# Development of a Transcriptional Regulation Bioinformatics Pipeline to Predict Co-Regulated Genes in Vascular Smooth Muscle Cell Phenotypic Transitions During Atherosclerosis


***Mahima Reddy*[1], Vlad Serbulea[1], Sohel Shamsuzzaman[1], Anita Salamon[1], Rupa Tripathi[1], Clint Miller[2] , Giuseppe Mocci[3], Johan Björkegren[4], Gary Owens[1]**

[1]Robert M. Berne Cardiovascular Research Center, Division of Cardiovascular Medicine, University of Virginia School of Medicine; [2]Center for Public Health Genomics, University of Virginia School of Medicine; [3]Department of Medicine Huddinge, Karolinska Institutet; [4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

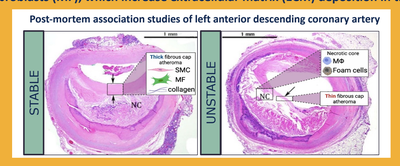


## Abstract

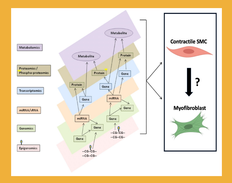
The rupture of vulnerable or advanced atherosclerotic lesions contributes to myocardial infarctions and strokes, the leading causes of death globally. During all stages of atherosclerosis, vascular smooth muscle cells (VSMCs) are recruited as a major source of plaque cells within lesions, where they undergo phenotypic modulation and give rise to both plaque-stabilizing and destabilizing phenotypes. Because most atherosclerosis studies test the functional role of single transcription factors, genes, and proteins in VSMCs, many aspects of how VSMC plasticity can be clinically exploited to prevent the rupture of vulnerable atherosclerotic lesions remain unclear. Systems biology approaches have emerged as a promising avenue for characterizing cellular dynamics at the organismal level, but they remain under-utilized in the field of cardiovascular medicine. This study proposes a novel systems biology pipeline that integrates experimental and computational tools with large datasets measuring the transcriptomes and epigenomes of phenotypically transitioning VSMCs. Our pipeline allows users to identify candidate transcription factors that co-regulate mouse or human gene sets controlling specific VSMC phenotypic transitions during atherosclerosis. The computational algorithms developed in this workflow integrate an array of publicly available databases, such as the Ensembl genome database and the JASPAR transcription factor database. To demonstrate the translational potential of our pipeline, we used it to test the hypothesis that the murine mesenchymal marker stem cell antigen-1 *Sca1* (*Ly6a*) has a human ortholog, which has not been previously validated as a target in clinical therapies. To test the hypothesis that there are sets of genes co-regulated with *Ly6a* in human VSMCs which, when expressed in concert, promote a distinct phenotypic state marked by *Ly6a* in mice, we integrated genomic and ATAC-seq analyses of VSMCs to identify murine genes with human orthologs that bind *Ly6a*-specific transcription factors. Specifically, we predict that the transcription factors FOS, EGR1, JUN, JunB and Atf3 co-regulate sets of human genes with *Ly6a*. While prior studies have associated these transcription factors with VSMC proliferation and phenotypic modulation, we specifically predict these *Ly6a*-specific transcription factor co-regulate the following human gene sets: *TNS1*, *THSD4*, *RCAN2*, *PTPRD*, *CPNE4*, *COL15A1*, and *ABLIM1* (FOS and Atf3). To validate these predictions, we will determine if these transcription factors and their co-regulated genes are expressed in single cell RNA-sequencing analyses of human carotid arteries and can be detected with immunofluorescence staining of human coronary artery sections. We will also use siRNA to target these candidate transcription factors and genes in cultured VSMCs. The results of our pipeline will help identify novel transcriptional regulatory networks in VSMCs that can be therapeutically targeted for the treatment of advanced atherosclerosis.

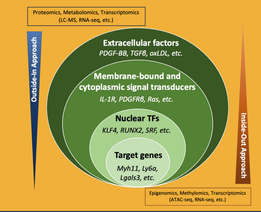

## Introduction

- Rupture or erosion of unstable atherosclerotic plaques is the underlying cause of myocardial infarction and stroke-related mortality, the leading causes of death worldwide.[1]
- A stable plaque has a low ratio of macrophages (MΦ) to vascular smooth muscle cells (VSMCs) and myofibroblasts (MF), which increase extracellular matrix (ECM) deposition in the fibrous cap.[2]




- VSMCs de-differentiate during atherosclerosis to take on both plaque stabilizing phenotypes (collagen producing myofibroblast-like cells) and plaque destabilizing phenotypes (i.e. macrophage-like cells).[3]
- Over the decade, many mechanisms have been identified in mice that play roles in VSMC-derived plaque stability, but we do not understand the complete regulatory states of VSMCs.
- A lack of this understanding limits the ability of identified mechanisms to be directly exploited clinically to improve plaque stability in humans.




## Hypothesis

**Conserved cis/trans elements coordinately regulate sets of genes that confer specific cellular (VSMC) phenotypes.**

## Approach

### *Strategy for prediction of transcription factors (TFs) co-regulating gene sets*

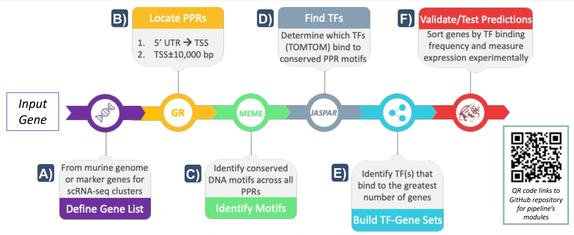
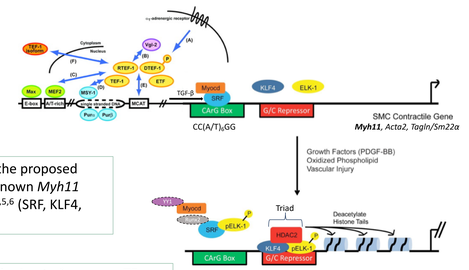



**Figure 1:** The proposed pipeline consists of 6 modules, which retrieve data from publicly available bioinformatics databases and are uniquely integrated in algorithms written in R. **A)** In the first step, the user selects an input gene(s) for which they would like to predict candidate regulatory transcription factors (TFs) and co-regulated genes. While genes co-regulated with the input gene can be predicted from an organism's whole genome (default gene list), the user can limit the analysis to a smaller gene list (i.e. marker genes for a transcriptomic cluster). **B)** To identify the pool of candidate TFs, the user defines the parameters for the proximal promoter region (PPR), which is a certain number of nucleotide base pairs (i.e. 10,000 base pairs) upstream and downstream from the input gene's transcriptional start site (TSS). The R package GenomicRanges is used to convert the PPR parameters, in the form chromosomal coordinates, to a string of DNA nucleotides. **C)** Known TF motifs (exact binding sites) are identified in the PPRs of the input gene and gene list and are then sorted by frequency of occurrence. **D)** TFs known to bind the most frequently occurring motifs are identified from the JASPAR transcription factor database. **E)** Sets of candidate transcription factors and their co-regulated genes (including the input gene) are characterized by a high percentage of shared known transcription factor binding sites. **F)** These initial predictions can be validated experimentally by using ATAC-seq to assess TFs' ability to bind accessible chromatin and by measuring genes' mRNA expression with RNA-seq.

## *Proof of Concept*



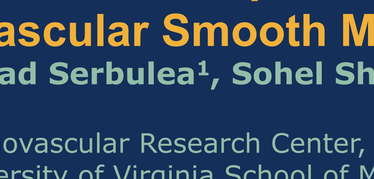

**Goal**: Determine if the proposed pipeline identifies known *Myh11* gene co-regulators[4,5,6] (SRF, KLF4, ELK-1, MEF2B, etc.)

**Hypothesis**: A set of common TFs co-regulate *Myh11* and other contractile SMC genes[7] (*Acta2*, *Tagln, and Cnn1*)

### Step 1: Define input gene of interest (*Myh11*)


Because the regulatory mechanisms controlling the expression of the smooth muscle contractile marker Myosin Heavy Chain 11 (*Myh11*) have been extensively studied in mice, murine *Myh11* was chosen as the input gene of interest (positive control) to test whether the proposed pipeline identifies transcription factors (TEF-1, Max, MEF2, MSY1, SRF, KLF4, ELK-1, etc.) that are known to co-regulate *Myh11* and other contractile genes like *Acta2* and *Tagln*.

### Step 2: Identify all TFs that bind the PPR of Myh11

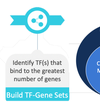



| Chrom # | Chrom Start | Chrom End | MatrixID | score | strand | TF Name | TF binding sequence |
|---|---|---|---|---|---|---|---|
| chr16 | 14289628 | 14289644 | MA0083.3 | 470 | - | SRF | ACTACCATACAAGGGCA |
| chr16 | 14289628 | 14289644 | MA0083.3 | 523 | + | SRF | ACTACCATACAAGGGCA |
| chr16 | 14291161 | 14291177 | MA0083.3 | 434 | + | SRF | TTGTCCTTATAAGAGAT |
| chr16 | 14281533 | 14281542 | MA0511.2 | 265 | + | RUNX2 | ACTCCCACA |
| chr16 | 14281622 | 14281634 | MA0039.4 | 311 | + | KLF4 | AGTCCACACACAT |

**Figure 2:** JASPAR identified 30,968 unique TF binding sequences in the PPR (Chr16: 14281365-14301365) of *Myh11*. 769 unique TFs were identified to bind at least one of these binding sequences, yielding 75,532 total combinations of *Myh11*-specific TFs and binding sequences.

Amongst the 769 candidate TFs, known *Myh11* transcriptional regulators, such as SRF and RUNX2, were identified as having the ability to bind the *Myh11* PPR.

### Step 3: Identify all mouse genes with PPRs that can bind Myh11-specific TFs




A)

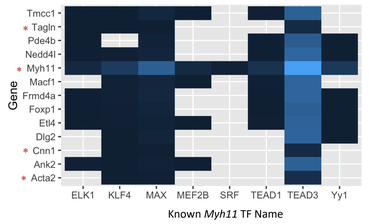
| TF Name | TF binding sequence | Mouse genes with PPRs that have Myh11 TF sequence |
|---|---|---|
| SRF | TTGTCCTTATAAGAGAT | **Myh11**, Efcab12, Oas1e |
| KLF4 | AGTCCACACACAT | Btnl1, Gm10486, Gm16405, Gm10488, Gm10230, Slco1a5, Casq1, Dpp6, Gk2, Mdfic, **Myh11**, Pea15a, Pecam1, Dusp1, Onecut2, St8sia6, Milr1, C1qtnf12, Ergic1, C1qtnf2, Slxl1, Cnnm4 |

**Figure 3:** The 20 kb DNA nucleotide sequences for the PPRs of all genes in the mouse genome were identified in the pipeline's second module. **A)** Murine genes able to bind *Myh11*-specific TFs (i.e. SRF or KLF4) were determined by an exact nucleotide sequence match between the TF's binding sequence and the genes' PPR sequences. **B)** To identify candidate genes co-regulated with *Myh11*, all murine genes were ranked by the percentage of *Myh11*-specific TFs able to bind their PPRs. Here, the top ten genes whose PPRs share the greatest number of *Myh11*-specific TFs and binding sites are shown. Despite containing <10% of all possible *Myh11*-specific TF binding sites in their PPRs, previously validated contractile genes (*Acta2*, *Cnn1*, *Tagln*) were also identified as sharing *Myh11* regulatory TFs. **C)** This heat map shows the diversity in the number of binding sites for *Myh11*-specific TFs in both the novel and previously validated contractile genes' PPRs*.

B)

| mgi_symbol | Frequency | % of Myh11 TFs that bind mouse gene PPR | Number of unique Myh11 TFs |
|---|---|---|---|
| Myh11 | 75532 | 100.00 | 769 |
| **Macf1** | 24892 | 32.96 | 546 |
| **Frmd4a** | 23589 | 31.23 | 489 |
| **Tmcc1** | 23452 | 31.05 | 530 |
| **Ank2** | 22632 | 29.96 | 498 |
| **Et4** | 22351 | 29.59 | 498 |
| **Foxp1** | 22315 | 29.54 | 489 |
| **Foxp1** | 22315 | 29.54 | 489 |
| **Nedd4l** | 22213 | 29.41 | 501 |
| ... | .... | ... | .... |
| Acta2 | 7263 | 9.62 | 272 |
| Cnn1 | 7098 | 9.40 | 347 |
| Tagln | 4645 | 6.15 | 268 |

C)




## *Proof of Concept*

### Step 4: Validate *Myh11*-associated TFs that co-regulate genes with *Myh11*

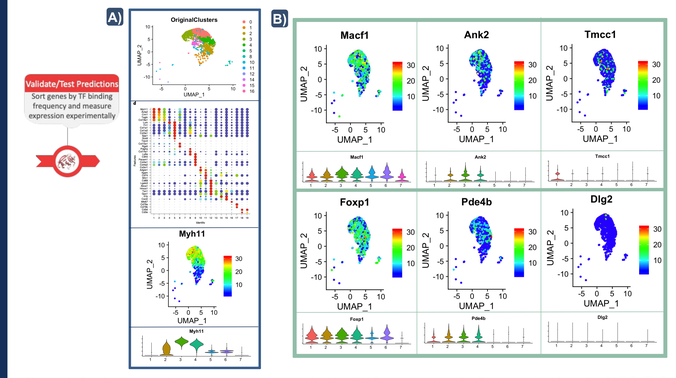
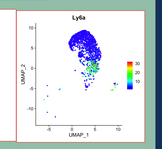



**Figure 4:** Candidate genes' co-expression is depicted in this UMAP presentation of scRNA-seq data from *Pdgfrb*$^{SMC\text{-}WT/WT}$ mice.[8] **A)** scRNA-seq analysis of advanced BCA lesions from *Pdgfrb*$^{SMC\text{-}WT/WT}$ identifies 19 distinct cell clusters. Clusters 1-7 are *eyfp* positive (SMC-derived). UMAP representation of *Myh11* expression is seen in clusters 1-2. **B)** UMAP representation of the expression of 6 candidate genes (*Macf1*, *Foxp1*, *Ank2*, *Pde4b*, *Tmcc1*, *Dlg2*) correlate with *Myh11* expression, supporting these genes' co-regulation with *Myh11*. At the bottom of each UMAP, violin plots show the corresponding gene's expression in SMC-clusters (clusters 1-7)

## *Experimental Application*

**Problem**: A subset of differentiated *Myh11*+ medial VSMCs (Figure 4a; UMAP on the right) express the murine mesenchymal marker stem cell antigen-1 *Sca1* (*Ly6a*). Within murine atherosclerotic lesions, *Ly6a*+ SMC-derived cells participate in injury-repair processes, such as migration, proliferation, and the phenotypic transition to plaque-stabilizing (ECM-producing) myofibroblast-like cells (unpublished).

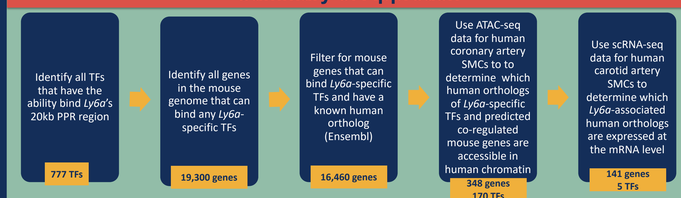
- **Unfortunately, murine *Ly6a* does not have a known human ortholog, which can be tested as a target in human clinical therapies.**




**Hypothesis**: There are sets of genes co-regulated with *Ly6a* in human VSMCs which, when expressed in concert, promote a distinct phenotypic state marked by *Ly6a* in mice.

**Goal**: Use the TRBP to identify human genes whose mouse orthologs are co-regulated by TFs that are able to modulate murine *Ly6a* expression.

### Summary of Approach




### Validation of human genes predicted to be co-regulated with *Ly6a* and its associated TFs

A)

| mgi_symbol | Human Homolog | Frequency | % of Ly6a TFs that bind mouse gene PPR | HCA scRNA-seq_cluster |
|---|---|---|---|---|
| Ptprd | PTPRD | 15752 | 22.23 | SMC8 |
| Sorbs2 | SORBS2 | 14807 | 20.90 | SMC1 |
| Mecom | MECOM | 14243 | 20.10 | SMC8 |
| Ablim1 | ABLIM1 | 14115 | 19.92 | SMC9 |
| Sorbs1 | SORBS1 | 12788 | 18.05 | SMC1 |
| Tsc22d1 | TSC22D1 | 12340 | 17.41 | SMC3 |
| Mbnl1 | MBNL1 | 11404 | 16.09 | SMC8 |
| | | | | SMC1, |
| Lpp | LPP | 11323 | 15.98 | SMC2 |
| Rcan2 | RCAN2 | 11310 | 15.96 | SMC2 |
| Tns1 | TNS1 | 11125 | 15.70 | SMC1 |
| Cpne4 | CPNE4 | 10937 | 15.43 | SMC8 |
| Sulf1 | SULF1 | 10916 | 15.40 | SMC5 |
| Thsd4 | THSD4 | 10071 | 14.21 | SMC8 |
| Zfhx3 | ZFHX3 | 9940 | 14.03 | SMC8 |
| Gpm6b | GPM6B | 9422 | 13.30 | SMC6 |
| | | | | SMC3, |
| Col15a1 | COL15A1 | 9410 | 13.28 | SMC5 |
| Enpp2 | ENPP2 | 9252 | 13.06 | SMC8 |
| Actn4 | ACTN4 | 9248 | 13.05 | SMC2 |
| Fgf14 | FGF14 | 9188 | 12.97 | SMC3 |
| | | | | SMC1, |
| Ppp1r12b | PPP1R12B | 9056 | 12.78 | SMC2 |

B)

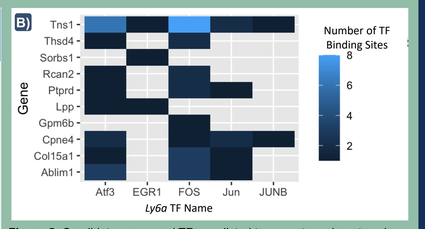


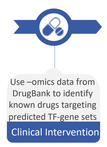

**Figure 5:** Candidate genes and TFs predicted to promote a phenotype in human SMCs similar to Ly6a in murine SMCs are shown. **A)** The top 20 (out of 141) mouse genes that share the greatest percent of *Ly6a* TF binding similarity and that have human orthologs with accessible chromatin and mRNA expression (last column) in human SMCs are presented in this table. **B)** This heat map shows the top candidate mouse genes (with known human orthologs) that bind the five *Ly6a*-specific TFs (Atf3, EGR1, FOS, Jun, and JUNB) validated in human SMCs.

## Summary

1. **The proposed pipeline (TRBP) identifies both known genes (*Acta2*, *Tagln*, and *Cnn1*) and novel genes (*Macf1*, *Ank2*, *Foxp1*, and *Pde4b*) that are co-regulated with the contractile SMC marker *Myh11* and its associated TFs (TEAD1/TEF-1, TEAD3, KLF4, MAX, and Yy1).**
   - **Novel genes *Macf1* and *Ank2* are associated with the actin cytoskeleton, supporting their co-expression with *Myh11*.**
2. **The murine *Ly6a*-specific TFs FOS and Atf3 are predicted to co-regulate *Tns1*, *Thsd4*, *Rcan2*, *Ptprd*, *Cpne4*, *Col15a1*, and *Ablim1*, which have known orthologs that are expressed in human SMCs.**
3. **A complex network of TFs co-regulate multiple genes conferring VSMC contractile and plastic phenotypes.**

## Future Directions

- **Approach 1:** Knock down predicted *Ly6a*-associated TFs and co-regulated genes in murine and human SMCs
  - Simultaneously treat cultured eyfp+ murine SMCs, in addition to HCASMCs, with multiple siRNAs targeting Atf3, FOS, *Tns1*, *Thsd4*, *Rcan2*, *Ptprd*, etc. in *Ly6a*-stimulating conditions (i.e. PDGF-BB supplementation)
  - Current Expectation: Reduction in proliferative/migratory capacities in both human and murine SMCs, in addition to reduced *Ly6a* gene expression in murine SMCs
- **Approach 2:** Identify clinically approved drugs that target predicted co-regulatory TF-gene sets
  - Integrate DrugBank pharmaco-transcriptomics and proteomics data as a module in the current TRBP
    - *Couple "outside-in" approach with current "inside-out" pipeline*
  - Administer candidate drug(s) to cultured SMCs (and eventually Myh11-ERT2$^{Cre}$ eYFP ApoE$^{-/-}$ mice), measuring ability of SMCs to participate in injury-repair processes during atherosclerosis
  - Current Expectation: Loss of *Ly6a*+ phenotype in SMCs after treatment with a known drug (i.e. valproic acid)




- **Approach 3:** Account for nucleotide sequence variability in TF binding sites
  - Compute a "conservation score" that accounts for the frequencies of nucleotides in TF binding sequences (instead of using exact PPR-TF binding sequence matches)
  - Current Expectation: More TFs are predicted to co-regulate common gene sets


## Acknowledgements

***Lab Management:*** *Rebecca Deaton, Laura Shankman*

***Computational Support:*** *UVA Research Computing (Rivanna)*


## Contact


Mahima Reddy msr3nf@virginia.edu

twitter.com/MahimaReddy11

linkedin.com/in/mahima-reddy


## References


1. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019.
2. Tavora, F., Cresswell, N., Li, L., Fowler, D. & Burke, A. Frequency of Acute Plaque Ruptures and Thin Cap Atheromas at Sites of Maximal Stenosis. *Arq Bras Cardiol*. 2010.
3. Shankman L., Gomez D, Cherepanova O.A., Salmon M., Alencar G.F., Haskins R.M., Swiatlowska P., Newman A.A.C., Greene E.S., Straub A.C., Isakson B.E., Randolph G.J., Owens G.K. KLF4-dependent phenotypic modulation of smooth muscle cells has a key role in atherosclerotic plaque pathogenesis. *Nature Medicine*. 2015
4. Yoshida, T., 2008. MCAT elements and the TEF-1 family of transcription factors in muscle development and disease. *Arterioscler Thromb Vasc Biol*. 2008.
5. Katoh, Y., Molkentin, J.D., Dave, V., Olson, E.N. and Periasamy, M., MEF2B is a component of a smooth muscle-specific complex that binds an A/T-rich element important for smooth muscle myosin heavy chain gene expression. *Journal of Biological Chemistry*. 1998.
6. Huang, J. and Parmacek, M.S.. Modulation of smooth muscle cell phenotype: The other side of the story. *Circulation research*. 2012.
7. Sun, Q., Chen, G., Streb, J.W., Long, X., Yang, Y., Stoeckert, C.J. and Miano, J.M.. Defining the mammalian CArGome. *Genome research*. 2006.
8. Newman, A.A., Serbulea, V., Baylis, R.A., Shankman, L.S., Bradley, X., Alencar, G.F.,... & Owens, G. K. Multiple cell types contribute to the atherosclerotic lesion fibrous cap by PDGFRβ and bioenergetic mechanisms. *Nature metabolism*. 2021


## Contact


Mahima Reddy msr3nf@virginia.edu

twitter.com/MahimaReddy11

linkedin.com/in/mahima-reddy