Expanding the Membership of a Stem Cell-Related Blood Transcriptional Module Containing Biomarkers of Cardiovascular Function by a Novel "Bottom-up" Bioinformatics Approach

Honors Research Thesis

Presented in partial fulfillment of the requirements for graduation with honors research distinction in Biomedical Studies in the undergraduate colleges of The Ohio State University

By

Taylor Kantor

The Ohio State University

April 2014

Project Advisor:
Professor Nicanor I. Moldovan, PhD
Department of Internal Medicine, College of Medicine,
Davis Heart and Lung Research Institute

Table of Contents

Abstract

Background. The circulating stem/progenitor cells (CSPCs) system comprises a variety of rare cells with mixed primitive and differentiation characteristics. So far, this system has eluded traditional methods of characterization. We recently showed that quantitative real-time PCR, the most sensitive and reliable gene expression analysis tool available, consistently detects a comprehensive battery of gene markers in all tested peripheral blood mononuclear cell samples from adult human subjects. Moreover, network representation of these genes' co-variation across samples organizes them into two modules of the hierarchical transcriptional network. One module, termed the 'cardiovascular module' (CVM), was inversely related with the age, vascular stiffness and blood pressure of healthy blood donors, and highly depressed (but paradoxically more compact) in hypertensive patients. However, the targeted definition of CVM composition required an expansion to include more potential members. These could be extracted from high-throughput, unbiased methods such as gene microarrays.

Approach and Results. We verified the ability of Affymetrix gene chips, available on public databases, to sense the CVM-associated rare transcripts, and found that only two (out of 15) were detectable. This oriented us towards gene chips hybridized with samples from children, where the frequency of CSPCs is known to be higher, and found that 5 CVM members were detectable. Implementing the 'guilt-by-association' (GBA) principle, we used these markers as 'seed genes' to generate co-variation lists. These were organized as network 'neighborhoods' which were then fused in increasingly higher gene communities that maintained their modular organization around their hub (i.e., 'seed') gene. This 'bottom-up' approach to reconstitute a network is both original and promising, as it reveals a large crop of candidate genes with meaningful known roles in the cardiovascular field (currently being validated by qRT-PCR). We also compared our method with the traditional "top-down" Weighted Gene Correlation Network Analysis (WGCNA) method, as well as a direct 'Clique Mining' protocol applied to an expanded list of stemness genes. Neither method produced modules enriched in CVM genes, even when the source data were blood samples from patients recovering from burn injury, a condition expected to stimulate CSPC release from bone marrow. However, the latter method did suggest the existence of a module containing the CVM gene Notch4 as being mobilized in burn patients, and one organized around OLR1/oxidized LDL receptor with a more widespread occurrence.

Conclusions. Here, we demonstrate a new method to expand a transcriptional network by detection of candidate members with particular relevance for the analysis of rare transcripts, such as CSPC markers. This method has the potential to be applied for cell isolation, diagnostic, prognostic, and treatment purposes in a variety of CSPC and other cell-dependent medical conditions.

Introduction

Circulating stem/progenitor cells (CSPCs) are promising biomarkers of an organism's resilience and of its ability to repair and maintain the cardiovascular system [1]. In spite of the progress in the basic biology of these cells, the translational applications for diagnostics, prognostics, and therapy are still in development. This system's complexity resides in the rarity of CSPCs, their heterogeneity, and the complicated relationships among themselves and their target tissues. In response to this limitation, we were recently able to show the CSPC system can be directly assessed by gene expression analysis, without the need of cell isolation (i.e., while present within the population of peripheral blood mononuclear cells, PBMCs) (*PlosONE*, in press). We proposed this method as transcriptome organization in modules reflecting common structural and functional patterns has become increasingly useful for cell characterization [2]. Because more comprehensive, high throughput methods, such as gene arrays, have insufficient sensitivity, a panel comprising 45 genes containing the most used markers of primitivity and differentiation were directly analyzed via quantitative real-time PCR (qRT-PCR) in peripheral blood mononuclear cell (PBMC) samples from normal and hypertensive individuals. Among these, 15 genes were found to be organized as a blood transcriptional network module, whose members were inversely correlated with age, blood pressure, and vascular stiffness of donors (Fig. 1), as would be expected from CSPCs themselves. In addition, the members of this module, termed the 'cardiovascular module' (CVM), co-varied in their expression patterns, leading to a high degree of gene network interconnectivity. Moreover, the expression of members of the CVM were dramatically decreased in hypertensive patients, while unexpectedly, the connectivity among them was stronger, indicative of a more primitive character of the CSPC population in circulation.

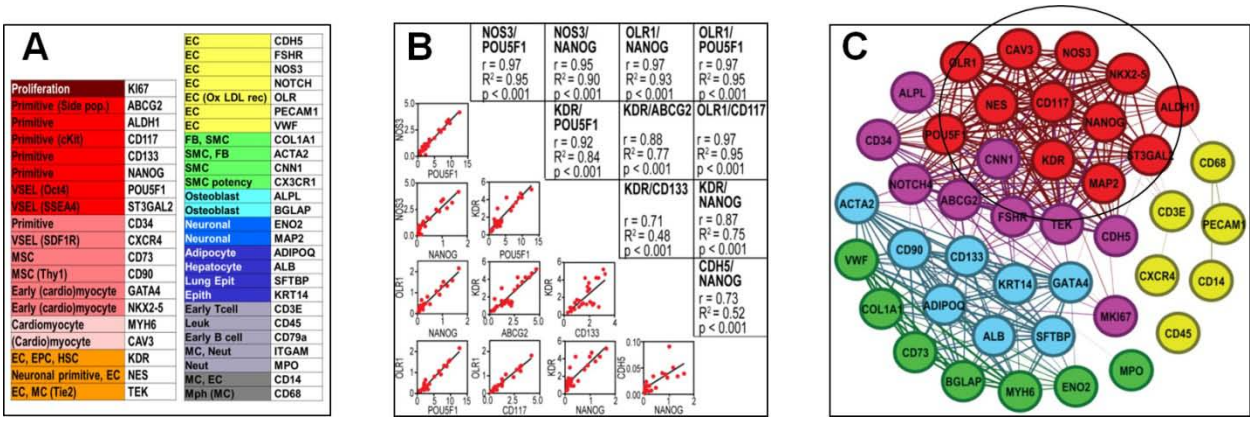

Figure 1. Defining the core network module with cardiovascular function ('cardiovascular module', CVM) by PCR analysis A. Panel of primitive (left column) and differentiation (right column) tested genes. **B.** Example of co-variation of primitive and cardiovascular genes in PBMCs from 27 healthy human subjects. **C.** Co-variation expression matrix of genes in PBMCs from normal subjects. A module composed of 15 genes called CVM (encircled) were found to be inversely related to blood donors' age, vascular stiffness and central blood pressure parameters, suggestive of a vascular protective role, as proposed of CSPCs. The goal of the current project is to expand this network using the 'guilt-by-association' (GBA) principle (from Moldovan et al., *PlosONE*, in press).

Given the focused selection of the initial CVM membership (limited to known CSPC markers), determination of more candidate members of the actual CVM became a major focus. Thus, **the goal of this project was to identify additional members of the CVM, possibly by using high-throughput, unbiased methods, such as gene microarrays.** To this end, we

proposed to take advantage of the same covariation-based gene expression analysis method, also known in bioinformatics as the 'guilt-by-association' (GBA) principle [3]. This indicates that a gene's expression occurs in tandem, coordinated with others which serve associated functions. Thus, any current member of the CVM module could be used as a 'seed' gene to identify potential members, which could be then verified by more stringent criteria.

To this end, we analyzed 503 Affymetrix microarrays from the Gene Expression Omnibus (GEO) public database (http://www.ncbi.nlm.nih.gov/geo/) hybridized with RNA of normal human PBMCs. Studies included both adults (where detection of primitive genes was low), and children (where primitive genes were better expressed). Normalized data were first analyzed using a "top-down" bioinformatics approach through the standard Weighted Gene Correlation Network Analysis [4] (WGCNA, Diagram 1). However, due to limitations of this method, we constructed a "bottom-up" approach using the co-variation method (i.e. GBA). This approach identified a large number of potential candidates, many having known roles associated with stemness, differentiation, angiogenesis, neovascularization, and/or cardiovascular diseases or repair. Networks were constructed with these potential members to identify distinct modules and the connectivity within them. In addition, studies involving burn victims (likely to induce massive mobilization of CSPCs) and pregnant women suffering from preeclampsia (a condition associated with deficiencies in CSPCs) were used for functional validation of the network constructs and of module members.

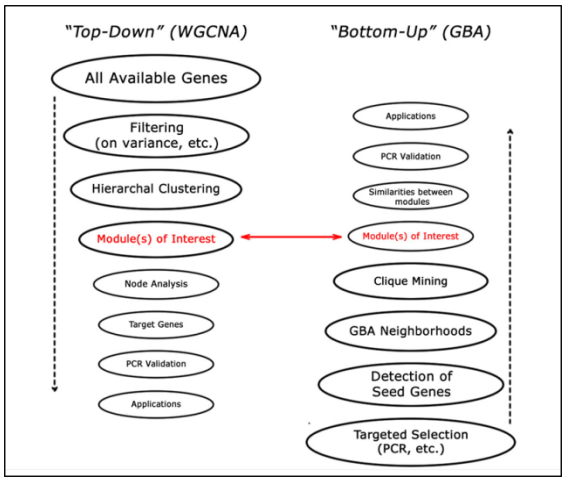

Diagram 1. Comparison of two methods to determine functionally relevant transcriptional network modules. Left, standard Weighted Gene Covariation Network Analysis (WGCNA); right, our original 'guilt-by-association' (GBA)-based approach.

Less than half of the tested primitive genes reliably measured by qRT-PCR were detectable on gene chips, their frequency being inversely related to age of the blood donor. However, using the "bottom-up" approach on chips from normal children, we found that endothelial nitric oxide synthase (NOS3) was strongly associated with low density lipoprotein receptor adaptor protein 1 (LDLRAP1), calcium/calmodulin kinase 4 (CAMK4), and epoxy hydrolase 2 (EPHX2), while megakaryoblastic leukemia (translocation) 1 (MKL1) highly co-varied with NANOG. NOTCH4 was found to be highly correlated with tetratricopeptide repeat domain 7A (TTC7A) and ZFAT zinc finger 1 (ZFAT1) in both children and adult studies. In addition, OLR1 - an essential gene for cardiovascular function - had remarkably consistent correlations between all four analyzed conditions (described later in detail). From this module, we selected for validation matrix metallopeptidase 8 (MMP8), lactotransferrin (LTF), lipocalin 2 (LCN2), and carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6).

In conclusion, we developed a method of detecting a new, collective, systemic biomarker for the elusive CSPCs and other currently unidentified systemic modules. Our expanded CVM could be used to track responses to injury, identify patients at risk for developing, or already experiencing, cardiovascular and other diseases, and give physicians new tools for performing diagnoses and assess treatment outcomes.

Assumptions of the study:

This study is based on the following assumptions:

i) *The CSPC system* represents the same cells, whether they are involved in cardiovascular maintenance or repair (including the blood vessels themselves) or of other organs; this principle, supported by a large body of experimental evidence [5], provides the flexibility to study it longitudinally (from fetuses to the elderly) and to consider as perturbations a variety of pathologies (hypertension, myocardial infarction, stroke, preeclampsia) or repairing-related conditions, such as post-injury response (acute blunt trauma, burning, etc); proportions may vary, the degree of differentiation may change, but overall, the system remains the same.

ii) *The origin of gene expression covariation* (which should be distinctly analyzed if positive or negative) could be explained either *functionally* (at single-cell level: co-transcription, co-regulation by common transcription or epigenetic factors, co-localization on chromatin, etc;) or *quantitatively* (at cell population level: if two proportionally-expressed genes positively co-vary in individuals of comparable chronological age, this could be due to variations in the frequency of carrier cells (this assumption is similar to the one used to detect circulating tumor cells by PCR).

Methods

Studies Used for Analysis. From the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) public database, we assembled a collection of Affymetrix GeneChip® Human Genome U133 Plus 2.0 arrays (platform GPL570) that analyze PBMCs from healthy control subjects of various ages (GSE21942 (15 samples)[6], GSE27034 (18)[7], GSE14642 (20)[8], GSE11761 (20)[8], GSE46480 (98)[9], GSE8507 (17)[10], GSE10041 (23)[11]), as well as from burn victims (GSE19743 (63-controls, 57 early/late burn victims)[12]). In addition, we used one study (GSE23025 (56)[13]) that analyzed $CD34^+$ cells on the same platform in order to compare detection levels of the original gene panel of the 45 most used primitive and differentiation genes. One study, solely based on children (GSE13501 (59)[14]), was analyzed separately, as children have been shown to contain higher levels of CSPCs in peripheral blood.

Presence Score Detection. Presence scores (detection levels) were determined using Affymetrix Expression Console and performing the MAS5 normalization. For genes represented by several probe sets, the probe set with the largest presence score and signal (expression value) was kept for analysis. 'Presence Score' was calculated as a percentage of all arrays within a study. Because genes with 'Marginal' scores were found only on a small number of arrays, all of these were considered 'Absent'.

SCAN Normalization. Microarrays were normalized using the Single-Channel-Array-Normalization (SCAN) method [15], to allow for cross-study comparisons and eliminate batch effects. This method normalizes each chip to itself rather than to the group of chips being analyzed. Thus, adding additional microarrays to any individual study had no effect on the normalization procedure. The intensity values obtained from this normalization were then used for co-variation analysis.

Weighted Gene Correlation Network Analysis (WGCNA). WGCNA [16] was used to identify modules containing important sets of primitive and differentiation genes. Graphs of scale independence, mean connectivity, and dendrograms exhibiting hierarchal clustering were constructed using the R program. Analysis was performed on children, using three separate gene listings: a) genes with highest coefficient of variation in the sample; b) NIH *Stem Cells Interest Group* listing of over 300 primitive and differentiation genes; and c) genes collected by co-variation analysis of the original CVM (Pearson Correlation > 0.4).

Matrix and Network Construction. Pearson correlation and Matrix analysis was performed using Partek Discovery Suite. Clique Mining analysis was performed using Matlab Software as well as Ohio Supercomputer Services (OSC). To mine patterns from the gene co-expression matrix data, we followed the network mining and merging workflow described by Xiang et al. [17]. First, we converted a gene co-expression dataset into an unweighted graph by creating an edge between any two genes with an absolute correlation value greater than 0.6. After the graph was created, we applied the Bron-Kerbosch algorithm [18] to generate all maximal cliques. We then applied the network merge approach [17] to these cliques under various density thresholds, depending on the study, which guaranteed that each resulting sub-network induced a sub-matrix with an average correlation value greater than the threshold used on the original gene co-expression matrix. Finally, we visualized the discovered sub-networks using Gephi [19] (https://gephi.org/).

Isolation of PBMCs for PCR. Blood was collected into a BD Vacutainer K2 EDTA (BD Bioscience, Franklin Lakes, NJ), diluted 1:1 with Hanks balanced salt solution (HBSS) (Invitrogen Life Technologies, Grand Island, NY), layered onto one volume of Lymphocyte Separation Medium (Cellgro Mediatech Inc., Manassas, VA), and centrifuged at 700 x g for 30 min at room temperature. The mononuclear cells were collected, diluted 1:1 with washing buffer (PBS supplemented with 2 mM EDTA and 2% FBS) (Gemini Bio-Products, West Sacramento, CA), and centrifuged at 300 x g for 7 min. The pellet was resuspended in 5 mL of 0.8% ammonium chloride solution (STEMCELL Technologies Inc., Vancouver, Canada) for 5 min to lyse any remaining erythrocytes. To remove as many platelets as possible, the cells were washed two more times as described above with two volumes of washing buffer.

RNA extraction and qRT-PCR. Total RNA was isolated using the RNeasy Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's protocol, tested for quality, and stored at -80ºC until use. The VILO kit (Life Technologies/Invitrogen, Grand Island, NY) was used to reverse transcribe 150 ng of total RNA. Primers (Qiagen) were diluted 1:20 with molecular-grade water, and 5 µL/well were added to 384-well plates using a Biomek® FX Laboratory Automation Workstation (Beckman Coulter, Inc., Brea, CA). The plates were left to dry overnight in a sterile hood and stored covered at -20ºC until use. qRT-PCR was performed using SYBR Green (Qiagen) and a 7900HT Real-Time PCR System (Life Technologies/Applied Biosystems, Foster City, CA) operated in standard mode. All of the runs contained a dissociation step. The samples were amplified in duplicate in a total volume of 5 µL. The results will be expressed as the relative copy number (RCN), defined as $RCN = 2^{-\Delta Cq} \times 100$, where ΔCq is the difference Cq(target) – Cq(reference) [20]. As a reference for normalization, we used the median Cq values of three endogenous controls (beta-2 microglobulin, GAPDH and RPL13).

Data analysis. ANOVA, t-test, Mann-Whitney test, and various correlation statistics (Pearson correlation, linear regression, principal component analysis, hierarchical clustering, Cronbach's alpha) were performed using JMP 10.0.2 (SAS Institute, Inc., Cary, NC), Partek Discovery Suite v. 6.4 (Partek 010 Inc., St. Louis, MO), and Microsoft Excel 2010 programs. In all of the statistical analyses, $p < 0.05$ was considered significant.

Results

Microarrays are routinely used to look for differences between different sets of samples, including differences in transcription levels of genes, identifying covariation between different genes or gene sets, and a wide variety of other applications. However, these studies are typically performed on well-known cell types, and the transcriptional analysis can only be performed on genes which are reliably detected on the gene chips themselves. Microarrays are notorious for their inability to detect rare gene transcripts in samples (measured by the 'Presence Score" i.e. the frequency of detected signal in replicate chips; usually a gene is considered reliably detected if it is detected in more than 50% of the replicates) [21].

Our initial PCR study detected the expression of an original panel of 45 primitive and differentiation genes in PBMCs isolated from peripheral blood in every tested individual. However, we wanted to verify the detectability of the same genes in microarrays due to their known sensitivity issues.

1. Assessment of sensitivity of microarrays to detect genes with low expression. (Table 1)

$CD34^+$ Dataset: An initial analysis using the MAS5 normalization method was performed to determine the detection levels of our original panel of 45 primitive and differentiation genes. To gauge the sensitivity of the microarrays for these rare, primitive genes, a study using microarrays analyzing $CD34^+$ cells [GSE23025 (n=56 samples)] was used. This study showed the majority of the primitive genes were detected by microarrays, but of those that were not, many were members of the CVM. Overall, only 4 members (OLR1, NOTCH4, cKIT/CD117, and ABCG2) were reliably detected (Presence Score > 0.5).

Adult Dataset: We then compiled microarrays from seven different studies [GSE21942 (15), GSE27034 (18), GSE14642 (20), GSE11761 (20), GSE46480 (98), GSE8507 (17), GSE10041 (23), GSE19743(63)], totaling to 274 microarrays analyzing PBMCs from a diverse population of healthy controls. From this, it was identified approximately half of the original panel of 45 genes could not be detected reliably (Presence Score < 0.5). In addition, none of the CVM genes were detected on all chips, and only two CVM genes (OLR1 and NOTCH4) had reliable detection levels (Presence Score > 0.5). Because of this, we sought to identify new populations which would contain higher levels of CSPCs in circulation and could therefore increase the detection levels of CVM genes so they may be analyzed.

Children Dataset: We selected a study focused on children [GSE13501(59)] as it has been shown children have higher levels of stem and progenitor cells in circulation. After performing a MAS5 normalization and analyzing detection levels of the original panel of genes, we found many of the variably detected, primitive genes in adults increased their detection levels in children, while the majority of the genes which were not detected at all in adults remained undetected in children as well, supporting both hypotheses of low sensitivity in microarrays

while having higher levels of CSPCs in children. Most notably, however, many of the variably detected CVM genes increased in detection levels, accounting for 5 CVM genes (NOS3, NANOG, MAP2, OLR1, and NOTCH4) to be reliably expressed.

Burn Dataset: Final detection analysis was performed on a study using burn victims at two time stages (early – 1-10 days, and late – 11-49 days) (GSE19743). This study was used as damage from burn injury would elucidate a response of increasing stem and progenitor cells in circulation in order to accommodate to the needed repairs of the cardiovascular system. In addition, unlike a study on a particular cardiovascular disease, the majority of these patients were not on drugs, nor did they have other conditions which may induce changes in their gene expression levels. Thus, this allowed for us to study this condition without worry of outside influence. Unexpectedly, the detection levels of the majority of the CVM members were decreased on microarrays from patients while other genes (i.e. OLR1) increased in detection. This unexpected result of OLR1 increasing in detection levels, while all other detected CVM genes decreased in detection, suggests a more complex relationship exists within this system.

Table 1. Detection of select primitive and differentiation genes on Affymetrix GeneChip microarrays.

Affymetrix ID	Gene Symbol	RefSeq Transcript ID	Presence Score (Detection Levels)				
			CD34+	Adult	Children	Early Burn	Late Burn
208204_s_at	CAV3	NM_001234	0%	0%	0%	0%	0%
1570276_a_at	GATA4	NM_002052	0%	0%	0%	0%	0%
209351_at	KRT14	NM_000526	0%	0%	0%	0%	0%
218678_at	NES	NM_006617	0%	0%	0%	0%	0%
206578_at	NKX2-5	NM_001166175	0%	0%	0%	0%	0%
217711_at	TEK	NM_000459	0%	0%	0%	0%	0%
213869_x_at	THY1	NM_006288	0%	0%	0%	2%	0%
207175_at	ADIPOQ	NM_001177800	25%	15%	0%	11%	16%
209543_s_at	CD34	NM_001025109	88%	9%	0%	0%	2%
211201_at	FSHR	NM_000145	0%	0%	2%	0%	0%
214468_at	MYH6	NM_002471	0%	5%	2%	0%	0%
203951_at	CNN1	NM_001299	9%	1%	2%	0%	2%
202112_at	VWF	NM_000552	71%	3%	2%	0%	7%
214837_at	ALB	NM_000477	4%	3%	3%	4%	4%
217430_x_at	COL1A1	NM_000088	39%	32%	7%	4%	2%
204677_at	CDH5	NM_001114117	2%	0%	10%	0%	0%
214532_x_at	POU5F1**	NM_001159542	0%	1%	12%	0%	2%
203934_at	KDR	NM_002253	16%	16%	12%	12%	5%
215783_s_at	ALPL	NM_000478	36%	82%	15%	100%	98%
209735_at	ABCG2**	NM_001257386	73%	10%	24%	9%	11%
205051_s_at	KIT**	NM_000222	98%	24%	44%	12%	7%
204304_s_at	PROM1**	NM_001145847	100%	28%	61%	21%	42%
205247_at	NOTCH4	NM_004557	64%	67%	61%	54%	42%
212022_s_at	MKI67**	NM_001145966	18%	1%	64%	23%	65%
210004_at	OLR1	NM_001172632	89%	64%	64%	82%	100%
225540_at	MAP2**	NM_001039538	21%	8%	78%	16%	19%
220184_at	NANOG**	NM_024865	11%	0%	83%	0%	2%
203949_at	MPO**	NM_000250	95%	24%	83%	75%	93%
200974_at	ACTA2	NM_001141945	98%	99%	92%	98%	100%
229093_at	NOS3**	NM_000603	0%	0%	93%	0%	0%
206956_at	BGLAP	NM_001199661	84%	92%	93%	79%	81%
214354_x_at	SFTPB**	NM_000542	13%	21%	98%	7%	11%
203507_at	CD68**	NM_001040059	39%	84%	98%	82%	74%
201313_at	ENO2**	NM_001975	91%	70%	100%	25%	19%
217650_x_at	ST3GAL2**	NM_006927	96%	84%	100%	88%	91%
212224_at	ALDH1A1	NM_000689	96%	93%	100%	39%	33%
205456_at	CD3E	NM_000733	25%	99%	100%	93%	93%
1555779_a_at	CD79A	NM_001783	38%	99%	100%	91%	86%
203939_at	NT5E	NM_001204813	55%	99%	100%	81%	65%
205898_at	CX3CR1	NM_001171171	61%	100%	100%	100%	100%
201743_at	CD14	NM_000591	88%	100%	100%	96%	96%
205786_s_at	ITGAM	NM_000632	96%	100%	100%	100%	100%
217028_at	CXCR4	NM_001008540	100%	100%	100%	100%	100%
208982_at	PECAM1	NM_000442	100%	100%	100%	100%	100%
212587_s_at	PTPRC	NM_001267798	100%	100%	100%	100%	100%
201891_s_at	B2M	NM_004048	100%	100%	100%	100%	100%
212191_x_at	RPL13	NM_000977	100%	100%	100%	100%	96%
213798_s_at	CAP1	NM_001105530	100%	100%	100%	100%	100%
217398_x_at	GAPDH	NM_001256799	100%	100%	100%	100%	100%

Table 1: Poor sensitivity of microarrays for detecting genes with low transcripts. Affymetrix microarrays from listed GSE studies were analyzed. The table contains 'presence scores' on Affymetrix GeneChips® of the selected gene panel (see Fig. 1). Highlighted (yellow) are members of the CVM module. Stars (**) represent genes whose presence levels were higher in children than adults. Most CVM genes were not present in a study where CD34+ cells were pre-selected (of note, even CD34 had an incomplete representation).

Conclusion: Our results indicate that the sensitivity of gene microarrays poses a severe problem in detecting and analyzing genes associated with rare cell types (i.e., CSPCs). This validates our reasoning for performing the initial analysis on the original panel of 45 primitive and differentiation genes. That being said, we were able to reliably detect some members of the CVM on microarrays and could utilize these genes as 'seeds' to identify important modules and potential extension candidates. In addition, we identified populations in which the seed genes, and potentially other genes associated with CSPCs, were better detected and could use these populations for more robust results.

2. Attempt to detect the localization of CVM genes in a WGCNA-detected module.

Next, we performed a "top-down" bioinformatics procedure known as Weighted Gene Correlation Analysis (WGCNA, Diagram 1) to identify new genes as candidates for CVM expansion [4]. Our approach consisted of identifying modules which already contained the appropriate seed genes (i.e. genes belonging to the CVM) by an established method for microarrays. A normalization method known as Single-Channel-Array Normalization (SCAN) was performed on the microarrays [15]. This normalization method has been shown to be more accurate by eliminating 'batch effects' and allowed us to include additional studies in our analysis due to its innovative normalization procedure.

To start, we isolated 8,000 genes with the highest coefficients of variation on microarrays from the study which had the greatest detection of the CVM genes (GSE13501). When identifying genes of the CVM, we found none of them were listed among these 8000 most varying genes. We then expanded the gene listing used to the 12,000 most varying. Only one CVM gene, OLR1, was included in this set.

Because of this, we decided to use a more targeted method for network construction. The more targeted method we selected was to combine the guilt-by-association approach with the WGCNA analysis. The purpose was to use GBA to aid in identifying genes of interest using genes which are already relevant (i.e. the seed genes). Using the GBA principle, we isolated 2,755 genes which co-varied (Pearson correlation > 0.4) with at least three of the five CVM members. Thus, each of these genes had multiple connections with the CVM module, even before performing the WGCNA analysis. We then attempted to identify the β value for the Scale Free Topology Model (signed R^2) and found β values to reach a threshold of 0.8 (as suggested in the paper; Equation 1: $a_{ij} = \text{power}(s_{ij}, \beta) = |s_{ij}|^{\beta}$) were very large (>30). Thus, both the Scale Independence and Mean Connectivity graphs were largely right-skewed. There are multiple potential reasons for this, the most likely being the large interconnectivity between the isolated genes due to their common correlations with three of the five CVM members. This prevented us from moving further with this analysis as such high β values lead to virtually all values used in the adjacency matrix to be approximately 0 (as seen by the Equation 1), and much of the useful information would be lost.

We then attempted to use a listing of over 300 stem and progenitor cell-related genes suggested by the NIH's *Stem Cells Interest Group*, including our CVM genes (only used genes with detection scores > 0.5). Again, we were unable to move forward with this analysis as the β values were extremely high, most likely from the large interconnectivity between the stem and progenitor cell-related genes, the limited variance in intensity values associated with the majority of these genes, and the limited number of genes in the analysis.

We concluded the WGCNA method could not help identify the localization of CVM seed genes in a relevant module. Thus, our goal of identifying important modules and genes associated with them could not be ascertained due to the gene isolation procedures as well as the functions contained within the analysis.

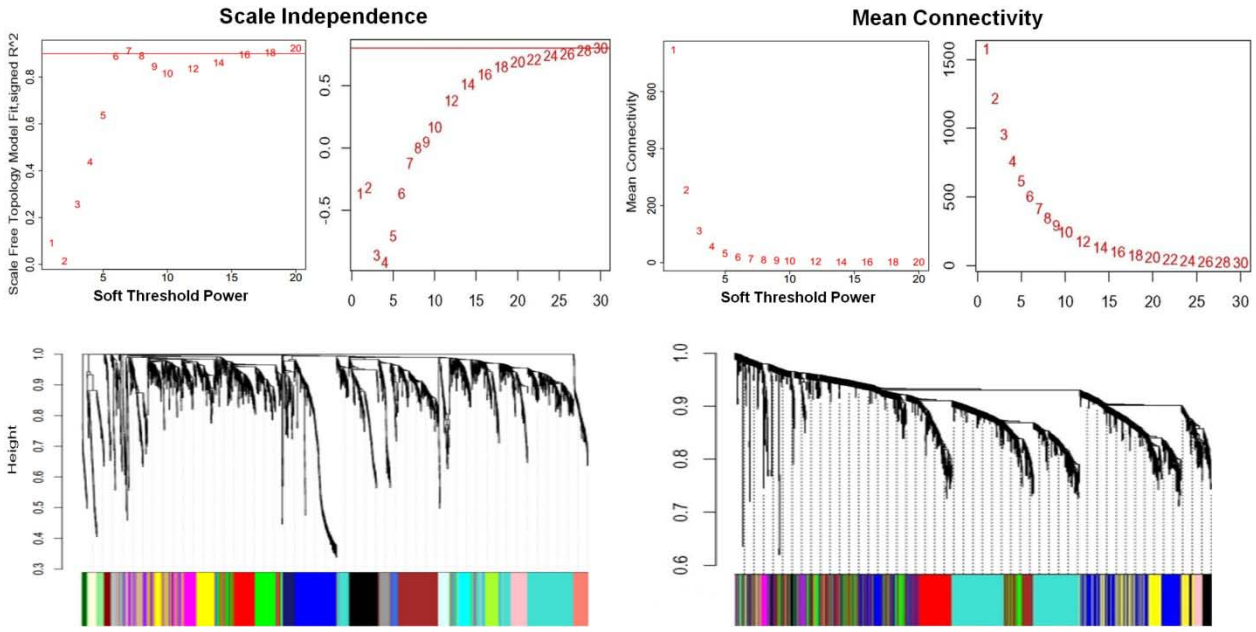

Figure 2. WGCNA attempt to identify additional CVM candidate genes. Scale-Free Topology Graph, Mean Connectivity, and Cluster analysis of genes co-varying with at least three CVM members, performed by Weighted Gene Covariation Network Analysis (WGCNA). This was done using 12,000 of the most variable genes, a listing of over 300 stem and progenitor cell-related genes suggested by the NIH as well as 2,755 genes gathered from the GBA method which were correlated with ($R > 0.4$) to at least 3 of the 5 present CVM genes (Presence > 0.5). Left side represents typical results of WGCNA method while the right side represents results from our GBA analysis using WGCNA. Note that higher beta values lead to a loss of information as the Pearson correlations approach zero. Higher mean connectivity values as well as higher beta values required to reach Scale Free Topology would not allow for this analysis to be used for stemness-related genes in peripheral blood. None of the identified modules contained CVM genes.

3. Attempt to localize CVM in modules obtained by 'Clique Mining'. Next, we reverted to our original method of module detection and construction known as 'Clique Mining', developed by our collaborator, Dr. Kun Huang's, team [17]. Using the burn study (GSE19743), because of its functional relevance to CSPC mobilization, we again used the gene list of stem and progenitor cell-related genes from the NIH's *Stem Cells Interest Group*, while also adding our CVM genes. After isolating detected genes (Presence Score > 0.5) from this listing, we created similarity matrices for the control group and for early and late burn stages. This was followed by network construction via Clique Mining to identify modules and module membership. One module in particular (seen in blue) was largely similar between the two stages of burn patients (though there were no similarities with the control group). Of the CVM genes, NOTCH4 was found to be associated with this module in the early stage victims (excluded from late stage due to Presence score < 0.5 – View Table 1). However, NOTCH4's connectivity and integration within this module was very low, forcing us to move to an even more targeted method of analysis.

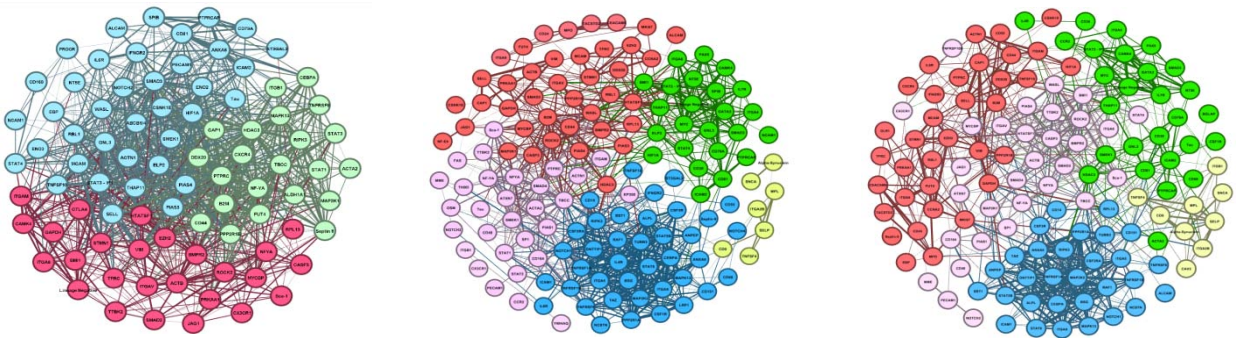

Figure 3. Covariation-based network structure of an extended panel of primitive and differentiation genes occurring in PBMCs in response to burn injury (based on the extended gene panel compiled by NIH's *Stem Cells Interest Group*)**.** Network detection of genes with a Presence Score > 0.5 via Clique Mining was performed on over 300 suggested stem and progenitor cell-related genes. Networks were created for controls **(A)**, early burn victims (1-10 days) **(B)**, and late burn victims (11-49 days) **(C)**. **A**. No modules were common between the controls and burn victims, as is established by the different colored modules. **B**-**C**. However, similar modules did arise between the early and late burn victims. Two modules had a majority of their members present in both groups (blue and yellow) while the red and green modules had more limited similarities between groups.

In conclusion, none of the seed CVM genes were vastly connected in modules detected in healthy controls**,** early stage burn victims (1-10 days) and late stage burn victims (11-49 days) with genes suggested by the NIH's Stem Cells Interest Group. However, we did detect a module which seemed to play an important role in response to burn injury, and this module did contain, though with limited connectivity, a key gene of the CVM, NOTCH4. This module was later analyzed in more depth because of its functional relevance to cardiovascular repair and because of the inclusion of NOTCH4 within its module membership.

4. Reconstruction of CVM extensions by a 'bottom-up approach' (GBA followed by Clique Mining).

Because of the inability to detect modules containing CVM genes using standard gene network analysis approaches, we conceived a new method to serve this purpose. The GBA principle was used to identify genes which exhibited the greatest co-variation with our CVM seed genes. Thus, we created 'neighborhoods', or gene listings of the most co-varying genes, for each of the detected CVM genes. By default, this identified the CVM gene as being connected to every member of the module, *thus identifying the seed gene as a hub of the module.* Neighborhoods were then compared between studies for each seed gene and within studies between the seed genes to detect interconnectivity between them.

a. Children study. Initial analysis was performed on GSE13501 as it had the most detected CVM genes. We isolated 'neighborhoods', or sets of the largest co-varying genes, for each of the detected CVM genes (Pearson Correlation > 0.7). The end result was a set of neighborhoods for NOTCH4, NANOG, NOS3, and OLR1 which contained possible CVM extension candidates to add to the previously PCR created module (MAP2 is not included as none of its correlates were above the set threshold). Each of these neighborhoods can be seen as modules with their 'seed' gene (or original CVM gene) as the hub of the module. We then merged these modules together to elucidate the connectivity that existed between the modules (as it would be expected for them to be interconnected due to the known interconnectivity of the seed genes validated by PCR) and performed Clique Mining analysis. Indeed, members of NOS3's, NOTCH4's, and NANOG's neighborhoods were largely interconnected (edge weight > 0.6) with one another, as is exhibited by the blue-purple color gradient between their members (Figure 4). In addition, NOTCH4's and NOS3's neighborhoods were positively correlated with one another while being *negatively* correlated with NANOG's neighborhood (data not shown). However, seed CVM genes themselves shared no correlations with one another. Interestingly, OLR1 and its neighborhood was defined as a distinct module with no connectivity with other neighborhoods unless lowering the edge weight threshold (edge weight < 0.5).

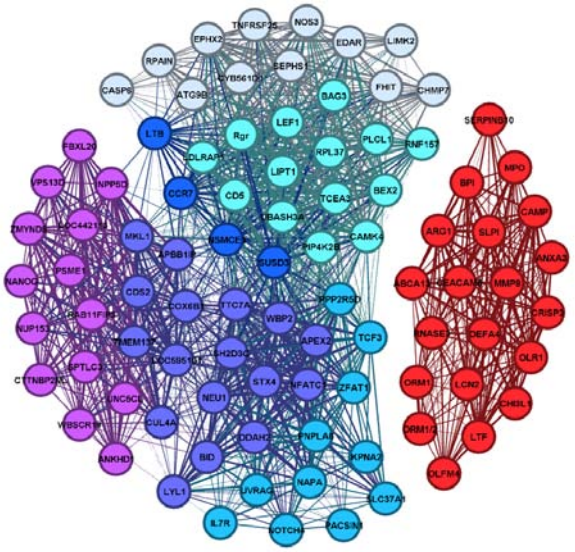

Figure 4: Network reconstruction from the covariation 'neighborhoods' of CVR seed genes in the children study. Network creation and module detection was done by performing regression analysis followed by a Clique Mining visualization on Affymetrix GeneChips® based on RNA isolated from PBMCs of healthy children. The red module consists of the seed gene OLR1 and its most highly correlated 'neighbors'. Additional seed genes, NOTCH4 (bottom), NANOG (left), and NOS3 (top), are also represented along with their most highly correlated neighbors (Pearson correlation > 0.7). Genes on the gradient of purple to blue are highly interconnected with one another as established by the large density of edges between them. OLR1's neighborhood has no connections with the others at this edge threshold (w = 0.6).

In addition, this neighborhood exhibited the highest connectivity and largest strength in connectivity between its members. However, even when lowering the threshold, connections could only be seen with LCN2 (a characteristic which later led us to include it for PCR validation). Each of the candidates belonging to the four neighborhoods was functionally verified for relation to stemness, differentiation, angiogenesis, neovascularization, cardiovascular disease, and/or cardiovascular repair in literature. In accordance with the GBA principle, the vast majority of these genes were found to have at least a minor direct or indirect association with these criteria.

Adult study. To confirm these results, we analyzed the adult study composed of 274 microarrays from seven different studies to detect if similar neighborhoods would be detected. As already observed, only two CVM genes, OLR1 and NOTCH4, were detected at reliable levels

(Presence Score > 0.5). A similar analysis to that on the children study, GSE13501, was run to isolate neighborhoods for OLR1 and NOTCH4. OLR1 had overall lower correlation strengths (only one gene, MPO, above threshold 0.7), potentially explained by lower detection levels. This forced us to lower the threshold to 0.6 for neighborhood creation. We found the majority of its neighborhood members constituted the module seen in the children study. To identify module members which may have been left out, we lowered the threshold to 0.4 ($p < 0.05$) and found the vast majority of members (14) seen in GSE13501 were also seen in the adult study (only SLPI, CHI3L1, ORM1, ORM2 and ANXA3 did not fit this criterion). Because of this, we decided to test if the connection strength between members of the children neighborhoods were still high in adults by performing a Clique Mining analysis. Indeed, although the connection strengths were lower, there was still a high level of interconnectivity among these genes.

NOTCH4's neighborhood, unlike OLR1's, was dissimilar between the children and adult studies. The neighborhood was constructed using a threshold of 0.7 for the adult correlation while only using genes that had a correlation of at least 0.6 in the children study. Thus, only genes which had a relatively high correlation in each study were used, though these were not the most correlated genes within each study. Only 2 members, TTC7A and ZFAT1, were corresponding between the children and adult neighborhoods of NOTCH4 (both had Pearson correlation > 0.7 in both studies). This led to their selection for PCR verification.

Similar to the analysis on neighborhoods compiled for children, neighborhoods for the adult group were merged to look for interconnectivity between them. Again, however, NOTCH4's and OLR1's neighborhood shared no connectivity (edge weight = 0.6) unless the edge weight threshold was lowered (as seen in Figure 5, at 0.4). At this threshold, very few genes from OLR1's neighborhood correlated with NOTCH4's, but again, LCN2 had the most connections between the two modules.

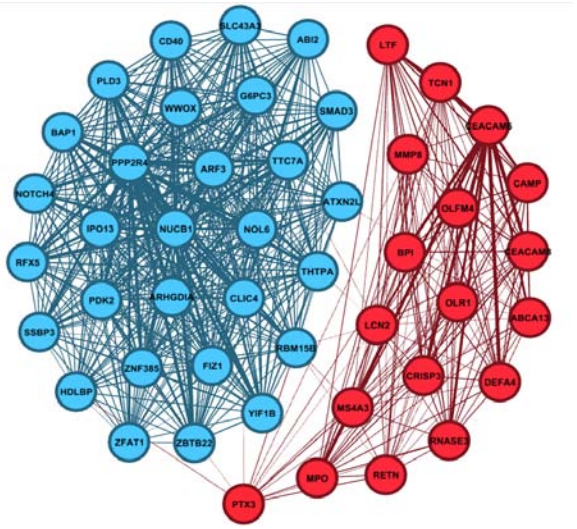

Figure 5. Network reconstruction from the covariation 'neighborhoods' of CVR seed genes in the Adult study. Network creation and module detection was done as in Fig. 4. Colors corresponding to NOTCH4's and OLR1's neighborhoods were kept consistent as NOTCH4 and OLR1 were the only seed genes used in the adult analysis (Presence Score > 0.5). While most of NOTCH4's members are vastly dissimilar from the children network, many of OLR1's neighbors remained highly correlated in the adult study. Edge strength was lowered (w = 0.4) to show the consistent limited connections between the two modules.

Because of the relative consistency of OLR1's neighborhood, we decided to look for OLR1 and its correlated genes in literature to find any significant functional relevance. We found a study on preeclampsia[22], a condition characterized by high blood pressure and protein levels in urine of pregnant women, which contained the majority of OLR1's neighborhood, in addition to OLR1 itself. This study identified genes exhibiting the largest negative fold change in expression levels between controls and early and late-onset preeclampsia patients, rather than looking at co-variation. Thus, in addition to the similarities seen between the children and adult studies, a functional relevance to cardiovascular

function could now be attributed to this module. Furthermore, our "bottom-up" analysis shows the additional information that can be obtained by such studies on disease pathologies and their effect on gene expression values.

c. Burn study. This finding of OLR1's consistent network in relation to a cardiovascular disease led us to reanalyze GSE19743 for relation of these genes to burn victims. We took the normalized expression levels from controls and compared them to early and late stage burn patients. We then looked to identify genes which exhibited the largest fold change (threshold set at 1, or 10x expression level change). We decided to limit our search to only those genes which are a part of OLR1's neighborhood in children / adult as well as genes which exhibited the largest fold change in preeclampsia due to their known functional relevance. Not surprisingly, virtually all of OLR1's related genes were identified as exhibiting the largest fold change in burn victims, though OLR1 itself did not. Even more interesting, rather than exhibiting a *negative* fold change, as was seen in preeclampsia, a condition in which there is a defect in the repair mechanisms, these genes exhibited the largest *positive* fold change in burn victims, a condition in which the repairing mechanisms would be "activated", fitting our hypothesis of OLR1 and its neighborhood as having a repairing mechanism in the cardiovascular system.

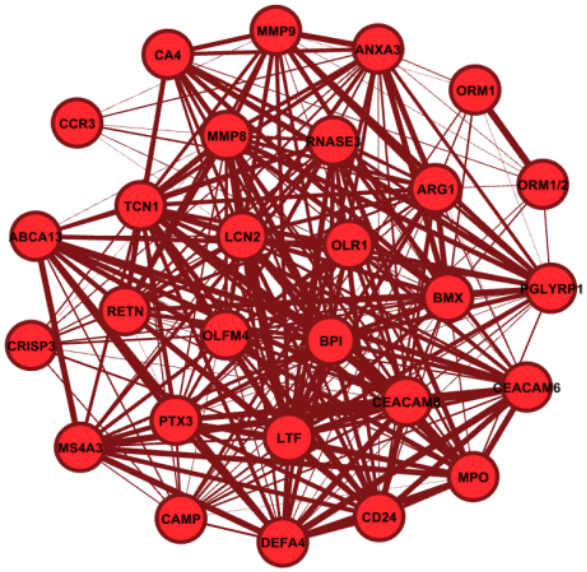

Figure 6: Network reconstruction from the co-variation 'neighborhoods' of OLR1 seed genes in the Burn study. 'Seed' genes were required to have a presence score of at least 50%. Genes exhibiting the largest fold change in expression value from controls to early and late term burn victims were recorded. These were compared to genes which had high fold change in expression level from the preeclampsia study. Genes with large fold changes (>1 on log scale) in both studies, as well as genes with large fold changes that were represented in OLR1's network in children, were kept for analysis. Module representation shows much higher overall connectivity and strength in connectivity in OLR1's module, though OLR1 is slightly less connected.

In addition to this finding, we isolated genes which had prevalence in the children, adult, and/or preeclampsia studies and also exhibited a large fold change (>1) in burn victims (we also included OLR1 for comparison). We then ran an 'all in one' Clique Mining analysis on this set of genes using microarrays from controls, early stage, and late stage burn victims. Yet again, we found all of these genes were highly interconnected with one another and OLR1 was moderately connected within the network. In addition, the connectivity of this network was higher in burn victims as compared to the study on healthy, adult controls (data not shown).

d. Pathway analysis. As a final analysis, we decided to analyze OLR1's neighborhood using the Ingenuity Pathways Analysis (IPA) program (http://www.ingenuity.com/) to identify the pre-established signaling pathways the genes in our modules are known to be associated with. Of the top associated functions, cardiovascular system development and function was the second most prevalent. In addition, this module occurred as having upstream regulators of

lipopolysaccharide and CEBPA, a transcription factor known to be associated with the differentiation of hematopoietic cell types.

In addition, we did the same analysis on the module containing NOTCH4 from the original burn study. Encouragingly, the second highest associated function was again cardiovascular system development and function. Even more interesting, upstream regulators of this module were the same as OLR1's, i.e. lipopolysaccharide and CEBPA. Thus, not only is there co-variation within, and functional significance of, OLR1's network, but there is an additional relevance in control of its expression via a common transcription factor, CEBPA, with another member of the CVM and genes correlated with it.

Genes with Large Fold Change				Burn Study			Preeclampsia Study			
Entrez	Affymetrix ID	Gene Symbol	Function	1-10 Day (Early) Fold Change (Compared to Controls)	11-49 Day (Late) Fold Change (Compared to Controls)	11-49 Day Fold Change (Compared to Early)	Earlyn onset Fold Change (Compared to Controls)	p-value	Late onset Fold Change (Compared to Controls)	p-value
4317	207329_at	MMP8	Plays an important role in plaque angiogenesis	**2.63**	**3.32**	0.68	-	-	1.99	0.037
4057	202018_s_at	LTF	Highly expressed in non-healing wound exudates	**2.36**	**3.09**	0.73	-	-	2.03	0.001
3934	212531_at	LCN2	An independentrenal predictor of incidence of AKI after surgical abdominal aortic aneurysm (AAA) repair	**2.30**	**3.20**	0.89	1.58	0.098	1.85	0.001
10562	212768_s_at	OLFM4	Marker of intestinal stem cells	**1.96**	**2.39**	0.43	-	-	1.88	0.077
10321	207802_at	CRISP3	Specific neutrophil granules protein, secreted as an extracellular matrix component	**1.51**	**1.12**	-0.39	-	-	1.83	0.052
671	205557_at	BPI	Antibacterial activity against Gram-negative bacteria	**1.43**	**2.43**	**1.00**	2	0.021	2.01	0.005
1669	207269_at	DEFA4	Neutrophil protein with antimicrobial activity against Gram-negative bacteria	0.55	**2.66**	**2.11**	2.36	0.011	2.49	0.001
4680	211657_at	CEACAM6	Implicated in cell adhesion, cellular invasiveness, angiogenesis, and inflammation	0.45	**2.42**	**1.96**	2.1	0.041	1.99	0.023
6037	206851_at	RNASE3	Eosinophil major basic protein with pro-angiogenic effects	0.54	**1.42**	0.89	-	-	1.89	0.027
154664	1553605_a_at	ABCA13	Markers of hMSC	0.18	**1.16**	0.98	1.56	0.086	-	-
820	210244_at	CAMP	Induces angiogenesis via PGE2-EP3 signaling in endothelial cells	0.59	**1.06**	0.47	-	-	1.66	0.002
4318	203936_s_at	MMP9	Involved in mobilization of hematopoietic progenitor cells from bone marrow as well as embryonic development, reproduction, and tissue remodeling through ECM degradation	**2.56**	**2.56**	-0.01	1.62	0.039	-	-
8993	207384_at	PGLYRP1	Binds to peptidoglycan of bacteria involved in human atherosclerotic lesions. Increasing levels associated with coronary artery calcium and aortic wall thickness	**1.69**	**2.12**	0.43	1.59	0.007	-	-
762	206209_s_at	CA4	Encodes an isozyme expressed on luminal surfaces of pulmonary and other capillaries and proximal renal tubules. Related to Ocular Hypotension	**1.54**	**1.37**	-0.17	1.52	0.040	-	-
660	206464_at	BMX	Plays a critical role in TNF-induced angiogenesis, and implicated in the signaling of TEK and FLT1 receptors, 2 important receptor families essential for angiogenesis	**1.54**	**1.61**	0.07	1.5	0.033	-	-
56729	220570_at	RETN	Resistin levels correlate with oxidative stress and myocardial injury in cardiac surgery patients. It may serve as a useful biomarker for ischaemia-reperfusion injury	**1.53**	**1.63**	0.09	-	-	1.53	0.060
6947	205513_at	TCN1	Influences human vitamin B-12 levels which have shown to be implicated in congenital heart disease via the folate metabolism pathway	**1.47**	**2.29**	0.82	1.75	0.012	1.68	0.008
1088	206676_at	CEACAM8	Leukocyte activation marker. Increased leukocyte activation is correlated with coronary artery disease, plaque destabilization, and vascular cell dysfunction	**1.18**	**2.86**	**1.68**	1.9	0.051	2.09	0.004
1232	208304_at	CCR3	CCR3-dependent chemokine interactions regulate endogenous migration of CD34+ progenitors from bone marrow to ischemic but not to normal myocardium	-1.01	-0.90	0.10	1.84	0.000	1.64	0.001
260429	1552348_at	PRSS33	Serine, protease predominantly expressed in macrophages	-1.02	-0.47	0.55	-	-	1.73	0.087
100133941	216379_x_at	CD24	Interacts with P-selectin which mediates rapid rolling of leukocytes over vascular surfaces during the initial steps in inflammation	0.70	**1.90**	**1.19**	1.63	0.039	1.52	0.033
932	210254_at	MS4A3	Hematopoietic stem cell cycle regulator	-0.05	**1.79**	**1.84**	1.83	0.096	1.63	0.085
5806	206157_at	PTX3	Modulates inflammatory processes, angiogenesis, atherosclerotic lesion development, and ECM formation. Released by vascular wall cells as an imflammatory marker	0.44	**1.18**	0.74	1.57	0.045	1.69	0.002
369	206177_s_at	ARG1	Reduces nitric oxide production and impairs endothelial function	**2.68**	**2.74**	0.05	-	-	-	-
306	209369_at	ANXA3	Potential angiogenic mediator	**2.21**	**2.31**	0.10	-	-	-	-
5004/5005	205041_s_at	ORM1/2	Acute phase reactant and bimodal regulator of angiogenesis	**1.00**	**1.06**	0.06	-	-	-	-
4353	203949_at	MPO	Leukocyte activation marker. Increased leukocyte activation is correlated with coronary artery disease, plaque destabilization, and vascular cell dysfunction	0.39	**1.61**	**1.22**	-	-	-	-
5004	205040_at	ORM1	Acute phase reactant and bimodal regulator of angiogenesis	0.98	**1.16**	0.18	-	-	-	-
1116	209395_at	CHI3L1	Expressed by macrophages, chondrocytes, and vascular SMC; it is a potent angiogenic factor	-0.97	0.03	**1.00**	-	-	-	-

Table 2. Genes with the largest fold change in burn and preeclampsia patients, as compared to prevalence in children vs. adults. Orange = Prevalent in burn victims, preeclampsia, and present in the children network; Yellow = Prevalent in burn victims and preeclampsia, but not present in children; Red = Prevalent in burn victims and present in the children network, but not prevalent in preeclampsia.

e. **Final conclusions of study on microarrays:** Many of the issues with this study arose from the lack of sensitivity of microarrays, especially in analyzing rare cell type populations whose transcriptome may not be well represented. Because our analysis relies on detection of genes associated with CSPCs, we were limited in the genes we could analyze for co-variation (as can be seen by Table 1). However, we were able to analyze five CVM genes in children, as well as two of these genes in healthy adult populations and a study using burn victims.

After running a guilt-by-association analysis on these detected genes, we were able to create neighborhoods of highly correlated genes for each of the CVM seeds. In addition, we analyzed these gene listings and found high levels of interconnectivity between neighborhoods and, in the case of OLR1, between studies. In addition, an analysis on expression fold change gave insight into OLR1's module and attributed to it a functional relevance to cardiovascular disease pathologies and repair functions (i.e. preeclampsia and response to burn injury).

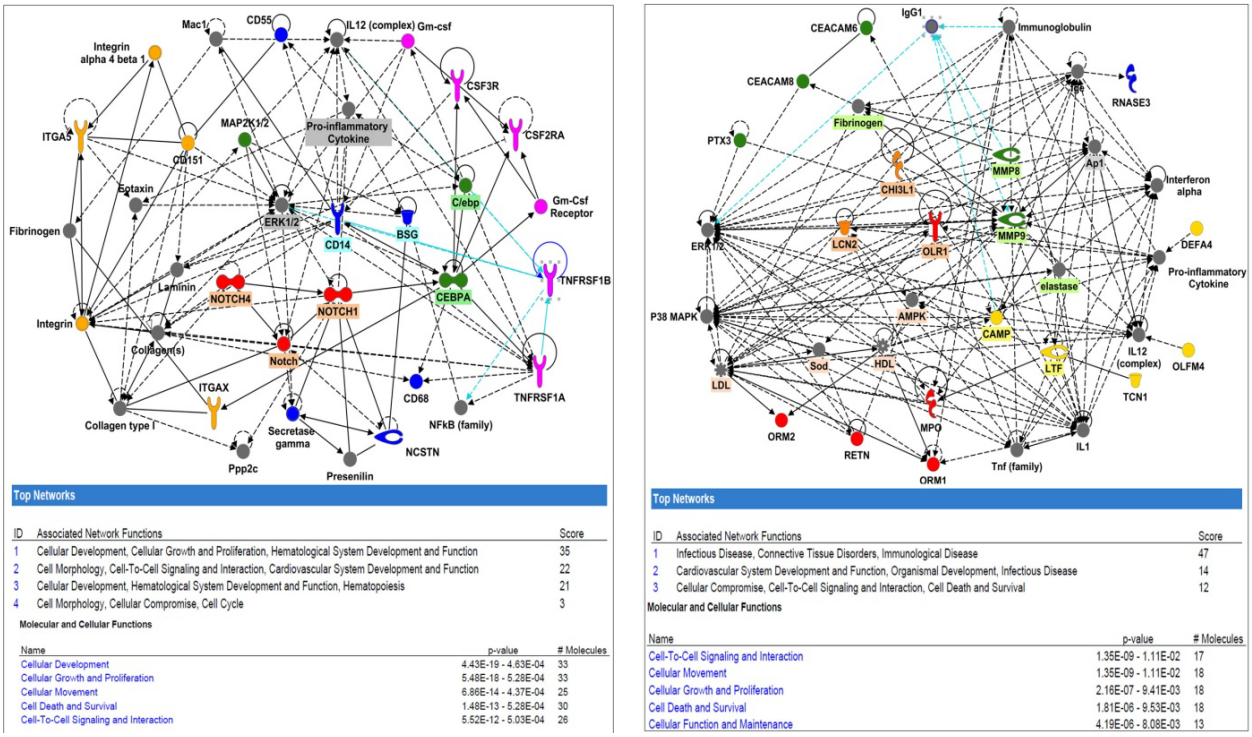

Figure 7: Signaling pathways of two gene clusters involved in response to burn injury, constructed with Ingenuity® Pathways analysis. (A) Pathways associated to 'blue cluster' (obtained by direct targeted network representation, Fig. 3 B,C). (B) Pathways associated to 'red cluster' obtained by GBA method. Of note, although the composition is different, both address similar functions and both are controlled by the transcription factor CCAAT/enhancer-binding protein alpha (CEBPA).

These insights led us to select multiple genes for PCR verification, to validate or exclude their inclusion as new members of the CVM. Of all of the genes identified to highly co-vary with the detected CVM seed genes, we selected low density lipoprotein receptor adaptor protein 1 (LDLRAP1), calcium/calmodulin kinase 4 (CAMK4), and epoxy hydrolase 2 (EPHX2), all of whom highly co-varied with NOS3 and had functional relevance to cardiovascular maintenance; megakaryoblastic leukemia (translocation) 1 (MKL1), which highly co-varied with NANOG ; NOTCH4 was found to be highly correlated with tetratricopeptide repeat domain 7A (TTC7A) and ZFAT zinc finger 1 (ZFAT1) in both children and adult studies, though functional validation of these genes to cardiovascular function was only moderately achieved as little is known about these genes. Finally, an essential gene to cardiovascular function, OLR1, had remarkably consistent correlations between all four conditions looked at. Of the genes constantly represented in its module, matrix metallopeptidase 8 (MMP8), lactotransferrin (LTF), lipocalin 2 (LCN2), and carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6) were selected for validation. These genes are discussed in more detail below.

Discussion

Within the circulation lies a wide variety of cells including red blood cells, leukocytes, and a set of very rare cells known as circulating stem or progenitor cells (CSPCs). Much as the erythrocytes supply oxygen to the body and leukocytes help maintaining the immune system of an individual, CSPCs play a role in the maintenance and repair of tissues and the cardiovascular system. These CSPCs are characterized by their primitive cell traits along with their capacity to differentiate into a wide variety of specialized cell types. Because their frequency in circulation and presumed functions (as tested *in vitro*) are inversely related to the severity of cardiovascular risk factors, CSPCs are considered promising biomarkers of the organism's resilience and/or ability to respond to injury. However, despite the progress in the basic biology of these cells, the translational applications for diagnostic, prognostic, and therapeutic effects are still in development. The cause of this delay lies in the rarity of CSPCs, their heterogeneity, and the complexity of relationships among themselves and the target tissues.

A prime example of these problems lies in the simple classification of these cells, which, to this day, is still not well defined [23]. Many studies use molecules such as CD34, CD90, or CD117, all seen as putative markers of primitivity, to identify or isolate these cells from circulation. However, it is well known that these common markers do not yield a true and thorough sample of stem and progenitor cells [24]. In addition, much of the obtainable information could be lost, as considering only a single gene or a small subset of genes, which prevents the analysis of gene interactions. Thus, CSPCs represent a cellular system that consistently has resisted the attempts to reliably describe it in a coherent, yet simple enough manner to be translationally useful.

In response to this issue, we proposed CSPCs could be directly assessed in blood samples by transcriptional analysis (gene expression) of PBMCs. This method is more robust, compared to single/few markers analysis, as the transcriptome of an undifferentiated cell at one stage of differentiation will be different from the transcriptome of a cell at a slightly more differentiated stage. Thus, the overall transcriptional activity of a cell is extremely sensitive to its differentiation status, much more so than any single gene. By performing a transcriptional analysis on a wide range of genes rather than a small set, as is used in most studies, a more accurate representation and analysis of these specific CSPCs can be obtained.

Based on these premises, we created a sensitive gene expression array, using quantitative real-time polymerase chain reaction (qRT-PCR), capable of detecting the presence of CSPCs among PBMCs, as a well-defined transcriptional module composed of primitive and cardiovascular differentiation genes (*in press*). Out of 45 tested genes representing the most used markers of primitivity and differentiation, we found 15 genes which strongly co-vary and inversely depend on the vascular stiffness of blood donors, as well as their age and blood pressure. We suggested this gene cluster, termed the Cardiovascular Module (CVM) because of its protective role, could be used as a new systemic biomarker for cardiovascular health.

The next logical step was to explore the actual, full composition of the human CVM by using an unbiased method. Here we show how this can be performed through microarray analysis of gene expression by an original method and using a public database (Gene Expression Omnibus (GEO)). Using it, we identified genes which highly co-varied with members of the CVM from our PCR study, a principle known as "guilt-by-association" (GBA).

Sensitivity of microarrays: However, a key limitation we encountered was the much lower sensitivity of microarrays, as compared to PCR, for obtaining expression values or even

detecting the presence of the genes of interest. This inherently limits the ability to run co-variation analyses on the data. We thus set a standard of detection > 0.5 in order for a gene from the original panel to be reliably analyzed by co-variation.

Such limitations were first noted when analyzing a study on $CD34^+$ isolated cells hybridized to microarrays. This study was used to gauge the sensitivity of microarrays to a particular cell type known to be associated with CSPCs in relation to our CVM. Of the original panel representing the 45 most used markers of primitivity and differentiation, the majority were not reliably detected by microarrays (Presence Score < 0.5). In addition, approximately half of the CVM genes were not detected at all (Presence Score = 0), and none of the CVM members were detected by 100% of the chips. Furthermore, even CD34 was detected on only 88% of the microarrays from the study, further confirming the lack of sensitivity of microarrays. However, four genes (ABCG2, KIT, NOTCH4, and OLR1) were reliably detected on these microarrays.

Encouraged by this result, we analyzed a set of 274 microarrays hybridized with RNA from studies on PBMCs from healthy controls (similar to our PCR model). These arrays were in seven different databases, representing diverse populations and age groups (referred to as the 'Adult study' or database). Our data shows detection of the majority of the original panel of genes was again below acceptable levels for analysis (Presence Score < 0.5), and detection of the CVM members decreased from the $CD34^+$ study, leaving only two, OLR1 and NOTCH4, being reliably detected for further analysis.

Because our CVM was inversely related with age, and because studies have shown children have increased levels of stem and progenitor cells in circulation, we turned our attention to a study focused on PBMCs isolated from healthy children (GSE13501). In accordance with our PCR data, detection of the majority of members of the original panel increased, including several members of the CVM. Our data shows CVM genes NOTCH4, OLR1, MAP2, NANOG, and NOS3 were now reliably detected.

As a model for vascular repairing mechanisms, we also selected for analysis a study on burn patients based on several premises: i) overall, CSPCs represent the same cellular system, irrespective of the specific organ-level function or pathology; ii) burn injury response could trigger an increase of CSPCs in circulation due to increased need for repair; iii) the majority of victims had no additional disease pathologies, as would be expected in a particular cardiovascular disease population (i.e. a hypertensive patient with diabetes); iv) the majority of victims would not be under prescription medications which could alter gene expression, as would be seen with patients diagnosed with a cardiovascular disease. These premises allow for the data to be analyzed without additional variables of consideration, solely focusing on the repairing mechanisms elicited by the response to injury.

Comparing detection values of burn victims to the healthy adult population described earlier, our data show detection of some CVM members actually *decreased* in response to burn injury (except for OLR1). We assume that the reduced detection of the CVM genes, possibly results from an amplified *retention* at injury site of CSPCs from blood. This interpretation is based on comparison with the results of our PCR assay applied to blunt trauma patients, and from work in our lab with experimental animals (mice with skin wounds, or implanted with retrievable scaffolds as 'cell traps'). CSPCs migrating into circulation from the bone marrow would thus be quickly extracted at the site of injury in order for repair processes to occur. This could occur faster than the marrow produces the cells, resulting in the observed depletion (decreased detection) of the CVM genes in circulation. At the same time, we assume that OLR1 and its neighborhood would increase their detection levels, as the signature of a different

functional network to which this gene serves as a hub, with enhanced transcriptional activity in response to injury (discussed later).

Alternatively, it could have been expected for a particular cell type with functional relevance in healing injuries to appear in the blood shortly after burn injury. Many of the primitive CVM genes, however, were decreased, while OLR1 was increased. This could indicate the majority of the CVM members represent a more primitive cell type, responsible for injury repair. In victims of burn injury, however, a more differentiated cell type, one whose functional relevance is more closely related to OLR1 and its neighborhood, could possibly increase in circulation, representing an "activated" cell type directly responding to injury. The more primitive members of the CVM would still be expected to be expressed in these cell types, but at a lower transcriptional rate.

In conclusion, our data shows that the sensitivity of microarrays, though lower than that of PCR, is still enough to detect some of the genes associated with the CSPCs, including some members of the CVM, in both healthy controls and burn patients. Using these detected genes, we were able to run covariation analyses and obtain valuable information leading to potential candidates for CVM expansion.

Weighted Gene Correlation Network Analysis (WGCNA): In order to identify new associates of CVM, we hoped we could apply a standard method in the field, WGCNA, to obtain relevant PBMCs transcriptional modules, and then to simple verify whether CVM modules were present in any of them. This empirical attempt could have been a shortcut to bring us to the desired result, in a manner recognized by the bioinformatics community. As suggested in Diagram 1, this would have been a "top-down" approach, by considering all genes present on microarrays, and then identifying important modules of interest based on their relation to a particular variable (in this case, presence of a CVM gene within the module). Encountering the sensitivity issues of microarrays in detecting members of the CVM, we selected for WGCNA analysis the children dataset as exhibiting a better representation of CVM members.

Following the recommended steps of the method [16], we began by first isolating the genes with the largest variance within the dataset (8000 in total). However, none of the CVM members remained for analysis when using this arbitrary cutoff. We then expanded the selection to the 12,000 most varying genes, but even so only OLR1 remained in the analysis.

Because our goal was to expand the CVM using *multiple* members as anchoring points, we needed to devise another approach. To this end, we identified the genes which had Pearson correlations > 0.4 with detected members of the CVM (NOTCH4, OLR1, MAP2, NANOG, and NOS3). We then *intersected* these gene associations, looking for genes which were correlated with at least three of the five CVM seed genes. This approach identified 2,755 genes which could be analyzed using the WGCNA method. These genes and their correlation values were used to create Scale Independence and Mean Connectivity graphs, which is the next step in the protocol, enabling appropriate β value selection for adjacency matrix construction. Our data, however, showed a large right-skew in each of these graphs, as compared to constructions seen in the original method paper [16]. These large right-skews led to the selection of β values > 30 in order to reach a threshold of 0.8 for scale free topology model fit, signed R^2, as suggested [25]. As seen by equation 1 (page), such high values for β lead to a loss of information as even large similarity matrix values are converted to nearly 0 for the adjacency matrix. The most likely cause of this large right-skewness is the large interconnectivity of the isolated genes, inherent in their method of isolation.

In one more attempt to use the WGCNA method, we isolated a set of over 300 genes proposed by the NIH's *Stem Cells Interest Group*, to which we added the members of the CVM. This gene panel could serve as an extension of our initial set (Fig. 1), using a larger selection of known stem and progenitor cell primitive and differentiation markers. In addition, genes with a presence score > 0.5 were used in this analysis. In doing so, however, we could not select an appropriate β value to reach a threshold of 0.8 on the Scale Independence graph, as no value for β approached this threshold. Unlike the previous analyses which used over 2,000 genes, this analysis was limited to under 200, likely causing the inability to reach the threshold for exhibition of Scale Free Topology.

Nevertheless, this method offered valuable insights. Often such analyses as WGCNA are used to identify important modules and gene members constituting them. Our study suggests that such "top-down" approaches may not elucidate all information contained within a dataset. In addition, our data shows such an approach is hindered by the lack of sensitivity of microarrays when analyzing genes associated with rare cell types. This is exemplified mainly in the isolation step of the genes to be analyzed, leaving out important genes (i.e. the CVM genes) associated with particular cell types (i.e. CSPCs).

Localization of the CVM in modules created from expression patterns in burn patients: In addition to using the WGCNA method as a "top-down" approach to obtain PBMCs-derived modular transcriptional structures where CVM could possibly be located, we also used our original 'Clique Mining' method for direct transcriptional network reconstruction. We applied it first to burn patients because of its relationship to cardiovascular repair and to the response to injury. Again, based on the NIH's *Stem Cells Interest Group* gene listing of over 300 primitive and differentiation genes, alongside our original panel of genes, we sought to identify modules which could contain members of the CVM.

In this study, healthy controls were compared with early (1-10 days) and late (11-49 days) stage burn victims. We created network representations from the NIH gene listing, only including the primitive and differentiation genes which were reliably detected (Presence Score > 0.5). Although no similar modules occurred between network representations of controls and burn victims (indicating the dramatic changes in blood transcriptional activity after this major injury), one of the new modules remained relatively consistent between the two time frames in patients. Moreover, the strength in connectivity within this module remained high. Also importantly, this module contained the CVM seed gene NOTCH4 in the early stage network representation (not included in late stage, as the detection level went less than 50% presence call). It is important to note that although NOTCH4 was a member of this module, its connectivity to the other members was relatively low, as was its strength. Still, because of its presumed function in response to burn injury and its relation to the CVM seed gene NOTCH4, this module was later analyzed in IPA for its potential functional relevance (discussed below).

Development of an alternative 'bottom-up' method to reconstruct gene modules with CVM members as hub genes:: The demonstrated inability of the standard gene network analysis approaches tested so far to detect modules containing CVM genes led us to conceive a new method serving this purpose: we combined the 'guilt-by-association' (GBA) principle to identify genes which exhibited the greatest co-variation with our CVM seed genes, with the 'clique mining' network visualization tool [17] to obtain the network organization of this gene community . Thus, we created 'neighborhoods', or gene listings of the most highly co-varying genes, for each of the detected CVM genes for multiple studies. By default, this method identifies the CVM gene as being connected to every member of the module, *establishing up*

front the seed gene as a hub of the new module. Neighborhoods were then compared between studies to identify similarities, and within studies to find interconnectivity between these CVM neighborhoods.

Children Study: The first dataset used for this analysis was the children study, because of its characteristics of having the most of the CVR genes present, as well as the greatest detection of the other primitive genes of the panel. The five reliably detected CVM genes, NOTCH4, OLR1, MAP2, NANOG, and NOS3, were used to create 'neighborhoods' of their most highly correlated genes. This was done by running a "many-to-one" correlation analysis for each of the CVM seed genes. After creation of these neighborhoods, a hard threshold of Pearson correlation 0.7 was used to isolate only those genes which exhibited the largest degree of co-variation. Because of this threshold, MAP2 was left out of further analysis as none of its correlates were above the threshold.

We then merged these neighborhoods together to highlight the connectivity that might exist between them (as it would be expected to be interconnected, due to the known interconnectivity of the seed genes) and performed Clique Mining analysis (note: the CVM genes were not correlated with one another above the threshold of 0.7). As exemplified by Figure 4, neighborhoods of NOTCH4, NANOG, and NOS3 were highly interconnected (edge weight = 0.6) with one another with NOTCH4's and NOS3's neighborhoods being *positively* correlated with one another while both were *negatively* correlated with NANOG's neighborhood. OLR1's neighborhood remained its own distinct module. Only when lowering the edge weight threshold (edge weight < 0.5) would OLR1's module share connections with the other neighborhoods. However, the only gene from OLR1's module which did so was LCN2, and its connectivity was extremely limited. In addition, OLR1's neighborhood exhibited the highest level of internal interconnectivity and strength of connectivity.

As a form of pre-validation of the success of the GBA principle, each of the candidates belonging to the four neighborhoods were functionally verified for relation to stemness, differentiation, angiogenesis, neovascularization, cardiovascular disease, and/or cardiovascular repair in literature. Encouragingly, the vast majority of these genes were found to have at least some direct or indirect relevance for these criteria. Because of this result, each gene could be seen as a potential candidate for CVM expansion.

Adult Study: To verify the network representation seen in the children study, the present CVM genes of the adult dataset (OLR1 and NOTCH4) were analyzed in the same manner. Pearson correlations for the highest co-varying genes with OLR1 were lower in the adult study than in children (only one gene, MPO, was above the threshold of 0.7). This is potentially explained by the lower detection levels of the genes associated with OLR1's neighborhood. Because of these lower correlation values, the threshold was lowered to 0.6. Our analysis of OLR1's neighborhood quickly revealed the *majority of the most highly correlated genes in the adult study were the same genes seen in the children study*. We then lowered the threshold value to 0.4 ($p < 0.05$) and identified genes which were highly correlated in both studies (14 genes; only SLPI, CHI3L1, ORM1, ORM2 and ANXA3 did not fit these criterion). We then ran a clique mining analysis on these genes, using the adult expression levels, and found that, although the strength of connectivity was decreased, the overall connectivity of the module remained mostly intact.

NOTCH4's module was mostly inconsistent between the children and adult data sets. Only two genes, TTC7A and ZFAT1, were the same between the two studies (Pearson correlation > 0.7). This later led to their selection for PCR verification. NOTCH4's

neighborhood was selected in a different fashion than OLR1's as it had greater correlation values. We required any gene within the neighborhood to be highly correlated in both children (Pearson correlation > 0.6) and adult (Pearson correlation > 0.7). Thus, only genes which had a relatively high correlation in each study were used, though these were not the *most* correlated genes within each study.

Neighborhoods of OLR1 and NOTCH4 were then merged together and Clique Mining analysis was performed to detect if there were connections between them. Similar to the children study, no connections could be seen at the original threshold (edge weight = 0.6), but when lowered (edge weight = 0.4), connections could be seen. These connections were mainly with the same gene as in the children study, LCN2, again promoting its selection for PCR validation.

Because of the relative consistency of OLR1's neighborhood, we selected OLR1 and its correlated genes in literature to find any significant functional relevance. In doing so, a study on preeclampsia was found [22], a condition characterized by high blood pressure and kidney problems in pregnant women related to deficiencies in CSPCs, which contained the majority of OLR1's neighborhood, in addition to OLR1 itself. This study identified genes exhibiting the largest negative fold change in expression levels between controls and early and late-onset preeclampsia patients, rather than looking at covariation. Thus, in addition to the similarities seen between the children and adult studies, a functional relevance to cardiovascular function could now be more confidently attributed to this OLR1 module. In conclusion, our "bottom-up" analysis shows the additional information that can be obtained by such studies on disease pathologies and their effect on gene expression values.

Burn Study: To include an analysis on a pathology which would elucidate a repairing response, we also applied our "bottom-up" analysis on the burn study dataset. Here, only OLR1 was analyzed, due to its common correlations found between children and adult, as well as the functional relevance for preeclampsia. Rather than performing an initial step of identifying genes which exhibited the largest co-variation with OLR1, we identified first the genes which exhibited the largest *fold change* in expression values between controls and early/late stage burn victims (expression difference > 1, or 10x fold change) to compare results to the preeclampsia study. We pre-selected genes based on their prevalence in the children, adult, and preeclampsia studies. Clique mining analysis was run on genes which exhibited a large fold change in burn victims and were prevalent in either the children/adult studies or in preeclampsia (we also included OLR1 as it was not one of the genes which exhibited a large fold change in burn victims). Although the genes analyzed were not isolated by covariation methods, the modular structure of OLR1's original neighborhood, as well as new members added to it from the preeclampsia study, was characterized by a high level of interconnectivity and strength of connectivity.

Even more interesting, rather than exhibiting a *negative* fold change in expression level, as was seen in preeclampsia, a condition in which there is a defect in repair mechanisms, this pre-selected set of genes exhibited the largest *positive* fold change in burn victims, a condition in which the repairing mechanisms would be "activated", fitting our hypothesis of OLR1 and its neighborhood as having a maintenance and repairing function in the cardiovascular system. In addition, the connectivity within OLR1's network seems to be increased in burn victims as compared to normal, healthy controls, suggesting alterations in the interactions of these genes in response to injury.

Functional Attributions to Identified Modules via IPA: OLR1's consistent neighborhood was imported into Ingenuity Pathway Analysis (IPA) to check for known

functional significance and associated signaling pathways attributed to this module. Among the functions related to genes contained within the module, 'cardiovascular system development' and function' was among the highest. In addition, noted 'upstream regulators' of this module were lipopolysaccharide and CEBPA, a transcription factor known to be associated with the differentiation of hematopoietic cell types [26].

NOTCH4's module from the original burn study on the NIH's *Stem Cells Interest Group* gene listing was also IPA-tested, due to its seemingly functional relevance in response to burn injury. Similar to OLR1, functional relations to 'cardiovascular system development and function' were high, and among upstream regulators of this module were again lipopolysaccharide and CEBPA. Thus, not only is there now covariation in and functional significance of OLR1's network, but there is an additional relevance in control of its expression via a common transcription factor, CEBPA, with another member of the CVM and genes correlated with it. In addition, this attributes to the success of our 'bottom-up' approach to elucidate modules of interest by combining GBA with Clique Mining.

Gene Selection for PCR Validation: The various methods described above led to the selection of genes to validate via PCR their membership in the CVM. To this end, from OLR1's network, we selected CEACAM6, MMP8, LTF, and LCN2[4]. Each of these genes were prevalent in all four studied conditions of OLR1 and were among its most highly correlated genes in each study and exhibited some of the greatest fold changes in both burn and preeclampsia. Of final note, each of these genes went through a pre-selection process in which they had to show known functional relevance to a cardiovascular disease condition and/or method of cardiovascular repair, whether it be direct or indirect. Although the majority of OLR1's module members fit these criteria, we decided to only select the above four so additional neighborhoods could be selected from. CEACAM6 is a gene associated with myelofibrotic transformation, release of hematopoietic cells into circulation, and neovascularization [27]. MMP8 is a secreted extracellular degradation protein which is associated with atherosclerotic plaque formation and angiogenesis [28]. LTF is a gene known to upregulate the expression of KDR, one of the original CVM genes, and induce endothelial cell proliferation and migration [29]. Finally, LCN2, the only gene from OLR1's neighborhood which exhibited connectivity to other neighborhoods, is a gene associated with the development of multiple cardiovascular diseases [30].

From NOTCH4, we selected TTC7A and ZFAT1. These were genes which exhibited large Pearson correlations (> 0.7) in both the children and adult studies. Little is known about TTC7A [31], so this gene is more speculative, but verification may lead to insight into its role with cardiovascular maintenance. ZFAT1 is a gene with suggested roles in hematopoiesis and angiogenesis [32]. We consider that these genes play more primitive functions in CSPCs as the seed gene they relate to, NOTCH4, is a very primitive gene, especially in comparison to the other seed genes being analyzed.

From NOS3, we selected LDLRAP1, EPHX2, and CAMK4. These genes were solely selected from the results obtained from the children's study and were the three most highly correlated genes to NOS3 (Pearson correlations > 0.7). In addition, these genes were among the most highly interconnected genes between the three neighborhoods of NOTCH4, NOS3, and NANOG. LDLRAP1 is a gene highly associated with hypercholesterolemia [33], a condition in which high levels of cholesterol are found in the blood, typically leading to atherosclerosis. EPHX2 plays a pivotal role in neovascularization and tissue repair, in part by the activation of

endothelial nitric oxide synthase (NOS3) [34]. CAMK4 is another gene shown to play a role in nitric oxide synthase activity, thereby regulating blood pressure [35].

From NANOG's neighborhood, we selected MKL1 (only one gene was selected from this neighborhood, due to limited space on the PCR arrays). MKL1 has been shown to control myofibroblast activation and fibrosis in response to the rennin-angiotensin system {Small, 2010 2865 /id}. In addition, it plays a role in vascular remodeling after myocardial infarction. Finally, after selecting the possible CVM extension candidates, a validation procedure will be performed using the same PCR methods as were used in the previous study.

Conclusions

Upon our first attempts to identify whether or not the original panel of primitive and differentiation markers was detected on microarrays, we found an extreme lack of sensitivity to genes associated with the rare CSPC population we were attempting to analyze. Based on our original PCR study, these genes were well detected in the same cell population (i.e. PBMCs), establishing the fact that PCR is much more sensitive, leading to a heavy loss of information when using microarrays for similar analysis. Although this may be the case, valuable information may still be obtained from microarrays, though this information may only be the "tip of the iceberg", only seeing information which the microarrays can detect. In this particular case, we were able to identify several genes which qualified as potential candidates to our original Cardiovascular Module (CVM) based on their correlations to the original CVM genes, functional relevance, relations to cardiovascular diseases and conditions, and genetic control mechanisms (i.e. common transcription factors).

These results suggest our method of the 'guilt-by-association' (GBA) principle-based bottom-up network reconstruction approach can be applied in order to reconstruct a transcriptional network on gene microarrays. In doing so, we identified new candidate genes as members of a pre-existing gene module containing rare primitive and differentiation markers. Thus, our method allowed for the expansion of the original CVM to contain additional genes identified via microarray analysis. These candidates were selected from 'neighborhoods' of the most highly covarying genes to the original CVM genes which were detected on microarrays (i.e. NANOG, NOS3, NOTCH3, and OLR1).

In addition, this method identified a cluster of genes rich in cardiovascular implicated genes (Fig. 8) expressed in PBMCs as co-variants of OLR1 (oxidized LDL receptor, LOX-1). This cluster recurrently occurs on Affymetrix microarrays from normal children and adults, in response to burn injury, and is among the most highly modified genes in preeclampsia. Alongside the spatial correlations seen (i.e. correlations seen among normal, healthy controls, between individuals), this cluster follows a coordinated pattern of gene expression following response to burn and trauma injury, increasing in expression immediately after injury, achieving its highest level of expression between days 5 and 10, and maintaining an above normal expression level over 100 days post-injury. Thus, OLR1 and its cluster of highly correlated genes also maintain temporal correlations in certain pathological conditions, suggesting they may play a direct role in the response to injury.

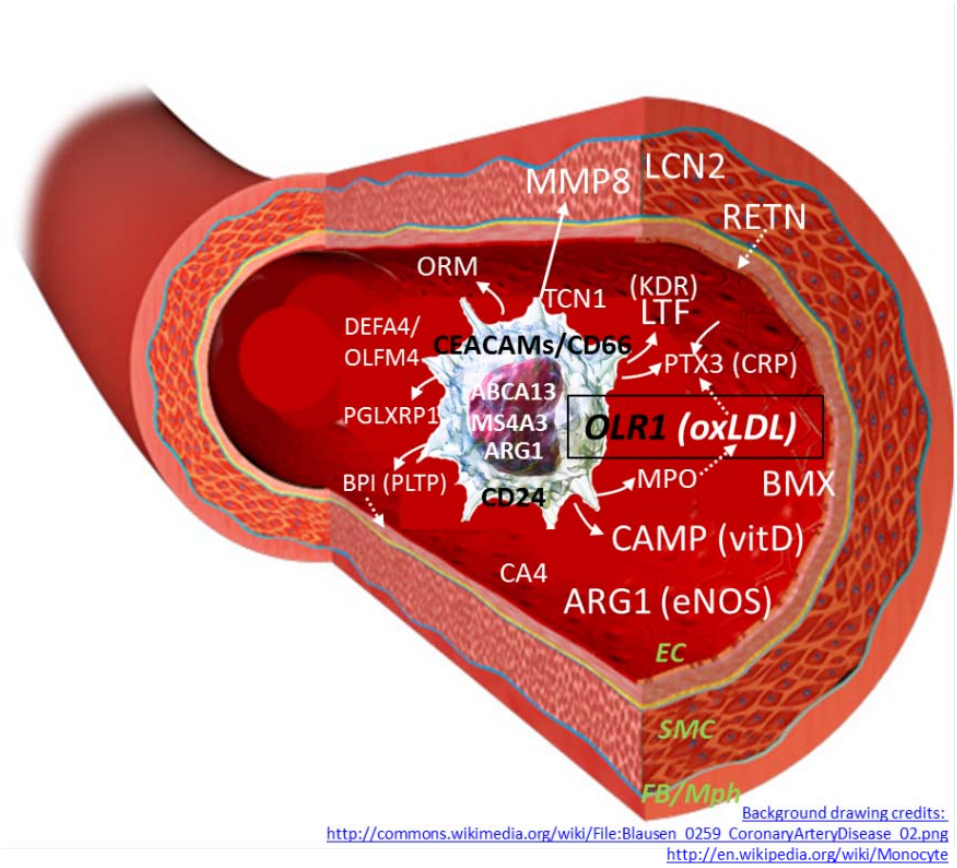

Figure 8: Cardiovascular roles of OLR1 module members. The meaning of these genes is given in Table 2. Projection of their functions over the vascular wall structure, as seen interacting with a generic mononuclear progenitor, suggests their common involvement in vascular protection and/or atherosclerotic plaque formation. Arrows indicate secreted proteins; dashed arrows indicate direct influences; in parenthesis are highlighted indirect relationships. EC=endothelial cells; SMC=smooth muscle cells; FB=fibroblasts; Mph=macrophages.

In contradiction with results seen with OLR1 and its cluster, a coordinate reduction, or going undetectable, of other gene clusters (i.e. the original panel of genes) can be seen during physiological or pathological perturbations (i.e. burn injury or cardiovascular disease). This illustrates a complex response of the transcriptional landscape in peripheral blood to injury and repair of the vascular system. This complex relationship likely stems from the release of stem and progenitor cells from the bone marrow into circulation, followed by the recruitment of these same cells to the site of injury. In addition, these cells would likely experience modulation of their gene expression based on the conditions in which they are surrounded. Thus, they may be 'activated' towards a line of differentiation which plays a more prevalent role in the repair and maintenance of the cardiovascular system (i.e. OLR1 and its cluster of genes), giving a proposed reasoning as to why OLR1 and its cluster increase in expression while other primitive and differentiation markers decrease in expression.

Finally, our method of the clique mining approach identified potential transcription factors which may play a role in the putative control of gene modules which co-varied with our original CVM genes. NOTCH4's related module detected in the original burn study, as well as OLR1's main cluster of highly co-varying genes both had upstream regulators of LPS

(lipopolysaccharide) and the transcription factor CEBPA. This finding retroactively brings credit to the hypothesis that co-expression is derived from functional linking, thereby validating the GBA principle. In addition, this method identified a transcription factor which may play a yet unknown role in the response to injury and cardiovascular perturbations.

With these results, a secondary panel of 10 genes has been selected for PCR validation. These 10 genes are the following: low density lipoprotein receptor adaptor protein 1 (LDLRAP1), calcium/calmodulin kinase 4 (CAMK4), and epoxy hydrolase 2 (EPHX2) – selected because of their correlations with NOS3; megakaryoblastic leukemia (translocation) 1 (MKL1) – selected because of its correlation with NANOG; tetratricopeptide repeat domain 7A (TTC7A) and ZFAT zinc finger 1 (ZFAT1) – selected because of their correlations with NOTCH4; and matrix metallopeptidase 8 (MMP8), lactotransferrin (LTF), lipocalin 2 (LCN2), and carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6) – selected because of their correlations with OLR1 and prevalence in burn/trauma and preeclampsia. These genes were among the most highly co-varying genes to the original CVM members and all had a known functional role with cardiovascular maintenance/repair and/or cardiovascular disease.

These selected genes will be validated for their correlations with the entire panel of CVM members, as well as the original cardiovascular health markers used (i.e. age, blood pressure, BMI, and vascular stiffness). Those which fit the initial criteria will be added to the CVM, for future analysis of cardiovascular conditions using our original PCR method.

Acknowledgements:

This research was funded by the Arts and Sciences Honors Research Scholarship. Additional aid for the project was provided by Dr. Leni Moldovan, Dr. Yang Xiang, Ryan O'Neill, and Andrew Yates. Judges for the thesis were Dr. David Stetson, Dr. Kun Huang, and Dr. Rene Anand. Special thanks to Dr. Anand for providing additional resources to enhance the efficacy of the paper.

References

1 Asahara T, Murohara T, Sullivan A, Silver M, van der Zee R et al. (1997) Isolation of putative progenitor endothelial cells for angiogenesis. Science 275: 964-967.

2 de JS, Boks MP, Fuller TF, Strengman E, Janson E et al. (2012) A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. PLoS One 7: e39498.

3 Klomp JA, Furge KA (2012) Genome-wide matching of genes to cellular roles using guilt-by-association models derived from single sample analysis. BMC Res Notes 5: 370.

4 Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

5 Chaudhury H, Goldie LC, Hirschi KK (2011) Vascular precursor cells. Genes Cancer 2: 1081-1084.

6 Kemppinen AK, Kaprio J, Palotie A, Saarela J (2011) Systematic review of genome-wide expression studies in multiple sclerosis. BMJ Open 1: e000053.

7 Masud R, Shameer K, Dhar A, Ding K, Kullo IJ (2012) Gene expression profiling of peripheral blood mononuclear cells in the setting of peripheral arterial disease. J Clin Bioinforma 2: 6.

8 Radom-Aizik S, Zaldivar F, Jr., Leu SY, Cooper DM (2009) Brief bout of exercise alters gene expression in peripheral blood mononuclear cells of early- and late-pubertal males. Pediatr Res 65: 447-452.

9 Herman NM, Grill DE, Anderson PJ, Miller AD, Johnson JB et al (2014) Peripheral blood mononuclear cell (PBMC) gene expression in healthy adults rapidly transported to high altitude.

10 Holland SM, DeLeo FR, Elloumi HZ, Hsu AP, Uzel G et al. (2007) STAT3 mutations in the hyper-IgE syndrome. N Engl J Med 357: 1608-1619.

11 Dusek JA, Otu HH, Wohlhueter AL, Bhasin M, Zerbini LF et al. (2008) Genomic counter-stress changes induced by the relaxation response. PLoS One 3: e2576.

12 Zhou B, Xu W, Herndon D, Tompkins R, Davis R et al. (2010) Analysis of factorial time-course microarrays with application to a clinical study of burn injury. Proc Natl Acad Sci U S A 107: 9923-9928.

13 Li L, Li M, Sun C, Francisco L, Chakraborty S et al. (2011) Altered hematopoietic cell gene expression precedes development of therapy-related myelodysplasia/acute myeloid leukemia and identifies patients at risk. Cancer Cell 20: 591-605.

14 Barnes MG, Grom AA, Thompson SD, Griffin TA, Pavlidis P et al. (2009) Subtype-specific peripheral blood gene expression profiles in recent-onset juvenile idiopathic arthritis. Arthritis Rheum 60: 2102-2112.

15 Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH et al. (2012) A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics 100: 337-344.

16 Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4: Article17.

17 Xiang Y, Fuhry D, Kaya K, Jin R, Catalyurek UV et al. (2012) Merging network patterns: a general framework to summarize biomedical network data. Netw Model Anal Health Inform Bioinforma 1: 103-116.

18 Bron C (1973) Algorithm 457: finding all cliques of an undirected graph. Commun ACM 16: 575-577.

19 Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media .

20 Gavrilin MA, Bouakl IJ, Knatz NL, Duncan MD, Hall MW et al. (2006) Internalization and phagosome escape required for Francisella to induce human monocyte IL-1beta processing and release. Proc Natl Acad Sci U S A 103: 141-146.

21 Etienne W, Meyer MH, Peppers J, Meyer RA, Jr. (2004) Comparison of mRNA gene expression by RT-PCR and DNA microarray. Biotechniques 36: 618-6.

22 Chaiworapongsa T, Romero R, Whitten A, Tarca AL, Bhatti G et al. (2013) Differences and similarities in the transcriptional profile of peripheral whole blood in early and late-onset preeclampsia: insights into the molecular basis of the phenotype of preeclampsiaa. J Perinat Med 41: 485-504.

23 Timmermans F, Plum J, Yoder MC, Ingram DA, Vandekerckhove B et al. (2009) Endothelial progenitor cells: identity defined? J Cell Mol Med 13: 87-102.

24 Richardson MR, Yoder MC (2011) Endothelial progenitor cells: quo vadis? J Mol Cell Cardiol 50: 266-272.

25 Mason MJ, Fan G, Plath K, Zhou Q, Horvath S (2009) Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics 10: 327.

26 Rosmarin AG, Yang Z, Resendes KK (2005) Transcriptional regulation in myelopoiesis: Hematopoietic fate choice, myeloid differentiation, and leukemogenesis. Exp Hematol 33: 131-143.

27 Hasselbalch HC, Skov V, Larsen TS, Thomassen M, Riley CH et al. (2011) High expression of carcinoembryonic antigen-related cell adhesion molecule (CEACAM) 6 and 8 in primary myelofibrosis. Leuk Res 35: 1330-1334.

28 Pirila E, Korpi JT, Korkiamaki T, Jahkola T, Gutierrez-Fernandez A et al. (2007) Collagenase-2 (MMP-8) and matrilysin-2 (MMP-26) expression in human wounds of different etiologies. Wound Repair Regen 15: 47-57.

29 Kim CW, Son KN, Choi SY, Kim J (2006) Human lactoferrin upregulates expression of KDR/Flk-1 and stimulates VEGF-A-mediated endothelial cell proliferation and migration. FEBS Lett 580: 4332-4336.

30 Furuya F, Shimura H, Yokomichi H, Takahashi K, Akiyama D et al. (2013) Neutrophil gelatinase-associated lipocalin levels associated with cardiovascular disease in chronic kidney disease patients. Clin Exp Nephrol .

31 Avitzur Y, Guo C, Mastropaolo LA, Bahrami E, Chen H et al. (2014) Mutations in Tetratricopeptide Repeat Domain 7A Result in a Severe Form of Very Early Onset Inflammatory Bowel Disease. Gastroenterology 146: 1028-1039.

32 Tsunoda T, Shirasawa S (2013) Roles of ZFAT in haematopoiesis, angiogenesis and cancer development. Anticancer Res 33: 2833-2837.

33 Soutar AK, Naoumova RP (2004) Autosomal recessive hypercholesterolemia. Semin Vasc Med 4: 241-248.

34 Hou HH, Hammock BD, Su KH, Morisseau C, Kou YR et al. (2012) N-terminal domain of soluble epoxide hydrolase negatively regulates the VEGF-mediated activation of endothelial nitric oxide synthase. Cardiovasc Res 93: 120-129.

35 Santulli G, Cipolletta E, Sorriento D, Del GC, Anastasio A et al. (2012) CaMK4 Gene Deletion Induces Hypertension. J Am Heart Assoc 1: e001081.