Multiplexed lineage tracing for high-dimensional single-cell analysis of smooth muscle cell differentiation gene control in atherosclerotic plaques

Emily T. Condlin

Human Biology Distinguished Majors Program

May 2021

Advisors

Eli R. Zunder, PhD
Principal Investigator of Zunder Lab
Associate Professor
Department of Biomedical Engineering

Corey Williams
PhD Candidate in Biomedical Engineering

Michael P. Timko, PhD
Lewis & Clark Chair Professor of Biology
Director, Human Biology Distinguished Majors Program
University of Virginia

This thesis is submitted in partial fulfillment of the degree requirements for the Distinguished Majors Program in Human Biology.

Honor Pledge: On my honor as a student at the University of Virginia, I neither given nor received unauthorized aid on this thesis.

ABSTRACT

Ischemic heart disease, which occurs when coronary arteries are narrowed and obstruct blood flow to cardiac muscle, is the leading cause of death in the United States and worldwide.[1,2] This narrowing is most often caused by atherosclerotic plaque buildup on arterial walls. Throughout atherogenesis, activated vascular smooth muscle cells (VSMCs) have been observed to undergo dedifferentiation and can later differentiate into dangerous phenotypes.[8,15] Knockout studies have demonstrated that *Klf4*, *Oct4*, *Runx2,* and *TGFb* play roles in the dedifferentiation of VSMCs and subsequent differentiation into high-risk phenotypes; however, the precise function of each gene remains unclear. Combinatorial effects are also unknown. We used multiplex lineage tracing to permanently tag gene loci based on unique markers followed by single-cell mass cytometry to understand the heterogeneity of cell types in plaques and the genomic expression histories of those cells. Multiplex lineage tracing was chosen because it allows for combinatorial analysis of all genes involved. This thesis project focused specifically on designing and incorporating the lineage tracing tools into target gene loci. Successful incorporation of these tools allowed for progression of the project and confirmed that our target genes are good candidates for this study. Overall, this long term project is still in its early stages; however, the experiments outlined here confirm that our system design is effective and should continue to be pursued. The long term goal of this study is to provide data to aid in development of gene therapies to control dedifferentiation, differentiation, and/or lysis of heterogeneous atherosclerotic plaques. Furthermore, our findings will hopefully clarify whether treatments that aim to strengthen the fibrous cap to deter lysis or ones that hinder phenotype switching of VSMCs in the first place will be the most effective means of treatment.

TABLE OF CONTENTS:

ABBREVIATIONS:

-HF: High Fidelity
°C: degrees celsius
AAV: Adeno-associated virus
ABC: ATP-binding cassette
AboB: Apolipoprotein B
ADB: agarose dissolving buffer
Apo E -/-: Apo E deficient
Apo E: apolipoprotein E
APR: acute phase response
BFP: Blue fluorescent protein
bp: base pair(s)
CreERT2: Cre recombinase fused to Estrogen ligand-binding domain (ERT2)
CRISPR: Clustered Regularly Interspaced Short Palindromic Sequences
CVD: cardiovascular disease
CyTOF: mass cytometry by time-of-flight
DNA: Deoxyribonucleic Acid
RNA: Ribonucleic Acid
ESC: embryonic stem cell
FDB: familial ligand-defective apoB-100
FHBL: Familial hypobetalipoproteinemia
GFP: green fluorescent protein
GRCm38: Genome Reference Consortium Mouse Build 38, synonymous with mm10
gRNA: guide RNA
HDL: high density lipoprotein
HDR: Homology Directed Repair
hDTR: Human diphtheria toxin receptor
HMG-CoA: β-Hydroxy β-methylglutaryl-CoA
i-GONAD: Improved-Genome editing via Oviductal Nucleic Acids Delivery
IACUC: Institutional Animal Care and Use Committee
IDT: Integrated DNA Technologies
IgF-1: insulin-like growth factor 1
kB: kilobase
Klf4: Kruppel-like factor 4; KLF4 gene
LB broth: Luria-Bertani broth
LDL: low-density lipoprotein
LDLR: LDL receptor
mESC: mouse embryonic stem cell
Mm10: Mus Musculus; synonymous with GRCm38
mRNA: messenger RNA
NCBI: National Center for Biotechnology Information
NeoR: neomycin resistance
NeoR: Neomycin resistance
NIH: National Institutes of Health
Oct4: Octamer-binding Transcription Factor 4

PAM: protospacer adjacent motif
PBS: Phosphate-buffered saline
PCR: polymerase chain reaction
PCSK-9: Proprotein convertase subtilisin/kexin type-9
Poly(A) tail: Polyadenine tail
pregnant mare's serum gonadotropin (PMSG) and human chorionic gonadotropin
Runx2: RUNX Family Transcription Factor 2
SMC: smooth muscle cell
TFG-β: transforming growth factor-β
TGF-β: Transforming Growth Factor
TNF: tumor necrosis factor
UMAP: Uniform Manifold Approximation and Projection
USCS: University of California Santa Cruz GEnome Institute
UTR: untranslated region
VLDL: very low density lipoprotein
vSMC: vascular smooth muscle cells
YFP: Yellow fluorescent protein
μL: microliter

BACKGROUND

Atherosclerosis is the most prevalent and fatal disease in the US and worldwide.[1,2] An expansive body of scientific evidence has established that hyperlipidemia, excess levels of low-density lipoprotein, lipid oxidation and immune responses, are the driving physical forces behind atherogenesis and its critical complications.[1] Although the physical events of atherogenesis and atherosclerosis have been determined and clarified, many of the genomic drivers of these events remain a mystery. Moreover, despite the intense and widespread effects of this disease, we lack highly effective, long-term therapies for treatment of this disease.

The events of atherogenesis and atherosclerosis:

Atherogenesis, or the emergence of fatty plaques in the arteries, begins when excess LDL particles in the bloodstream enter lesions in a dysfunctional endothelial lining of arterial cell walls. Serum levels of LDL are proportional to levels of LDL that enter arterial endothelial cells. Injurious events adjacent to the endothelial barrier, including hemodynamic forces, toxins, excess lipid particles, hypertension and inflammation, may cause endothelial dysfunction which weakens the lining, allowing for increased permeability and lipid accumulation in layers of the subendothelial space.[2] Hemodynamic factors include disrupted blood flow patterns and low shear stress. Shear stress is the tangential frictional force upon the arterial walls created by blood flow. In straight, non-obstructed vasculature, shear stress is high because blood maintains a unidirectional, laminar flow pattern. Arterial branching leads to disturbed laminar flow, causing flow oscillation and lowered shear stress against the endothelial walls. This alters the barrier's morphology to exhibit increased permeability and inflammation. Studies have shown that these hemodynamic effects cause a predilection for

atherosclerotic plaque formation at bifurcations and arterial branch points. [5,6] The endothelial lining's integrity can also be decreased by lesion formation, most often initiated by irritation or stress to the barrier as a result of hyperlipidemia, high blood pressure, high glucose levels or toxins from cigarettes. [1,7] These lesions also lead to inflammation and increased barrier permeability, allowing for accumulation of LDL in the subendothelial space. Endothelial cells do possess some protective functions to offset the effects of toxins or increased pressure, including secretion of nitric oxide (NO) to induce vasodilation and anti-inflammatory effects. Physical exercise and diets low in LDL can also decrease hemodynamic factors and lesions formation. Still, over time, these protective functions may not be sufficient to avoid induction of endothelial dysfunction.

Once endothelial dysfunction occurs, excess LDL particles can deposit in the tunica intima layer of the arterial wall and bind to proteoglycans, causing LDL to accumulate in the tunica intima (the innermost tunica (layer) of an artery or vein). Elevated arterial wall stress caused by hypertension causes increased proteoglycan synthesis, which promotes LDL-proteoglycan binding and accumulation. In the subendothelial space, metalloproteinases and reactive oxygen species can induce oxidation of LDL particles bound to proteoglycans. Once oxidized, LDL particles can no longer leave the tunica intima layer.[1] Exposure to toxins, including those from cigarettes, increases production of reactive oxygen species in endothelial cells, leading to increased LDL oxidation.[7] The presence of oxidized LDL causes endothelial cells to express white blood cell receptors or adhesion molecules on their surface.[1,8] These receptors signal for immune cell (monocyte and lymphocyte) translocation to the lesion site. Monocytes arriving at the lesion differentiate into macrophages, which take up the excess LDL in the tunica intima layer and become foam cells once saturated with

lipids.[1,8] These foam cells then release proinflammatory chemokines and recruit more monocytes. As foam cells become oversaturated with lipids they die and release proinflammatory reactive oxygen species along with their lipid contents, attract neutrophils, and lead to activation of more endothelial cells which produce adhesion molecules to signal for more immune cell recruitment.

At this point, a fibrous cap begins to form between the arterial vessel's lumen and the plaque core. When vascular tissues are severely damaged, they release cytokines, tumor necrosis factors (*TNF*), and growth factors which induce vSMC migration into the lumen. Some of these migrated vSMCs dedifferentiate and differentiate into fibrous cap cells. These cells, along with collagen-rich fiber tissues, macrophages (and foam cells) and T lymphocytes synthesize an extracellular matrix around the plaque and lead to formation of the mature atherosclerotic bulge into the vessel lumen.[8]

Previous studies suggest that foam cells may secrete growth factor IgF-1, which has been identified as a probable factor in inducing migration of vSMCs from the tunica media to the tunica intima and into the fibrous cap, as well as acting as a vSMC antiapoptotic and mitogenic factor.[3] IgF-1 is known to express in atherosclerotic tissues. However, low levels of IgF-1 are associated with increased risk of ischemic events and although IgF-1 is correlated with atherosclerotic events, it may possess a protective or anti-ischemic function.[4] Numerous studies have also illustrated that this protective function is materialized by stabilization of the fibrous cap and increased cap to core ratio in plaques. Activated vSMCs migrate and undergo phenotype switching to aid in formation of the fibrous cap, which stabilizes plaques and deters thrombosis, clot formation, and ischemic events. vSMC

apoptosis is thought to contribute to fibrous cap weakening and plaque rupture. Thus, IgF1's vSMC migratory and antiapoptotic functions may be integral to plaque stabilization.

The precise drivers of the dedifferentiation and differentiation pathways of vSMCs in atherosclerotic plaque formation are still largely unknown and are the main focus of this study. It is known that activated vSMCs dedifferentiate into stem cell-like cells and myofibroblast-like cells that can then differentiate into mesenchymal stem cell-like cells which then differentiate into osteoblast phenotype. Osteoblast phenotype vSMCs are thought to produce collagen and lead to plaque calcifications.

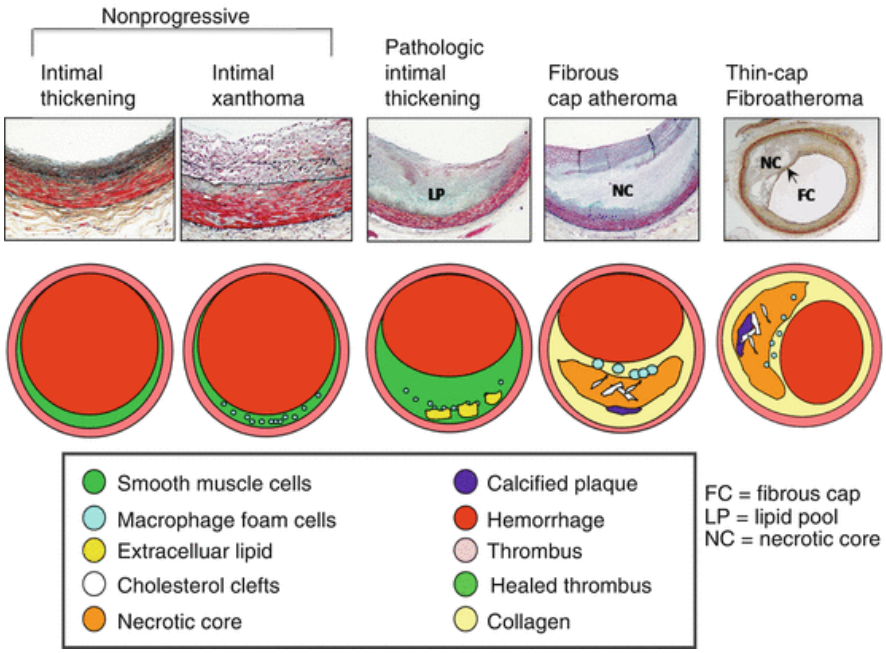


Figure 1: Progression of atherogenesis within arterial lumen[51]

Causes and genetic mechanisms of hyperlipidemia:

Many factors cause fatty streaks to progress through atherogenesis and form mature atherosclerotic plaques. A major risk factor is pathogenic levels of cholesterol and

lipoproteins, known as hyperlipidemia or hyperlipoproteinemia. High cholesterol levels can be caused by cholesterol synthesized in the body or ingested in the diet. Although it can lead to atherosclerosis and heart disease, cholesterol, when balanced correctly, is a crucial part of cell membranes and a major steroid precursor.[9] Normally, cells synthesize about 80% of the body's cholesterol while about 20% comes from the diet.[9] When ingested, pancreatic lipase breaks down fats, then bile salts and lecithin aid in emulsification of the hydrophobic lipid molecules.[9,10] Bile salts form micelles with fat molecules to transport them through intestinal microvilli. Cholesterol combines with phospholipids, triglycerides, and protein carriers to form lipoproteins called chylomicrons.[9,10] The outer envelope of lipoproteins are composed of phospholipids, proteins, and cholesterol while the inner core is mainly triglycerides and cholesterol esters. Chylomicrons move through lymphatic vessels and into the bloodstream where muscles and adipose tissue extract triglyceride molecules, leaving chylomicron remnants[9]. The liver takes up these chylomicron remnants and forms VLDL, which move back into circulation where its triglycerides are metabolized by muscles and adipose tissue. VLDL remnants are metabolized to LDL, which is the lipoprotein with the highest proportion of cholesterol, often called "bad cholesterol" [10,11]. Chylomicron remnants, VLDL, and LDL act as pro-atherogenic lipoproteins while HDL is anti-atherogenic (atheroprotective).[11]

Nascent HDL particles are synthesized in the liver and intestine. HDL aids in reverse cholesterol transport, which allows for the constant cholesterol turnover cells require, by efficiently carrying excess insoluble cholesterol and phospholipids from cells to the liver.[9,10] Hepatocytes in the liver are the only cells capable of degrading and eliminating cholesterol. If the hepatocytes cannot eliminate LDL and cholesterol quickly enough because of excess levels or other environmental or genetic factors, hyperlipidemia ensues[9].

Endogenous lipid production and clearance rates are governed by the concentration of LDL receptors on the surface liver cells. High LDL receptor activity signals that plasma LDL levels are sufficiently high and causes increased uptake from the bloodstream as well as a decreased rate of VLDL to LDL synthesis. Conversely, low LDL receptor activity causes less uptake from the bloodstream and increased synthesis of LDL molecules. If cholesterol levels in the cell decrease below the cellular limit, regulatory elements are activated which send transcription factors to the nucleus to stimulate genes including those which code for the LDL receptor and HMG-CoA reductase, a crucial enzyme in cholesterol synthesis [12]. On the other hand, if cholesterol levels are sufficiently high, these pathways are not activated and no transcription of the genes occur. High intracellular LDL levels can also trigger ubiquitination and degradation of LDL receptors.[12] One gene involved in LDL receptor targeting and degradation is *PCSK-9*. Mutations in *PCSK-9* that cause loss of function are atheroprotective and cause increased LDL receptor activity and decreased plasma LDL. On the other hand, gain of function *PCSK-9* mutations are atherogenic and cause decreased LDL receptor activity and increased plasma LDL levels.

Genetic risk factors for development of atherosclerosis:

The overall risk of cardiovascular disease has been found through extensive research to be dynamic influenced by both environmental and genetic factors.

Genetic factors have been studied through twin studies, genome-wide scans, single nucleotide polymorphism (SNP) investigations, and knockout studies. A family history of cardiovascular disease is now considered a risk factor independently associated with risk of developing the disease.[13] Twin studies suggest that coronary atherosclerosis and ensuing

ischemic events have a heritability between 38% and 57%.[13] There are two major gene-type foci in these studies: genes that control lipid homeostasis mechanisms and genes that control plaque formation or stability. Since it is likely that many genes act together to effect atherosclerotic phenotypes, many studies utilize a candidate gene approach in which several candidates for involvement in atherogenesis are examined for association with atherosclerotic or ischemic events.[13] Often, this approach has used haplotype blocks with low histories of recombination (i.e., allelic variations inherited together) and tagged SNPs to assess the combinatorial effects of various polymorphisms and distinct haplotypes. This technique allows researchers to focus more acutely on genes imparting high genetic risk rather than individually analyzing the huge number of allelic variations that exist within genomes. A multiplexed tracing system with a similar goal of combinatorial association has been used in this study.

Familial hypercholesterolemia (FH) is a condition that can impart extremely high risk of atherosclerosis and ischemic events. FH often presents with elevated LDL cholesterol as a result of deficient LDL receptor (LDLR) activity.[13] Absence or dysfunction of LDLRs hinder hepatocyte uptake and breakdown of plasma LDL cholesterol. Liver LDLR concentration is especially impactful on plasma LDL concentration because the number of LDLRs determines the rate at which hepatocytes can take up plasma LDL. Conversely, the number of LDLRs present is dictated by cellular cholesterol content. If there is high lipid content in the cell, surface LDLRs decrease to hinder further lipid uptake while low cellular cholesterol increases the number of surface level receptors.[12] The LDLR encoding gene is housed on chromosome 19 and has been identified to have 800 allelic variations, categorized into 5 classes. Class 1 variants are unable to produce any immunoprecipitable LDLR and are

referred to as receptor-negative mutations.[13] Classes 2-5 are called receptor-defective mutations because they result in immunoprecipitable LDLR but the LDLR produced has defective function. FH heterozygous individuals express half the number of functional LDLR seen in wild type variations, which has been observed to cause plasma LDL cholesterol concentrations to double. Individuals who are FH homozygous are rare and express few to no functional LDLR, causing LDL cholesterol concentrations up to eight times higher than wild type. FH homozygotes have been reported to suffer ischemic events prior to the age of 20.[13,14]

Mutations in apolipoproteins that act as ligands for functional LDLR can also cause increased plasma LDL cholesterol concentrations. Apolipoprotein E (Apo E) is housed on LDL particles and acts as a ligand for LDLR, therefore playing a major role in LDL uptake by the liver[12]. Apo E,which has been mapped to chromosome 19, has 3 genetic variants, Apo E2, E3, and E4. E3 and E4 are good ligands for LDLR while Apo E2 is poorly recognized by receptors. Thus, homozygosity for Apo E2 causes a strong predisposition for familial dysbetalipoproteinemia, a form of hyperlipidemia.[9] In addition, complete Apo E deficiency (Apo E -/-) poses an extreme risk for atherosclerotic plaque formation. Apo E deficiency is employed as a disease model in this study.

Another gene that affects lipid homeostasis is *APOB*, which codes for apolipoprotein B (Abo B) and is located on chromosome 2.[13] The Apo B acts as the main apolipoprotein in chylomicrons and LDL particles. *APOB* mutations can result in hyperlipidemia or hyperlipidemia, depending on their location. Familial hypobetalipoproteinemia (FHBL) is a genetic disorder that occurs when an *APOB* mutation results in truncated Apo B protein molecules, which decreases plasma LDL concentrations.[13] As such, FHBL causes a hypolipidemic, atheroprotective phenotype. Conversely, familial ligand-defective apoB-100

(FDB) occurs when an *APOB* mutation results in Apo B LDLR ligands with a defective C-terminus.[13] This hinders LDL and chylomicron particles with the defective ligands from entering through LDLR into hepatic cells to be broken down. FDB therefore results in a hyperlipidemic phenotype and has been shown to increase risk of atherosclerosis and CVD.[13]

HDL related mutations also increase risk of atherosclerosis. ATP-binding cassette (ABC) membrane proteins are transporters on the surface of HDL particles that mediate uptake of cholesterol for delivery to hepatocytes by hepatocyte scavenger receptor SCARB1.[13] ABC Subfamily A Member 1 (ABCA1) is an imperative cholesterol pump in the hepatic LDL removal pathway and is located on chromosome 9.[13] *ABCA1* mutations cause reduced serum HDL concentrations and lead to development of Tangier disease and premature atherosclerosis. While *ABCA1* polymorphisms do not have a vast impact on plasma lipoprotein level, mutations are still considered a high risk because, independent of lipoprotein concentrations, they are associated with increased occurrence and severity of ischemic events. More research is needed to understand the exact mechanisms of this outcome.[13] Although rare, individuals with multiple *ABCG5* or *ABCG8* mutations can also endure deleterious increases in sterol (including cholesterol) absorption through the intestinal lining and decreases in excretion of sterols.[13] These genetic disorders can result in premature atherogenesis and CVD.

Despite the evident impact that lipid concentrations have on atherogenesis and ischemic event frequency, genes that influence other stages of atherosclerosis can also cause variations in individual risk. Many who lack abnormal lipid and lipoprotein concentrations develop CVD or atherosclerosis from other causes, namely, inflammation.[13,15] As aforementioned, monocyte and macrophage immune cell recruitment plays a major role in

atherogenesis and plaque progression. Genes that control monocyte or macrophage recruitment are thought to have major impacts on risk for development of atherosclerosis. While many genes and activation pathways are involved in atherosclerotic inflammatory processes, growth factors, toll-like receptors, and calcification are some of the most prominent and, therefore, will be focused on in this current exploration.

Transforming growth factor-β (TGF-β) is involved in proliferation and differentiation regulation in many cell types including VSMCs. TGF-β concentration is known to vary between individuals because of many polymorphisms in the gene locus. Increased plasma TGF-β has been shown to increase risk of atherosclerosis. It has also been observed that TGF-β released from platelets stimulates collagen synthesis that, along with its stimulation of VSMC proliferation and differentiation, leads to establishment of an extracellular matrix and thickening of a fibrous cap, which is protective against plaque thrombosis and ischemic events. Collagen synthesis is especially dangerous because it leads to hardening of the plaque and cap, creating a stiffer plaque expansion into the lumen. *TGF-β* is not the only gene that plays a role in calcification or hardening of the plaque: overexpression of genes coding for osteopontin (OPN) and osteoprotegerin (OPG) have also been identified as factors in the development of increases calcification in plaques and creation of highly dangerous atherosclerotic states. OPN has been demonstrated in numerous atherosclerotic mechanisms including migration and proliferation of VSMCs, endothelial cells, and macrophages. OPN has been specifically implicated in calcification of coronary artery plaques and ischemic strokes. Studies have shown that symptomatic (i.e. ischemic) plaques express abundant OPG, a member of the tumor necrosis factor (TNF) family, far more often than do asymptomatic ones. Therefore, OPN is also thought to play a major role in plaque instability.

A final gene that has been shown to greatly increase risk of atherosclerosis is *Toll like receptor 4* (*TLR4*). *TLR4* has been implicated in angiogenesis at plaque sites, recruitment of proinflammatory factors, and other signaling involved in plaque progression. *TLR4* is upregulated proportionally with increased levels of oxygenated LDL and has been observed to increase macrophage recruitment to plaque sites. Furthermore, *TLR4* has a major role in initiating macrophage apoptosis within the intima layer.

Environmental risk factors for development of atherosclerosis

Risk of atherosclerosis is complex and most likely determined by the combinatorial effects of individual genotype and environment. Genetic disorders, diseases and polymorphisms are only able to capture a portion of those factors that increase individual risk; therefore, complex, non-genetic factors and genetic factors that interact with individual environments must be addressed as well[26]. Chronic inflammation, chronic stress, hypertension, sedentariness, diabetes, obesity, smoking, ageing, and inability to access healthy diet sources may all increase individual predispositions to development of CVD and atherosclerosis. Population level studies have shown that stress due to structural violence and racism is an environmental factor that not only impacts inflammation and atherosclerosis but also many other health events including maternal and infant mortality and birth outcomes, diabetes, depression and anxiety.

Food deserts, which are low-income areas with scarce access to healthy food and grocery stores, have been shown to increase risk of CVD. Limited access may be defined in various ways including a lack of nearby grocery stores, no restaurant options outside of fast food, no affordable options outside fast food, and whether food access hours are correlated

with the hours that local residents work. If healthy food is not available when one can access it, or is not available at all, cheaper, more accessible fast food options will be used instead. Unfortunately, these fast, cheap, available options are often extremely high in cholesterol, trans fats, sugars and dangerous additives which increase cardiovascular risk. In addition, fast food options often lack adequate fiber concentrations. Since fat aids in absorbing cholesterol while fiber hinders its absorption, a high-fat-low-fiber diet poses a distinct risk of developing hypercholesterolemia[28].

Of note, a cross-sectional study found that limited food access by itself did not implicate increased risk, but rather did so in association with low income.[27] Access to community and individual resources that correlate with income level, such as recreational spaces, community walkability, healthcare opportunities, education, physician-patient relationship and health insurance, can act as determinants in cardiovascular health status[27]. Along with diet and food access, these factors limit ability to get exercise and receive cardiovascular healthcare. Furthermore, even if healthcare is accessed, lifestyle changes are often the first treatment option presented by physicians; yet, after considering the aforementioned factors, it becomes apparent that if an individual lives in a community with little recreational or walkable spaces, limited healthy food access, and has scarce free time, such lifestyle changes may not be possible.

Many cross-sectional and longitudinal studies have been performed to quantify this relationship between income and CVD risk. The highest risk population has been identified as individuals who experience an income drop over at least 6 years compared to those with a steady income over this same timeframe.[29] Populations with an income drop experience a significant increase in risk of incident CVD within the next 17 years. Lowered risk of CVD

was identified in populations that experience income increase over at least 6 years[29]. Nonetheless, overall low income (even without a decrease) remains a major risk factor for ischemic events and development of CVD. Using the weighted Duke index and controlling for clinical and familial cardiovascular risk factors, it was found that the highest income groups have a significantly lower risk of atherogenesis and subsequent ischemic events compared to low income groups, which experience a significantly increased risk.[30] One study found that the risk of developing atherosclerosis increased threefold in the lowest income category compared to the highest income category in the study[31]. A linear increase in odds ratio was observed with decreasing income.[31] It is of note, however, that while high income groups are less likely to form atherosclerosis plaques, if lesions do form in individuals within either group, no correlation has been established between income and lesion severity.[30] These findings emphasize the impact that economic status and the impact access to resources can have on CVD and atherosclerosis outcomes. As such, consideration of income and income changes are important to note in treatment plans.

Another factor that has been proven to increase risk of death due to CVD and atherosclerosis is race. In the United States, African American citizens experience an increased risk of bad disease outcomes when compared to white American citizens and first-generation African immigrant citizens. Causal factors of this increased risk are thought to include increased inflammation stemming from chronic stress and decreased trust of medical systems stemming from past structural violence and traumas.[32, 33] Expansive evidence has shown that repeated acute episodes of stress and chronic stress both induce acute phase response (APR). APR is an innate immune response in which the hematopoietic and hepatic systems are activated and secrete plasma proteins and cytokines to prepare the

body to fight infection, inflammation, or trauma.[34] Repeated bouts of APR increases baseline cytokine and immune cell levels and therefore overall inflammation.[32] Increased inflammation subsequently induces stress and APR, creating a cycle of increasing inflammation and cardiovascular risk. Liver, endothelium and fat deposits have all been implicated as loci of origin of cytokines induced in APR, which further highlights the cyclical nature and interconnectedness of APR and atherosclerosis.[32]

While non-Hispanic black American citizens had a lower incidence of CVD than white American citizens (9.5% and 11.5%, respectively) in 2019, non-Hispanic black Americans experienced a significantly higher number of deaths due to heart disease than White Americans (age-adjusted rates were 208 per 100,000 and 168.9 per 100,000 in 2017, respectively).[35] Additionally, non-Hispanic black Americans experience a 13.4% higher rate of hypertension, a 9.2% higher rate of obesity and a 6.6% increase in diabetes compared to their white American counterparts. As such, non-Hispanic black Americans not only have an increased risk of death with CVD but also possess increased rates of risk factors for development of CVD.

A cross-sectional study of the prevalence of hypertension, diabetes, obesity, high cholesterol, physical inactivity and smoking in first generation African Immigrants to the United States and African Americans who were born in the United States further illustrates the disproportionate effects that African Americans experience.[36] This study partially exposed the extent to which sustained environmental factors, such as racism and structural violence experienced throughout life, may exacerbate potential endogenous risks. It was found that African Americans had higher rates of hypertension (32% versus 22%), obesity (70% versus 61%), diabetes (10% versus 7%), smoking status (19% versus 4%) and high

cholesterol (5% versus 4%) compared to their African immigrant counterparts regardless of sex and after age-standardization (Figure 2). CVD prevalence was, overall, lower in African immigrants than in African American populations. Both populations did demonstrate significant heterogeneity in disease status, presentation, and risk factors. These findings also displayed the importance of disaggregation of data and clinical treatment by country of origin.[36]

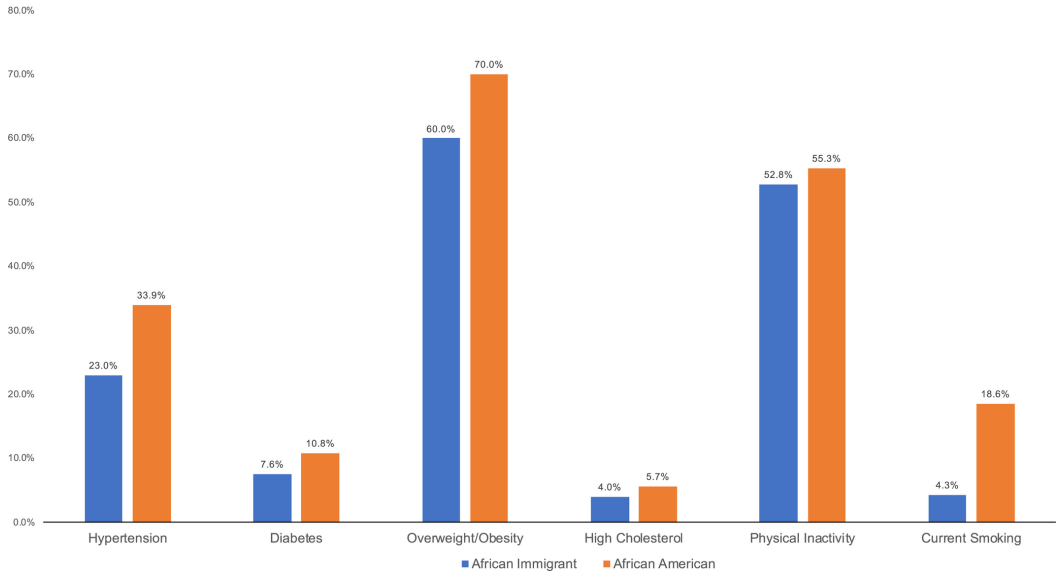


Figure 2: Presentation of CVD risk factors in African Americans versus African immigrants[36]

As is evident upon examination of environmental factors, atherosclerosis risk becomes incredibly heterogenous and individualized based on genetic and environmental exposure variations. When considering treatment options, these combinatorial effects must be considered in completeness.

Current Treatment Options:

Despite the vast effects of CVD and atherosclerosis, treatment options remain incomprehensive and are often accompanied by side effects that hinder patients from experiencing their ideal quality of life. Current secondary treatments, which center around surgical procedures and pharmaceutical regiments, are employed when modifications to diet, physical activity, and smoking are insufficient.[37]

Surgical options like coronary artery bypass, atherectomy, angioplasty, or stent placements may acutely relieve pressure on the vascular system and mediate symptoms; yet, they cannot provide long term or ongoing treatment to mitigate formation of new plaques or growth of already established ones. Surgery and procedures can also become increasingly expensive (especially if patients lack insurance coverage) and may come with extensive recovery periods post-operation.

Though pharmaceutical options have proved successful in many ways, they do require strict dedication to a regimen and may induce intense side effects in patients. Options ranging from LDL lowering treatments to cholesterol absorption inhibitors to proprotein convertase subtilisin/kexin type-9 inhibitors to triglyceride lowering therapies to HDL enhancing therapies are currently available and all target different facets of atherogenesis.[37,38]

LDL lowering therapies are a type of lipoprotein modification treatment, the most common of which are HMG-CoA reductase inhibitors, more commonly called statins.[38] Statins induce recruitment of LDLRs to hepatocyte cell surfaces, which, as mentioned above, allows LDL particles in the bloodstream to be taken up at an increased rate, thus decreasing plasma LDL concentrations. Statins have been shown to decrease risk of cardiovascular

events by 22% for each mmol/L that plasma LDL concentrations are reduced.[38] Current guidelines suggest that those with cholesterol levels greater than 4.8 mmol/L should consider the genetic and environmental risk factors that apply to them in order to decide whether they should begin statins. It is suggested that anyone with LDL concentrations of 10.5 and over should begin taking statins.[39,40] In clinical trials, it has been found that statins are able to lower LDL levels to the goal concentration of 1.8 mmol/L using statins. In clinical realities though, only 48.5% of people on statins achieved LDL concentrations that low after 2 years of use.[38] A major contributing factor for this finding likely includes non-compliance with the pharmaceutical regimen for various reasons including side effects and inability to pay for doses. In one study 38% patients noted that their high-dose statin treatments caused extreme side effects including myalgia (muscular pain) which deterred them from participating in their everyday activities. 4% of patients on high-dose statin treatment said that myalgia effects were so severe they could not get out of bed.[38] Some patients may be classified as "statin intolerant" when they are unable to continue statin treatment due to intolerable side effects or when blood tests provide evidence of abnormal liver or muscle function. Partial statin intolerance is limited to inability to tolerate specific doses or drugs while complete intolerance terminates all statin use.[42] Furthermore, statin use in patients with already elevated blood sugar or predisposition to diabetes has been shown to increase risk of Type II diabetes mellitus.[38,40]

Another pharmaceutical approach uses cholesterol absorption inhibitors to reduce concentration of dietary cholesterol absorbed. Ezetimibe, a major cholesterol absorption inhibitor, has been shown to decrease plasma LDL concentrations by 15-20% when used as a monotherapy and an additional 15-20% when used in conjunction with statin therapy.

Cholesterol absorption inhibitors are most often used when statins are unable to achieve target plasma LDL concentrations on their own.[38]

Proprotein convertase subtilisin/kexin type-9 (PCSK-9) inhibitors are a newer pharmaceutical therapy in comparison to statins and were first introduced in 2007. PCSK-9 in the liver normally binds to LDLR and induces the receptors' degradation. PCSK-9 inhibitors disable this pathway and thereby allow more LDLRs to remain on hepatocyte cell surfaces.[38] Maximization of LDLR on cell surfaces allows for more LDL absorption and therefore lower plasma LDL concentrations. PCSK-9 inhibitors have been demonstrated to reduce plasma LDL by a mean of 51% after 6 months of treatment in clinical trials. Adverse effects of PCSK-9 inhibitors include flu-like symptoms, nasopharyngitis, abdominal discomfort and myalgia.[41] One study found that 40.3% of those who experienced adverse effects from a PCSK-9 inhibitor drug discontinued use because of it.[41]

Other pharmaceuticals in use include HDL increasing therapies. While these therapies pose promising advantages in theory, current offerings have not proven to lead to a reduction in cardiovascular or ischemic events. Many researchers think that this is an area to identify future drug classes to aid in treatment of patients who do not demonstrate sufficient decreases in plasma LDL with other therapies or who demonstrate statin intolerance.

A major concern of pharmaceutical treatments is that treatment regimens must be consistently adhered to in order to remain effective. With the necessity of strict adherence comes increasing monetary costs of long-term drug treatment. In addition, surgical and pharmaceutical therapies may not be sufficiently effective and may cause increasing physical costs associated with side effects and recovery time. There also exist costs to the individual

and society stemming from loss of productivity due to disease outcomes. The other major current treatment path, lifestyle and behavioral change, remains even more unpredictable and possibly becomes impossible for certain individuals and populations upon consideration of socioeconomic factors and resource availability. As such, it becomes increasingly problematic that the only available therapies outside of lifestyle change include strict pharmaceutical regimens and surgery.

Future Treatment Options: Gene Therapies:

Gene therapies compose a growing research and treatment field in which personalized and effective medical pursuits can be used to target disease-specific gene drivers. Atherosclerosis gene drivers have been investigated; however, to date, there are no approved genetic therapies for any types of CVD.

One CVD gene therapy, which uses CRISPR editing technology and partially deactivating Cas9 to knock out *PCSK-9* function, is furthest along in development and testing. The treatment, delivered by adeno-associated viruses (AAVs), was recently given to a cohort of 14 monkeys, 7 of whom received the treatment and 7 a placebo.[43] All 14 monkeys survived and those receiving the experimental technology presented with a mean 89% decrease in PCSK-9 levels and 59% decrease in plasma LDL levels.[43] Since this experiment was completed in June 2020, it is not yet possible to determine the long term treatment outcomes.[43]

Our approach aims to better understand a different area of CVD, namely, atherosclerotic plaque formation. VSMC specific conditional knockout studies have been performed to determine the genes involved in differentiation of VSMCs into various

phenotypes indicated in plaque formation.[44] *KLF4* gene (*Klf4*) knockout mice demonstrated formation of atherosclerotic lesions that were smaller in size and more stable than wild type. In addition, reduced concentrations of macrophage-like and mesenchymal stem cell-like cells derived from vSMC pathways were observed.[44]

In order to understand the purpose of *Klf4*, singleplex lineage tracing of *Klf4* in knockout and wildtype mice has performed.[44] It was found that, in addition to its atherogenic functions, *Klf4* aids in tissue repair after vascular traumas.[44] VSMCs express Klf4 when remodeling vascular tissue after reperfusion injuries, or injuries that occur when blood and oxygen supply reprofuse an area of vasculature after an anoxic period. In addition, concentration of *Klf4*-mediated VSMC derived macrophage-like cell phenotypes were seen to increase during repair phases.[44] In *Klf4* knockout mice, incidence of heart failure following reperfusion injuries was heightened.[44] *Klf4* knockout mice also demonstrated cardiac dilation, leading to lowered and oscillating shear stress, which, as aforementioned, leads to increased permittivity and inflammation of the endothelial barrier, a major step in atherogenesis. *Klf4* therefore plays a critical role in transition of VSMCs into heterogeneous plaque components and lesion pathogenesis.[45]

Knockout studies of another gene of interest, *Oct4*, showed converse effects to the *Klf4* knockouts.[46] *Oct4* knockout mice presented with increased lesion sizes and thinner fibrous caps indicating decreased plaque stability.[46] In some mice, fibrous cap formation seemed almost entirely absent and plaques contained less collagen and more lipid concentration (which indicted higher risk of thrombosis) compared to plaques in wildtype mice on the same diet and in the same environment. *Oct4* knockout mice were also observed to express a higher rate of plaque hemmorhage.[46] Pathways that were enhanced in the *Klf4*

knockout were seen to be minimized in *Oct4* knockouts almost completely, highlighting the antithetic function of these two genes.

In the same study that explored the role of *Oct4*, the previous function of *Runx2* was further analyzed.[46] *Runx2* acts as an osteogenic marker as it plays a major role in differentiation of mesenchymal stem cells into osteoblast phenotype cells. This holds a distinct importance in vSMC differentiation in atherosclerotic plaques because it allows for tracing of the cells from their dedifferentiated mesenchymal stem cell-like stage into the high-risk osteoblast stage.[46,47]

The final gene studied in this project is *Transforming Growth Factor*-β (*TGF–β*). TGF-β, as mentioned above, plays a major role in atherogenesis by mobilizing fibroproliferation in response to damaged tissues. *TGF-β* codes for a protein which acts as a ligand for TGF-β protein super family peptides. There are two main sources of TGF-β: secretion from platelets and activation of reservoirs within matrices. TGF-β orchestrates fibroproliferative responses after its release by initiating repair cell chemotaxis and regulating immune and inflammatory responses as matrix production occurs.[48] *TGF-β* also regulates formation of fibrotic scarring (fibrosis) by producing antiproliferation or apoptotic signals for fibrotic cells once sufficient fibroproliferation has been achieved at a lesion or damage site.[48]

Within atherosclerotic plaques and advanced stage lesions, TGF-β induces calcification of the artery.[48] Calcifications within vascular tissue, lesions and plaques may result in various outcomes depending on the size, shape, and location of the calcification. In general, the extent of calcification is inversely correlated with macrophage infiltration into

the lesion, and positively correlated with asymptomatic (i.e. non-ischemic) disease state. That is to say, as calcification increases, macrophage infiltration decreases, and the likelihood of being asymptomatic increases.[49] Still, if multiple calcification or calcifications on the outer surface of the plaque (rather than within the core) exist, extent of protrusion into the lumen and risk of thrombosis both increase.[49] Therefore, while calcification is often protective, extensive or superficial calcification may increase ischemic risk.

TGF-β knockout studies have confirmed that dysfunctions or absences of *TGF-β* pathways promote inflammation, fibrosis, and atherosclerosis. In atherosclerotic states, it has been shown that TGF-β secretion is downregulated in vascular walls and lesion cell response to TGF-β signaling is reduced. On the other hand, smooth muscle cell-like (SMC-like) phenotype cells within the lesion matrix do release active TGF-β. This upregulation in SMC-like cells has the initial purpose of allowing repair cells to function and remedy vascular damage. However, it is thought that this progress proceeds unchecked and thus the adverse effect of fibrosis is allowed to ensue. It becomes clear then that *TGF-β* is not a gene that can be accurately labeled as atherogenic or atheroprotective. As such, many studies have called for an increase in research on the molecular mechanisms that control temporal and spatial TGF-β activation and secretion.

Combinatorial Approach

As is evident from the above review of atherosclerosis and CVD risk factors, no one gene or environmental factor can be isolated as the driver of atherogenesis in an individual. As such, a multiplex, combinatorial approach was used in this study to visualize activation of four genes in one disease state model. Performing a multiplex approach provided many

advantages over performance of multiple singleplex approaches because it allowed for visualization of variations in activation of the four target genes, *Klf4, Oct4, Runx2, and TGF-β* on the single cell level. A multiplex approach allowed an increased number of data points to be collected and analyzed from the same sample, removing variables (including human and handling errors) that inevitably arise when multiple experiments are performed.

Aims of this Study:

The knockout studies discussed above have demonstrated that the genes *Klf4*, *Oct4*, and *Runx2* play a role in the vSMC phenotype switching that occurs when smooth muscle cells dedifferentiate into heterogeneous atherosclerotic plaques and differentiation into osteoblast phenotypes. However, the precise function of each of these genes in vSMC phenotypic switching remains unclear. Previous studies have demonstrated that since atherosclerosis involves so many cell types and circulating blood serum particles, it is difficult to recapitulate *in vitro*. Therefore, we used an ApoE -/- mouse disease model mouse, which, along with a high fat diet, ensures development of atherosclerosis in mice.

We used a multiplex lineage tracing approach to permanently tag gene loci based on unique markers for each gene. Mass cytometry of these tagged cells can show the heterogeneity of cell types in a plaque as well as the molecular history of gene expression in those cells. Mass cytometry (using CyTOF) is advantageous in this project because it is able to measure and sort over 40 markers simultaneously on the single-cell level. Since the main goal of this project is to determine the role of multiple genes involved in vSMC phenotype switching during atherosclerosis and plaque thrombosis, a multiplex lineage tracing system is preferred over a singleplex one. This is due to the multiplex system's ability to assemble

many markers into a combinatorial analysis of gene involvement, therefore removing extraneous variables which would surface if multiple single gene analyses were compiled instead.

In this study, we created a multiplex lineage tracing system to demonstrate the specific roles of the four genes, *Klf4*, *Oct4*, *Runx2* and *TGF-β*, in vSMC dedifferentiation and differentiation pathways during atherogenesis and in atherosclerotic states. *Oct4*, *Runx2*, and *TGF-β* are candidates being pursued in the Zunder lab; however, this thesis project specifically focused on the design and insertion of *Klf4* lineage tracing components.

METHODS

PART I: SYSTEM DESIGN:

The first step of this project was to design a lineage tracing system that would function efficiently with our specific gene and cell-type targets. The project design to target genes of interest involves knocking recombinase elements in at the gene loci of interest. These recombinases will act on recombinase recognition sites to allow for expression of fluorescent proteins only when genes of interest are actively expressed. Recombinase will be knocked-in for *Klf4* using a plasmid, 2AC Neo2, designed in the Zunder Lab. Knocked-in recombinase elements, which include Cre, Flp, VCre, and SCre, will be separated from the gene's coding sequence by a 2A self-cleaving peptide sequence (P2A). The knock-in will be done using homology directed repair (HDR). HDR uses homology arms complementary to

the DNA that flanks the target knock-in locus. Double stranded breaks are induced in the DNA, and the homology arm sequences, which flank the sequence to be knocked in, are added. The homology arms match with the DNA around the insert site and when the double stranded breaks are repaired, the knock-in sequence is incorporated (*Figure 3).*[74]

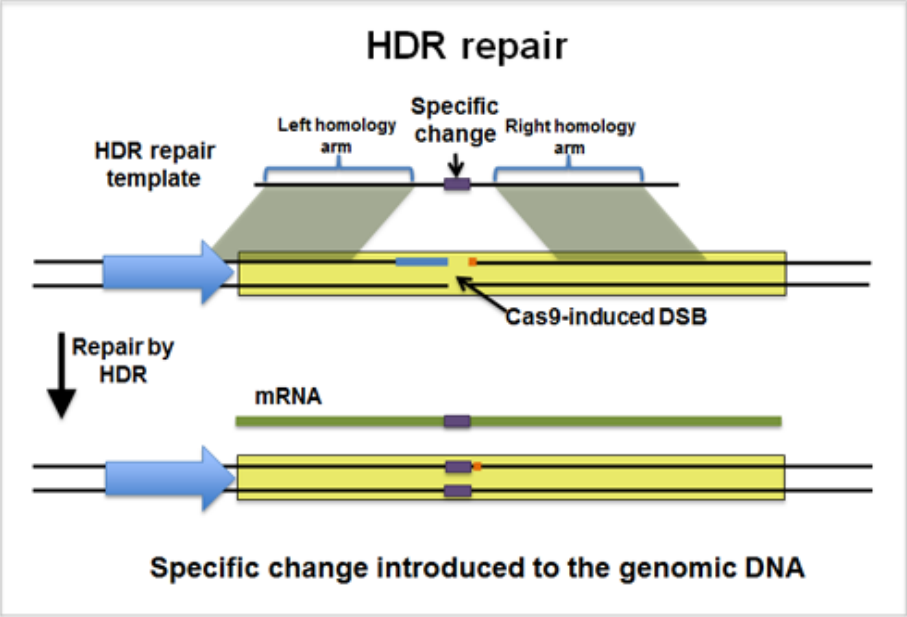


Figure 3: Illustration of homology directed repair (HDR) technology.[74]

Fluorescence expressing reporter motifs in our system rely on Cre-LoxP and analogous recombination elements. Cre is a recombinase that was first isolated from bacteriophages. When Cre is expressed, if it encounters LoxP sites that are unidirectional (facing the same way), Cre will cut out the intermediate sequence (*Figure 4).* Flp/FRT, VCre/VLoxp, and SCre/SLoxP systems work in this same way; however, each recombinase, Flp, VCre, and SCre, only acts upon its respective recombinase sites, FRT, VLoxP, and SLoxP.

In order to trace gene expression, reporter motifs with fluorescent markers will be inserted downstream of constitutively active promoters located at the ROSA26 locus. Since four genes are being traced in this system, four reporter motifs, each associated with a unique fluorescent protein, will be used. Use of four separate promoters allows for orthogonal lineage tracing of the individual genes of interest because each reporter is able to express its respective fluorescence independently of the activity of the other motifs. Ubiquitous expression of reporter motifs is essential to system function so that whenever target genes are expressed, fluorescence will be expressed. To achieve this, each reporter motif was inserted downstream of a ubiquitous promoter and all four reporter motifs were inserted at the ROSA26 locus using pROSA26-1 plasmids. The ROSA26 locus is used because of its ubiquitous chromatin accessibility.

As mentioned, a different color fluorescent marker is associated with each gene (*Figure 4)*. Fluorescent proteins GFP (green), mKate2 (red), mCerulean (blue), and PhiYFP (yellow) were used as markers for *Klf4, Oct4, Runx2* and *TGFb*, respectively. Each of these fluorescent markers has a unique epitope tag attached to allow for measurement of expression via mass cytometry. Upstream of each fluorophore, there is a "negative" fluorescent marker, which is a mutant form of PhiYFP (mPhiYFP). mPhiYFP, derived from the Brainbow system, acts as a scaffold upon which metal-bound epitope tags can bind to allow for detection of negative markers using CyTOF.[75] The mPhiYFP motif upstream of GFP, mKate2, mCerulean, and PhiYFP, are flanked by LoxP, FRT, VLoxP and SLoxP sites, respectively. mPhiYFP, which is followed by a poly(A) tail, acts as a stop cassette so that when it is upstream of a fluorophore, the fluorophore is not transcribed. Therefore, only when Cre, Flp,VCre, or SCre are expressed, will the negative marker be removed to allow for

expression of the fluorescent proteins (*figure 4)*. The negative tags for each reporter motif have small nucleotide variations that preserve amino acid sequence. This allows markers to be identified using RNAseq while all are still recognized as epitope scaffolds during CyTOF analysis.

The plasmid used for insertion of reporter motifs, pROSA26-1 is a publicly available, validated ROSA26 plasmid. It contains an ampicillin resistance cassette for positive selection and a DTA diphtheria toxin negative selection marker. In the Zunder Lab, the pROSA26-1 plasmid was modified using synthetic gene sequences to include an mRNA splice acceptor and neomycin resistance cassette, resulting in the pRosa26-SA-NeoR plasmid. The splice acceptor strategy was used to allow the maximum number of unique promoters to be used in our reporters.[76]

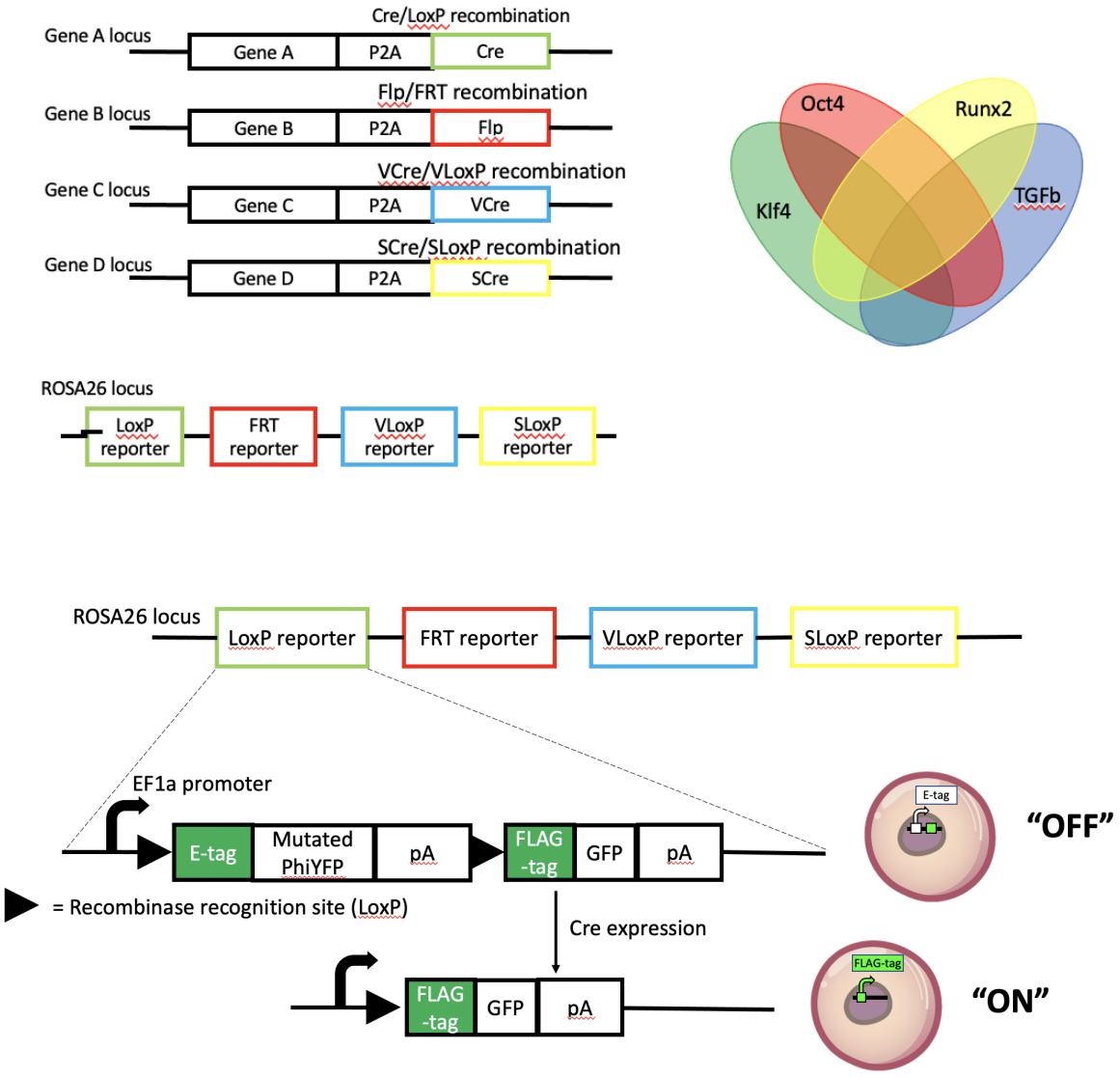


Figure 4: Outline of lineage tracing technique: Top Left: genes of interest will have P2A recombinase elements inserted after the gene's final exon and before the stop codon. Top Right: simplified visualization of ideal combinatorial lineage tracing in results, which depict gene involvement and overlap. Middle: reporter sequences inserted after the ubiquitous ROSA26 gene locus. Bottom: visualization of negative marker mPhiYFP floxed by unidirectional loxP sites. The first sequence here, with two LoxP sites, is the reporter motif state prior to Cre expression. This signifies the "OFF" or negative state because mPhiYFP is expressed but GFP is not. The second sequence in this section illustrates the reporter motif when Cre is expressed: the mPhiYFP motif is removed, only one LoxP site remains, and GFP is able to express, signifying the "ON" state.

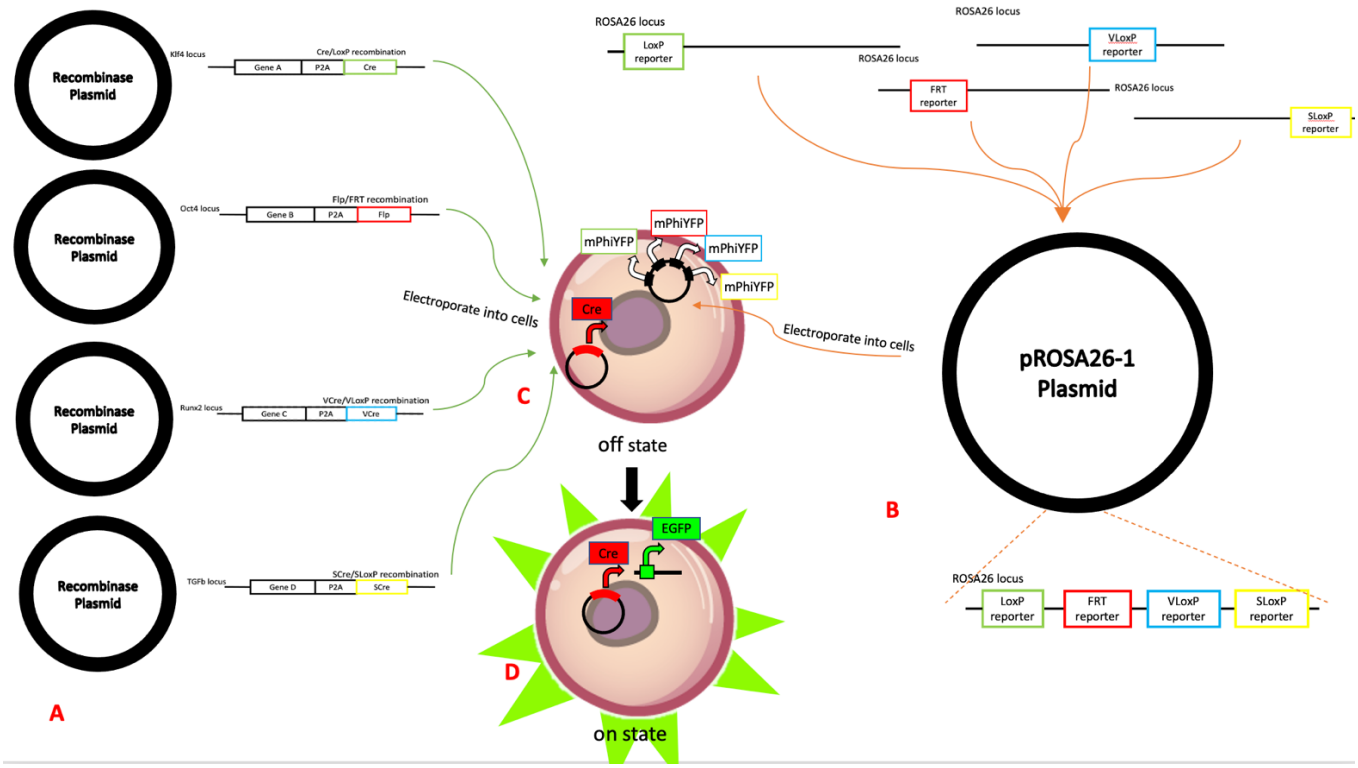


Figure 5: Concept Map of lineage tracing protocol: A: as described above, recombinases were knocked-in downstream of genes of interest using HDR. Each gene will be knocked into its own plasmids and replicated separately before insertion into mESCs. Plasmid 2AC Neo2 was used to knock in the lineage tracing elements for Klf4, the gene of interest in this thesis. B: All four reporter motifs, each of which acts as a tag for a gene of interest are inserted into a pROSA26 targeting plasmid after constitutively active promoters. DNA from the plasmid is isolated and, using nucleofection, is inserted into mESC cells. C: the cell is in the "OFF" state for Cre, the recombinase inserted after Klf4. As such, we see mPhiYFP, not GFP. Absence of Cre expression is signified by the Cre motif's position outside of the nucleus. D: After Cre is expressed in a cell, the cell transitions into its "ON" state, expressing EGFP (enhanced GFP) because active Cre expression led to removal of the negative cassette between the LoxP sites.

Recombinase Gene Targeting:

The plasmid 2AC-Neo2, was designed in the Zunder Lab and contains necessary lineage-tracing elements for *Klf4*. Its name is derived from its composition, with 2A referencing the P2A region, C referencing the Cre region, and Neo referencing the region of neomycin resistance. Upon insertion, *Klf4* homology arms containing the gene's final coding exon will be followed immediately by P2A, Cre, followed by an ERT2 sequence, Neomycin resistance flanked by unidirectional attB/P sites for excision with PhiC31, the stop codon, the

3’ homology arm and diphtheria toxin catalytic domain (hDT) behind a ubiquitous promoter (PGK in this case) (*Figure 6*). The P2A sequence inserted between the Klf4 and Cre coding sequences allows for the Cre protein to be translated in conjunction with the Klf4 sequence and then cleaved to allow for individual expression of both proteins. This allows for the best fidelity of recombinase expression to Klf4 protein expression without potential confounds from regulation at the DNA or RNA level. Cre recombinase then is able to act upon LoxP sites associated with reporter motifs.

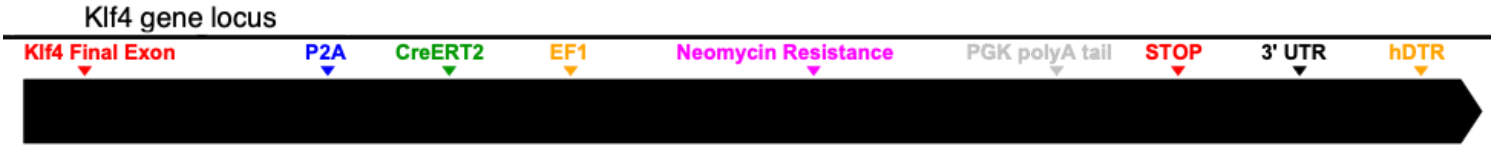


Figure 6: Klf4 gene locus after insertion into the 2AC-Neo2 plasmid at the target locus

Neomycin resistance (NeoR) acts as the positive selection marker, allowing only cells that incorporate the insert into their genome to survive treatment with neomycin. The hDT motif downstream of the 3’ UTR acts as a negative selection marker. If incorporated, diphtheria toxin catalytic domains cause cells to synthesize diphtheria toxin and therefore kill themselves. Thus, the *hDT* gene can be used as a negative selection marker because it is unlikely to insert in the intended HDR. If recombination occurs outside of the target homology arms during nucleofection of mESCs with the plasmid, the *hDT* gene will be inserted and diphtheria toxin will be synthesized. Therefore, only cells with off-target inserts will be eliminated from the population due to diphtheria toxin.

System Validation in mESCs:

In order to validate the system described above, it was essential to confirm that it could be used to trace genes known to be involved in already identified differentiation pathways. This step was performed in parallel with the experimental steps below. Validation involved a proof-of-concept experiment using genes outside of the VSMC family to confirm our procedure's efficacy. The genes *IslI, Sox2, Pax6, GATA4,* and *Nkx2-5* were used because they have been proven to become active or increase expression during differentiation stages. These genes therefore acted as good candidate genes to test the lineage tracing system.[57,58,59] Insertion of reporters was validated by PCR to confirm presence at the correct genomic locus. Reporter function was validated by transfecting recombinases and detecting epitope tags and fluorophores by flow cytometry

PART II: COMPLETED METHODS

Lineage Tracer Homology Arm Amplification:

To perform lineage tracing, knocking-in P2A-recombinase sequences at gene-specific endogenous loci using homology directed repair was attempted for the *Klf4* locus.

We first identified the stop codon of *Klf4* using NCBI, UCSC, and Uniprot *Mus musculus* (house mouse) genome data. [18,19,20] Since knock-ins will be done via homology directed repair (HDR), Benchling software was used to design about 1 kb long homology arms 5' and 3' of the *Klf4* final exon.[21]

Benchling software was used to analyze the target gene's homology arms to create viable primers for PCR amplification of the sequence.[21] Primers were analyzed and chosen

based on melting temperature (minimum of 63°C), GC content, homodimer ΔG and secondary structure. Homodimer ΔG, is a measure of stability of secondary structures, and thus describes how likely those secondary structures are to interfere with binding for PCR. If ΔG magnitude is too great, primers have a higher tendency to bind to themselves than to the target sequence. ΔG target range was set to between 0 kcal and -10 kcal. Homodimer secondary structure was also considered; if hairpins or otherwise stable secondary structures exist, it may inhibit primer binding to the target sequence and result in null results. Of note, secondary structures involved the 3' base pairs were considered more detrimental than secondary structures elsewhere.

Primers with acceptable ΔG values and homodimer secondary structures were identified and synthesized commercially via Integrated DNA Technologies (IDT). Next, PCR amplification of the 3' arm was carried out with and without the Q5 GC enhancer following standard PCR Q5 polymerase procedure using gradient PCR. BsiWI-HF and PacI restriction enzyme recognition sites were added by PCR. After amplification, the DNA fragment was digested with BsiWI-HF and PacI restriction enzymes and the subsequent digestion product was extracted and ligated into the plasmid 2AC Neo2.

Identification of efficient primers and subsequent PCR amplification attempts failed for the 5' arm. Therefore, Gibson assembly through IDT was employed to incorporate verified primer sequence complements onto the *Klf4* target sequence. Preliminary attempts at a *Klf4* 5' homology arm HiFi G-Block failed despite several attempts. A 14 base pair (bp) run of adenines was hypothesized to be a factor in the failed synthesis. To reduce G-block complexity score, we split the sequence within its run of adenines to create two separate G-Blocks. On the inside ends of the G-block (i.e., the ends that had been split to reduce

complexity score) we used BsaI to join the sections of the homology arm and preserve the run of adenines. Golden gate ligation techniques allowed us to ligate the two fragments simultaneously. A validated primer sequence was also added to allow for amplification of the G-block. Finally, recognition sites for Type IIS restriction enzyme BsaI-HF were added to the outermost ends of the G-block sequences to be used for ligation of the 5' arm into our 2AC Neo2 plasmid.[55]

Upon receiving the G-block parts from commercial vendor IDT, gradient PCR amplification was performed on the synthesized sequences following standard PCR Q5 polymerase procedure to optimize the gene fragment. Gel electrophoresis was run to confirm expected band size.

Insertion of Reporter Motifs:

Next, reporter motifs were inserted at the ROSA26 locus using restriction enzymes. Reporter sequences were first inserted individually into separate pRosa26-SA-NeoR plasmids to test reporter motif function. Double digest restriction digest protocols were followed for each reporter motif (LoxP, FRT, VLoxP, or SLoxP) as well as for the four Rosa26 targeting vector backbones these reporters were inserted into. The restriction digest was done using equal concentrations of MluI-HF and AscI restriction enzymes. After digestion, gel electrophoresis and gel extraction of target fragments from each agarose gel was performed using Zymo Research DNA Recovery Kit for each reporter motif and its backbone plasmid.[54] Next, ligation of each reporter motif into a pRosa26-SA-NeoR backbone was performed according to standard ligation protocol, using NEB Quick Ligase and a 1:10 molar ratio of vector: insert.

Next, Zymo E. coli Transformation protocol was followed to transform the plasmids containing single reporter motifs into E. coli. The transformed E.Coli were spread onto warmed culture plates containing Ampicillin and incubated to select only for bacteria that underwent successful transformation. Surviving colonies were selected for manually and cultured to expand.

Zyppy Miniprep kits (from Zymogen) were used to extract inserted plasmid DNA from E. coli cells.[54] Analytical digests were designed using available restriction enzyme pairs to differentiate successful ligation form parental plasmid or reverse insertion.[20] AscI and EcoRV were used for analysis of pRosa26-LoxP, pRosa26-FRT and pRosa26-VLoxP and resulted in bands that distinguished whether the target sequence inserted in the correct orientation. XhoI and NruI resulted in distinguishable bands for pRosa26-SLoxP (*Appendix A*). Analytical digests were then performed for each reporter sequence using double digest protocols. Gel electrophoresis was performed and bands were visualized for analysis.

To combine all four reporter motifs into one plasmid, a cloning strategy using compatible sticky ends was employed. We used MluI and AscI, which both have CGCG sticky end sequences. In our system, MluI was used 5' of each reporter sequence and AscI was used 3'. Parental plasmids are opened using just the AscI site, so reporters can be added sequentially to the 3' end of the series and kill the 5' AscI/MluI site. In SloxP, there exists a second AscI cut site within the promoter sequence so BbsI was used to create compatible sticky ends instead of AscI.

After each step, analytical digests and sequencing were performed to demonstrate that motifs had been successfully inserted into the Rosa26 plasmid and no mutations had occurred. Analytical digests were designed and performed to ensure application of chosen

restriction enzymes for each reporter resulted in bands that clearly distinguished whether the target sequence had inserted in the correct orientation, the reverse orientation, or not at all.

Rosa26-reporter plasmids were transfected via nucleofection into E14TG2a mESCs. After nucleofection, cells were allowed to settle for 2 days, then positive selection was performed by adding neomycin to the plates. Selection was done for one week. Over this week, if off-target or over-extended HDR occurred, presence of the diphtheria toxin catalytic domain caused these cells to make diphtheria toxin and die. Colony selection was then done manually.

Colonies were then screened both genomically and functionally. Genomic screening involved PCR amplification followed by gel electrophoresis analysis to ensure that the DNA inserted in the correct locus and orientation.

Functional screening involves confirming that reporters express tags before and after transfection with recombinase. This is done by transfecting and without Cre, Flp, SCre and VCre and testing whether fluorescence can be detected when recombinases are present as well as testing for the presence of the expected epitope tags and fluorophore by flow cytometry.

After the preliminary genomic sequencing, the colonies were treated with phiC31 recombinase, which targets the unidirectionally attB/P flanked neomycin resistance cassette and removed it. After removal of this resistance cassette, PCR and gel electrophoresis were again performed to confirm presence of expected bands.

To validate the reporters by flow cytometry, antibody concentrations for epitope tags needed to be titrated. Antibody concentrations for LoxP and FRT negative marker epitope tags had been previously validated in the Zunder Lab but SLoxP and VLoxP negative marker

binding antibodies needed to be validated. To achieve this, Cytobank software was used to determine the most efficient titrations of antibodies. We first made expression plasmids for each epitope tag using pcDNA3.1 backbone, linking the epitope tag to mPhiYFP. These plasmids were transiently transfected into HEK293T cells using Turbofect transfection. Cells were fixed for 10 minutes with 1.6% PFA, permeabilized for 10 minutes with -20°C methanol, and stained with intracellular antibody for 1 hour at room temperature (20-25°C).

OLLAStag was titrated for the VloxP negative reporter and T7tag was titrated for the SloxP negative reporter. The OLLAStag antibody conjugated to a dye and T7tag purified antibody were purchased. T7Tag antibody-dye conjugation was performed by mixing 50 µg of purified antibody with NHS ester dye in 100 µM NaHCO3. The NHS ester dye was used at three concentrations: 60 ng/µL, 30 ng/µL, and 15 ng/µL. As such, the T7tag antibody needed to be titrated across these different dye concentrations. Dye concentrations were denoted as T7Tag hi, T7Tag mid, and T7Tag lo for the 60 ng/µL, 30 ng/ µL, and 15 ng/µL concentrations, respectively. For each of the four antibodies being studied, antibody concentrations 0.01/100 µL, 0.03/100 µL, 0.1/100 µL, 0.3/100 µL, 1.0/100 µL, 3.0/100 µL and 10.0/100 µL were analyzed for maximal separation of positive and negative fluorescent signal. For each antibody at each concentration, Cytobank was used to draw gates which allow for "positive" fluorescent tags to be grouped and delineated from "negative" fluorescent tags.[64] For each concentration, the following equation was used to calculate separation indices, which determine which concentration of each antibody results in clearest separation of positive and negative signals:

$$Separation\ Index\ =\ \frac{(MedianPositive)-(MedianNegative)}{((84\%\ negative-MedianNegative))/0.995}.$$ [56,64]

Murine Aorta Phenotyping by Mass Cytometry:

In anticipation of using this lineage tracing system in atherosclerotic mice, we analyzed a mass cytometry dataset collected in the Zunder lab. The cell types present were determined by UMAP dimensionality reduction in R ("RunUMAP") and Leiden clustering in Python (*Appendix C*).

Wild type UMAP clusters were created using R Studio. Cells from *Mus musculus* (house mouse, C57/Bl6) aortas were used. These mice were fed on chow and showed no evidence of atherogenesis, established atherosclerotic plaques, or CVD. C57/Bl6 aortic cells were isolated from mice following NIH and IACUC policies.

PART III: METHODS PLANNED AND IN PROGRESS:

Due to COVID-19, wet-lab time was severely limited and the following protocols were temporarily postponed. They are expected to be completed in the next few months. These methods are included, written as planned, not only because part of this thesis project involved designing protocols for future use, but also because they are integral to a complete understanding of the experimental processes and goals.

Lineage Tracer Incorporation into Plasmids and E. coli:

Once both the 3' and 5' arm homology arms have been amplified, restriction digests of both *Klf4* homology arm sequences and the 2AC Neo2 plasmid template are performed in preparation for ligation of the homology arms into our plasmid. Where available, optimized NEB High Fidelity (HF) restriction enzymes were used to reduce star activity. The 3' arm will be inserted first. A double digest protocol should be performed first on the 3' arm of *Klf4*

using equal concentrations of each restriction enzyme. Then a double digest reaction is used to digest the template 2AC Neo2 in preparation for ligation of the 3' arm.[55]

After restriction digestion of the 3' homology arm and plasmid are complete, gel purification will be performed to isolate the 3' arm and plasmid DNA fragments for ligation. Gel extraction from each gel is then performed using Zymo Research DNA Recovery Kit after excising fragments from the agarose gel.[54] After excision, purified 3' homology arm and 2AC Neo2 backbone DNA will be obtained.[54]

The purified 3' *Klf4* homology arm is then ligated into the prepared 2AC Neo2 plasmid with standard NEB reactant components and protocols. A 1:10 molar ratio of vector: insert should be used. The plasmid is then transformed into E. Coli using Zymo E. coli Transformation protocol.[54] Transformed E.Coli are spread onto warmed culture plates containing Ampicillin and incubated to select only for bacteria that underwent successful transformation. Surviving colonies are selected for manually and cultured to expand.

Before restriction digest and ligation of the 5' arm G-block into the plasmid is performed, the two parts of the G-block must be conjoined. A Golden Gate ligation using NEB Golden Gate Kit will be used. Golden gate ligation includes both restriction enzymes and ligase and therefore can perform digest and ligate steps in the same reaction.

Once the conjoined 5' arm G-block is achieved, a single enzyme restriction digest protocol using BsaI HFv2, a high fidelity type-IIS restriction enzyme, can be used on both the 2AC Neo2 plasmids that have already been ligated with the 3' homology and the G-blocks. After digestion of the G-block and plasmid, gel electrophoresis is performed on the DNA fragments and gel extraction is performed to isolate the vector and insert DNA components.

Ligation of the 5' arm into the 2AC Neo2 plasmid (which was already ligated with the 3' homology arm) is then performed. After this ligation, 2AC Neo2 plasmids are expected to contain both the 3' and 5' homology arms of *Klf4* flanking our lineage tracing components (*Figure 6)*. Plasmids will be transformed, selected for using antibiotics, and expanded, as before. Analytical digests and sequencing will be performed to confirm results.

When future *in vivo* insertion of these motifs is performed, ethical considerations will include handling of mice, which will be done in accordance with the University of Virginia's Animal Care and Use Committee (ACUC) and NIH Guidelines for Care and Use of Laboratory Animals.

RESULTS

PART I: SYSTEM DESIGN:

A major aspect of this project involved designing and understanding an efficient and effective system capable of lineage tracing multiple genes simultaneously. The overall goal was to add permanent, gene specific markers that allow for identification of single-cell-level expression histories in atherosclerotic plaques. This lineage tracing information can be coupled with cellular expression data to give a detailed picture of cell phenotypes and their expression histories. Our lineage tracing system and protocols were designed and implemented as described in methods.

PART II: EXPERIMENTAL RESULTS

Lineage Tracer Homology Arm Amplification:

Homology arm primers used in this HDR attempt were designed to be about 1 kB 5' of the stop codon for the 5' homology arm and 1 kB 3' of the stop codon for the 3' homology arm.

To amplify the *Klf4* 3' homology arm, two primer pairs, which were identified to have the best ΔG and homodimer structure (GC contents of 55-59%) were tested. Primer pair B, (Figure 5, top right) provided the most distinct bands at 67°C annealing temperature with no GC enhancer. This primer pair was used for amplification of the 3' arm.

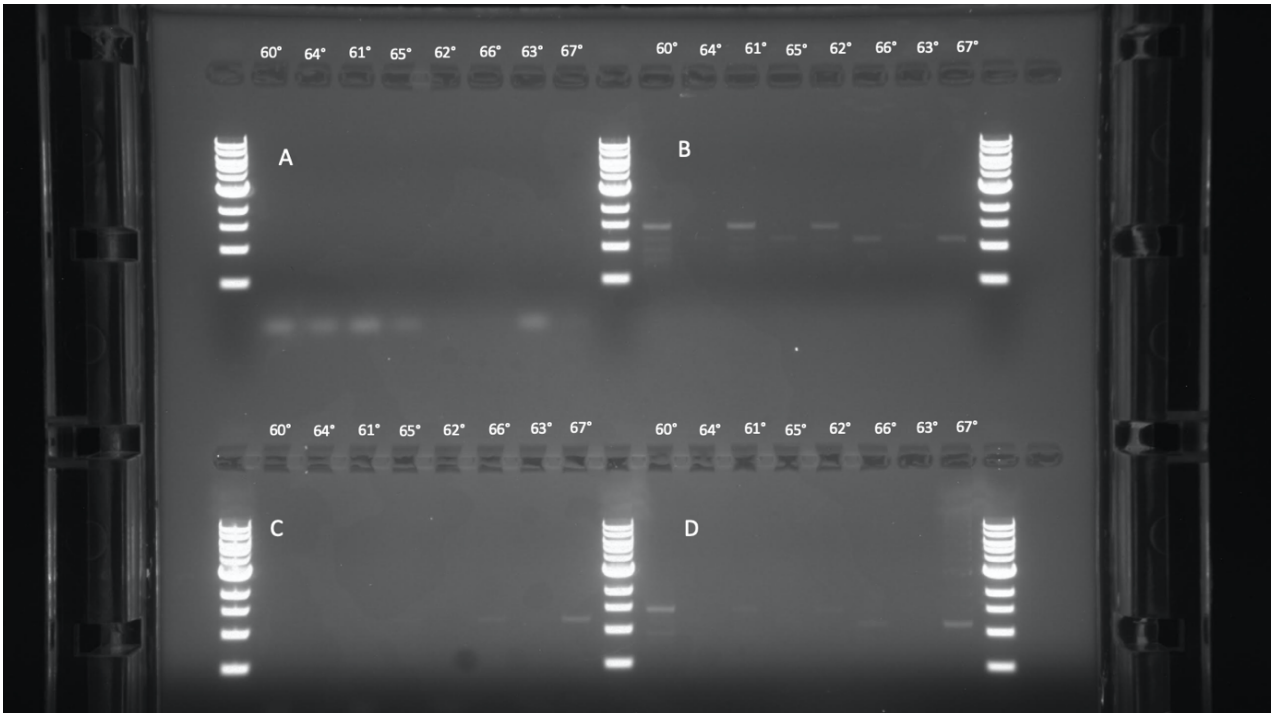


Figure 7: PCR results to identify best primers to use for amplification of Klf4's 3' homology arm. PCR temperature was run on a gradient spanning 8°C with mid 63°C. Well loading was done in the following order with 1 referring to the lowest and 8 to the highest temperatures: 1, 5, 2, 6, 3, 7, 4, 8. Gel layout is: A: Primer pair A, no GC enhancer, B: Primer pair B, no GC enhancer, C: Primer pair A + GC enhancer, D: primer pair B +GC enhancer. : Primer pair B, no GC enhancer at 67°C was chosen as the most efficient.

Genomic PCR of the 5' homology arm failed because PCR and gel electrophoresis produced no visible bands. We hypothesize that this was due to unsuitable primers. Homodimer ΔG, for both the forward and reverse primers for the *Klf4* 5' homology arm were found to be outside the acceptable range, measured at -11.4 kcal and -15.4 kcal, respectively, signifying that primers bound to themselves with a very high tendency and formed stable secondary structures.

Gibson assembly through commercial DNA synthesis via IDT HiFi G-Block was approached as the next attempt for isolation of the 5' homology arm. Gibson Assembly is an exonuclease-based DNA synthesis technique in which up to 15 fragments of DNA can be combined using a 5' nuclease to generate long overhangs, which are completed using a polymerase and ligated with DNA ligase.[22] Using this technique, novel sequences can be established through combination of DNA fragments. HiFi G-Blocks produced by IDT are double-stranded fragments of DNA that can be assembled using Gibson Assembly. HiFi G-Blocks allow for assembly of larger and more complex sequences than traditional Gibson Assembly methods and were therefore chosen for this project. Sequence complexity scores, as determined by IDT online software, were promising for HIFI G-block synthesis over standard G-blocks. In this case, HiFi G-Block assembly provided a method for adding validated primers and BsaI-HF recognition sites onto the 5' arm to allow for amplification. Addition of BsaI-HF cut and recognition sites was advantageous because we were able to establish validated cut sites directly outside of our target sequence.

Preliminary attempts at a *Klf4* 5' homology arm HiFi G-Block failed despite several attempts. I hypothesized that synthesizing two separate G-Block fragments, by splitting the run of adenines, would sufficiently reduce the complexity for successful synthesis. Splitting

the sequence shortened the run of adenines to 6 in G-Block Part I and 8 in G-Block Part II. Splitting the sequences also effectively reduced the sequence's tendency to form hairpins and allowed for synthesis of the entire homology arm. Thus, we successfully lowered the G-block's complexity score enough to allow for synthesis of the G-Block while creating distinct, complementary sticky ends along the split site that allow the two parts to insert into the plasmid as one conjoined sequence in the correct orientation. Previously validated primers were included on the ends of the G-Block to allow for amplification and multiple ligation attempts with a restriction digest and ligate protocol using BsaI-HFv2.

After the split G-Blocks were synthesized, PCR amplification and gel electrophoresis were performed on each fragment. PCR amplification was successful, as shown in *Figure 7.*

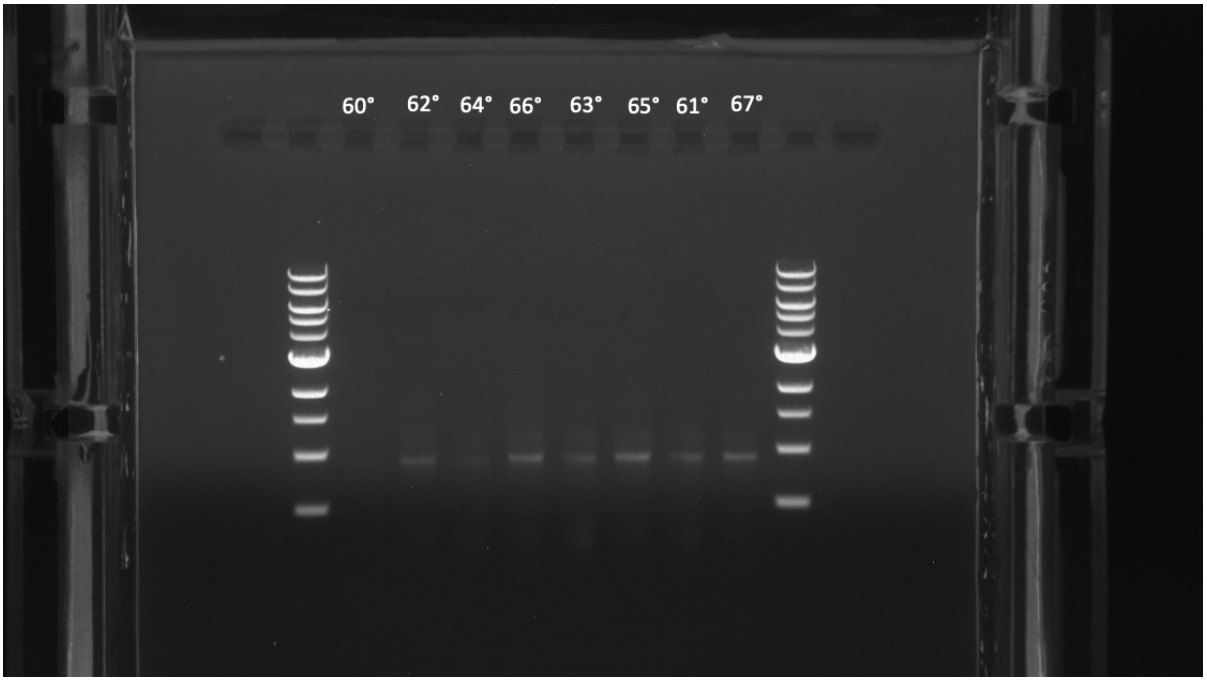


Figure 8: Gel electrophoresis results after PCR amplification of Klf4 5' Homology Arm G-Blocks I and II. PCR was run on a gradient spanning 8°C with mid 63°C. Well loading was done in the following order with 1 referring to the lowest and 8 to the highest temperatures: 1, 3, 5, 7. 2. 4, 6, 8.

Once the *Klf4* target homology arm insertion fragments are obtained and prepared for ligation, restriction digests using restriction enzymes PacI and BsiW-HF to insert the 3' arm followed by use of Type IIS restriction enzyme BsaI-HFv2 to insert the 5' arm will be performed. Areas of concern for efficacy of this digest and ligate process include the ability to discern whether the gene fragment has inserted into the plasmid, if the gene fragment inserted in the correct orientation, and whether methylation on the DNA will interfere with restriction enzymes. Presence and orientation of insertion will be determined using analytical digests. Methylation is used by bacteria as a natural mechanism of protection against restriction enzyme damage to their genomes. As such, methylation was investigated in depth as a possible area of complexity. CpG and Dcm methylation were of interest for the *Klf4* fragment because they are methylation types which have high potential to inhibit the restriction enzymes used in our target area of the *Klf4* gene and have been shown to impair restriction digest efficacy. Though CpG methylation has been demonstrated to block BsiW-HF and BsaI digestion, CpG methylation is only present in eukaryotic genomes and is not conserved after cloning into a bacterial host. Since all cloning in this project was done in E. Coli, this type of methylation was not of concern for us. Dcm methylation, on the other hand, had potential to interfere with BsaI in our system. Dcm methyltransferases methylate carbon 5 at the second cytosine position of CCAGG and CCTGG motifs. Dcm methylation is thought to have evolved as a protective mechanism to cleave by restriction endonucleases. Benchling Software and NEB databases were used to analyze the *Klf4* sequence to determine whether methylation within the gene would complicate successful transformation of the DNA into bacteria. Dcm methylation sites are present in the target area of the *Klf4* gene, they

do not directly overlap with restriction enzyme sites, though methylation sites do exist adjacent to cut sites. If, in the future, methylation is suspected as a source of problematic *Klf4* function or targeting, transforming the 2AC Neo2 + *Klf4* homology arm plasmid into Dcm-/Dcm- mutated E.Coli will be attempted to eliminate methylation blocking.

Reporter Component Validation:

Each reporter motif was first validated independently before placing them into series. Each was transformed with the pRosa26-1 empty vector and treated with recombinase. Expected reporter status was observed.

Analytical digests were designed using Benching Virtual Digest software to determine whether reporter motif insertion after digestion and ligation steps was successful (*Appendix A)*. Restriction enzymes were chosen that produced gel electrophoresis results in which the empty backbone and reverse insertion outcomes could be differentiated from the correct, forward insertion. Of note, some of the restriction enzymes chosen resulted in bands for the empty backbone and reverse insertion which could not be distinguished. However, this was acceptable because both empty and reverse results reflect undesired plasmid products.

In practice, analytical digest results showed that one plasmid demonstrated bands which matched those expected for insertion of the LoxP reporter motif while two did for VloxP and three did for SloxP. For FRT, no bands which correlated with correct insertion of the reporter motif were observed. Insertion attempts for the FRT reporter motif will be repeated. If no correct insertion is observed again, other approaches, including application of

different restriction enzymes and analysis of possibly problematic methylation will be applied.

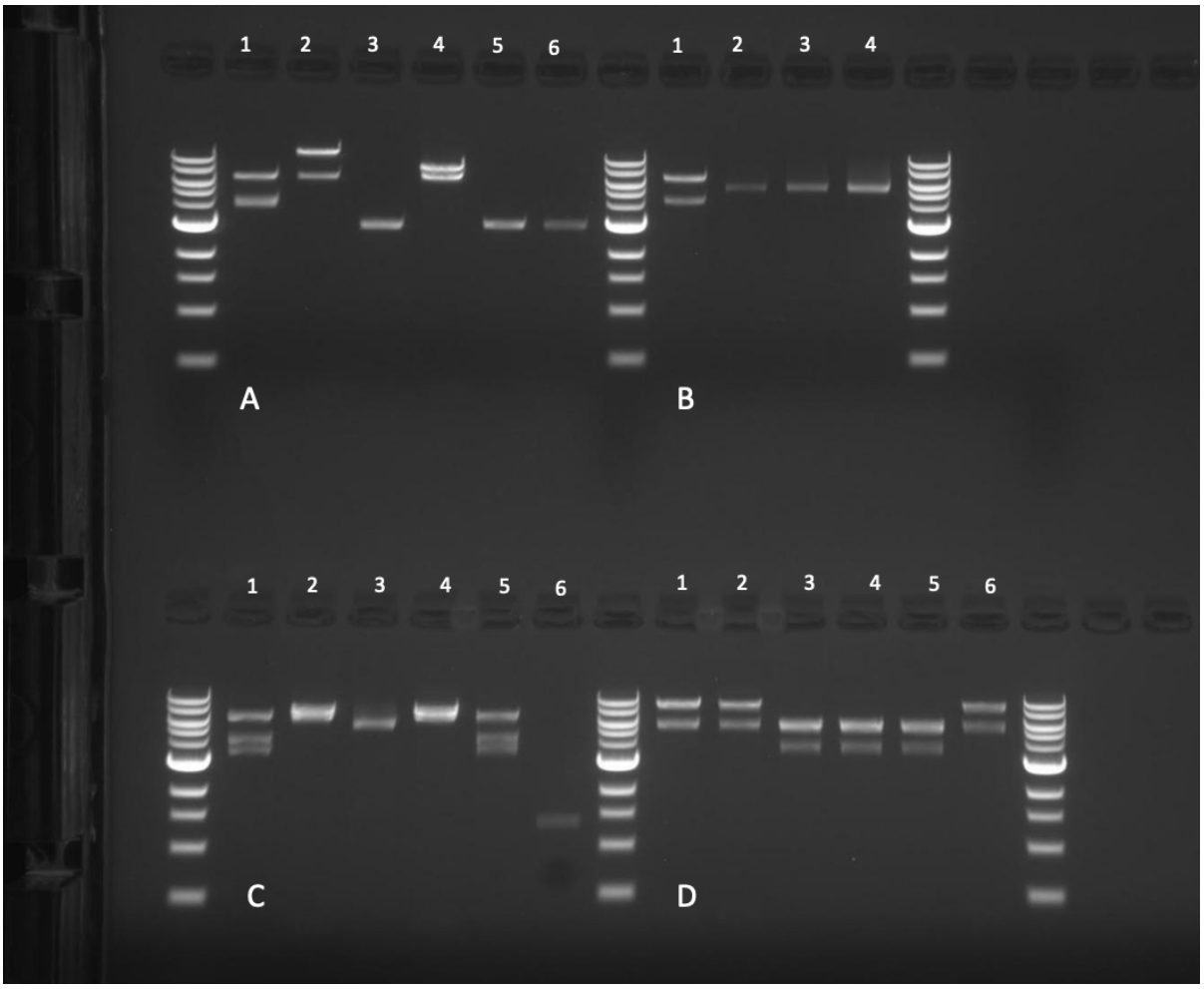


Figure 9: Results of analytical digest for reporter motif insertion into pRosa26-SA-NeoR backbone. A: LoxP, well 1 showed evidence of a plasmid with reporter motif inserted correctly; B: FRT, no wells showed a reporter motif inserted in the correct orientation; C: VloxP, wells 1 and 5 showed evidence of a plasmid with reporter motif inserted correctly; D: SloxP, well 2, 3, and 4 showed evidence of a plasmid with reporter motif inserted correctly. See Appendix A for expected bands.

Insertion of Reporter Motifs at Rosa26 Locus

Next, attempts were made to insert the validated reporter motifs in series at the Rosa26 locus. Previous transformation techniques failed to allow for read-through transcription of all reporter motifs; therefore, we attempted an approach that utilized

compatible restriction enzymes, which have different recognition sites but complementary sticky ends. We used MluI and AscI, which both have CGCG sticky end sequences but differing recognition sites. When MluI and AscI sticky ends match, recognition sites for both enzymes are eliminated. Therefore, the reporter motifs are inserted and the genomic sequence is conserved while restriction enzyme recognition sites are "killed."[51,52] "Killing" the recognition site ensures that when restriction enzymes are applied again for addition of subsequent reporter sequences, combined MluI and AscI sites will no longer be recognized and already inserted motifs will not be cut again.

A major concern when using this approach is that if there exist AscI or MluI cut sites within the reporter sequences or their promoters, the reporter sequences will be disrupted. This occurs in SloxP's promoter sequence, which contains a second AscI restriction site in its promoter. Therefore, BbsI, which has a sticky end compatible with MluI, was used instead of AscI[51]. While this resolves the issue of SloxP's second AscI cut site, another issue arises because BbsI is a Type IIS restriction enzyme with its recognition site upstream (3') of its cut site.[52] This means that when BbsI's compatible sticky ends pair with the downstream (5') end of SloxP, its recognition site is lost. Consequently, BbsI cannot be used again to attach another reporter sequence after SLoxP; SLoxP must then always be the last reporter sequence in the series. LoxP, VLoxP and SloxP have been successfully sequenced; however, validation of FRT insertion is still in progress. We are hopeful that results will be achieved in the coming weeks.

Antibody Titration for Flow Cytometric Reporter Validation

To validate reporter function by flow cytometry, we needed to first titrate the best staining concentration of fluorescent antibodies.

Substoichiometric and excess antibodies want to be avoided. Substoichiometric concentrations occur when adding more antibodies allows for increased binding, meaning that there exists unbound epitopes and too little antibodies. Substoichiometric antibody concentrations can result in such low signals that it becomes difficult to distinguish positive cells. Excess antibody occurs when all epitopes are bound and excess antibodies remain unbound. This causes increased off-target binding outside of the intended epitope, leading to increased positive signal from signal-negative cells. Optimal binding concentration is achieved when there is a low amount of unbound antibody and signal is at a maximum.

Cytobank software was used to draw gates around positive and negative signals.[64] Negative signals were defined as those identical to the empty vector sample and positive signals were defined as cells outside the population identical in the empty vector sample (*Figure 10).* There often exists a small amount of "leakage" outside the negative gate of the empty vector, usually from previous antibody titration tests. As such, negative gates were drawn to the border of the main island cluster. As mentioned, antibodies for LoxP and FRT reporters were previously titrated, so I analyzed OLLAStag, the negative signal for the VloxP epitope and T7tag, the negative signal for the SloxP epitope. Positive gates were found to contain a higher number of positive cells for VloxP and SloxP trials as compared to empty vector trials, as expected, which provided confirmation to proceed with analysis of these samples.

Separation indices were calculated for each concentration of each antibody (not including empty vector control trials), with the highest separation indices indicating the most effective concentrations. T7tag lo concentration 0.1/100 µL, T7tag mid concentration 0.1/100 µL, T7tag hi 3.0/100 µL concentration, and OLLAStag 0.01/100 µL concentration were found to be the optimal binding concentrations for these antibodies (*Appendix B).* Between T7tag lo, mid and hi variations, T7tag hi was observed to have the highest separation index and therefore T7tag bound to high dye concentration at a concentration of 3.0 µL is most likely the optimal binding state for this antibody.[64,56]

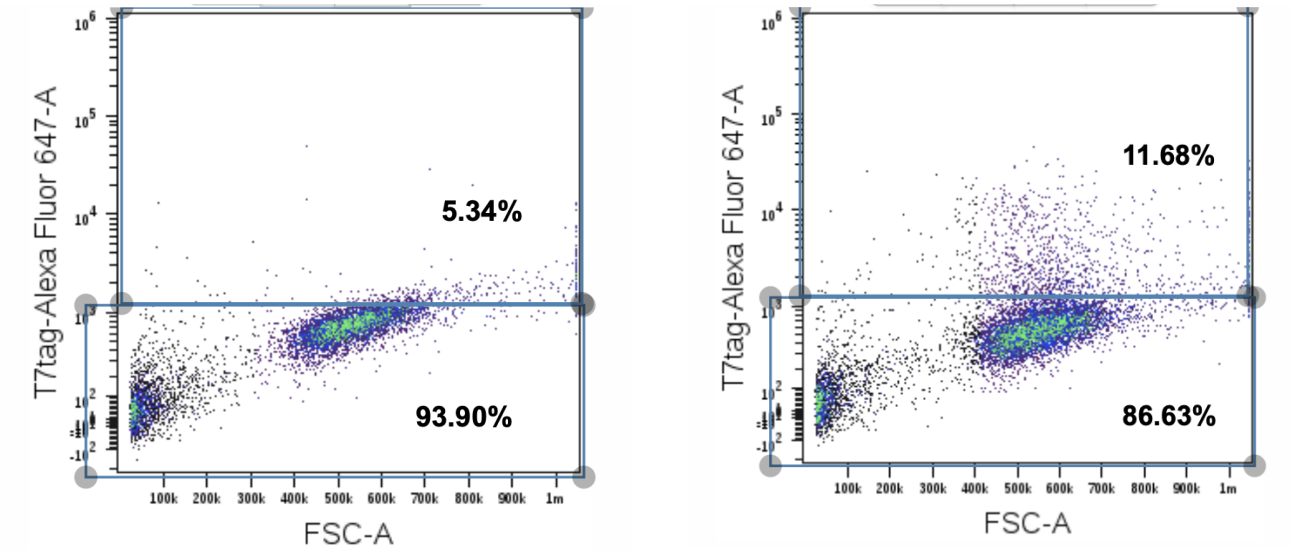


Figure 10: Cytobank gating to determine antibody concentration: Left: Empty vector gating for 0.01µL T7tag-mid antibody concentration. The negative gate is drawn to the edge of the main island and contains 93.9% of the signal. In this empty vector control, only 5.34% of the signal is "positive." Right: VloxP gating for 0.01 µL T7tag-mid antibody concentration. The negative gate, drawn to the same boundary as the control negative gate only contains 86.63% of this signal in this case, though the positive gate contains 11.63%.[64]

Lineage Tracing Analysis:

The lineage tracing system will be used in conjunction with cell profiling by CyTOF. CyTOF antibody panels involve substantial validation to ensure the cell types of interest can be observed using a proposed antibody panel. Here, I analyzed the heterogeneity of cell types expressed in the aortas of C57/Bl6 wild type mice by clustering expression data and creating UMAPs. Clustering allows for high-dimensional single-cell expression data to be aggregated by cell type, making genomic expression patterns in certain cell types evident. This clustering data can then be plotted onto UMAPs, where each color represents a different cell type and proximity of clusters indicates similarity in expression.

Using code written in the Zunder Lab, a heatmap of the wild type aortic cells was made (*Codes in Appendix C) (Figure 11).* Heatmaps are a major step in aggregation of data based on cell type and expression patterns. Tile colors for each gene are representative of gene expression levels and cluster colors, which will later appear in the UMAPs, are noted for each gene. Heatmap tiles can then be clustered and visualized on UMAPs, using code written in Python. This allows for visualization of emerging patterns in expression and makes obvious cells whose expression profiles are most similar (*Figure 12).*

The success of this experiment was important in two ways. Firstly, successful clustering and visualization of expression patterns of these cells using our code confirmed that our panel works effectively. Secondly, it created the heatmap and UMAP that will be useful in analyzing future disease state aortic cells from ApoE -/- mice. The comparison of this wildtype gene expression heatmap and UMAP clustering to ones synthesized for disease state data sets will make the changes in expression pattern between wild type and atherosclerotic states clear. In order to create clear single-cell level genetic histories,

heatmaps and UMAP clustering data will be synthesized at frequent intervals throughout the disease state and compared to the wild type expression data. This will allow for synthesis of single-cell level timelines of gene expression in atherosclerotic plaques.

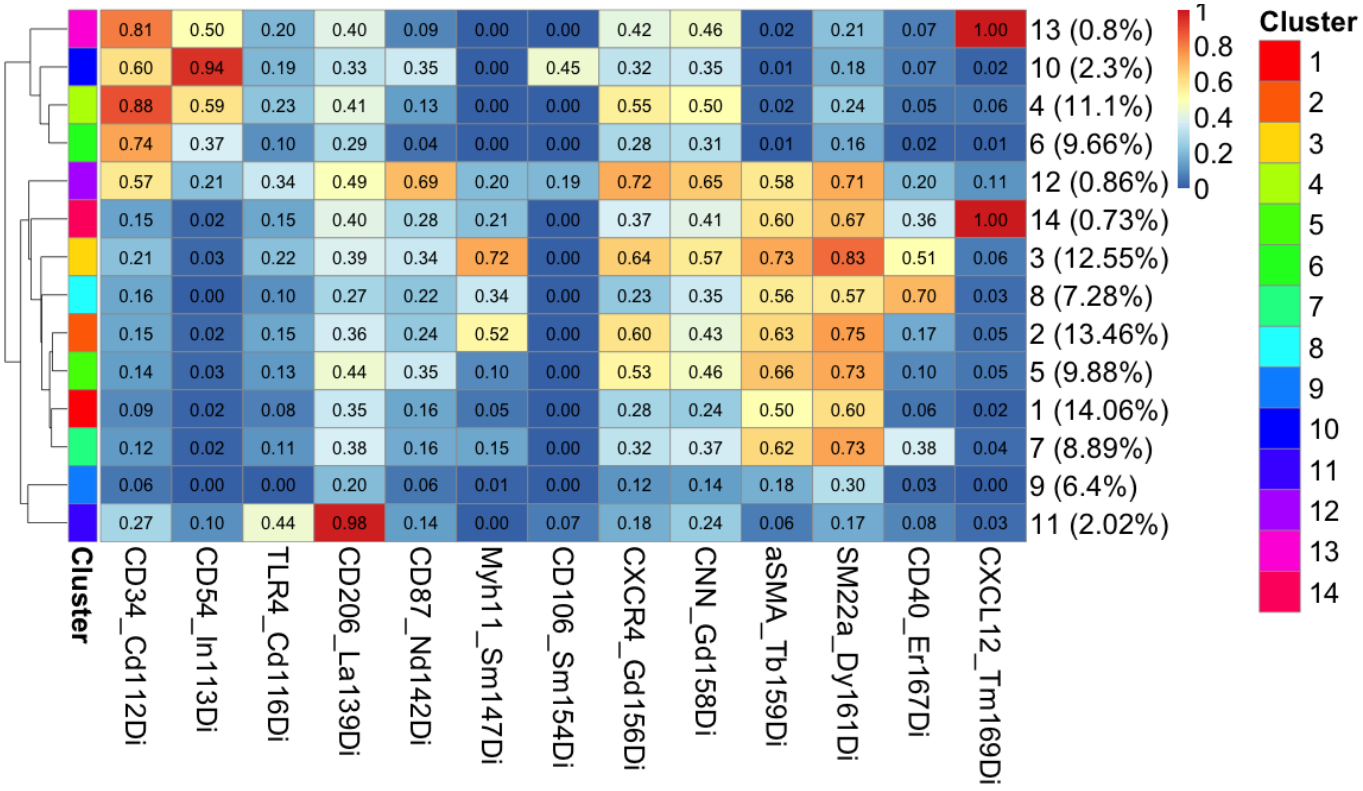


Figure 11: Heatmap from CyTOF data of wild type aortic cells of C57/Bl6 mice. Tile coloring represents gene expression level. Clusters are marked according to the color table on the right and displayed according to patterns of similar expression in the UMAP cluster map in Figure 12.

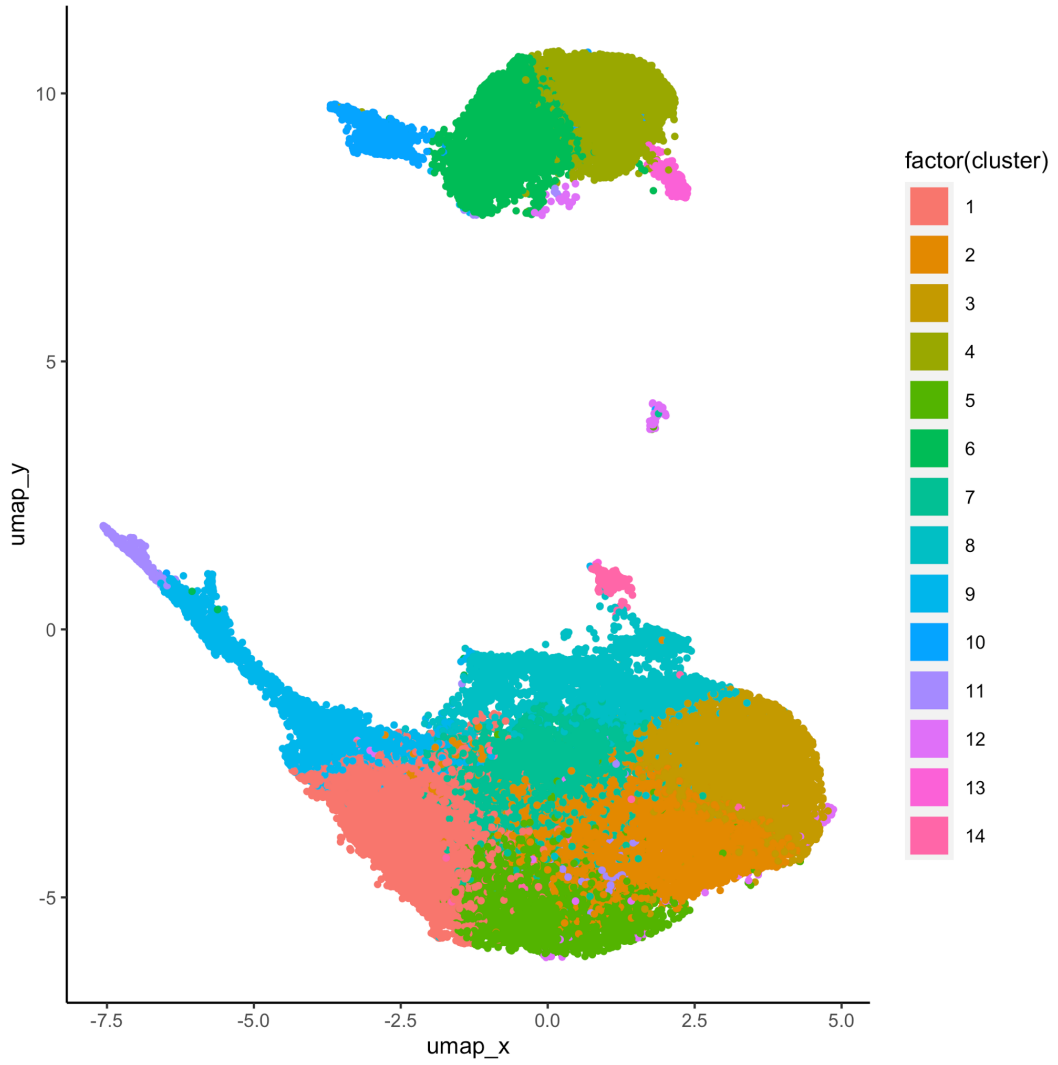


Figure 12: UMAP clustering derived from heatmap data shown in Figure 11. Clustering allows for visualization of expression patterns between cell types. Cell islands closer together on the UMAP have more similar expression and islands further apart of less similar expression. This wildtype clustermap will be compared to disease state cluster maps to analyze changes in cell expression.

PART III:

Lineage Tracing Plasmid Cloning:

Lineage tracer incorporation protocols were designed, written, and expected to be performed before the end of the 2021 academic year. Due to delays, these protocols were temporarily postponed.

Insertion of homology arms into plasmids is the essential step in gene incorporation. The 2AC Neo2 plasmid allows for selection to be performed after attempted integration. Isolation of plasmids also allows for target sequences to be transformed into E. coli, replicated, isolated from E. coli, nucleofected into mESCs, and eventually incorporated into mouse models. As such, this step, though temporarily stalled, is a crucial step in lineage tracing systems.

Transfection of Reporter DNA into Mammalian Stem Cells

DNA preparation for transfection can affect transfection efficiency and cell viability following transfection. Minipreps are used to achieve extraction of DNA from the E. coli cells that plasmids were housed in. Though maxipreps provide a higher DNA concentration and have lower endotoxin than minipreped DNA, it was found that sufficient DNA yield could be obtained through minipreps. Still, maxipreps have been considered through commercial vendor Biozilla and may be attempted to improve HDR efficiency if necessary in the future.

Next, nucleofection through electroporation and HDR will be attempted. Nucleofection program, which affects the voltage and wave profile of the electroporation, is the optimizable step in this procedure. In addition, linearization of plasmids prior to electroporation allows for increased efficiency while smaller plasmids, which yield shorter linearized strands, insert even more efficiently into cells. All experiments are first done with a GFP expression construct to measure transfection efficacy.

Initial attempts at transfection have been attempted with recombinase plasmids, but no colonies were observed, necessitating an adjustment in protocol to achieve results. Other

methods including cutting away more bacterial genome before linearizing plasmids, CRISPR or TILLING CRISPR will be attempted to achieve results and solidify a successful transfection procedure.

After successful transfection is achieved, a 5-day selection using neomycin will be performed and colony selection will be done manually then expanded. After validation of successful transfection using PCR and sanger sequencing via commercial vendor, resistance cassettes will be removed using Phi-c 31 recombinase, which will cut out the NeoR motif between the unidirectional attB/P motifs. Removal of positive selection markers is important so they do not perturb expression of the target genes and selection of multiple genes to perform the multiplex tracing is allowed. Genomic screening using PCR and gel electrophoresis should be done to confirm correct target gene insertion before and after removal of neomycin cassettes.

DISCUSSION

Atherosclerosis is a potentially life-threatening expression of CVD, which remains the most common cause of death worldwide.[1,2] While various pharmaceutical and surgical therapy options exist, many find procedure recoveries and strict, long-term drug regimens unsustainable.[38,39] Furthermore, individuals who possess major gene drivers of atherosclerosis or are statin intolerant often find that pharmaceutical therapies are insufficient to reduce their plasma LDL levels. As such, treatment of atherosclerosis would greatly benefit from breakthroughs in new, highly effective therapies and individualized medicine. While this

project does not directly address invention of new therapies, establishment of clear gene involvement, especially on a temporal scale, is a crucial primary step in conceptualizing gene therapy targets and mechanisms.

In this thesis, I aimed to establish lineage tracing elements for use in a mouse stem cell line for *Klf4* as well as achieve incorporation of reporter elements for *Klf4, Oct4, Runx2*, and *TGF-ß.* My work was conducted in parallel with work on tracing elements for *Oct4, Runx2,* and *TGF-ß* performed in the Zunder Lab.

While significant progress was made during this thesis project on the design and implementation of effective lineage tracing elements for the target genes, this study is still in its early stages. The success of preliminary tracing attempts provides support for the efficacy of our model. While many knockout studies of *Klf4* have been done, the gene is known to be difficult to target for sequence editing and knock-ins.[62] Therefore, our successful amplification and ligation results are quite promising in terms of maintaining *Klf4* as a future gene target.

Once lineage tracing elements are successfully established in a mouse population, single-cell lineage histories can be established. Lineage histories are an essential component of gene therapy design because, while knockout studies can establish involvement of genes, they are unable to analyze the nuances of gene expression. Combinatorial lineage tracing in atherosclerosis allows for identification of single-cell level gene expression at a point in time within plaques. By combining multiple successive disease state time point gene expression patterns, an overall expression timeline can be synthesized for individual cell types. Thus, lineage tracing is able to define exactly when *Klf4* is driving specific cells to dedifferentiate into high-risk phenotypes in the atherosclerotic disease state. This expression lineage

information provides uniquely specific guidance for the design of gene therapies and individualized medicine. This information is so essential because ubiquitous editing of *Klf4* is not a therapeutic option due to its system-wide expression. *Klf4* is known to have incredibly diverse functions, including cornea homeostasis, maintenance of ectopic skin impermeability, regulation of number and type of white blood cells, and maintenance of germinal epithelium for sperm production in males. Total long-term inhibition of *Klf4* would obviously result in deleterious effects in *Klf4* dependent pathways in the eyes, skin, bone marrow, and testes.

In the future, this project will hopefully be able to identify temporal and cell type gene involvement for *Klf4*, Oct4, Runx2, and *TGF-ß* in atherosclerotic plaques. Successful temporal targeting to inhibit *Klf4* at time points identified and spatial targeting to limit inhibition only to cells involved in atherosclerosis will define ideal gene therapies.

Future Experimental Directions:

The overarching goal of this lineage tracing study is incorporation of lineage tracers *in vivo* in disease model C57/Bl6 mice. Atherosclerosis disease model mice are Apo E -/- (Apo E deficient) . As mentioned in *Background*, Apo E is a lipoprotein on LDL particles, which acts as a ligand for LDLRs and is integral in uptake of LDL into the liver. Apo E -/- organisms possess an extreme risk for development of atherosclerotic plaques and are, for that reason, used as the disease model in this study to ensure expression of genes involved in atherosclerosis.

A major area of investigation has included determining which technique of *in vivo* lineage tracer incorporation is best suited for insertion of our stem cells into mice. Blastocyst

injection, 8-cell injection and i-GONAD are validated embryo editing methods which have been identified as potential approaches.

Blastocyst and 8-cell injections require a cultured mESC line to be injected, for which we would use E14TG2a cells. At this point, our plasmids, which will contain the *Klf4* motifs described in this thesis project as well as, *Oct4, TGFb* and *Runx2* motifs, will have been transfected into E14TG2A using the same technique outlined in *Methods* for *Klf4* transfection. In blastocyst injection or 8-cell injection techniques, stem cells containing our lineage tracing motifs will be inserted via nucleofection by microinjection into cells in the early-embryonic blastocyst phase (blastocyst injection) or in the very early 8-cell stage (8-cell injection). For these methods, combination of the transgene and genome occurs via homology directed repair (HDR) paired with CRISPR/Cas9 editing technology.

Eight-cell injection and blastocyst injection work in similar manners; however, 8-cell injection is done when the embryo is in the eight-cell phase, before it reaches the blastocyst (70-100 cells) phase. Using this method, between 3 and 5 of mESCs from our E14TG2a stem cell line will be injected into an embryo in the 8-cell phase. These extremely early embryos are difficult to harvest, so purchase of frozen, uncompact eight-cell embryos is common. A laser is used to punch a miniscule hole in the embryo and insert the mESCs.[65] For both injection methods, transfected zygotes must be surgically implanted into pseudo-pregnant female recipients. Delivery of pups can be expected after 20 days. 10 days after delivery of pups, a tail biopsy should be performed in order to genotype the mice and identify transgene founders.[65]

While blastocyst injection and eight-cell injection provide straightforward and validated methods for targeting sequence incorporation into the mouse genome, a

CRISPR-Cas9 system called iGONAD may provide an efficient alternative approach that eliminates breeding and in vitro egg injection procedures. iGONAD or "improved genome-editing via oviductal nucleic acids delivery" is a technique in which CRISPR elements Cas9, mRNA, and gRNA, are injected into pregnant mice's oviducts 1.5 days after conception.[66] The CRISPR components and lineage tracing motifs are then electroporated into eggs as they travel through the oviducts. Founder mice can be identified after 6 weeks using this technique.[66]

All three stem cell incorporation approaches explored have many advantages and drawbacks. Blastocyst injection has the advantage of being a well-established and validated technique. Many highly regarded commercial vendors and labs can aid in the process and no proof of concept is necessary.[67] In addition, 8-cell injection is a relatively established method and has many other advantages. Since the embryo is so young at the 8-cell stage, when multiple stem cells are injected, the eight existing cells act as support cells and the embryo does not make its own stem cells, instead using the injected ones as its own.[67] This eliminates variation that exists when performing blastocyst injections caused by presence of both the inserted and endogenous stem cells. While eight-cell injection maintains only one stem cell line, as mentioned, chimeric pups resulting from blastocyst injection will incorporate both the endogenous and injected stem cell lines. Therefore, chimeric mice from blastocyst injection must be bred to isolate a population of mice with our target motifs incorporated on an Apo E -/- background. (Wenhao Xu, PhD; oral communication over Zoom video chat; March 25, 2021)

i-GONAD offers other advantages, including allowing for the earliest incorporation of mESCs, elimination of chimeras and mosaic mice, no requirement for breeding

post-insertion, female mice do not need to be sacrificed, and microinjection is not needed.[68] Elimination of microinjection removes error that can arise from missed microinjection timing windows.[66] Timing of microinjection in blastocyst injection and 8-cell injection is a major factor because, if the optimal DNA integration window is missed, the pronuclei of the injected zygotes may not have sufficient time to prepare to fuse and initiate the primary cell divisions.[67] If injection is attempted after 13 hours, integration may not occur, or the pups may be mosaic mice (only certain cell populations within the mouse contain the transgene). Still, i-GONAD is accompanied by some major limitations, including that it is not a validated and established technique. This means that not only will the system have to be validated but troubleshooting resources for complexities which arise will be incredibly limited.[68,69]

At this point, we plan to first attempt 8-cell injection; however, if this technique is not successful, we have identified i-GONAD as the next best option.

Limitations:

A major limitation for this study was laboratory access because of the COVID-19 pandemic. This caused the projected course and the study to be entirely readjusted, leading to setbacks in research timelines. In addition, new online software and coding techniques had to be learned in order to continue pursuing research results outside of the lab setting.

Scientific limitations included inexact replication of atherosclerosis between different animal models. While mouse presentation of CVD is similar to human presentation and provides a close characterization, there do exist significant differences in physiology and genetic drivers that should be considered when applying results to human disease states.

Future Applications:

Future directions of this study revolve around establishment of lineage tracers in Apo E -/- mice for long term lineage tracing studies and eventual application of lineage tracing findings to therapeutic settings.

As outlined above, within the next 2-3 months, insertion of mESC lines with incorporated lineage tracing elements will be inserted *in vivo* into Apo E-/- mice. This disease model will allow for the first atherosclerotic disease state CyTOF results to be collected and analyzed via comparison to wild type expression patterns of our target genes. As previously mentioned, expression analysis at as many time points as possible is essential to synthesizing comprehensive single-cell level expression lineages and therefore to creating an exhaustive timeline of gene involvement.

Future applications of this lineage information will hopefully involve gene therapies. One promising approach to gene therapy has been the use of Adeno-Associated Viruses (AAVs) as a vector for delivery of target insertions and gene editing tools like CRISPR/Cas9 to somatic cells. Dr. Ali Güler, an expert in *in vivo* gene targeting at UVA, was consulted to discuss important cardiovascular-specific considerations when designing AAV mediated therapies.

In previous approaches using traditional Adenoviruses, dangerous immune responses limited realization of designed gene therapies. AAVs are non-enveloped viruses that can be engineered to reduce or eliminate mounting of immune responses and therefore are a promising alternative for gene therapy delivery.

10 AAV serotypes exist. Serotypes AAV1-AAV10 differ by capsid proteins. Capsid protein expression çan be matched to cell type targets in which it will be most effective. This

increases spatial control because, if serotype is chosen correctly, viral vectors will have an increased tendency to enter target cells. Despite established reductions in immune responses to AAV therapies, immune reactions may still surface upon injections. Serotype choice has been shown to be an effective mechanism for further reduction or elimination in immune response. Disease-specific immune states should be considered when choosing serotypes. For example, as was discussed in *Background*, atherosclerosis involves a significant increase in cytokine and proinflammatory activity. As such, it has been found that CVD monkeys treated with AAV2 therapies, which in many cases lead to lowest immune responses, have mounted immune responses to treatment overtime. Creation of a chimeric AAV2/8 viral vector, however, was able to greatly reduce immune response and allow monkeys in the trial to tolerate treatment well.[43] Therefore, when designing future gene therapies for treatment of atherosclerosis, chimeric AAVs should be considered to increase treatment tolerance and efficacy (Ali Güler, PhD; oral communication over Zoom video chat; April 5, 2021).

Two limitations to AAV therapies include that AAVs have a limited capacity for transportation of genetic material and can only target one site at a time. AAVs cannot hold more than 4.7 kB of DNA. CRISPR is an effective editing strategy to use in conjunction with AAVs because Cas9 and gRNA sequences are very small and therefore leave more space for gene targets and inserts to be transported. In addition, each AAV is only able to target one cell at a time. This means that while multiple gRNAs could be included in an AAV to edit more than one gene within one cell type, multiple genes in different cell types must be targeted individually with separate AAVs. This may become an important consideration when working towards gene therapies for CVD because cell type and gene involvement in this disease is so heterogeneous.

A final AAV consideration posed by Ali Güler in our consultation was that local versus global injections must be considered. Global injections in mouse models are often injected into the base of the tail. Blood flow is used to circulate vectors through tissues. Local injections on the other hand have more efficacy when cell types are localized in a small area. For atherosclerosis, aortic angiography may provide an effective local injection site to allow for AAVs to be majorly concentrated in aortic and arterial vascular tissues.

Ethical and Societal Implications

Bioethical considerations of future applications of research findings center around equitable uses and distribution of gene editing technology. As was discussed in *Background,* African American citizens as well as those with decreasing income and low income experience disproportionately high rates of CVD and atherosclerosis.[33,34,35] Therefore, when eventually designing experimental trials in clinical settings, it is essential to ensure that populations represented within trials are proportional to those affected by the target disease. This consideration is important not only on a genetic and physical level but also as a measure of acceptance and trust of the therapeutic approach.

Another major area of equitability is defined by disproportionate access to new therapies. Gene therapy, as a newly emerging complex form of medicine, can be incredibly expensive. Price points for currently available gene therapies range from $425,000-$2.1 million per treatment.[70] Since atherosclerosis is most common in individuals with low or decreasing income, considerations must be made for how these populations will access such expensive therapies. Still, many argue that despite high per-treatment costs, gene therapies,

as a one- or few-time treatment option, have the ability to improve patient quality of life to such an extent that it outweighs these staggering costs.[70] Multiple studies have demonstrated that the incredibly high price of one-time gene therapies still amounts to higher life-time treatment costs than surgical or pharmaceutical alternatives.[71] However, one 2019 study found the majority of the public is willing to overlook this cost analysis: 78% of study participants (N=1,000) felt that high one-time gene therapy costs were "worth it" because they are seen by many as a "cure" rather than treatment.[73]

Societal considerations also come into view when gene therapy prices are analyzed. Currently, health resources and insurance companies are structured around the American health care norm of long term, repeated costs and treatments. As more high-cost one-time treatments surface, insurance and health resource structures may have to readjust accordingly.[71,73] While this may benefit those pursuing gene therapies, such restructuring will likely be accompanied by major budgetary shifts. Therefore, if such readjustments do occur, equitable redistribution of insurance coverage must be a priority.

Finally, the societal goal of medicine revolves around increasing quality of life by decreasing symptoms with the fewest side effects. If equitable and widespread access to gene therapies can be achieved, this treatment option has the potential to decrease symptoms, side effects and reliance on medication.

ACKNOWLEDGEMENTS:

First, I would like to thank Corey Williams, my primary mentor in this thesis project. He guided me through countless new experimental processes and taught me so much about the scientific process. Without Corey's patience and commitment this project would not have

been possible. I would also like to thank Dr. Eli Zunder, my lab PI, for allowing me to be part of this project and supporting my scientific and research pursuits. Third, I would like to thank Dr. Michael Timko for his leadership in this Human Biology community, his willingness to answer questions at all times, and his support of each of us in our academic endeavors. Fourth, I would like to thank Emily Puleo for introducing me to the Zunder lab - without her recommending me to Corey, this project would not have been possible. Fifth, I would like to thank Ali Güler and Wenhao Xu for providing their time and expert opinions for this thesis.

BIBLIOGRAPHY

1. Rafieian-Kopaei M, Setorki M, Doudi M, Baradaran A, Nasri H. Atherosclerosis: process, indicators, risk factors and new hopes. *Int J Prev Med*. 2014;5(8):927-946.
2. Cahill PA, Redmond EM. Vascular endothelium – gatekeeper of vessel health. *Atherosclerosis*. 2016;248:97-109. doi:10.1016/j.atherosclerosis.2016.03.007
3. von der Thüsen JH, Borensztajn KS, Moimas S, et al. IGF-1 has plaque-stabilizing effects in atherosclerosis by altering vascular smooth muscle cell phenotype. *Am J Pathol*. 2011;178(2):924-934. doi:10.1016/j.ajpath.2010.10.007
4. Juul A., Scheike T., Davidsen M., Gyllenborg J., Jorgensen T. Low serum insulin-like growth factor I is associated with increased risk of ischemic heart disease: a population-based case-control study. *Circulation.* 2002;106:939–944.
5. Ogunrinade, O., Kameya, G.T. & Truskey, G.A. Effect of Fluid Shear Stress on the Permeability of the Arterial Endothelium. *Annals of Biomedical Engineering* 30, 430–446 (2002). https://doi.org/10.1114/1.1467924
6. Cybulsky MI, Marsden PA. Effect of Disturbed Blood Flow on Endothelial Cell Gene Expression. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2014;34(9):1806-1808. doi:10.1161/atvbaha.114.304099
7. Mullick AE, McDonald JM, Melkonian G, Talbot P, Pinkerton KE, Rutledge JC. Reactive carbonyls from tobacco smoke increase arterial endothelial layer injury. *Am J Physiol Heart Circ Physiol*. 2002;283(2):H591-H597. doi:10.1152/ajpheart.01046.2001
8. Libby P. Vascular biology of atherosclerosis: overview and state of the art. *The American Journal of Cardiology*. 2003;91(3):3-6. doi:10.1016/s0002-9149(02)03143-0
9. Cohen DE. Balancing cholesterol synthesis and absorption in the gastrointestinal tract. *J Clin Lipidol*. 2008;2(2):S1-S3. doi:10.1016/j.jacl.2008.01.004
10. Omega-3 fatty acids. University of Maryland Medical Center. http://www.umm.edu/altmed/articles/omega-3-000316.htm. Updated August 5, 2015. Accessed September 28, 2017.
11. Publishing HH. How it's made: Cholesterol production in your body. Harvard Health. https://www.health.harvard.edu/heart-health/how-its-made-cholesterol-production-in-your-body. Accessed April 15, 2021.
12. Feingold KR. Introduction to Lipids and Lipoproteins. [Updated 2021 Jan 19]. In: Feingold KR, Anawalt B, Boyce A, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK305896/
13. Biros E, Karan M, Golledge J. Genetic variation and atherosclerosis. *Curr Genomics*. 2008;9(1):29-42. doi:10.2174/138920208783884856

14. Goldstein JL, Brown MS. *Familial hypercholesterolemia: The Metabolic Basis of Inherited Disease.*6th. New York: McGraw Hill; 1989. [Google Scholar]
15. Libby P. Inflammation in atherosclerosis. *Nature.* 2002;420:868–874. [PubMed] [Google Scholar]
16. Haskins RM, Nguyen AT, Alencar GF, et al. Klf4 has an unexpected protective role in perivascular cells within the microvasculature. *Am J Physiol Heart Circ Physiol*. 2018;315(2):H402-H414. doi:10.1152/ajpheart.00084.2018
17. Shankman, L., Gomez, D., Cherepanova, O. *et al.* KLF4-dependent phenotypic modulation of smooth muscle cells has a key role in atherosclerotic plaque pathogenesis. *Nat Med* 21, 628–637 (2015). https://doi.org/10.1038/nm.3866
18. UCSC Genome Browser Home. https://genome.ucsc.edu/. Accessed April 15, 2021.
19. UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. UniProt Consortium. UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. https://www.uniprot.org/. Accessed April 15, 2021.
20. Msc musculin [Mus musculus (house mouse)] - Gene - NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=17681. Accessed April 15, 2021.
21. Cloud-Based Informatics Platform for Life Sciences R&D. Benchling. https://benchling.com/. Published April 6, 2021. Accessed April 15, 2021.
22. Biolabs NE. Gibson Assembly®. NEB. https://www.neb.com/applications/cloning-and-synthetic-biology/dna-assembly-and-cloning/gibson-assembly. Accessed April 15, 2021.
23. Biolabs NE. Dam-Dcm and CpG Methylation. NEB. https://www.neb.com/tools-and-resources/selection-charts/dam-dcm-and-cpg-methylation. Accessed April 15, 2021.
24. Libby, P. Inflammation in atherosclerosis. *Nature* 420, 868–874 (2002). https://doi.org/10.1038/nature01323
25. Cho A, Haruyama N, Kulkarni AB. Generation of transgenic mice. *Curr Protoc Cell Biol*. 2009;Chapter 19:Unit-19.11. doi:10.1002/0471143030.cb1911s42
26. Biros E, Karan M, Golledge J. Genetic variation and atherosclerosis. *Curr Genomics*. 2008;9(1):29-42. doi:10.2174/138920208783884856
27. Kelli HM, Hammadah M, Ahmed H, et al. Association Between Living in Food Deserts and Cardiovascular Risk. *Circ Cardiovasc Qual Outcomes*. 2017;10(9):e003532. doi:10.1161/CIRCOUTCOMES.116.003532
28. Lattimer JM, Haub MD. Effects of dietary fiber and its components on metabolic health. *Nutrients*. 2010;2(12):1266-1289. doi:10.3390/nu2121266
29. Wang SY, Tan ASL, Claggett B, et al. Longitudinal Associations Between Income Changes and Incident Cardiovascular Disease: The Atherosclerosis Risk in

Communities Study. *JAMA Cardiol*. 2019;4(12):1203-1212. doi:10.1001/jamacardio.2019.3788
30. Rabi DM, Edwards AL, Svenson LW, Graham MM, Knudtson ML, Ghali WA. Association of Median Household Income With Burden of Coronary Artery Disease Among Individuals With Diabetes. *Circulation: Cardiovascular Quality and Outcomes*. 2010;3(1):48-53. doi:10.1161/circoutcomes.108.840611
31. Diez-Roux AV, Nieto FJ, Tyroler HA, Crum LD, Szklo M. Social inequalities and atherosclerosis. The atherosclerosis risk in communities study. *Am J Epidemiol*. 1995;141(10):960-972. doi:10.1093/oxfordjournals.aje.a117363
32. Black PH. The inflammatory response is an integral part of the stress response: Implications for atherosclerosis, insulin resistance, type II diabetes and metabolic syndrome X. *Brain, Behavior, and Immunity*. 2003;17(5):350-364. doi:10.1016/s0889-1591(03)00048-5
33. Skloot R. *The Immortal Life of Henrietta Lacks*. Toronto, Ontario; 2020.
34. McGloughlin S, Padiglione AA. Host defence mechanisms and immunodeficiency disorders. *Oh's Intensive Care Manual*. 2014. doi:10.1016/b978-0-7020-4762-6.00067-9
35. Health, United States Spotlight Racial and Ethnic Disparities in Heart Disease. Center for Disease Control. https://www.cdc.gov/nchs/hus/spotlight/HeartDiseaseSpotlight_2019_0404.pdf. Published April 2019.
36. Turkson‐Ocran RAN, Nmezi NA, Botchway MO, et al. Comparison of Cardiovascular Disease Risk Factors Among African Immigrants and African Americans: An Analysis of the 2010 to 2016 National Health Interview Surveys. *Journal of the American Heart Association*. 2020;9(5). doi:10.1161/jaha.119.013220
37. Bergheanu SC, Bodde MC, Jukema JW. Pathophysiology and treatment of atherosclerosis : Current view and future perspective on lipoprotein modification treatment. *Neth Heart J*. 2017;25(4):231-242. doi:10.1007/s12471-017-0959-2
38. Bergheanu, S.C., Bodde, M.C. & Jukema, J.W. Pathophysiology and treatment of atherosclerosis. *Neth Heart J* 25, 231–242 (2017). https://doi.org/10.1007/s12471-017-0959-2
39. Reena Pande MD. Cholesterol and statins: it's no longer just about the numbers. Harvard Health Blog. https://www.health.harvard.edu/blog/cholesterol-and-statins-its-no-longer-just-about-the-numbers-201311136868 . Published June 9, 2020. Accessed April 15, 2021.
40. Statins: Are these cholesterol-lowering drugs right for you? Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/statins/art-20045772. Published March 14, 2020. Accessed April 15, 2021.
41. Gürgöze MT, Muller-Hansma AHG, Schreuder MM, Galema-Boers AMH, Boersma E, Roeters van Lennep JE. Adverse Events Associated With PCSK9 Inhibitors: A

Real-World Experience. *Clin Pharmacol Ther*. 2019;105(2):496-504. doi:10.1002/cpt.1193
42. Fitchett DH, Hegele RA, Verma S. Statin Intolerance. *Circulation*. 2015;131(13). doi:10.1161/circulationaha.114.013189
43. Kolata G. A 'Cure for Heart Disease'? A Single Shot Succeeds in Monkeys. The New York Times. https://www.nytimes.com/2020/06/27/health/heart-disease-gene-editing.html. Published June 27, 2020. Accessed April 15, 2021.
44. Haskins RM, Nguyen AT, Alencar GF, et al. Klf4 has an unexpected protective role in perivascular cells within the microvasculature. *Am J Physiol Heart Circ Physiol*. 2018;315(2):H402-H414. doi:10.1152/ajpheart.00084.2018
45. Shankman, L., Gomez, D., Cherepanova, O. *et al.* KLF4-dependent phenotypic modulation of smooth muscle cells has a key role in atherosclerotic plaque pathogenesis. *Nat Med* 21, 628–637 (2015). https://doi.org/10.1038/nm.3866
46. Alencar GF, Owsiany KM, Karnewar S, et al. Stem Cell Pluripotency Genes Klf4 and Oct4 Regulate Complex SMC Phenotypic Changes Critical in Late-Stage Atherosclerotic Lesion Pathogenesis. *Circulation*. 2020;142(21):2045-2059. doi:10.1161/circulationaha.120.046672
47. Osteogenesis Markers. www.rndsystems.com. https://www.rndsystems.com/research-area/osteogenesis-markers. Accessed April 15, 2021.
48. *Toma I, McCaffrey TA. Transforming growth factor-β and atherosclerosis: interwoven atherogenic and atheroprotective aspects. Cell Tissue Res. 2012;347(1):155-175. doi:10.1007/s00441-011-1189-3*
49. Shi X, Gao J, Lv Q, et al. Calcification in Atherosclerotic Plaque Vulnerability: Friend or Foe?. *Front Physiol*. 2020;11:56. Published 2020 Feb 5. doi:10.3389/fphys.2020.00056
50. Yahagi K., Davis H.R., Joner M., Virmani R. (2015) Atherosclerosis, Introduction and Pathophysiology. In: Jagadeesh G., Balakumar P., Maung-U K. (eds) Pathophysiology and Pharmacotherapy of Cardiovascular Disease. Adis, Cham. https://doi.org/10.1007/978-3-319-15961-4_25
51. Biolabs NE. AscI. New England Biolabs: Reagents for the Life Sciences Industry. https://international.neb.com/products/r0558-asci#Product%20Information. Accessed April 15, 2021.
52. Biolabs NE. MluI. New England Biolabs: Reagents for the Life Sciences Industry. https://international.neb.com/products/r0198-mlui#Product%20Information. Accessed April 15, 2021.
53. ZR Plasmid Miniprep - Classic. ZYMO RESEARCH. https://www.zymoresearch.com/products/zr-plasmid-miniprep-classic. Accessed April 15, 2021.

54. Zymoclean Gel DNA Recovery Kit. ZYMO RESEARCH. https://www.zymoresearch.com/collections/zymoclean-gel-dna-recovery-kits/products/zymoclean-gel-dna-recovery-kit. Accessed April 15, 2021.
55. Biolabs NE. Restriction Digest Protocol. NEB. https://www.neb.com/protocols/2018/07/30/restriction-digest-protocol. Accessed April 15, 2021.
56. Flow Cytometry Technology Notes. UWCCC Research. https://cancer.wisc.edu/research. Accessed April 15, 2021.
57. Mouse Gene Sox2. Mouse Gene Sox2 (ENSMUST00000099151.5) Description and Page Index. https://genome.ucsc.edu/cgi-bin/hgGene?org=Mouse&hgg_gene=ENSMUST00000099151.5&hgg_chrom=none&db=mm10. Accessed April 15, 2021.
58. Mouse Gene Nkx2-5. USCS Genome Browser (ENSMUST00000015723.4) Description and Page Index. http://genome.cse.ucsc.edu/cgi-bin/hgGene?org=Mouse&hgg_gene=ENSMUST00000015723.4&hgg_chrom=none&db=mm10. Accessed April 15, 2021.
59. GRCm38 - mm10 - Genome - Assembly - NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/. Accessed April 15, 2021.
60. CHOPCHOP. https://chopchop.cbu.uib.no/. Accessed April 15, 2021.
61. Editas Medicine. https://www.editasmedicine.com/. Published September 17, 2020. Accessed April 15, 2021.
62. Alt-R CRISPR Custom Guide RNAs: IDT. Integrated DNA Technologies. https://www.idtdna.com/pages/products/crispr-genome-editing/custom-guide-rna?utm_source=google&utm_medium=cpc&utm_campaign=ga_crispr_cas&utm_content=ad_group_grna&gclid=Cj0KCQjwyN-DBhCDARIsAFOELTlKBJReEVQAIbf4MVBrzHdBTGVbBvDIn7LYMqmY97MSHbU4KU1oR9gaAmyGEALw_wcB. Accessed April 15, 2021.
63. T7 Endonuclease I (T7E1) Assay Protocol for CRISPR/Cas9 gRNA validation. GeneMedi. https://www.genemedi.net/i/t7-endonuclease-assay. Accessed April 15, 2021.
64. Cytobank. Welcome to Cytobank. https://community.cytobank.org/cytobank. Accessed April 15, 2021.
65. DeChiara TM, Poueymirou WT, Auerbach W, Frendewey D, Yancopoulos GD, Valenzuela DM. Producing Fully ES Cell-Derived Mice from Eight-Cell Stage Embryo Injections. *Methods in Enzymology*. 2010:285-294. doi:10.1016/s0076-6879(10)76016-x
66. (2005) Pseudopregnant Mouse. In: Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer, Berlin, Heidelberg . https://doi.org/10.1007/3-540-29623-9_8423

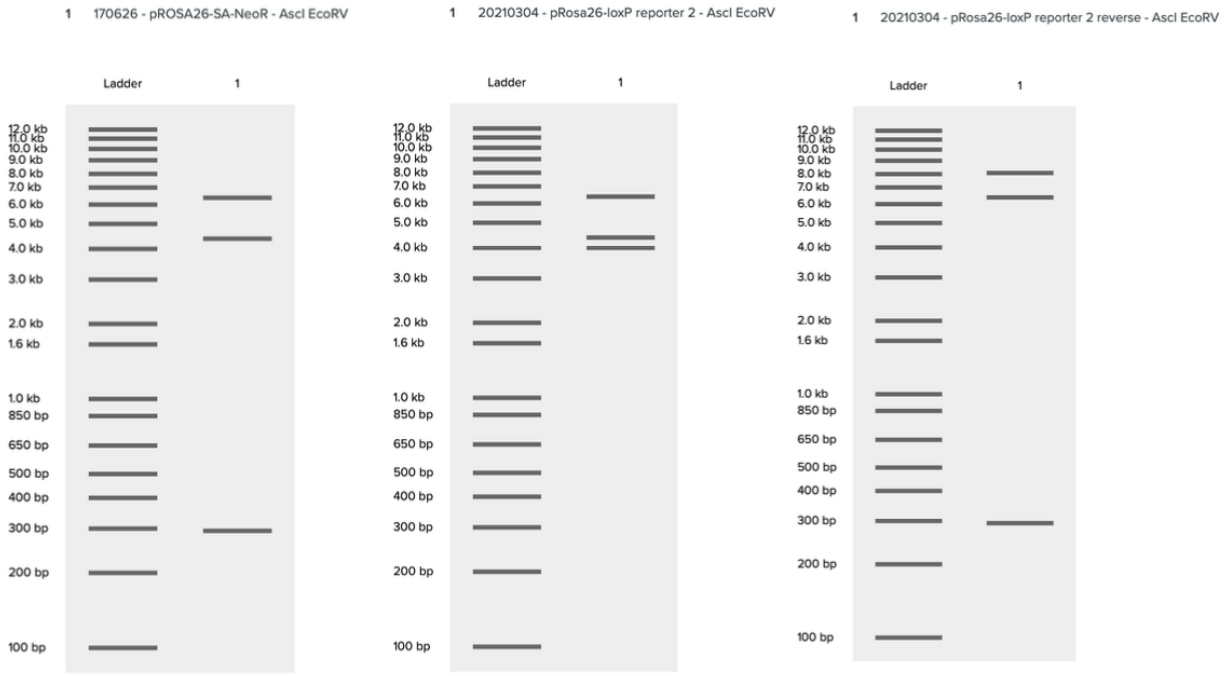
67. Gurumurthy CB, Sato M, Nakamura A, et al. Creation of CRISPR-based germline-genome-engineered mice without ex vivo handling of zygotes by i-GONAD. *Nat Protoc*. 2019;14(8):2452-2482. doi:10.1038/s41596-019-0187-x
68. Ohtsuka, M., Sato, M., Miura, H. *et al. i*-GONAD: a robust method for *in situ* germline genome engineering using CRISPR nucleases. *Genome Biol* 19, 25 (2018). https://doi.org/10.1186/s13059-018-1400-x
69. Gurumurthy, C.B., Sato, M., Nakamura, A. *et al.* Creation of CRISPR-based germline-genome-engineered mice without ex vivo handling of zygotes by *i*-GONAD. *Nat Protoc* 14, 2452–2482 (2019). https://doi.org/10.1038/s41596-019-0187-x
70. Wong CH, Li D, Wang N, Gruber J, Conti R, Lo AW. Estimating the Financial Impact of Gene Therapy*. medRxiv. https://www.medrxiv.org/content/10.1101/2020.10.27.20220871v1.full. Published January 1, 2020. Accessed April 15, 2021.
71. Joshua T. Cohen James D. Chambers Madison C. Silver Pei-Jung Lin Peter J. Neumann. Putting The Costs And Benefits Of New Gene Therapies Into Perspective: Health Affairs Blog. Health Affairs. https://www.healthaffairs.org/do/10.1377/hblog20190827.553404/full/. Published September 4, 2019. Accessed April 15, 2021.
72. Gene therapies offer breakthrough results but extraordinary costs. Massachusetts Municipal Association (MMA). https://www.mma.org/gene-therapies-offer-breakthrough-results-but-extraordinary-costs/. Published March 18, 2020. Accessed April 15, 2021.
73. National Voter Poll: Pacific Research Institute . Center for Medical Economics and Innovation. https://medecon.org/wp-content/uploads/2019/06/LPS-PRI-Omnibus-Poll-May-2019-MQ-1-1.pdf. Published June 7, 2019. Accessed April 15, 2021.
74. CRISPR Guide. Addgene. http://www.addgene.org/guides/crispr/. Accessed May 11, 2021.
75. Gearing M. Evolution of Brainbow: Using Cre-lox for Multicolor Labeling of Neurons. Addgene blog. https://blog.addgene.org/evolution-of-brainbow-using-cre-lox-for-multicolor-labeling-of-neurons. Accessed May 11, 2021.
76. Friedrich G, Soriano P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes & Development*. 1991;5(9):1513-1523. doi:10.1101/gad.5.9.1513

APPENDICES:

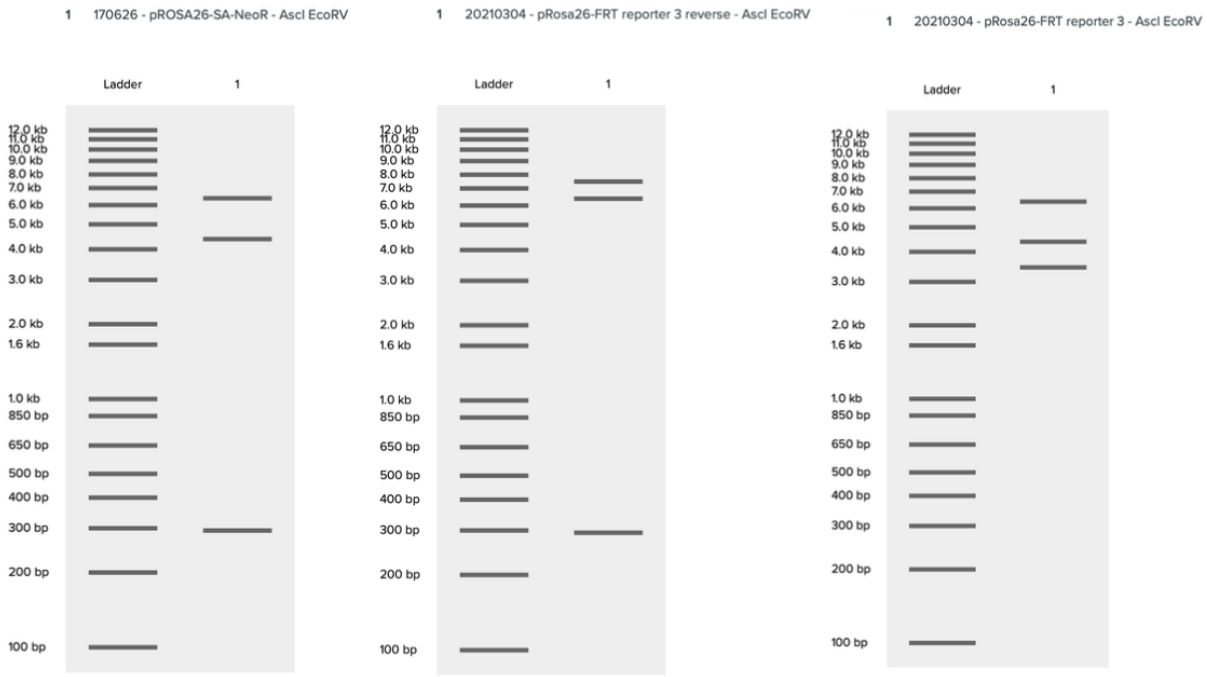
Appendix A: Analytical digest expected bands for reporter motifs:

Backbone Correct Insertion Reverse Insertion

LoxP: AscI/EcoRV



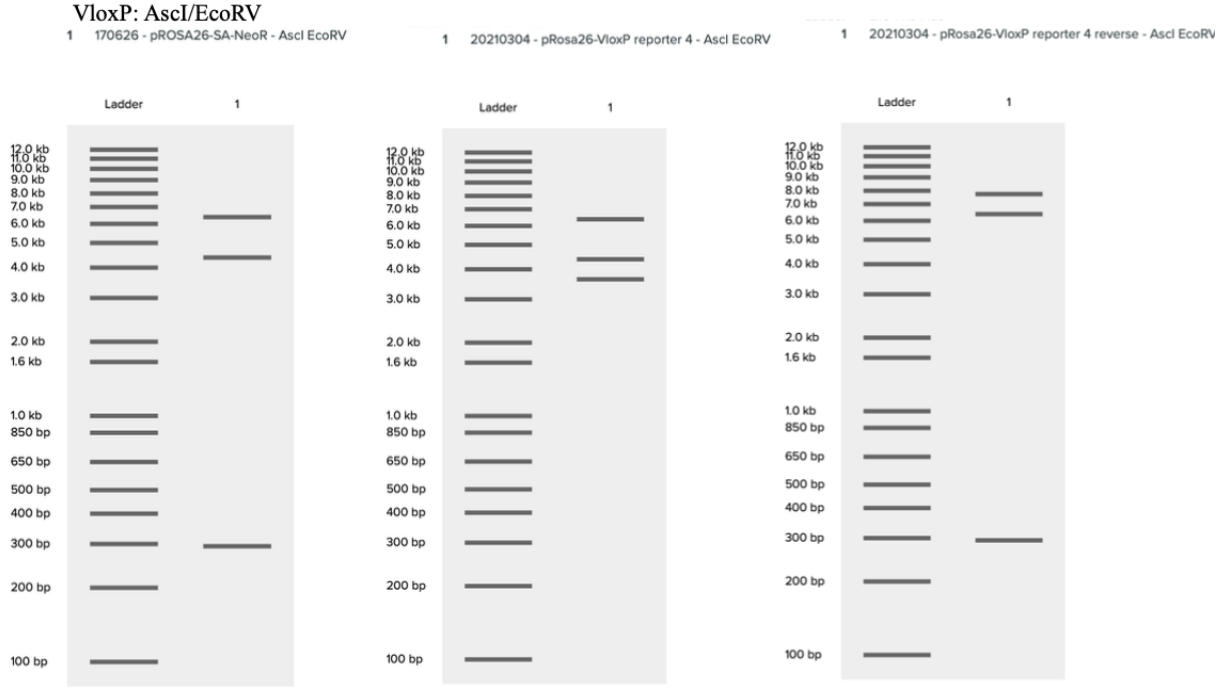
FRT: AscI/EcoRV



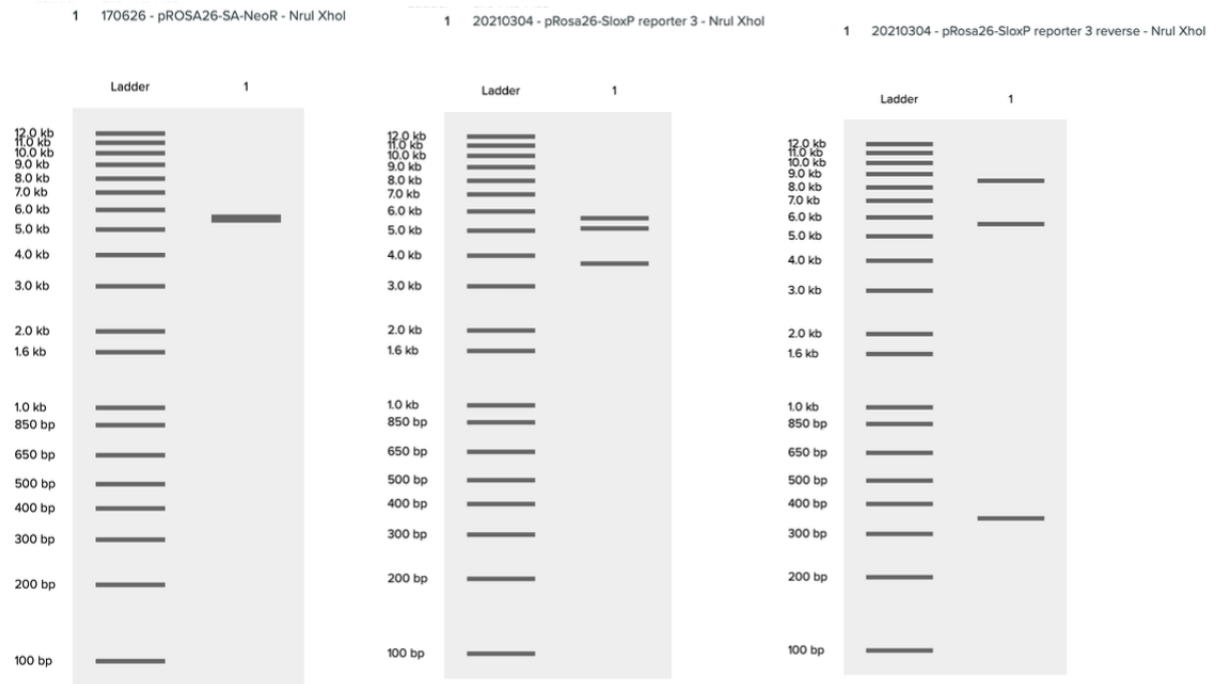
Appendix A, ctd.

Backbone | Correct Insertion | Reverse Insertion

VloxP: AscI/EcoRV



SloxP: XhoI/NruI



Appendix B: *Calculated Separation Indices for antibody titration analysis in functional reporter motif analysis.*

	Concentration	Median Pos.	Median Neg.	84th percentile Neg.	Separation Index
t7tag lo	0.01	1501	441	636	5.408717949
for Sloxp	0.03	1548	473	692	4.88413242
	0.1	2557	664	943.4	6.741356478
	0.3	3506	927	1321	6.512956853
	1	3920.5	1378	1950.72	4.417145376
	3	6807	3112	4089.64	3.760612291
	10	14429.5	5307	7361.12	4.418869151
t7tag mid	0.01	1400.5	409	593	5.361644022
for Sloxp	0.03	1520	453	649	5.416658163
	0.1	2663	607	887.6	7.290520314
	0.3	3475.5	907	1303	6.453680556
	1	3830.5	1304	1844.48	4.651175807
	3	7184	1964	2786.96	6.311242345
	10	11248.5	4015	5543.4	4.7090634
t7tag hi	0.01	1487	455	655	5.1342
for Sloxp	0.03	1470	550	749.8	4.581581582
	0.1	2111	811	1155.76	3.75188537
	0.3	5458	1477.5	2088.08	6.486615186
	1	7574	2365	3459.72	4.734502887
	3	16318	2833	4207.88	9.759088066
	10	36669	6631	10182.8	8.414834732
OllasTag	0.01	4654	227	490.24	16.73326622
for VLoxP	0.03	5433	338	656	15.94190252
	0.1	9728	1010	1690.52	12.74673779
	0.3	--	--	--	N/A
	1	40266	5081	7775	12.9952023
	3	45376	9698	14497.52	7.396491733
	10	68397	17178	24824.36	6.664988962

$$\frac{(\boldsymbol{Median\ Pos}) - (\boldsymbol{Median\ Neg})}{(\boldsymbol{84\%\ Neg}) - (\boldsymbol{Median\ Neg})/_{\boldsymbol{0.995}}}$$

Appendix C: *R Studio codes written in the Zunder Lab by Corey Williams. Used in R to create heatmaps and UMAP clusters.*

Metadata code:

```
#Corey Williams, University of Virginia
#16 Feb, 2020
#Make metadata csv for URD
#Sets up metadata to have columns for file, genotype, and stage of each cell
#
```

Since this code had not yet been published, it was redacted for the print version of this thesis

Umap.R code:

```
#Corey Williams, University of Virginia
#16 Feb, 2020
#Make metadata csv for URD
#Sets up metadata to have columns for file, genotype, and stage of each cell
#
```

Since this code had not yet been published, it was redacted for the print version of this thesis

Heatmaps code:

```
#Corey Williams, University of Virginia
#21 Jul, 2019
#Adapting code from Nowicka et al., F1000 Res., 2016 to make heatmaps

rm(list = ls())
.libPaths( c( .libPaths(), "~/local/R_libs/") )

library(ZunderPipelineFunctions)
library(matrixStats)
library(dplyr)
library(RColorBrewer)
library(pheatmap)

INPUT.FOLDER <- getwd()
CONCATTRANSFORMED.FILENAME <- "/Concat_Transformed.csv"
CLUSTERS.FILENAME <- "/clusters.csv"
PANEL.FILENAME <- "/panel.csv"
FONTSIZE <- 16 #base fontsize for plot
FONTSIZE_NUMBER <- 10 #fontsize for numbers displayed in cells
FONTSIZE_CLUSTERING <- 24 #base fontsize for plot showing clustering vars
FONTSIZE_NUMBER_CLUSTERING <- 15 #fontsize for numbers displayed in cells in clustering var plot
```

```
#load exprs matrix
expr <- as.matrix(read.concat.transformed(INPUT.FOLDER,CONCATTRANSFORMED.FILENAME))

#load cluster vector
cell_clustering1 <- read.clusters(INPUT.FOLDER,CLUSTERS.FILENAME)

#load panel
panel <- read.panel(INPUT.FOLDER,PANEL.FILENAME)

#get clustering vars
clustering.vars <- get.clustering.annotate(panel)

#make color palette
color_clusters <- palette(rainbow(max(cell_clustering1)))
color_clusters <- palette(rainbow(max(cell_clustering1))) #unclear why, but need this line twice

## data transformation
rng <- colQuantiles(expr, probs = c(0.01, 0.99))
expr01 <- t((t(expr) - rng[, 1]) / (rng[, 2] - rng[, 1]))
expr01[expr01 < 0] <- 0
expr01[expr01 > 1] <- 1

plot_clustering_heatmap_wrapper <- function(expr, expr01,cell_clustering, color_clusters, cluster_merging =
NULL,fontsize_heatmap,fontsize_number_heatmap){

  # Calculate the median expression
  expr_median <- data.frame(expr, cell_clustering = cell_clustering) %>%
    group_by(cell_clustering) %>% summarize_all(funs(median))
  expr01_median <- data.frame(expr01, cell_clustering = cell_clustering) %>%
    group_by(cell_clustering) %>% summarize_all(funs(median))

  # Calculate cluster frequencies
  clustering_table <- as.numeric(table(cell_clustering))
  clustering_prop <- round(clustering_table / sum(clustering_table) * 100, 2)

  # Sort the cell clusters with hierarchical clustering
  d <- dist(expr_median[, colnames(expr)], method = "euclidean")
  cluster_rows <- hclust(d, method = "average")

  expr_heat <- as.matrix(expr01_median[, colnames(expr01)])
  rownames(expr_heat) <- expr01_median$cell_clustering

  # Colors for the heatmap
  color_heat <- colorRampPalette(rev(brewer.pal(n = 9, name = "RdYlBu")))(100)
  legend_breaks = seq(from = 0, to = 1, by = 0.2)
  labels_row <- paste0(expr01_median$cell_clustering, " (", clustering_prop ,
                       "%)")

  # Annotation for the original clusters
  annotation_row <- data.frame(Cluster = factor(expr01_median$cell_clustering))
  rownames(annotation_row) <- rownames(expr_heat)
  color_clusters1 <- color_clusters[1:nlevels(annotation_row$Cluster)]
  names(color_clusters1) <- levels(annotation_row$Cluster)
  annotation_colors <- list(Cluster = color_clusters1)

  # Annotation for the merged clusters
  if(!is.null(cluster_merging)){
    cluster_merging$new_cluster <- factor(cluster_merging$new_cluster)
    annotation_row$Cluster_merging <- cluster_merging$new_cluster
    color_clusters2 <- color_clusters[1:nlevels(cluster_merging$new_cluster)]
    names(color_clusters2) <- levels(cluster_merging$new_cluster)
```

```
  annotation_colors$Cluster_merging <- color_clusters2
 }

 pheatmap(expr_heat, color = color_heat, cluster_cols = FALSE,
         cluster_rows = cluster_rows, labels_row = labels_row,
         display_numbers = TRUE, number_color = "black",
         fontsize = fontsize_heatmap, fontsize_number = fontsize_number_heatmap,  legend_breaks = legend_breaks,
         annotation_row = annotation_row, annotation_colors = annotation_colors)

}

dev.off() #not sure why this is needed, but was having trouble making two plots without
png("Heatmap.png",width = 1162,height = 664)
plot_clustering_heatmap_wrapper(expr = expr,
                          expr01 = expr01,
                          cell_clustering = cell_clustering1,color_clusters = color_clusters,fontsize_heatmap =
FONTSIZE,fontsize_number_heatmap = FONTSIZE_NUMBER)
dev.off()

png("Heatmap_clustering.png",width = 1162,height = 664)
plot_clustering_heatmap_wrapper(expr = expr[,clustering.vars],
                          expr01 = expr01[,clustering.vars],
                          cell_clustering = cell_clustering1,color_clusters = color_clusters,fontsize_heatmap =
FONTSIZE_CLUSTERING,fontsize_number_heatmap = FONTSIZE_NUMBER_CLUSTERING)
dev.off()
```

Clustering code:

```
#Corey Williams, University of Virginia
#15 Jul, 2019
#Plot colored by expression of markers

print("Start PlotByCluster.R")

rm(list = ls())
.libPaths( c( .libPaths(), "~/local/R_libs/") )

library(ZunderPipelineFunctions)
library(ggfortify)

print("libraries loaded")

## Input parameters
================================================================================
INPUT.FOLDER <- getwd()
OUTPUT.FOLDER <- INPUT.FOLDER
LAYOUT.FILENAME <- "/UMAP_layout.csv"
CLUSTERS.FILENAME <- "/clusters.csv"
CLUSTER.PLOT.FILENAME <- "Clusters.png"
POINT.SIZE <- 1
ANNOTATE.CLUSTERS <- FALSE #(TRUE if you want to print cluster numbers on output plot)

print("input parameters loaded, reading needed files")

## Read needed files
================================================================================
layout.in <- read.layout(INPUT.FOLDER,LAYOUT.FILENAME)
clusters.in <- read.clusters(INPUT.FOLDER,CLUSTERS.FILENAME)

print("needed files read, prepping data to plot")
```

```
## Prep dataframe for plotting
=========================================================================
plotting.df <- as.data.frame(cbind(layout.in,clusters.in))
colnames(plotting.df) <- c("umap_x","umap_y","cluster")

print("data ready to plot, plotting")

## Save plots colored by each marker
=================================================================
#set output folder
setwd(OUTPUT.FOLDER)
#loop through variables to plot
if (ANNOTATE.CLUSTERS) {
  ggsave(CLUSTER.PLOT.FILENAME,plot = ggplot(plotting.df,aes(x=umap_x,y=umap_y,
                                          color=factor(cluster))) +
          geom_point(size = POINT.SIZE) + theme(panel.grid.major = element_blank(),
                                        panel.grid.minor = element_blank(),
                                        panel.background = element_blank(),
                                        axis.line = element_line(colour = "black")) +
          annotate("text",
            x = plotting.df[match(unique(plotting.df[,3]),plotting.df[,3]),1],
            y = plotting.df[match(unique(plotting.df[,3]),plotting.df[,3]),2],
            size = 5,
            label = as.character(unique(plotting.df[,3]))) +
        guides(colour = guide_legend(override.aes = list(shape=15, size=8))),
        #^^should work for changing size/shape of legend elements... might have to tweak size per preference
        height = 7,width = 7)
} else {
  ggsave(CLUSTER.PLOT.FILENAME,plot = ggplot(plotting.df,aes(x=umap_x,y=umap_y,
                                          color=factor(cluster))) +
          geom_point(size = POINT.SIZE) + theme(panel.grid.major = element_blank(),
                                        panel.grid.minor = element_blank(),
                                        panel.background = element_blank(),
                                        axis.line = element_line(colour = "black")) +
        guides(colour = guide_legend(override.aes = list(shape=15, size=8))),
        #^^should work for changing size/shape of legend elements... might have to tweak size per preference
        height = 7,width = 7)
}

print("data plotted and file outputted")
print("End PlotByCluster.R")
```