Genetic and epigenetic factors fine-tune \textit{TGFB1} expression within the osteoarthritic articular joint

Sarah J. Rice, PhD*, Jack B. Roberts, MRes, Maria Tselepi, MRes, Abby Brumwell, BSc, Julia Falk, MSc, Charlotte Steven, BSc, John Loughlin, PhD*

Biosciences Institute, Newcastle University, International Centre for Life, Newcastle upon Tyne, NE1 3BZ, United Kingdom.

sarah.rice@ncl.ac.uk
john.loughlin@ncl.ac.uk
Abstract

Objective: Osteoarthritis (OA) is an age-related disease characterised by articular cartilage degeneration. It has a large heritability and genetic screens have identified single nucleotide polymorphisms (SNPs) marking genomic risk loci. One such locus is marked by G>A SNP rs75621460, downstream of \( TGFB1 \). This gene encodes TGF-β1, the correct expression of which is essential for cartilage maintenance. We have used a combination of human patient samples (n=319) and a chondrocyte model to characterise the impact of rs75621460 in multiple articular joint tissues.

Methods: Patient samples were genotyped and DNA methylation (DNAm) levels quantified by pyrosequencing. Gene reporter and electrophoretic mobility shift assays were used to determine differential nuclear protein binding to the region. The functional impact of DNAm upon \( TGFB1 \) expression was tested using targeted epigenome editing.

Results: We identified that rs75621460 is located within a \( TGFB1 \) enhancer, and that the OA risk A-allele alters transcription factor binding, decreasing enhancer activity. Protein complexes binding to A (but not G) induced DNAm at flanking CG-dinucleotides. Strong correlations were observed between patient DNAm levels and \( TGFB1 \) expression, the direction of which was opposing between cartilage and synovium. This demonstrated biological pleiotropy in the impact of the SNP within different tissues of the articulating joint.

Conclusion: The OA risk SNP rs75621460 impacts \( TGFB1 \) expression by modulating the function of a gene enhancer. We propose a mechanism by which the SNP impacts enhancer function, providing novel biological insight into one mechanism of osteoarthritis genetic risk, which may facilitate the development of future pharmacological therapies.
Introduction

Transforming growth factor-β (TGF-β) signalling plays vital developmental and homeostatic roles in mammalian cell differentiation, proliferation, and extracellular matrix (ECM) production. TGF-β signalling is widespread in mammalian tissues and the effects can be cell-type specific, displaying distinct, and sometimes paradoxical, effects. There are three human TGF-β isoforms (TGFβ-1, 2, and 3), encoded by physically distinct genes (TGFB1, 2, and 3, respectively), which are differentially expressed. Missense mutations within all three genes are clinically relevant in skeletal and connective tissue disorders including otosclerosis (TGFB1), Camurati-Engelmann disease (TGFB1), Marfan and Loeys-Dietz syndrome (TGFB2 and TGFB3). A role for aberrant TGF-β signalling in common musculoskeletal pathologies is supported by the genetic association of single nucleotide polymorphisms (SNPs) residing at chr19p13.2, the genomic location of TGFB1, with a spectrum of phenotypes including osteophytosis, osteoporosis, cleft palate, and, most recently, osteoarthritis (OA).

OA is an age-related, degenerative disease of the articulating joints, affecting over 40 million Europeans. The disease hallmark is the thinning and loss of articular cartilage, often accompanied by a low-grade synovial inflammation within the affected joint. This leads to chronic impairment of joint function, with a resultant increased risk of premature death due to secondary co-morbidities. A typical clinical end-point is surgical replacement of the affected joint. Currently, there are no disease-modifying OA drugs and novel treatments are urgently required.

The causes of primary OA are complex. Yet, with an estimated heritability of ~50%, genetic influences contribute highly to disease susceptibility. Genome-wide association studies (GWAS) have revealed the highly polygenic nature of OA and, over 90 significant association signals have been reported. Risk variants are often intergenic and thought to operate by mediating differential expression of their target genes. This places OA in the “enhanceropathy” category of common diseases, in which subtle but detrimental changes in gene expression through aberrant activity of DNA regulatory elements, or “enhancers”, contribute to disease progression.

In 2019, an OA risk signal was reported at chr19q13.2, marked by intergenic SNP rs75621460 (G>A; minor allele frequency (MAF), 0.03). The SNP lies 2.4kb downstream of TGFB1 and has >99% probability of being the single causal variant at this locus. In this study, we investigate rs75621460 and the encompassing region of DNA for regulatory activity.
Furthermore, we quantify genetic variation and epigenetic modifications within the region and measure the impact upon expression of *TGFB1* in multiple human joint tissues.

**Methods**

**In silico analysis of the locus**

An *in-silico analysis* of the locus was performed using ROADMAP chromatin state data\(^{21}\), RNA-sequencing (RNA-seq) data from hip OA and neck-of-femur (NOF) fracture cartilage\(^{22}\), and ATAC-sequencing data from knee OA cartilage\(^{23}\). *P*-values for RNA-seq data were calculated using a Wald test within the DESeq2 package. The ROADMAP 18-state model utilises 6 histone post-translational modifications to assign one of eighteen chromatin states to cell-specific epigenomes and was used here to identify potential regulatory function in two cell types: E006, embryonic stem cell-derived mesenchymal stem cells (MSCs); and E049, bone marrow-derived cultured chondrocytes. Analysed knee articular cartilage ATAC-seq data was downloaded directly from GEO (accession GSE108301)\(^{23}\). Population allele frequencies of rs75621460 were taken from LDlink.

**Luciferase reporter analysis**

A 553bp region encompassing rs75621460 was amplified from pooled blood DNA, cloned into the pGL3-Basic firefly reporter vector (Promega), and sequenced to identify clones with the ancestral G- or derived A-allele at rs75621460. Tc28a2 immortalised chondrocytes were seeded into a 96-well plate 24h prior to transfection with the relevant pGL3-promoter luciferase vector construct (100ng) and pRL-TK Renilla vector (1.5ng) using Fugene HD transfection reagent (Promega). After 24h, cells were lysed and luciferase activity was measured by GloMax Navigator (Promega). For each well, luciferase activity was normalised to that of Renilla as previously described\(^{24}\).

**Electrophoretic Mobility Shift Assay (EMSA)**

Nuclear protein was extracted from Tc28a2 cells as previously described\(^{25}\). For each allele of rs75621460, forward and reverse single stranded DY682-labeled oligonucleotides (Eurofins), spanning 15bp each side of the SNP, and encompassing CpG2 were annealed to generate double-stranded probes (Table S1). Four probe combinations were generated containing either the G allele or A allele at rs75621460 that were unmethylated or methylated at CpG2. Reactions were carried...
out as previously described\textsuperscript{25,26}. For supershift assays, 2\(\mu\)g of the indicated antibody was added to the binding reaction (Table S2).

**CRISPR-Cas9**

The CHOPCHOP CRISPR Design Tool\textsuperscript{27} was used to design guide RNA (gRNA) sequences which were predicted to have low off-target effects, a GC content between 40 and 70\%, and with a high targeting efficiency immediately upstream (gRNA1) and downstream (gRNA2) of rs75621460. The selected gRNAs created an 84bp deletion encompassing rs75621460 (Table S3). Guide sequences were cloned into the CRISPR-Cas9 vector, pSpCas9n(BB)-2A-Puro (PX462), which was a gift from Feng Zhang (Addgene plasmid 62987)\textsuperscript{28}. Constructs were nucleofected into Tc28a2 chondrocytes in 6-well plates as previously described in detail\textsuperscript{24}. Cas9-expressing cells were selected using 1\(\mu\)g/ml puromycin for 24h. Post-selection, 5\(\times\)10\(^5\) cells were pelleted in 15ml tubes and cultured in chondrogenic medium consisting of high glucose Dulbecco's modified Eagle's medium (Lonza BE12-614, 4.5g/L glucose, SLS) containing, 10 ng/ml TGF-\(\beta\)3 (PeproTech), 100 nM dexamethasone, 40 \(\mu\)g/ml proline, 50 \(\mu\)g/ml ascorbate-2-phosphate (Sigma-Aldrich), 1\(\times\) ITS-1 premix (Corning), 2mM l-glutamine (Thermo-Fisher) and 100U/ml penicillin-streptomycin solution (Thermo-Fisher). After 14 days, nucleic acids were extracted using NucleoSpin TripPrep (Machery-Nagel). Deletion of the target region was confirmed by Sanger sequencing (Source Bioscience).

**Gene expression analysis**
cDNA was reverse transcribed from total RNA using the Superscript IV standard protocol (Invitrogen) after an initial 15-minute treatment with 1 unit of amplification grade DNaseI (Invitrogen). Gene expression was measured by quantitative reverse transcription–polymerase chain reaction (qRT-PCR) using pre-designed TaqMan assays (Integrated DNA Technologies). Gene expression was quantified using TaqMan chemistry, normalised to housekeeping genes 18\(S\), HPRT1 and GAPDH and expressed as 2\(^{-\Delta\text{ct}}\) as described previously\textsuperscript{29}.

**Patient samples and extraction of nucleic acids**

Human tissue samples were obtained from patients undergoing hip or knee joint replacement surgery due to end-stage OA or NOF fracture. Arthroplasty was conducted at the Newcastle upon Tyne NHS Foundation Trust hospitals. The Newcastle and North Tyneside Research Ethics
Committee granted ethical approval for the collection, with each donor providing verbal and written informed consent (REC reference number 14/NE/1212). Further details of the patient samples used in this project are provided in Table S4.

RNA was extracted from cartilage by TRIzol-chloroform (Life Technologies) separation, following which the RNA was purified from the aqueous phase using the RNeasy Mini Kit (Qiagen). Both DNA and RNA were extracted from whole blood and synovium using the EZNA DNA/RNA Isolation kit (Omega Bio-Tek). For genotyping DNA was used directly. For methylation analysis, 500ng DNA was bisulphite converted using the EZ DNA methylation kit (Zymo Research).

Pyrosequencing

PyroMark Q24 Advanced (Qiagen) was used to genotype all patient DNA samples as previously described\textsuperscript{24}. Pyrosequencing was also used to quantify DNAm at six CpGs flanking rs75612460 following bisulphite conversion of DNA (EZ DNA Methylation Kit, Zymo). Each sample was amplified in duplicate. Samples were excluded from the analysis if the replicates differed by $>5\%$. Assays were designed using PyroMark assay design software 2.0 and primer sequences are listed in Table S5.

Lucia Reporter Assay

A 546bp region containing either the G or the A allele of rs75621460 was amplified and cloned into the pCpG-free-promoter-Lucia vector (Invivogen). Primer sequences are listed in Table S6. Clones were transformed into competent GT115 cells (Invivogen) according to the manufacturer’s protocol. Plasmids were methylated or mock-methylated \textit{in vitro} using \textit{M.SsII} (NEB). Successful methylation was determined by digest with methylation-sensitive \textit{SmaI} (New England Biolabs). Tc28a2 cells were transfected with 100ng of pCpG-free-promoter construct, along with 10ng of the pGL3-promoter vector (Promega) and luminescence measurements were made as described above.

Epigenome modulation with dCas9

Guide RNA sequences (Table S3) were cloned into the DNMT3a-dCas9 vector (Addgene, 71666)\textsuperscript{30}. Following nucleofection, Tc28a2 monolayers were cultured for 24h and DNMT3a-dCas9 expression was confirmed by GFP expression. Cells were passaged twice at a density of 1:5.
and expanded each time to 90% confluency. At each passage, cells were isolated for extraction of nucleic acids (Purelink, Thermo Fisher).

Statistical analyses
Genotype and methylation correlations were calculated using Kruskal-Wallis testing. For Lucia reporter assays we corrected for multiple comparisons using the methods of Holm-Sidak or Dunn, as specified in the figure legends. Changes in gene expression following Cas9 modulation were calculated using paired t-tests. AEI and DNAm relationships were determined using linear regression analysis. The exact details of all statistical tests is provided in the relevant figure legend. All tests were performed in GraphPad Prism 8.3.1.

Results
The region encompassing rs75621460 is a gene enhancer
OA risk SNP rs756215460 is an intergenic variant at chromosome 19q12, positioned between CCDC97 and TGFB1 (Fig.1a). ChIP-seq data from mesenchymal stem cells (MSCs) and differentiated chondrocytes, along with ATAC-seq from OA knee chondrocytes indicate that the SNP resides within a chromatin-accessible region with post-translational histone modifications H3K27ac (yellow) and H3K27me3 (red), indicating that this region possesses regulatory function (Fig.1a). TGFB1 expression was significantly (P<0.01) upregulated in OA hip cartilage. No significant change (P>0.05) in CCDC97 expression was observed (Fig.1b).

We cloned the 550bp accessible chromatin region into a luciferase reporter vector. The ancestral G-allele construct conferred a 2.7-fold increase in luciferase activity (Fig.1c). The derived A-allele (OA-risk) also demonstrated regulatory activity (1.6-fold), which was significantly lower (P<0.05) than that of G.

A multiple sequence alignment revealed that the G-allele is highly conserved in mammals (Fig.1d). Within human populations, the A-allele emerged at a frequency >1% only in Europeans. The minor allele frequency (MAF) within European subpopulations ranged from 1.01% (FIN) to 4.04% (CEU) (Fig.1e).

Differential allelic protein binding occurs at rs75621460
We used electrophoretic mobility shift assays (EMSAs) to characterise proteins binding to rs75621460. This revealed several complexes with a greater binding affinity to the G-allele than A
(Fig.2a, arrows 1 and 4). Furthermore, proteins were identified which exclusively bound to one of the two alleles (Fig.2a arrows 2-3 and 5-6). Unlabelled probes were added to the reaction at increasing concentrations (Fig.2b and c). The unlabelled A-probe was unable to strongly compete for binding for the higher molecular weight complexes bound to the G-probe (Fig.2b). However, some lower molecular weight complexes were outcompeted by increasing concentrations of the unlabelled A-probe, indicated by two arrows on Figure 2b. The unlabelled G-probe was able to compete for binding of all protein complexes to the labelled A-probe, with only one exception, indicated with an arrow on Figure 2c. The TRANSFAC database predicted four transcription factors which differentially bind to the alleles of rs75621460: SP1, MAZ, KLF17, and ETF (Fig.2d). All four were predicted to bind exclusively to the G-allele. EMSA was performed using antibodies raised against the four proteins (Fig.2e). A supershifted band was observed in the presence of the SP1 antibody. This complex (indicated by an arrow) was bound to both alleles, however with a greater abundance at the G probe (Fig.2e). These combined EMSA results indicate that the G-allele binds proteins with greater affinity than the A-allele, and that distinct protein complexes bind to the region in chondrocytes, determined by the allele carried at rs75621460.

**TGFB1 is the gene target of the rs75621460 enhancer**

We deleted an 84bp region of the enhancer encompassing rs75621460 from the genome of Tc28a2 immortalised chondrocytes using CRISPR-Cas9 and a pair of gRNAs (gRNA 1 and 2) (Fig.3a). No change in CCDC97 expression was measured (P=0.12) following deletion of the region (Fig.3b). However, a significant decrease in TGFB1 expression was observed in Tc28a2-Δ84, in which mean expression was 0.48 of that measured in wild-type cells (P=0.003).

**Methylation quantitative trait locus (mQTL) analysis of rs75621460**

The deletion introduced in Tc28a2-Δ84 cells encompassed six CG-dinucleotides (CpGs), positions at which eukaryotic DNA can be methylated. This included a single upstream CpG (CpG1), and five downstream CpGs (CpG2-6) (Fig.4a). We investigated whether DNAm at these CpGs was modulated by SNP genotype. Due to the low MAF at rs75621460, we screened 206 human hip and knee cartilage samples to identify sufficient individuals carrying the A allele for analysis and identified 190 major allele homozygotes (GG), and 16 heterozygotes (GA).

We quantified cartilage DNAm at the 6 CpGs and stratified values by SNP genotype. All homozygous (GG) individuals investigated (n=93-101 across the six CpGs) were hypomethylated
(DNAm <10%) at the region. Significant correlations, marking methylation quantitative trait loci (mQTLs), were identified at all CpGs. Upstream of the SNP (CpG1), the difference in median DNAm was small, yet significant (0.8%, P=8.0x10^{-5}). However, at the downstream CpGs, genotype had a much greater influence upon the DNAm values: CpG2, 14.9% increase (P=1.3x10^{-20}); CpG3, 5.8% (P=9.5x10^{-18}); CpG4, 4.1% (P=1.2x10^{-16}); CpG5, 5.0% (P=1.2x10^{-18}); CpG6, 5.3% (P=9.0x10^{-19}) (Fig.4b). Amongst heterozygous patients, mean DNAm was higher in knee samples than in hip at all CpGs, significantly so at CpG2 (P=0.02), CpG5 (P=0.005), and CpG6 (P=0.003) (Fig.4c). Only two of the heterozygous hip cartilage samples were NOF controls, therefore it was not possible to investigate relationships between disease status and DNAm.

We analysed DNA from knee synovium (n=55-61) to test for mQTLs in a distinct joint tissue (Fig.4d). Mean DNAm was higher in synovium than in cartilage (27.2%, compared to 5.6% at CpG2). Significant mQTLs (P<0.0001) were identified at all six CpGs in synovium (Fig.4d). A systemic effect was investigated by analysing whole blood samples but no significant mQTLs (P=0.14 – 0.39) were identified (Fig.4e).

**Methylation and expression correlations (meQTLs) are present in heterozygote patients**

We tested whether rs75621460 genotype and TGFB1 expression correlated in cartilage (n=31) and synovium (n=28) (Fig.S1). No significant expression QTLs (eQTLs) were observed in either tissue (P=0.45-0.53).

In samples for which both DNA and RNA were available, we tested for correlations between DNAm and TGFB1 expression (methylation and expression QTLs, meQTLs). In cartilage, data were analysed together (n=31), and also by joint site (hip, n=14; knee, n=17). Across both tissues, homozygous patient samples (GG) showed no significant meQTLs (P>0.09, Fig.4f). Conversely, very strong correlations were observed amongst the heterozygote samples. In cartilage, this was dependent upon the joint site from which cartilage was taken, with stronger meQTLs measured in knee (r^2=0.47 - 0.99) than in hip (r^2=0.01 - 0.65) (Fig. S2a). In knee cartilage and synovium, the strongest effect was observed for CpG1 (upstream of the SNP), where r^2 values were 0.99 and 0.90, respectively (Fig.4g). In both knee tissues, increasing DNAm at CpG1 tightly correlated with decreasing gene expression across a very narrow range of methylation values <2.7% (Figs.4f and S2a).

Correlations between DNAm and TGFB1 expression were observed at the five downstream CpGs (Fig.4g). In knee cartilage, a very strong meQTL (r^2=0.92) operated at CpG2
(Fig. S2a). Strikingly, the impact of rs75621460 upon DNAm at the downstream CpGs was paradoxical in the two distinct knee joint tissues. In cartilage, increasing DNAm correlated with increased TGFβ1, whereas in synovium the opposite effect was seen (Fig.4f).

Heterozygous DNAm at CpG2 was stratified by DNAm at CpG1 (to identify correlations between CpGs upstream and downstream of the SNP) and at CpG3 (to identify correlations between CpGs downstream of the SNP). In synovium, positive correlations were observed between CpG1 and CpG2 ($r^2=0.84$) and between CpG2 and CpG3 ($r^2=0.97$) (Fig.S2b). However, in cartilage, correlations were observed at the downstream CpGs ($r^2=0.92-0.94$), but not between CpG1 and CpG2 ($r^2=0.04-0.31$), which are physically separated by rs75621460 (Fig.S2c). This validates the observations made in the meQTL analysis and suggests that in cartilage, distinct mechanisms regulate DNAm upstream and downstream of the SNP.

The detected meQTLs were the strongest in cartilage, the central tissue in OA pathogenesis. We therefore continued to use a chondrocyte model for subsequent downstream analyses.

**DNA methylation in the enhancer attenuates activity**

We next investigated whether DNAm at the CpGs flanking rs75621460 have a functional impact upon enhancer activity. The enhancer was cloned into a CpG-free reporter vector and expressed in Tc28a2 cells in either an unmethylated or methylated state. Methylation of the cloned region resulted in a significant reduction in enhancer activity in constructs containing both the G- (P=0.004) and A-allele (P=0.019) (Fig.5a), demonstrating that DNAm influences chondrocyte enhancer activity independently of rs75621460 genotype.

We repeated the EMSA, this time including probes that were methylated at CpG2, the sole CpG contained within the probe sequence. We compared nuclear protein binding to both alleles in the unmethylated or methylated state. The six bands of interest that were previously identified (Fig.2a) are highlighted (Fig.5b). All of these protein complexes were able to bind to methylated probes (Fig.5b). Interestingly, for bands 1 and 4, methylation of the A-probe appeared to recover protein binding (Fig.5b).

We conducted a supershift assay using the methylated probes and also included an antibody for DNMT3a with the aim of detecting recruitment of a de novo DNA methylating enzyme by proteins bound to the A-allele. However, the only visible shift identified using this panel of antibodies was in the lanes containing anti-SP1, which was able to bind to both the
unmethylated and methylated probes (Fig.5c and d). These EMSA data, along with our reporter assay data indicate that methylation of the region attenuates activity of the enhancer. However, DNAm at the single most proximal CpG to rs75621460 (CpG2) does not prevent the binding of proteins adjacent to the SNP, including SP1.

Modulation of the epigenome using DNMT3a-dCas9

Finally, we investigated whether DNAm flanking the SNP could functionally impact TGFB1 expression in the absence of the derived A-allele. We used a DNMT3a-dCas9 fusion protein for targeted editing of DNAm at the six CpGs in Tc28a2 cells, which are homozygous (GG) at rs75621460. Five gRNAs (gRNA3-7) were designed to target the region (Fig.S3a). DNMT3a-dCas9 was expressed alone (non-targeting control), or along with one of the five gRNAs, and DNAm was quantified over three cell passages (Fig.S3b). Four of the five guides (gRNAs4-7) successfully increased DNAm at one or more CpGs, an effect which was lost passively (Fig.S3b). As gRNA3 did not modulate DNAm at any of the targeted six CpGs, it was not included in subsequent experiments.

The four gRNAs were individually co-expressed with DNMT3a-dCas9 in Tc28a2 cells (Fig.6a). Additionally, two combinations of gRNA pairs were used: gRNAs 4 and 6, and gRNAs 5 and 7. Targeted editing of DNAm with 3/4 single gRNAs significantly decreased TGFB1 expression: gRNA4 (0.80-fold, P=0.039), gRNA6 (0.77-fold, P=0.019), and gRNA7 (0.80-fold, P=0.004) (Fig.6b). gRNA5 increased DNAm at CpG 1 (5.5%), CpG2 (8.3%), and CpG6 (30.7% increase), but methylation at these CpGs alone was not sufficient to significantly alter TGFB1 expression (P=0.203) (Fig.6b). The use of the gRNAs 4, 6, and 7 alone and in combinations all successfully edited DNAm at one or more of CpGs 3, 4, and 5, significantly decreasing TGFB1 expression (Fig.6b). This suggests that increased DNAm at any of these three CpGs can impact binding of proteins 20-43bp downstream of the SNP, further modulating TGFB1 expression beyond the effects conferred by genotype alone.

Discussion

TGF-β1 has a well-established role in OA pathophysiology, however this is the first study to identify an interplay between genetic and epigenetic regulation of TGFB1 expression in the context of disease risk. We have characterised an intergenic TGFB1 enhancer within the
articulating joint, at which the alleles of an OA risk SNP impact upon DNAm and regulate \textit{TGFB1} \textit{in vivo}.

We confirmed the SNP region as an \textit{in vitro} enhancer at which the rs75621460 OA-risk A-allele reduces enhancer activity compared to the highly-conserved ancestral G-allele. The conservation of the G-allele amongst distinct human populations and throughout mammalian evolution illustrates the importance of the G for protein binding and enhancer function. EMSA analysis supported this data, showing that different alleles at rs75621460 could bind distinct proteins. The emergence of the A-allele in European populations implicates a selection advantage resulting from population-specific pressure, yet that this selection also simultaneously confers a detriment to cartilage health in older age, a phenomenon known as antagonistic pleiotropy. Additionally, we identified that the transcription factor SP1, which has previously been shown to play a role in \textit{TGFB1} regulation\textsuperscript{31}, binds to complexes at both alleles. Deletion of the region in chondrocytes confirmed \textit{TGFB1} as the enhancer gene target.

The absence of eQTLs in patient samples was perhaps unsurprising due to our modest sample size. Interindividual variability in gene expression often necessitates sample sizes involving hundreds of patients for the detection of significant genotype-expression correlations\textsuperscript{32,33}. A complementary approach for eQTL analysis, which greatly increases sensitivity, involves measuring allelic imbalance between the expression of gene transcripts, and has been widely applied to investigations of OA risk loci\textsuperscript{24,34-46}. We were unable to utilise this approach here due to the low MAF and the absence of a suitable \textit{TGFB1} transcript SNP. However, we have demonstrated how the use of a secondary endophenotype, DNAm, can provide a more sensitive approach to investigate the impact of genotype upon gene expression within an individual.

We identified mQTLs at six CpGs in two tissues of the articulating joint, indicating that genetic and epigenetic interplay at the locus contributes to disease aetiology as observed at other OA risk loci\textsuperscript{37}. The very small range of DNAm values over which correlations with \textit{TGFB1} expression occur potentially suggest the effects operate in a sub-population of chondrocytes within the tissue. \textit{In vitro} methylation of the enhancer reduced the activity of both alleles in a reporter assay, and EMSA analysis indicated that DNAm at CpG2 impacted upon protein binding. The SP1 antibody bound to both methylated and unmethylated probes, consistent with previous reports into binding of this transcription factor\textsuperscript{38,39}. We further identified that a targeted increase of DNAm at CpGs3-6 could reduce \textit{TGFB1} expression in the absence of the rs75621460 A-allele.
It has previously been documented that regulatory SNPs can confer tissue-specific effects upon genes, resulting in biological pleiotropy. At this locus, the directly opposing effects in cartilage and synovium are the result of a shared effect of a single variant rather than the colocalisation of two distinct effects. This emphasises that whilst integration of epigenetic data is a useful post-GWAS tool, functional analyses in appropriate disease models are imperative to elucidate tissue-specific pathological mechanisms.

We propose a molecular mechanism of \textit{TGFB1} regulation in cartilage, as follows (Fig. S4). Substitution of the highly-conserved G-allele at rs75621460 alters the consensus sequence for protein binding. In the presence of the G-allele, a protein complex with strong transcriptional activity binds to the sequence (Fig. S4a). This complex does not modulate DNAm, hence there is no correlation between methylation and gene expression. Additional transcription factors bind downstream, further enhancing \textit{TGFB1} expression. A distinct protein complex binds to the A-allele (both complexes share SP1), which confers lower levels of transcriptional activation (Fig. S4b). This complex recruits modifiers of DNAm, namely DNMTs, which independently methylate CpGs up- and downstream of the SNP. Methylation of downstream CpGs prevents binding of the downstream proteins, further suppressing expression. Patients with the A-allele therefore have higher levels of DNAm at flanking CpGs, accompanied by lower levels of \textit{TGFB1} expression. However, since the protein complex binding to the A-allele does enhance \textit{TGFB1} expression, albeit at a lower level than the G-allele complex, and also induces methylation, there is a positive correlation between DNAm and \textit{TGFB1} expression.

We speculate that in synovium, where a paradoxical correlation was observed, tissue-specific proteins which have a repressive effect upon \textit{TGFB1} could bind to the A-allele (Fig. S4c). In both tissues, the OA risk A-allele results in attenuated enhancer activity, and decreased \textit{TGFB1} expression.

Elucidating the mechanism of \textit{TGFB1} expression in synovium was not within the scope of this study and requires further investigation. Furthermore, the impact of the SNP upon downstream TGF-β signalling remains unknown. The SNP resides within a region of open chromatin in fibroblast-like synoviocytes. This knowledge, along with our data, indicates that the region is also utilised to regulate \textit{TGFB1} in synovium. The \textit{TGFB1} enhancer is an interesting focus for future studies, especially in the context of inflammatory joint diseases such as rheumatoid arthritis. Additionally, the use of single-cell technologies could identify subpopulations of cells within joint tissues in which these mechanisms operate. Furthermore,
novel techniques for targeted subnuclear proteomics profiling provide a promising tool to identify the exact proteins modulating $TGF\beta 1$ expression in distinct tissue types$^{43}$.

TGF-β has been well-studied in the context of OA pathophysiology$^{44,45}$. In healthy cartilage, TGF-β acts as an anabolic factor to stimulate the synthesis of ECM proteins, conveying a chondroprotective effect against mechanical loading in a healthy joint$^{46,47}$. Active TGF-β1 and subsequent downstream signalling prevents hypertrophic differentiation of chondrocytes and tissue degeneration$^{48,49}$. TGF-β expression decreases with age, exposing chondrocytes to ECM degradation by catabolic factors such as interleukin-1 (IL-1) and matrix metalloproteases (MMPs)$^{50}$. However, it has been shown that supplementation of an OA joint with TGF-β can lead to chondrocyte hypertrophy and synovial fibrosis$^{51}$. It is therefore highly plausible that a genetic deficit in $TGF\beta 1$ expression conferred during development and younger adulthood by the A-allele at rs75621460, although relatively small, could lead to a breakdown in cartilage integrity over time. Consequently, TGF-β is a promising target for therapeutic intervention in OA$^{52}$.

The use of personalised therapeutics to treat OA is not yet a reality$^{53}$. For this advancement, complete understanding of the molecular mechanisms contributing to pathogenic subtypes is required. A precision medicine approach in OA demands a deeper understanding between genetics and disease, and aetiology-based classifications$^{54}$. In OA, there is increasing evidence for distinct pathways which determine disease subtypes, ultimately presenting with the endotype of cartilage loss, yet requiring diverse therapeutics$^{55}$. The investigation of compounds which can regulate TGFβ-1 are a promising first-step into disease-modifying treatments for OA.

**Acknowledgements**

Patient tissue was provided by the Newcastle Bone and Joint Biobank supported by the NIHR Newcastle Biomedical Research Centre (BRC) awarded to the Newcastle upon Tyne NHS Foundation Trust and Newcastle University. We thank the surgeons and research nurses at the Trust for providing us with access to these samples.

**Conflict of interest**

The authors have no conflicts to declare.

**Funding**

This article is protected by copyright. All rights reserved
This work was supported by Versus Arthritis (grant 20771), by the Medical Research Council and Arthritis Research UK as part of the MRC-Arthritis Research UK Centre for Integrated Research into Musculoskeletal Ageing (CIMA, grant JXR 10641), by The John George William Patterson (JGWP) Foundation and by the European Union’s Seventh Framework Program for research, technological development and demonstration under grant agreement number No. 305815 (D-BOARD). Funds to purchase PyroMark Q24 Advanced were provided by The Ruth and Lionel Jacobson Charitable Trust.
References


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved
Fig. 1. The functional polymorphism rs75621460 falls within a TGFB1 enhancer.

(A) Position of rs75621460 on chromosome 19 (red). A schematic diagram CCDC97 and TGFB1, along with chromatin state data (ROADMAP epigenome database) from mesenchymal stem cells (E006) and cultured chondrocytes (E049): red, transcription start site; green, transcription; orange, active enhancer. Chromatin accessibility in human knee chondrocytes is represented by ATAC-seq peaks.

(B) Expression of CCDC97 and TGFB1 in cartilage from OA and NOF fracture patients. TPM, transcripts per million. Bars represent mean (±SEM).

(C) Luciferase reporter analysis in Tc28a2 chondrocytes. Mean luciferase activity (n=6) of the G- or A-allele normalised against a control vector. P-values were calculated using a Wilcoxon matched-pairs signed rank test. *, P=0.03.

(D) Position of rs75621460 (highlighted red) within a 30-vertebrate Multiz alignment (UCSC Genome Browser, hg19). A derived human A-allele occurs at an otherwise highly conserved ancestral G base in the region.

(E) Percentage frequency of the A-allele at rs75621460 in European subpopulations taken from (LDlink). The allele counts are listed next to their representative bar. FIN, Finnish in Finland; TSI, Toscani in Italy; GBR, British in England and Scotland; IBS, Iberian in Spain; CEU, Utah residents from North and West Europe.

Fig. 2. Electrophoretic mobility shift assay (EMSA) analyses of rs75621460 in Tc28a2 cells.

(A) Labelled probes containing rs75621460 with the G or A-allele were incubated with nuclear protein extract from Tc28a2 chondrocytes. Arrows indicate complexes which visibly bind more strongly to the G-allele (1 and 4) or those which exclusively bind to the G-allele (2 and 3) or the A-allele (5 and 6).

(B) Increasing concentrations of unlabelled G-allele and the A-allele competitor were added to the EMSA reactions containing cell nuclear protein extract and a fluorescent G-allele probe. Lower molecular weight complexes that were outcompeted by increasing concentrations of the unlabelled A-probe are labelled with arrows.

(C) Increasing concentrations of unlabelled G-allele and A-allele competitor were added to the EMSA reactions containing cell nuclear protein extract and a fluorescent A-allele probe. One complex was not outcompeted by the unlabelled G-probe (indicated with an arrow).
(D) Analysis of differential transcription factor binding to the G and A-allele at rs75621460 using the transcription factor database, TRANSFAC.

(E) Supershift experiment with antibodies targeting SP1, MAZ, KLF17, and ETF compared to no antibody (control) to the EMSA reaction containing the G or A-allele probe. The arrow indicates the position of supershifted complexes.

Fig. 3. CRISPR-Cas9 deletion of rs75621460 in Tc28a2 chondrocytes reduces TGFB1 expression.

(A) The 84bp deletion (Δ84) introduced into the cells using a pair of gRNAs (gRNA 1 and 2) was confirmed by Sanger sequencing (upper) and compared to the wild-type sequence (lower). The deleted region is highlighted in yellow. The positions of the two gRNAs are indicated with arrows.

(B) Gene expression of CCDC97 and TGFB1 in Tc28a2-Δ84 cells normalised to the mean of the wild-type controls. P values were calculated using a paired t-test, n=6 biological replicates, each consisting of three technical repeats.

Fig. 4. Analysis of DNA methylation and TGFB1 expression in patients.

(A) Schematic diagram of the region surrounding rs75621460

(B) Hip and knee cartilage DNAm stratified by rs75621460 genotype. GG, n=93-116; GA, n=16. Horizontal lines, median and interquartile range. P-values were calculated using Mann-Whitney testing. *, P<0.05; **, P<0.01; ****, P<0.0001.

(C) Cartilage DNAm stratified by genotype and joint site. Yellow, knee; grey, hip. Adjusted P-values were calculated using multiple t-tests and corrected using the method of Holm-Sidak.

(D) Knee synovium DNAm stratified by genotype. GG, n=47; GA, n=8. Horizontal lines, median and interquartile range. P-values were calculated using Mann-Whitney testing.

(E) Blood DNAm stratified by genotype. GG, n=18; GA, n=5. Horizontal lines, median and interquartile range. P-values were calculated using a Mann-Whitney testing. All P-values >0.05.

(F) DNAm in hip (n=14) and knee (n=17) cartilage samples (GG, n=23, GA, n=8), and in knee synovium (GG, n=19, GA, n=7) stratified by TGFB1 expression, expressed as 2^Δct. Correlations were calculated using simple linear regression. r² values are displayed along with the line of best fit where r²>0.5.

(G) Simple linear regression r² of DNAm and TGFB1 expression in cartilage (yellow, GG, n=23;
GA, n=8) and synovium (pink, GG, 19; GA, n=7).

**Fig. 5. DNA methylation surrounding rs75621460 affects enhancer activity in vivo.**

(A) Lucia reporter assay in Tc28a2 chondrocytes. Mean Lucia activity (n=12) of the enhancer containing either the unmethylated or methylated G- or A-allele normalised against a unmethylated or methylated control vector. Horizontal lines indicate the median and interquartile range, whiskers represent the maximum and minimum. Adjusted $P$-values were calculated using multiple t-tests with Holm-Sidak correction. **, $P=0.004$; *, $P=0.018$.

(B) Labelled probes containing rs75621460 with the G-allele or A-allele and with CpG2 unmethylated (open circle), or methylated (closed circle) were incubated with Tc28a2 nuclear protein extract. Arrows indicate complexes which visibly bind more strongly to the G-allele (1 and 4) or those which exclusively bind to the G-allele (2 and 3) or the A-allele (5 and 6).

(C) Labelled probes containing rs75621460 G-allele with CpG2 unmethylated or methylated were incubated with antibodies raised against SP1, MAZ, KLF17, ETF, or DNMT3a (all 2μg) and Tc28a2 nuclear protein extract. The arrow indicates the SP1-supershifted complex.

(D) Labelled probes containing rs75621460 A-allele with CpG2 unmethylated or methylated were incubated with antibodies raised against SP1, MAZ, KLF17, ETF, or DNMT3a (all 2μg) and Tc28a2 nuclear protein extract. The arrow indicates the SP1-supershifted complex.

**Fig. 6. DNMT3a-dCas9 methylation of the enhancer in Tc28a2 cells.**

(A) Schematic diagram showing the structure of the DNMT3a-dCas9 construct used in this investigation, and the physical distance of the 6 targeted CpGs from the SNP (rs75621460) along with the position of the 4 gRNAs used for modulating the epigenome: gRNA 4 (orange), gRNA 5 (purple); gRNA 6 (green); gRNA 7 (maroon).

(B) Tc28a2 percentage methylation levels at the 6 CpGs surrounding the SNP in no guide controls (black, dashed line) or using a gRNA targeting the region (colours as described above, solid line) (n=6 biological replicates, each with two technical repeats). Combinations of gRNAs were also used: gRNA 4+6 and gRNA 5+7 (light blue, solid line). Subsequent changes in $TGFBI$ expression were also measured following targeted editing of methylation, and gene expression was normalised to that in no-guide controls (n=6 biological repeats, each with 3 technical repeats). $P$ values were calculated using paired t-tests following testing of control values for normality (D’Agostino and Pearson, $P=0.37$).
Supplementary Figure 1. Expression quantitative trait locus (eQTL) analysis in patient cartilage and synovium samples.

Normalised *TGFB1* gene expression in cartilage (yellow) and synovium (pink) stratified by genotype at rs75621460. Horizontal lines in the box indicate the median and interquartile range, and whiskers represent the maximum and minimum, P values were calculated using a Mann-Whitney test. No significant differences were observed.

Supplementary Figure 2. Correlations between DNA methylation values at neighbouring CpGs.

(A) DNA methylation values measured in hip (GG, black, n=10; GA, grey, n=4) and knee (GG, black, n=13; GA, yellow, n=4) cartilage samples. Values are displayed stratified by *TGFB1* expression, expressed as $2^{-\Delta\text{ct}}$. Correlations were calculated using a simple linear regression analysis and $r^2$ values are displayed along with the line of best fit where $r^2>0.5$.

(B) Synovium DNA methylation values in heterozygous patients at CpG2 stratified by DNA methylation at CpG1 (upstream of the SNP, $r^2=0.84$) and at CpG3 (downstream of the SNP, $r^2=0.97$). Statistical analysis was performed using simple linear regression (n=8).

(C) OA hip (grey), NOF (green), and OA knee (yellow) DNA methylation values in heterozygous patients at CpG2 stratified by DNA methylation at CpG1 (H, $r^2=0.31$; K, $r^2=0.04$) and at CpG3 (H, $r^2=0.94$; K, $r^2=0.92$). H, OA hip (red); K, OA knee (blue); NOF (green). Statistical analysis was performed using simple linear regression analysis (n=8).

Supplementary Figure 3. DNMT3a-dCas9 editing of methylation in Tc28a2 cells is lost passively through cell division.

(A) Schematic diagram showing the structure of the DNMT3a-dCas9 construct used in this investigation, and the physical distance of the 6 targeted CpGs from the SNP (rs75621460) along with the position of the 5 gRNAs screened for use in epigenome modulation: gRNA 3 (dark blue), gRNA 4 (orange), gRNA 5 (purple); gRNA 6 (green); gRNA 7 (maroon).

(B) Tc28a2 percentage methylation levels at the 6 CpGs surrounding the SNP in no guide controls (black) or using a gRNA targeting the region (colours as described above) (n=4 biological replicates, each consisting of 2 technical replicates). Measurements of methylation were made 24h
post-editing (P0), and at two subsequent passages (P1 and P2, respectively).

**Supplementary Figure 4. Proposed mechanism of TGFβ1 regulation by rs75621460.**

(A) A complex of proteins, which includes SP1, is able to bind to the highly conserved motif encompassing the SNP. This motif includes the G-allele at rs75621460. DNA in the region is not methylated (open circles at CpGs1-6), and additional proteins bind downstream (at CpGs3-5), which further enhances TGFβ1 expression. No correlation exists between cartilage DNAm and expression of TGFβ1.

(B) A distinct complex of proteins, which also includes SP1, is able to bind to the A-allele at rs75621460. This complex has weak transcriptional activation potential and recruits multiple enzymes which can methylate DNA, DNA methyltransferases (DNMTs). DNA in the region is methylated (closed circles at CpGs1-6), and as a result, additional proteins cannot bind downstream. Within cartilage, the more copies of the protein complex which bind to DNA, the higher tissue-wide DNAm and expression of TGFβ1 become. As a result, a positive correlation is seen between DNAm and expression of TGFβ1.

(C) In synovium a repressor complex of proteins binds to the A-allele at rs75621460. This complex recruits DNMT, leading to methylation of flanking CpGs (closed circles at CpGs1-6). Within the synovium, the more copies of the protein complex which bind to DNA, the higher tissue-wide DNAm and lower expression of TGFβ1 become. As a result, a negative correlation is seen between DNAm and expression of TGFβ1.
A

B

CCDC97

TGFB1

P=0.003