Characterization of variants associated with Cerebral Small Vessel Disease identifies a functional SNV in Versican

Jae-Ryeon Ryu, PhD[1], Ankita Narang, PhD[1], Quentin LeGrand, PhD[2], David-Alexandre Trégouët, PhD[2], Stephanie Debette, MD,PhD[2,3], Sarah J. Childs PhD[1, *]

1. Department of Biochemistry and Molecular Biology and Alberta Children's Hospital Research Institute, University of Calgary, 3330 Hospital Drive NW, Calgary AB, T2N 4N1
2. Bordeaux Population Health Research Center, Inserm U1219, Université de Bordeaux, Bordeaux, France
3. Institut du Cerveau (ICM), Paris Brain Institute, INSERM U1127, UMR CNRS 7225 Paris, Sorbonne Université, Assistance Publique des Hôpitaux de Paris, Paris, France.

***Correspondence:** Sarah Childs, schilds@ucalgary.ca

Short title: Characterizing functional variants associated with CSVD

Abstract:

Single nucleotide variants in the non-coding genome can significantly alter disease risk, but identifying the function of SNVs is a challenge. Increasing numbers of variants of unknown significance have been associated with the risk of Stroke, Cerebral Small Vessel Disease and burden of White Matter Hyperintensities, but without biological validation, the significance of these findings is uncertain. We use a Multiplexed Parallel Reporter Assay (MPRA) to assay the function of 44 SNVs associated with Cerebral Small Vessel Disease and stroke in *EFEMP1, FOXF2, HAAO, KCNK3, NMT1, OPA1, RASL12, STAT3 and VCAN*. A functional SNV in versican was further probed using transcriptional reporter and Chromatin Immunoprecipitation assays. We identify 26 allele-specific enhancers where a SNV increases or decreases transcriptional activity. We validate rs13176921 as an SNV that significantly modulates an intronic enhancer in the matrisome protein Versican and show it affects mRNA levels of the gene. We identify that the transcription factor NKX3.1 binds to the region of this SNV. Thus, using MRPA we were able to gain mechanistic insight into which GWAS-identified SNVs modulate transcriptional activity, and validate a functionally important SNV in Versican and provide a mechanism by which it controls its transcription.

Author Summary:

Genome wide association studies identify many single nucleotide variant changes associated with diseases such as stroke and Cerebral Small Vessel disease, but not all the variants lead to disease. In this work we screen for variants that control transcription and therefore expression of adjacent genes. We find that a human variant associated with CSVD in the young and elderly decreases the expression of the extracellular matrix protein Versican. We use molecular biology experiments to determine the mechanism by which the variant changes expression and determine that the variant changes a transcription factor binding site.

Introduction

Small vessels (arterioles, capillaries, and venules) provide blood flow to support cognitive tasks through an organized distribution system. The smallest vessels with the slowest flow are critical for delivery of oxygen and nutrients to tissues. For this reason, damage to small vessels can lead to significant tissue consequences. Cerebral small vessel disease (SVD) results in poor brain blood flow and leads to ~20% of stroke and 45% of vascular dementias (1). SVD originates from blood vessels and secondarily affects neurons. The global lifetime risk of ischemic stroke in humans is 18.3% (2), and ~25% of ischemic stroke is due to SVD. SVD is thought to be a lifespan disease that begins during development and worsens across the lifespan (3).

Brain blood vessels are critical for maintaining the blood-brain barrier (BBB) through establishment of tight and adherens junctions between endothelial cells. A tight endothelium requires direct contact with vascular smooth muscle cells (vSMCs) and pericytes(4), providing physical support, extracellular matrix (ECM) deposition and contractility(5). Maturation of blood vessels requires coordination of reciprocal signalling among endothelium, mural cells, and ECM. The majority of vascular ECM is secreted by vascular mural cells (6) particularly during development. Mechanistically, many SVD-associated genes are members of the matrisome (6), the group of proteins that comprise or modify ECM.

Genome-wide association studies (GWAS), have identified monogenic (7, 8), and single nucleotide variants (SNVs) that are associated with higher risk of SVD (3, 9, 10) in known and novel genes associated with MRI-markers of SVD, such as white matter hyperintensity volume, the most common neuroimaging feature of SVD (3, 9, 11, 12) in aged populations, and novel biophysical diffusion MRI techniques (neurite orientation and dispersion and density imaging (NODDI)) to identify white matter microstructure differences in young adults(13). The role of these SNVs in modulating gene expression or vascular stabilization is unknown. Most SVD-associated SNVs are in either intergenic or intragenic regions in predicted cis-candidate regulatory elements (CCRE) enhancers on ENCODE(14). Thus, both the variants themselves and many of the genes they are located near, are currently of unknown significance with respect to the SVD mechanism.

In this study, we hypothesize that SVD-associated variants are in transcriptional enhancers, located in intragenic and intergenic regions of the non-coding genome. While enhancers can be

bioinformatically identified using epigenetic signatures and transcription factor (TF) binding, the precise mechanisms by which they regulate gene expression, the specific residues required for TF binding, and the target genes they control must be determined through functional experiments.

Multiplexed Reporter Assays (MPRAs) are unbiased assay of functional enhancer activity, testing the activity of alternative SNVs in a cell line-based assay (15). Using an MPRA, we determine allele-specific enhancer activity near variants located in stroke and SVD associated loci: *EFEMP1*, *FOXF2, HAAO, KCNK3*, *NMT1, OPA1, RASL12, STAT3* and *VCAN* (3, 9, 11, 13, 16, 17). For validation, we focus on intronic variants in Versican (*VCAN*) and identify rs13176921 as a critical variant that disturbs a conserved residue of the NKX3.1 binding site, suggesting that rs13176921 promotes *VCAN* gene expression via NKX3.1 binding.

Results

A multiplexed reporter assay identifies enhancers

In vitro Multiplexed Reporter Assays (MPRAs) report unbiased detection of allele-specific enhancer activity. We modified the original MPRA design(18) to analyze uncharacterized stroke-associated variants from GWAS studies. For this medium-scale screen, we selected 44 variants in 8 human genes associated with stroke (and subtypes), SVD MRI-markers, vascular risk factors (HDL- and LDL-cholesterol, triglycerides) or related diseases, such as venous thromboembolism or coronary artery disease (Supp. Table 3)(3, 9, 11, 12, 19, 20) The variants are located in intergenic or intronic regions adjacent to the following genes: *EGF Containing Fibulin Extracellular Matrix Protein 1/Fibulin 3 (EFEMP1), Forkhead Box F2 (FOXF2), 3-Hydroxyanthranilate 3,4-Dioxygenase (HAAO), Potassium Two Pore Domain Channel Subfamily K Member 3 (KCNK3), N-Myristoyltransferase 1 (NMT1), OPA1 Mitochondrial Dynamin Like GTPase (OPA1), RAS Like Family 12 (RASL12), Signal Transducer And Activator Of Transcription 3 (STAT3)* and *Versican (VCAN)*.

A pool of 968, 293bp oligonucleotides (Supp. Table 4), were synthesized. The oligo design contains forward adaptors, 204 bp of genomic context with major or minor alleles at the center, and barcodes (Fig. 1). Each oligo was synthesized with 1 of 11 different barcodes to avoid any barcode sequence bias and cloned upstream of a minimal promoter and GFP (Fig. 1). In this way, any enhancement or

repression of GFP expression levels can be quantitated through comparison of barcodes from DNA (transfection efficiency) and RNA for the same barcode (expression levels). Correlated expression 11 oligos with the same core sequence but different barcodes provide built-in redundancy (Supp. Table 4). As a positive control, we included known allele-specific enhancer variants in *FOXF2* we have previously validated using luciferase, ChIP and CRISPR assays(20).

As proof of concept, GFP expression in transfected cells confirmed the varied ability of the oligos to confer enhancer activity (Supp. Fig. 1). DNA and RNA were isolated from 5 biologically independent replicate transfections (Fig. 1A). To ensure that sequencing reads were from transfected material and not endogenous genomic DNA and RNA, we sequenced an amplified fragment containing the 3' end of GFP and the barcode bracketing the test oligo sequence. The sequence was bioinformatically trimmed, and non-specific amplification products that were not exactly 181 bp were removed. After this quality control step, an average of 83.62% of the unique reads were mapped in both DNA and RNA samples (Supp. Table 5). The raw counts from the five RNA and DNA samples show correlation of between 0.99 and 1.00 at the DNA level and 0.97 and 1.00 at the RNA level, suggesting high concordance between transfections (Supp. Fig. 2, Supp. Fig. 3). The majority of the 968 barcodes were represented in the sequencing. Only 21 (2.17%) were completely lacking from sequencing, while 146 were found at low count levels (less than 5 reads in both DNA and RNA). This suggests that the diversity of the library was well represented in the transfections (Supp. Fig. 4).

Using Bioconductor mpralm (21), we identified 26 allele-specific enhancers that met a significant false discovery rate (FDR) of <0.05, and a further 4 allele-specific enhancers that meet a suggestive FDR of P<0.1 (Supp. Table 6). In summary, MPRA analysis of stroke- and SVD-associated variants identified novel allele-specific enhancers suggesting potential functions in transcriptional regulation of the underlying genes.

Previously identified FOXF2 variants are allele-specific enhancers in the MPRA

To confirm that the MPRA identifies bone fide enhancers, we included previously characterized SVD-associated variants in *FOXF2*. We previously showed that a switch from C to T in rs74564934 results in a ~50% reduction in enhancer activity using a cell-based assay(20). We also included rs12204590, a lead variant associated with all stroke (16) but had no allele-specific effects in previous cell-based

assays. The MPRA shows that rs74564934, rs4959130 and rs7766042 show significant allele-specific activity, with the minor allele decreasing expression, agreeing with our experimental data (Fig. 2). In agreement with our published work, rs12204590 does not show allele-specific activity. The MPRA allowed us to predict that multiple transcription factor sites may be lost and gained at each of the SNVs using the MotifBreakR algorithm(22) (Supp. Fig. 5).

Novel allele-specific enhancers in other SVD and Stroke-associated genes

We assayed allele-specific for additional stroke-associated variants. We identify 3 allele-specific enhancers in *EGF-Containing Fibulin-Like Extracellular Matrix Protein 1 (EFEMP1*, also known as *Fibulin-3*) in the MPRA (Supp. Fig. 6). Rs376525 strongly decreases expression in the MPRA, and several transcription factor sites are predicted to be lost and gained at each of the SNVs using the MotifBreakR algorithm (Supp. Fig. 7).

One SNV near the 3-hydroxyanthranilate 3,4-dioxygenase (HAAO) gene showed significant allele-specific enhancer activity (Supp. Fig. 8). HAAO has modest expression in developing mural cells but is expressed higher in adult mural cells macrophages of the brain, and the kidney and arteries of adults. rs11679640 is predicted to strongly decrease binding of multiple retinoic acid orphan receptors, that act as transcription factors (ROR-α, -β, and -γ, encoded by *RORA*, *RORB*, and *RORC* respectively (Supp. Fig. 9)).

In the MPRA, we identified 12 allele-specific enhancers for *KCNK3* that either increase or decrease its expression (Supp. Fig. 10). 10 of the 12 allele-specific enhancers increase reporter expression in our assay suggesting that they may be located in transcriptional repressors, especially rs1275978 and rs1275979, which show strong allele-specific expression differences. Several transcription factor sites are predicted to be lost and gained at each of the SNVs using the MotifBreakR algorithm (Supp. Fig. 11).

RAS Like Family 12 (RASL12) is expressed in vascular mural cells, arteries, and vascular smooth muscle cells in human sequencing databases but is poorly studied (Supp. Fig 12). *RASL12* has two allele-specific enhancers that increase reporter expression (Supp. Fig. 12). Although rs12449945, an intronic variant within *RASL12,* is associated with white matter hyperintensity burden volume(11), the presence of the minor allele showed no significance in our study. Several transcription factor sites are predicted to be lost and gained at each of the SNVs using the MotifBreakR algorithm (Supp. Fig. 13).

Three other genes tested, *NMT1, OPA1 and STAT3,* had no allele-specific enhancers in our MPRA.

Novel allele-specific enhancers in Versican

Versican (*VCAN*) is a chondroitin sulfate proteoglycan (*CSPG2*) with roles in primordial ECM deposition in development. In zebrafish single cell sequencing, *vcana* and *vcanb* (two orthologs of human *VCAN*) are expressed in developing vascular mural cells (Fig. 4). In the first trimester developing human brain, *VCAN* is expressed in vascular mural cells. In adult brains, *VCAN* is expressed in both vascular cells, arteries, and brain oligodendrocyte precursor cells, as well as arteries (Fig. 4). Thus, *VCAN* is expressed in locations consistent with a role in SVD.

We evaluated allele-specific enhancer activity for 8 *VCAN* variants in LD (r^2 >0.88) with the lead variant rs67827860 associated with WMH and NODDI in GWAS studies. All variants are in intron 12. Four variants have significant allele-specific enhancer activity (rs17205972, rs13176921, rs12653308, and rs3852188), while rs13164785 is suggestive (Supp. Table 4; Fig. 3). Our analysis therefore identifies novel intronic enhancers for *VCAN*. Several transcription factor sites are predicted to be lost and gained at each of the SNVs using the MotifBreakR algorithm, including the transcription factor NKX3.1 (Supp. Fig. 14).

Allele-specific enhancers in VCAN do not alter alternative splicing

VCAN is a highly spliced transcript, with major splice forms V0, V1, V2, and V3. It is possible that disease-associated variants could affect splicing, even though all genomic variants tested in our MPRA are in an intron downstream of the alternative exons, and it is unlikely. We used qPCR to assay *VCAN* splice form expression in IMR-32, a neuroblastoma line robustly expresses all V0-V3 variants. We note

that the SNVs in IMR-32 are of the major allele in this line, and this is the only cell line we identified with expression of all *VCAN* splice forms (Supp. Fig. 15.). To assess intron 12 variant effects on splicing vs. gene expression, we deleted a ~2 kb intronic region containing 3 significant allele-specific enhancers in IMR-32 cells by simultaneous transfection of two CRISPR guides and Cas9 (Fig. 5A). We single-cell cloned two CRISPR modified lines and sequenced breakpoints around the deletion, confirming they both had similar sequences (Fig 5B). Using qPCR, we show that *VCAN* V3 expression is significantly reduced when the 2kb enhancer is deleted (Fig 5C). However, there were no changes to alternative splicing of *VCAN* V0, V1 and V2 using qPCR. Our data suggests that intron 12 promotes *VCAN* V3 expression but does not modulate splicing (Fig 5C).

Identification of a functional SNV in intron 12 of VCAN

In the 2kb intron 12 region we deleted there are 3 disease-associated SNVs. rs10052710 is associated with the Neurite Density Index (NDI) metric (13), rs17205972 is associated with WMH volume (11) , and rs13176921 is associated with NDI in four different brain regions (13). All SNVs are in LD. Allele-specific enhancer activity was non-significant for rs10052710, and while it was significant for both rs17205972 and rs13176921. rs13176921 shows a larger magnitude of fold change between major and minor alleles, and thus we focused on this SNV for validation. Using SNP2TFBS (23) and MotifBreakR (22) to predict transcription factor binding sites disrupted by rs13176921, a strong binding site for the transcription factor NKX3.1 was identified, where the G>A change is at a highly conserved position, predicted to abolish binding to the minor allele (Fig 5F). NKX3.1 is a transcription factor that powerfully induces a mural cell precursor phenotype, which intrigues us because of the relationship between mural cells, ECM, and vascular stability (24, 25). rs13176921 is located in a region of open chromatin as shown by snATACseq in the CATlas (26) (Fig 3C), in both oligodendrocyte precursor cells and vascular smooth muscle cells. To specifically test that the SNV (G-A) changes enhancer activity we used a luciferase assay, cloning two identical 300 bp oligonucleotides, varying only at the SNV position, ahead of a minimal enhancer and luciferase (Fig 5D). This showed that the 300 bp region around rs13176921 is a strong transcriptional enhancer. Further, presence of the minor allele (G), as opposed to the major allele (A), significantly reduces luciferase activity confirming the results of the MPRA assay in an independent assay (Fig 5E). Finally, we show that NKX3.1 binds to this region using Chromatin Immunoprecipitation (Fig 5G), suggesting it could regulate this enhancer

sequence.

Discussion

Disease-associated SNVs are a window into the function of the genome. GWAS and downstream bioinformatic analysis have become increasingly sophisticated but provide incomplete insight into variant biological function. Of the 44 variants tested in our MPRA assay, 26 had allele-specific enhancer properties suggesting that disease-associated variants arising out of highly powered GWAS studies identify DNA sequences capable of modulating gene expression. Mechanistically, several genes are members of the human matrisome [18] and have strong expression in the aorta/artery[19]. including EFEMP1 (Fibulin 3), *VCAN* (Versican), and *RASL12* a poorly studied gene expressed in a similar pattern to *EFEMP1* and *VCAN*. On the other hand, *HAAO, NMT1*, *OPA1*, *STAT3* and *KCNK3* are more broadly expressed, and are not part of the matrisome. Their connection to SVD and stroke likely occurs via a mechanism other than by ECM regulation.

We included *FOXF2* stroke-associated variants(20) that had been tested using cell-based assays as a positive control. We previously identified that ETS1 bound to the major allele of rs74564934 and presence of the minor allele resulted in ~50% expression(20). We extend this work here by identifying additional allele-specific enhancers around rs7766042 and rs4959130. The three SVD-associated *FOXF2* variants are spaced 10-20 kb apart and each is in a different ENCODE CCRE, suggesting complex regulation of *FOXF2* from multiple enhancers in this intergenic region. Thus, we identify new *FOXF2* enhancers in this study.

Versican is of particular interest because variants associate with MRI-markers of SVD in young adults and the elderly(11, 13). We show that intron 12 of *VCAN* has at least 4 transcriptional enhancers that

positively regulate *VCAN* expression. Presence of the minor (risk) variant decreases reporter expression in vitro. Deletion of the enhancer, and luciferase reporter assays specifically testing rs13176921 demonstrate the functional role of the major allele in positively regulating *VCAN* expression. Further, we demonstrate the transcription factor NKX3.1 binds to the enhancer surrounding Rs13176921. NKX3.1 binding is predicted to be completely abolished by presence of the minor allele. Although MotifBreakR predicts that rs13176921 could bind other transcription factors such as DNMT1, there is no predicted effect between DNMT1 binding to major and minor allele, it was not studied further here. However, additional dissection of enhancer architecture would be valuable.

rs13176921 is present globally in an average of 23% of humans (dbSNP). Versican is a matrisome component and an ECM protein. Developmental ECM is essential for establishing the mechanical and signaling environment of cells. Many ECM proteins have long half-lives and are largely synthesized in development, setting up initial patterns of vascular stability for the lifespan. Versican, together with hyaluronan, a highly hydrated polysaccharide, forms a soft 'provisional matrix' (27) in development. Versican is a large protein that can be modified by chondroitin sulfate proteoglycan on alternatively spliced exons present in the V0, V1 and V2 splice forms. The core protein splice variant, V3, is not glycated and is expressed in provisional matrix promoting tropoelastin synthesis and elastin deposition(28, 29). Non-elastic, stiff blood vessels cannot respond to changes in pulsatile blood flow, resulting in hypertension, a risk factor that enhances SVD. *VCAN* is also expressed in the aorta, arteries, vascular mural cells, and oligodendrocytes in developing and adult humans and oligodendrocytes after injury(30).

We identified allele-specific enhancers in other stroke and SVD-associated genes. *EFEMP1 (Fibulin 3)* is a secreted glycoprotein expressed by connective tissues that interacts with elastic fibers. While

EFEMP1 binds the EGFR receptor, it also has roles in cell migration, adhesion, formation, and homeostasis of elastic tissues. *EFEMP1* is strongly expressed in arteries and binds to vascular smooth muscle cells. *EFEMP1* is processed by *HTRA1* (the CADASIL2 gene causative of SVD(31, 32)). We identified 3 allele-specific enhancers for *EFEMP1*.

HAAO is an intracellular dioxygenase involved in the synthesis of quinolinic acid. Quinolinic acid is involved in neuroinflammation, suggesting a mechanistic link to SVD. HAAO expression has not previously been reported in mural cells of the brain, but Human Protein Atlas data suggests it is enriched in mural cells and macrophages. Rs11679640 was identified as a genome-wide significant locus for WMH intensity burden (33). MotifBreakR predicts that the minor allele of Rs11679640 will strongly reduce the binding of retinoic acid receptors; this prediction has not yet been experimentally validated.

Intergenic and intronic variants near the *Potassium Two Pore Domain Channel Subfamily K Member 3 (KCNK3)* are associated with cerebrovascular events in a multiancestry study (34), with all stroke(19), pulmonary arterial hypertension and increased proliferation, vasoconstriction, and inflammation in smooth muscle cells(35), and autosomal dominant pulmonary hypertension(36). *KCNK3* is expressed in smooth muscle cells. Interestingly, 9 of the *KCNK3* variants we tested significantly increased enhancer activity, one as high as 8-fold. GTEX shows that transcripts in pancreas and heart derived from individuals with the minor alleles of rs1275985, rs1275986, rs1275923, rs1275978, rs1275979, and rs1275980, have higher expression of transcripts than those with the major allele, supporting our data that the minor alleles of these SNVs may increase enhancer activity in tissues.

RASL12 is a poorly studied Ras-like GTPase expressed in mural cells. Functionally, it blocks YAP1 nuclear localization(37). *RASL12* also binds *RASSF1A*, recruiting it to microtubules(37). As *RASSF1A*

interacts with Hippo signaling, it is possible that *RASL12* may act in smooth muscle cells via Hippo pathway promoting vascular smooth muscle cell differentiation (38). *RASL12* has strong expression in human embryonic brain vascular mural cells, human adult vascular mural cells and astrocytes, along with strong expression in human arteries suggesting a plausible role in SVD. We identified 2 intronic allele-specific enhancers in *RASL12*.

In summary, MPRA assays distinguish variants with cis-regulatory enhancer activity (i.e. functional) vs. variants that are simply in LD with important variants. This information is important for stratifying risk in individuals, particularly for those who would benefit from preventative therapies. For instance, hypertension is a comorbidity that increases the burden of neurological damage in individuals who carry high risk SVD-associated variants(11). Theoretically, the long latency between detectable brain changes on MRI in young adults, and SVD and stroke in later life gives a window to offer lifestyle changes and therapeutics for those at greatest risk.

Limitations: A limitation of the MPRA design is that the relatively short oligo may be shorter than typical mammalian cis-regulatory enhancers that are 100-1000 bp (39) and may require 'collective occupancy' (40). Finally, the assay was run in a cell line that does not exactly match the context of gene expression due to the need for high transfection efficiency of a library(15). Nonetheless, the speed and breadth of the MPRA can rapidly identify functional SNVs for further dissection with functional studies in an appropriate genomic context.

Materials and methods

Primers and reagents are found in Supplemental Tables 1 and 2.

MPRA library oligonucleotide synthesis

A pool of 968 oligonucleotides (Supp. Table 3) was synthesized containing 102 bp of sequence upstream and downstream of the variant, the major or minor variant, a multi-enzyme cloning site, a 10 bp barcode and a reverse adaptor (Twist Bioscience South San Francisco, CA). To amplify the library, 12 independent PCR reactions were set up containing 1 ng the synthesized MPRA oligos, 10 µl of 5X Phusion HF buffer, 200 µM dNTPs, 1-unit Phusion HF DNA polymerase (New England Biolabs), 0.5uM MPRA oligo-f and MPRA oligo-r (Supp. Table 1). The PCR-amplified MPRA oligos were pooled and purified with NucleoSpin Gel & PCR clean-up column (Macherey-Nagel).

MPRA oligo library assembly

10µg of pMPRAv3:Δluc:ΔxbaI (Addgene) was digested with SfiI (NEB) at 50C overnight and gel extracted (NucleoSpin Gel & PCR clean-up column, Macherey-Nagel). To create the MPRA oligo library, the MPRA oligos and digested pMPRAv3:Δluc:Δxbal were assembled using NEBuilder HiFi DNA Assembly (NEB). The reaction was purified with NucleoSpin PCR clean-up column. 150ng of the assembled library was transformed into NEB 10-Beta high efficiency competent cells (NEB). Cells were cultured in 4 x50 ml LB with Ampicillin at 37C for 6 hours to amplify the library, pooled and plasmid DNAs purified (NucleoBond Xtra Midi EF, Macherey-Nagel).

MPRA oligo:GFP library assembly

10µg of MPRA oligo library was digested with AsiSI (NEB) at 37C overnight and gel-extracted (NucleoSpin Gel & PCR clean-up) and PCR amplified with MiniP-GFP-f and MiniP-GFP-r primers18

(Supp. Table 1). The MPRA library incorporating GFP was created using 2µg of MiniGFP amplicon and 1µg of AsiSI linearized MPRA oligo library and assembled using NEBuilder HiFi DNA Assembly at 50C for 90 min, then purified with NucleoSpin PCR clean-up column, eluted and digested to remove remaining uncut vector by incubation with 25U AsiSI, and 5U RecBCD Plasmidsafe nuclease (NEB), 10µg BSA, 1 mM ATP in a 100µl reaction at 37 C overnight followed by NucleoSpin PCR clean-up column.

The eluted MPRA oligo:GFP library was transformed into NEB 10-Beta high efficiency competent cells, split into 4 tubes, and recovered in 1 ml of NEB 10-Beta/Stable Outgrowth Medium, incubated at 37 C for 1 hour. The mixture was diluted to >1X106 CFU/ml. MPRA library complexity was assessed by sequencing 12 random colonies. 3 independent replicates of four 100 ml flasks of LB-Ampicillin were incubated at 37C for 6 hours to prepare the library for transformation. Plasmid preparations were pooled and normalized to 1µg/µl.

Transfection of MPRA oligo:GFP library assembly

HEK 293 cells were seeded at 1.5X106 cells per 100 mm plate in D10 medium (DMEM supplemented with 10 %FBS) 24hr before transfection. On the day of transfection, cells at 60-70% confluency were transfected with 10µg of MPRA oligo:GFP library assembly and 20 µl of Lipofectamine 3000 (Life Technologies). The transfected cells were incubated for 24hr in D10 medium. The transfection was repeated independently 5 times.

Total RNA and DNA extraction

Total RNA and DNA were extracted using TRIzol (Invitrogen) as per the manufacturer's protocol. Total RNA was extracted and purified using Zymo-Spin-IC column (RNA clean& concentrator-5) with on-column DNase digestion. A second DNase treatment was performed on the purified RNA using 5 µl of

Turbo DNase (Life Technologies) in 300 µl of total volume for 1 hour at 37C followed by a Zymo-Spin IC column.

cDNA, product amplification, and sequencing

First-strand cDNA synthesis was from DNase treated RNA using a primer specific to the 3' GFP (RT-GFP_ 2 specific primer; Supp. Table 1) and SuperScript IV Reverse Transcriptase (Invitrogen) following the manufacturer's protocol. Single-stranded cDNA was purified by NucleoSpin Gel & PCR clean-up column.

To add Illumina sequencing adaptors, we performed eight independent 50 µl PCR reactions: 10 ng of template DNA (5 DNA samples and 5 single-stranded cDNA), 25 µl of 2X NEBNext UltraII Q5 MM (NEB), 0.5 µM of Illumina adaptor_ GFP_R and Illumina adaptor_ GFP_F primers (Supp. Table 1). The PCR amplified samples were gel-extracted to remove primers and dNTPS using NucleoSpin Gel & PCR clean-up column. Samples were sequenced by 250bp paired ends sequencing without fragmenting the amplicons on an Illumina MiSeq (Azenta, NJ).

MPRA data processing: Mapping of DNA and RNA barcodes

Forward and reverse fastq reads were mapped to barcodes separately using locate module of SeqKit (v2.8.0). According to the experimental design, reverse transcribed barcodes were used as a reference to align with forward reads (R1) and while forward barcodes were used a reference to align with reverse reads (R2). To reduce false positives, no mismatches were allowed for mapping. Secondly, mapped reads of length 181 were selected where barcodes were mapped to specific positions in both forward and reverse reads. Read counts for barcodes were generated separately for forward and reverse reads per replicate using custom R script (https://github.com/ankita86/MPRA_Childs_Lab).

We kept the highest barcode count either from reverse or forward read for downstream analysis. We summed barcode counts per allele or element to generate aggregate counts of DNA and RNA.

Identification of allele-specific enhancers

To identify allele-specific enhancers, paired aggregated counts of major and minor alleles or elements for both DNA and RNA was provided as an input to mpralm function of mpra bioconductor R package. This linear based model assesses the differential activity of alleles. To construct the linear model, we passed following parameters to the mpralm function: mpralm(object = mpraset, design = design, aggregate = "none", normalize = TRUE, block = block_vector, model_type = "corr_groups", plot = TRUE). A SNP with an FDR-adjusted p-values $P<0.05$ (calculated using a multiple-testing correction method to reduce the false discovery rate) is defined as an active allele-specific enhancer and those with P-value between 0.05 and 0.01 are defined as suggestive allele-specific enhancers19.

Cell lines:

HEK293 (ATCC CRL-1573) and IMR-32 (gift from Dr. A. Narendran) were grown in D10 culture media (DMEM, Gibco 11995073) supplemented with 10% fetal bovine serum (FBS, Gibco 10082147) at 37oC with 5 % CO2. Genomic DNA was isolated using Purelink Genomic DNA mini kit (Invitrogen K182001). The STR genotype of the cells was confirmed by profiling at The Centre for Applied Genomics (TCAG) (Toronto, Canada).

Testing rs13176921 minor allele activity using luciferase assays:

Two 300 bp oligonucleotides containing either the major (A) and minor (G) alleles of rs13176921 were synthesized (Twist Biosciences; Supp. Table 1) and cloned into the pGL4.23GW-Luciferase reporter vector (Addgene). HEK 293 cells (2X104 cells/100µl D10 culture medium) were seeded on white clear 96 well plate (Greiner Bio-One). The following day, 100 ng of enhancer-Luc plasmid, 5ng of pTK-renilla

as an internal control and 0.3µl of Lipofectamine 3000 Transfection Reagent were added in10 µl Opti-MEM (ThermoFisher) medium per well and incubated for 2 days before the medium was aspirated. 30µl DPBS (Gibco) and 30µl of Dual-Glo reagent (Promega) was added into each well and incubated for 30 min at room temperature. Luciferase activity was measured using Dual-Glo Luciferase SpectraMaxL (Molecular Devices) before 30µl of Stop-Glo solution added and incubated for 30 min at room temperature and Renilla activity measured. Relative luciferase activity is luciferase activity/ Renilla activity.

Testing NKX3.1 binding to the region around rs13176921

H1 cells expressing NKX3.1 (following 20) that had been differentiated to induced mural progenitor cells (iMPCs) were pelleted (kind gift from Dr. Manna Abubaker) at 760 × g for 5 min at 4°C, resuspended with media, cross-linked with 1% formaldehyde (ThermoFisher) with shaking for 10 min at room temperature before treatment with 125mM Glycine (Sigma) for 5 min at room temperature. Cells were pelleted at 760g for 5 min at 4oC to pellet cells. Cross-linked cells were lysed with ChIP Lysis buffer (50mM TrisHCl pH 8, 10mM EDTA pH 8.0, 1% SDS, 1X Complete Protease inhibitor cocktail tablets (Roche), 1 mM PMSF (Sigma) for 10 min on ice. The lysate was sonicated for 25-30 cycles of 30 seconds ON, 30 seconds OFF at High Power using a Bioruptor (Diagenode). The sonicated lysate was centrifuged at 16,000g at 4 oC for 20 minutes and the sheared chromatin was transferred to new tubes.

Each 500ug of sheared chromatin-protein complex was incubated with antibody targeting anti-NKX3.1 (Cell Signaling) and normal rabbit Ig G (EMD Millipore) as a negative control, respectively at 4oC overnight. After immunoprecipitation, 30µl of pre-adsorbed Dynabeads protein G (Invitrogen) were added and rotated further for 6hr at 4 oC. The Antibody/protein/chromatin/bead complex was

washed 4x with RIPA buffer for 10 min at 4 oC and 2x with TE buffer before reverse-crosslinking at 65oC overnight, while rotating using a Thermomixer at 500rpm. DNA was purified by DNA purification columns (Omega). The relative enrichment of specific regions in precipitated DNA was measured by quantitative PCR (qRT-PCR) using PowerUp SYBR green master mix (Applied Biosystems) and a QuantStudio6 PCR machine.

Knock out CRISPR sgRNA mutagenesis, clonal selection, and qPCR validation

All knock out CRISPR target oligos (Supp. Table 1) were designed with the ends: 5'-caccg (BbsI) N20-3' and 5'-aaac (BbsI) N20-c-3'. Annealed oligos were inserted into pSpCas9 (BB2)-2A-GFP (pX458) vector (Addgene) 21digested with FastDigest BpiI (BbsI) (ThermoFisher) as above. CRISPR sgRNA plasmids were verified by sequencing. Transfection and clonal analysis were as reported 22.

To sequence VCAN knockout clones, total RNA was isolated using TRIzol (Invitrogen) and purified with Zymo RNA clean and concentrator-5 (Zymo Research). For cDNA synthesis, 1ug total RNA mixed with 4µl of LunaScript RT mix (NEB) in a 20µl reaction (25°C for 2 min, 55°C for 10 min at, 95°C for 1 min). cDNA was used for real time quantitative PCR (qPCR) to determine gene expression using custom primers (Supp. Table 1). qPCR was performed using PowerUp SYBR Green Master Mix (Applied Biosystems) on a BioRad Opus 96 (BioRad). Cycle conditions followed 95°C for 3 min, 40 cycles of 95 °C for 10 s and 60 °C for 30 s. Fold increase or decrease in gene expression were determined relative to control cells using the ΔΔCT method.

Statistical analysis

All experimental replicates are noted in the results, and are typically a minimum of 3 biological replicates, and averaging 3 technical replicates (qPCR). Experimental statistics were calculated in GraphPad Prism using tests noted in figure legends.

Data and code availability:

The datasets generated during this study are available in the Supplemental Tables. The sequencing data is deposited in GEO with the accession: GSE316826. The code generated during this study is available at https://github.com/ankita86/MPRA_Childs_Lab.

Acknowledgements

Salary support for bioinformatic analysis (AN) is supported by the Alberta Children's Hospital Research Institute. This work was initiated from a Cumming School of Medicine at the University of Calgary Research Enhancement fund grant to SJC. Funding to SJC was from a Canadian Institutes of Health Research Project Grant (PJT-203822). SD acknowledges support from the French National Research Agency and France 2030 (ANR-18-RHUS-0002, RHU-SHIVA; ANR-23-IAHU-0001, IHU-VBHI), Prix Burrus-FRM and NRJ-neurosciences, EU Horizon 2020 (grant No 754517). SD and QLG acknowledge support from the Fondation Recherche Alzheimer.

References

1. Cannistraro RJ, Badi M, Eidelman BH, Dickson DW, Middlebrooks EH, Meschia JF. CNS small vessel disease: A clinical review. Neurology. 2019;92(24):1146-56.
2. Global, Regional, and Country-Specific Lifetime Risks of Stroke, 1990 and 2016. New England Journal of Medicine. 2018;379(25):2429-37.
3. Bordes C, Sargurupremraj M, Mishra A, Debette S. Genetics of common cerebral small vessel disease. Nat Rev Neurol. 2022;18(2):84-101.
4. Armulik A, Genové G, Mäe M, Nisancioglu MH, Wallgard E, Niaudet C, et al. Pericytes regulate the blood–brain barrier. Nature. 2010;468(7323):557-61.
5. Siekmann AF. Biology of vascular mural cells. Development. 2023;150(16):dev200271.
6. Joutel A, Haddad I, Ratelade J, Nelson MT. Perturbations of the cerebrovascular matrisome: A convergent mechanism in small vessel disease of the brain? J Cereb Blood Flow Metab. 2016;36(1):143-57.
7. Gould DB, Phalan FC, van Mil SE, Sundberg JP, Vahedi K, Massin P, et al. Role of COL4A1 in small-vessel disease and hemorrhagic stroke. N Engl J Med. 2006;354(14):1489-96.
8. Joutel A, Corpechot C, Ducros A, Vahedi K, Chabriat H, Mouton P, et al. Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. Nature. 1996;383(6602):707-10.
9. Duperron M-G, Knol MJ, Le Grand Q, Evans TE, Mishra A, Tsuchida A, et al. Genomics of perivascular space burden unravels early mechanisms of cerebral small vessel disease. Nature Medicine. 2023;29(4):950-62.
10. Mishra A, Chauhan G, Violleau M-H, Vojinovic D, Jian X, Bis JC, et al. Association of variants in HTRA1 and NOTCH3 with MRI-defined extremes of cerebral small vessel disease in older subjects. Brain. 2019;142(4):1009-23.
11. Sargurupremraj M, Suzuki H, Jian X, Sarnowski C, Evans TE, Bis JC, et al. Cerebral small vessel disease genomics and its implications across the lifespan. Nat Commun. 2020;11(1):6285.
12. Persyn E, Hanscombe KB, Howson JMM, Lewis CM, Traylor M, Markus HS. Genome-wide association study of MRI markers of cerebral small vessel disease in 42,310 participants. Nat Commun. 2020;11(1):2175.
13. Le Grand Q, Tsuchida A, Koch A, Imtiaz M-A, Aziz NA, Vigneron C, et al. Diffusion imaging genomics provides novel insight into early mechanisms of cerebral small vessel disease. Molecular Psychiatry. 2024.
14. Kagda MS, Lam B, Litton C, Small C, Sloan CA, Spragins E, et al. Data navigation on the ENCODE portal. Nat Commun. 2025;16(1):9592.
15. Easterlin R, Ahituv N. Lineage-specific regulatory evolution: insights from massively parallel reporter assays. Curr Opin Genet Dev. 2025;93:102372.
16. Chauhan G, Arnold CR, Chu AY, Fornage M, Reyahi A, Bis JC, et al. Identification of additional risk loci for stroke and small vessel disease: a meta-analysis of genome-wide association studies. The Lancet Neurology. 2016;Jun 15 (7):695-707.
17. Mishra A, Malik R, Hachiya T, Jürgenson T, Namba S, Posner DC, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. Nature. 2022;611(7934):115-23.
18. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016;165(6):1519-29.
19. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018;50(4):524-37.
20. Ryu JR, Ahuja S, Arnold CR, Potts KG, Mishra A, Yang Q, et al. Stroke-associated intergenic variants modulate a human FOXF2 transcriptional enhancer. Proc Natl Acad Sci U S A. 2022;119(35):e2121333119.
21. Myint L, Avramopoulos DG, Goff LA, Hansen KD. Linear models enable powerful differential activity analysis in massively parallel reporter assays. BMC Genomics. 2019;20(1):209.
22. Coetzee SG, Hazelett DJ. motifbreakR v2: expanded variant analysis including indels and integrated evidence from transcription factor binding databases. Bioinformatics Advances. 2024;4(1).
23. Kumar S, Ambrosini G, Bucher P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res. 2017;45(D1):D139-d44.

24. Lee U, Zhang Y, Zhu Y, Luo AC, Gong L, Tremmel DM, et al. Robust differentiation of human pluripotent stem cells into mural progenitor cells via transient activation of NKX3.1. Nat Commun. 2024;15(1):8392.
25. Ahuja S, Adjekukor C, Li Q, Kocha KM, Rosin N, Labit E, et al. The development of brain pericytes requires expression of the transcription factor nkx3.1 in intermediate precursors. PLoS Biol. 2024;22(4):e3002590.
26. Zu S, Li YE, Wang K, Armand EJ, Mamde S, Amaral ML, et al. Single-cell analysis of chromatin accessibility in the adult mouse brain. Nature. 2023;624(7991):378-89.
27. Wight TN. Provisional matrix: A role for versican and hyaluronan. Matrix Biol. 2017;60-61:38-56.
28. Merrilees MJ, Lemire JM, Fischer JW, Kinsella MG, Braun KR, Clowes AW, Wight TN. Retrovirally mediated overexpression of versican v3 by arterial smooth muscle cells induces tropoelastin synthesis and elastic fiber formation in vitro and in neointima after vascular injury. Circ Res. 2002;90(4):481-7.
29. Wight TN, Day AJ, Kang I, Harten IA, Kaber G, Briggs DC, et al. V3: an enigmatic isoform of the proteoglycan versican. American Journal of Physiology-Cell Physiology. 2023;325(2):C519-C37.
30. Asher RA, Morgenstern DA, Shearer MC, Adcock KH, Pesheva P, Fawcett JW. Versican is upregulated in CNS injury and is a product of oligodendrocyte lineage cells. J Neurosci. 2002;22(6):2225-36.
31. Hara K, Shiga A, Fukutake T, Nozaki H, Miyashita A, Yokoseki A, et al. Association of HTRA1 mutations and familial ischemic cerebral small-vessel disease. N Engl J Med. 2009;360(17):1729-39.
32. Lin MK, Yang J, Hsu CW, Gore A, Bassuk AG, Brown LM, et al. HTRA1, an age-related macular degeneration protease, processes extracellular matrix proteins EFEMP1 and TSP1. Aging Cell. 2018;17(4):e12710.
33. Verhaaren BF, Debette S, Bis JC, Smith JA, Ikram MK, Adams HH, et al. Multiethnic genome-wide association study of cerebral white matter hyperintensities on MRI. Circ Cardiovasc Genet. 2015;8(2):398-409.
34. Verma A, Huffman JE, Rodriguez A, Conery M, Liu M, Ho YL, et al. Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. Science. 2024;385(6706):eadj1182.
35. Antigny F, Hautefort A, Meloche J, Belacel-Ouari M, Manoury B, Rucker-Martin C, et al. Potassium Channel Subfamily K Member 3 (KCNK3) Contributes to the Development of Pulmonary Arterial Hypertension. Circulation. 2016;133(14):1371-85.
36. Ma L, Roman-Campos D, Austin ED, Eyries M, Sampson KS, Soubrier F, et al. A novel channelopathy in pulmonary arterial hypertension. N Engl J Med. 2013;369(4):351-61.
37. Dhanaraman T, Singh S, Killoran RC, Singh A, Xu X, Shifman JM, Smith MJ. RASSF effectors couple diverse RAS subfamily GTPases to the Hippo pathway. Sci Signal. 2020;13(653).
38. Manderfield LJ, Aghajanian H, Engleka KA, Lim LY, Liu F, Jain R, et al. Hippo signaling is required for Notch-dependent smooth muscle differentiation of neural crest. Development. 2015;142(17):2962-71.
39. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biology. 2021;22(1):108.
40. Jindal GA, Farley EK. Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev Cell. 2021;56(5):575-87.
41. Sur A, Wang Y, Capar P, Margolin G, Prochaska MK, Farrell JA. Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. Dev Cell. 2023;58(24):3028-47.e12.
42. Crouch EE, Bhaduri A, Andrews MG, Cebrian-Silla A, Diafos LN, Birrueta JO, et al. Ensembles of endothelial and mural cells promote angiogenesis in prenatal human brain. Cell. 2022;185(20):3753-69.e18.
43. Ponten F, Jirstrom K, Uhlen M. The Human Protein Atlas--a tool for pathology. J Pathol. 2008;216(4):387-93.
44. Consortium TG, Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318-30.

Figures

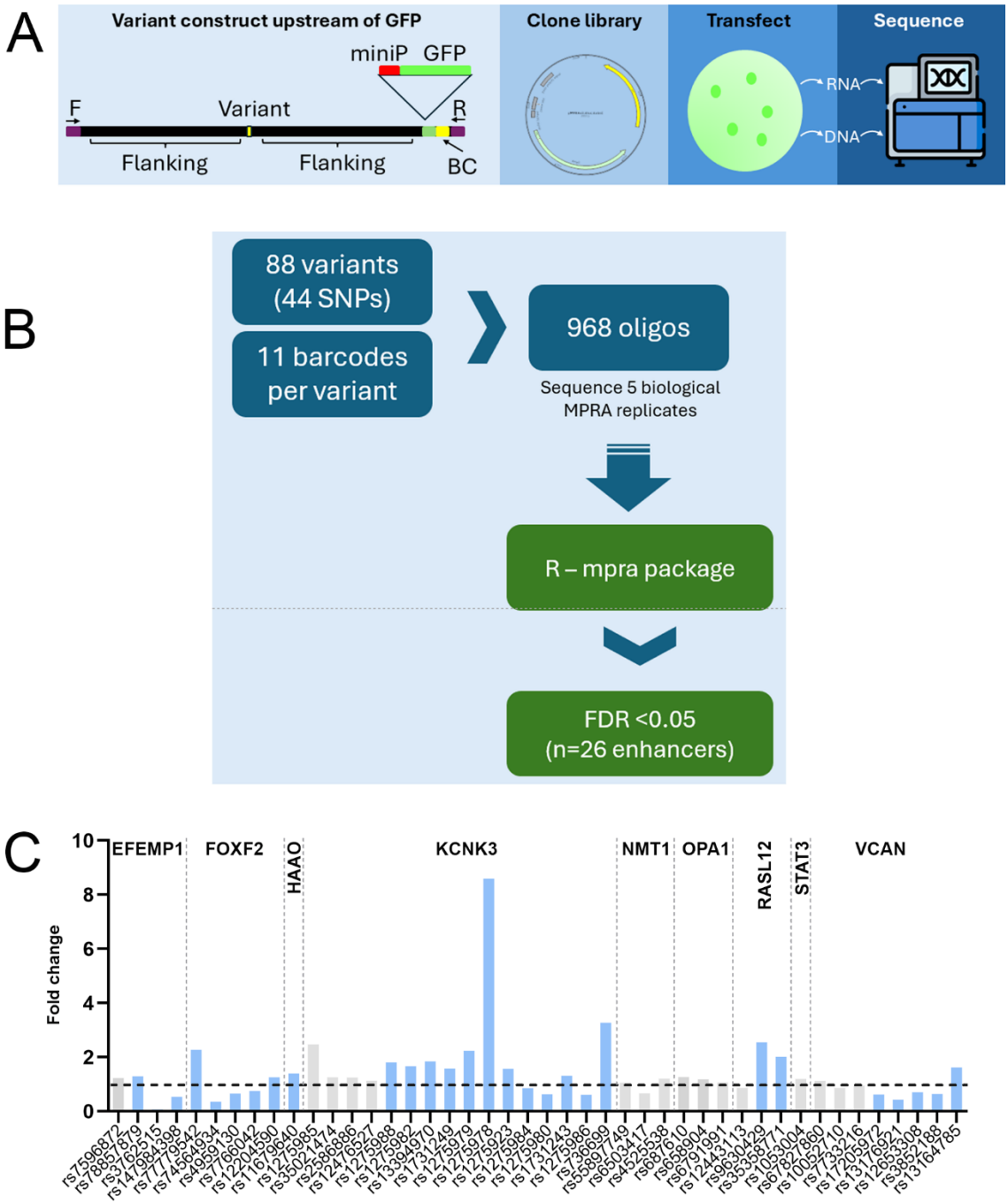
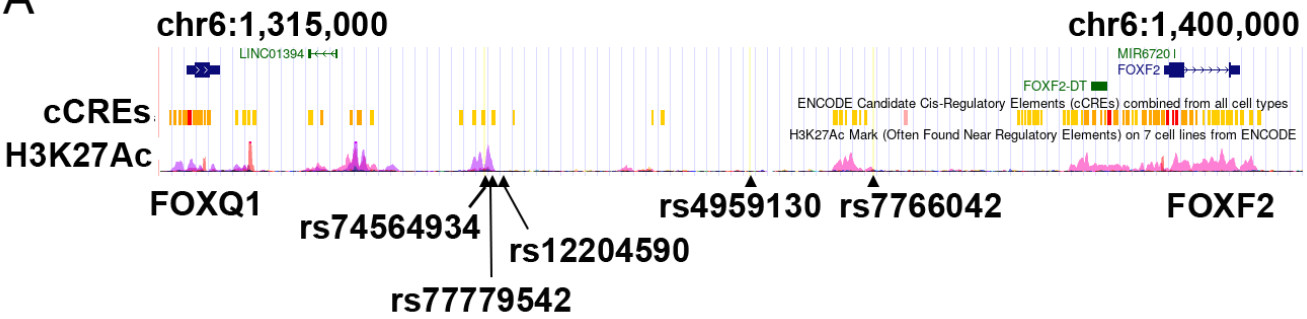


Figure 1: MPRA methodology. A) Schematic of the MPRA synthetic oligonucleotide including a forward primer (F), 102 bp of flanking sequence before, and 102 bp of flanking sequence after, the variant followed by a miniPromotor (miniP)- GFP construct, a 10 bp barcode and reverse primer (R). The oligo library was cloned into the pMPRAv3:Δluc:ΔxbaI vector and transfected into HEK293 cells. DNA and RNA was isolated separately and sequenced using MiSeq. B) Bioinformatic workflow for the MPRA analysis showing how allele-specific enhancers were identified. C) Fold change in expression for alelle specific enhancers in the indicated genes. Blue bars show significant allele-specific activity.

A



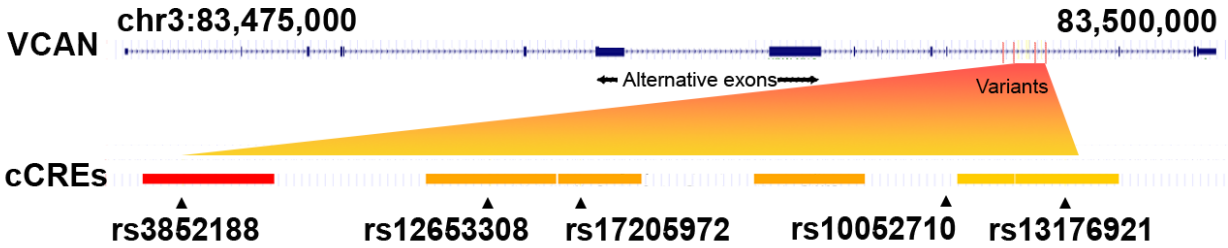
B

SNP	logFC	adj.P.Val	Allele-specific enhancer
rs7766042	-0.41	0.02047	Significant
rs12204590	0.33	0.03331	Significant
rs77779542	1.18	0.00000	Significant
rs74564934	-1.49	0.00004	Significant
rs4959130	-0.62	0.00026	Significant

Figure 2: The MPRA identifies known allele-specific enhancers in FOXF2

A) Intergenic region between FOXQ1 and FOXF2 showing location of variants that are significant allele-specific enhancers from the MPRA. B) Table showing the log fold change and adjusted p value for the expression difference between major and minor alleles (allele-specific enhancers).

A



B

SNP	logFC	adj.P.Val	Allele-specific enhancer
rs13176921	-1.22	0.00040	Significant
rs17205972	-0.69	0.00000	Significant
rs3852188	-0.65	0.02845	Significant
rs12653308	-0.50	0.00067	Significant
rs13164785	0.69	0.06698	Suggestive

C

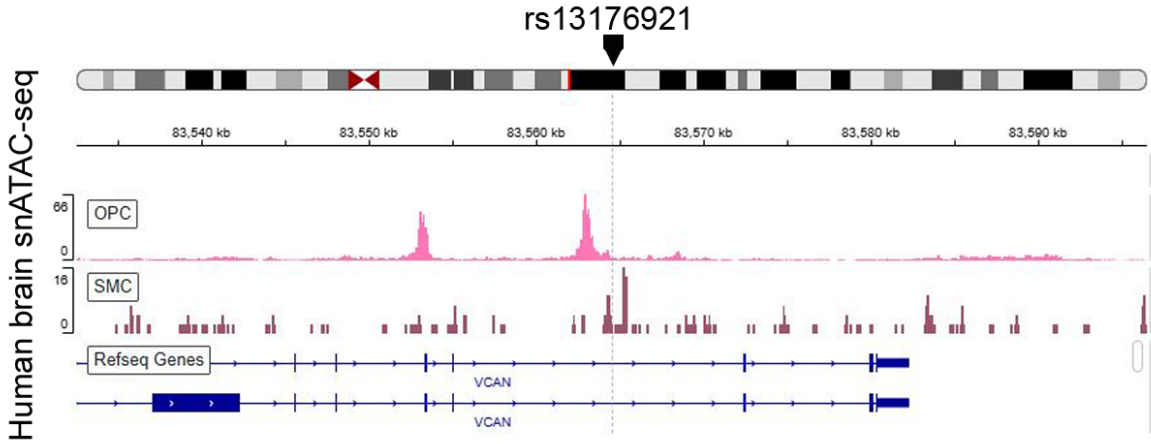


Figure 3: The MPRA identifies allele-specific enhancers in VCAN

A) Map of the VCAN gene with enlargement of the region of the variants identified in our study. The orange trapezoid shows how variants map to intron 12 of VCAN, in Candidate Cis-Regulatory Elements (cCREs) from ENCODE (red and yellow bars). B) Table showing log fold change an adjusted P value as determined by the MRPA assay for significant allele-specific enhancers. C) Representative human brain snATAC-seq data from CATlas(26) around rs13176921 in oligodendrocyte precursor cells (OPC) and smooth muscle cells (SMCs) showing open chromatin in the region of this variant.

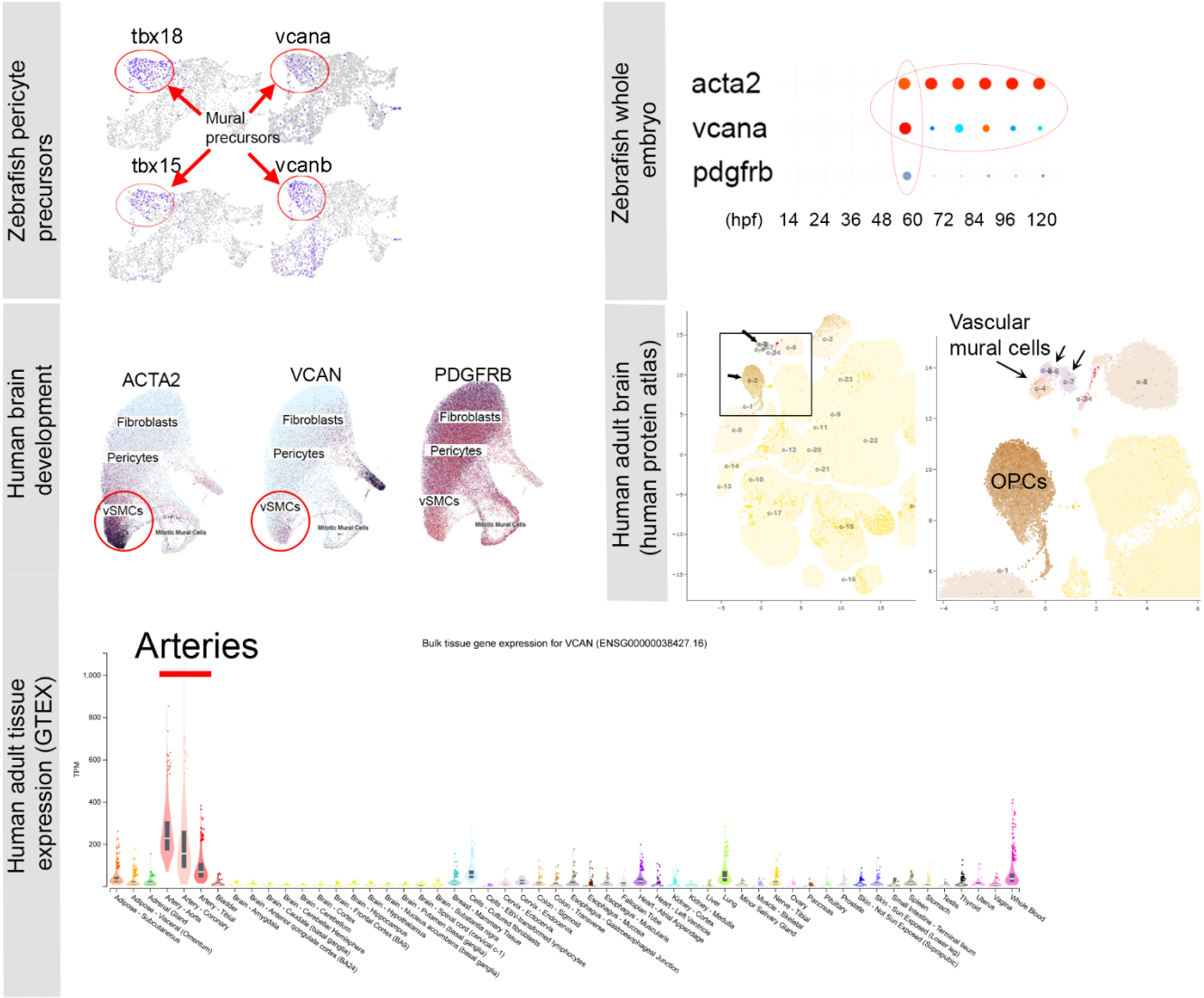


Figure 4: Expression of Versican mRNA in zebrafish and human

In zebrafish, *versican-a (vcana)* and *versican-b (vcanb)* are both expressed in FACs-sorted pericyte precursors(25). In zebrafish whole embryo, *vcana* is co-expressed in *acta2*-expressing vSMCs and 60 hpf *pdgfrb*-expressing pericytes(41). In human first trimester sorted vascular cells from the brain, *VCAN* is co-expressed with both *ACTA2* and *PDGFRB*(42). In the human adult brain, VCAN is strongly expressed in oligodendrocyte precursors (OPCs), and in vascular mural cells, especially fibroblasts(43). In human adult tissues from the GTEX database, VCAN is strongly expressed in the aorta and arteries(44).

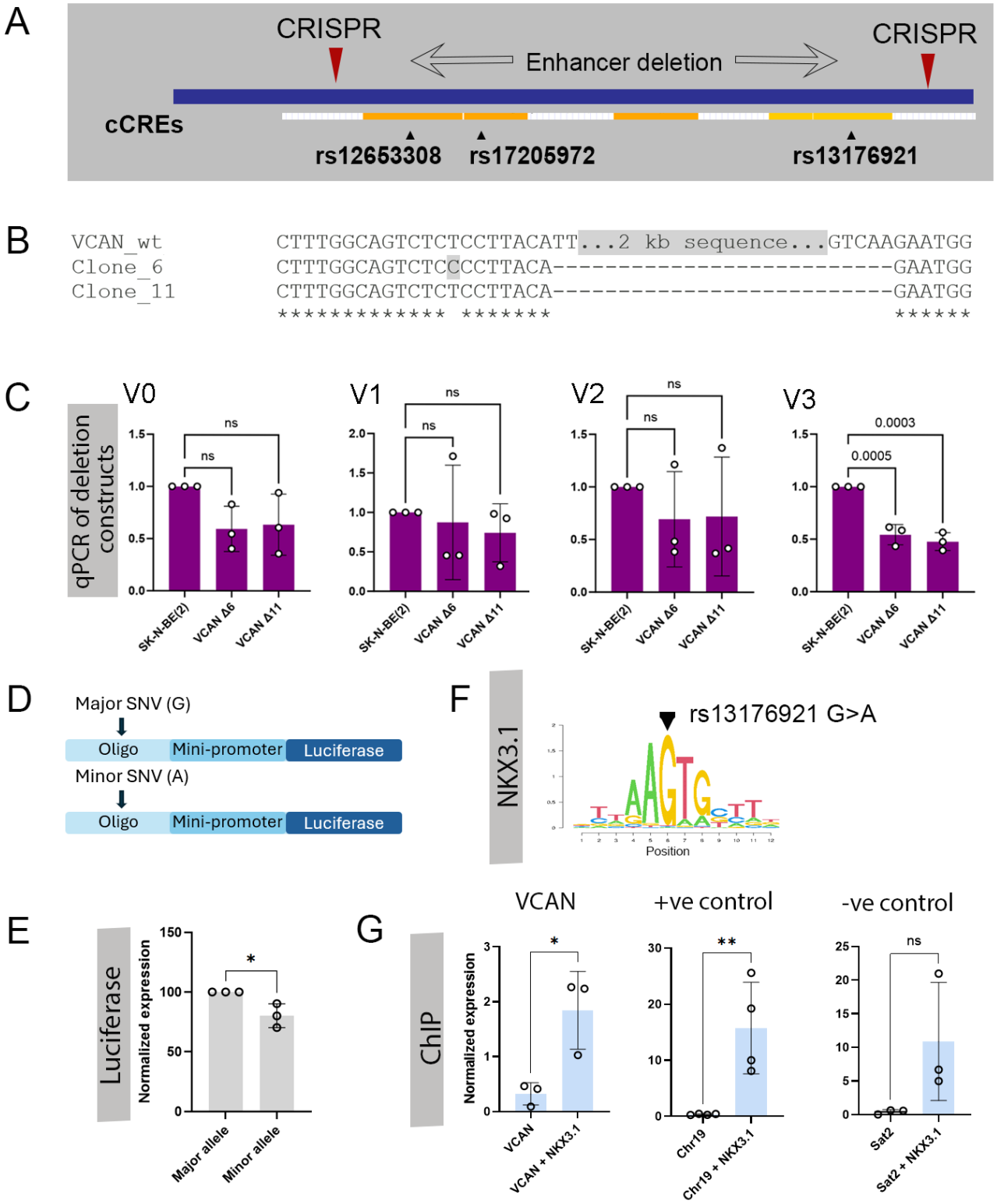


Figure 5: Modulation of a novel VCAN enhancer leads to decreased expression. NKX3.1 binds to the region around rs13176921. A) Schematic of where a 2 kb intronic enhancer region was deleted using two CRISPRs. B) Sequences of independent clones with the 2 kb region deleted. C) qPCR of endogenous VCAN in the deleted clones showing that VCAN V3 is significantly downregulated when the enhancer is deleted but there are no significant changes in V0, V1 and V2. D) Schematic of

constructs with the major and minor alleles of rs13176921 in a 300bp oligo. E) Luciferase activity is decreased when the minor allele is present, relative to the major allele. N=3 biological replicates, with 3 technical replicates. F) Position weight matrix for NKX3.1 from JASPAR showing location of rs13176921 SNV. G) ChIP for NKX3.1 in the region of rs13176921, a positive control genomic region known to bind NKX3.1 (Chr 19) and a negative control region (Sat2). Statistics used a One Way ANOVA with Dunnett's multiple comparison test, or the Student's t-test.