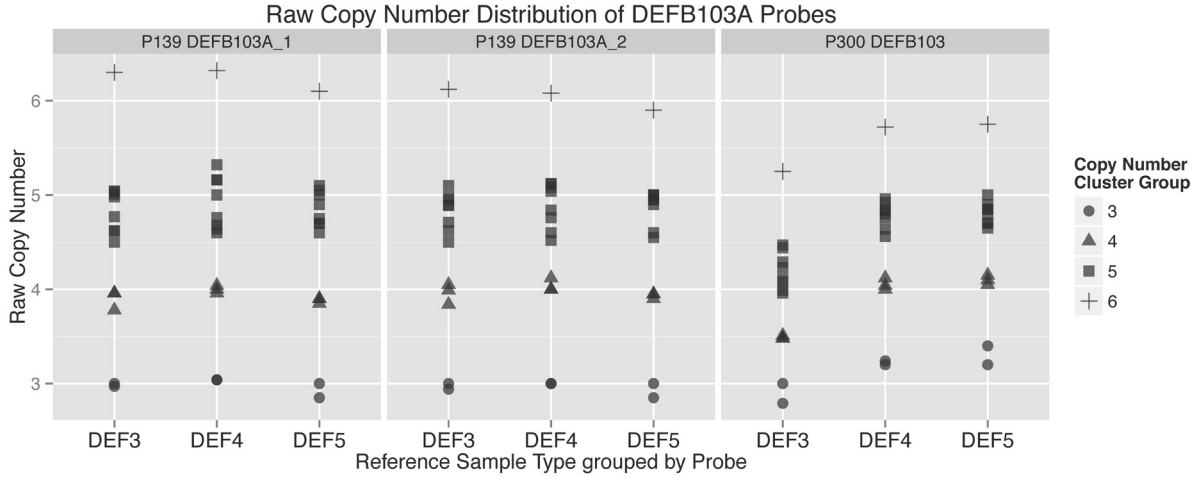
The Coffalyser.net software (http://wiki.coffalyser.net) was used to perform fragment and comparative analysis. Initially, no reference samples were indicated, and samples with average signal across all probes were selected as references for each of the four 96 sample runs. Fragment analysis was then performed for a second time with default settings. Samples with poor reference probe quality and reproducibility were removed before comparative analysis was performed for a second time. Final ratios and standard deviations were analyzed with the R statistical software package[57] and the following modules (ggplot2[58], reshape[59], & gridExtra[60])

3.2.5 Copy Number Calling

For probes that lacked reference samples with known copy number levels, discrete copies were not calculated. Instead, the default ratio thresholds (0.7, 1.3), defined by MRC-Holland and based on a 2-copy reference sample (MLPA Results Interpretation – V02.2;11-02-2010)[61], were used to identify individuals who exhibited a gain (>1.3) or loss (<0.7) of copy number. The interquartile range (IQR) of each probe was also compared to that of the reference probes to identify experimental probes with potential CNV that did not cross the default threshold.

When a reference sample containing a known amount of copies was available for a probe, copy number was initially called by multiplying the ratio with the number of copies in the reference then rounding to get a discrete copy. When this method produces inconsistent results, copy number was typed by applying a *k*-means clustering algorithm to the raw copy number calls[62].

Additional reference samples with differing copies of the gene of interest served to verify group calls when they clustered with their expected group (Figure 3.1).



The P300/RCT MLPA assay was extended to include the shorter DEFB103A probe from the P139 Defensin assay. Both assays were run simultaneously on 14 samples, including 3 Coriell samples previously typed for DEFB103A. Each Coriell sample was set as a reference in separate analyses and the resulting ratios were converted to raw copy numbers by multiplying them with the known number of copies in the reference. The column titles indicate each of the 3 DEFB103A probes analyzed while the shape of each point represents the copy number called by k-means clustering. Most copy number groups clustered around a whole integer value, with the exception of the 5-copy group. This group skewed towards the 4-copy group, and at times a few samples crossed the halfway point between the two groups resulting in incorrect copy number calls when raw copies were simply rounded to whole integers. Use of a k-means clustering algorithm allowed correct calls each time. All of the raw copy number calls seen for the DEFB103A probe in the 3-copy reference sample using the P300/RCT probe set were lower than expected. The run for this sample was of poor quality and therefore ratios generated using it had the potential to fluctuate. Even so, cluster analysis still resulted in proper calls.

Figure 3.1: Reproducible Typing of Copy Number Calls depends on the Distribution of Raw Copy Numbers around Whole Integer Values

3.3 POLYMERASE CHAIN REACTION (PCR)

3.3.1 Primer Design

Primers were designed using the Primer3 design tool in MacVector v11.0.2 utilizing default parameters. The reference sequences of genes in this study were acquired from GenBank and used in MacVector to generate a list of potential primer pairs for each region of interest. The resulting primer pairs were further analyzed to identify primers with similar annealing temperatures that formed no self 3'-dimer, hairpin, or self duplexes and generated PCR amplicons that ranged from 300-600bp.

3.3.2 Primary Amplification

Initial PCR products for both Fluorescence Polarization and Sanger Sequencing were generated in the same fashion in which optimal amplification conditions were first identified before amplifying all samples.

3.3.2.1 Optimization

Magnesium titrations and temperature gradients were performed simultaneously for each set of primers to identify their optimal magnesium concentration and annealing temperature for PCR. To perform this optimization, 4 individual mastermixes containing a range of magnesium were made using a polymerase reagent kit (Applied Biosystems AmpliTaq Gold or Sigma Taq DNA Polymerase #D4545-250UN), 1X TE primer mix containing both the forward and reverse primers ordered as oligos from Integrated DNA Technologies (Coralville, Iowa), dNTPs

(Invitrogen), sterile distilled water and control DNA. Enough mastermix was made to perform 10 individual reactions per Magnesium concentration and Table 3.1 lists the reagent volumes per individual sample.

Table 3.1: Magnesium Gradient Mastermix Setup per Individual Tube

		Volume per PCR Reaction Tube (uL)			
	Stock	**Magnesium Concentration**			
Reagent	**Concentration**	**1.0mM**	**1.5mM**	**2.0mM**	**2.5mM**
Buffer	*10X*	1.0	1.0	1.0	1.0
$MgCl_2$	*25mM*	0.4	0.6	0.8	1.0
Primer Mix	*50mM each*	0.2	0.2	0.2	0.2
dNTPs	*25mM*	0.1	0.1	0.1	0.1
Taq Polymerase	*5U/uL*	0.1	0.1	0.1	0.1
Sterile Distilled H_2O		7.2	7.0	6.8	6.6
DNA	*5ng/uL*	1.0	1.0	1.0	1.0
	Total Volume	10	10	10	10

Each mix was then aliquotted at 10uL/tube into separate 0.2mL 8-tube strips (ISC Bioexpress, Kaysville UT) that were placed in an Eppendorf Mastercycler Gradient where the gradient cycling program amplified each sample of the strip at a separate annealing temperature ranging from 57°C to 63°C in the PCR reaction. Touchdown PCR cycling programs were design to enhance the specificity of amplification when regular Taq was used. These programs begin 7°C above the recommended annealing temperature and decrease 0.05°C degrees every cycle until cycle 15 where they remain at their optimal annealing temperature for the remainder of the PCR. As touchdown programs would not harm amplification with hot-start Taq Polymerases such as AmpliTaq Gold, they were used for all of our PCRs. We also opted to omit the extended initial

heat denature step for AmpliTaq Gold and allow the enzyme to be activated slowly throughout the PCR in a time released fashion. Following amplification, PCR products are analyzed using gel electrophoresis as documented in Section 3.3.2.2.

3.3.2.2 Gel Electrophoresis

Verification of a successful PCR amplification was performed using gel electrophoresis. If the PCR product was greater than 100bp, a 2% agarose gel was made by combining 5g of GenePure LE Agarose powder [ISC BioExpress #E-3120-500] with 250mL of 0.05X TBE Buffer (10X TBE concentrate [AMRESCO] diluted with distilled water) in a 500mL Pyrex bottle and microwaved at 50% power for 5 minutes. If the PCR product was less than or close to 100bp, a 3% GenePure HiRes agarose gel [ISC BioExpress #E-3115-500] was made. The powder for this agarose is finer and allows clearer separation of shorter PCR products. GelRed [Biotium #41003], a non-toxic stain for nucleic acids was then added to the melted agarose that was then allowed to cool for 5-10 minutes. At this point, the agarose was poured onto a gel tray in a casting stand and combs were positioned onto the plate before it was placed in the refrigerator to cool. Once the gel solidified the combs are carefully removed to ensure no damage to the wells and the gel tray was placed in the electrophoresis tank [Enduro Gel XL Electrophoresis System] filled with 0.05X TBE buffer. Each sample, prepared by taking 1uL of PCR product and adding it to 6X loading dye (0.06% Bromophenol Blue, 0.06% Xylene Cyanole FF, 1.5% Ficoll-400), was loaded to individual wells in the submerged gel plate with the first well of each row containing 5uL of the ΦX174 DNA ladder (12.5ng/uL NEB ΦX174-HaeIII Digest in 1X loading dye). The samples were then migrated through the gel from the negative electrode to positive at 100V (400mA) for 30 minutes. Once the Bromophenol Blue dye has migrated near to the edge of the gel, it was imaged on an Alpha Innotech Red Imaging System [Protein Simple] using a

UV transilluminator to activate the fluorescent GelRed stained DNA allowing visualization of the PCR products as well as the DNA ladder. For our singled sized PCR products only one band is expected and the DNA ladder is used to confirm a successful PCR by estimating product size in relation to the ladder.

3.3.3 Post-PCR Clean up

For any PCR products that are going to be further amplified, the additional primers and dNTPs need to be removed so that only the desired secondary reaction amplifies. This was performed using our standard EXOSAP clean-up step. The Alkaline Phosphatase (Roche rAPid Alkaline Phosphatase, used in place of Shrimp Alkaline Phosphatase) and Exonuclease I (NEB) Mastermix shown in Table 3.2 is added directly to the PCR product tubes in a volume equal to that of the initial PCR. The tubes are then placed into the thermocycler where they are held at 37°C for 1hr before a 15min heat kill at 85°C to inactivate the enzymes.

Table 3.2: EXOSAP Mastermix

Reagent	Concentration	Volume
		uL per sample
rAPid Alk Phos Buffer	*10X*	1.0
rAPid Alk Phos Enzyme	*1U/uL*	1.0
Exo I	*20U/uL*	0.05
Sterile Distilled H_2O		7.95
	Total Volume	10

3.3.4 Sanger Sequencing

Sanger sequencing was performed using the Applied Biosystems BigDye Terminator v3.1 Cycle Sequencing kit (Life Technologies, Carlsbad, CA; Cat #4337457). Initial PCR is performed in the same manner as explain in the Primary Amplification methods section on page 28 while the technique for removing excess primers and dNTPs is found in the Post-PCR Clean up section on page 31. Following clean up, 5uL aliquots of the product were transferred into two optical plates (ABI MicroAmp) serving as the forward and reverse sequencing plates. Two separate mastermixes consisting of the 5X Sequencing Buffer, BigDye Terminator v3.1 Ready Reaction Mix and a 1uM dilution of primer (Initial PCR forward or reverse primer unless otherwise noted) were aliquotted to the wells of their corresponding plate at a volume of 5uL per well (Table 3.3). The plates were then sealed with silicone sealing mats and placed into a double 96-well block ABI GeneAmp PCR System 9700 and amplified using the preloaded BigDye program (Table 3.4). Following amplification, the sequencing products were Ethanol precipitated as per our standard protocol listed in Section 3.3.5 and taken to the Genomics and Proteomics Core Laboratories (GPCL) for capillary electrophoresis runs on an ABI 3730xl following resuspension in 10uL of formamide. The resulting .ab1 files were then analyzed using Gene Codes Sequencher DNA sequencing software v5.2. Ambiguous bases were determined to be either true SNPs or sequencing noise. After identification of the SNP of interest in the sequence, the height and area of each allele were analyzed by eye and with the PolySNP software package to compare to the findings in other genotyping assays.

Table 3.3: BigDye PCR Mastermix

Reagent	Concentration	Volume per Sample (uL)
BigDye Sequencing Buffer	*5X*	2.0
1:500 Primer	*1uM*	2.5
BigDye Terminator 3.1 Reaction Ready Mix		0.5
		5.0

Table 3.4: BigDye PCR Cycling Conditions

Step	Temperature	Time
	°C	*Min:Sec*
1	96.0	1:00
2	50.0	0:05
3	60.0	4:00
Repeat Steps 1-3 25 times		
6	4.0	hold

3.3.5 Ethanol Precipitation

In order to concentrate the products of the sequencing reactions, we performed ethanol precipitations. This was accomplished by placing the final sequencing product (max volume of 20uL) into a sequencing plate then adding 5uL of 250mM EDTA and 60uL of 100% ethanol.

The plate was then sealed with a silicon plate mat, inverted to mix, covered in foil and allowed to incubate for 15min so that the EDTA could bind free dye-labeled bases that were not incorporated into the sequencing product. Following incubation, the plate was cooled to 4°C while being spun at 2500g for 30 minutes. The PCR product was now attached to the bottom of the plate, which was inverted and spun to 185g into a stack of paper towels to remove the supernatant. To further clean up the PCR product, a wash with 60uL of 70% EtOH was performed and the plate was then again spun at 4°C but at 1650g for 15 minutes. The plate was then inverted again to remove the supernatant before being lightly covered with foil and air-dried for 30 minutes. Once dry, the plate was sealed with film and covered with foil. It should be noted that if the product being precipitated does not contain any light sensitive components then the precipitation can be performed without foil.

3.4 REAL-TIME QUANTITATIVE PCR

3.4.1 Single Nucleotide Polymorphism Assays

Genotyping of SNPs by real-time PCR was performed using Applied Biosystems TaqMan SNP Genotyping Assays (both ready-made and custom designed) in combination with Applied Biosystems Genotyping Mastermix. For the custom assays, primers and minor groove binding probes were designed using Primer Express V3.0 in our laboratory to have greater control over the primer positions. These probes were then ordered as complete assays through Life Technologies' (formerly Applied Biosystems) Custom TaqMan Assay Design Tool or ordered as individual MGB probes that were then paired with their IDT-ordered primers. The 40x assay

mixes that were prepared in the laboratory using 1X TE, where primers had a final concentration of 900nM and MGB probes had 200nM, were subaliquoted after preparation to reduce the amount of freeze thaw cycles.

The pre-made assays are supplied at a concentration of 40x (also subaliquoted after first thaw) and both types of assays were diluted in the genotyping mastermix described in Table 3.5. This mix was then aliquotted using an Eppendorf Combitip Plus in 25uL volumes to each well of a white well plate [GeneMate, 96 well, Low Profile, white, 0.2ml PCR Flat Top plates/T-3184-W] before 1uL of DNA [5ng/uL] was added by an 8-channel pipette. The plate was sealed with 2 mil thick optically transparent sealing film [Excel Scientific/TS-RT2-100/ThermalSealRT/Non-Sterile], spun down at 1000rpm for 1min, and amplified using an Eppendorf Mastercycler realplex[4] using the program listed in Table 3.6.

Table 3.5: Real-Time PCR SNP Genotyping Mastermix

Reagent	20x Assay	40x Assay
ABI Genotyping Mastermix	12.5	12.5
Assay Mix	1.25	0.63
Sterile Distilled H2O	11.25	11.87
	Total Volume	25

Table 3.6: PCR Cycling Conditions for TaqMan SNP Genotyping

Step	Temperature °C	Time Min:Sec	Measuring Point
1	95.0	10:00	
2	92.0	0:15	
3	55.0	0:15	
4	60.0	1:00	yes
Repeat Steps 2-4 80 times			
5	60.0	2:00	yes

Lid temperature = 105 °C, Temperature mode = Standard, Application Type = Quantification

While the program does not need to run for 80 cycles for all assays, we set it for more cycles than needed to ensure amplification is not prematurely ended before the reagents are exhausted. Once the assay was run for the first time, the number of cycles could be shortened based on the cycle where the assay plateaued. Analysis to generate a SNP genotype depends upon the cleanness of the genotyping. For each allele, the assay contains a separate probe labeled with either FAM or VIC fluorescent dye. The software for the RealPlex generates baseline corrected data for each of these reads after it calculates optimal baseline for each individual read for each sample. When the genotyping data is clean, the endpoint (last cycle of the run) for each dye will have clear separation of curves between samples with increased fluorescence indicating presence of the allele over samples that ran near to baseline that lack the allele. By plotting the FAM by the VIC fluorescence, four distinct groups will cluster in the plot. Those individuals who are homozygous for the SNP will only have an increase along one axis, those who are heterozygous will have an increase along both axes**,** while the samples that failed to run during the assay will

cluster near the origin of the plot at 0,0. If the genotyping data is not clean, there will be no distinct separation of the normalized fluorescent signal for an allele but rather a spread. In this case, instead of plotting the endpoint, we opt to plot an earlier cycle where all samples had fluorescent signal that was increasing exponentially.

3.4.2 TaqMan Expression Analysis

Verification of the top differentially expressed transcripts during the transcriptome analysis was carried out using ABI TaqMan Gene Expression Assays on a select set of RNA. Initial verification involved selecting 15 individuals with the most extreme expression levels for the following genes; ABCA1, CD8a, CD8b, HDC, and CPA3. RNA consisted of the same samples used for the Transcriptome analysis with the exception that these samples were not treated with GLOBINclear.

3.4.2.1 Synthesis of cDNA

The ABI High Capacity RNA-to-cDNA kit was used to synthesize cDNA from PAXgene derived RNA. Approximately 500ng of stock RNA, concentrations determined by Agilent 2100 Bioanalyzer system, was used in the reaction listed within Table 3.7. The mastermix containing only buffer and enzyme was aliquoted in an 11uL volume to each well containing 500ng of RNA brought up to 9uL using nuclease-free water. Low concentration samples had less than 500ng of RNA that varied based on the max volume (9uL) of RNA allowed within the 20uL cDNA reaction. This reaction was conducted in an ABI GeneAmp PCR System 9700 at 37°C for 1hr

then 95°C for 5min to heat kill the enzymes followed by a 4°C hold. The newly generated cDNA was stored in -20°C to ensure stability for expression analysis.

Table 3.7: Mastermix for cDNA Synthesis

Reagent	Sample	
	Individual (uL)	16 (uL)
2X RT Buffer	10.0	160.0
20X RT Enzyme	1.0	16.0
RNA and Nuclease-Free H_2O	9.0	
	Total Volume	20

3.4.2.2 TaqMan Gene Expression Assays

Gene expression assays were conducted following the TaqMan Gene Expression Assay protocol (PN4333458N) with TaqMan Universal Mastermix II containing UNG. A single pre-designed gene expression assay was ordered for each gene identified in the transcriptome along with the GAPDH endogenous control assay and stored at -20°C. Assays were subaliquoted in 50uL volumes to reduce the amount of freeze thaw cycles as the efficiency of the assays are known to decrease as the number of thaws increases. For each sample, 20uL of the mastermix Table 3.8 containing probes for both the gene of interest and endogenous control was added to a well in an Eppendorf plate (twin.tec semi-skirted PCR Plate 96 – 951020346) then 1uL of cDNA was mixed in using an 8-channel Eppendorf pipette. Each sample was ran in triplicate on the Eppendorf Realplex^4 using the cycling conditions in Table 3.9.

Table 3.8: Real-Time Expression Assay Mastermix

Reagent	Individual Sample	8 Samples Triplicate
20X Gene Expression Assay	1.0	27.0
20X GAPDH Assay	1.0	27.0
2X Mastermix	10.0	270
Sterile Distilled H_2O	8.0	216
	Volume Aliquotted per Sample	20.0
cDNA		1.0
	Total Volume	21.0

Table 3.9: TaqMan Expression Assay Cycling Conditions

Step	Temperature °C	Time Min:Sec
1	50.0	2:00
2	95.0	10:00
3	95.0	0:15
4	60.0	1:00

Repeat Steps 3-4 40 times (Max of 80)

The Eppendorf RealPlex software was used to analyze the resulting data. A sample identified as having the lowest expression for each of the HT-12 transcripts was selected as the reference

sample and set as the calibrator sample for relative comparison using the delta delta CT method. Resulting expression values were then plotted against their matching HT-12 values to observe if individuals had the same degree of expression levels.

3.5 TRANSCRIPTOME ANALYSIS

The Illumina Human HT-12 v.4 whole-genome expression array enables genome-wide expression analysis through probes that detect >47,000 transcripts for well-characterized genes, gene candidates and splice variants. The 500 RNA samples for this assay were those derived from the PAXgene kit described in Section 3.1.2.2. Once extracted, the RNA was processed at the University of Pittsburgh's Genomics and Proteomics Core Laboratory where samples were quantified then purified with GLOBINclear before cDNA was synthesized for transcript expression analysis. Illumina GenomeStudio V2011.1 was subsequently used to generate raw transcript data from the BeadChip image files prior to exporting a report containing sample and control probe profiles that would be further processed within the R statistical software package. For statistical analysis of HT-12 data, the following Bioconductor R modules were used; ArrayQualityMetrics[63], lumi[64], limma, GOstats, ggplot2[58], and gplots. Lumi was used to transform (log or VST) and normalize (quantile or RSN) the data set while at the same time enabling annotation of each sample with categorical data. To ensure that the data being analyzed was of the best quality, we processed our expression dataset using ArrayQualityMetrics to identify outliers to the dataset. The samples that were identified to be of poor quality (n=53) were marked in the dataset to be excluded from further analysis. As an additional QC step, we used the ComBat module in R to correct for batch effects among the different array chips.

This clean expression set was then subsetted using variables within the categorical data to include two separate groups for comparison and followed by analysis with limma. Using the command lmFIT, a linear model was fitted to each gene to estimate fold change and standard errors. This was followed up with the eBayes command that applies empirical Bayes smoothing to the standard errors. The final result is then displayed using the command topTable to display the statistics for the top 10 differentially expressed genes. If significant comparisons were identified based on the corrected p-value then all of the transcripts with $p<0.05$ were further analyzed. This was accomplished by plotting heatmaps in ggplot2 in order to visualize up- and down-regulation of genes per sample as well as running Gene Ontology analysis within GOstats to identify the pathways that the differentially expressed genes fall within.

3.6 NANOSTRING CNV ANALYSIS

A NanoString nCounter custom CNV CodeSet was designed containing the 16 RCT genes analyzed in our MLPA assay along with 20 genes of interest for CNV, 4 genes with CNV, 2 genes without CNV and another 8 genes/locations that illustrated potential CNV. In total 351 samples were analyzed by NanoString including 267 of the experimental samples previously typed by MLPA. DNA was processed and analyzed using a NanoString Technologies nCounter system at the University of Pittsburgh Genomics and Proteomics Core Laboratories. The NanoString nSolver Analysis Software (v1.1) was used to generate normalized ratios from raw counts. For CNV analysis of MLPA genes, 34 of the original MLPA samples were set as RCT references in nSolver. The ratios generated in this software were exported, rearranged using

Excel and Filemaker Pro then analyzed within the R statistical software package. Analysis within R involved plotting the ratios using ggplot2 & gridExtra along with reshape to arrange the data in the format necessary for the preferred plots.

4.0 AIM 1: COPY NUMBER VARIATION AND RCT GENES

Individuals infected with HIV-1 exhibit changes in serum lipid levels seen as hypercholesterolemia and hypertriglyceridemia[9, 19, 21, 65]. Following antiretroviral therapy, lipid levels remain skewed for many patients, as LDL-cholesterol (LDL-C) and triglycerides increase while HDL-cholesterol (HDL-C) remains lowered[9, 19, 21, 65-67]. Previous studies have shown that this dyslipidemic profile is associated with greater risk for cardiovascular disease (CVD), myocardial infarction and atherosclerosis in HIV-positive individuals[18, 19, 21, 23, 68, 69].

As some in the HIV-1 infected population have begun to reach the age where CVD risk is increased and the effect of HIV infection on this risk in unknown, there is a need to understand the mechanisms behind therapy-associated lipid dysfunction. The prevalence of dyslipidemia is high, but not all-inclusive, among the HIV-positive population suggesting that genetic factors potentially have a role[70]. Studies have already illustrated a broad genetic impact on lipids, as lipid levels and CNV risk vary based on ethnic background in HIV uninfected populations[26-28]. We have recently shown that biogeographical ancestry was significantly associated with lipid levels in a cohort of MSM, and that European ancestry results in a more atherogenic phenotype even after controlling for HIV and therapy[25].

Furthermore, several genome-wide association studies (GWAS) have identified polymorphisms associated with CVD risk[71-80], many of which are present in genes involved in cholesterol metabolism and transport. One particularly relevant set of genes is that of the reverse cholesterol transport (RCT) pathway, which directly influences lipid levels. Polymorphisms in genes of this pathway, and in those directly interacting with it, contribute to the variance of lipid levels, and also alter expression levels of some of the genes themselves[29, 30, 35, 36, 70, 80-83]. For instance, expression levels of *LDLR* can be modified by mutations in the proprotein-convertase subtilisin-kexin type 9 gene (*PSCK9)*, ultimately resulting in altered levels of LDL-C[81, 84, 85]. Individuals with loss of function mutations in *PCSK9* show decreased amounts of LDL-C while those with gain of function mutations have increased amounts[34-38].

In addition to posttranslational protein regulation such as that seen with *PCSK9*, protein levels of RCT gene products could also be influenced by copy number variation (CNV). This type of genetic variation includes duplications, deletions, and inversions of DNA segments greater than 50bp in size[86-89]. Previous studies on CNV in the *CCL3L1* and *DEFB4* genes illustrate that an increase in transcriptionally available copies of a gene not only results in increased expression levels but also increases in protein levels directly proportional to the number of copies[39-42]. Such variation in one or a few RCT genes has the potential to alter the functionality of this lipid metabolism pathway dramatically, and thereby influence serum HDL and LDL levels. Yet, while there have been a number of studies investigating the association of SNPs within these genes to lipid levels[29, 30, 34, 35, 80, 82, 83], little has been documented related to their CNV, apart from reports on rare structural variation in the *LDLR* gene associated with Familial Hypercholesterolemia [43-49] and the occasional reported variant in *LPL*, *ABCA1*, and *LIPC*[43,

44]. Furthermore, the Database of Genomic Variants, a compilation of structural variation in healthy control sample genomes, contains only rare CNV encompassing the RCT genes[50].

Here, we designed a study employing custom Multiplex Ligation-dependent Probe Amplification (MLPA) and NanoString probes to screen for CNV in 16 RCT associated genes (Table 4.1) in participants from the Multicenter AIDS Cohort Study (MACS), to identify if CNV is present, the degree to which it varies, and whether it has an association with the abnormal lipid metabolism observed in HIV-positive individuals undergoing antiretroviral therapy.

Table 4.1: Reverse Cholesterol Transport (RCT) Pathway Genes Selected for Analysis

Gene Name	Symbol	Chromo-some	Function	Ref
Scavenger Receptor Class B, Member 1	SRBI	12	Plasma membrane receptor for HDL that mediates transfer of cholesterol to and from HDL.	29
Apolipoprotein C-III	APOC3	11	Very low density lipoprotein that inhibits lipoprotein lipase and hepatic lipase delaying triglyceride-rich particle catabolism	30
Apolipoprotein A-I	APOA1	11	Major protein component of HDL and a cofactor of LCAT. Defects in APOA1 results in HDL deficiencies	29,30
Apolipoprotein E	APOE	19	Main apoprotein of chylomicron and essential for catabolism of triglyceride-rich lipoprotein constituents	29,30
Phospholipid Transfer Protein	PLTP	20	Lipid transfer protein that transfers phospholipids from triglyceride-rich lipoproteins to HDL	29
Hepatic Lipase	LIPC	15	Triglyceride hydrolase and ligand/bridging factor for receptor mediated lipoprotein uptake.	29,30
Lecithin-cholesterol Acyltransferase	LCAT	16	Extracellular cholesterol esterifying enzyme that esterifies cholesterol for transport.	29,30
Apolipoprotein A-IV	APOA4	11	Potent activator of lecithin-cholesterol acyltransferase	30
Lipoprotein Lipase	LPL	8	Triglyceride hydrolase and ligand/bridging factor for receptor mediated lipoprotein uptake.	29,30
Endothelial Lipase	LIPG	18	Regulates circulating levels of HDL and acts has phospholipase activity.	29
Low Density Lipoprotein Receptor	LDLR	19	Cell surface protein involved in receptor-mediated endocytosis of LDL	29,30
Cholesteryl ester transfer protein	CETP	16	Transfers cholesteryl esters between lipoproteins	29,30
Apolipoprotein A-V	APOA5	11	Component of high density lipoprotein	30
Apolipoprotein B	APOB	2	Main apolipoprotein of chylomicrons and low density lipoproteins	30
ATP-binding cassette, sub-family A, member 1	ABCA1	9	Membrane associated protein that functions as a cholesterol efflux pump in the cellular lipid removal pathway	29,30
Apolipoprotein C-II	APOC2	19	Plasma lipid-binding protein that activates lipoprotein lipase	30

4.1 SAMPLE DEMOGRAPHICS

Using the 2005 clinic measurements and the NCEP/ATP III report criteria[51] (HDL-C $\leq$40 mg/dL or $\geq$60 mg/dL; LDL-C $\leq$100 mg/dL or $\geq$130 mg/dL), 366 suitable MACS participants were identified, of which 320 yielded sufficient DNA for MLPA analysis. The demographic data for these 320 samples are summarized in Table 4-2. We identified 23 samples with an atheroprotective phenotype [HDL-C $\geq$60 mg/dL and LDL-C $\leq$100 mg/dL] and 7 samples with an atherogenic phenotype [HDL-C $\leq$40 mg/dL and LDL-C $\geq$160 mg/dL]. Those with the atherogenic lipid profile had a higher mean BMI, plus higher total cholesterol and triglyceride levels when compared to those who had the atheroprotective phenotype. Age among all lipid groups was similar, with a median age of 48 (IQR: 47-49). BMI was higher in the uninfected individuals within each grouping and, with the exception of the atheroprotective group, the mean BMI of most groups ranged from overweight to borderline obese. Total cholesterol levels increased with increasing lipid levels for both HDL-C and LDL-C groups, while Triglyceride levels decreased with increasing HDL-C.

Table 4.2: Demographics and Descriptive Characteristics of Study Participants

		Atheroprotective HDL >60 mg/dL & LDL <100 mg/dL		Atherogenic HDL <40 mg/dL & LDL>160 mg/dL		HDL-C						LDL-C							
						≤40 mg/dL		40-60 mg/dL		≥60 mg/dL		≤100 mg/dL		100-130 mg/dL		130-160 mg/dL		≥160 mg/dL	
		HIV -	HIV +	HIV -	HIV+	HIV -	HIV +	HIV -	HIV +	HIV -	HIV +	HIV -	HIV +	HIV -	HIV +	HIV -	HIV+	HIV -	HIV +
	n	8	15	3	4	52	91	73	43	22	38	48	76	38	44	38	24	19	14
	AGE (years)	45	42	43	46	47	49	48	48	48	47	47	47	49	50	47	48	49	49
	BMI	25.2	22.3	34.0	29.8	30.2	26.0	27.4	25.5	26.0	23.4	28.6	24.9	28.0	25.7	28.0	24.0	28.7	26.2
Mean	HDL (mg/dL)	73.7	79.4	35.7	31.2	34.2	30.8	49.8	48.1	67.5	77.6	47.2	43.4	47.5	42.5	44.1	53.0	50.6	52.5
	LDL (mg/dL)	87.8	76.7	170.0	188.8	116.9	98.4	121.3	116.2	115.7	112.4	79.0	76.4	115.7	114.0	144.5	143.0	173.7	178.6
	TCHOL (mg/dL)	181.0	184.9	241.0	254.8	193.0	171.2	195.5	192.3	206.1	216.6	156.5	152.8	195.3	190.3	219.3	227.7	253.3	257.8
	TRIG (mg/dL)	97.9	144.9	176.0	175.5	209.5	222.2	127.2	142.3	114.5	136.4	152.4	168.7	162.8	173.3	155.4	158.5	145.3	135.8
# on Therapy	No Therapy	-	4	-	1	-	26	-	5	-	9	-	19	-	9	-	6	-	3
	Monotherapy	-	0	-	0	-	0	-	1	-	0	-	0	-	1	-	0	-	0
	Combination	-	2	-	0	-	6	-	0	-	2	-	7	-	1	-	0	-	0
	Potent ART	-	9	-	3	-	59	-	37	-	27	-	50	-	33	-	18	-	11
Therapy Adherence	100%	-	5	-	2	-	22	-	14	-	11	-	18	-	12	-	7	-	7
	95-99%	-	3	-	1	-	35	-	20	-	11	-	30	-	18	-	8	-	3
	<75%	-	3	-	0	-	5	-	3	-	7	-	7	-	3	-	3	-	1
	NA	-	4	-	1	-	29	-	6	-	9	-	21	-	11	-	6	-	3
BGA	AEA	1	11		1	7	10	18	7	4	16	7	18	12	6	5	6	3	1
	AsEA	1	3			1	3	2	1	1	4	2	5			2	1		
	EA	6	1	3	3	42	77	52	35	17	17	37	52	25	38	31	17	16	12
	NA					2	1	1			1	2	1	1					1
	Group Total	23		7		143		136		60		124		82		62		33	

BMI, body mass index; HDL-C, High Density Lipoprotein Cholesterol; LDL-C, Low Density Lipoprotein Cholesterol; TCHOL, Total Cholesterol; TRIG, Triglycerides; Monotherapy, single nucleoside reverse transcriptase inhibitor; Combination, two or more nucleoside reverse transcriptase inhibitors; Potent ART, two or more nucleoside reverse transcriptase inhibitors with a proteaseinhibitor or a nonnucleoside reverse transcriptase inhibitor; BGA, Biogeographical Ancestry; AEA, African/European ancestry; EA, European ancestry; AsEA, Asian European ancestry.

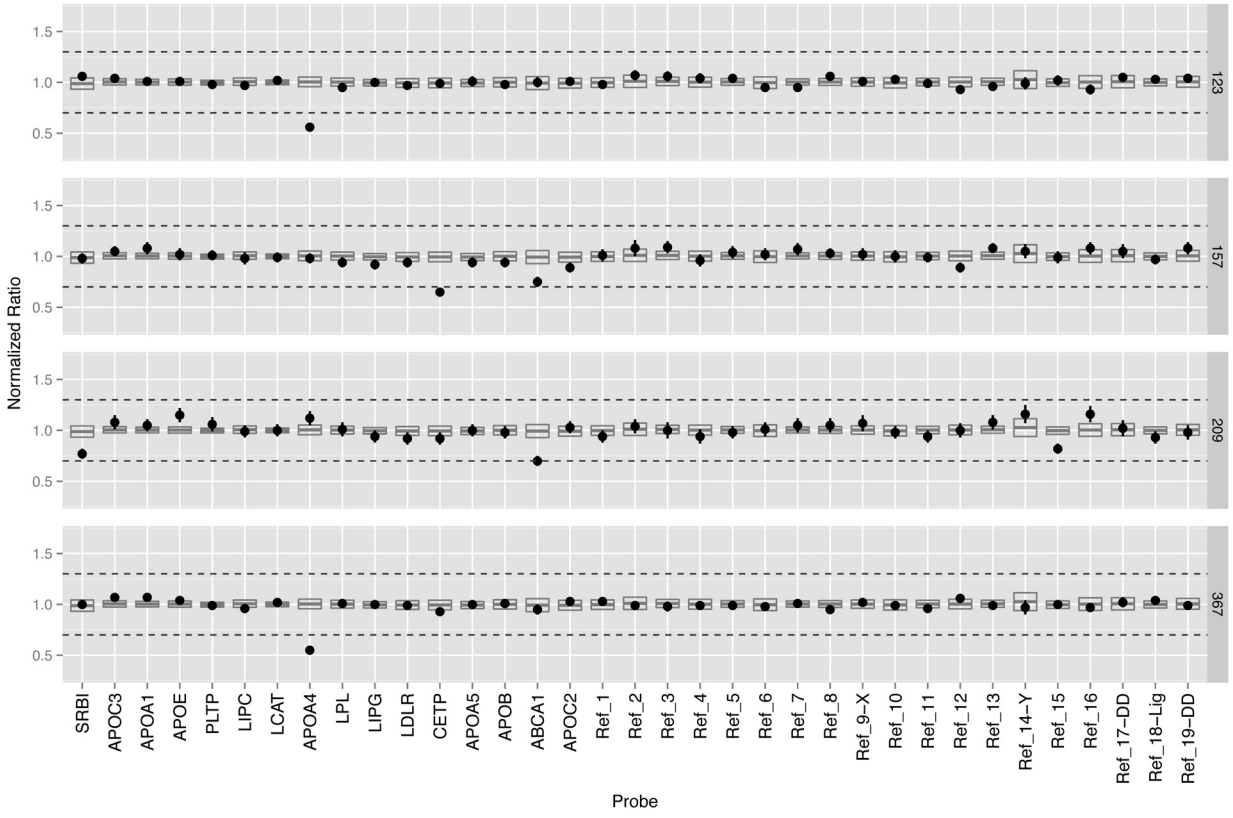
Of the individuals who were HIV-positive during 2005, over 70% were receiving a type of antiretroviral therapy (Potent ART, combination, or monotherapy defined according to the DHHS/Kaiser panel criteria[90]). Of those who reported therapy use, around three quarters had over 95% adherence. We also compared the distribution of Biogeographical Ancestry (BGA), recently determined for these samples[25], for the sample subset. The majority of samples in each group were those of European ancestry, followed by those of mixed African-European ancestry, and a few samples with Asian-European ancestry. However in the HIV-positive groups with high HDL-C ($\geq$60 mg/dL) and in the atheroprotective group, samples with African-European ancestry were in the majority.

4.2 MULTIPLEX LIGATION DEPENDENT PROBE AMPLIFICATION

As reference samples with known copies of the RCT genes are not available, we identified experimental samples whose normalized peak height for each probe was similar to the sample set mean height. Using these samples as references, the coffalyser.net software calculated the probe ratios for each sample relative to the reference samples. Assuming that the most frequently observed ratio corresponded to two copies per diploid genome, ratios above 0.7 and below 1.3 are considered to be within the normal range of two copies[61]. Anything outside of these thresholds was identified as an outlier with potential CNV.

Of the 16 RCT pathway associated genes screened, only three (APOA4, CETP, and ABCA1) showed any signs of CNV, and in each case the CNV was extremely rare. For each of these genes, a few individuals showed ratios that crossed or were at the lower threshold (Figure 4.1).

None of the RCT genes had CNV that passed the upper ratio bound of 1.3, suggesting that no sample showed gains in copy number. Table 4.3 lists normalized ratios of the three genes for samples with losses along with the median HDL-C and LDL-C levels for a minimum of 8 visits surrounding the 2005 visit.



Copy number ratios are shown for the four individuals that had detectable CNV. Probes representing the RCT genes are on the left of the figure while reference probes (Ref_1 – Ref_16), ligation controls (Ref_18), and denaturation controls (Ref_17, Ref_19) are on the right. The dots show the copy number ratios of each probe for each individual. The box plots represent the 95% confidence interval of each probe ratio derived from the entire sample set. Arbitrary thresholds at 1.3 and 0.7 are represented by the dotted horizontal lines. Points that fall within these thresholds are considered to have a copy number ratio of 1.0.

Figure 4.1: Copy Number Variation is Exceedingly Rare for Reverse Cholesterol Transport Pathway Genes.

Two samples (123 & 367) had a loss of *APOA4* copy number (with normalized ratios of 0.56 and 0.55, respectively) while sample 157 had a loss of *CETP* copy number with a ratio of 0.65. Sample 209 had a normalized ratio for *ABCA1* that fell on the 0.7 threshold indicating a possible loss. The standard deviations for each of these outlying probes were relatively small ($<\pm 0.05$) indicating that the decrease in ratio observed was likely genuine. When MLPA was performed for a second time on these 4 samples, the observations were consistent with the first run. There was no association between any of these losses and lipid levels (Table 4.3), although this observation is not conclusive due to the small number of samples involved.

Table 4.3: Normalized Ratios of RCT Pathway CNV Probes that Showed Significant Departure from Unity.

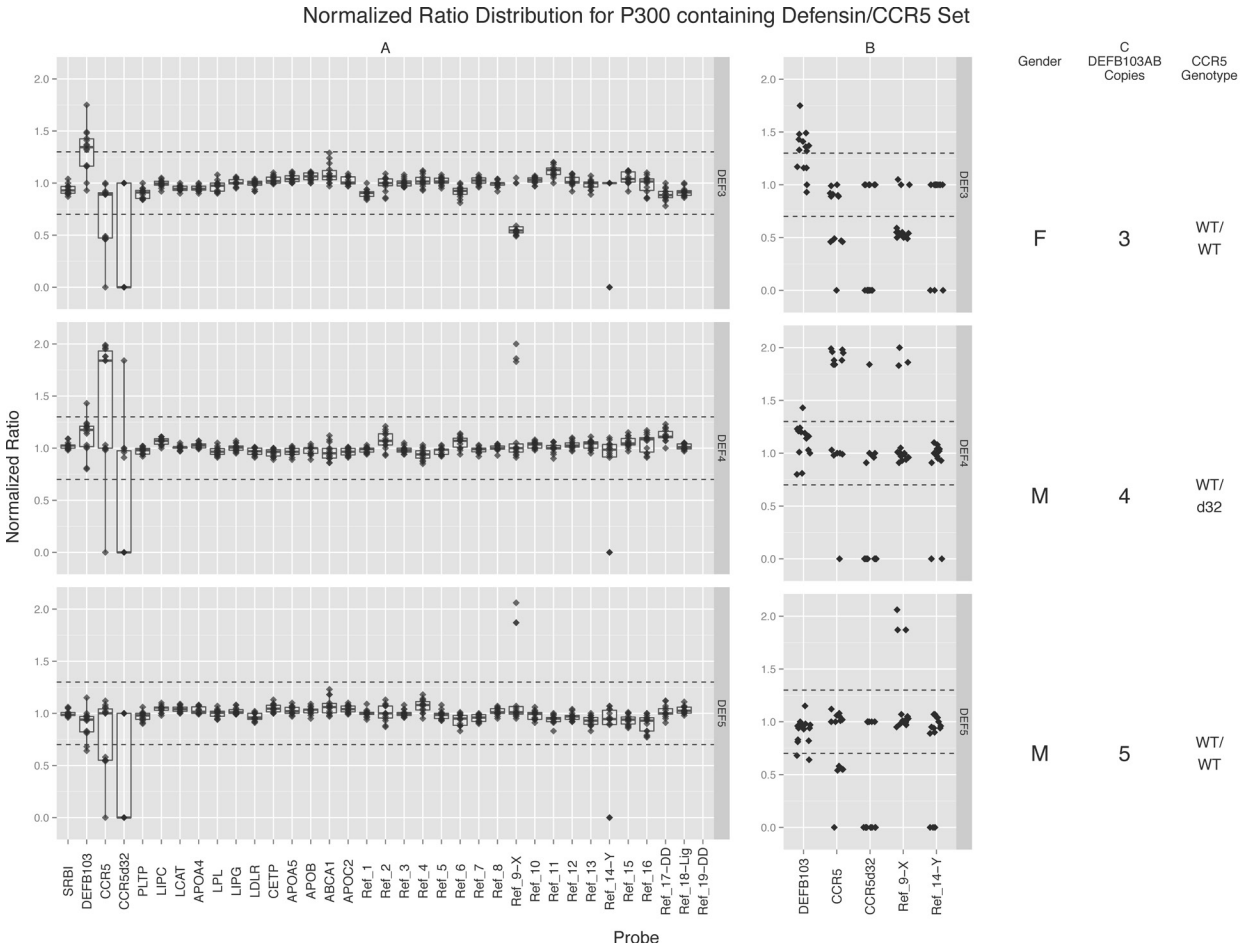
	Lipid Levels		Normalized Ratios			
Sample	**LDL-C [mg/dL]**	**HDL-C [mg/DL]**	**ABCA1**	**APOA4**	**CETP**	**SRBI**
123	147 (132.25-162.25)	55 (51-68.3)	1	**0.56**	0.99	1.06
157	114 (103-129)	30 (24.8-38.1)	0.75	0.98	**0.65**	0.98
209	136.5 (124.25-148.5)	38 (36-40.4)	**0.7**	1.12	0.92	0.77
367	69 (49-76)	39.2 (35.8-54.1)	0.95	**0.55**	0.93	1.0

The first two columns list median serum HDL-C and LDL-C levels from a minimum of 8 visits for that individual. Within the brackets is the IQR range for those lipid levels. Probes that crossed or were on the 0.7 ratio threshold are indicated with (*).

The interquartile ranges (IQR) for most of the RCT gene probes were narrow (0.04-0.09), and similar to those of the 2-copy reference probes. This tight clustering of ratios around the mean of each probe further suggests that CNV is not common in the RCT genes. The only probes with wider IQRs were *ABCA1* and *SRBI*. While their IQRs were slightly broader than the other RCT probes, this spread of the ratio distribution was also seen in the properly functioning reference probes suggesting that this was within the normal range of our experiment (data not shown).

As “gold standard” referents containing known copy numbers of each RCT gene are not available, and the P300 reference probe set includes probes with a maximum number of 2 copies, we developed a quality control assay to ensure that our MLPA protocol was capable of picking up a range of CNV above 2 copies. Through use of control samples with known amounts of CNV, the MRC-Holland P139 Defensin MLPA assay, and our P300/RCT assay extended to include probes for variants in the *DEFB103A* and *CCR5* genes, we were able to verify that our assay can identify a range of CNV even when the sample mean is used to define a referent as illustrated in Figure 4.2.

In this extended assay, one probe was designed to detect the DEFB103A gene, known to show widespread CNV in humans[52-54]. Probes were also designed to detect the wild-type and the Δ32 forms of the CCR5 gene[91]. By using these probes, together with reference samples who’s CNV for them was known, we could determine the ability of the default ratio thresholds to detect CNV when the number of copies in the reference sample is unknown. We studied this in a subgroup of 14 individuals (11 males and 3 females; 5 references of known CNV and 9 samples chosen at random from the MACS study set). We set each of our 3 Coriell reference samples as the referent in separate analyses and used the remaining samples to validate the copy number groups identified.



*Column A shows the ratios obtained from each analysis using the entire probe set while column B contains the subset of probes that show CNV. Each row in the figure represents the results from each separate analysis using a different reference sample of known CNV and CCR5 genotype. The CCR5 genotype, DEFB103A copy number, and gender of the referent is shown in column C. Arbitrary thresholds at 1.3 and 0.7 are represented by the dotted horizontal lines. Points that fall inside these thresholds are considered to have the same copy number as the reference sample when a 2 copy reference sample is used. Each dot represents an individual person, the hinges of the box and whisker plots indicate the first and third quartiles of the observed range, and each whisker extends to the furthest value from the median that is within 1.5 * IQR of the hinge.*

Figure 4.2: Probes and Reference Samples Demonstrating a Range of CNV

The referent for the analyses in the top row is a female who is homozygous for the wild-type (wt) allele of the CCR5 gene and has a validated DEFB103A copy number of 3. In this case, the presence of one single additional DEFB103A copy in a test sample is sufficient to cross the threshold line and be detectable as a copy number variant. In fact, the 4-copy and 5-copy samples appear distinct from each other. This referent has two copies of the wt allele of CCR5 and no

copies of the Δ32 allele: thus, with the wt allele probe the samples that are also homozygous wt/wt have a copy ratio of 1, the samples with the heterozygous wt/Δ32 genotype have one copy of the wt allele, giving a CCR5 wt probe copy ratio of 0.5, and the sample that is homozygous Δ32/Δ32 has no copies of the wt allele and gives a wt copy ratio of zero. The ratios obtained with the Δ32 probe show the errors that occur when the referent does not contain any copies of a probe target. In this case, the algorithm cannot discriminate between samples with one copy of the target present and samples with more than one copy present. All of the homozygous wt/wt samples here give a copy ratio of zero, as they do not contain the target for the Δ32 probe, but both the heterozygous wt/Δ32 and homozygous Δ32/Δ32 samples give ratios of 1. Lastly, this referent is female, and therefore has two copies of the X chromosome control probe, and zero copies of the Y chromosome control probe. The male samples thus give a ratio of 0.5 with the X chromosome probe and the female samples give a ratio of 1. With the Y chromosome probe, the females have a probe ratio of zero and the males have a ratio of 1.

The referent for the analysis in the middle row is a male, with 4 copies of DEFB103A, who is heterozygous wt/Δ32 at the CCR5 gene. In this case, both the 3-copy and 4-copy DEFB103A samples fall within the threshold lines and cannot be discriminated from each other, but the ratio seen with a 5-copy sample is greater than the threshold and can be identified. The referent has one copy each of the wt and Δ32 alleles of CCR5, thus the CCR5 genotypes of the samples can easily be determined: with the wt probe, wt/wt homozygotes have a ratio of 2, wt/Δ32 heterozygotes have a ratio of 1, and the Δ32/Δ32 homozygote has a ratio of 0. The reverse relationship is seen with the Δ32 probe. The referent is male, and has one copy each of the X and

Y chromosome probe targets. Male samples have one copy of the X chromosome target whereas females have two, and males have one copy of the Y chromosome target and females have zero.

The referent for the analysis in the bottom row is a male, with 5 copies of DEFB104A, who is homozygous for the wt allele of CCR5. In this case, both the 4-copy and 5-copy DEFB103A samples fall within the thresholds and cannot be discriminated from each other, but the 3-copy samples have a probe ratio below the threshold. As this referent lacks the target for the Δ32 probe, the same erroneous ratios are seen as in the top row for this probe, but as it has two copies of the wt probe the wt/wt, wt/Δ32, and Δ32/Δ32 can once again be identified. The sample is a male and therefore has one copy each of the X and Y control probes, allowing the sex of the samples to be identified as in the row above.

These reference samples and probes show that our assay is highly sensitive to detecting CNV when the referent has one or two copies of probe target, and is still sensitive when the referent has three copies. Sensitivity begins to decline when the referent has more copies, but samples that differ from the referent by two or more copies can still be distinguished. Also, the spread of the ratio distributions for the DEFB103A probe with true CNV was noticeably larger than that of the 2-copy reference probes and our RCT probes. The IQR did decrease as the copy number of the referent increased but it still remained larger than that of the 2 copy genes even when the largest DEFB103A copy number referent was used. These results suggest that our RCT gene probes have accurately detected the CNV present in our samples, but that this CNV is limited in scope.

4.3 SANGER SEQUENCING

To determine whether the rare loss for the three RCT genes identified during MLPA reflected true CNV or problems with probe binding, Sanger sequencing was performed to examine the probe binding site for those individuals who showed losses, plus several control individuals who showed no changes in copy number. We determined that individuals who showed a loss in signal for *APOA4* were heterozygous for a rare Single Nucleotide Polymorphism (SNP), rs185210669, located 1 base from the ligation site for that probe. This mutant allele fails to bind the MLPA probe, leading to impaired ligation and decreased MLPA signal. The other genes (*ABCA1* and *CETP*) contained no SNPs within their ligation sites.

4.4 CNV CONFIRMATION BY NANOSTRING

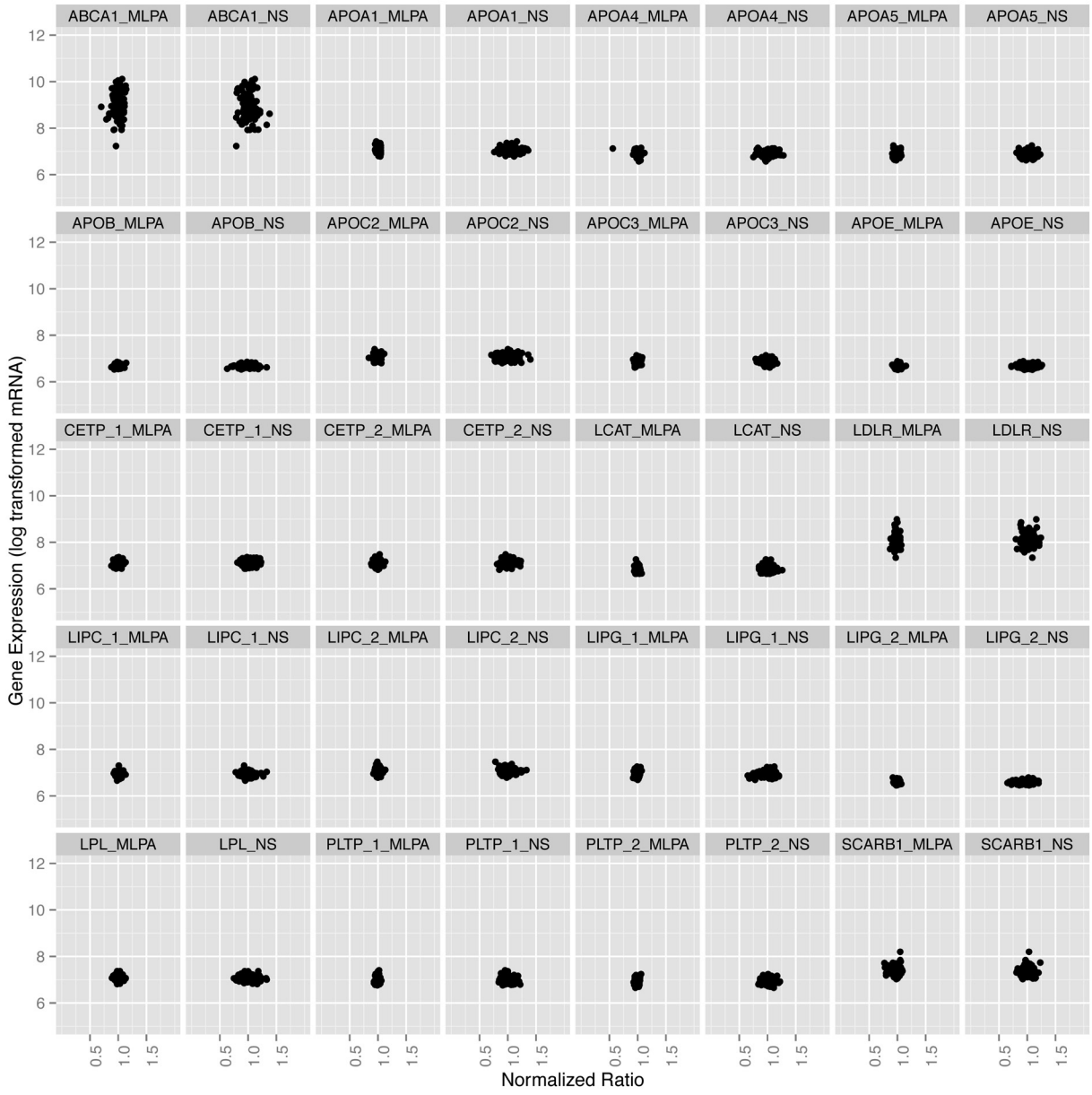
We confirmed our findings by using a custom NanoString assay to measure CNV of the RCT genes for 267 of the samples analyzed by MLPA. The CNV ratios generated mirrored those seen with MLPA (data not shown). We replicated the loss in copy number of *CETP* in sample 157 (copy number ratio of 0.58) but did not observe the losses for *ABCA1* in sample 209 (1.06) or *APOA4* for both samples 123 and 367 (1.10 & 1.04). As the MLPA-derived ratio for *ABCA1* in sample 209 fell on the threshold value of 0.7, it is likely that this sample does not in fact have a true loss in copies. It is also possible that MLPA probe used for *ABCA1* is picking up a rare small CNV that is not detected by the NanoString probe, as the probes for these assays bind in different

regions of the gene. The loss observed in *APOA4* only by MLPA is attributed to the ligation-site SNP identified in the MLPA probe.

4.5 EXPRESSION ANALYSIS

We also determined the expression levels of the RCT genes in our study, using data extracted from a whole-genome transcription dataset obtained using the Illumina HT-12 platform. Gene expression levels on 127 samples were compared to both MLPA and NanoString CNV ratios. As expected, comparisons of MLPA- and NanoString-generated CNV ratios to log transformed mRNA expression levels yielded no significant associations as illustrated in Figure 4.3.

In this section of our study, we developed a sensitive custom MLPA assay for copy number detection of 16 genes within the reverse cholesterol pathway. We were able to illustrate and verify that CNV is exceedingly rare for RCT genes as only 1 sample, in only 1 gene (*CETP*), had verified CNV. And as expected, there were no expression changes associated with the amount of RCT genes that we detected. As this number of CNV containing samples and genes were extremely low in our dataset, we were able to conclude that copy number variation is not common in RCT genes and therefore does not play a role in the dyslipidemia associated with antiretroviral use in HIV-1 infected individuals.



Normalized copy number ratios obtained by MLPA and NanoString are plotted on the x-axis while log transformed mRNA expression levels are plotted on the y-axis. If expression level data were available for multiple splice variants of the same gene, they were each plotted against their available CNV ratios, with the different variants represented by "_#" following the gene name. Data are shown for the 127 individuals for whom both CNV and expression data were available.

Figure 4.3: Expression Levels of RCT Genes are not associated with CNV Levels

During this analysis we were also able to use ratio thresholds, interquartile ranges and ratio distributions to determine gains and losses of copies. While we were not able to demonstrate this using our RCT genes, use of genes with known CNV allowed us to verify our methods used to determine that the RCT genes are present in 2 copies within the genome. As whole genome sequencing becomes more readily available and used to analyze commercially available DNA samples, such as those in the HapMap set, then read depth of those results will be available to determine exact copies within referent samples allowing more precise determination of copy number than use of ratio thresholds.

However, we also showed that even when you expect to be able to determine discrete copy numbers for a probe due to use of a referent of known copies, assay variation can hinder calling. A novice investigator must be wary of simply multiplying by their referent then rounding to obtain discrete calls as assays where the raw calls do not cluster around whole integer values will result in improper calling. Despite this, simple plotting of the raw calls are capable of illustrating the true CNV pattern that can then be called using k-means clustering.

5.0 AIM 2: TRANSCRIPTOME VARIATION

In order to identify genes that are differentially expressed in individuals exhibiting the atheroprotective versus atherogenic phenotype, we collected blood samples from the Pittsburgh center of the MACS during a 1 year period stretching from August 2010 to July 2011. Initially all participants who came to the clinic were sampled but after 50 individuals had blood collected for a second time, we opted to collect from only those who had not yet provided a sample. Blood was collected in PAXgene tubes (PreAnalytix/Qiagen, Valencia CA) to stabilize the intracellular RNA of each individual before extraction. RNA samples were taken to the Genomics and Proteomics Core Laboratories (GPCL) at the University of Pittsburgh for transcriptome analysis. Here they were quantified and processed into cDNA for use on the Illumina Human HT-12 v.4 whole-genome expression array. Because samples from participants who are HIV positive have excessive amounts of Beta globin mRNA, which would decrease the sensitivity to detect other transcripts, the GLOBINclear kit (Life Technologies, Carlsbad CA) was used to remove the majority of α- and β-globin transcripts from all samples prior to cDNA synthesis. The resulting image files were processed using Illumina's GenomeStudio V2011.1 software, where bead summary data was exported without background correction due to previous study findings that indicated such correction was disadvantageous. The raw data were then analyzed within the R statistical software and modules where it was pre-processed (transformed

and normalized) with the Lumi module, analyzed using the Limma module, and visualized by using the ggplot2 module.

Of the 528 runs performed during whole transcriptome analysis, 437 represent whole blood-derived RNA from unique individuals while the remainder were biological and technical replicates used to verify the proper functioning of the Illumina Human HT-12 assay. Through use of the R module ArrayQualityMetrics during the initial analysis of the transcriptomes, 53 of the 528 runs were identified to be outliers from the dataset. These individuals were removed from the raw expression set, which was then transformed and normalized for a second time using the R module Lumi. Correction for batch effects using the R module ComBat was performed next to account for any chip-to-chip variation despite the data being generated with chip sets ordered all at one time. Differentially expressed transcripts were determined by using an empirical Bayes analysis with a linear model framework on subsets of the expression set. These subsets were selected using clinical values collected during the patient's visit that included HIV status, CD8 counts, viral load, low density lipoprotein cholesterol, high density lipoprotein cholesterol, triglycerides and total cholesterol. For the lipoprotein clinical values, we opted to use the mean of 5 visits around the date of collection rather than the individual value from that date. This was due to the fluctuation of lipid values for some of the participants that could add more noise to the analysis. Once differential transcripts were generated with significant p-values, they were visualized using heatmaps that group individuals with similar expression patterns together. These transcripts were then analyzed for Gene Ontology to identify the relationships between the significant transcripts by illustrating the pathways they fall within.

5.1 ATHEROPROTECTIVE VS ATHEROGENIC

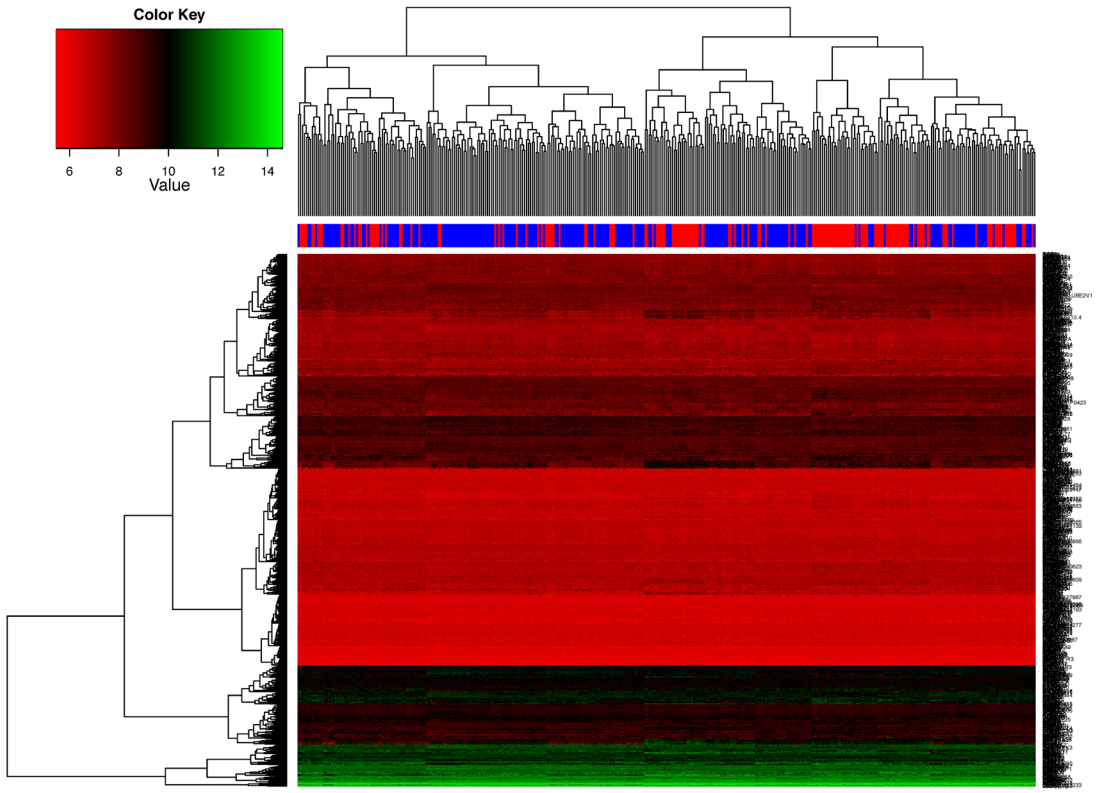
The first comparison we chose to make was that of individuals identified as having the atherogenic versus atheroprotective phenotypes. Those in the atheroprotective group (n=23) had LDL-C levels at or below 100 mg/dL **and** HDL-C levels at or above 60 mg/dL while those in the atherogenic (n=2) group had LDL-C levels at or above 160 mg/dL **and** HDL-C levels at or below 40 mg/dL. Due to the small sample size in the atherogenic group, the p-values for the top differentially expressed transcripts were not significant after correction for multiple tests.

5.2 HIV STATUS

We next analyzed HIV status of individuals with European ancestry in the expression set breaking them into groups of those infected with HIV-1 (n= 100; not including known LTNPs) and those negative for the virus (n=165). We ran this comparison only on the European ancestry as data from our lab had previously shown that lipid level variation is associated with biogeographical ancestry as well as with the known risk factors. For this reason we wanted to ensure that we could subset by ancestry and still be able to detect significant differences in gene expression before running our lipid comparisons on individuals of European ancestry that would be more likely to have an atherogenic phenotype[25]. This was accomplished as the HIV status comparison yielded 433 transcripts with p-values less than 0.05, the top twenty-five of which are listed in Table 5.1 and illustrated in Figure 5.1.

Table 5.1: Top 25 Transcripts for EA HIV Status Comparison

	TargetID	DEFINITION	logFC	Ave Expr	t	P.Value	adj.P.Val	B
1	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	-0.537	7.601	-12.449	2.05E-28	9.69E-24	52.254
2	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	-0.603	7.232	-12.111	2.98E-27	7.03E-23	49.744
3	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	-0.734	10.954	-11.152	5.30E-24	8.35E-20	42.711
4	CD8A	CD8a molecule (CD8A), transcript variant 1, mRNA.	-0.699	8.232	-11.035	1.30E-23	1.53E-19	41.871
5	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	-0.712	10.992	-10.567	4.55E-22	4.30E-18	38.521
6	MCOLN2	mucolipin 2 (MCOLN2), mRNA.	-0.468	8.659	-9.671	3.47E-19	2.73E-15	32.270
7	ARRDC4	arrestin domain containing 4 (ARRDC4), mRNA.	-0.278	7.136	-9.444	1.79E-18	1.21E-14	30.726
8	LAG3	lymphocyte-activation gene 3 (LAG3), mRNA.	-0.563	7.955	-9.269	6.21E-18	3.67E-14	29.552
q9	CD40LG	CD40 ligand (CD40LG), mRNA.	0.314	7.431	8.634	5.26E-16	2.76E-12	25.369
10	LOC644695	PREDICTED: hypothetical LOC644695 (LOC644695), mRNA.	-0.280	7.012	-8.358	3.42E-15	1.62E-11	23.603
11	HS.553068	BX103476 NCI_CGAP_Lu5 cDNA clone IMAGp998C053946, mRNA sequence	-0.266	7.175	-8.302	4.99E-15	2.14E-11	23.248
12	CST7	cystatin F (leukocystatin) (CST7), mRNA.	-0.350	10.577	-8.140	1.47E-14	5.78E-11	22.231
13	LOC197135	PREDICTED: hypothetical LOC197135, transcript variant 5 (LOC197135), mRNA.	-0.387	8.091	-8.098	1.94E-14	6.86E-11	21.970
14	FLJ33590	hypothetical protein FLJ33590 (FLJ33590), mRNA.	-0.262	7.205	-8.091	2.03E-14	6.86E-11	21.925
15	PATL2	PREDICTED: misc_RNA (PATL2), miscRNA.	-0.256	7.151	-7.758	1.79E-13	5.63E-10	19.878
16	TSHZ2	teashirt zinc finger homeobox 2 (TSHZ2), mRNA.	0.312	7.310	7.569	5.98E-13	1.76E-09	18.742
17	CACNA1I	calcium channel, voltage-dependent, T type, alpha 1I subunit (CACNA1I), transcript variant 2, mRNA.	0.327	7.607	7.558	6.39E-13	1.77E-09	18.679
18	FBLN7	fibulin 7 (FBLN7), mRNA.	0.205	6.725	7.416	1.57E-12	4.11E-09	17.835
19	VCAM1	vascular cell adhesion molecule 1 (VCAM1), transcript variant 1, mRNA.	-0.144	6.013	-7.289	3.46E-12	8.59E-09	17.090
20	PATL2	PREDICTED: misc_RNA (PATL2), miscRNA.	-0.378	8.536	-7.270	3.89E-12	8.93E-09	16.979
21	TNIP3	TNFAIP3 interacting protein 3 (TNIP3), mRNA.	-0.165	6.386	-7.265	4.02E-12	8.93E-09	16.949
22	TIGIT	T cell immunoreceptor with Ig and ITIM domains (TIGIT), mRNA.	-0.221	6.749	-7.259	4.16E-12	8.93E-09	16.916
23	LFNG	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase (LFNG), transcript variant 1, mRNA.	0.168	10.795	7.144	8.46E-12	1.74E-08	16.248
24	RCAN2	regulator of calcineurin 2 (RCAN2), mRNA.	-0.213	6.216	-7.120	9.77E-12	1.92E-08	16.113
25	AKR1C3	aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II) (AKR1C3), mRNA.	0.394	7.808	7.114	1.01E-11	1.92E-08	16.078



Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples are HIV positive (red) or HIV negative (blue). In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all 433 significant differentially expressed transcripts for the HIV status comparison of individuals with European Ancestry.

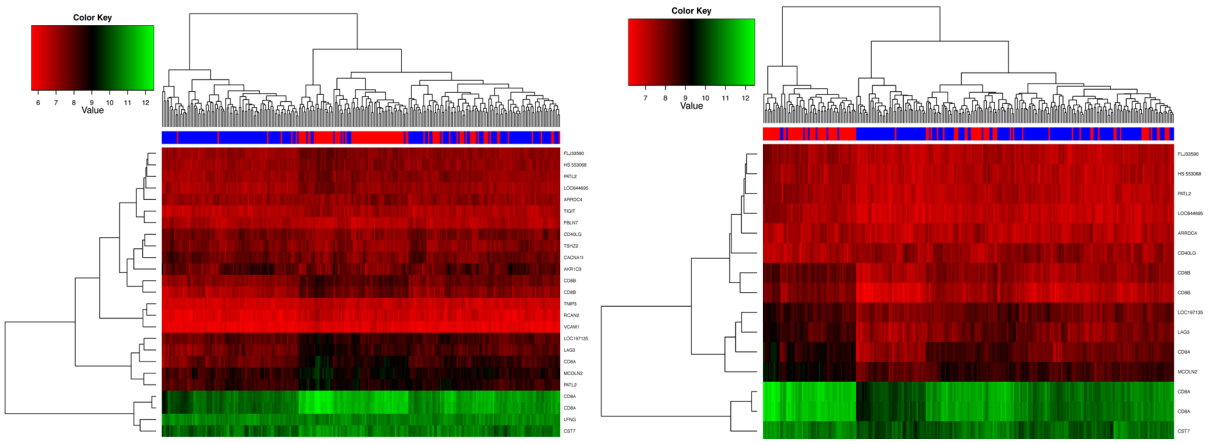
Figure 5.1: Heatmap of Top Transcripts for HIV Status Comparison

Within this table, multiple transcript variants of CD8 isoforms alpha and beta (n=5) were in the top ten most significant transcripts, showing increased expression in HIV positive individuals compared with seronegatives. Other immune associated transcripts were also identified including LAG3, CD40LG, VCAM1, and TIGIT. As the discrete blocks of altered expression are not easily visible in this comparison, we generated additional heatmaps with decreasing

amounts of the most significant genes to observe distinct clustering. The expression levels of the top 25 (p-value < $1.92x10^{-08}$) and top 14 transcripts (p-value < $5x10^{-10}$) are visualized in Figure 5.2's heatmaps.

A. Top 25 Transcripts

B. Top 14 Transcripts



Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples are HIV positive (red) or HIV negative (blue). In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains differentially expressed transcripts for the HIV status comparison of European Ancestry. Plot A displays the heatmap for the top 25 transcripts while Plot B displays the top 14 illustrating ($P < 5^{-10}$) that as a larger number of significant genes are included the clustering pattern of differential expression becomes clearer.

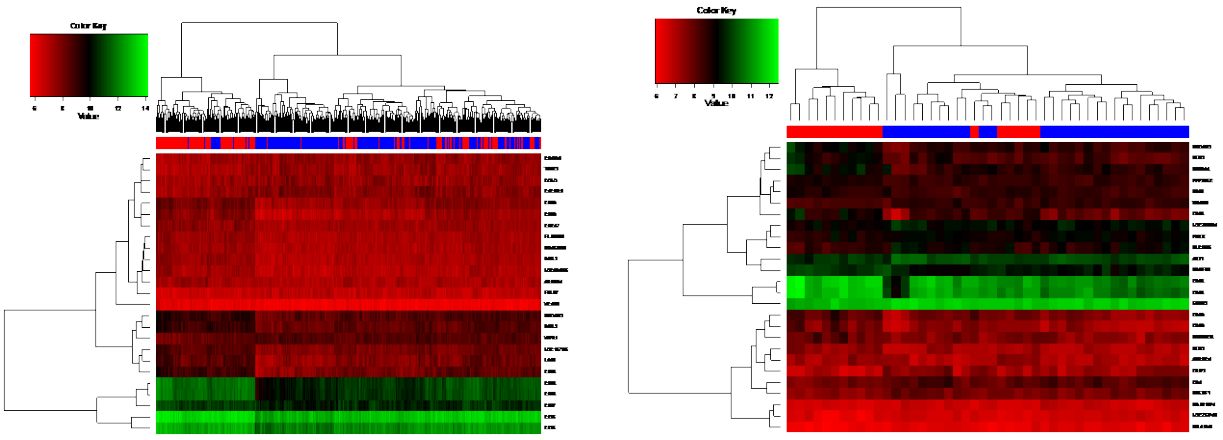
Figure 5.2: Heatmap of Most Significant Transcripts for HIV Status Comparison

The same analysis utilizing all of the ancestry groups together also contained a similar list of these top transcripts but in a slightly rearranged order (Figure 5.3). As the European ancestry individuals represent at least 68% (AEA n=52, EA n=269, AsEA n=5, & NA n=67) if not more of the dataset, their transcript differences, along with actual infection associated differences, likely account for this observation. This is supported in part by the top table for the African European ancestry samples (Table 5.3) as it is comprised of a mix of those top transcripts

observed in the European ancestry as well as new transcripts. There are also clear differences present in the heatmaps for these other ancestry group comparisons as well.

A. All Ancestries

B. AEA Ancestry



Columns represent individual samples while rows are the differentially expressed transcripts. The red and blue colored bar above the heatmap indicates which samples are HIV positive (red) or HIV negative (blue). In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains differentially expressed transcripts for the HIV status comparison of European Ancestry. Plot A displays the heatmap for the top 25 transcripts for all ancestries while Plot B displays the top 25 transcripts for the AEA ancestry. All Ancestries combined: HIV+: n=156, HIV-: n=223; AEA Ancestry: HIV+: n=17, HIV-: n=29 Note: As the AEA ancestry contained only 2 transcripts with $p<0.05$, the top table and heatmap for it were generated using the top 25 transcripts to illustrate that the EA samples compare more closely to the entire data set.

Figure 5.3: Top Transcript Heatmaps for HIV Status Comparison in Additional Ancestries

Table 5.2: Top 25 Transcripts for All Ancestry of HIV Status Comparison

	TargetID	DEFINITION	logFC	Ave Expr	t	P.Value	adj.P.Val	B
1	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	-0.50	7.63	-13.91	6.65E-36	3.14E-31	69.60
2	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	-0.55	7.26	-13.77	2.52E-35	5.96E-31	68.32
3	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	-0.71	10.97	-13.26	2.73E-33	4.29E-29	63.83
4	CD8A	CD8a molecule (CD8A), transcript variant 1, mRNA.	-0.67	8.25	-12.99	3.24E-32	3.83E-28	61.45
5	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	-0.70	11.02	-12.64	7.16E-31	6.76E-27	58.48
6	MCOLN2	mucolipin 2 (MCOLN2), mRNA.	-0.48	8.68	-12.17	5.03E-29	3.96E-25	54.40
7	ARRDC4	arrestin domain containing 4 (ARRDC4), mRNA.	-0.26	7.12	-10.59	3.65E-23	2.46E-19	41.44
8	LAG3	lymphocyte-activation gene 3 (LAG3), mRNA.	-0.53	7.97	-10.50	7.82E-23	4.62E-19	40.71
9	CD40LG	CD40 ligand (CD40LG), mRNA.	0.31	7.43	10.29	4.27E-22	2.24E-18	39.08
10	LOC197135	PREDICTED: hypothetical LOC197135, transcript variant 5 (LOC197135), mRNA.	-0.39	8.08	-10.06	2.79E-21	1.32E-17	37.27
11	PATL2	PREDICTED: misc_RNA (PATL2), miscRNA.	-0.26	7.15	-9.81	2.10E-20	9.00E-17	35.34
12	LOC644695	PREDICTED: hypothetical LOC644695 (LOC644695), mRNA.	-0.27	7.02	-9.61	9.86E-20	3.88E-16	33.85
13	PATL2	PREDICTED: misc_RNA (PATL2), miscRNA.	-0.40	8.52	-9.46	3.17E-19	1.15E-15	32.73
14	CST7	cystatin F (leukocystatin) (CST7), mRNA.	-0.34	10.59	-9.35	7.64E-19	2.58E-15	31.89
15	FBLN7	fibulin 7 (FBLN7), mRNA.	0.21	6.72	9.14	3.81E-18	1.20E-14	30.35
16	HS.553068	BX103476 NCI_CGAP_Lu5 cDNA clone IMAGp998C053946, mRNA sequence	-0.23	7.17	-9.00	1.05E-17	3.09E-14	29.38
17	CACNA1I	calcium channel, voltage-dependent, T type, alpha 1I subunit (CACNA1I), transcript variant 2, mRNA.	0.32	7.62	8.99	1.12E-17	3.10E-14	29.31
18	FLJ33590	hypothetical protein FLJ33590 (FLJ33590), mRNA.	-0.24	7.21	-8.99	1.18E-17	3.11E-14	29.26
19	VCAM1	vascular cell adhesion molecule 1 (VCAM1), transcript variant 1, mRNA.	-0.15	6.02	-8.75	7.03E-17	1.75E-13	27.55
20	CCL5	chemokine (C-C motif) ligand 5 (CCL5), mRNA.	-0.35	12.30	-8.64	1.52E-16	3.59E-13	26.81
21	TSHZ2	teashirt zinc finger homeobox 2 (TSHZ2), mRNA.	0.30	7.32	8.60	2.07E-16	4.65E-13	26.51
22	CDCA7	cell division cycle associated 7 (CDCA7), transcript variant 1, mRNA.	-0.24	7.39	-8.57	2.63E-16	5.65E-13	26.28
23	COLQ	collagen-like tail subunit (single strand of homotrimer) of asymmetric acetylcholinesterase (COLQ), transcript variant VIII, mRNA.	0.27	7.43	8.47	5.42E-16	1.11E-12	25.59
24	CCL5	chemokine (C-C motif) ligand 5 (CCL5), mRNA.	-0.30	13.18	-8.26	2.45E-15	4.81E-12	24.15
25	VIPR1	vasoactive intestinal peptide receptor 1 (VIPR1), mRNA.	0.23	8.45	8.16	4.74E-15	8.95E-12	23.51

Table 5.3: Top 25 Transcripts for AEA Ancestry of HIV Status Comparison

	TargetID	DEFINITION	logFC	Ave Expr	t	P.Value	adj.P.Val	B
1	CD8A	CD8a molecule (CD8A), transcript variant 1, mRNA.	-0.82	8.29	-5.66	7.35E-07	0.023	5.13
2	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	-0.56	7.69	-5.57	9.92E-07	0.023	4.88
3	LOC727748	PREDICTED: similar to scratch homolog 1, zinc finger protein (Drosophila) (LOC727748), mRNA.	0.24	6.50	5.16	4.30E-06	0.068	3.68
4	MCOLN2	mucolipin 2 (MCOLN2), mRNA.	-0.57	8.66	-4.96	8.34E-06	0.079	3.14
5	PDXK	pyridoxal (pyridoxine, vitamin B6) kinase (PDXK), mRNA.	0.30	9.05	4.95	8.89E-06	0.079	3.08
6	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	-0.74	11.01	-4.91	1.01E-05	0.079	2.98
7	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	-0.74	11.08	-4.82	1.35E-05	0.088	2.74
8	AKT1	v-akt murine thymoma viral oncogene homolog 1 (AKT1), transcript variant 2, mRNA.	0.31	10.02	4.78	1.55E-05	0.088	2.63
9	ARRDC4	arrestin domain containing 4 (ARRDC4), mRNA.	-0.33	7.09	-4.76	1.69E-05	0.088	2.55
10	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	-0.53	7.33	-4.73	1.86E-05	0.088	2.48
11	HIST3H2A	histone cluster 3, H2a (HIST3H2A), mRNA.	-0.40	7.63	-4.63	2.65E-05	0.114	2.18
12	EGLN2	egl nine homolog 2 (C. elegans) (EGLN2), transcript variant 3, mRNA.	0.21	11.69	4.55	3.42E-05	0.122	1.97
13	HS.561874	cDNA clone IMAGE:4794367	0.19	6.62	4.48	4.31E-05	0.122	1.78
14	PATL2	PREDICTED: misc_RNA (PATL2), miscRNA.	-0.33	7.11	-4.48	4.38E-05	0.122	1.77
15	PATL2	PREDICTED: misc_RNA (PATL2), miscRNA.	-0.56	8.45	-4.48	4.40E-05	0.122	1.77
16	SREBF1	sterol regulatory element binding transcription factor 1 (SREBF1), transcript variant 1, mRNA.	0.29	7.67	4.45	4.76E-05	0.122	1.70
17	PPP2R5C	protein phosphatase 2, regulatory subunit B', gamma isoform, transcript variant 4, mRNA.	-0.23	8.59	-4.44	4.94E-05	0.122	1.67
18	CD4	CD4 molecule (CD4), mRNA.	0.35	8.03	4.43	5.17E-05	0.122	1.63
19	SLC16A5	solute carrier family 16, member 5 (monocarboxylic acid transporter 6) (SLC16A5), mRNA.	0.44	8.86	4.43	5.19E-05	0.122	1.63
20	KIAA1143	KIAA1143 (KIAA1143), mRNA.	0.19	6.29	4.41	5.39E-05	0.122	1.60
21	SH2D1A	SH2 domain protein 1A, Duncan's disease (lymphoproliferative syndrome), mRNA.	-0.54	8.78	-4.41	5.43E-05	0.122	1.59
22	CRIP2	cysteine-rich protein 2 (CRIP2), mRNA.	0.46	7.20	4.40	5.69E-05	0.122	1.55
23	WDR68	WD repeat domain 68 (WDR68), transcript variant 2, mRNA.	0.22	8.48	4.32	7.28E-05	0.149	1.35
24	RPA3	replication protein A3, 14kDa (RPA3), mRNA.	-0.22	8.55	-4.31	7.63E-05	0.150	1.31
25	LOC730994	PREDICTED: similar to NACHT, leucine rich repeat and PYD (pyrin domain) containing 1, transcript variant 1 (LOC730994), mRNA.	0.29	9.28	4.28	8.40E-05	0.157	1.23

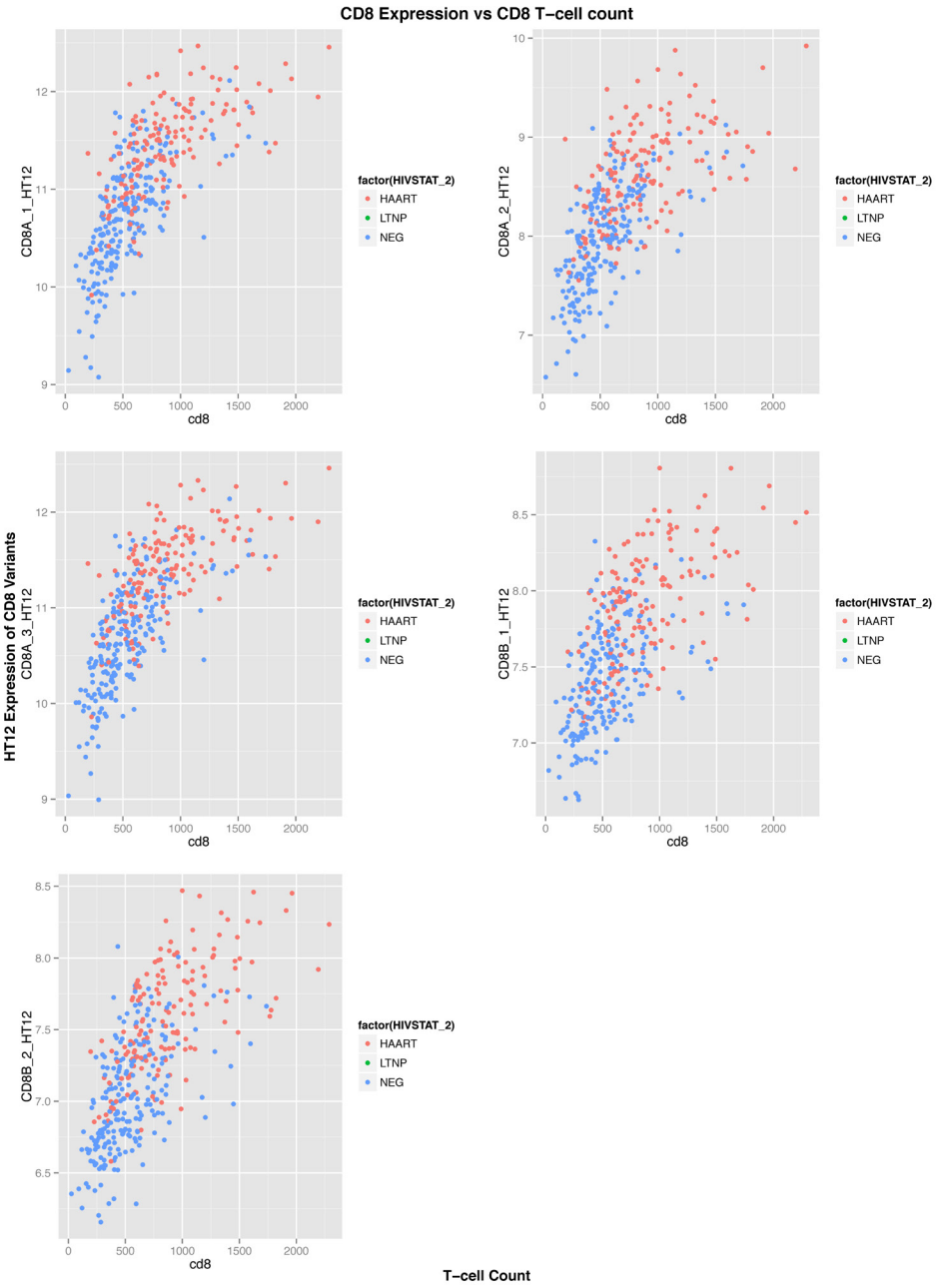


Figure 5.4: Expression of CD8 is Higher among HIV-positive Individuals

When we plotted the CD8 counts for each individual against the expression values for each of the CD8 variants, the HIV-positive individuals had higher CD8 counts and expression levels (Figure 5.4). There was not a clear distinct break between the HIV status groups but the majority

of individuals for each did cluster together. Because CD8 was observed to be among the top transcripts in all sets of comparisons, it was a candidate for further investigation.

5.3 CD8 COUNT

As the top rated transcripts for HIV status comparisons included multiple variants of CD8, and we had clinical measurements of CD8 T-cell counts taken for each individual at the same visit as their PAXgene blood sample collection, we next performed a comparison of CD8 counts. This was done by subsetting the European ancestry expression set used above into two smaller sets containing only those that were HIV negative and those that were HIV positive but not LTNP. For each subset we calculated the cutoffs for the top [HIV(Q4 = , n=39) & NEG Q4 = , n=55)] and bottom quartiles [HIV(Q1 = , n=39) & NEG Q1 = , n=55)] of CD8 counts and then subsetted those expression sets so that they contained only those extreme ends of the CD8 ranges. As expected, the individuals with higher CD8 counts had higher expression of CD8 transcript variants independent of HIV status. Overall, the HIV-negative individuals compared for CD8 had 1752 significant transcripts in 89 pathways including RNA catabolism, mRNA metabolism, ribosome biogenesis, wound healing, and translational termination. Conversely, the HIV positive individuals had 681 significant transcripts in 16 pathways including those involved in regulating the immune system, ribosome biogenesis, and RNA processing metabolism. The top transcripts with $p<0.0005$ are illustrated in Figure 5.5 for HIV positive and Figure 5.6 for HIV negative subsets while the top ten transcripts for both groups are listed in Table 5.4. The majority of top transcripts shown in these figures and table had increased expression for the top quartile group with distinct blocks of expression observed in the heatmaps.

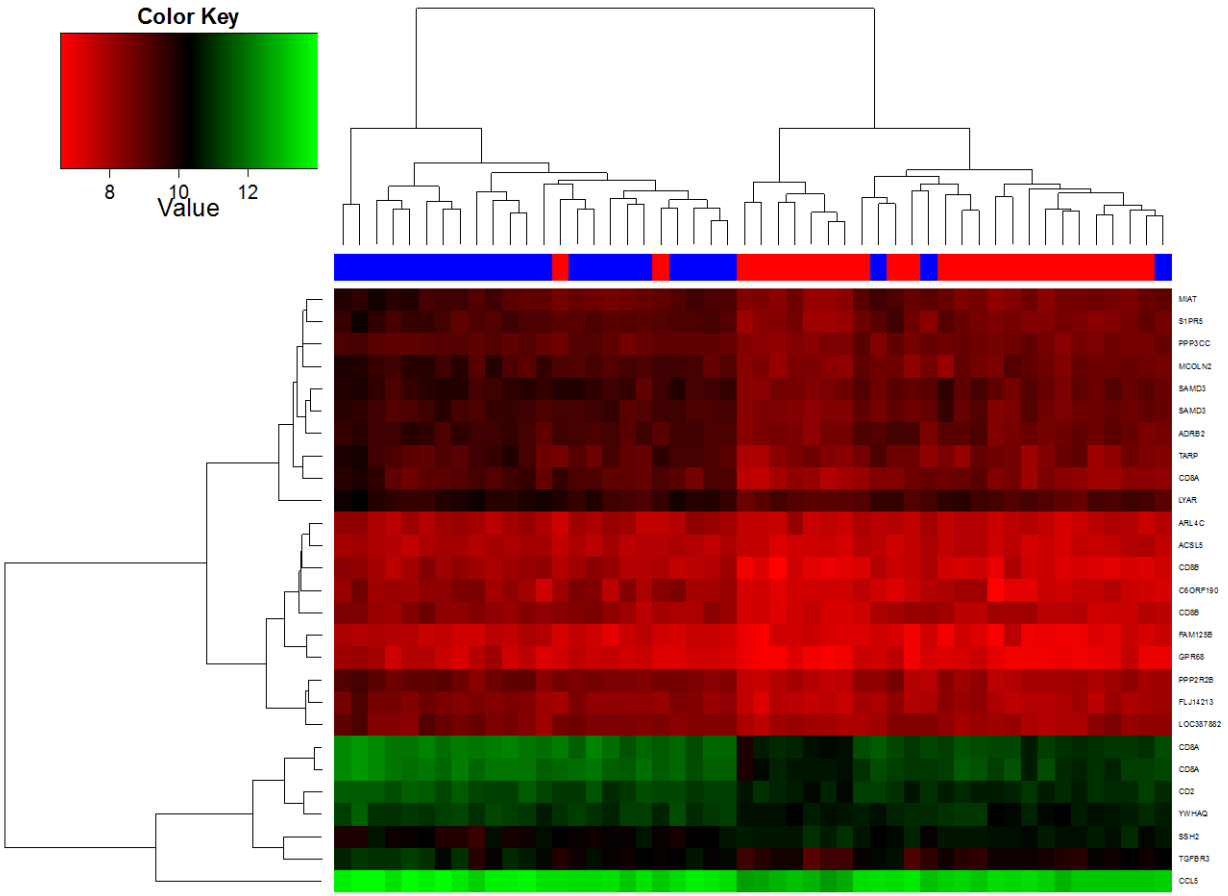
Table 5.4: CD8 Quartile Comparison for HIV Status Subsets

HIV- positive Samples CD8 Quartile Comparison

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	0.70	7.58	9.20	9.70E-13	4.58E-08	17.88
2	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	0.87	11.44	8.57	9.78E-12	2.31E-07	15.83
3	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	0.83	11.39	8.21	3.75E-11	5.91E-07	14.63
4	CD2	CD2 molecule (CD2), mRNA.	0.48	11.05	7.55	4.62E-10	5.45E-06	12.38
5	FLJ14213	protor-2 (FLJ14213), mRNA.	0.56	8.08	7.17	1.91E-09	1.81E-05	11.10
6	PPP2R2B	protein phosphatase 2 (formerly 2A), regulatory subunit B, beta isoform	0.65	8.27	7.03	3.33E-09	2.35E-05	10.60
7	MIAT	myocardial infarction associated transcript (non-protein coding) (MIAT), non-coding	0.62	8.89	7.00	3.72E-09	2.35E-05	10.50
8	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	0.57	7.90	6.98	3.98E-09	2.35E-05	10.44
9	MCOLN2	mucolipin 2 (MCOLN2), mRNA.	0.65	8.95	6.87	5.96E-09	3.13E-05	10.07
10	S1PR5	sphingosine-1-phosphate receptor 5 (S1PR5), mRNA.	0.63	8.88	6.78	8.41E-09	3.97E-05	9.76

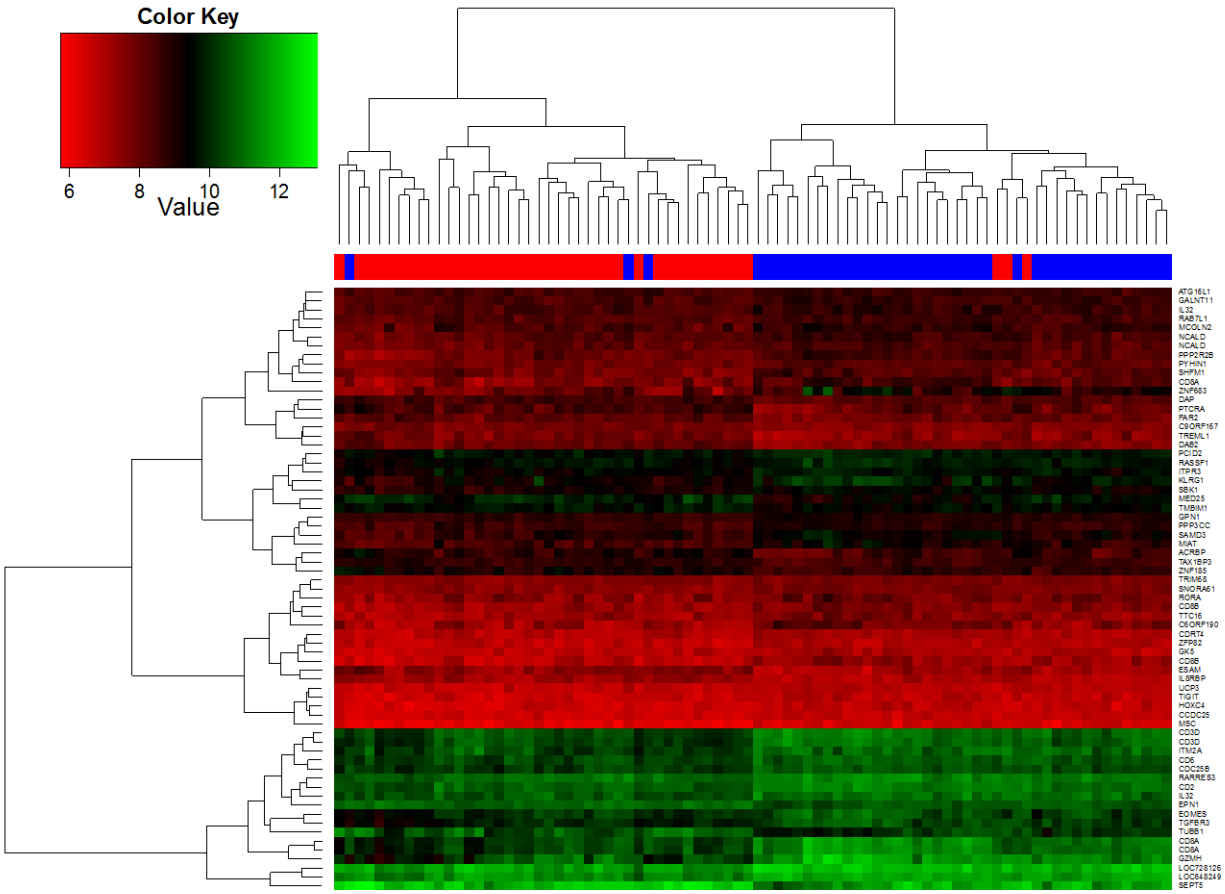
HIV-negative Samples CD8 Quartile Comparison

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	1.04	10.69	12.11	1.74E-20	8.23E-16	34.28
2	CD8A	CD8a molecule (CD8A), transcript variant 2, mRNA.	1.00	10.66	11.39	4.98E-19	1.18E-14	31.29
3	CD8A	CD8a molecule (CD8A), transcript variant 1, mRNA.	0.85	7.95	9.90	5.54E-16	8.72E-12	24.96
4	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	0.59	7.00	9.07	2.80E-14	3.31E-10	21.39
5	CD8B	CD8b molecule (CD8B), transcript variant 5, mRNA.	0.48	7.38	8.37	7.83E-13	7.39E-09	18.34
6	ZNF683	zinc finger protein 683 (ZNF683), transcript variant 2, mRNA.	1.11	8.44	7.46	5.70E-11	4.49E-07	14.40
7	GALNT11	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-	0.33	8.49	7.05	3.80E-10	2.56E-06	12.65
8	CD3D	CD3d molecule, delta (CD3-TCR complex) (CD3D), transcript variant 2, mRNA.	0.55	10.69	6.83	1.03E-09	5.95E-06	11.73
9	CDC25B	cell division cycle 25 homolog B (S. pombe) (CDC25B), transcript variant 2, mRNA.	0.36	10.37	6.81	1.13E-09	5.95E-06	11.64
10	MIAT	myocardial infarction associated transcript (non-protein coding) (MIAT), non-coding	0.58	8.75	6.75	1.48E-09	6.99E-06	11.40



Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the bottom (red) or top (blue) quartiles. In the upper left corner is the legend that indicates the range of the $\log_2$ transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains only the top transcripts ($P < 0.0005$) out of a total of 681 significant differentially expressed transcripts for CD8 comparison in HIV-positive samples

Figure 5.5: CD8 Quartile Comparison for HIV-positive Samples



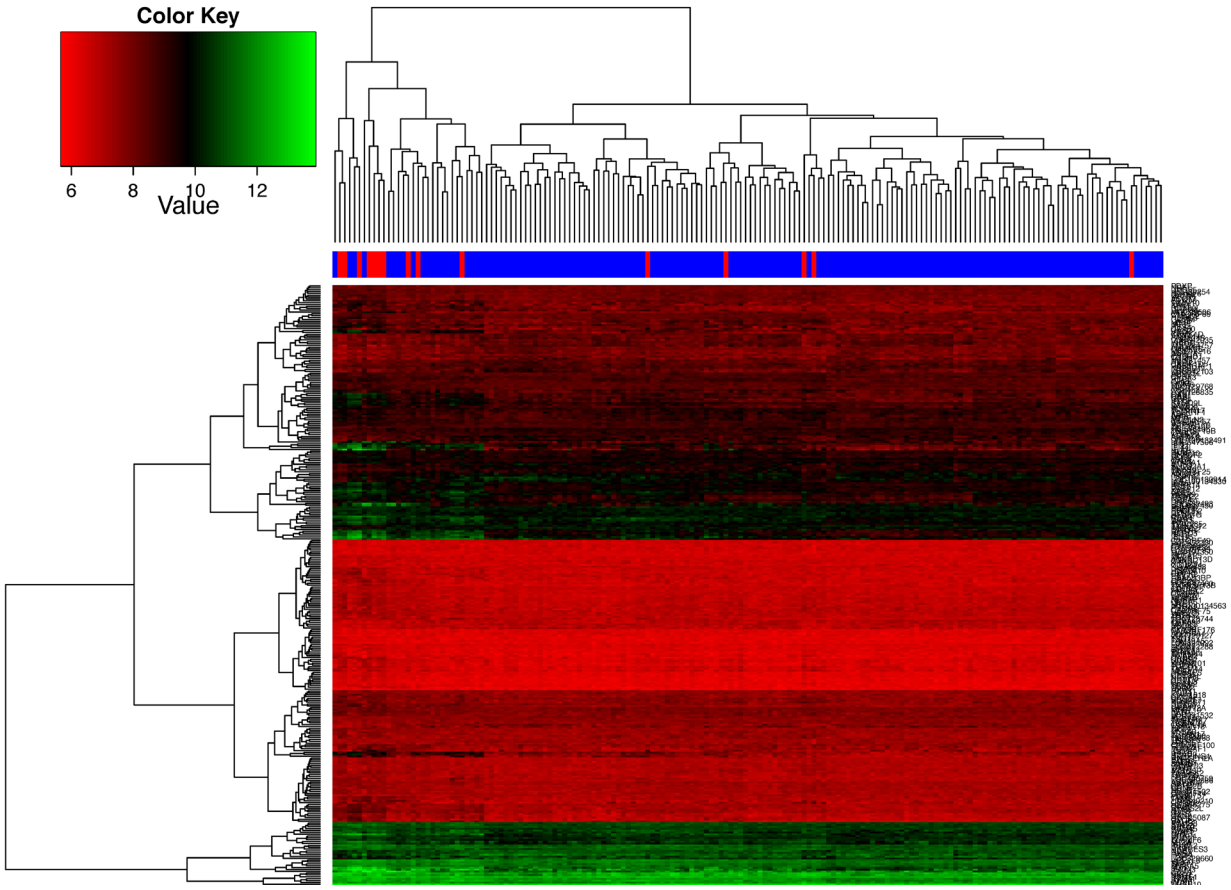
Columns represent individual samples while rows are the differentially expressed transcripts. The red and blue colored bar above the heatmap indicates which samples fall in the bottom (red) or top (blue) quartiles. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains only the top transcripts ($P < 0.0005$) out of a total of 1752 significant differentially expressed transcripts for the CD8 comparison in HIV-negative samples.

Figure 5.6: CD8 Quartile Comparison for HIV-negative Samples

5.4 VIRAL LOAD

The MACS currently calculates viral load utilizing a Roche kit with an ultra-sensitive method which is molecular-based therefore providing fast and accurate results. The lower limit of detection for this method is 50 copies/mL so participants with values below this level were

considered to have undetectable levels of the viral RNA. The groups used in this comparison were generated from HIV positive individuals who were considered to have undetectable viral load levels (40 copies/mL or below) and those who had high viral load levels (200 copies/mL or above). This comparison yielded 421 significant transcripts in 107 pathways that primarily involved immune response, viral control and cell cycle regulation. The heatmap for this comparison is shown in Figure 5.7.



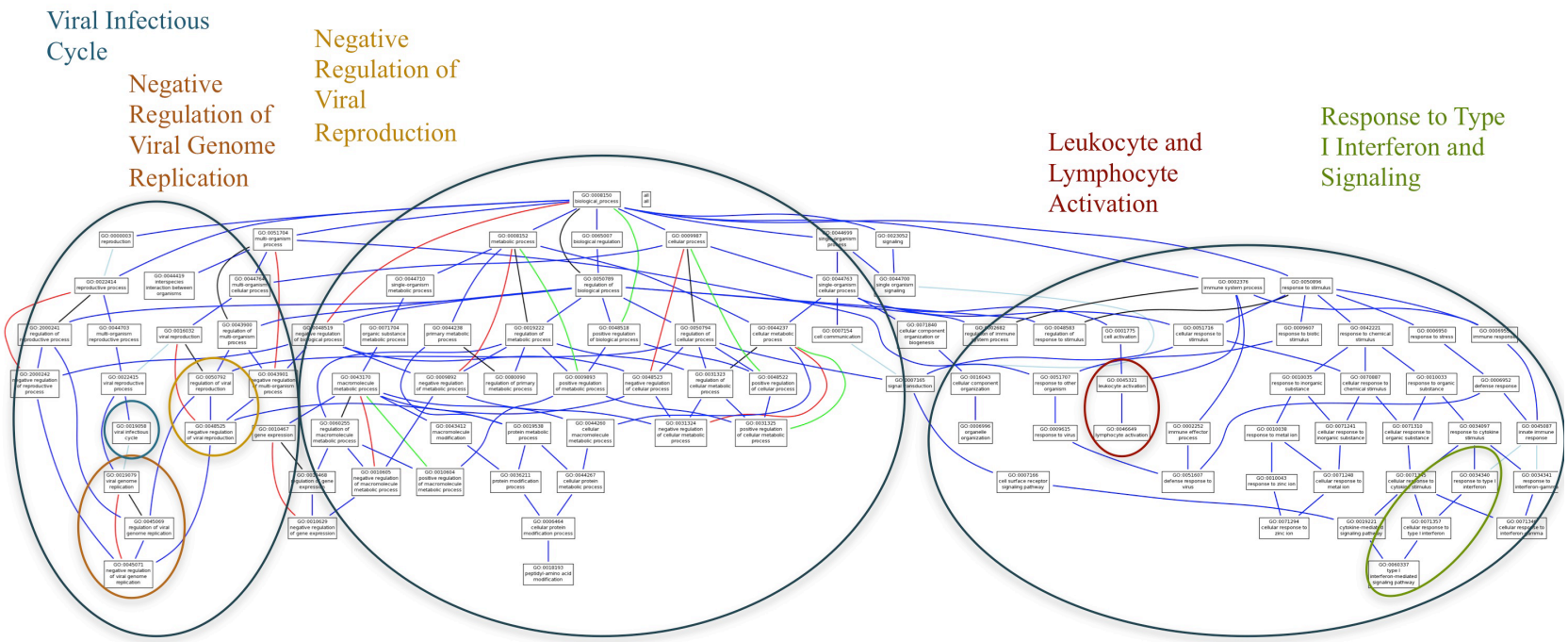
Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the low (red) or high (blue) viral load group. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all 421 significant transcripts ($P < 0.05$) differentially expressed transcripts for viral load comparison in HIV-positive samples.

Figure 5.7: Viral Load Comparison

Because the number of individuals with low viral loads greatly outnumbered those with high levels, we wanted to determine if more similar sample sizes could lead to a tighter visual representation of the differential expression. To do this, we took the individuals with high viral load from the first comparison along with the first 15 samples with undetectable virus. This comparison resulted 10 significant transcripts (Table 5.5) in 49 pathways (Figure 5.8) including many of those observed in the first comparison as well as pathways associated with negative regulation of gene expression. It was also observed that the heatmap (Figure 5.9) groupings resulted in the viral load groups clustering together more efficiently.

Table 5.5: Viral Comparison with Smaller Group Size

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	LOC728835	PREDICTED: similar to cytokine, transcript variant 3 (LOC728835),	-0.78	8.62	-5.96	9.97E-07	0.04445	5.08
2	AGPAT9	1-acylglycerol-3-phosphate O-acyltransferase 9 (AGPAT9), mRNA.	0.72	8.90	5.46	4.41E-06	0.04445	3.85
3	PDPK1	3-phosphoinositide dependent protein kinase-1 (PDPK1), transcript variant 1,	0.45	7.66	5.43	4.87E-06	0.04445	3.76
4	POP4	processing of precursor 4, ribonuclease P/MRP subunit (S. cerevisiae) (POP4),	-0.30	7.52	-5.34	6.30E-06	0.04445	3.55
5	IL6R	interleukin 6 receptor (IL6R), transcript variant 1, mRNA.	0.62	7.82	5.33	6.58E-06	0.04445	3.51
6	RXRA	retinoid X receptor, alpha (RXRA), mRNA.	0.44	10.82	5.32	6.76E-06	0.04445	3.49
7	EIF5A	eukaryotic translation initiation factor 5A (EIF5A), mRNA.	-0.51	9.82	-5.29	7.38E-06	0.04445	3.41
8	PADI4	peptidyl arginine deiminase, type IV (PADI4), mRNA.	0.80	11.10	5.25	8.36E-06	0.04445	3.31
9	SPATS2L	spermatogenesis associated, serine-rich 2-like (SPATS2L), transcript variant 2,	-0.66	7.28	-5.22	9.04E-06	0.04445	3.24
10	RBPJ	recombination signal binding protein for immunoglobulin kappa J region (RBPJ),	0.47	8.26	5.21	9.41E-06	0.04445	3.21



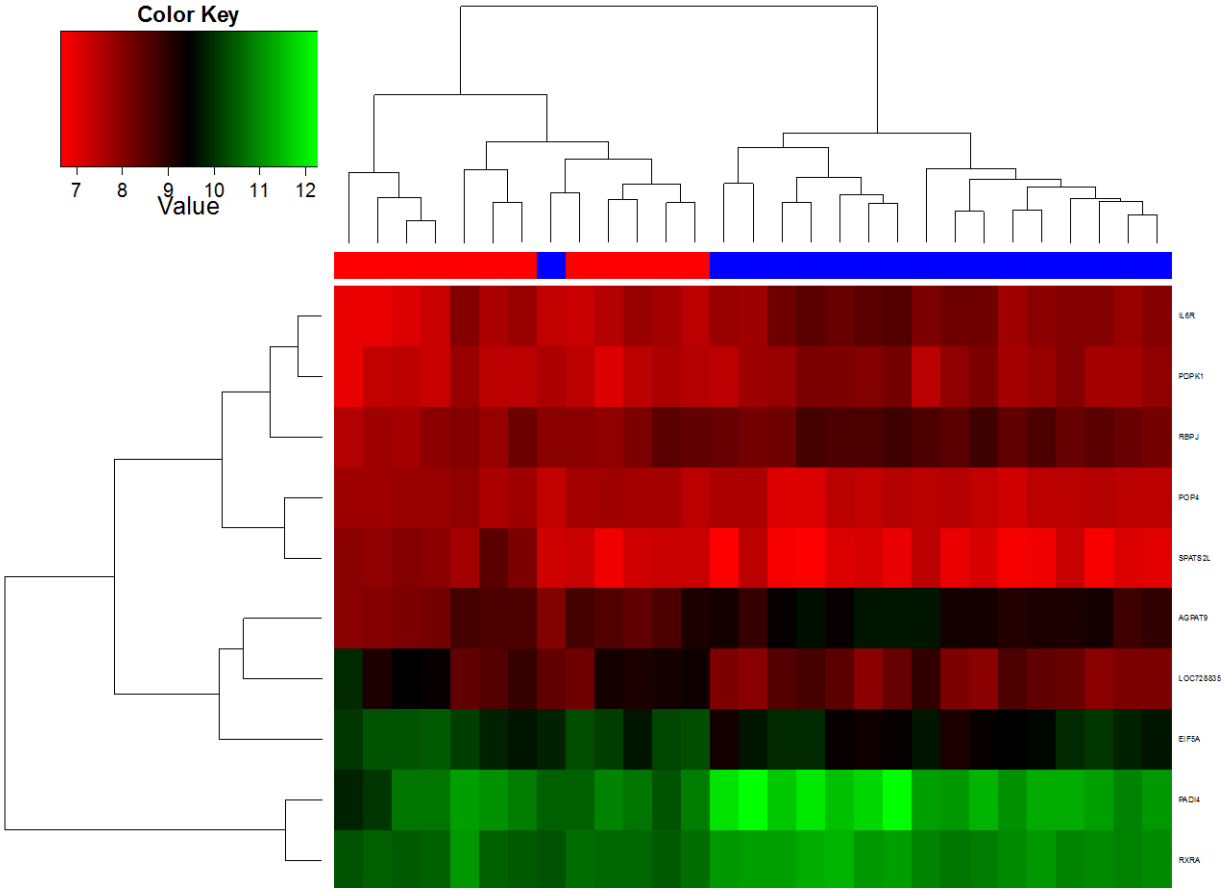
Viral Replication Control
(Negative regulators of genome replication and viral)

Metabolic Processes
(Regulation of Biological, Cellular and Metabolic Processes resulting in peptidyl-amino acid modification)

Immune Response
(Cell Activation, Type I Interferon Activity, Response to Virus, etc)

Individual Pathways are depicted by the boxes while the broad groups of related pathways are circled.

Figure 5.8: Pathway Analysis for Viral Load Comparison



Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the low (red) or high (blue) viral load group. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all 10 significant transcripts ($P < 0.05$) differentially expressed transcripts for the smaller group size viral load comparison in HIV-positive samples.

Figure 5.9: Viral Load Comparison with Smaller Group Size

5.5 EXTENDED LIPID COMPARISONS

As our studies of HIV-related phenotypes showed that microarray analysis of PAXgene-derived whole blood RNA could indeed detect transcriptome-level differences, we began to compare a

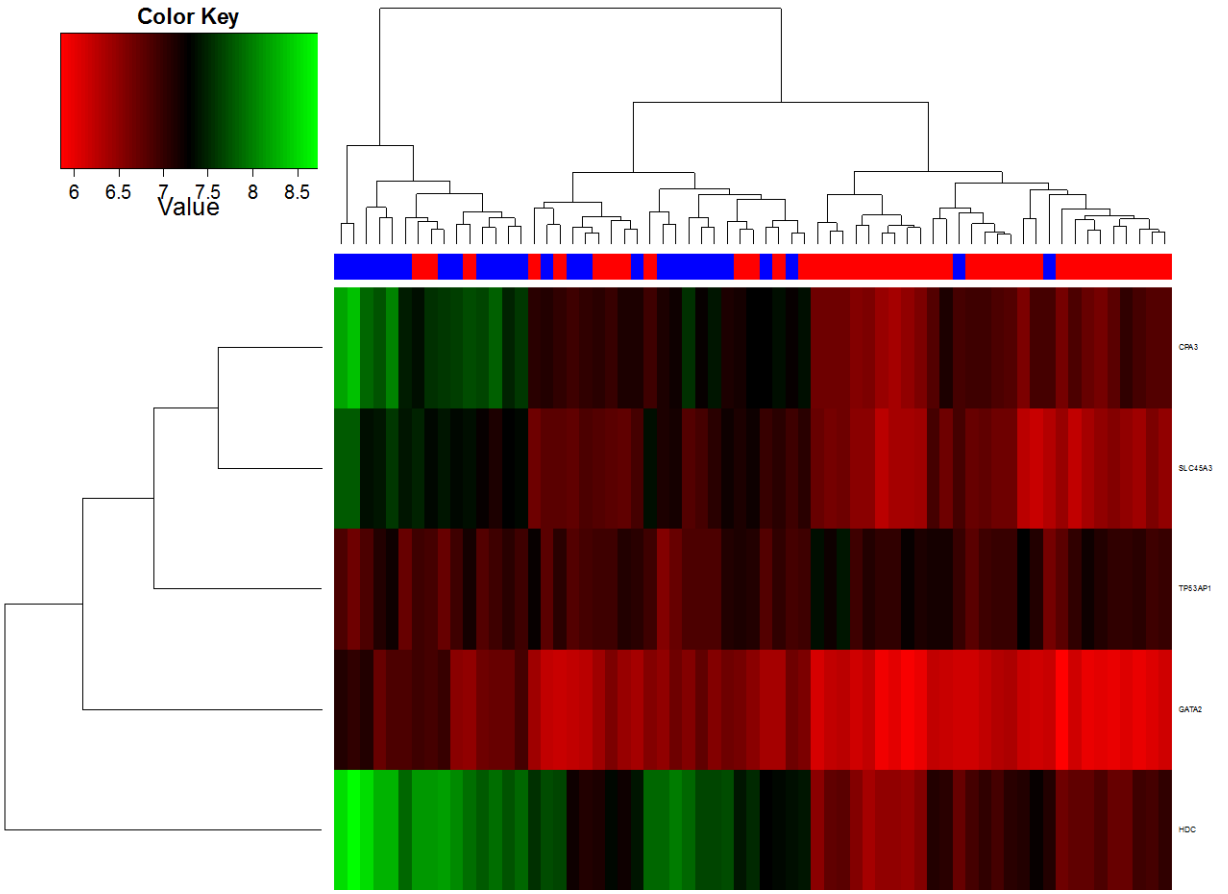
broader range of lipid levels beyond the extreme atherogenic and atheroprotective phenotypes. Our initial comparisons were to contrast the clinically relevant extremes for HDL-C and LDL-C while comparing the top and bottom quartiles for triglycerides and total cholesterol. Out of these, only HDL-C and triglycerides had differentially expressed transcripts that were significant.

5.5.1 High Density Lipoprotein

For the HDL comparison we used all of the samples available regardless of geographical ancestry and selected individuals with HDL-C levels that were considered high [>60mg/mL (n=26)] or low [<40mg/mL (n=39)] based on the NCEP Adult Treatment Panel III. This yielded 5 significant transcripts that are listed in Table 5.6 and shown in the heatmap in Figure 5.10. Increased expression was generally seen for individuals with higher HDL levels, although these transcripts do not appear to be directly related to cholesterol metabolism. When this dataset was further divided into groups based on HIV status, the comparisons in the HIV-positive individuals yielded no significant transcripts while the HIV negative had one transcript (HDC) with a p-value of 0.0058. Ancestry groups were not compared for HDL as the sample sizes became too small for comparison after subgrouping by Lipid levels and HIV status.

Table 5.6: Low vs High HDL-C Top Differentially Expressed Transcripts

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	HDC	histidine decarboxylase (HDC), mRNA.	-1.48	7.44	-11.69	6.53E-10	3.08E-05	7.85
2	GATA2	GATA binding protein 2 (GATA2), mRNA.	-0.75	6.47	-9.53	1.65E-08	0.0004	6.34
3	SLC45A3	solute carrier family 45, member 3 (SLC45A3), mRNA.	-0.76	6.88	-8.46	9.94E-08	0.0016	5.38
4	CPA3	carboxypeptidase A3 (mast cell) (CPA3), mRNA.	-0.81	7.07	-8.25	1.43E-07	0.0017	5.17
5	KLF2	Kruppel-like factor 2 (lung) (KLF2), mRNA.	-0.51	12.78	-6.18	7.41E-06	0.0700	2.71



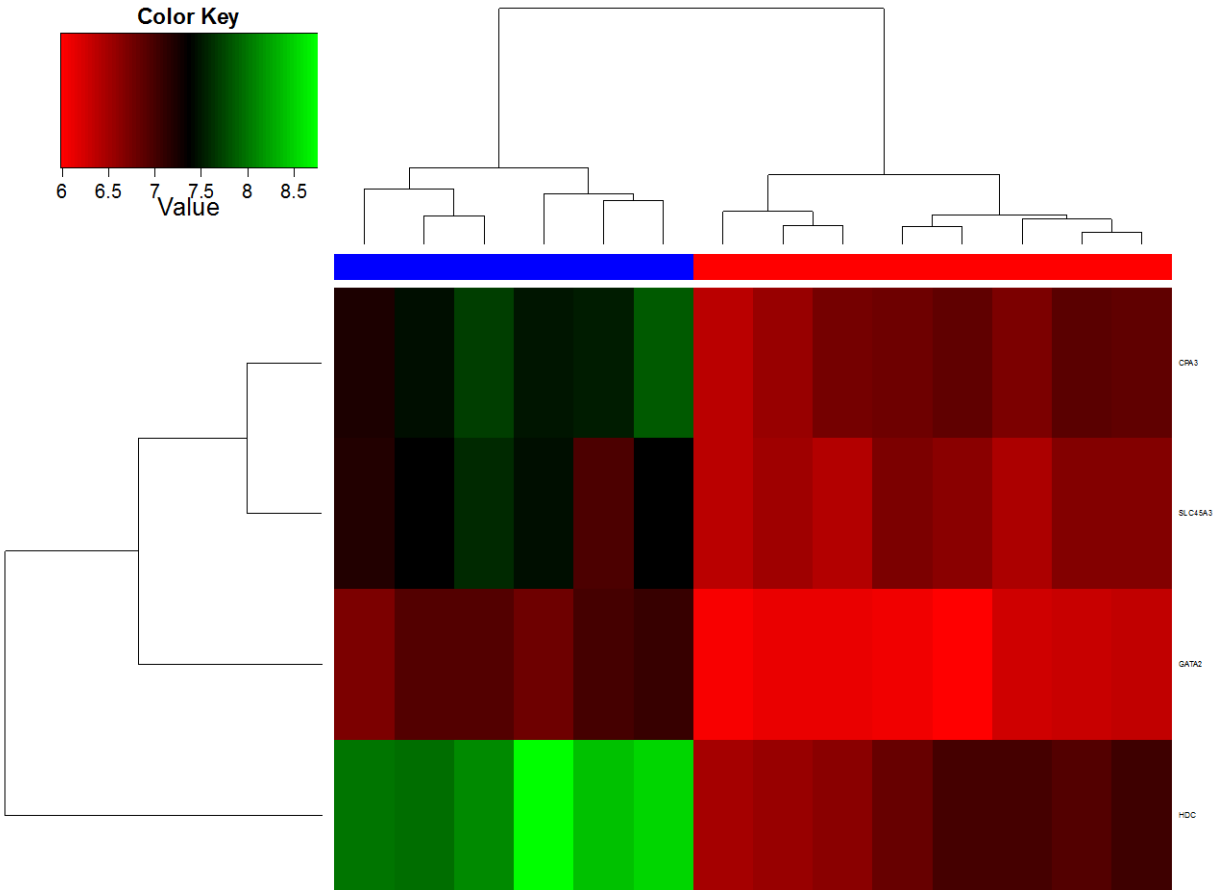
Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the low (red) or high (blue) HDL group. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all 5 significant transcripts ($P < 0.05$) out of 27 total differentially expressed transcripts for the HDL-C comparison in all samples

Figure 5.10: High vs Low HDL-C Comparison

Based on the initial findings for our HDL-C comparison, we decided to use more stringent criteria for HDL-C levels, using a lower cutoff of <30mg/mL and an upper cutoff of >70mg/mL in order to identify those transcripts with the greatest association with HDL-C levels. As expected, we observed 4 transcripts that were also present in the top ten transcripts from the first HDL-C comparison but this time the differences were more significant. This can be seen in Table 5.7 and is clearly depicted in the heatmap in Figure 5.11. Gene Ontology pathway analysis was then performed on these 4 transcripts to reveal 56 pathways. These pathways can be broken down into smaller groupings of metabolic processes, protein dimerization, protein targeting to the ER, mRNA catabolism, ATP synthesis, and Control of Viral Replication. The relationship between these transcripts and the HDL-C phenotype is not immediately clear.

Table 5.7: Stringent Low vs High HDL-C Top Differentially Expressed Transcripts

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	HDC	histidine decarboxylase (HDC), mRNA.	-1.48	7.44	-11.69	6.53E-10	3.08E-05	7.85
2	GATA2	GATA binding protein 2 (GATA2), mRNA.	-0.75	6.47	-9.53	1.65E-08	0.00039	6.34
3	SLC45A3	solute carrier family 45, member 3 (SLC45A3), mRNA.	-0.76	6.88	-8.46	9.94E-08	0.00157	5.38
4	CPA3	carboxypeptidase A3 (mast cell) (CPA3), mRNA.	-0.81	7.07	-8.25	1.43E-07	0.00169	5.17



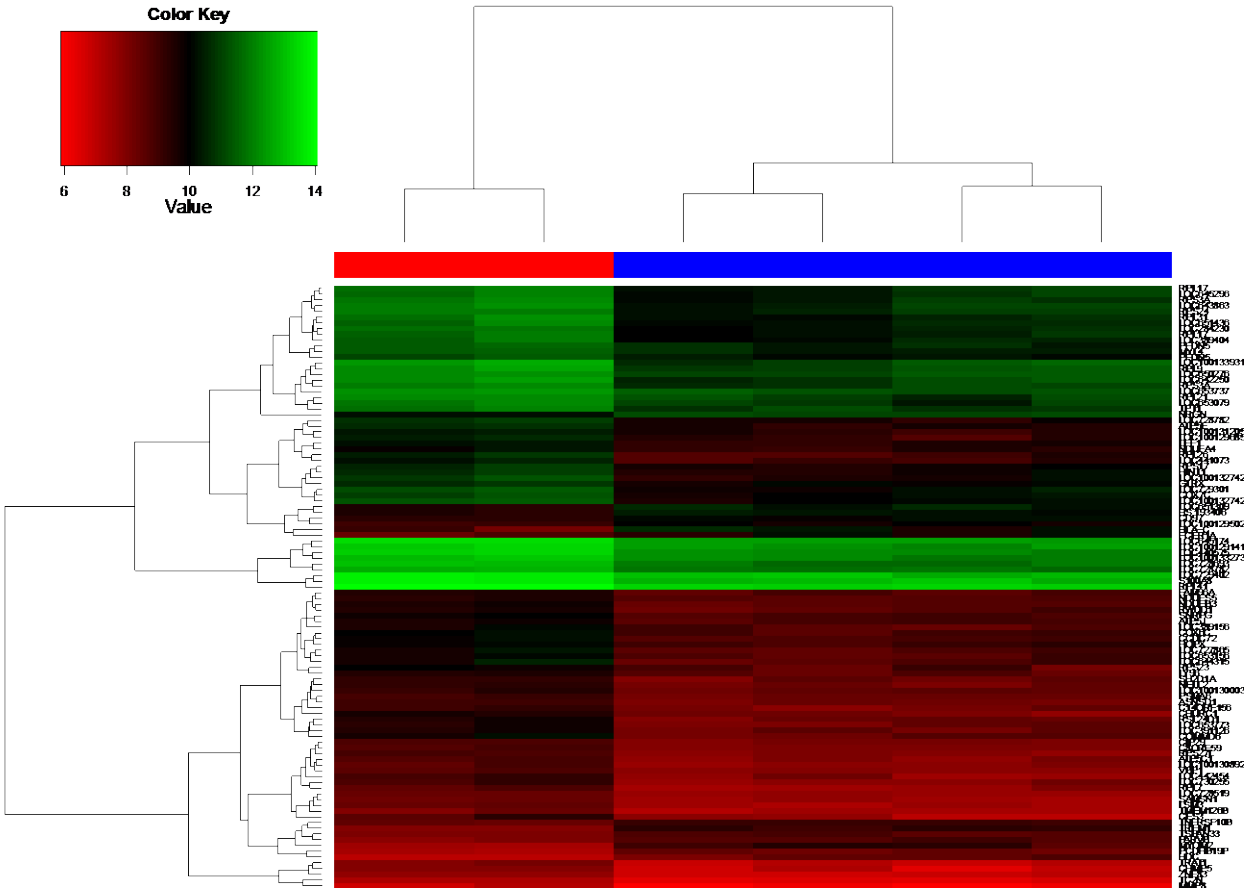
Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the very low (red) or very high (blue) HDL group. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all 4 significant transcripts ($P < 0.05$) differentially expressed transcripts for the HDL-C comparison in all samples.

Figure 5.11: Heatmap for Stringent Low vs High HDL-C comparison (All Serogroups)

The stringent HDL-C dataset was then subdivided by HIV status. Comparisons resulted in HIV-positive samples having no significant transcripts while the HIV-negative samples differed by 105 transcripts between their two HDL groups. The top 25 genes from the HIV-negative comparison are shown in Table 5.8 while the heatmap in Figure 5.12 illustrates distinct differences for the extreme HDL-C groups of this comparison. Once again, the top transcripts observed contained histidine decarboxylase but lacked all of the previous transcripts (*GATA2*, *SLC45A3*, & *CPA3*) identified in the comparison containing all samples.

Table 5.8: Stringent Low vs High HDL-C Top Differentially Expressed Transcripts for HIV-negative Samples

	Transcript	DEFINITION	logFC	Ave Expr	t	P.Value	adj. P.Val	B
1	TREM1	triggering receptor expressed on myeloid cells 1 (TREM1), mRNA.	-1.32	8.79	-9.93	1.98E-07	0.0068	6.71
2	CHURC1	churchill domain containing 1 (CHURC1), mRNA.	1.50	8.55	9.61	2.86E-07	0.0068	6.43
3	LOC440575	PREDICTED: hypothetical LOC440575 (LOC440575), mRNA.	1.20	12.53	9.12	5.19E-07	0.0071	5.98
4	CCDC72	coiled-coil domain containing 72 (CCDC72), mRNA.	1.20	9.24	8.89	6.99E-07	0.0071	5.75
5	LOC728693	PREDICTED: misc_RNA (LOC728693), miscRNA.	1.16	12.19	8.56	1.06E-06	0.0071	5.42
6	RPL31	ribosomal protein L31 (RPL31), transcript variant 1, mRNA.	1.60	10.91	8.50	1.15E-06	0.0071	5.36
7	LOC645174	PREDICTED: misc_RNA (LOC645174), miscRNA.	0.98	12.79	8.49	1.17E-06	0.0071	5.34
8	LOC728782	PREDICTED: similar to ribosomal protein L21 (LOC728782), mRNA.	1.04	12.08	8.43	1.27E-06	0.0071	5.28
9	ATP5E	ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit (ATP5E), nuclear gene encoding mitochondrial protein, mRNA.	1.21	9.89	8.34	1.42E-06	0.0071	5.19
10	LOC100132742	PREDICTED: hypothetical protein LOC100132742, transcript variant 2 (LOC100132742), mRNA.	1.43	10.42	8.23	1.64E-06	0.0071	5.08
11	RSL24D1	ribosomal L24 domain containing 1 (RSL24D1), mRNA.	1.35	8.70	8.17	1.79E-06	0.0071	5.00
12	MYOM2	myomesin (M-protein) 2, 165kDa (MYOM2), mRNA.	-1.59	8.53	-8.09	1.98E-06	0.0071	4.92
13	COMMD6	COMM domain containing 6 (COMMD6), transcript variant 1, mRNA.	1.51	8.91	8.05	2.11E-06	0.0071	4.87
14	LOC284230	PREDICTED: similar to mCG7611 (LOC284230), mRNA.	1.42	10.85	7.93	2.46E-06	0.0071	4.75
15	HDC	histidine decarboxylase (HDC), mRNA.	-1.38	7.95	-7.93	2.48E-06	0.0071	4.74
16	RPS27L	ribosomal protein S27-like (RPS27L), mRNA.	0.98	8.20	7.91	2.53E-06	0.0071	4.72
17	NDUFB3	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 3, 12kDa (NDUFB3), mRNA.	1.09	8.93	7.91	2.54E-06	0.0071	4.72
18	PCDHB19P	protocadherin beta 19 pseudogene (PCDHB19P), non-coding RNA.	-1.02	8.01	-7.87	2.70E-06	0.0071	4.67
19	COX6C	cytochrome c oxidase subunit VIc (COX6C), mRNA.	1.23	9.30	7.78	3.04E-06	0.0076	4.57
20	RPL26	ribosomal protein L26 (RPL26), mRNA.	1.63	9.51	7.62	3.80E-06	0.0086	4.39
21	NRGN	neurogranin (protein kinase C substrate, RC3) (NRGN), mRNA.	-0.92	10.80	-7.62	3.82E-06	0.0086	4.39
22	NELL2	NEL-like 2 (chicken) (NELL2), mRNA.	1.06	8.58	7.57	4.12E-06	0.0086	4.32
23	SNRPG	small nuclear ribonucleoprotein polypeptide G (SNRPG), mRNA.	0.99	9.20	7.56	4.17E-06	0.0086	4.31
24	LOC653737	PREDICTED: hypothetical LOC653737 (LOC653737), mRNA.	1.01	11.65	7.45	4.83E-06	0.0088	4.19
25	LOC391126	PREDICTED: misc_RNA (LOC391126), miscRNA.	1.27	8.79	7.45	4.85E-06	0.0088	4.19



Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the very low (red) or very high (blue) HDL group. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all 4 significant transcripts ($P < 0.05$) differentially expressed transcripts for the HDL-C comparison in HIV-negative samples.

Figure 5.12: Heatmap for Stringent Low vs High HDL-C comparison (HIV-negative)

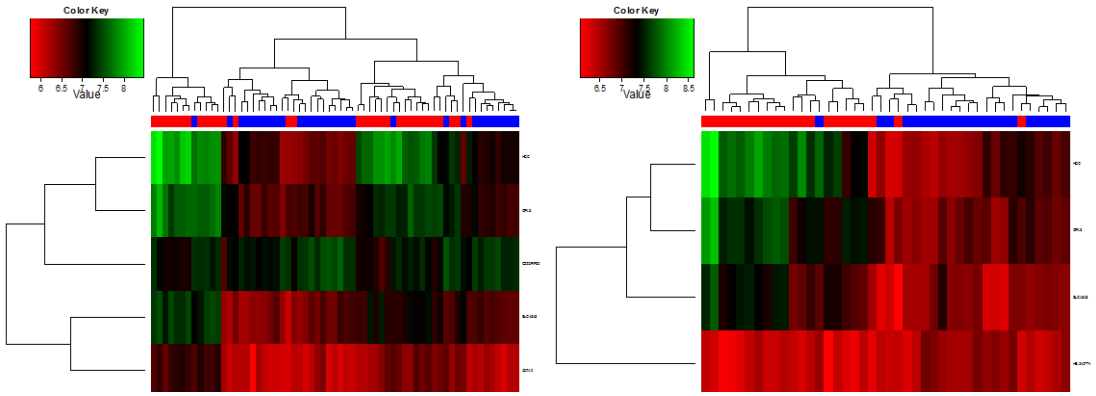
5.5.2 Triglycerides

Triglyceride measurements obtained on the same visit day as the collection of the PAXgene blood sample were used for our next set of comparisons. Groups were defined on the basis of the top and bottom quartile triglyceride values collected. Unlike the HDL comparison, we were able

to get significant differentially expressed transcripts for both the complete set of HIV-positive samples and its subset of individuals with European ancestry. The comparisons for both yielded exactly the same top 3 differentially expressed transcripts (Table 5.9) and yielded a cleaner cluster pattern for the European subset (Figure 5.13). In total, 5 transcripts in 49 pathways were observed to be differentially expressed for the triglyceride groups derived from all the HIV-positive samples while 4 transcripts in 34 pathways were observed for those of European Ancestry. Both comparisons included pathways that involved protein degradation, protein transport/targeting to the ER, gene expression/translation, RNA metabolism/biosynthesis and viral genome expression. The African European Ancestry subset was too small to yield any significant transcripts.

A. All HIV+ Samples

B. EA HIV+ Samples



Columns represent individual samples while rows are the differentially expressed transcript. The red and blue colored bar above the heatmap indicates which samples fall in the bottom quartile (red) or top quartile (blue) for triglyceride levels. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all the top significant transcripts differentially expressed transcripts for the triglyceride comparison in all HIV-positive samples (Plot A) and the European subset of HIV-positive samples (Plot B)

Figure 5.13: Triglyceride Comparison Heatmaps

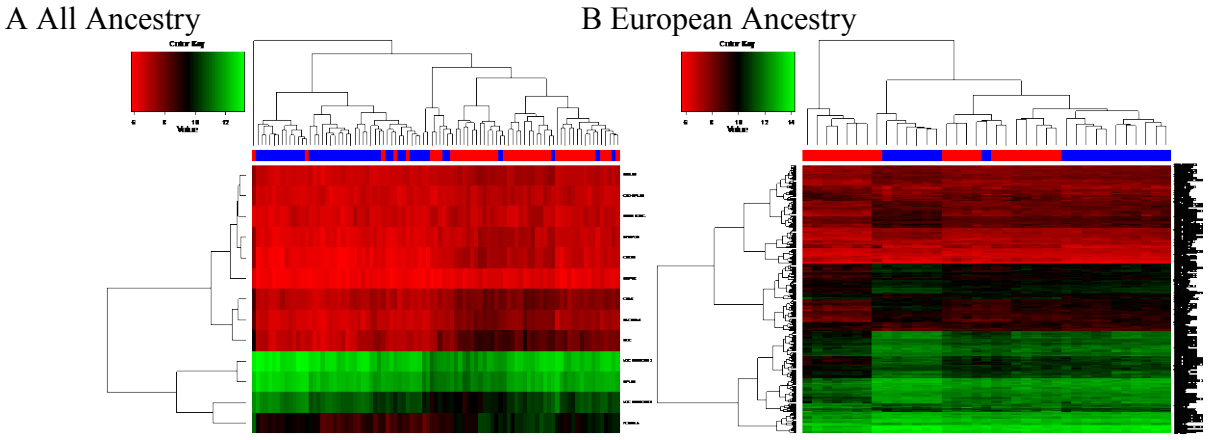
Table 5.9: Triglyceride Comparison Top Transcripts for HIV-positive Samples

All HIV-positive Samples

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	CPA3	carboxypeptidase A3 (mast cell) (CPA3), mRNA.	-57.30	173.09	-7.20	5.09E-10	2.40E-05	9.13
2	HDC	histidine decarboxylase (HDC), mRNA.	-102.84	194.65	-6.54	7.97E-09	0.00019	7.25
3	SLC45A3	solute carrier family 45, member 3 (SLC45A3), mRNA.	-38.68	147.08	-6.34	1.85E-08	0.00029	6.66
4	GATA2	GATA binding protein 2 (GATA2), mRNA.	-20.42	118.49	-6.08	5.44E-08	0.00064	5.91
5	RPL41	ribosomal protein L41 (RPL41), transcript variant 2, mRNA.	4185.24	14336.94	4.95	4.85E-06	0.04579	2.73

European Ancestry HIV-positive Samples

	Transcript	DEFINITION	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	CPA3	carboxypeptidase A3 (mast cell) (CPA3), mRNA.	-0.62	7.13	-6.77	1.89E-08	0.00089	6.62
2	HDC	histidine decarboxylase (HDC), mRNA.	-0.83	7.22	-6.10	1.94E-07	0.00457	5.04
3	SLC45A3	solute carrier family 45, member 3 (SLC45A3), mRNA.	-0.51	6.86	-5.47	1.72E-06	0.02420	3.52
4	HS.242774	cs14c01.y1 Human Retinal pigment epithelium/choroid cDNA	0.23	6.43	5.42	2.05E-06	0.02420	3.40



Columns represent individual samples while rows are the differentially expressed transcripts. The red and blue colored bar above the heatmap indicates which samples fall in the bottom quartile (red) or top quartile (blue) for triglyceride levels. In the upper left corner is the legend that indicates the range of the log_2 transformed expression values that correspond to the color bar. Green indicates higher expression while red indicates lower. This graphical comparison contains all the top significant transcripts differentially expressed transcripts for the triglyceride comparison in all HIV-negative samples (Plot A) and the European subset of HIV-negative samples (Plot B)

Figure 5.14: Heatmap of Top Transcripts for Triglyceride Comparisons in HIV-negative Samples

Table 5.10: Triglyceride Comparison Top Transcripts for HIV-negative Samples

All HIV-negative Samples

	Transcript	DEFINITION	logFC	Ave Expr	t	P.Value	adj.P.Val	B
1	HDC	Homo sapiens histidine decarboxylase (HDC), mRNA.	-1.16	7.55	-10.42	1.95E-17	9.20E-13	25.66
2	CPA3	Homo sapiens carboxypeptidase A3 (mast cell) (CPA3), mRNA.	-0.75	7.32	-10.27	3.97E-17	9.38E-13	25.08
3	GATA2	Homo sapiens GATA binding protein 2 (GATA2), mRNA.	-0.62	6.54	-8.92	3.23E-14	4.21E-10	19.59
4	SLC45A3	Homo sapiens solute carrier family 45, member 3 (SLC45A3), mRNA.	-0.73	7.05	-8.90	3.57E-14	4.21E-10	19.51
5	MS4A2	Homo sapiens membrane-spanning 4-domains, subfamily A, member 2.	-0.36	6.84	-7.34	7.02E-11	6.26E-07	13.17

European Ancestry HIV-negative Samples

	Transcript	DEFINITION	logFC	Ave Expr	t	P.Value	adj.P.Val	B
1	HDC	Homo sapiens histidine decarboxylase (HDC), mRNA.	-1.42	7.40	-9.86	1.87E-12	8.84E-08	16.87
2	CPA3	Homo sapiens carboxypeptidase A3 (mast cell) (CPA3), mRNA.	-0.96	7.26	-9.37	8.09E-12	1.91E-07	15.64
3	SLC45A3	Homo sapiens solute carrier family 45, member 3 (SLC45A3), mRNA.	-0.97	6.97	-9.19	1.42E-11	2.24E-07	15.16
4	GATA2	Homo sapiens GATA binding protein 2 (GATA2), mRNA.	-0.80	6.52	-8.73	5.80E-11	6.85E-07	13.97
5	MS4A2	Homo sapiens membrane-spanning 4-domains, subfamily A, member 2.	-0.50	6.80	-7.47	3.28E-09	3.10E-05	10.47

In order to identify if the Triglyceride-associated expression differences observed for the HIV-positive samples reflected any interactions with the virus, comparisons were also performed for the HIV-negative individuals. Analysis of HIV-negative individuals, and the subset of those with European ancestry, yielded 15 and 263 differentially expressed transcripts (Figure 5.14), respectively (once again, the sample size of the African Ancestry group was not large enough to yield any significant results.). The top 5 transcripts listed in Table 5.10, identical for both ancestry groupings, also include 4 transcripts observed for the HIV-positive comparison (Figure 5.14). This would suggest that these 4 genes (*HDC*, *CPA3*, *GATA2* & *SLC45A3*) have an association with Triglyceride levels independent of HIV status.

5.5.3 Differential Expression Verification

To verify that the genes identified as top transcripts during these comparisons were indeed differentially expressed, we performed TaqMan Real-Time expression assays for the following transcripts; CD8A, CD8B, ABCA1, HDC and CPA3. In order to conserve RNA, we selected 16 individuals that fell in the highest or lowest expression bracket for each transcript and performed reverse transcription to generate cDNA. Our initial run verified expression differences for most of the transcripts differentially expressed for the HT-12. This was not a completely linear relationship but those samples that had higher expression for the HT-12 remained high while those that were low remained low (Figure 5.15). Transcripts that did not perfectly match for verification were those for CD8. As there are multiple variants of CD8A and CD8B within the HT-12, the two probes we ordered for TaqMan expression analysis are not likely to pick up each variant equally and therefore will not have a direct linear relationship. This could be overcome by ordering multiple TaqMan probes for each but as the sequence of those probes are not available, we are not at liberty to run multiple CD8 transcripts. This however did not completely impede our ability to identify that the majority of samples within each assay had similar patterns of up and down regulation of expression.

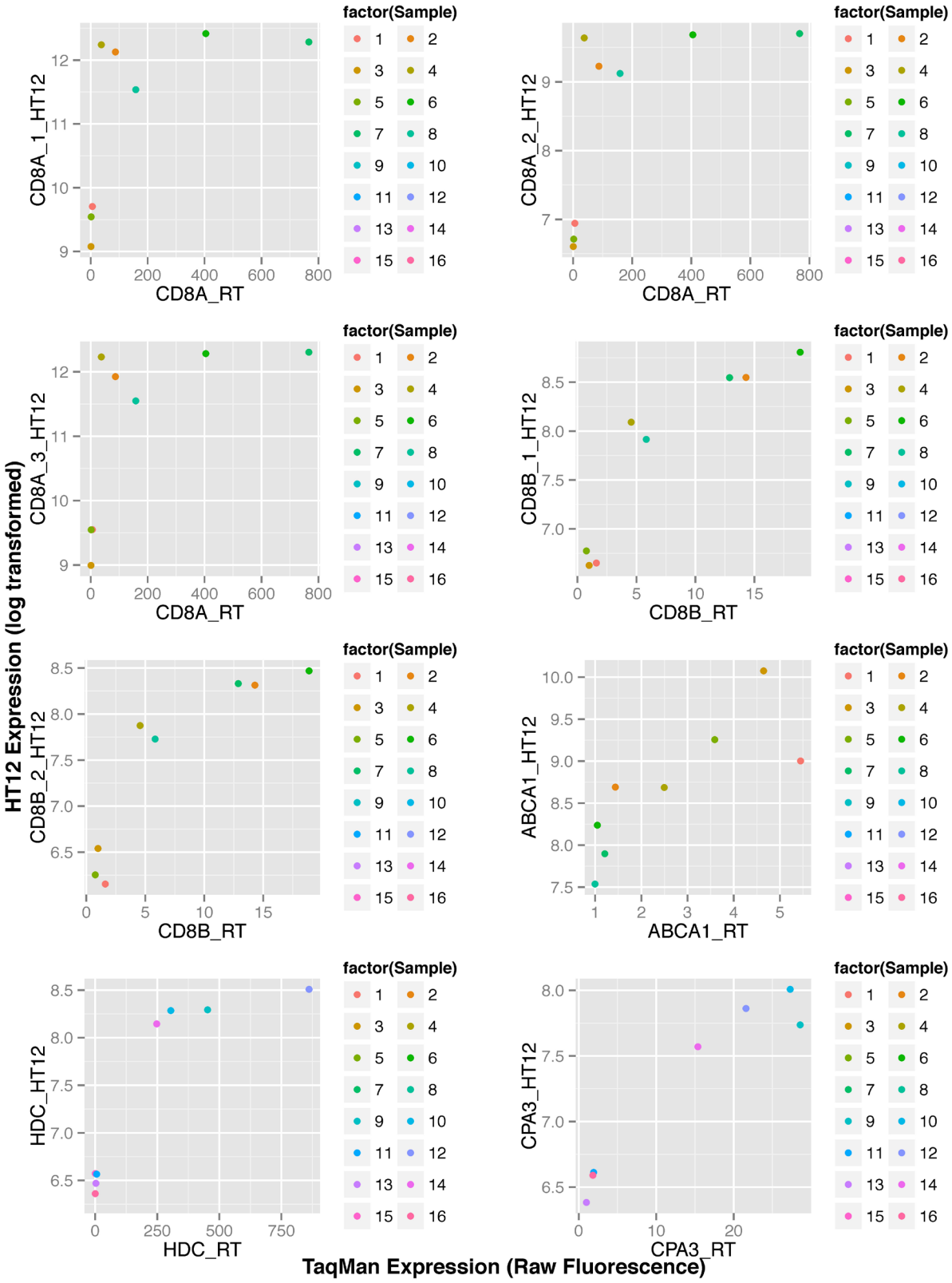


Figure 5.15: Expression Confirmation Analysis of Differentially Expressed Genes

Our initial transcriptome analysis illustrates that we are able to identify significant differentially expressed genes in RNA extracted from whole blood collected into PAXgene tubes. The number of transcripts varied according to the criteria used for sample selection: a large number of significantly differentially expressed transcripts were detected when HIV status was used as our outcome variable, but much smaller numbers, and in some cases no significantly differentially expressed transcripts, were detected when lipid-related phenotypes were used in our comparisons. The reason for this is not yet known but might be associated with data used to subgroup individuals. Despite the small number of transcripts, once pathway analysis was performed, all of the transcripts identified fell into pathways relevant to the comparison being performed. For instance, viral load comparisons yielded numerous pathways associated to regulation of viral replication, response to interferon signaling and defense response to viruses while HDL-C and triglyceride comparisons led to pathways involving protein degradation and gene expression and translation. We were also able to observe general trends among groups when the p-values were adjusted for the heatmaps such that they displayed the most significant transcripts. For instance, when the criteria used to group samples in the HDL-C comparison was adjusted to select only individuals with the most extreme HDL-C levels, the top genes identified were all observed to decrease in expression for the low HDL-C group compared to those with high levels. We have also been able to identify genes that warrant further investigation for associations to HDL-C and triglyceride levels, as transcripts for *HDC*, *CPA3* and *GATA2* were all observed in the top significant transcripts for both.

6.0 AIM 3: IMPACT OF CNV ON SNP GENOTYPING

A single nucleotide polymorphism (SNP) is a mutation at a single base location in the genomic sequence. If the SNP is located within the coding region of a gene, it can have many effects, such as the substitution of one amino acid for another or the introduction of a premature stop codon. A SNP located in the regulatory regions of a gene, such as the promoter, enhancer, or intron/exon splice sites, can potentially influence the level of that gene's expression. Alternatively, many SNPs have no effect on gene expression; although some may be valuable measure of disease risk though linkage disequilibrium with other, functional SNP. Once a SNP is identified, genotyping it within a population involves calling the allele located in each chromosome copy by assays that are designed to pick up each base. As most genes are present twice in the human genome (once on each parental chromosome) the resulting data represents one of three possible genotypes: two copies of one allele (homozygosity); one copy of each allele (heterozygosity), or two copies of the second allele (homozygosity again). Because of this, SNP genotyping assays in the past were chosen based on the conditions and designs that gave the clearest 3 genotypes. If 3 distinct groups failed to be produced, additional assays such as Sanger sequencing were performed to clearly identify the genotypes. Any deviation from the normal expectation of the genotyping or sequencing assays was deemed to be noise from assay artifacts rather than a potential genetic variation. However now that more information is present regarding copy number variation and the existence of pseudogenes, this "noise" has the potential to be much more. In a recent study

from our lab, several SNPs were observed to have unusual clustering patterns during analysis with a custom designed Illumina GoldenGate assay on 1,945 MACS samples. Closer inspection of their genotype distributions, for example those shown in Figure 6.2, revealed the presence of more than the 3 distinct genotype groups expected. These clustering patterns suggest the possibility of CNV interfering with SNP genotyping as concluded by the clustering patterns. For the SNPs examined by GoldenGate, 3 genotype clusters were expected based on the number of possible combinations of alleles from each single copy of the gene on each of the two chromosomes, When multiple copies of a gene are present on each of the two chromosomes, there can still be 3 possible genotypes (provided that a third allele doesn't appear as a new point mutation in a copy) but the alleles present for the individuals in the heterozygous group are no longer at a 1:1 ratio. Depending on the sensitivity of the assay, the amount of each allele present in the sample would pull that individual's data point away from the heterozygous group and towards the homozygous group for its dominant allele (the allele in highest frequency among copies). In assays that are extremely sensitive and accurate, individuals with the same ratio of alleles will group with each other resulting in the formation of multiple subgroups between the homozygotes and heterozygotes. This is what appears to be present for the GoldenGate SNPs with unusual clustering.

6.1 AIM 3 METHODS

For genotyping our SNPs of interest, we employed three commonly used methods for analyzing SNPs individually; Fluorescence Polarization, Sanger Sequencing, and Life Technologies TaqMan SNP Genotyping Assays.

6.1.1 Sample Selection

Samples for this section of our study are described in detail within Section 3.1.1.1. There was a total of 2104 DNA samples available for analysis, including those from the 2005 cross-sectional study of genetic impacts on HAART-associated dyslipidemia described in Chapter 4 and Matthew Nicholaou's 2012 manuscript[25], as well as DNA samples from some individuals only present in the transcriptome study described in Chapter 5. These were available as both genomic and whole-genome amplified DNA (whole-genome amplification was performed in our laboratory by Dr. Matt Nicholaou, using the Illustra GenomiPhi V2 DNA Amplification kit [GE Healthcare Biosciences]). As whole-genome amplification is a commonly used technique to increase the amount of DNA available for low yield samples, and allow low concentration samples to be brought up to a usable concentration, it was important for us to determine how this amplification may affect the quantities of alleles present for genotyping. We used GenomiPhi DNA for all of the samples used in the Illumina GoldenGate SNP assay, as those samples were

specifically used for this. From the available genomic DNA samples, we took a subset of 373 samples for copy number analysis using NanoString technologies (described in Section 3.6). Of these, 192 were analyzed using all 3 assays for various SNPs of interest. In addition to these samples, representative 96-well plates from the entire 2005 cross-sectional set were selected for each SNP to be analyzed for all of the assays.

6.1.2 SNP Selection

SNPs for analysis were selected using several criteria. First, we aimed to find common SNPs, defined as those with Minor Allele Frequency (MAF) greater than 0.05, in regions of CNV by using the Database of Genomic Variants browser to scan each chromosome for high frequency SNPs within areas of structural variation encompassing genes with a potential lipid metabolism. This yielded a few genes of interest but due to its time-consuming nature we adopted other approaches to identify SNPs that were more likely to have CNV interference based on previous study findings. This involved SNPs from our Illumina GoldenGate assay that formed multiple groups within the heterozygous individuals as well as SNPs in genes and regions that were well documented in the literature as having CNV (Table 6.1).

Table 6.1: SNPs Selected for Analysis

SNP	Gene	MAF (HapMap-CEU)	Chromosome	CNV	Selection Basis
rs2271072	*FABP3*	0.42	1	None	DGV Scan
rs7037117	Near *TLR4*	0.208	9	Potential	GoldenGate
rs4352264	Near *EFEMP1*	0.487	2	Potential	GoldenGate
rs1828283	*CCL3L1*	N/A	17	Documented	Literature
rs2220067	Near *MRGPRX1*	0.518	11	Documented	Literature
rs917015	*CCL16*	0.216	17	None	DGV
rs3789864	*DEFB103A*	N/A	8	Documented	Literature
rs2477240	*DEFA1B*	N/A	X	Documented	Literature
rs2373961	*Near NOS3*	0.372	7	Potential	GoldenGate
rs6703462	*Factor 5*	0.042	1	Potential	GoldenGate

The SNPs are listed by their Reference SNP ID, also known as an rs number. If the SNP falls in or near a gene, it is listed in the gene column. When available, the minor allele frequency (MAF) is given for the CEU HapMap sample set. The CNV column lists if that SNP falls in a region of documented CNV from the literature, potential CNV identified in our GoldenGate assay or no CNV initially selected by a scan of the Database of Genomic Variants.

Table 6.2: Initial PCR Reaction Conditions for each SNP

SNP	Annealing Temp	MgCl2 Concentration	Alleles
rs2271072	*60*	0.6 uL of 25mM	C/G (rev)
rs7037117	*60*	0.6 uL of 25mM	A/G (fwd)
rs4352264	*60*	0.6 uL of 25mM	C/T (fwd)
rs1828283	*60*	0.6 uL of 25mM	C/G (fwd)
rs2220067	*58*	0.6 uL of 25mM	C/T (fwd)
rs917015	*60*	0.6 uL of 25mM	A/G (fwd)
rs3789864	*60*	0.6 uL of 25mM	A/G (fwd)
rs2477240	*60*	0.8 uL of 25mM	A/G (fwd)

6.1.3 Fluorescence Polarization

Genotyping using this method involves first amplifying a region of 300-500bp surrounding the SNP by standard PCR, as per Section 3.3.2, in a total volume of 10uL. The light-sensitive nature of this approach requires the use of black PCR plates, such as the Eppendorf Twin-Tec 96 black well plate. A list of optimal initial PCR conditions for each SNP can be found in Table 6.2.

Following successful PCR, excess primers and dNTPs are removed using our standard PCR clean up protocol described in Section 3.3.3. Once a clean product is generated, the single base extension PCR step is performed utilizing: an internal primer that binds directly next to the SNP; GE Thermo Sequenase Polymerase (Affymetrix 78500 1KT /GE E79000Y); and 2 different fluorescent labeled chain-terminating dideoxynucleotides (ddNTPs) specific to each of the alleles of the SNP. The mastermix for FP is shown in Table 6.3 while the cycling conditions are shown in Table 6.4.

Table 6.3 : FP Reaction Mixture

Reagent	Concentration	Volume per Sample (uL)
Buffer	*10X*	1.0
Internal Primer	*10 uM*	1.0
Dye Mix	Table 6.6	0.05
Thermo Sequenase	*4U/uL*	0.05
Sterile Distilled Water		7.90
		10

Table 6.4: FP PCR Cycling Conditions

Step	Temperature °C	Time Min:Sec
1	94.0	1:00
2	94.0	0:10
3	52.0	0:30
	Repeat Steps 2-3 40 times	
5	72.0	0:10
6	10.0	hold

When both the forward and reverse internal primers are being tested at one time, the initial PCR and clean-up steps use double the volume, which is then split between two black well plates for the FP PCR step. Depending on the genotype at the location of the SNP, one or both of the chain terminating bases will be incorporated for an individual by the Thermo Sequenase enzyme. FP discriminates the presence of an allele based on the incorporation of the its specific dye labeled ddNTP onto the internal primer, which allows the fluorophore to emit polarized light following excitation with polarized light. When the fluorophore is attached to a larger mass, the primer in this case, it rotates slower than its relatively free spinning counterpart attached to only the ddNTP. By exciting the PCR product with plane-polarized, the slower moving dye attached to the extension product will remain stationary enough to emit more polarized light than the faster rotating single bases that emits mostly depolarized light. We used the LJL Analyst HT to excite and detect emissions for PerkinElmer ddNTPs attached to the fluorophores R110 and TAMRA (Table 6.5).

Table 6.5: FP Dideoxynucleotide-5'-Triphosphate Catalog Numbers

	Catalog Number		
ddNTP	**R110**	**TAMRA**	**Unlabeled**
ddATP	NEL494001EA	NEL474001EA	77110
ddCTP	NEL493001EA	NEL473001EA	77112
ddGTP	NEL495001EA	NEL475001EA	77114
ddTTP	Use ddUTP		77116
ddUTP	NEL492001EA	NEL472001EA	Use ddTTP
	PerkinElmer		Affymetrix (77126)

Table 6.6: FP Dye Mixes for Single Base Extension

Standard Dye Mix Chart		Recipe for 1:16 Dye Mix G/T Example	
SNP	**Dye Combination**	*ddNTP*	*Amount (uL)*
G/T	*R110 U/TAMRA G*	*0.1 mM ddCTP*	*16*
A/C	*R110 C/TAMRA A*	*0.1 mM ddATP*	*16*
A/G	*R110 A/TAMRA G*	*0.1 mM ddGTP*	*15*
C/T	*R110 U/TAMRA C*	*0.1 mM ddTTP*	*15*
G/C	*R110 C/TAMRA G*	*R110 ddUTP*	*1*
A/T	*R110 U/TAMRA A*	*Tamra ddGTP*	*1*
			64

Emissions are detected in the parallel and perpendicular fields, which the Analyst HT uses to generate a ratio of these values that is then used to create a measure of polarization, expressed as an mP value. When the different dyes for each allele are read, two different mP values are generated that are plotted against each other in a scatterplot with one dye along the x-axis and the other along the y-axis. Individuals with a homozygous genotype will have a much larger mP value for one dye alone while those that are heterozygous will have large values for both dyes. When multiple samples are genotyped, the three possible genotypes will cluster together into groups on the scatterplot of mP values.

6.1.4 Quantitative Analysis of Sanger Sequencing Genotypes

Quantification of allele amounts for each SNP in this section (Aim3) of the study was performed by calculating the area of each allele at the position of the SNP using the perl script, polySNP. Two reference sequences were generated in a FASTA file containing each base of the SNP of interest centrally located in a 1kb span. If other SNPs were present in the GenBank reference of this sequence, they were not represented, as we were only interested in gathering the area for the

SNPs analyzed in this Aim. The standard script, as found on http://www.nybg.org/files/scientists/dlittle/polySNP.html, was used with the reference_file representing our FASTA reference sequence, the trace_file being an individual .ab1 sequencing file and the –p 0 option was selected to use PHRED for base calling and trimming the sequence along with determining the peak area measurement. As have not calculated a standard curve for our sequences, we left that option out. Because we were running multiple sequence files for analysis, we generated an additional script containing multiple polySNP usage scripts with one set up for each individual sequence run.

polySNP -r reference_file -t trace_file [-s standard_curve_file] [-l] [-a] [-p 0|1|2|3] [-c 0.XX]

6.1.5 CNV Analysis

We quantified CNV using NanoString Technologies as described in the main methods section on page 41). As reference samples with known copies were only available for our *DEFB103* probe, we generated only relative amounts of CNV to the sample used as a reference rather than absolute for all the other probes.

6.2 AIM 3 RESULTS

The initial method for SNP selection involved screening the first version of the Database of Genomic Variants (DGV) for genes possibly associated to lipid metabolism, located in regions of CNV and containing a high frequency SNP (Minor Allele Frequency >0.05). However, the early version of the DGV often lacked complete records for each CNV, missing key data such as frequency and sample size. Because of this, our initial method for identifying SNPs to analyze the effect of CNV on their genotyping was not as successful as anticipated. One such example was rs2221072 located in the *FABP3* gene. Within the DGV this gene fell in a block of complex variation (Figure 6.1) but lacked adequate frequency data.

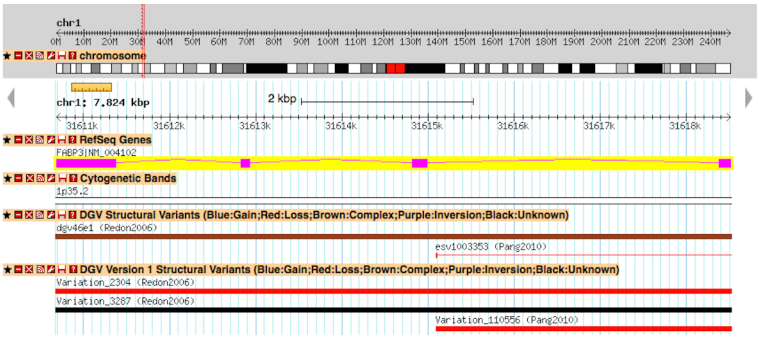


Figure 6.1: Copy Number Variation illustrated on Database of Genomic Variants

Upon testing the FP primer set for this SNP, we failed to see any interference impacting the ability to genotype and instead observed one of the better group separation patterns. This clean genotyping was further confirmed by sequencing and TaqMan SNP genotyping assays, while TaqMan CNV assays and NanoString confirmed lack of CNV, allowing us to use this SNP as

negative control for CNV (Figure 6.3). Because this screening method of identifying SNPs was dependent on the completeness of the data in the DGV, we focused more on SNPs that formed more than 3 distinct groups on our custom Illumina GoldenGate assay, together with genes and regions containing CNV documented in the literature (Table 6.1).

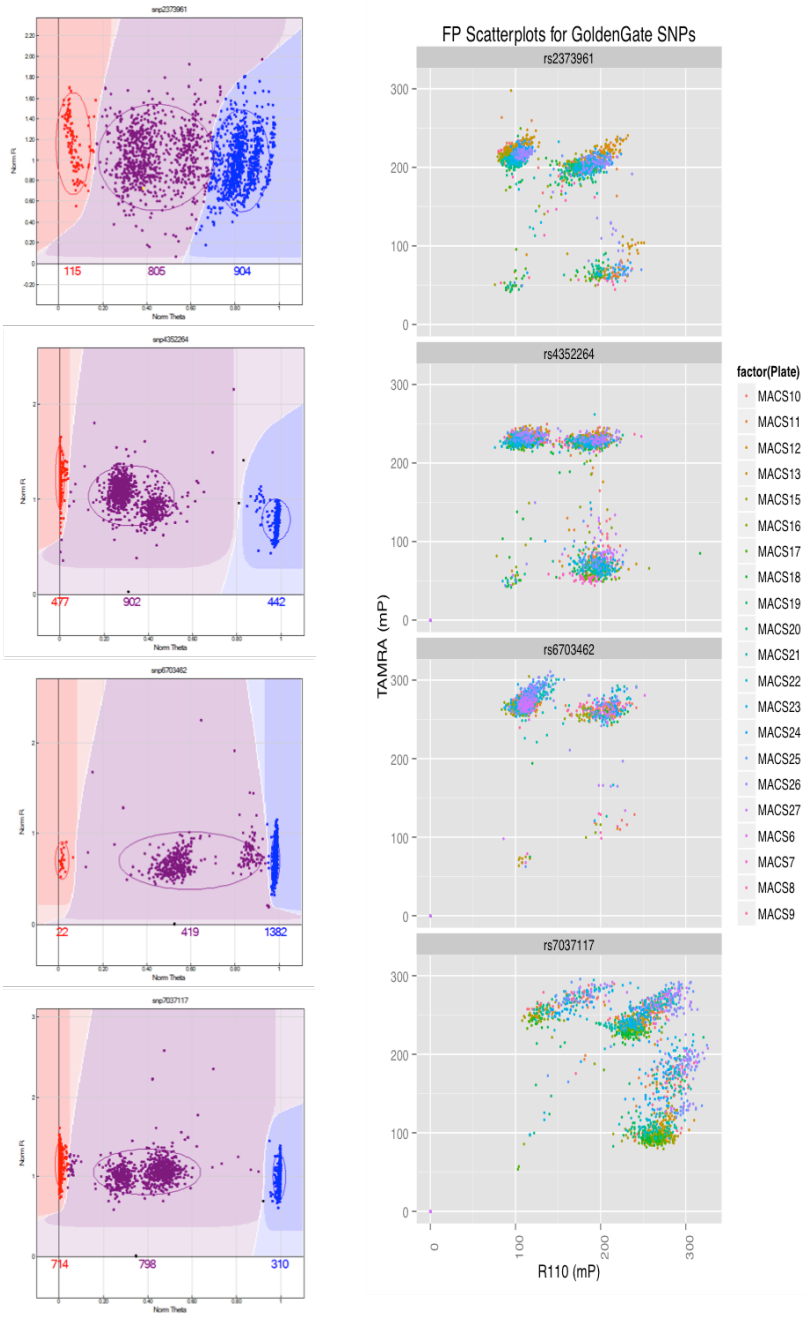
6.2.1 GoldenGate SNPs

We began by investigating a small number of SNPs (rs7037117, rs4352264, rs2373961, rs6703462) from the GoldenGate assay. These particular SNPs proved difficult to call due to formation of multiple subgroups during genotype calling. Upon analysis with FP, we observed a small amount of noise between genotype groups for the majority of SNPs but not the extent seen for the GoldenGate. As we used the same whole-genome amplified samples as those analyzed with the Illumina assay, DNA variation should not play a role in the difference seen for any of these SNPs analyzed.

Because we were concerned with identifying SNPs that could demonstrate how CNV affects SNP genotyping, we further investigated the two GoldenGate SNPs (rs4352264 & rs7037117) that had the most promising FP results. We also compared two assays for rs2373961 as its Real-Time PCR SNP genotyping assay was available from ABI. Of the SNPs analyzed, rs7037117 was observed to have multiple subgroupings in FP analysis (Figure 6.3). It also had reduced separation for the "A" allele where individuals who were homozygous had more signal for the "G" allele than what was expected as background noise alone. This is readily seen in plots from individual plates but is obscured when combining the separate reads from multiple plates into one plot. This increase of "G" allele signal was also observed in the TaqMan SNP genotyping

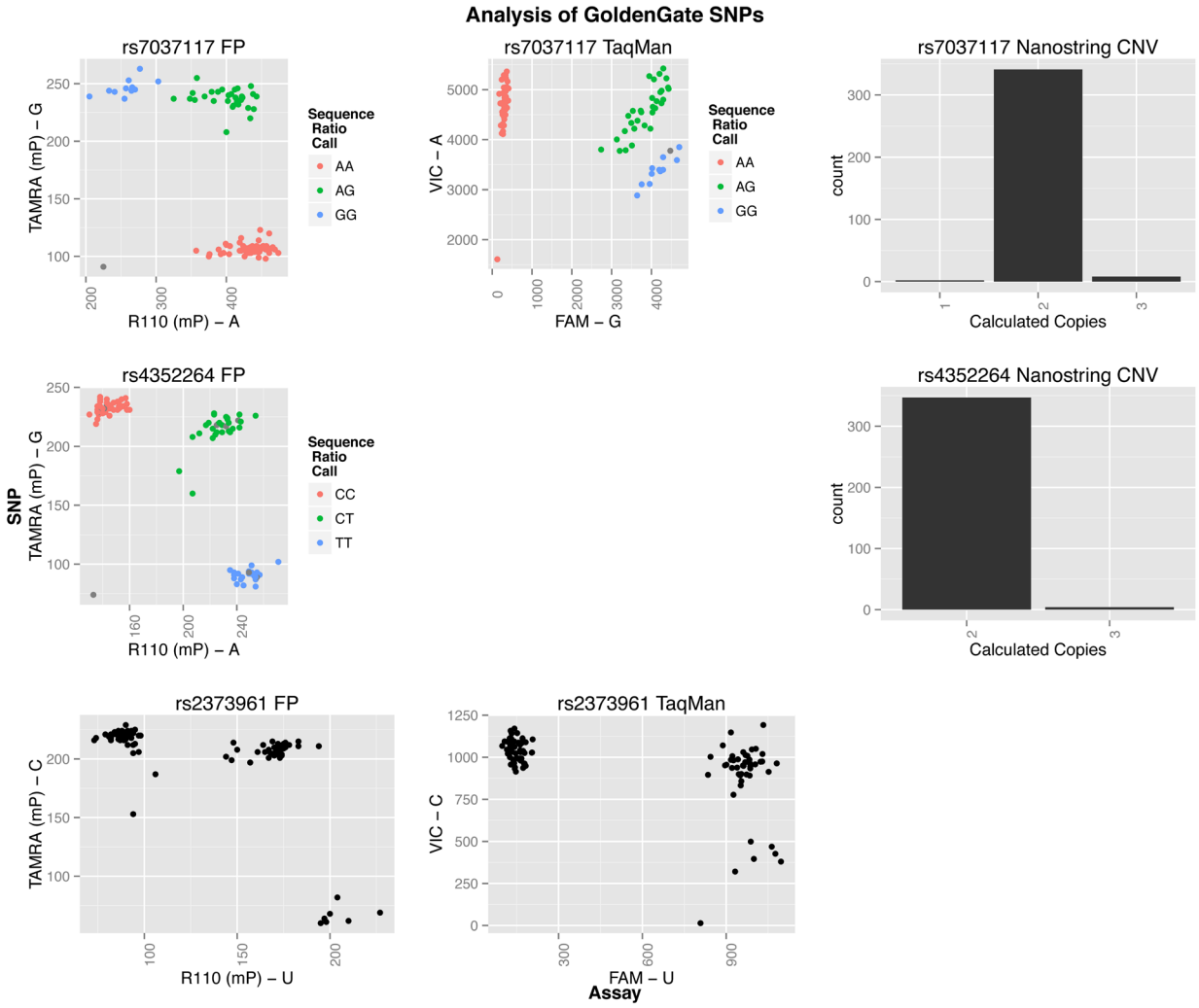
assay for rs7037117. The other SNP, rs4352264, while not having the extent of variation similar to that of rs7037117, had a wider spread around each of its genotype groups as well as some scatter along the G allele axis (Figure 6.2).

When these two SNPs were analyzed using sequencing, the apparent noise from both FP and the GoldenGate was not visible (data not shown). Each of the alleles had clean peaks of equal size suggesting equal amounts of each allele. When the regions around these SNPs were analyzed for CNV using a TaqMan CNV assay as well as NanoString technologies, a range of copies was not observed, but rather the vast majority of samples contained only 2 copies. The SNP rs2373961 also failed to replicate the noise observed during the GoldenGate assay as the plots for both FP and TaqMan SNP assays had clear separation between the genotype groups. Therefore the issues with the genotyping assays for these SNPs are likely to be within the assays themselves. This could be due to other SNPs located within the primer and probe binding sites, the GC content of the binding sites, specificity of the primers and probes as well as their ability to not form hairpins. And in the case of a multiplex assay such as the GoldenGate, there is also the possibility of probe to probe interaction that could reduce the quantity of probe available for binding the genomic sequence.



The left column of plots were generated during analysis of the GoldenGate SNPs by the Illumina BeadStudio software. Each plot is a different SNP with the clusters representing genotype groups. Samples that were homozygous for either the wildtype or mutant alleles are shown in the red and blue sections while the heterozygous samples are observed in the purple. These four SNPs were selected because they contained multiple subgroups within the heterozygous genotype. The column of plots on the right are those same SNPs from the left but genotyped using Fluorescence Polarization. Some of the SNPs still have unusual genotyping patterns but not to the extent of that seen in the GoldenGate assay.

Figure 6.2: GoldenGate SNPs Analyzed by Fluorescence Polarization



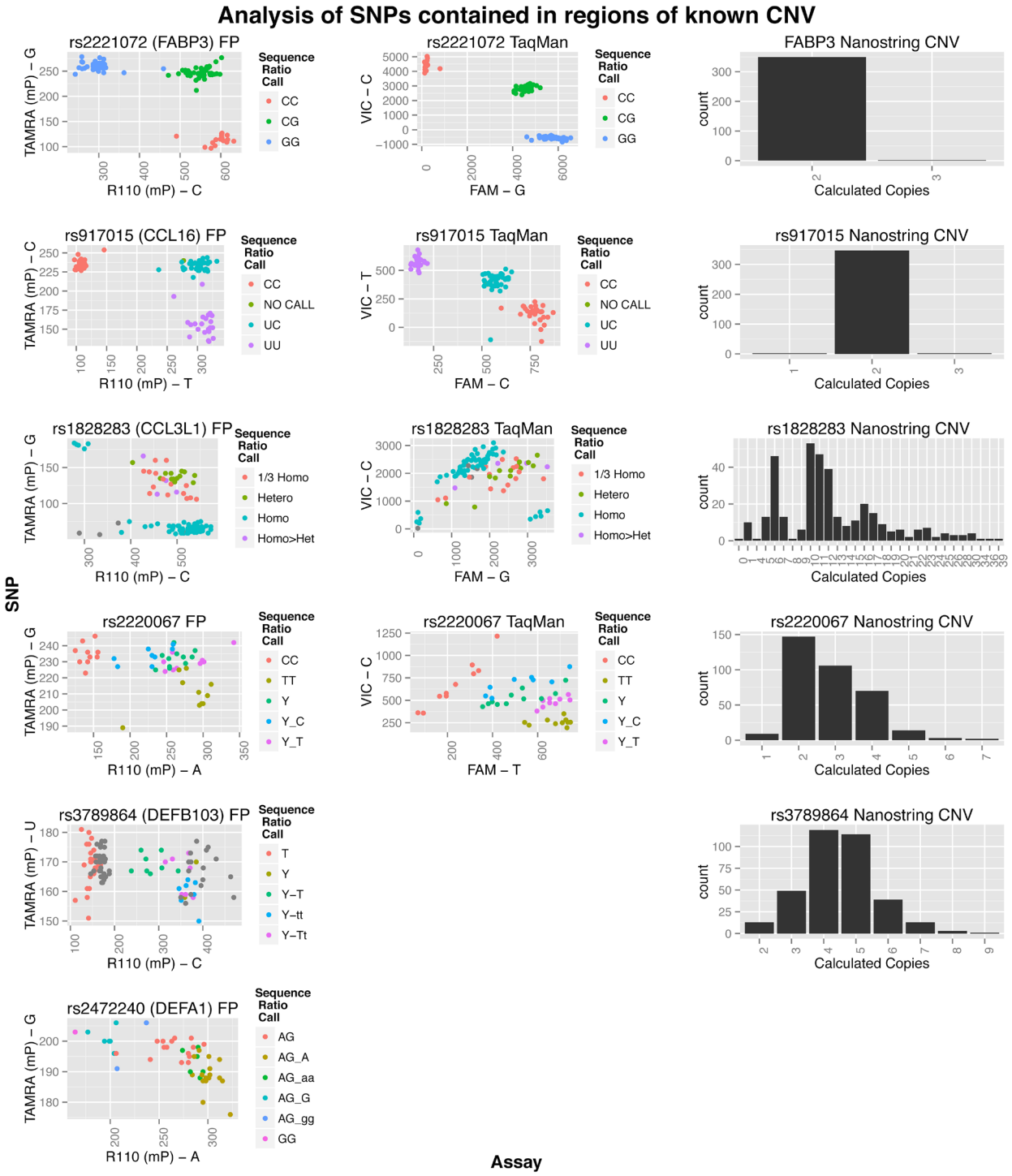
Plots for FP and SNP TaqMan are representative plates containing 96 samples. The NanoString CNV plot contains 384 samples. When sequencing data was available, the plots were color-coded based on the genotype obtained during sequencing. In some cases the alleles plotted for the FP or TaqMan will be the compliment to that of the sequencing results depending on if all the assays were typed on the forward or compliment strand of DNA.

Figure 6.3: SNP Genotype Analysis of GoldenGate SNPs

6.2.2 SNPs in Regions of Known CNV

While some of the GoldenGate SNPs had unusual genotyping patterns, these patterns do not in fact represent CNV interference on SNP genotyping. For this reason, we examined SNPs within genes and regions of known CNV. These included SNPs within *CCL3L1*, *DEFB103AB*, and *DEFA1*, as well as a SNP known to fall in an area of CNV (rs2220067 located near *MRGPRX1*). We also added another CNV negative gene, *CCL16*, for comparison in addition to *FABP3*. Use of these SNPs gave us an accurate view of the impact that a range of copies could have on SNP genotyping by FP. Rather than having discrete clustering groups, SNPs within regions of CNV had a continuous spread of genotype calls from one group to the other. This pattern was observed in the TaqMan SNP genotyping assays with individual samples locating to similar positions as seen for FP. Figure 6.4 shows the range of genotypes present for the SNPs containing CNV while also illustrating the clear clustering for those without CNV.

For some SNPs, such as rs3789864 and rs2477240, the continuous range of genotyping was less pronounced for one allele due to lower frequencies of the minor allele. In the case of rs3789864, which was able to be designed for FP, sequencing and CNV through NanoString, production of a custom TaqMan Assay failed several times, potentially due to the lack of homozygous mutant individuals. Because rs2477240 also had a MAF so low that the homozygous mutant allele group was missing as well, we opted to exclude this SNP from further study, as it was likely to generate similar problems with TaqMan. Regardless of the frequency of the alleles, the

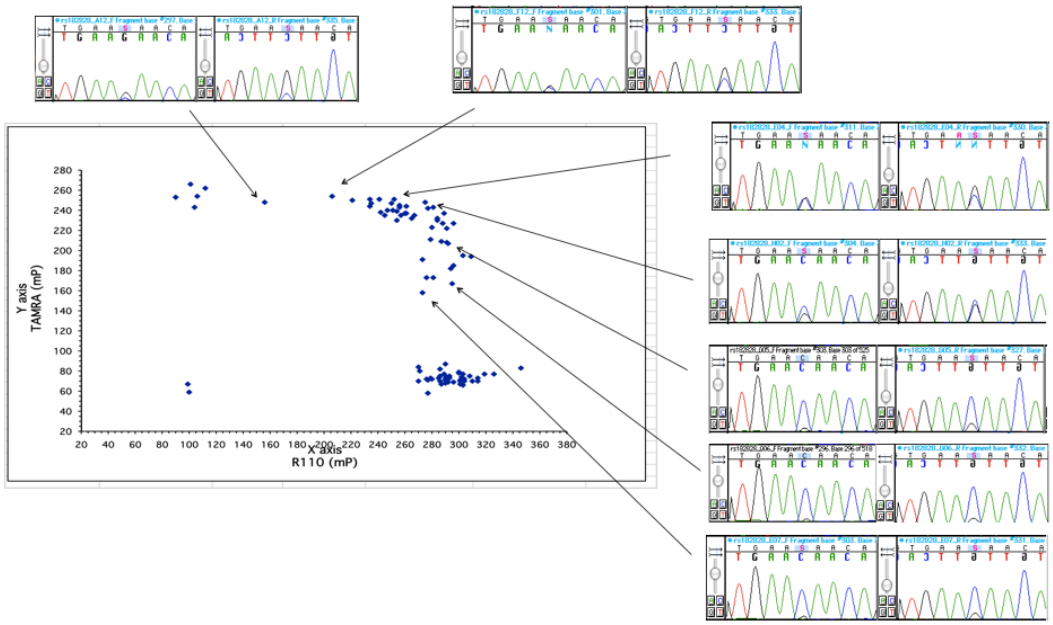


Plots for FP and SNP TaqMan are representative plates containing 96 samples. The NanoString CNV plot contains 384 samples. When sequencing data was available, the plots were color-coded based on the genotype obtained during sequencing. In some cases the alleles plotted for the FP or TaqMan will be the compliment to that of the sequencing results depending on if all the assays were typed on the forward or compliment strand of DNA. The columns contain the assays while the rows show data for a single SNP.

Figure 6.4: SNP Genotyping Comparison by Assay and CNV Amount

continuous spread of alleles observed for SNPs in regions of CNV makes it impossible to determine a cutoff between groupings. And if a cutoff were determined, giving a discrete label to this continuous occurrence of alleles would improperly call the actual genotype and any risk associated with it.

Further confirmation that the heterozygous individuals identified outside of the discrete genotype cluster regions were indeed representative of a range of allele ratios resulting from CNV, rather than assay artifacts, was provided by Sanger sequencing. SNPs present within regions of CNV were observed to have varying heights/areas of each allele at the SNP location relative to the amounts expected based on location in the genotyping plots. An example of this is shown in Figure 6.5 for rs1828283 where select individuals along the continuous spread of alleles have their corresponding sequencing trace for that allele shown. As points move from a homozygous genotype group to the heterozygous group, they are observed to gain more of their secondary allele as its peak at the location of the SNP increases in size. Further illustration of this is shown in Figure 6.4 where the individuals are color-coded based on ratios of their sequenced genotypes as determined by polySNP. To simplify the genotypes of those in CNV regions, the range between alleles was broken into subgroups for coding.



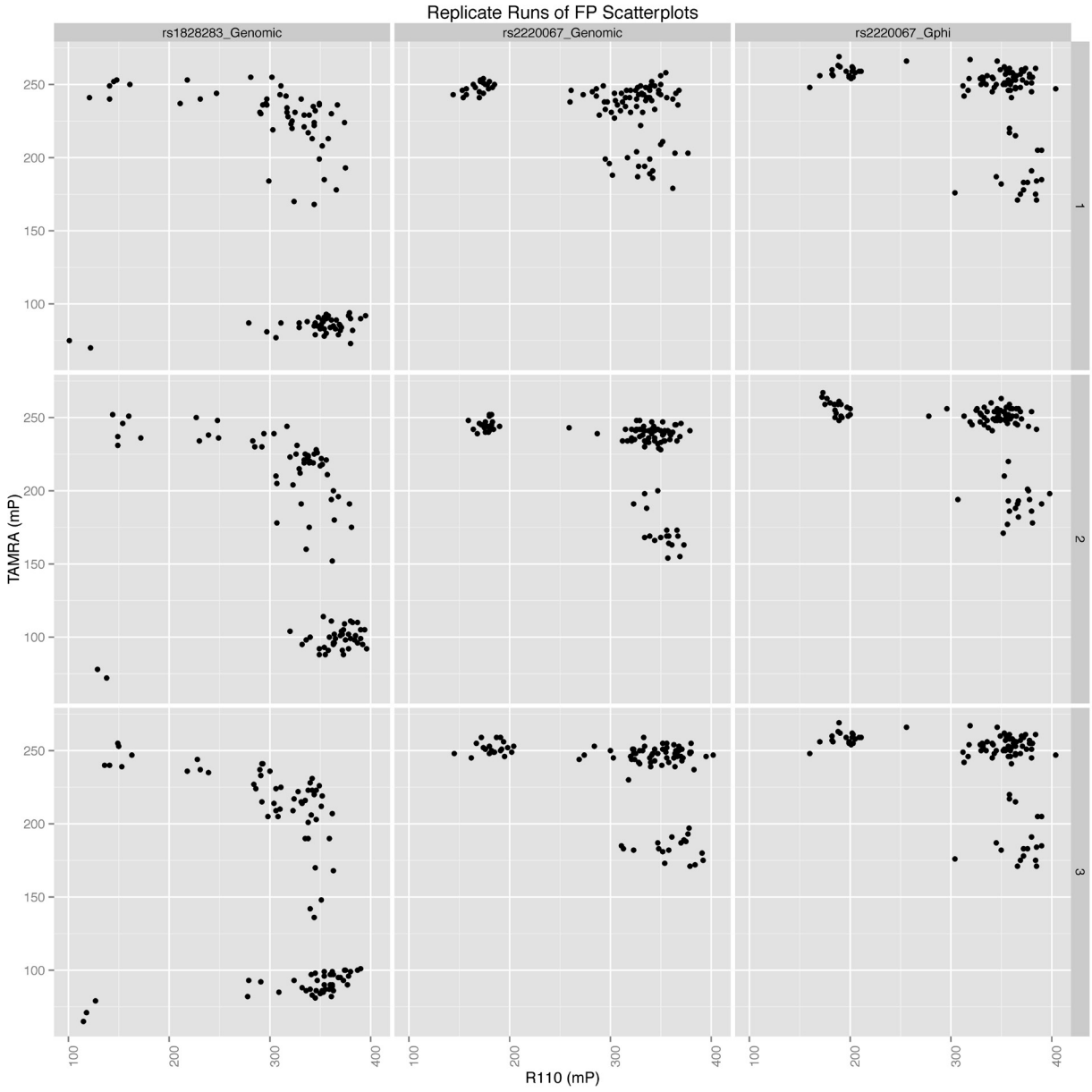
The FP data is plotted in the large section of the figure with representative sequencing trace files shown around the outer border. As the location of a heterozygous sample shifts closer to the homozygous cluster, the ratio of alleles for that individual shift as well such that the peak for the allele seen in the homozygous cluster now becomes larger in size.

Figure 6.5: Direct Comparison of Sequencing Data to FP Scatterplot

6.2.3 Other Assay Factors that can Influence SNP Genotyping

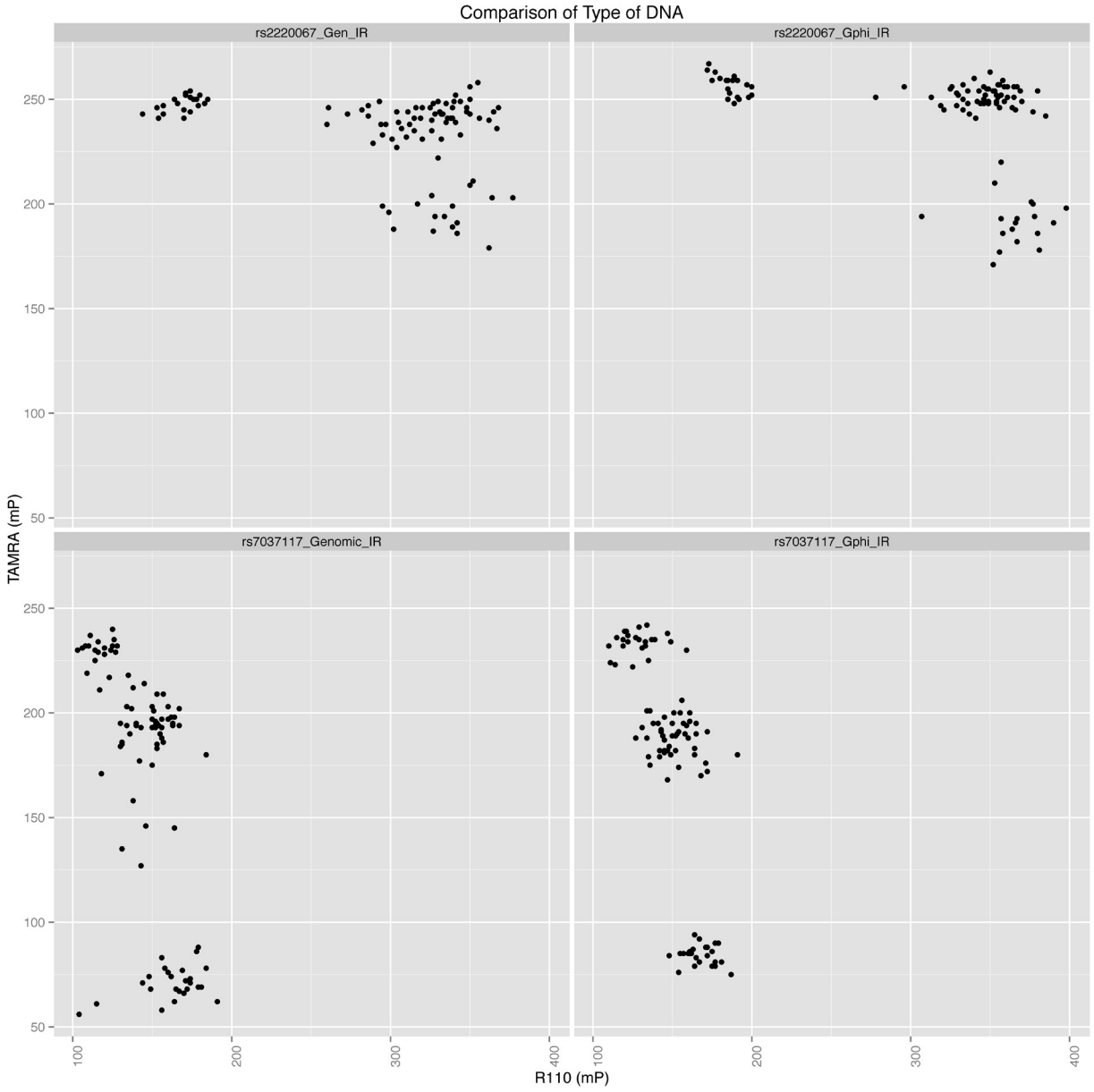
While the results above illustrate an association between the noise observed while SNP genotyping in presence of CNV, we identified that some variation between runs of the same samples could occur due to altering aspects of the assays. To determine if varying any of these parameters (DNA type, internal primer, dye mix, and type of polymerase) could lead to widespread genotyping problems, we tested the most easily altered aspects of our genotyping

assays focusing primarily on FP as it is the most easily altered assay. First, reproducibility of the FP assay was tested by running two of the SNPs (rs1828283 & rs2220067) in triplicate. The scatterplots for each of the runs as shown in Figure 6.6 were very similar with individuals remaining in nearly the same locations between runs. Samples that did move stayed within the same genotype groupings and never jumped from one extreme to another. Obviously it is not expected that each run would be an exact duplicate of others as this three step assay has room for slight variation between runs due to occurrences of evaporation, pipetting variation, differences between thermocyclers, and adjustments of the LJL Analyst parameters without our knowledge. Even so, FP performed consistently between runs. During this analysis, we also tested two separate types of DNA for rs2220067 as we not only use genomic DNA samples in our lab but also DNA that is whole-genome amplified. While it would be ideal to use genomic DNA for all assays, some of the samples available to us are limited due to the individual leaving the study either by choice or progression of illness. Because the reduced availability of some samples, as well as low DNA concentrations, could have hindered our ability to run all the necessary assays, we performed whole genome amplification on those samples, as well as all of those in the cross section for continuity. The comparison of DNA type for rs2220067 illustrated that genomic amplification slightly subdues the range seen for genomic DNA. It does not alter the overall distribution but rather softens it causing groups to be more compact.



Columns represent each SNP/DNA used for the replicate runs while rows represent each of the triplicate runs for those SNPs. Each point is an individual sample out of the 96 samples per run.

Figure 6.6: Replicate Runs of Individual SNP Assays



Along the rows of this plot are the SNPs while the columns represent DNA type. Genomiphi samples are represented by the titles with "_Gphi" while genomic are represent by "_Genomic"

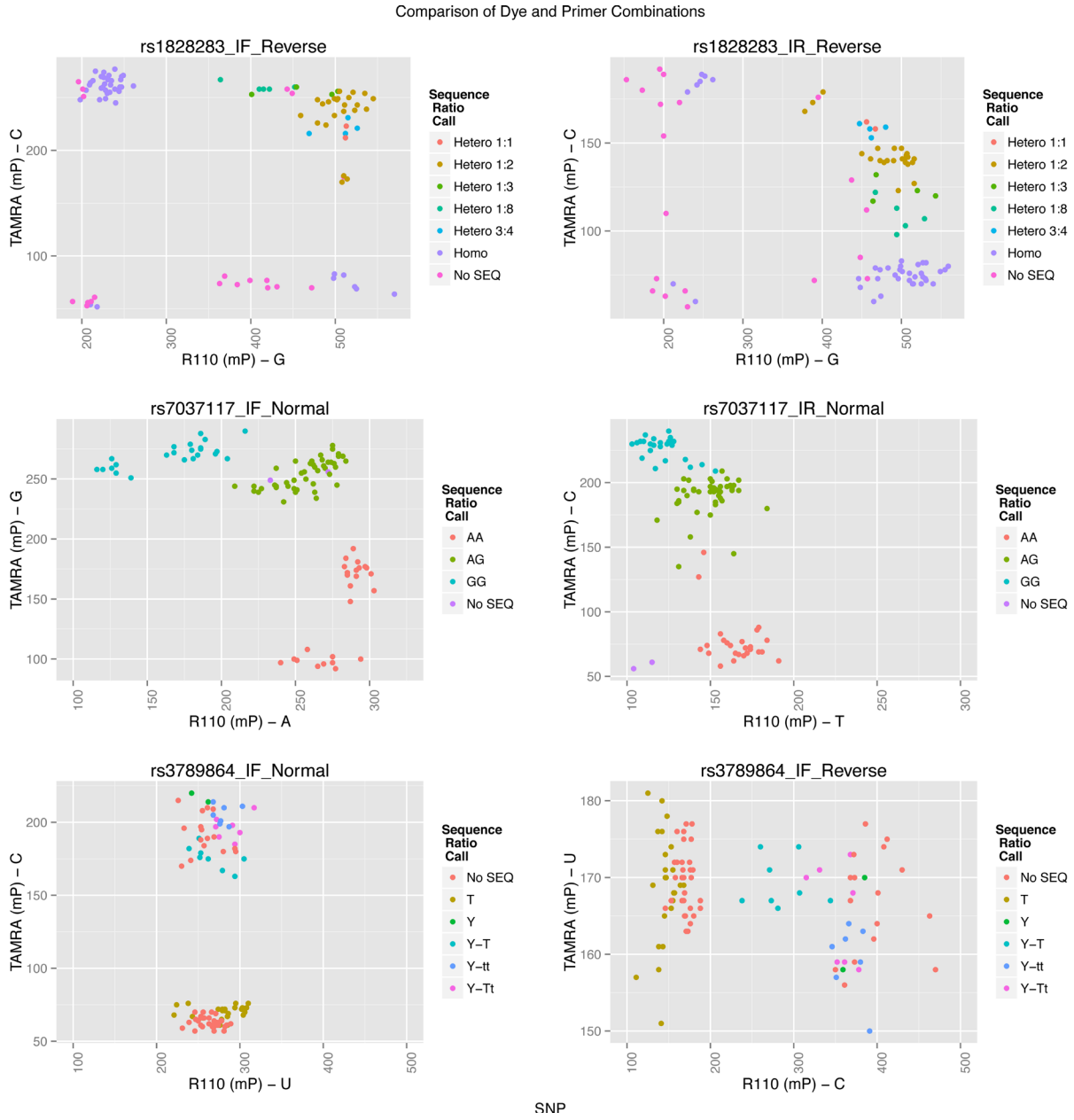
Figure 6.7: Impact of DNA Type on FP Genotyping Assay

Because of this observation, we decided to test other SNPs and assays for DNA variation. Independent of assay type, the GenomiPhi DNA clustered tighter and had reduced amount of noise between clusters (Figure 6.6 & Figure 6.7). With this finding, it is key to use genomic

DNA, when available, for assays within regions of CNV as genomic amplification can marginally reduce the natural range of variation present.

The next combination of assay conditions we wanted to compare was that of the FP internal primers and dye combinations. Although only one primer is used in the single base extension assay, both the forward and reverse internal primers are generally manufactured, to identify the primer that provides the best separation for genotyping groups. Because these primers anneal directly next to the SNP of interest on the DNA sequence, their design is limited to the ~50bp region around it. If other SNPs are located within an internal primer sequence or the GC content is not ideal, these primers cannot be interchanged with ones that bind another location as their 3' end needs to bind directly next to the SNP. It is these inherent variations surrounding the SNP that often times lead to one of the internal primers failing to bind properly resulting in poor genotyping patterns. When comparing different primer combinations, we observed mixed results. Often, one primer would generate clearer results than the other, but this varied depending on the SNP. For instance, in Figure 6.8, the separation for rs1828283 is relatively similar for the different primers but the spread of heterozygous individuals for the internal forward primer is tighter than that of the reverse. In other cases, separation may be greatly reduced as in the case of rs7037117 where separation of genotype groups went from bad to undesirable when using different internal primers. This is observed in the heterozygous group in the reverse primer assay where the homozygous individuals merge with the heterozygous making it impossible to call either of the individuals in those groups accurately.

Similarly, the different dye labeled dideoxynucleotide (ddNTP) combinations can yield varying results for genotyping plots. That is why the standard ddNTP mixes (Table 6.6) used for FP consist of just one set of dye labeled bases for each type of SNP present. These combinations were selected as they have consistently yielded the cleanest separation of genotype groups over many years of use. While that kind of separation is key for a SNP containing equal amounts of alleles for their heterozygotes, it is not optimal for heterozygotes with a range of allele ratios. This was observed with the different dye mixes used for rs3789864 where the Normal dye-labeled ddNTP mix merged all of the heterozygous individuals together despite a range of allele ratios (Figure 6.8). When the reverse dye mix (alleles attached to R110 and TAMRA are opposite to those attached in the standard) is used, the degree of separation observed by sequencing is also seen in the FP. We observed during this set of altered assay conditions that general methods for selecting clean genotyping results for normal SNPs are not suitable for selecting conditions for SNPs in regions of CNV. While this presents the possibility of improper SNP genotyping, it can be countered by sequencing the samples with the same primers for the FP prior to selection of assay conditions. This is because sequences of the FP regions when using the same amplicon PCR primers as those used in FP generates a consistent representation of the variation of alleles present.



In this plot, the different dye and primer runs are represented in the columns for each of the SNP runs represented in the rows. Each point in the plots is color-coded based on its sequencing call. Heterozygotes with uneven ratios are grouped into smaller subsets to better illustrate their ratio amount vs plot location. When a reverse primer is used the complementing dye mix is also used. This is why in some plots one group of homozygous individuals swaps places with another.

Figure 6.8: Ability of Altering Dye and Internal Primers to Hinder Ability to Genotype

Within this section of our study, we were able to show that when CNV is present it has a distinct impact on the ability to call a SNP properly, using the standard assumption that only 3 genotypes will be present. While methods for SNP detection are still dependent on prior knowledge of the genome, use of any genotyping methods with probes based on reference genome sequences need to be scrutinized during initial design. Performing multiple validations of the accuracy of the genotype calls is necessary as well as a method such as Sanger sequencing to validate the assay further. While the presence of CNV will generally create more noise and less separation between the groups being genotyped, there is a chance that unknown limitations of the assay, such as variation due to use of different dye mixes, can negate the appearance of any noise, which in turn would result in failure to observe the true ratio of alleles present. However, as whole genome sequencing technology is becoming more readily available, both SNP genotyping and CNV calls will be achieved in one assay as read depth of the sequencing would allow determination of CNV and also give amounts of each allele present for proper SNP genotyping.

7.0 DISCUSSION

7.1 RCT GENE COPY NUMBER VARIATION

The contribution of host genetic variation to the development of the CVD-associated side effects seen in response to antiretroviral therapy is still not fully understood. We have previously studied the roles of Biogeographical Ancestry[25], and of individual SNPs on this, but no studies have been done to date on the impact of quantitative genetic variation such as CNV on this process. To address this, we developed an MLPA assay, and used it to measure CNV in genes within the reverse cholesterol transport (RCT) pathway[29, 30]. As extreme HDL and LDL abnormalities are observed in only a subset of HIV-positive patients receiving anti-retroviral therapy and experiencing dyslipidemia[70], the susceptibility to these severe lipoprotein changes is likely to have a genetic component.

While previous studies have already found an association between sequence variation in genes within, and influencing, the RCT pathway and lipoprotein levels, we theorized that CNV in the RCT pathway could play a role in these extreme lipoprotein phenotypes. A region of duplication encompassing an entire gene and its regulatory regions has the capability to alter expression and protein levels in a manner directly proportional to the amount of copies present. Such a

relationship is observed for the *CCL3L1* and *DEFB4* genes, where increases in gene products correspond to copies present for each gene up to a plateau point[39, 42]. Even though this type of variation has the potential to influence lipid metabolism, the available information on whole gene CNV in the genes of the RCT pathway is limited to a few select genes (*LDLR*, *LPL*, *ABCA1*, and *LIPC*) [43, 44].

Data currently available within the Database of Genomic Variants[92] show a limited amount of rare CNV present for the RCT genes. The documented CNV that is there consists primarily of insertions and deletions within the genes, rather than whole gene variation. Thus, it is unlikely that CNV for these genes is common in the general population, but our strategy here was to combine a population-based screen with a focused investigation of individuals with extreme lipid phenotypes (strongly atheroprotective *vs.* strongly atherogenic). Our hypothesis was that CNV encompassing the full length of a RCT gene would result in an increased or decreased amount of transcribed protein product directly proportional to the amount of copies present, thus impacting serum cholesterol levels. Further, we wished to investigate whether individuals with CNV in the central range would have normal lipid levels while those whose CNV was in the outermost edges of the range would have a dysregulated lipid metabolism leading to the extreme lipid profiles.

Our results in this study identified apparent rare loss variants in 3 of the RCT genes. Out of 267 individuals and 16 genes studied with two different CNV assay procedures, *CETP* showed a loss in a single individual, and two genes (*ABCA1* and *APOA4*) showed apparent copy number losses with MLPA. The loss of copy number seen for *APOA4* was determined to be due to a SNP in the ligation site of its MLPA probe while the loss seen for *ABCA1* is suspected to be not genuine as

its loss fell upon the threshold. The small standard deviation seen, along with the reproducibility of the significant loss of signal during additional MLPA runs, indicates that the loss for the *CETP* probe was valid. Coupled with the tight clustering around the normalized ratio of 1.0 for the non-outlying points of all of the RCT probes, these results strongly suggest that whole gene CNV is not present in the RCT genes at anything above very low levels, and is therefore not likely to be a major influence on lipid levels in either the normal population or those infected with HIV and receiving antiretroviral therapy.

These findings are consistent with previous reports of limited structural variation in the RCT genes, as presented in the Database of Genomic Variants[50, 92]. Within this database, deletions that included whole genes were observed for *CETP*. The *ABCA1* gene was observed to have a wide variety of insertions and deletions within its bounds, including 5 losses in the region of our MLPA probe, although none of them encompass the entire gene. All of these reported variants were extremely rare, with only a few individuals having the variant in studies containing several thousand participants. For those with higher frequencies, the study sizes were too small to conclude that a common variant was observed.

We also compared the ratios obtained with both CNV assays to gene expression data available for a subset of our samples. None of the genes for which we had available CNV data showed variation in expression level. This further suggests that it is unlikely that significant CNV is present in these genes that might affect expression.

In this section of our study we were able to develop a sensitive MLPA assay that can accurately detect CNV when it is present. Using this assay we have illustrated that whole gene CNV is present only at very low levels in the RCT genes, and is not a major factor in the development of HAART-associated dyslipidemia. Thus, other host genetic influences exist and need to be identified before we are able to understand fully the ways in which host, viral, and therapeutic environmental factors interact to determine the outcome of antiretroviral therapy in HIV-positive individuals.

7.2 TRANSCRIPTOME ASSOCIATED WITH LIPID LEVELS

With the rapid advances made in whole genome sequencing platforms and the increased accessibility to whole exome sequencing among the scientific community, there is an increased interest to simultaneously examine expression profiles along with sequence variation. As whole-transcriptome assays simultaneously analyze expression levels of all transcripts in the human genome within a single experiment, they are prime candidates to be used in conjunction with exome sequencing. While in theory this is practical, the execution of such a study can be problematic due to the starting material needed for transcriptome analysis. As gene expression varies in a tissue-specific manner, studies, such as ours, examining lipid levels would require liver samples for expression analysis. Unfortunately, such tissue samples are not readily available from the study population for various reasons; safety of the participants, desire not to donate, or unavailability due to deceased status. For this reason, a surrogate source of RNA is needed. As blood collection from study participants is easily obtained, lacks major risks and is not as impacted by donation denial, PBMCs within the serum are a suitable candidate for tissue

surrogates. Studies have already illustrated that whole blood can serve as a suitable source for expression analysis in a variety of settings including investigating expression changes in cardiovascular disease. Such investigations have resulted in identification of novel biomarkers and predictors of disease outcome. Furthermore, expression analysis in leukocytes has identified differentially expressed genes involved in lipid metabolism and inflammatory response when investigating lipid levels.

We aimed to determine if RNA derived from whole blood collected in a PAXgene tube would enable differential expression analysis among study participants with atherogenic and normal lipid profiles to be performed, to expand the existing expression analysis results present for cardiovascular disease in the MACS.

Indeed, we were able to identify differentially expressed transcripts within whole blood for each of our comparisons prior to correction for multiple tests. Once the p-values were adjusted, the significance of those transcripts dropped profoundly. Despite this, we were still able to identify significant transcripts for comparisons involving infection status, CD8 counts, viral load, HDL-C and triglycerides.

The resulting transcripts for comparisons among infection status generated numerous transcripts associated with immune response, which is as expected. The most significantly associated transcripts were those of the CD8 isoforms. As HIV-1 infection results in chronic immune activation and exhaustion of CD8 T-cells, an increase in CD8 expression in those infected with HIV when compared to uninfected controls is logical. This increased expression is likely a

product of higher CD8 counts among the HIV infected as we were able to show that individuals with increased CD8 counts also were observed to have increases in the same CD8 isoforms differentially expressed in the HIV status comparison. Yet, it is still unclear what is behind the increased absolute counts of CD8 cells. Increases of CD8 T-cells have been observed during the natural course of infection[93]. Additionally CD8 T-cells are also observed to have increased turnover, activation and proliferation[94, 95]. This could explain the increased expression observed. However, therapy has been shown to reduce proliferation and result in the increase of CD4 T-cell counts for some[96]. But as the majority of our HIV-infected individuals had higher CD8 counts than the negatives, we have failed to confirm previous findings. This could be due to use of absolute counts of CD8 T-cells rather than percentages or the CD4/CD8 ratio. Because cellular composition of the serum is altered during infection, we may be observing higher CD8 counts as a result of this. Furthermore, as the CD4/CD8 ratio has been illustrated to be a predictor of higher morbidity and mortality[96], further analysis of our CD8 expression data in context of the CD4/CD8 ratios will determine if our increased CD8 expression is a marker for low CD4/CD8 ratio levels or merely representative of the amount of CD8 T-cells in the serum.

When contrasting individuals with viral load and lack thereof, various transcripts related to immune response and regulation of viral replication within the host were observed. Within the transcripts involved in immune response were various ones involved in the Type I interferon response. Considering Type I interferons are antiviral immune modulating cytokines, the expression differences observed are feasible for follow up.

On the other hand, differentially expressed transcripts identified to be associated with lipid levels initially did not seem to have a direct connection to cardiovascular disease or the comparisons at hand. Within the comparisons that gave us significant transcripts (HDL-C and Triglycerides), the top transcripts identified across all comparisons contain HDC, GATA2, SLC45A3 and CPA3. HDC, the gene for histidine decarboxylase, was observed at the top transcript for HDL-C comparisons. Initially, the catalytic enzyme of histamine synthesis does not appear to be associated with lipid metabolism or CVD risk but closer scrutiny of the literature reveals the role of histamine in atherosclerosis.

Levels of HDC have been shown to increase in the aorta during atherosclerosis progression.[97] Also, histamine has been shown to be produced by HDC expressing cells that can be found in atherosclerotic lesions including mast cells, foam cells, monocytes/macrophages and T-cells [98-100]. Histamine levels have also been observed to be increased in individuals with stable coronary artery disease and acute coronary syndrome[101]. Additionally, increased incidence of atherosclerosis among those likely to have elevated levels of histamine (sufferers of common allergies and asthma) has been observed in epidemiological studies[102-104]. These previous findings illustrate a strong association between histamine and atherosclerosis.

There are 4 membrane receptors for histamine of which the H1 (HH1R) and H2 (HH2R) are expressed on aortic vessels[105]. HH1R has also been observed to be expressed on vascular endothelial cells, foam cells and smooth muscle cells in atherosclerotic lesions[106]. And signaling through this receptor has been shown to cause smooth muscle proliferation as well as

increased vascular permeability[105, 107]. These two actions are key factors in progression of atherosclerosis and indicate the extent to which histamine can influence that progression.

However, various studies involving mouse models have illustrated that the impact of HDC and histamine on atherosclerotic lesions can be altered. Knockout out studies involving two different methods of inducing atherosclerosis (physically induced vascular injuries or hyperlipidemia-induced) found that knocking out HDC resulted in a reduction of atherosclerotic lesions. For the physically induced atherosclerosis, mice with carotid ligations were observed to have increased HDC expression in those arteries in comparison to non-ligated arteries[105]. They also were observed to have thicker intimas with markers for smooth muscle cells and macrophages as well as increased levels of histamine indicating active plaque formation. The HDC knockout mice on the other hand were observed to have decreased intima thickness in comparison to wildtype. This phenomenon could be reversed when HDC knockout mice received wild type bone marrow transplants. During the hyperlipidemia-induced atherosclerosis, double knockout mice were used (APOE-/- and HDC-/-) [100]. Within the APOE knockout mice, serum histamine levels were increased as well as expression of HDC and the histamine H1 and H2 receptors at the site of the atherosclerotic lesions. Double knockout mice had no such increases and also were observed to have smaller lesion areas despite having high serum cholesterol levels. They also were observed to have high HDL levels in comparison to controls. These two studies together illustrate that HDC and histamine together are associated with atherosclerotic lesions.

Yet another study illustrated that the histamine H1 receptor promotes atherosclerotic lesion formation. This study utilized the APOE knockout mice and treated them with HH1R or HH2R

antagonists while feeding a high cholesterol diet to observe that inhibiting the H1 receptor resulted in fewer atherosclerotic lesions. This inhibition went from 40% to 60% when they created a double knockout (APOE-/- and HH1R-/-). And when the double knockout received a bone marrow transplantation from an APOE knockout mouse, the amount of lesions remained reduced indicating that the H1 receptor present on the vascular cells is needed to promote atherosclerosis. They also found that the H1 receptor was associated with increased aortic inflammation and permeability[108]. From these combined mouse models and human studies findings, evidence suggests that HDC, histamine and the H1 receptor are equally important in the etiology of atherosclerotic lesions.

In addition to all of the literature on histamine, the 2010 Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study by the Wellcome Trust identified the same transcripts (HDC, CPA3, GATA2, and SLC45A3) as some of their top associations with Triglycerides, HDL-C and APOB. To account for differences in expression due to variable amounts of cell types in the whole blood, they added already identified cell specific covariates into their model and the transcripts remained equally as significant[109]. As these findings were present in a population not infected with HIV-1, this would suggest that our top differentially expressed transcripts are relevant to the larger population of individuals with cardiovascular disease. It also serves to validate our findings for further follow up studies on these transcripts within the HIV-1 infected population currently receiving therapy.

Based on the data in the literature, our identification of HDC as a top differentially expressed transcript associated with lipid levels further substantiates the potential role of histamine in

atherosclerosis. As we compared only lipid values, this will need to be followed up by identifying if these transcripts are still significant among individuals with evidence of arterial thickening and advanced atherosclerosis.

Within this section of the study as a whole, we have shown that whole blood is a valid surrogate for transcriptome analysis, as we identified numerous genes differentially expressed that were associated with HIV- infection and lipid levels. While some of these initially did not have a direct link to the comparisons at hand, further investigation has led to connections that were previously unimaginable. Follow up studies on the top transcripts may prove to find additional methods resulting in metabolic alterations that were not anticipated before this study.

7.3 SNP GENOTYPING INTERFERENCE BY CNV

As the general public becomes more aware and interested in the impact of genetics on disease risk, testing platforms will crossover from the scientific to the commercial market. This means that individuals with very little scientific background will view reports indicating risk associated with their genetic makeup. As laypersons to such scientific reports, there is the likelihood that the results will often be over interpreted as not a possible risk but as a definite conclusion that the disease outcome will occur.

Such scenarios have already begun to play out involving the 23andme genetic testing service that not only provided ancestry reports but also health reports indicating if the individual tested had mutations (SNPs in this case) associated with certain disease outcomes. Due to the apparent

medical nature of these reports, the FDA mandated the cessation of generation of the reports. But, this brings up a valid concern regarding the use of single nucleotide polymorphisms as risk factors for clinical outcomes.

The use of SNPs for clinical measures requires that detection of the variation is an accurate rendering of the actual variation present. While this is feasible, there is the definite possibility that SNPs associated with risk fall within regions that are difficult to genotype (Alu repeats, high GC, regions of CNV). If proper design measures are not taken when initially designing an assay, the results derived during SNP genotyping may not truly represent the inherent risk associated with a specific genotype of that SNP. For instance, if a SNP falls within a region of copy number variation, the standard model of 3 equal genotypes expected for a SNP will no longer hold true. Based on the number of copies present containing each allele, the genotypes will now include a range of subgroups for the heterozygous individuals based on proportion of each allele. Now an individual typed as a heterozygote based on the presence of both alleles will be inaccurately genotyped unless they have equal proportions of each allele. Despite this, little is available in the literature involving the interaction of CNV with SNP genotyping, most likely because assays that fail to give clear results are often not developed despite high minor allele frequencies. Such is the case for TaqMan assays for all of the SNPs in regions of CNV that we investigated. While CNV assays were readily available, the SNPs of interest within those CNV assays were not. Additionally, while designing the custom SNP genotyping assays through their service, some of those continued to fail their design procedures most likely due to the assay's resulting data not meeting a minimum genotyping pattern. For this reason, we intended to investigate if common single SNP genotyping assays would be able to identify the range of

variation expected in the presence of CNV or if such assays would mask the true variation present.

In this section of our study, we illustrated that SNPs located within regions of CNV were likely to have genotyping patterns that made SNP genotypes difficult if not impossible to call. This was apparent across two of the three genotyping platforms we used. Sequencing alone continued to give clear genotyping calls because this method intrinsically involves evaluating each allele as opposed to genotyping patterns.

We also observed that genotype patterns vary based on alteration of assay conditions when the genotyping method is FP. In most cases the variations were minor, such as that seen with use of different types of DNA, but in other cases where internal primers or dye mixes were altered the consequences could completely change the conclusion. For instance, in the case of rs3789864 the original dye mix yielded two clean groups despite the SNP's location within DEFB103 (a gene with well documented CNV). When the reverse combination of dyes, where the base attached to R110 in the original mix is now swapped for the base attached to TAMRA originally and *vice versa*, the genotyping plot clearly shows the influence of CNV on the spread of genotypes. Sequencing verified that the range of genotypes was authentic and that the dye mix was likely to blame. As the standard dye mixes used for FP are combinations that were chosen because they yielded clean separation of groups when 3 genotypes were all that was expected, it is likely that other standard dye mixes could miss CNV interference as well.

From our observations, we recommend that when designing a single SNP genotyping assay, it is necessary to evaluate the region for CNV through websites such as the Database of Genomic Variants first. Design should also encompass running the genotyping assay in conjunction with other SNP genotyping assays, preferably sequencing of the region, to verify that calls are properly determined. And no longer should sequencing chromatographs that look noisy at the site of the SNP be ignored or disregarded as they may represent the actual genotype. That being said, the trend for genome analysis is moving towards whole genome sequencing as it becomes more readily available to the scientific community. This type of sequencing will negate having to run separate assays for CNV and SNP genotyping as the read depth of the sequence can be used to determine copy number and the sequence itself will indicate the SNPs present. However what we found will still be valid, as these assays will allow investigators to identify SNPs of interest that will likely be followed up in a larger sample size using single SNP detection methods. While these findings do not apply to SNPs within the RCT genes we studied, as those genes did not have CNV, they will still apply to other SNPs outside our area of focus.

7.4 SUMMATION

Within this study we investigated the relationship between copy number variation (CNV) and dyslipidemia associated with antiretroviral therapy use among individuals infected with HIV-1. We began by designing a custom multiplex ligation dependent probe amplification assay for 16 genes found within the reverse cholesterol transport pathway. Our accurate assay identified an extremely rare deletion within the *CETP* gene for one person. As none of the other genes varied from two copies, copy number variation of the reverse cholesterol transport pathway genes is not a factor in the lipid dysfunction observed during HAART therapy. It is Instead those genetic variations already identified in genes of and influencing this pathway as well as variants yet to be identified outside of it that influence the functioning of cholesterol metabolism and serum lipid levels.

To identify other possible factors that could play a role in lipid dysfunction, we analyzed the whole transcriptomes of individuals with lipid levels that fell within the extreme ends of the range. In doing so, we identified 4 top transcripts (HDC, CPA3, GATA2, & SLC45A3) that were differentially expressed for both HDL-C and triglyceride comparisons. The top transcript HDC, histadine decarboxylase, initially did not seem to be associated with lipid levels but further investigation of the literature led to the discovery that this gene and the product of its catalytic activity histamine have been extensively linked to atherosclerosis. As HDC levels have been associated with progression of atherosclerosis in the general population, it is likely that a similar

association will be identified in those infected with HIV-1. More research will have to be conducted on this gene in the MACS to determine if we identified a marker for increased risk of atherosclerosis in the HIV-infected population on therapy. At the very least, we have found possible markers for serum HDL and triglyceride variation as changes in both were associated with expression differences in HDC and the other genes identified. Additionally, this evidence indicates that the use of whole blood as a tissue surrogate is capable of identifying relevant genes for the analysis of lipid level variation. This was further confirmed by the results from viral load comparisons yielding differentially expressed genes involved in viral life cycle control and immune response against HIV-1.

We further examined the role of copy number variation within lipid metabolism by investigating how CNV can hinder the ability to genotype a single nucleotide polymorphism. As over 50% of the genetic variation for altered HDL-C and LDL-C levels are associated with SNP haplotypes within the reverse cholesterol transport pathway, it stands to reason that at some point genetic testing to type these variants will be involved in therapy[30]. Had CNV been present within these genes, it had the possibility to alter the effectiveness of risk assessment using these haplotypes based on our study's findings.

We observed that when a gene falls within a region of CNV, SNP genotyping requires a different approach to accurately quantify the amount of each allele present. When SNPs with no CNV were analyzed, three possible genotypes were present and clearly observed by plotting the data. The distinct groups of genotypes were lost when analyzing SNPs in the presence of CNV. Instead, we observed a spread of values for each allele of the SNP and were unable to identify

the edges of genotype groupings in the plotted data. This was observed for all SNP genotyping assays excluding sequencing. We also identified that within FP, altering assay conditions could result in loss of the ability to observe the range for each allele present in the CNV background. This would result in inaccurate genotyping for that individual and improper risk assessment.

7.5 PUBLIC HEALTH SIGNIFICANCE

Currently, the World Health Organization indicated that in 2013 approximately 35 million individuals across the globe were living with HIV-1. As the incidence of new infections remain higher than the number of those dying of dying of AIDS, this number will continue to grow. And, with antiretroviral therapy becoming more available to regions of the world that originally had little access to therapy, the incidence of therapy related dyslipidemia will increase. Therefore it is important to identify genetic factors that make some individuals more likely to experience adverse lipid altering effects of antiviral therapy.

In this study we identified that copy number variation of reverse cholesterol transport pathway genes does not play a role in the dyslipidemia associated with antiretroviral therapy. In doing so, focus is redirected towards the already identified variants within and influencing the RCT pathway along with other unknown variants yet to be identified.

During our transcriptome analysis, we identified a few genes whose expression levels could serve as potential markers for HDL and Triglyceride levels. In particular, *HDC* shows the most promise due to its extensive role in atherosclerosis in the general population. If further

investigation indicates an association between *HDC* gene expression and progression of atherosclerosis in the HIV-infected population then monitoring levels of this gene will allow physicians to alter treatment plans to reduce plaque buildup before it results in clinical manifestations. Depending on the future identification of how these genes play a role in lipid metabolism, it is possible that we have identified future targets of therapy for HIV-1 infected individuals with dyslipidemia.

In addition to typing the amount of CNV among the reverse cholesterol transport pathway genes, we also illustrated that copy number variation can inhibit the ability to properly genotype a single nucleotide polymorphism. With the growing interest of genetics among the general population and the increased use of polymorphisms for commercially available genetic testing products such as 23 and me, it is paramount that SNP testing is accurate. When a researcher or clinician (In the next 10yrs we will see genotyping in the medical sector.) is unaware that the SNP(s) they are typing falls within a block of CNV, they are likely to genotype the individuals tested into three distinct genotype groups rather the range of genotypes are present. In doing so, overestimations will cause undue anxiety for some patients while underestimations will give the false sense of security to those that need to pay closer attention to their condition. Basically, if the SNP genotyping being performed for risk assessment is not accurate then the test in general is irrelevant for the patient. However with more accurate SNP genotyping in place, physicians and patients may be able to identify risk factors early enough to alter the patient's lifestyle and treatment plan to prevent the unwanted outcome.

APPENDIX: PRIMERS AND PROBES

A.1 MLPA PROBE OLIGOS

Oligo Size (nt)	Sequence Name	Sequence[1]	% GC	Tm °C [50mM NaCl]	5' Phosphoryl -ation
46	SRBI_LPO	GGG TTC CCT AAG GGT TGG AAG TGG CCG TCT TGG GCT GGG CGT GTC T	63	74	N/A
50	SRBI_RPO	TCC TGC CTT CAC ACC ACT CGG CCC CAA TCT AGA TTG GAT CTT GCT GGC AC	56	72	In lab
48	APOC3_LPO	GGG TTC CCT AAG GGT TGG AGA AGC ACG CCA CCA AGA CCG CCA AGG ATG	60	73	N/A
52	APOC3_RPO	CAC TGA GCA GCG TGC AGG AGT CCC AGG TGT CTA GAT TGG ATC TTG CTG GCA C	58	72	In lab
50	APOA1_LPO	GGG TTC CCT AAG GGT TGG AGG CGG GGC AGG GGT GTT GGT TGA GAG TGT AC	62	73	N/A
54	APOA1_RPO	/5Phos/TGG AAA TGC TAG GCC ACT GCA CCT CCG CGG ATC TAG ATT GGA TCT TGC TGG CAC	56	72	IDT
54	APOE_LPO	GGG TTC CCT AAG GGT TGG ACA GGA AGA TGA AGG TTC TGT GGG CTG CGT TGC TGG	57	72	N/A

[1] Contains oligo sequence only without the universal primer sequences that would be located at the 5' end of the LPO and 3' end of the RPO.

58	APOE_RPO	TCA CAT TCC TGG CAG GTA TGG GGG CGG GGC TTG CTT CTA GAT TGG ATC TTG CTG GCA C	57	73	In lab
56	PLTP_LPO	GGG TTC CCT AAG GGT TGG AGA GTA GGA ATG CAG AGG GCG GAA GGG AGG GCA TCA GT	59	73	N/A
60	PLTP_RPO	AAG CCG ATG GAT GTG GGG ATG CTC AGA GTG GGT TTG ATC TAG ATT GGA TCT TGC TGG CAC	52	71	In lab
58	LIPC_LPO	GGG TTC CCT AAG GGT TGG ATC GGA GGC AGG TCC AGA GAC TTC GGT TCC TGG TGA TTT A	55	72	N/A
62	LIPC_RPO	AAC AGC CCC TAG TCA AGA GCA TGG CAC ACA ACA GAT GTT TCT AGA TTG GAT CTT GCT GGC AC	48	71	In lab
60	LCAT_LPO	GGG TTC CCT AAG GGT TGG AGA TGT GGT GAA CTG GAT GTG CTA CCG CAA GAC AGA GGA CTT	53	72	N/A
64	LCAT_RPO	CTT CAC CAT CTG GCT GGA TCT CAA CAT GTT CCT ACC CCT TGT CTA GAT TGG ATC TTG CTG GCA C	50	70	In lab
64	APOA4_LPO	GGG TTC CCT AAG GGT TGG AGG CGA GTG GTA TAC AAG CAG ACA AAG TCT TGC CGT GTA AAT GCC A	52	72	N/A
68	APOA4_RPO	AAT GTA ACG TGG CCT CCT TGT GCC CTT CCC CAC AGT GCC CTC TTC TCT AGA TTG GAT CTT GCT GGC AC	54	73	In lab
66	LPL_LPO	GGG TTC CCT AAG GGT TGG ACA AAA TAG CAG ATG TCA CTG AAG GAG AGC TCA GCG AGG GAG TGA TTG	52	71	N/A
70	LPL_RPO	/5Phos/ATT AAT AGC TGT ATT GAA AGG TGG GAG TCA GGT ACG GGG GAA GAG CGT CTA GAT TGG ATC TTG CTG GCA C	49	71	IDT
68	LIPG_LPO	GGG TTC CCT AAG GGT TGG AGA AAT GCC CAT GTA TGT GGA GCT AAG TGA GAC AGA GGG GTT GTC ATG CT	51	72	N/A

72	LIPG_RPO	TCA CTA TCC CCT TGT CCC ATG CTG CAA TCC GTT ATT TCA GAC GTG AGG ATC TAG ATT GGA TCT TGC TGG CAC	49	71	In lab
70	LDLR_LPO	GGG TTC CCT AAG GGT TGG AGG CTT ACG TAC GAG ATG CAA GCA CTT AGG TGG CGG ATA GAC ACA GAC TAT A	51	71	N/A
74	LDLR_RPO	GAT CAC TCA AGC CAA GAT GAA CGC AGA AAA CTG GTT GTG ACT AGG AGG AGG TCT AGA TTG GAT CTT GCT GGC AC	49	71	In lab
74	CETP_LPO	GGG TTC CCT AAG GGT TGG ATC TCA CCA CCT CTG CTG GCA CTG GTT GTC TCT TGC ACA TGG CTC CTT ACA ATC AA	53	73	N/A
78	CETP_RPO	AAT CAC ATC ATG CAA GTA ACG AGG GGG TAC ACA CGT GGT TTC CAC AGC TTA GGT ATC TAG ATT GGA TCT TGC TGG CAC	47	71	In lab
76	APOA5_LPO	GGG TTC CCT AAG GGT TGG AGA GGA CGC CCG CTG CAG TCC CCA GAA TCA AAG GAT GAT GTG GCG CAT CTA TGT TTC T	55	74	N/A
80	APOA5_RPO	/5Phos/TTG GAG AGT GTT GTA GGT CTG GAT TTG TAT GGG CAA TGT GTT TGT GCT TCG TGC GTG TCT AGA TTG GAT CTT GCT GGC AC	48	72	IDT
78	APOB_LPO	GGG TTC CCT AAG GGT TGG AGA GCA AGG GTT CAC TGT TCC TGA AAT CAA GAC CAT CCT TGG GAC CAT GCC TGC CTT TGA	53	73	
82	APOB_RPO	/5Phos/AGT CAG TCT TCA GGC TCT TCA GAA AGC TAC CTT CCA GAC ACC TGA TTT TAT AGT CCC CCT CTA GAT TGG ATC TTG CTG GCA C	48	71	IDT
80	ABCA1_LPO	GGG TTC CCT AAG GGT TGG ATT TCC AGA ACT TGG CTC CAG TCT GGT TGC TCG CCA TGA AGC ACT TAC AGA TAA ACC TCA TC	50	72	N/A
84	ABCA1_RPO	TTG GGC CAG TGC TTC CAT TTA CTG TCT CCT TTT GGC TTG CTT ATC CTT CCT TCT GCC TTC TTC TAG ATT GGA TCT TGC TGG CAC	48	72	In lab
82	APOC2_LPO	GGG TTC CCT AAG GGT TGG ACT GCC GTA CTT CCT CAT CTC CTA CGT GTG GAT GAT GAT ATT GTG CCC TGT GCA TGT TCT TCG T	51	72	N/A

86	APOC2_RPO	CAC CAA AAG TGC CTC TCT CAT AGA GCA GGT GAG AAC TCA GTG AGG AGA TGC AGG GAC ATG AGG TCT AGA TTG GAT CTT GCT GGC AC	51	72	In lab
57	DEFB103_RPO (MRC)	/5Phos/CAG ATC GGC AAG TGC TCG ACG CGT GGC CGA AAA TTC TAG ATT GGA TCT TGC TGG CAC	54	72	IDT
43	DEFB103_LPO (MRC)	GGG TTC CCT AAG GGT TGG ACT CAG CTG CCT TCC AAA GGA GGA A	56	70	N/A
50	CCR5_LPO	GGG TTC CCT AAG GGT TGG ACA TTA CAC CTG CAG CTC TCA TTT TCC ATA CA	48	69	N/A
54	CCR5_RPO	/5Phos/GTC AGT ATC AAT TCT GGA AGA ATT TCC AGA CTC TAG ATT GGA TCT TGC TGG CAC	43	66	IDT
62	CCR5_d32_RPO	/5Phos/TTA AAG ATA GTC ATC TTG GGG CTG GTC CTG CCG CTG CTT TCT AGA TTG GAT CTT GCT GGC AC	50	71	IDT

A.2 TAQMAN EXPRESSION AND SNP ASSAYS

ABI Catalog #	Gene/SNP	Assay Type	Custom Design	Primers
Hs00157914_m1	HDC	Expression	No	N/A
Hs00157019_m1	CPA3	Expression	No	N/A
Hs00233520_m1	CD8A	Expression	No	N/A
Hs00174762_m1	CD8B	Expression	No	N/A
Hs01059118_m1	ABCA1	Expression	No	N/A
4326317E	GAPDH	Expression	No	N/A
C___3219470_10	rs2373961	SNP Genotyping	No	N/A
C___29961120_10	rs7037117	SNP Genotyping	No	N/A
N/A	rs1828283	SNP Genotyping	**Yes**	**See A.3**
N/A	rs2220067	SNP Genotyping	**Yes**	**See A.3**
C____1496696_1_	rs2221072	SNP Genotyping	No	N/A
C___3275769_10	rs917015	SNP Genotyping	No	N/A
C___3201533_10	rs4352264	SNP Genotyping	No	N/A

Custom probe assays designed and made in our lab were mixed at a final concentration of 900nM primers and 200nM probes

A.3 PCR PRIMERS

SNP/Gene	PCR Type	Primer Type	Primer Name	Primer Sequence
rs2220067	FP	PCR F	**rs2220067_PF_new**	CAC CCA ATA CAG GAG CAC AC
rs2220067	FP	Internal R	**rs2220067 FP_R**	TGG TCA ATT TTA GAA GAA GTG CTA C
rs2220067	FP	Internal F	**rs2220067 FP_F**	GAG AAT ACA CAT TCT TCT CAG TGC C
rs2220067	FP	PCR R	**rs2220067 P_R**	TTG AAT GCA CTG TGG TCA GA
rs2472240	FP	PCR R	**rs2472240_PR_2**	CCA ATC TGA CCT CTG ACT GTG GG
rs2472240	FP	PCR F	**rs2472240_PF_2**	CAG CAA AGG AAA CAA TCA ACC G
rs2472240	FP	Internal F	**rs2472240_F_IF**	CAA AAA TCA AAC AAC CTG ATT GAA A
rs2472240	FP	Internal R	**rs2472240_F_IR**	AAA CCT ATT TAG GTC CTT GGT CTG T
rs3789864	FP	PCR F	**rs3789864_P_F_DEFB103**	CAA TTC TCT GCC TCA GCC TC
rs3789864	FP	PCR R	**rs3789864_P_R_DEFB103**	GGA CCA AGC AGG TTT GTT GT
rs3789864	FP	Internal F	**rs3789864_F_IF_DEFB103**	TAC AGG TGC CCG CCA CTG TGC CCA G
rs3789864	FP	Internal R	**rs3789864_F_IR_DEFB103**	CAT CTC TAC TAA AAA TAC AAA AAT T
rs2373961	FP	Internal F	**rs2373961 FP IF**	CTT TAC GGC AGG CTC AGC AGA AAA C
rs2373961	FP	Internal R	**rs2373961 FP IR**	TCT TGT CAC CTG GGC TGC CCT CCC T
rs2373961	FP	PCR F	**rs2373961 PCR F**	TAC TCA ATG CCT CCA GCC AGG TTG
rs2373961	FP	PCR R	**rs2373961 PCR R**	CAG AAA CTC CCA AAG GAA ATC CC
rs4352264	FP	Internal F	**FP_F_rs4352264**	ACC GCT TGA AGT CAT GGA AAC AAG A
rs4352264	FP	Internal R	**FP_R_rs4352264**	TTC ACA AAT AAA AGT TTA TTG TTGA

rs4352264	FP	PCR F	**P_F_rs4352264**	GCA ACC GCT TGA AGT CAT GG
rs4352264	FP	PCR R	**P_R_rs4352264**	TGG GGC TTT AGG TTC TTG CAC
rs7037117	FP	Internal F	**FP_F_rs7037117**	TCG GTT CCT TGA TCT TGT GTC TCC A
rs7037117	FP	Internal R	**FP_R_rs7037117**	AGA TGT AAG AGA GAG AGC AAG TGA T
rs7037117	FP	PCR F	**P_F_rs7037117**	TGG TTT AGT CTG GGC TGT TAG CG
rs7037117	FP	PCR R	**P_R_rs7037117**	AAA AGT GAG AGT TTG GGA CCT GC
rs2221072	FP	PCR F	**FABP3 PCRF**	GGC TTG GCT GAA AGA GCA GTA GTA AT
rs2221072	FP	PCR R	**FABP3 PCRR**	TTC CCC AGA AAG GCA GTA GTG G-3'
rs2221072	FP	Internal R	**FABP3 FP IR**	TAG TTT GGG TCA AAG GCT GTG T
rs2221072	FP	Internal F	**FABP3 FP IF**	GTC TGG ACA CTG GGC CAC AGA G
rs917015	FP	PCR F	**CCL16_2_PCRf**	TGT TTT TAC CCC CAT AGA GCC C
rs917015	FP	PCR R	**CCL16_2_PCRr**	CCC ACC ATT TGT GTT TCA CTC C
rs917015	FP	Internal F	**CCL16_2_I_f**	CGG TTC CTT GGC AAG TGT GAA TAA C
rs917015	FP	Internal R	**CCL16_2_I_r**	GCT GAG TGT CAA CTA CAA ATG ACT T
rs1828283	FP	Internal F	**rs1828283 FP IF**	GAC CTA GGG TGA GCT GGA GAG TGA A
rs1828283	FP	Internal R	**rs1828283 FP IR**	TGC TTA CTT CCC AGT GGG GTC TGT T
rs1828283	FP	PCR R	**rs1828283 PCR R**	CCC GAA GAG AAA AGA AGG AAG TTC
rs1828283	FP	PCR F	**rs1828283 PCR F**	AGA GAA TAA GCC CGA GTC ACA GC
rs6703462	FP	PCR F	**P_F_rs6703462**	AAG AAC ATA GGC TCT GGC ACC TC
rs6703462	FP	PCR R	**P_R_rs6703462**	AAA AAA TCT CCC CTT GAC CCT G
rs6703462	FP	Internal F	**FP_IF_rs6703462**	AGC AGA TTA GGG AAG GAA TAT AGG C
rs6703462	FP	Internal R	**FP_IR_rs6703462**	CTC TGT ATC ATT CTC TAC ATT TCT T

APOA4	SEQ	Reverse	**APOA4_R_SEQ**	CCT TCC CAA TCT CCT CCT TC
APOA4	SEQ	Forward	**APOA4_F_SEQ**	ACC TCA AAG TCC CAC CCT CT
CETP	SEQ	Reverse	**CETP_R_SEQ**	TGG TGG TGT TTG TCT GTG GT
CETP	SEQ	Forward	**CETP_F_SEQ**	CCC TGT CTT CCA CAG GTT GT
SRBI	SEQ	Reverse	**SRBI_R_SEQ**	AGG GCC AAC TGT AGG GAC TT
SRBI	SEQ	Forward	**SRBI_F_SEQ**	AGT GTG GGG ACT TAT GCC AG
ABCA1	SEQ	Reverse	**ABCA1_R_SEQ**	TCG AGG AAC TTT CAA GGC TG
ABCA1	SEQ	Forward	**ABCA1_F_SEQ**	CCA GCA ACA TAG GGG AGA AG
rs1828283	TaqMan_SNP	Forward	**rs1828283_F_TM**	TAA GAC ATC CAA GGG ACA GG
rs1828283	TaqMan_SNP	Reverse	**rs1828283_R_TM**	AGC CCC GAA GAG AAA AGA A
rs1828283	TaqMan_SNP	Probe 1	**rs1828283_ 1 probe**	6FAMTGGAGAGTGAAGAACAGMGBNFQ
rs1828283	TaqMan_SNP	Probe 2	**rs 1828283 _2probe**	VICTGGAGAGTGAACAACAGMGBNFQ
rs2220067	TaqMan_SNP	Forward	**rs2220067_F_TM**	TCC ACT CCA AAT TAG AGA ATA CAC ATT C
rs2220067	TaqMan_SNP	Reverse	**rs2220067_R_TM**	TTG CAT TTG CTG AGA AGT GTT TTA C
rs2220067	TaqMan_SNP	Probe 1	**rs2220067 1 probe**	6FAMTCTCAGTGCCATGT AGMGBNFQ
rs2220067	TaqMan_SNP	Probe 2	**rs2220067 2probe**	VICTCTCAGTGCCACGTAGMGBNFQ

BIBLIOGRAPHY

1. Gilbert PB, McKeague IW, Eisen G, Mullins C, Gu ye-NDiaye A, Mboup S, Kanki PJ: **Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal**. *Statist Med* 2003, **22**:573–593.

2. Reeves JD, Doms RW: **Human immunodeficiency virus type 2**. *Journal of General Virology* 2002, **83**:1253–1265.

3. Sharp PM, Hahn BH: **Origins of HIV and the AIDS Pandemic**. *Cold Spring Harbor Perspectives in Medicine* 2011, **1**:a006841–a006841.

4. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM: **Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes**. *Nature* 1999, **397**:436–441.

5. Keele BF: **Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1**. *Science* 2006, **313**:523–526.

6. Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F: **Breif Communications**. *Nat Med* 2009:1–2.

7. Hirsch VM, Olmstead RA, MurpheyCorb M, Purcell RH, Johnson PR: **An African Primate Lentivirus (Sivsm) Closely Related to Hiv-2**. *Nature* 1989, **339**:389–392.

8. Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, Greene BM, Sharp PM, Shaw GM, Hahn BH: **Human Infection by Genetically Diverse Sivsm-Related Hiv-2 in West Africa**. *Nature* 1992, **358**:495–499.

9. Grunfeld C, Pang M, Doerrler W, Shigenaga J, Jensen P, Feingold K: **Lipids, lipoproteins, triglyceride clearance, and cytokines in human immunodeficiency virus infection and the acquired immunodeficiency syndrome.** *Journal of Clinical Endocrinology & Metabolism* 1992, **74**:1045–1052.

10. Gordon DJD, Probstfield JLJ, Garrison RJR, Neaton JDJ, Castelli WPW, Knoke JDJ, Jacobs DRD, Bangdiwala SS, Tyroler HAH: **High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies.** *Circulation* 1989, **79**:8–15.

11. Joshi VV, Pawel B, Connor E, Sharer L, Oleske JM, Morrison S, Marin-Garcia J: **Arteriopathy in children with acquired immune deficiency syndrome.** *Pediatr Pathol* 1987, **7**:261–275.

12. Kabus D, Greco MA: **Arteriopathy in children with AIDS: microscopic changes in the vasa vasorum with gross irregularities of the aortic intima.** *Pediatr Pathol* 1991, **11**:793–795.

13. Safrin S, Grunfeld C: **Fat distribution and metabolic changes in patients with HIV infection.** *AIDS* 1999, **13**:2493–2505.

14. Gervasoni CC, Ridolfo ALA, Trifirò GG, Santambrogio SS, Norbiato GG, Musicco MM, Clerici MM, Galli MM, Moroni MM: **Redistribution of body fat in HIV-infected women undergoing combined antiretroviral therapy.** *AIDS* 1999, **13**:465–471.

15. Carr A, Samaras K, Burton S, Law M, Freund J, Chisholm DJ, Cooper DA: **A syndrome of peripheral lipodystrophy, hyperlipidaemia and insulin resistance in patients receiving HIV protease inhibitors.** *AIDS* 1998, **12**:F51–8.

16. Sam S, Haffner S, Davidson MH, D'Agostino RB, Feinstein S, Kondos G, Perez A, Mazzone T: **Hypertriglyceridemic Waist Phenotype Predicts Increased Visceral Fat in Subjects With Type 2 Diabetes**. *Diabetes Care* 2009, **32**:1916–1920.

17. Arsenault BJ, Lemieux I, Despres JP, Wareham NJ, Kastelein JJP, Khaw KT, Boekholdt SM: **The hypertriglyceridemic-waist phenotype and the risk of coronary artery disease: results from the EPIC-Norfolk Prospective Population Study**. *Canadian Medical Association Journal* 2010, **182**:1427–1432.

18. Janiszewski PM, Ross R, Despres J-P, Lemieux I, Orlando G, Carli F, Bagni P, Menozzi M, Zona S, Guaraldi G: **Hypertriglyceridemia and Waist Circumference Predict Cardiovascular Risk among HIV Patients: A Cross-Sectional Study**. *PLoS ONE* 2011, **6**:e25032.

19. Hadigan C, Meigs JB, Corcoran C, Rietschel P, Piecuch S, Basgoz N, Davis B, Sax P, Stanley T, Wilson PW, D'Agostino RB, Grinspoon S: **Metabolic abnormalities and cardiovascular disease risk factors in adults with human immunodeficiency virus infection and lipodystrophy.** *Clin Infect Dis* 2001, **32**:130–139.

20. Friis-Møller N, Sabin CA, Weber R, Monforte AD, El-Sadr WM, Reiss P, Thiebaut R, Morfeldt L, de Wit S, Pradier C, Calvo G, Law MG, Kirk O, Phillips AN, Lundgren JD, Lundgren JD, Weber R, Monteforte AD, Bartsch G, Reiss P, Dabis F, Morfeldt L, de Wit S, Pradier C, Calvo G, Law MG, Kirk O, Phillips AN, Houyez F, Loeliger E, et al.: **Combination antiretroviral therapy and the risk of myocardial infarction**. *N Engl J Med* 2003, **349**:1993–2003.

21. Friis-Møller N, Weber R, Reiss P, Thiébaut R, Kirk O, D'Arminio Monforte A, Pradier C, Morfeldt L, Mateu S, Law M, El-Sadr W, De Wit S, Sabin CA, Phillips AN, Lundgren JD, D:A:D Study Group: **Cardiovascular disease risk factors in HIV patients--association with antiretroviral therapy. Results from the DAD study.** *AIDS* 2003, **17**:1179–1193.

22. Asztalos BF, Schaefer EJ, Horvath KV, Cox CE, Skinner S, Gerrior J, Gorbach SL, Wanke C: **Protease inhibitor-based HAART, HDL, and CHD-risk in HIV-infected patients**. *Atherosclerosis* 2006, **184**:72–77.

23. Duprez DA, Kuller LH, Tracy R, Otvos J, Cooper DA, Hoy J, Neuhaus J, Paton NI, Friis-Møller N, Lampe F, Liappis AP, Neaton JD: **Lipoprotein particle subclasses, cardiovascular disease and HIV infection**. *Atherosclerosis* 2009, **207**:524–529.

24. Anastos KK, Lu DD, Shi QQ, Tien PCP, Kaplan RCR, Hessol NAN, Cole SS, Vigen CC, Cohen MM, Young MM, Justman JJ: **Association of serum lipid levels with HIV serostatus, specific antiretroviral agents, and treatment regimens.** *J Acquir Immune Defic Syndr* 2007, **45**:34–42.

25. Nicholaou MJ, Martinson JJ, Abraham AG, Brown TT, Hussain SK, Wolinsky SM, Kingsley LA: **HAART-Associated Dyslipidemia Varies by Biogeographical Ancestry in the Multicenter AIDS Cohort Study**. *AIDS Research and Human Retroviruses* 2013:130308073730004.

26. Johnson JL, Slentz CA, Duscha BD, Samsa GP, McCartney JS, Houmard JA, Kraus WE: **Gender and racial differences in lipoprotein subclass distributions: the STRRIDE study**. *Atherosclerosis* 2004, **176**:371–377.

27. Kuller LH: **Ethnic differences in atherosclerosis, cardiovascular disease and lipid metabolism**. *Curr Opin Lipidol* 2004, **15**:109–113.

28. D'Adamo E, Northrup V, Weiss R, Santoro N, Pierpont B, Savoye M, O'Malley G, Caprio S: **Ethnic differences in lipoprotein subclasses in obese adolescents: importance of liver and intraabdominal fat accretion**. *American Journal of Clinical Nutrition* 2010, **92**:500–508.

29. Morabia A: **Association of extreme blood lipid profile phenotypic variation with 11 reverse cholesterol transport genes and 10 non-genetic cardiovascular disease risk factors**. *Human Molecular Genetics* 2003, **12**:2733–2743.

30. Knoblauch H: **Haplotypes and SNPs in 13 lipid-relevant genes explain most of the genetic variance in high-density lipoprotein and low-density lipoprotein cholesterol**. *Human Molecular Genetics* 2004, **13**:993–1004.

31. Dedoussis GVZ, Schmidt H, Genschel J: **LDL-receptor mutations in Europe**. *Hum Mutat* 2004, **24**:443–459.

32. Garcia CK: **Autosomal Recessive Hypercholesterolemia Caused by Mutations in a Putative LDL Receptor Adaptor Protein**. *Science* 2001, **292**:1394–1398.

33. Marduel M, Carrié A, Sassolas A, Devillers M, Carreau V, Di Filippo M, Erlich D, Abifadel M, Marques-Pinheiro A, Munnich A, Junien C, Boileau C, Varret M, Rabès J-P: **Molecular Spectrum of Autosomal Dominant Hypercholesterolemia in France**. *Hum Mutat* 2010, **31**:E1811–E1824.

34. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH: **Sequence variations in PCSK9, low LDL, and protection against coronary heart disease.** *N Engl J Med* 2006, **354**:1264–1272.

35. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH: **Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9**. *Nature Genetics* 2005, **37**:161–165.

36. Guella II, Asselta RR, Ardissino DD, Merlini PAP, Peyvandi FF, Kathiresan SS, Mannucci PMP, Tubaro MM, Duga SS: **Effects of PCSK9 genetic variants on plasma LDL cholesterol levels and risk of premature myocardial infarction in the Italian population.** *The Journal of Lipid Research* 2010, **51**:3342–3349.

37. Benn M, Nordestgaard BG, Grande P, Schnohr P, Tybjaerg-Hansen A: **PCSK9 R46L, Low-Density LipoproteinCholesterol Levels, and Risk of Ischemic Heart Disease**. *JAC* 2010, **55**:2833–2842.

38. Abifadel M, Varret M, Rabès J-P, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, Derre A, Villeger L, Farnier M, Beucler I, Bruckert E, Chambaz J, Chanu B, Lecerf JM, Luc G, Moulin P, Weissenbach J, Prat A, Krempf M, Junien C, Seidah NG, Boileau C: **Mutations in PCSK9 cause autosomal dominant hypercholesterolemia**. *Nature Publishing Group* 2003, **34**:154–156.

39. Hollox EJE, Armour JALJ, Barber JCKJ: **Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster.** *The American Journal of Human Genetics* 2003, **73**:591–600.

40. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET: **Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes**. *Science* 2007, **315**:848–853.

41. Groth M, Wiegand C, Szafranski K, Huse K, Kramer M, Rosenstiel P, Schreiber S, Norgauer J, Platzer M: **Both copy number and sequence variations affect expression of human DEFB4**. *Genes Immun* 2010, **11**:458–466.

42. Townson JR, Barcellos LF, Nibbs RJ: **Gene copy number regulates the production of the human chemokine CCL3- L1**. *Eur J Immunol* 2002, **32**:3016–3026.

43. Lanktree M, Hegele RA: **Copy number variation in metabolic phenotypes**. *Cytogenet Genome Res* 2008, **123**:169–175.

44. Wang J, Ban MR, Hegele RA: **Multiplex ligation-dependent probe amplification of LDLR enhances molecular diagnosis of familial hypercholesterolemia.** *The Journal of Lipid Research* 2005, **46**:366–372.

45. Chmara MM, Wasag BB, Zuk MM, Kubalska JJ, Wegrzyn AA, Bednarska-Makaruk MM, Pronicka EE, Wehr HH, Defesche JCJ, Rynkiewicz AA, Limon JJ: **Molecular characterization of Polish patients with familial hypercholesterolemia: novel and recurrent LDLR mutations.** *J Appl Genet* 2010, **51**:95–106.

46. Goldmann R, Tichý L, Freiberger T, Zapletalová P, Letocha O, Soška V, Fajkus J, Fajkusová L: **Genomic characterization of large rearrangements of the LDLR gene in Czech patients with familial hypercholesterolemia**. *BMC Med Genet* 2010, **11**:115.

47. Futema M, Plagnol V, Whittall RA, Neil HAW, on behalf of the Simon Broome Register Group, Humphries SE, UK10K: **Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia**. *Journal of Medical Genetics* 2012, **49**:644–649.

48. Tosi I, Toledo-Leiva P, Neuwirth C, Naoumova RP, Soutar AK: **Genetic defects causing familial hypercholesterolaemia: Identification of deletions and duplications in the LDL-receptor gene and summary of all mutations found in patients attending the Hammersmith Hospital Lipid Clinic**. *Atherosclerosis* 2007, **194**:102–111.

49. Holla ØL, Teie C, Berge KE, Leren TP: **Identification of deletions and duplications in the low density lipoprotein receptor gene by MLPA**. *Clinica Chimica Acta* 2005, **356**:164–171.

50. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW: **The Database of Genomic Variants: a curated collection of structural variation in the human genome**. *Nucleic Acids Research* 2013, **42**:D986–D992.

51. National Cholesterol Education Program NCEP Expert Panel on Detection EATOHBCIAATPI: **Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report.** *Circulation* 2002, **106**:3143–3421.

52. Fode P, Jespersgaard C, Hardwick RJ, Bogle H, Theisen M, Dodoo D, Lenicek M, Vitek L, Vieira A, Freitas J, Andersen PS, Hollox EJ: **Determination of beta-defensin genomic copy number in different populations: a comparison of three methods.** *PLoS ONE* 2011, **6**:e16768.

53. Groth M, Szafranski K, Taudien S, Huse K, Mueller O, Rosenstiel P, Nygren AO, Schreiber S, Birkenmeier G, Platzer M: **High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes**. *Hum Mutat* 2008, **29**:1247–1254.

54. Armour JAL, Palla R, Zeeuwen PLJM, Heijer MD, Schalkwijk J, Hollox EJ: **Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats**. *Nucleic Acids Research* 2007, **35**:e19–e19.

55. Zhi J: **MAPD: a probe design suite for multiplex ligation-dependent probe amplification assays.** *BMC Res Notes* 2010, **3**:137–137.

56. Zhi J, Hatchwell E: **Human MLPA Probe Design (H-MAPD): a probe design tool for both electrophoresis-based and bead-coupled human multiplex ligation-dependent probe amplification assays**. *BMC Genomics* 2008, **9**:407.

57. R Core Team: *R: a Language and Environment for Statistical Computing*. Vienna, Austria; 2012.

58. Wickham H: *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York; 2009.

59. Wickham H: **Reshaping data with the reshape package**. *Journal of Statistical Software* 2007, **21**:1–20.

60. Auguie B: **gridExtra: functions in Grid graphics. R package version 0.9.1.** 2012

61. MRC-Holland: **Interpretation of MLPA Results.** 2010:1-7.

62. Field SF, Howson JMM, Maier LM, Walker S, Walker NM, Smyth DJ, Armour JAL, Clayton DG, Todd JA: **Experimental aspects of copy number variant assays at CCL3L1**. *Nat Med* 2009, **15**:1115–1117.

63. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.** *Bioinformatics* 2009, **25**:415–416.

64. Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing Illumina microarray**. *Bioinformatics* 2008, **24**:1547–1548.

65. Riddler SA: **Impact of HIV Infection and HAART on Serum Lipids in Men**. *JAMA: The Journal of the American Medical Association* 2003, **289**:2978–2982.

66. Fontas E, van Leth F, Sabin CA, Friis-Møller N, Rickenbach M, d'Arminio Monforte A, Kirk O, Dupon M, Morfeldt L, Mateu S, Petoumenos K, El-Sadr W, de Wit S, Lundgren JD, Pradier C, Reiss P, D:A:D Study Group: **Lipid profiles in HIV-infected patients receiving combination antiretroviral therapy: are different antiretroviral drugs associated with different lipid profiles?** *J Infect Dis* 2004, **189**:1056–1074.

67. Riddler SA, Li X, Otvos J, Post W, Palella F, Kingsley L, Visscher B, Jacobson LP, Sharrett AR: **Antiretroviral therapy is associated with an atherogenic lipoprotein phenotype among HIV-1-infected men in the multicenter AIDS cohort study**. *J Acquir Immune Defic Syndr* 2008, **48**:281–288.

68. Worm SW, Kamara DA, Reiss P, Kirk O, El-Sadr W, Fux C, Fontas E, Phillips A, D'Arminio Monforte A, De Wit S, Petoumenos K, Friis-Mller N, Mercie P, Lundgren JD, Sabin C: **Elevated triglycerides and risk of myocardial infarction in HIV-positive persons**. *AIDS* 2011, **25**:1497–1504.

69. Friis-Møller N, Reiss P, Sabin CA, Weber R, Monforte AD, El-Sadr W, De Wit S, Kirk O, Fontas E, Law MG, Phillips A, Lundgren JD, Grp DS: **Class of antiretroviral drugs and the risk of myocardial infarction**. *N Engl J Med* 2007, **356**:1723–1735.

70. Egaña-Gorroño L, Martínez E, Cormand B, Escribà T, Gatell J, Arnedo M: **Impact of genetic factors on dyslipidemia in HIV-infected patients starting antiretroviral therapy**. *AIDS* 2013, **27**:529–538.

71. Raman K, Chong M, Akhtar-Danesh GG, D'Mello M, Hasso R, Ross S, Xu F, Paré G: **Genetic Markers of Inflammation and Their Role in Cardiovascular Disease**. *CJCA* 2013, **29**:67–74.

72. McPherson R: **From Genome-Wide Association Studies to Functional Genomics: New Insights Into Cardiovascular Disease**. *CJCA* 2013, **29**:23–29.

73. Dubé JB, Hegele RA: **Genetics 100 for Cardiologists: Basics of Genome-Wide Association Studies**. *CJCA* 2013, **29**:10–17.

74. Roberts R, Stewart AFR: **Genes and Coronary Artery Disease**. *JAC* 2012, **60**:1715–1721.

75. Sawhney V, Brouilette S, Abrams D, Schilling R, O'Brien B: **Current genomics in cardiovascular medicine.** *Curr Genomics* 2012, **13**:446–462.

76. Swerdlow DI, Holmes MV, Harrison S, Humphries SE: **The genetics of coronary heart disease**. *British Medical Bulletin* 2012, **102**:59–77.

77. Patel RS, Ye S: **Genetic determinants of coronary heart disease: new discoveries and insights from genome-wide association studies**. *Heart* 2011, **97**:1463–1473.

78. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardissino D, Merlini PA, Berzuini C, et al.: **Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants**. *Nature Publishing Group* 2009, **41**:334–341.

79. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen W-M, Li Y, Scott LJ, Scheet PA, et al.: **Newly identified loci that influence lipid concentrations and risk of coronary artery disease**. *Nature Genetics* 2008, **40**:161–169.

80. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, Hirschhorn JN, Berglund G, Hedblad B, Groop L, Altshuler DM, Newton-Cheh C, Orho-Melander M: **Polymorphisms associated with cholesterol and risk of cardiovascular events.** *N Engl J Med* 2008, **358**:1240–1249.

81. Lopez D: **PCSK9: an enigmatic protease.** *Biochim Biophys Acta* 2008, **1781**:184–191.

82. Arnedo M, Taffé P, Sahli R, Furrer H, Hirschel B, Elzi L, Weber R, Vernazza P, Bernasconi E, Darioli R, Bergmann S, Beckmann JS, Telenti A, Tarr PE, Swiss HIV Cohort Study: **Contribution of 20 single nucleotide polymorphisms of 13 genes to dyslipidemia associated with antiretroviral therapy**. *Pharmacogenetics and Genomics* 2007, **17**:755–764.

83. Thompson A, Di Angelantonio E, Sarwar N, Erqou S, Saleheen D, Dullaart RPF, Keavney B, Ye Z, Danesh J: **Association of cholesteryl ester transfer protein genotypes with CETP mass and activity, lipid levels, and coronary risk.** *JAMA* 2008, **299**:2777–2788.

84. Lagace TA, Curtis DE, Garuti R, McNutt MC, Park SW, Prather HB, Anderson NN, Ho YK, Hammer RE, Horton JD: **Secreted PCSK9 decreases the number of LDL receptors in hepatocytes and inlivers of parabiotic mice**. *J Clin Invest* 2006, **116**:2995–3005.

85. Cameron JJ, Holla ØLØ, Ranheim TT, Kulseth MAM, Berge KEK, Leren TPT: **Effect of mutations in the PCSK9 gene on the cell surface LDL receptors.** *Human Molecular Genetics* 2006, **15**:1551–1558.

86. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nat Rev Genet* 2006, **7**:85–97.

87. Consortium T1GP, author C, committee S, Medicine PGBCO, BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, Technologies L, Max Planck Institute for Molecular Genetics, Science RA, Washington University in St Louis, Wellcome Trust Sanger Institute, Technologies AGA, Medicine BCO, BGI-Shenzhen, College B, Hospital BAW, Broad Institute of MIT and Harvard, Cardiff University, The Human Gene Mutation Database, Laboratory CSH, Universities CAS, European Bioinformatics Institute, Laboratory EMB, Illumina, Johns Hopkins University, Leiden University Medical Center, Technologies L, Louisiana State University, Max Planck Institute for Molecular Genetics, US National Institutes of Health, et al.: **A map of human genome variation from population-scale sequencing**. *Nature* 2011, **467**:1061–1073.

88. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, Macdonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, et al.: **Global variation in copy number in the human genome**. *Nature* 2006, **444**:444–454.

89. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, Carter NP, Church DM, Felsenfeld A, Guyer M, Lee C, Lupski JR, Mullikin JC, Pritchard JK, Sebat J, Sherry ST, Smith D, Valle D, Waterston RH: **Completing the map of human genetic variation**. *Nature* 2007, **447**:161–165.

90. Panel on Antiretroviral Guidelines for Adults & Adolescents: **Guidelines for using antiretroviral agents among HIV-infected adults and adolescents.**

91. Martinson JJ, Chapman NH, Rees DC, Liu YT, Clegg JB: **Global distribution of the CCR5 gene 32-basepair deletion.** *Nature Genetics* 1997, **16**:100–103.

92. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome**. *Nature Publishing Group* 2004, **36**:949–951.

93. Margolick JB, Munoz A, Donnenberg AD, Park LP, Galai N, Giorgi JV, Ogorman M, Ferbas J: **Failure of T-Cell Homeostasis Preceding Aids in Hiv-1 Infection**. *Nat Med* 1995, **1**:674–680.

94. Ribeiro RM: **Dynamics of CD4+ T cells in HIV-1 infection**. *Immunol Cell Biol* 2007, **85**:287–294.

95. Catalfamo M, Wilhelm C, Tcheung L, Proschan M, Friesen T, Park JH, Adelsberger J, Baseler M, Maldarelli F, Davey R, Roby G, Rehm C, Lane C: **CD4 and CD8 T Cell Immune Activation during Chronic HIV Infection: Roles of Homeostasis, HIV, Type I IFN, and IL-7**. *The Journal of Immunology* 2011, **186**:2106–2116.

96. Serrano-Villar S, Sainz T, Lee SA, Hunt PW, Sinclair E, Shacklett BL, Ferre AL, Hayes TL, Somsouk M, Hsue PY, Van Natta ML, Meinert CL, Lederman MM, Hatano H, Jain V, Huang Y, Hecht FM, Martin JN, McCune JM, Moreno S, Deeks SG: **HIV-infected individuals with low CD4/CD8 ratio despite effective antiretroviral therapy exhibit altered T cell subsets, heightened CD8+ T cell activation, and increased risk of non-AIDS morbidity and mortality.** *PLoS Pathog* 2014, **10**:e1004078.

97. Ohtsu H: **Histamine synthesis and lessons learned from histidine decarboxylase deficient mice**. 2010:21–31.

98. Sasaguri Y, Tanimoto A: **Role of macrophage-derived histamine in atherosclerosis--chronic participation in the inflammatory response--.** *J Atheroscler Thromb* 2003, **11**:122–130.

99. Higuchi S, Tanimoto A, Arima N, Xu H, Murata Y, Hamada T, Makishima K, Sasaguri Y: **Effects of histamine and interleukin-4 synthesized in arterial intima on phagocytosis by monocytes/macrophages in relation to atherosclerosis**. *FEBS Lett* 2001, **505**:217–222.

100. Wang KY, Tanimoto A, Guo X, Yamada S, Shimajiri S, Murata Y, Ding Y, Tsutsui M, Kato S, Watanabe T, Ohtsu H, Hirano KI, Kohno K, Sasaguri Y: **Histamine Deficiency Decreases Atherosclerosis and Inflammatory Response in Apolipoprotein E Knockout Mice Independently of Serum Cholesterol Level**. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2011, **31**:800–807.

101. Clejan S, Japa S, Clemetson C, Hasabnis SS, David O, Talano JV: **Blood histamine is associated with coronary artery disease, cardiac events and severity of inflammation and atherosclerosis**. *J Cell Mol Med* 2002, **6**:583–592.

102. Iribarren C: **Are patients with asthma at increased risk of coronary heart disease?** *International Journal of Epidemiology* 2004, **33**:743–748.

103. Knoflach M, Kiechl S, Mayr A, Willeit J, Poewe W, Wick G: **Allergic rhinitis, asthma, and atherosclerosis in the Bruneck and ARMY studies**. *Arch Intern Med* 2005, **165**:2521–2526.

104. Siegel D, Devaraj S, Mitra A, Raychaudhuri SP, Raychaudhuri SK, Jialal I: **Inflammation, Atherosclerosis, and Psoriasis**. *Clinic Rev Allerg Immunol* 2012, **44**:194–204.

105. Sasaguri Y: **Role of Histamine Produced by Bone Marrow-Derived Vascular Cells in Pathogenesis of Atherosclerosis**. *Circulation Research* 2005, **96**:974–981.

106. Takagishi T, Sasaguri Y, Nakano R, Arima N, Tanimoto A, Fukui H, Morimatsu M: **Expression of the histamine H1 receptor gene in relation to atherosclerosis.** *The American Journal of Pathology* 1995, **146**:981–988.

107. Langeler EG, Snelting-Havinga I, van Hinsbergh VW: **Passage of low density lipoproteins through monolayers of human arterial endothelial cells. Effects of vasoactive substances in an in vitro model**. *Arteriosclerosis, Thrombosis, and Vascular Biology* 1989, **9**:550–559.

108. Rozenberg I, Sluka SHM, Rohrer L, Hofmann J, Becher B, Akhmedov A, Soliz J, Mocharla P, Boren J, Johansen P, Steffel J, Watanabe T, Luscher TF, Tanner FC: **Histamine H1 Receptor Promotes Atherosclerotic Lesion Formation by Increasing Vascular Permeability for Low-Density Lipoproteins**. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2010, **30**:923–930.

109. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, Jousilahti P, Männistö S, Eriksson JG, Saarela J, Ripatti S, Perola M, van Ommen G-JB, Taskinen M-R, Palotie A, Dermitzakis ET, Peltonen L: **An Immune Response Network Associated with Blood Lipid Levels**. *PLoS Genet* 2010, **6**:e1001113.