Genetic and functional studies of osteoarthritis susceptibility at *COL11A1* and *GDF5*

Andrew William Dodd

Doctor of Philosophy

Institute of Cellular Medicine,
Newcastle University

August 2013

I wish to dedicate this thesis to my parents, William and Patricia Dodd. I thank my mother for all the love and affection she has given me throughout my 27 years and I thank my father for being the absolute inspiration in my life.

This has all been for you two.

Abstract

Genetic and functional studies of osteoarthritis susceptibility at *COL11A1* and *GDF5*

Osteoarthritis (OA) is the most common musculoskeletal disease and is characterised by joint pain and dysfunction resulting from the progressive focal loss of the articular cartilage of the joint. It is a multifactorial disease arising from the interplay between genetic and environmental risk factors, with a continuous distribution between the two extremes of predominantly genetic and predominantly environmental. Whilst several environmental risk factors for the disease are known, including body mass index (BMI), injury and occupation, known genetic risk factors are less well understood. Two major strategies have been employed in order to identify genetic factors conferring OA susceptibility: the investigation of candidate genes, which are principally chosen on the basis that the proteins that they code for are known to have a role in joint formation and maintenance, and genome wide association scans (GWASs), which search agnostically for disease associated polymorphic DNA variants. Within this thesis I report research on two genes highlighted using both strategies, although each is a compelling candidate. I report on the investigation of the functional effects of common polymorphism within *COL11A1*, which harbours a single nucleotide polymorphism (SNP) that showed evidence of association to OA via the arcOGEN GWAS. I also report on the search for and analysis of rare variants of *GDF5*, a gene previously discovered as harbouring OA susceptibility by candidate gene studies. I tested for an allelic expression imbalance (AEI) of *COL11A1* using RNA from patient cartilage but I did not detect a correlation between genotype at the GWAS OA associated SNP rs2615977 and the expression of the gene. I did however discover that genotype at a second polymorphism within *COL11A1*, rs1676486, did correlate with *COL11A1* AEI. rs1676486 has previously been reported to be associated with lumbar disc herniation. However, my analysis of the arcOGEN dataset revealed that this SNP is not associated with OA. To identify rare variants in *GDF5* that could impact upon OA susceptibility I sequenced the protein coding region of the gene, its untranslated regions, exon-intron boundaries and its proximal promoter in 962 OA cases and controls. Six novel and very rare variants with minor allele frequencies (MAFs) of ≤0.0006 were discovered and I confirmed the existence of known variants with common MAFs. The absence of variants with intermediate MAFs implies that the gene may have been subjected to a genetic bottleneck. One rare variant, within the proximal promoter of *GDF5*, was carried forward for functional analysis using luciferase reporter assays. I discovered that the novel A-allele of this variant increased the expression of the reporter plasmid relative to its common C-allele. I also demonstrated that this A-allele is able to counteract the reduced gene expression mediated by the T allele of the previously reported OA associated *GDF5* SNP rs143383. I subsequently performed electrophoretic mobility shift assays (EMSAs) and identified the transcriptional activator/repressor YY1 as the *trans*-acting factor binding differentially to the alleles of the variant. Overall, this thesis demonstrates that the OA association to *COL11A1* identified by GWAS is not caused by AEI of that gene within the articular cartilage of OA patients, and that the deep sequencing of current OA susceptibility loci to identify novel risk alleles is also a means to identify mechanisms to counteract the effects of susceptibility alleles that are already known.

Contents

Abbreviations

AEI	Allelic Expression Imbalance
APS	Ammonium Persulfate
arcOGEN	Arthritis Research Campaign Osteoarthritis Genetics (Funded by Arthritis UK)
ASP	Affected Sibling Pair
BMP	Bone Morphogenetic Protein
BSA	Bovine Serum Albumin
cDNA	Complementary Deoxribose Nucleic Acid
Ct	Cycle Threshold
DEPC	Diethylpyrocarbonate
DMEM	Dulbecco's Modified Eagle's Medium
DNA	Deoxribose Nucleic Acid
DTT	Dithiothreitol
DZ	Dizygotic
ECM	Extracellular Matrix
EMSA	Electrophoretic Mobility Shift Assay
EXO	Exonuclease
FBS	Foetal Bovine Serum
GDF5	Growth/Differentiation Factor 5
GWAS	Genome Wide Association Scan
LAR II	Luciferase Activating Reagent II
LD	Linkage Disequilibrium
LDH	Lumbar Disc Herniation
MAF	Minor Allele Frequency
mRNA	Messenger Ribose Nucleic Acid
MMPs	Matrix Metalloproteinases
MZ	Monozygotic
NP-40	Nonyl Phenoxypolyethoxylethanol
OA	Osteoarthritis
PBS	Phosphate-Buffered Saline
PCR	Polymerase Chain Reaction
qPCR	Quantitative PCR
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribose Nucleic Acid
RT	Reverse Transcription
SAP	Shrimp Alkaline Phosphatase
SNP	Single Nuclear Polymorphism
TBE	Tris-Borate EDTA
TGF	Transforming Growth Factor
THR	Total Hip Replacement
TJR	Total Joint Replacement
TKR	Total Knee Replacement
UTR	Untranslated Region

Chapter 1: Introduction

1.1 Osteoarthritis

Osteoarthritis (OA) is the most common musculoskeletal disease, affecting a large proportion of people above the age of 65, regardless of ethnicity, sex or geographic location (1, 2). Buckwalter 2004 (2) claim that in some populations more than 75% and as high as 90% of over 65s have OA in one or more of their joints. The economic cost of OA has been difficult to attribute, as traditionally OA has been grouped with other joint diseases or even with all musculoskeletal diseases as a whole. The economic burden of OA to the United States has been estimated at $60 billion per year and in terms of working disability was second only to ischemic heart disease in men over the age 50 years (2). In the United States a total of half a million joint replacements occur per year, and it has been estimated that every 90 seconds a joint is replaced in Europe as a consequence of OA (1). With the incidence of the disease predicted to double by 2020 the economic and social burden of OA will continue to escalate.

OA has commonly been considered a disease of the articular cartilage, a smooth frictionless material that covers the ends of bones in synovial joints. It has however been suggested that OA in actual fact is a disease of the entire synovial joint organ, affecting synovium, menisci, ligaments and subchondral bone (3, 4). Characterised by joint pain and dysfunction, due to the progressive loss of articular cartilage, the joint undergoes remodelling by attempting to repair the cartilage, forming new bone in the joint margins (osteophytes), and the thickening of the subchondral bone (sclerosis) (Figure 1.1) (3). It can affect any synovial joint; however it is common in the hand, foot, spine, and especially the knee and hip joints. Primary OA rarely manifests in the ankle, wrist, elbow or shoulder, whereas these joints are common sites for secondary - or post-traumatic - OA. Age is the most predominant risk factor of OA although genetic predisposition, obesity, female gender, increased bone density and joint laxity are all recognised factors (2), racial and ethnic differences also exist with a higher incidence of knee OA in black and Chinese people (5). As yet there is no cure for OA, with only analgesics or anti-inflammatory drugs available to palliate until disease progression necessitates a joint replacement (1). One method of assessing OA progression is radiographically, with a classification scale known as the Kellgren-Lawrence grading system (6). As only bone shows up on an x-ray, the area of the joint between the bones, where the articular cartilage lies, shows up as a space. As cartilage is lost the bones share an increased proximity and so a reduction in this joint space implies a reduction in articular cartilage. The Kellgren-Lawrence grading system has four grades ranging from Grade I: unlikely joint space narrowing and possible osteophytes; Grade II: possible joint space narrowing and small osteophytes; Grade III: definite joint space narrowing, mild sclerosis (hardening of the bones), possible deformation of the end of the bones, and multiple osteophytes; Grade IV: severe joint space narrowing, sclerosis and deformation of the end of the bones, and multiple large osteophytes.

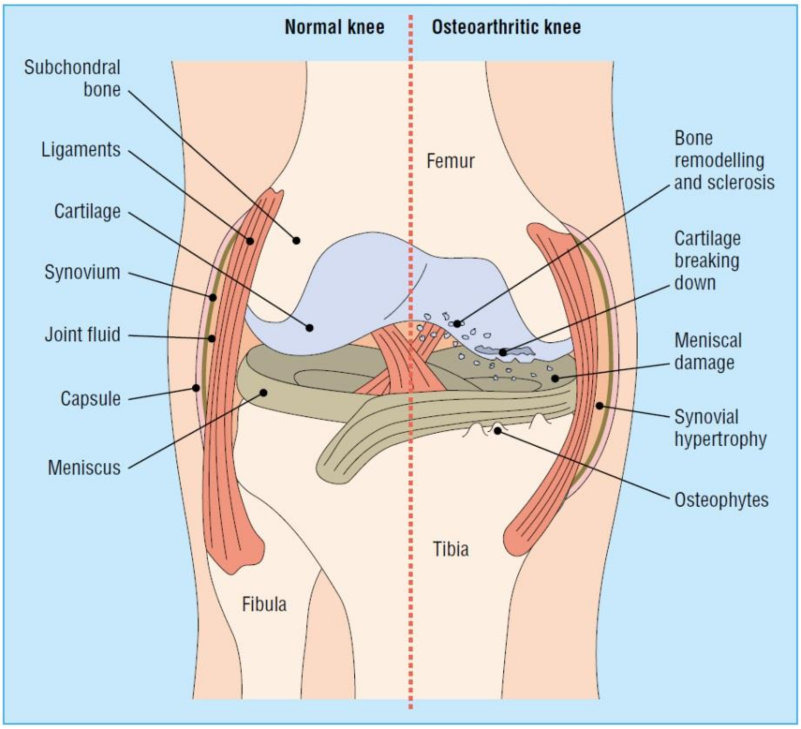


Figure 1.1 Diagram depicting the structure of the knee and the changes which occur in an osteoarthritic joint. Figure taken from *Osteoarthritis*, Hunter & Felson, 2006. Reference (3)

1.2 The synovial joint

A joint is the connection of two structural skeletal elements, and most of the joints in the body are synovial joints. Unlike fibrous or cartilaginous joints, synovial joints are made up of a collection of varied tissue types which include, or may include depending upon the site of the joint; articular cartilage, subchondral bone, ligament, tendon, synovium, synovial fluid, synovial membrane, fat pad and menisci all housed within a fibrous capsule (7). It is this intricate multi-tissue makeup that has led to the concept of the synovial joint being regarded as an organ, with disease affecting one tissue perpetuating into downstream effects upon other tissues within the joint organ (7).

Articular cartilage is a milky white, shiny, specialised material that acts as a smooth, slick, firm surface which allows low friction articulation of a joint and also cushions mechanical forces exerting between bones under load (1, 8–10). The thickness, cell density and matrix composition varies both within a joint and between joints but the overall task carried out remains constant, and at only a few millimetres thick at most, its ability to resist compression and distribute loading is a property that is ubiquitously devoid in any synthetic material (10). Cartilage is almost unique in the fact that it is maintained by only a single cell type – the chondrocyte – a highly differentiated and solitary cell (2). The chondrocyte surrounds itself with an extracellular matrix (ECM) and forms no cell-to-cell contacts, the cartilage is avascular and aneural and as a result chondrocytes depend largely on anaerobic metabolism (7, 10). Nutrients required by the chondrocytes for their matrix production and maintenance must diffuse through the synovial tissue and synovial fluid as well as through the ECM, restricting the size and charge of materials available to the chondrocyte. Although being the sole cell type found within cartilage, the chondrocyte contributes an almost negligible volume to the tissue, only approximately 1 per cent (10).

The ECM is made up of a highly ordered framework of macromolecules and water which provide cartilage with its aforementioned properties. The macromolecules of the ECM can be placed into two groups; collagenous and non-collagenous (9). The main collagenous component is type II collagen, with types VI, IX, X and XI also being present in lower abundance (9, 10). These collagens form a structured meshwork which gives the tissue cohesiveness and allows the entrapment of proteoglycans, the non-collagenous components of the ECM. Proteoglycans are a protein core with one or more glycosaminoglycan chains (long unbranched polysaccharides consisting of repeating disaccharides that contain an amino sugar) (10). The disaccharides contain negatively charged sulphate groups, resulting in long chains with negative charge and thus repelling any like charged molecules while attracting positively charged molecules. The proteoglycans are classified as large aggregating proteoglycan monomers, and small proteoglycans (10). Aggrecan molecules fill most of the space between the collagen networks, accounting for 90% of the proteoglycan mass. When a joint is under load, water is expelled from the cartilage as it is compressed. It is the hydrophilic properties of aggrecan monomers which draw water back into the matrix allowing cartilage to recover its shape and function (11).

The cartilage itself comprises of four layers, or zones: superficial, middle, deep and calcified (11) (Figure 1.2). The superficial zone, which is closest to the joint space, has the chondrocytes orientated with the collagen fibrils. Moving down into the middle zone, the chondrocytes become larger, rounder and like the fibrils more distributed throughout the ECM. The deep zone houses chondrocytes arranged in columns perpendicular to the surface, mirroring the collagen fibrils which extend down through the tide line – which separates the deep and calcified zones – into the calcified zone where they aid the anchoring of the cartilage to the subchondral bone (8, 11).

The synovial capsule, in tandem with ligaments, set the parameters of flexibility of the joint and ensure mechanical stability (12). The synovial membrane is made up of synoviocytes, which play a critical role in nourishing the chondrocytes. They also remove degraded ECM material from the synovial space (12). Synoviocytes are responsible for maintaining homeostasis within the joint by producing hyaluronic acid and lubricin, which acts as a lubricant between opposing cartilage surfaces (12, 13). The synovial fluid is the predominant medium through which nutrients, metabolites and oxygen diffuse from the synoviocytes to the chondrocytes embedded within the articular cartilage.

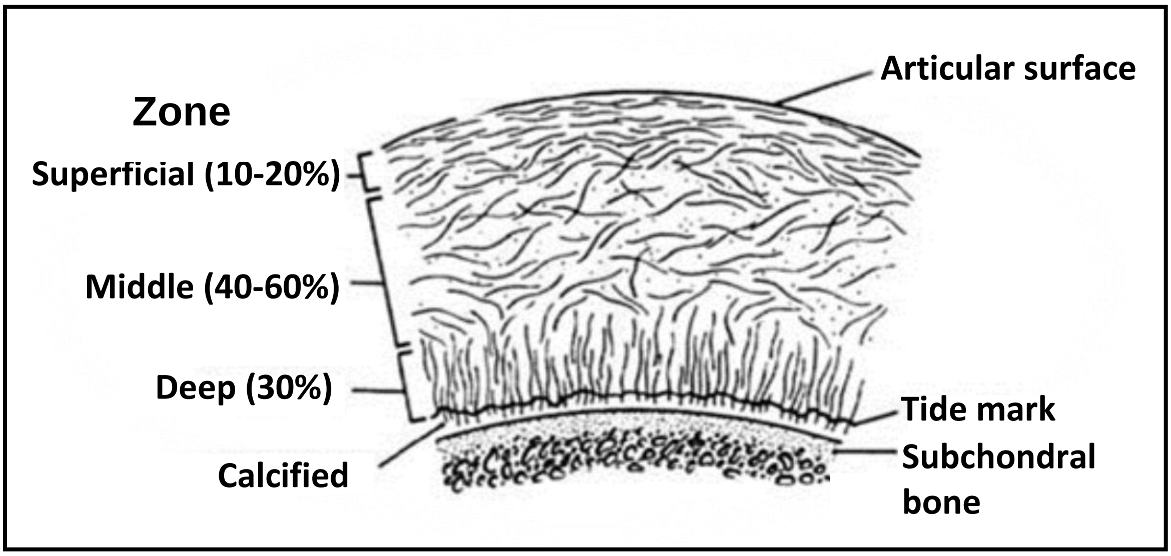


Figure 1.2 Schematic diagram depicting the four zones of articular cartilage. Percentage values show how much each zone contributes to the overall thickness of articular cartilage. Thin fibres represent the orientation of the collagen fibrils through the zones and small dots represent the distribution of chondrocytes. Image adapted from Orthoteers.org (14).

1.3 Bone growth

Other than the skull, endochondral ossification occurs in all parts of the skeleton, where cartilage is replaced by bone. This cartilaginous skeleton is formed during development when embryonic mesenchymal cells condense and differentiate into chondrocytes which go on to produce and secrete the components of the ECM (15). These chondrocytes proliferate and go on to expand the cartilaginous skeleton. An interzone forms as mesenchymal cells condense to form chondrocytes and it is this interzone which dictates the location of the joint space (16). Growth/differentiation factor 5 (GDF5) is believed to be expressed firstly throughout the mesenchymal condensations, stimulating chondrogenic differentiation and then later in development its expression becomes limited to the interzone, where it can promote apoptosis, segmentation of skeletal precursors, articular cartilage differentiation and tendon and ligament development (17). Loss of GDF5 function in mice results in a broader expansion of the interzone and its markers, but also of *Gdf5* expression itself, indicating that *Gdf5* regulates its own expression. The interzones' normal expansion can be partially restored by the addition of exogenous GDF5 (17).

Following the formation of the interzone, a process of cavitation and morphogenesis takes place where the opposing limb ends form interlocking structures (18). Following this, endochondral ossification occurs whereby chondrocytes firstly proliferate, secreting ECM around themselves. They then undergo hypertrophy, growing in size, arranged in columns that run parallel to the length of the limb, while down regulating the synthesis of type II collagen and initiating the expression of non-fibrillar type X collagen, a specific marker of hypertrophic chondrocytes. They also express matrix metalloproteinase 13, a collagenase capable of degrading type II collagen within the ECM. During late hypertrophy, the chondrocytes deposit vesicles into the matrix containing mainly calcium and phosphate allowing for mineralisation to occur. Finally ossification occurs by the vascular invasion of osteoblasts and osteoclasts. The ossification process occurs during development in the primary centre of ossification in the limb but also in the secondary centre of ossification, with the growth plate sandwiched between the two centres. Ossification at the growth plate continues to occur during adolescence allowing for further bone growth, it is here that all stages of endochondral ossification can be seen at once (Figure 1.3) (15).

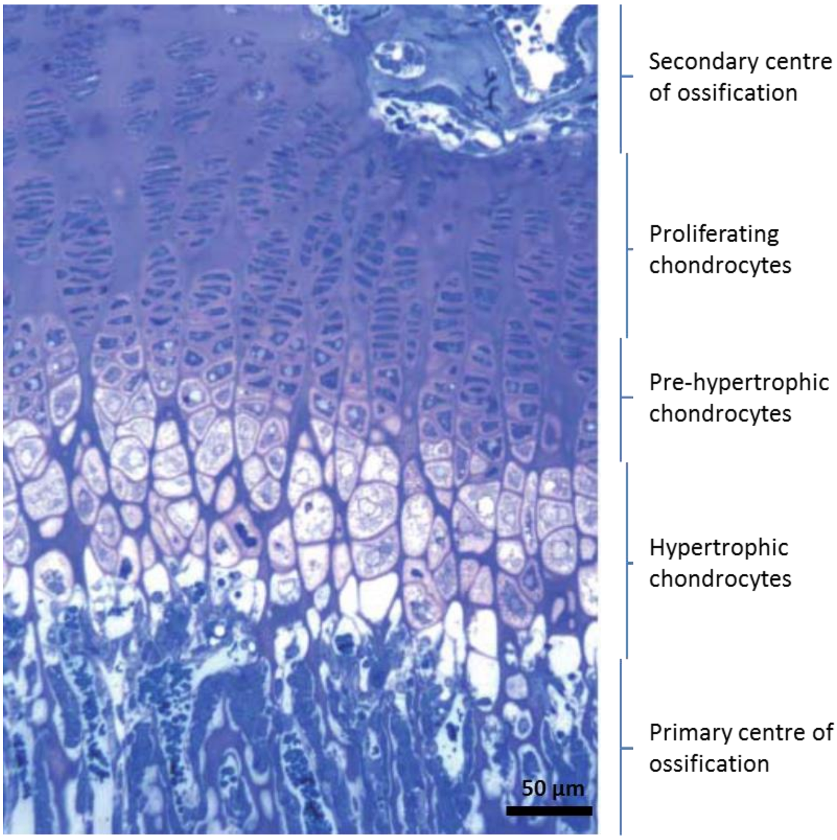

Secondary centre of ossification
Proliferating chondrocytes
Pre-hypertrophic chondrocytes
Hypertrophic chondrocytes
Primary centre of ossification
50 µm

Figure 1.3 The tibial growth plate from a three week old mouse. All stages of endochondral ossification can be seen here as chondrocytes proliferate and then become hypertrophic before undergoing cell death and ossification. Figure adapted from Mackie et al., 2011, reference (15).

1.4 Osteoarthritis pathobiology

Osteoarthritis arises from the destruction and failure of the articular cartilage extracellular matrix. An imbalance between applied mechanical stress, such as loading, and both the physical and chemical ability of the articular cartilage to resist this stress is generally the cause of primary OA, especially in the large load bearing joints (12). The disease is however not restricted to articular cartilage and should be considered as both a failure and a disease of the whole joint organ, not just the articular cartilage. All connective tissues, associated muscle groups and the nerve network along with the central nervous system may all have a role in the disease (4). With the primary symptom of OA being pain, the central nervous system is a key component, and indeed a therapeutic target and the characterisation of the joint as an organ is significant not only for physiology but also for joint pathology (12). A study looking into pain thresholds and inflammatory responses found that, in a small cohort of 26 OA patients and 33 age- and sex-matched controls, OA patients have a lower threshold to experimental pressure stimuli, and that as a result of this pain they have increased inflammatory responses to these stimuli (19). While the investigators commented on the large intra-individual variation within their small test groups, they concentrated on this as an inability to identify biomarkers but failed to touch on the fact that not only is there variation of pain thresholds between OA patients and healthy controls but there is a large variation between OA patients themselves. The investigators recorded a standard deviation of 43% in the pain threshold of leg pressure of their OA patients, but failed to highlight this (19).

As mentioned earlier, the synovial capsule, notably the synovial membrane, is a key structure in the homeostasis of the articular cartilage and of the joint as a whole. Taking this into account, it is understandable that this tissue can play a crucial role in disease progression, if not inception. The joint capsule, in tandem with the ligaments dictates the mechanical stability and flexibility of the joint. A feature of the disease in many late stage OA patients is reduced joint flexibility and therefore patient mobility. After pain, joint stiffening is the biggest concern for OA sufferers and capsular fibrosis is a major cause of this (12). Significant synovial pathology is associated with all cases of OA, both early and advanced (20, 21). This supports a possible link between the synovial reaction and the clinical symptoms of OA. However, whether these alterations are involved in disease initiation or merely progression is unknown, but an effect on the latter is accepted. The turnover of normal articular cartilage matrix produces molecular “detritus” and for example under wear and tear of a joint fragments of cartilage and bone may be sheared off and end up in the synovial membrane, when these “detritus” levels exceed physiological levels, the synoviocytes undergo activation and proliferation with a subsequent synovial hyperplasia in order to clear the synovial fluid of these fragments. An increase in the number of type A synoviocytes, which like macrophages are phagocytic, occurs in the synovial lining. This reaction is described as the initial event in osteoarthritic synoviopathy (12).

The subchondral bone is also important in the pathology of OA, although whether changes (such as sclerosis) to the tissue occur before changes to the articular cartilage or after the biomechanical attributes of cartilage have diminished is unclear (22). The formation of osteophytes – or osteochondrophytes – at the joint space margins is a clear indicator of OA. However there are contradictory opinions as to whether osteophytes are adaptive to cartilage degradation, by redistributing mechanical stresses away from damaged cartilage, or involved in cartilage destruction with some studies suggesting osteophyte formation positively correlates with cartilage destruction (1). At present there is no evidence that the removal of osteophytes slows or accelerates OA disease progression and pain (1).

Cartilage changes from its smooth, firm, milky white and glassy appearance to a soft, brown or yellowish appearance in OA, with a rough surface in early disease progression. Fibrillation and matrix loss occur in late stages until eventually the cartilage is worn through to the subchondral bone (12). The uncovering of the bone is due to a partnership of enzymatic degradation and mechanical wear of the cartilage. This causes the destruction of the collagen network and the depletion of the proteoglycan components of the matrix. Aggrecan loss, and its negative charge with it, is a distinctive marker of early stage cartilage degradation. However, collagen levels, which provide the rigidity of the matrix, remain constant until late disease stages (23). Even though the collagen levels remain relatively constant, it is the loosening of the network which is the hallmark of cartilage failure. This sparks a chicken and egg scenario; does the loss of proteoglycan or the loosening of the collagen network contribute to cartilage degradation? Either way a vicious cycle ensues with each criteria affecting the other. The most marked degradation occurs in the superficial zone and proximal to the chondrocytes, where there are enhanced levels of metalloproteinases responsible for matrix degradation. The chondrocytes undergo cell death or proliferate to compensate for those which have undergone apoptosis. Their phenotype changes drastically as the gene expression profile alters from that of healthy cartilage (12). The osteoarthritic chondrocyte population become vastly heterogeneous and zonal population changes arise. Osteoarthritic chondrocytes mimic those of foetal cartilage whereby type II collagen is up-regulated in relation to aggrecan expression, in order to manufacture new matrix (24). This is in stark contrast to normal mature cartilage where chondrocytes maintain a homeostasis by preserving matrix composition, primarily with regards to proteoglycan turnover and not type II collagen turnover (12). However, even in some late stage osteoarthritic cartilage, some chondrocytes remain anabolically active and still transcribe aggrecan and type II collagen mRNA (25), which supports the idea that chondrocytes in osteoarthritic cartilage are heterogeneous and that OA affects distinct populations or zones of chondrocytes rather than the global articular cartilage population. In early stage OA, key phenotypic changes occur in the superficial zone. Chondrocytes express genes which do not fit into the expression profile of normal chondrocytes. They express enzymes, cytokines and growth factors all involved in the degradation of the ECM. This may be due to epigenetic changes such as the demethylation of metalloproteinase genes (26). This phenotypic alteration propagates from the superficial zone down through the mid and deep zones, as more and more chondrocytes become degenerative (12). The ability to therapeutically halt such propagation would certainly been welcomed.

A classic feature of cell senescence, and the aging process, is that of the shortening of the telomeres, which has been exhibited in chondrocytes (27). However, due to the chondrocytes low replication rate it is unlikely that the cause of this telomere shortening is due to any form of replicative senescence (28). Instead telomere shortening in chondrocytes is more likely due to stress-induced senescence caused by oxidative damage and inflammation (29, 30). Cell senescence is not just the halting of cell replication but can also result in the alteration of gene expression and cellular phenotype. This phenotype is known as the senescent secretory phenotype (31) and is characterised by the increases in cytokines such as

IL-1, Matrix Metalloproteinases (MMPs) and growth factors. The expression of MMPs along with cytokines such as IL-1 is elevated in OA cartilage and this senescent secretory phenotype may link the ageing process directly with the onset and progression of OA (28).

Another age related change within cartilage is the accumulation of reactive oxygen species (ROS) such as super oxide and hydrogen peroxide which are normally controlled by anti-oxidants (32, 33). Evidence shows that levels of these anti-oxidants are lower in aged cartilage (34) and that of OA cartilage (35). The depletion of one anti-oxidant in particular – superoxide dismutase 2 – has been shown to lead to oxidative damage and mitochondrial dysfunction (36). This will have a negative effect on normal cell function and this oxidative stress as a result of aging can further push chondrocytes into a senescent phenotype and prompt the progression of OA (37).

1.5 The role of genetics in osteoarthritis

1.5.1 Epidemiological studies/Twin studies

Osteoarthritis is as multifactorial disease manifesting from the interplay of both genetic and environmental components, with a continuous distribution between the two extremes of predominantly genetic and predominantly environmental (38). The participation of genetic factors was first observed by Stecher & Hersh in the 1940s where they studied almost 70 000 adults and the heritability of hand OA (39). They observed a two to threefold increased risk of OA in the mothers and sisters of cases with hand OA. This study lay the foundations for expanded studies which began to suggest a multifactorial pattern of inheritance rather than a single gene defect with Mendelian inheritance (11). However this familial clustering with suggestion of genetic determinants had its detractors. An argument could be levelled that these families not only shared genetic material but also shared environments, implying that genetic factors had little or no role (11, 38).

A way of circumventing these criticisms is twin studies, where measures of the concordance of disease in monozygotic (MZ) and dizygotic (DZ) twins allows the separation of genetic factors and shared environmental factors (40). This also removed the potential confounder of age, which is the predominant risk factor for OA. Twin studies allow for quantification of genetic and environmental contributions due to MZ twins sharing the entirety of their genomes, such that any differences in disease concordance between them is environmental in origin. DZ twins however share half their genomes, on average, with each other and for that reason differences in disease concordance between these twins is the result of an interaction between both genetic and environmental factors (40). Tim Spector along with a number of colleagues has over the last 15 years looked into the heritability of OA at various sites through disease concordance in female MZ and DZ twins. Firstly, they assessed hand and knee OA both radiographically and clinical knee pain. They found that both in terms of a radiographic diagnosis and as a presence of joint pain that incidence was higher in MZ twins than in DZ twins. They estimated that the genetic factors accounting for radiographic OA of the hand and knee in women is between 39% and 65% (41).

In the year 2000, Spector and colleagues focused on joint space narrowing in hip OA using twin studies and found that the heritability was approximately 60% at this site (42). They also used twins to examine the genetic contribution to both cervical and lumbar spine disc degradation using magnetic resonance imaging. They found that the heritability of spinal disc degeneration was 73% at the cervical curve and 74% at the lumbar curve (43). Taken together these findings suggest that genetic factors account for approximately half of the variation in the heritability of OA, but individually they suggest that OA may not be a generalised disease affecting all joints in exactly the same manner. A study testing whether shared genetic influences were affecting hand, knee and hip OA in MZ and DZ twins, again involving Tim Spector, was performed in 2009. They looked at, and compared, three sites within the hand (the distal interphalangeal joint, the proximal interphalangeal joint and the carpometacarpal joints) along with the hip and the knee and found that the only genetic overlap present within OA was in the joints of the hand, predominantly in the two interphalangeal joint sites they tested (44). The investigators were unable to find evidence of a ‘generalised OA’ phenotype where common genetic factors were attributing to OA at multiple sites, and hypothesised that this could predominantly be due to differences in joint morphology. They detected an overlap within the joints of the hand, which due to their close proximity share similar mechanical stresses and traumas, with the greatest overlap being recorded between the distal and proximal interphalangeal joints which share the greatest morphology amongst of all the joints they tested. The large joints – the knee and hip joints – although they are subjected to similar stresses and loads, deal with these in differing ways as one is a hinge joint (although there is some lateral rotation during flexion) and the other is a ball joint.

Classification of OA in genetic studies is a cause for debate, with many investigators classifying OA by radiographic changes, which have been shown to have weak association with progressive symptomatic OA (38). Other studies have defined the phenotype as progressive OA requiring a total hip/knee replacement (THR/TKR) (38). However, this classification also stimulates debate as it selects for individuals with late stage disease who were willing to undergo major surgery to alleviate their pain and their mobility issues, and these factors may depend a great deal on varied pain thresholds as well as on varied confidence in medical procedures (45). In 2011, Kerkhof *et al.* attempted to redress the variation in definition of the OA phenotype by proposing a standardisation and a clear definition of what criteria were used when patients were recruited (46). For example, they suggest that instead of stating “patients had OA with a Kellgren/Lawrence scale of ≥2” the exact physical condition of the joint should be reported such as “at least two moderate definite osteophytes and possible joint space narrowing at the tibio-femeral joint”.

Despite the problems in defining the phenotypes of OA as well as the initiating and progressive molecular elements, there have been numerous successes in identifying susceptibility loci involved in OA by using candidate-gene studies, genome-wide linkage scans, and genome-wide association studies (9, 11, 47, 48).

1.5.2 The candidate gene approach

Candidate-gene studies in OA initially stemmed from a hypothesis-driven investigation of genes known to be mutated in rare diseases in which secondary OA forms a phenotypic component of a single gene defect disease, such as an osteochondrodysplasia (40). These rare, monogenic disorders, with high penetrance and autosomal dominance, highlighted genes that may also be involved in OA, with the assumption being that common DNA variants in the gene rather than highly penetrant rare mutations predispose an individual to primary OA (49, 50). Genes investigated included *COL2A1*, *COL1A1*, *COL9A1*, *COL11A2*, *CMP* (cartilage matrix protein), *VDR* (vitamin D receptor), *ERS1* (estrogen receptor α), *IGF-1* (insulin-like growth factor 1), *ACAN* (aggrecan) and *TGFB1* (transforming growth factor-β 1) (50). Candidates were tested by association analysis with most results being negative or inconclusive due to the small cohort sizes used, with any compelling results suggesting that only a small proportion of OA heritability could be attributed to polymorphism at these genes (49).

A Japanese group chose to test as a candidate gene *ASPN* for association to OA on the basis that its encoded protein can bind transforming growth factor β (TGF-β), blocking the growth factor's interaction with its receptor and modulating TGF-β induced chondrogenesis (51). ASPN was shown to be abundant in the articular cartilage of OA patients but was not abundant in unaffected individuals. They found a triplet repeat coding for a stretch of aspartic acid (D) residues in ASPN asporin to be associated with OA. The group found 10 alleles for this D-repeat, ranging from 10-19 aspartic acid residues. They found the D14 allele was associated with both hip- and knee-OA. Additionally, they identified that the D13 allele, coding for one less aspartic acid residue than the OA associated D14 allele, was elevated in their control group, suggesting that not only had they found an OA-susceptibility allele but also an OA-protective allele. They demonstrated that asporin is able to inhibit TGF-β signalling which then reduces the expression of type II collagen and aggrecan encoding genes, both of which are the fundamental macromolecules of cartilage ECM. When interrogating the functionality of the aspartic acid repeat alleles they noted that the OA-associated D14 allele was the strongest inhibitor of TGF-β signalling of all the D-repeat alleles and significantly less than the D13 encoded asporin (51).

Their data suggests that both a susceptibility allele and a protective allele exist amongst many neutral alleles all at the same locus. Those carrying the D14 OA associated allele produce less type II collagen and aggrecan than those carrying the neutral D-alleles, possibly allowing for a looser collagen network which permits the progression of OA, and those carrying the D13 allele produce more type II collagen and aggrecan presumably creating articular cartilage with a greater structural integrity.

Another candidate gene in which a reproducible association was found was *GDF5*, which codes for growth/differentiation factor 5 (52). GDF5, is a member of the transforming growth factor-β (TGF-β) superfamily of secreted proteins, and due to its pivotal role in joint formation, rare mutations in GDF5 cause several skeletal dysplasias, including type C brachydactyly and angel-shaped phalangoepiphyseal dysplasia (52). The OA association was to the T-allele of a SNP, rs143383, found in the 5′ UTR of *GDF5*. Miyamoto and colleagues demonstrated that this T-allele correlates with a reduction in GDF5 expression relative to that of the C-allele. These findings in an Asian population were replicated in a Caucasian population by Southam and colleagues (53). Further functional studies were performed which indicated that rs143383 was in of itself the functional cause of the association (54). Further association studies have demonstrated that the OA association to the T-allele of rs143383 is one of the most robust associations to OA across multiple populations (55–57).

1.5.3 Linkage analysis

There has been some great success stories in the candidate gene approach to OA susceptibility, such as with *ASPN* and *GDF5*, as well as complex diseases as a whole, but there is no doubt that the number of candidate gene investigations that returned a result of no association far, far outweigh the success stories. The selection of candidates is hampered by our incomplete understanding of the OA disease process, which potentially leaves many causative factors over-looked (11). Genome-wide linkage scans aim to scrutinize the genome agnostically, with no prior bias to current understanding. They aim to highlight areas of the genome that co-segregate with the disease phenotype, in closely related individuals. These studies are also hampered by the late onset of OA, which makes it difficult to include the parents of affected individuals in the linkage studies (58). To compensate for the difficulties in recruiting the parents of affected individuals, linkage analysis is often carried out with affected sibling pairs (ASPs), using microsatellites or variable number tandem repeat (VNTRs) polymorphisms as genetic markers.

A genome-wide linkage scan in the UK that defined the OA phenotype as that which required a total joint replacement at the hip or knee, studied families with at least one ASP (59). This study was enhanced by stratification of the ASPs by sex and by joint and more susceptibility loci were identified, one region (2q23-2q32), once further investigated through analysis of genes within the interval demonstrated association to females with hip OA to a functional non-synonymous SNP within the FRZB gene (60). *FRZB* encodes for secreted frizzled-related protein 3 (sFRP3) which acts as an antagonist to the Wnt signalling pathway demonstrating a role in both chondrogenesis and osteogenesis (61). Wnt activity allows for β-catenin to initiate transcription, Loughlin and colleagues suggested that non-synonymous mutation within sFRP3 reduces its ability to antagonise Wnt, therefore reducing the efficacy of the Wnt signalling pathway (60). The subsequent aberrant gene expression could lead to structurally compromised articular cartilage laid down during development that manifests as hip OA when mechanically challenged during adult life of females. (60). Another study demonstrated that the *FRZB* non-synonymous mutation was associated to hip OA and also in knee OA (62). However, when *FRZB* was assessed for an allelic expression imbalance (AEI) less than a quarter of individuals tested exhibited an imbalance in allelic output (63). Of these individuals tested, the fold difference between allelic expressions was an average of 1.19 suggesting that *cis*-acting regulatory elements within *FRZB* were not influencing the development of OA. When a large scale meta-analysis was performed the association to the two polymorphisms within *FRZB* was not replicated (56), demonstrating the need for large sample sizes to avoid false positives. A hypothesis that the OA associated alleles within *FRZB* are correlated with hip morphology has been tested in two studies with conflicting results, with one study finding no correlation (64) and a more recent study finding a correlation (65).

One robust association to OA susceptibility is that of *DIO2*, encoding the type II iodothyronine deiodinase (D2) enzyme which is able to activate thyroid signalling in growth plate cartilage (55). Investigators used a genome-wide linkage scan on affected sibling pairs allowing them to identify linkage on chromosome 14q32.11, and subsequent association analysis of SNPs in that region allowed them to identify an association between OA and the C-allele of rs225014 in *DIO2*. A further investigation reported that there is an increased level of D2 enzyme in OA cartilage compared to healthy cartilage by using immunohistochemistry (67). Further to this, they also reported that the C-allele of rs225014 correlates with an increased expression of *DIO2* relative to that of the T-allele of rs225014. As yet it is unknown if it is rs225014 or another polymorphism in LD with it that is the functional cause of this association.

1.5.4 Genome-wide association scans (GWASs)

Genome-wide linkage scans have provided broad genomic intervals that have been able to be examined further in a similar manner to candidate approaches but they have been succeeded by a more comprehensive, powerful and higher resolution study in genome-wide association scans.

Since the completion of the human genome project and the conception of the International HapMap Project (68), the latter investigating patterns and correlations of SNPs and providing linkage disequilibrium (LD) data, a vast number of SNPs have been discovered and catalogued. A GWAS compares the frequencies of SNPs within a cohort of affected individuals and a cohort of controls (either free of disease or representative of the population as a whole). By exploiting LD, the number of SNPs that need to be studied is a fraction of the total number of SNPs in the human genome. The drawback of this however is that once an associated SNP is discovered all polymorphisms which are in LD with the tagging SNP (tagSNP) must be interrogated, which may be many if the associated SNP resides in an area of high LD.

GWASs have been applied to many common complex diseases, with type 1 diabetes having benefited from over 40 loci being associated with the disease. However, many of these associated loci are still to be fine mapped to identify the functional variant within the LD block, with only a hand full of genes so far being pinpointed (69). In 2007 The Wellcome Trust Case Control Consortium performed a GWAS which genotyped half a million SNPs in approximately 2 000 patients each from seven common complex diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes) and a shared set of approximately 3 000 controls (70). They uncovered 24 association signals across the seven diseases with a significance of $p<5x10^{-7}$, and a further 58 loci with p-values between 10^{-5} and 10^{-7}. They demonstrated that the power of a GWAS depends upon the effect sizes being uncovered and the sample sizes used. Should one perform a GWAS with a smaller sample size, one should employ a more stringent p-value (ie. a lower p-value threshold) which is exactly the opposite of what many investigators of smaller GWASs implement.

Early GWASs in the OA field were underpowered due to small sample sizes and/or low numbers of cases and controls. In 2006, a GWAS was performed using 25 494 SNPs to look for association in 335 female knee OA patients and 335 controls (71). This study genotyped only

5% of the number of SNPs that the latest OA GWAS to be published. The association they uncovered failed to replicate in subsequent studies (72, 73). In 2008 a GWAS testing >100 000 SNPs on knee OA patients and controls was performed (74). Despite the increased number of SNPs genotyped, the numbers were still low with only 357 knee OA cases and 285 controls tested. Following replication of signals this climbed to 1534 knee OA cases and 2620 controls. They detected an association to a SNP proximal to *PTGS2* and demonstrated it to be expressed in articular chondrocytes from OA patients. The investigators performed AEI analysis on compound heterozygotes of the associated SNP and a transcript SNP within *PTGS2* and found the OA associated allele correlated with reduced expression in only two out of the six patients tested.

In 2010 a GWAS was performed on a cohort from the Netherlands and reported an association to knee and/or hand OA at a locus on 7q22 (75). By genotyping over half a million SNPs in 1 341 OA cases and 3 496 controls, with associated SNPs undergoing replication in some 14 938 OA cases and approximately 39 000 controls, this was the first GWAS looking for OA susceptibility loci with reasonably high power than the studies which went before it. A further investigation using additional cohorts confirmed this association to 7q22 (76). The most significant associations within the 7q22 locus lay in a 500 kb LD block containing six genes: *PRKAR2B*, encoding protein kinase-cAMP-dependent-regulatory-type II-β; *HBP1*, encoding HMG-box transcription factor 1; *COG5*, encoding component of oligomeric Golgi complex 5; *GPR22*, encoding G protein-coupled receptor 22; *DUS4L*, encoding dihydrouridine synthase 4-like; and *BCAP29*, encoding B cell receptor-associated protein 29. Due to the high LD of this 500 kb block the causal factor of the association could lie within any of these six genes. Raine and colleagues investigated the expression of the six genes in OA and disease-free cartilage (77). They found that five of the six genes (the exception being *GPR22*) were expressed at a lower level in OA cartilage than that of disease free origin, and that the carriage of the OA-associated alleles from the Dutch GWAS correlated with significantly reduced *HBP1* expression. HBP1 has been implicated in the Wnt-signalling pathway, suppressing Wnt-β-catenin-regulated growth (78) and as in the case of FRZB – mentioned earlier – this could impinge upon the integrity of the articular cartilage laid down.

A Japanese group performed a GWAS in early 2010 on 899 knee OA patients diagnosed via radiography and 3 396 controls from the Japanese population. They tested over half a million SNPs between cases and controls after replication and found two SNPs with a genome-wide significant ($p \leq 5.0X10^{-8}$) association to disease within a 340 kb region of the human leukocyte antigen (HLA) class II locus on chromosome 6 region 6p21.31 (79). The genes within this region are involved in the body's immune response and an association supports the concept that OA should not be considered as a non-inflammatory disease.

In late 2007, the Arthritis Research Campaign (now known as Arthritis Research UK) announced plans to fund a large well powered GWAS to identify OA susceptibility loci. This undertaking was named arcOGEN, standing for Arthritis Research Campaign Osteoarthritis Genetics. The arcOGEN GWAS was headed by 16 principal investigators from centres across the UK (80). Of the 7 410 unrelated OA patients, 80% of them were selected due to having end-stage OA requiring a total joint replacement of the knee or hip. The 11 009 controls used were either population-based, or OA-free controls. Population-based controls came from various sources such as the Type 1 Diabetes Genetics Consortium and the Wellcome Trust Case Control Consortium, whereas the OA-free controls stemmed from the TwinsUK cohort which used twins to study the heritability and genetics of age-related diseases (80, 81). One signal, that although did not reach genome-wide significance, showed a suggestive association to OA at stage 1 of the arcOGEN GWAS and remained significant following replication. That signal was to the T-allele of rs2615977, which resides on chromosome 1p21 within intron 31 of *COL11A1* ($p=1.2x10^{-5}$) (82). Again, while not genome-wide significant, an association within a gene coding for part of a cartilage collagen gains increased attention due to its critical role in cartilage. The potential significance of this gene to human disease is heightened due to a discovery that another polymorphism within *COL11A1* is associated with lumbar disc herniation in the spine (83). Type XI mutations, including those in *COL11A1*, typically affecting highly conserved glycine residues preventing the correct formation of the heterotrimer, are found in certain forms of Stickler syndrome, a rare and Mendelian disorder of severe OA that presents during adolescence (84, 85). Patients with Stickler syndrome also have abnormalities with their spine, including a narrowing of the intervertebral disc, further highlighting type XI collagen's critical role (83).

The most promising signals from the full arcOGEN study were replicated in up to 7 473 cases and 42 938 controls stemming from other European cohorts from Iceland, Estonia, the Netherlands and the UK. The arcOGEN GWAS identified five loci that associated with disease at genome-wide significance ($p\leq5.0X10^{-8}$) (Table 1.1) (80).

A key reason for the use of GWASs to identify factors contributing to disease aetiology is their agnostic and global nature. With the exact causes of OA unknown, then the use of techniques such as expression profiling would depend upon the cell types used. We know that the articular cartilage is primarily affected in OA but it would be naive to assume that the chondrocyte is the only cell that is the cause of OA. There is evidence that pathways identified by GWASs can be used as targets for therapies. In 2009, a GWAS reported an association between the IL12/IL23 pathway and Crohn's disease (86), treatment with anti-IL-12 p40 monoclonal antibody has a therapeutic effect upon the deregulation of T-helper 1 cells which underlies Crohn's disease (87).

Table 1.1 Signals from the arcOGEN GWAS which reached genome-wide significance following replication. * denotes the same signal due to an r^2=1 value. TJR = Total Joint Replacement, THR = Total Hip Replacement. Data taken from Table 2 of reference (80).

Signal	Chromosome	Nearest gene(s)	Stratum	Odds ratio	p value
rs6976*	3	*GLT8D1*	TJR	1.12	$7.24\text{x}10^{-11}$
rs1177*	3	*GNL3*	TJR	1.12	$1.25\text{x}10^{-10}$
rs4836732	9	*ASTN2*	THR-female	1.20	$6.11\text{x}10^{-10}$
rs9350591	6	*FILIP1; SENP6*	Hip	1.18	$2.42\text{x}10^{-09}$
rs10492367	12	*KLHDC5; PTHLH*	Hip	1.14	$1.48\text{x}10^{-08}$
rs835487	12	*CHST11*	THR	1.13	$1.64\text{x}10^{-08}$

1.6 Investigating osteoarthritis associated genes

Once OA associated genes have been identified these genes need to be investigated in detail to assess how polymorphism at the genes influences OA development. Important considerations include the effect on exon splicing, allelic expression imbalance (AEI) or the discovery of rare variants. *Cis*-regulatory elements play a major role in the total transcriptional output of the human genome, with functional *cis*-regulatory sites estimated to be heterozygous in over 40% of all genes in an average individual (88). Functional variants in regulatory regions, rather than coding regions, of a gene are difficult to elucidate from solely nucleotide sequences. The location of regulatory elements can reside as much as 1 Mb either side of the transcription unit, many of which may be within introns of neighbouring, biologically unrelated genes and alternative promoters or exons can confound this issue (89) however, a GWAS carried out to look for association to vitamin A deficiency, found a SNP 8 kb upsteam of the key enzyme that converts β-carotene to vitamin A, β-carotene 15,15′-monoxygenase (BCMO1) in what is likely to be an enhancer element (90). *Cis*-regulatory elements can be context specific and enhance expression in one cell type and repress expression in another (91, 92). It is the spatiotemporal characteristics of *cis*-regulatory elements which make nucleotide sequence analysis inadequate without functional investigation both in vitro and in vivo.

In vitro studies can use luciferase reporter plasmids transfected into cell lines, and luciferase activity can infer the activity of *cis*-regulatory elements when allelic variants are tested. *In vivo*, one can measure the allelic output, via mRNA levels, in order to quantify the extent of a particular *cis*-regulatory element. By doing this, one would assume that in the absence of any *cis*-regulatory polymorphism the transcriptional output of an autosomal gene would yield a ratio of 1:1. Should any AEI be detected within an individual, the implication would be that that individual is heterozygous at a *cis*-regulatory site which impinges upon transcription or mRNA stability. Each allele represents an internal control against changes in *trans*-acting factors and environmental factors, as each assay is performed within an individual and not between individuals (93).

1.7 The missing heritability in OA and complex disease

The heritability of radiographic OA has been estimated at around 50% (94), but as yet we only have several OA-associated loci all with small increments in risk, with the *GDF5* SNP rs143383 being one of the most robust associations to disease but with an odds ratio of 1.15 (56). This leads to question as to where the remaining 'missing' heritability lies. This is true for many complex diseases and traits, human height for example is estimated to have heritability of around 80%, and has more than 40 associated loci being identified, but this only amounts to only 5% of the phenotypic variance observed (95). One criticism of many early GWASs, into a number of complex diseases, is that imprecise phenotyping, coupled with inadequate control groups reduced the effect sizes but still enabled the identification of disease/trait associated variants (96). Some structural variations such as inversions, translocations, microsatellite repeats, insertions and complex rearrangements have been identified in rare Mendelian conditions but these types of variations have been largely unexplored in complex diseases and traits. GWASs have shown that complex disease cannot be explained by common variants with moderate effects (Figure 1.4), however they are still able to show us where rare or structural variants are likely to cluster (96).

The detection of rare causal variants is not possible using common tagSNPs in a GWAS as the weak correlations between the rare and common SNPs leave the study under-powered (97). It has been hypothesised that slightly deleterious SNPs in genes implicated in disease pathways are mainly responsible for the inter-individual variation in susceptibility (98). These SNPs have lower minor allele frequencies (MAFs) than common variants (which have MAFs of 1% or more), but they have higher MAFs than clearly deleterious SNPs (which have MAFs of 0.1% or less). Deleterious SNPs are in large numbers in the genome, with the rare allele disrupting gene function and dramatically increasing susceptibility. However their actions are not strong enough to be eradicated from the population (98, 99).

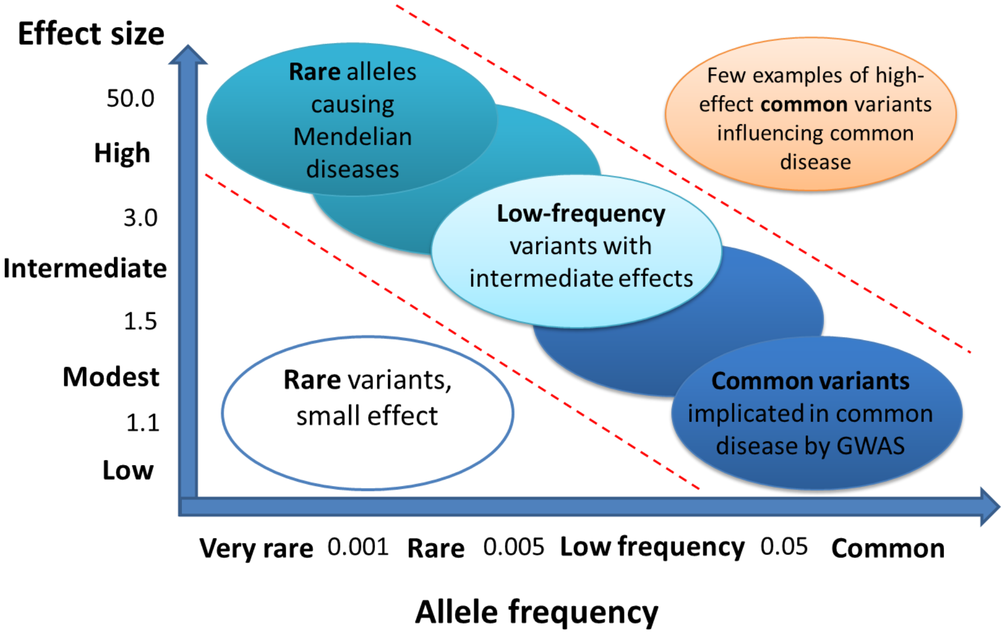

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian diseases
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effects
Rare variants, small effect
Common variants implicated in common disease by GWAS
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 1.4 Genetic variants by allele frequency and genetic effect. Most emphasis on susceptibility allele discovery has concentrated on those between the dotted lines. Figure adapted from (96).

The discovery of rare variants can be achieved by resequencing of candidate genes in a cohort of both cases and controls. If variants are expected to be slightly deleterious then they should be elevated in cases, where if they were protective they should be elevated in controls. For this reason it is important to resequence both groups, and not resequence cases and then merely genotype controls (97).

1.8 Growth/differentiation factor 5

Growth/differentiation factor 5 (*GDF5*), also known as cartilage derived morphogenic protein-1, is member of the GDF-subgroup of the bone morphogenetic protein (BMP) family (100), with the BMP family itself being a subfamily of the transforming growth factor β (TGFβ) superfamily of secreted proteins (47), which in a pro-anabolic process, assist joint homeostasis (53). GDF5's main role is the induction of chondrogenesis and the patterning of joints (100), with *GDF5* being expressed during development and in adulthood. GDF5 is active in all synovial joint tissues, being involved in tendon and ligament formation (101), as well as in the maintenance and repair of bone and cartilage (102). *GDF5* is located on a 4.88 kb stretch of chromosome 20q11.2 and has a 5′ UTR, a 3′ UTR and 2 exons either side of an intronic sequence, giving a 2 383 bp transcript. According to the Ensembl database (www.ensembl.org), there is a hypothetical transcript that codes for an extended 5′ UTR, creating 2,572 bp mRNA, however, there is no current experimental evidence to support the existence of this transcript.

GDF5 binds to membrane bound serine-threonine receptors, which upon ligand binding, transduce their signals via the SMAD signalling pathway, inducing the transcription of genes such as aggrecan and *COL2A1* (Figure 1.5) (103). GDF5 interacts with two forms of type I BMP receptor; BMPR-IA and BMPR-IB, although shows a higher affinity to BMPR-IB (104). Null mutations within the *Bmpr-1b* gene lead to viable mice carrying defects in bone and joint formation, mirroring those seen in *Gdf5* knock-outs (105) but *Bmpr-1a* knock-outs die in early embryogenesis (106). However, when a conditional knock-out of *Bmpr-1a* is introduced to circumvent the embryonic leathality under the control of a GDF5-Cre driver mice are viable but with age and physical challenge articular cartilage quickly wears, producing an OA-like phenotype which indicates this receptor is key for cartilage homeostasis and repair (103, 104). It is therefore plausible to suggest that despite a lower binding affinity for GDF5, it is BMPR-1A that should command greater attention in OA research. Wild-type GDF5 has a binding affinity for BMPR-IB 12-fold greater than that for BMPR-IA but by mutating Arg57 to an alanine the binding affinities become identical, indicating that the Arg57 of GDF5 that is the determinant for the binding specificity of BMPR-1B (107).

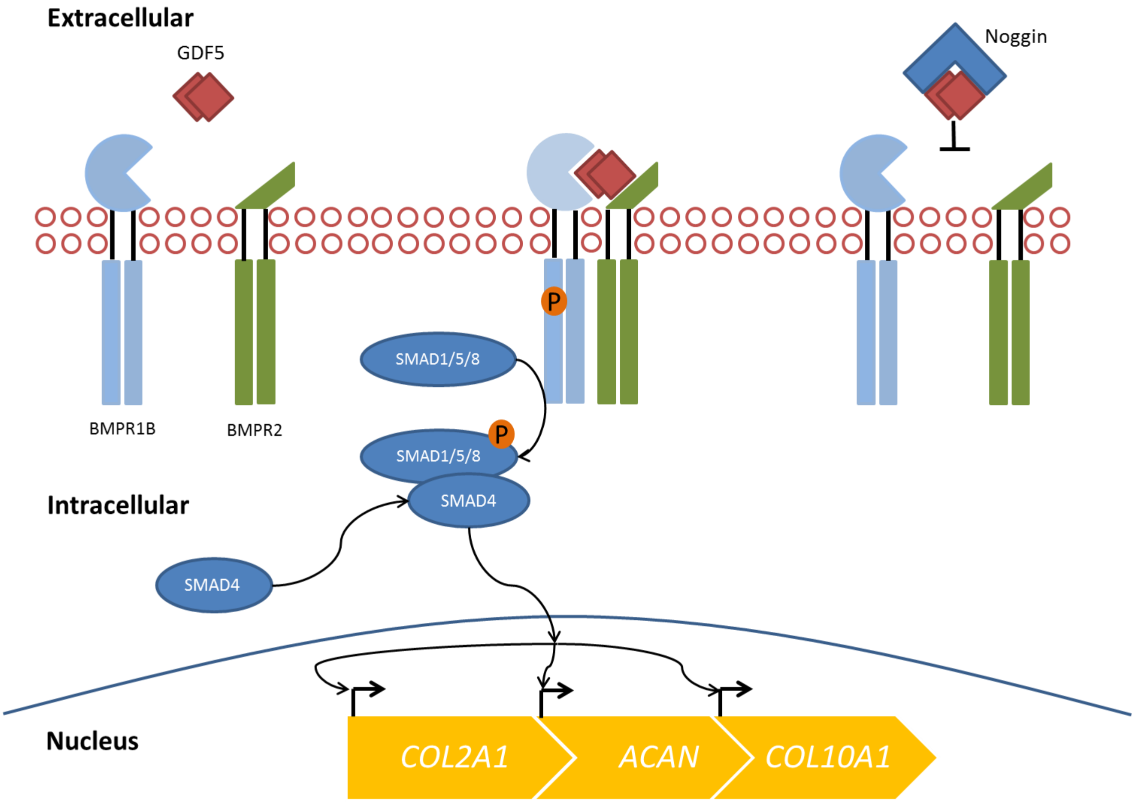


Figure 1.5 GDF5 signalling pathway. The GDF5 homodimer binds to its membrane bound receptor BMPR1A or BMPR1B (The higher affinity BMPR1B is shown). Upon binding and the localisation of the receptor complex, phosphorylation of SMADs 1/5/8 occurs which enables the recruitment of SMAD4. This SMAD complex then enters the nucleus and initiates the transcription of genes such as *COL2A1*, *ACAN* and *COL10A1*. The binding of the GDF5 homodimer to Noggin inhibits GDF5 from binding to its receptor complex. Figure adapted from (108).

A number of rare diseases are caused by penetrant mutations in GDF5, including brachydactyly type C (specific pattern of shortened fingers and toes) caused by one of five mutations (ΔG121, insertionG206, ΔG759, Arg301STOP and Arg438Cys), brachydactyly type A2 (shortened index finger and second toe), as well as Grebe and Hunter-Thompson type chondrodysplasias (caused by Cys400Tyr and a 22 base pair fame shifting insertion respectively) both of which are caused by mutation of the cysteine knot in the mature form of GDF5 (100). Other diseases include: DuPan (dwarfism or absence of bones within the limbs, bilateral absence of the fibula); angle-shaped phalangeopiphyseal dysplasia (short stature, hip dysplasia, severe bilateral vertical talus (rocker-bottom feet), premature vertebral end-plates); proximal symphalangism (ankylosis of proximal interphalangeal joints, fusion of carpal and tarsal bones, conductive deafness due to congenital ankylosis of the stapes) and multiple synoptosis syndrome 2 (Many skeletal abnormalities such as tarsal-carpal fusion, humeroradial synostosis, brachydactyly, proximal symphalangism, broad hemicylindrical nose and vertebral fusion) (109). GDF5 is synthesised as a 501 amino acid precursor protein, which then undergoes proteolytic cleavage at residue 381, releasing the mature, active, 120 amino acid C-terminal protein of GDF5 (100). It is this proteolytic cleavage which is the defining factor in rendering GDF5 active (100). Cleavage occurs at a conserved R(R/K)RR (RXXR) sequence.

GDF5 was discovered in the brachypodism (*bp*) mouse, where a spontaneous mutation within the gene results in the absence of proximal interphalangeal joints, joint fusions in the ankle and wrist and bone abnormalities associated with dislocations with the mutation causing loss of function of *GDF5* (110). Due to the homozygote brachypodism mouse *(Gdf5*$^{Bp\text{-}J/Bp\text{-}J}$) exhibiting such a severe developmental phenotype, one group bred the haploinsufficient (*Gdf5*$^{Bp\text{-}J/+}$*)* mouse where the phenotype shows no difference from that of the wildtype (111). Here, the investigators showed that when challenged by use of collagenase induced arthritis, the *Gdf5*$^{Bp\text{-}J/+}$ mouse demonstrated increased OA-like changes in the contralateral knee due to increased loading, and a higher degree of synovial hyperplasia when these mice were challenged on a treadmill. Analysis of the homozygote *Gdf5*$^{Bp\text{-}J/Bp\text{-}J}$ mouse revealed that the Achilles tendons of these *GDF5* insufficient mice are structurally weaker and pervade increased laxity in the joint, correlating with the phenotype of both chondrodysplasias Grebe and Hunter–Thompson in humans which is characterised by both joint laxity and dislocation (111, 112).

Miyamoto *et al.* sequenced the two exons of *GDF5* in 239 Japanese cases with symptomatic hip OA validated by radiography, and in 256 Japanese controls (52). Three common *GDF5* SNPs were identified and these were tested for association with hip OA. The three SNPs were found to be associated and this was confirmed in an independent case-control cohort of 761 cases and 728 controls. After combing the cohorts, Miyamoto and colleagues found that the strongest association ($p = 1.8 \times 10^{-13}$) was with a SNP in the 5′ UTR (rs143383, T/C) with the associated T-allele being present in 84% of cases and 74% of controls, with an odds ratio (OR) of 1.79 (95% CI 1.53-2.09). This SNP was also associated with knee OA in Japanese and Han Chinese populations (52).

When Southam *et al.* tested for association of rs143383 in Europeans, they genotyped the SNP in UK and Spanish knee/hip/hand OA cases (n = 2 487) and age matched controls (n = 2 047) and like the Asian study, the T-allele was elevated in OA cases (53). Whereas, Miyamoto *et al.* had used luciferase reporter assays to demonstrate that the T-allele of rs143383 led to reduced expression of GDF5, Southam *et al.* carried out an AEI analysis using RNA extracted from the articular cartilage of OA patients who were heterozygous for rs143383. An average 12% reduction in expression of the T-allele was observed, relative to the C-allele. From a molecular genetics perspective, comparisons can be made between the bp mouse and the effect of rs143383 in humans, with both systems confirming the a joint’s reliance upon sufficient levels of GDF5 (113).

Egli *et al.* (2009) carried out a more comprehensive interrogation of the functionality of rs143383 with the identification of a candidate trans-acting factor, deformed epidermal autoregulatory factor-1 (DEAF1), which binds differentially to the two alleles of the SNP (54). They found that AEI of *GDF5* was occurring in a joint-wide capacity and that rs143383 was not the only *cis*-acting factor effecting *GDF5* expression, with another SNP in the 5′ UTR, rs143384, also having an effect on the expression of the gene. Their luciferase reporter assays showed that a down regulation of the T-allele of rs143383 only occurred when in the presence of the T-allele of rs143384. Egli et al. also found that a SNP in the 3′ UTR, rs56366915, had an independent and partly additive effect upon *GDF5* allelic imbalance.

1.9 Collagen fibrillogenesis

Callagens are the most abundant proteins in the body, coded for by more than 30 genes, they are responsible for providing the structural components of the tissues of the body (114). These collagen structures can be both fibrils such as types I, II and III or sheets like type IV (115, 116). Collagens contain a specific amino acid sequence of Glycine – X –Y with X usually being a proline residue every three residues and every seventh Y being a hyroxyproline (115). Type II collagen is the most abundant of the collagens found in articular cartilage and is synthesised within chondrocytes as a procollagen trimer, with large N- and C-propeptides at their termini. The procollagen is secreted from the cell and undergoes propeptide cleavage and the trimers are then assembled into large collagen fibrils via covalent lysyl oxidase cross-linking at their telopeptide regions (Figure 1.6) (117). Type XI collagen is also a cartilage specific fibrillar collagen but it does not form its own fibrils, instead it associates with type II collagen molecules to create heterotrimers.

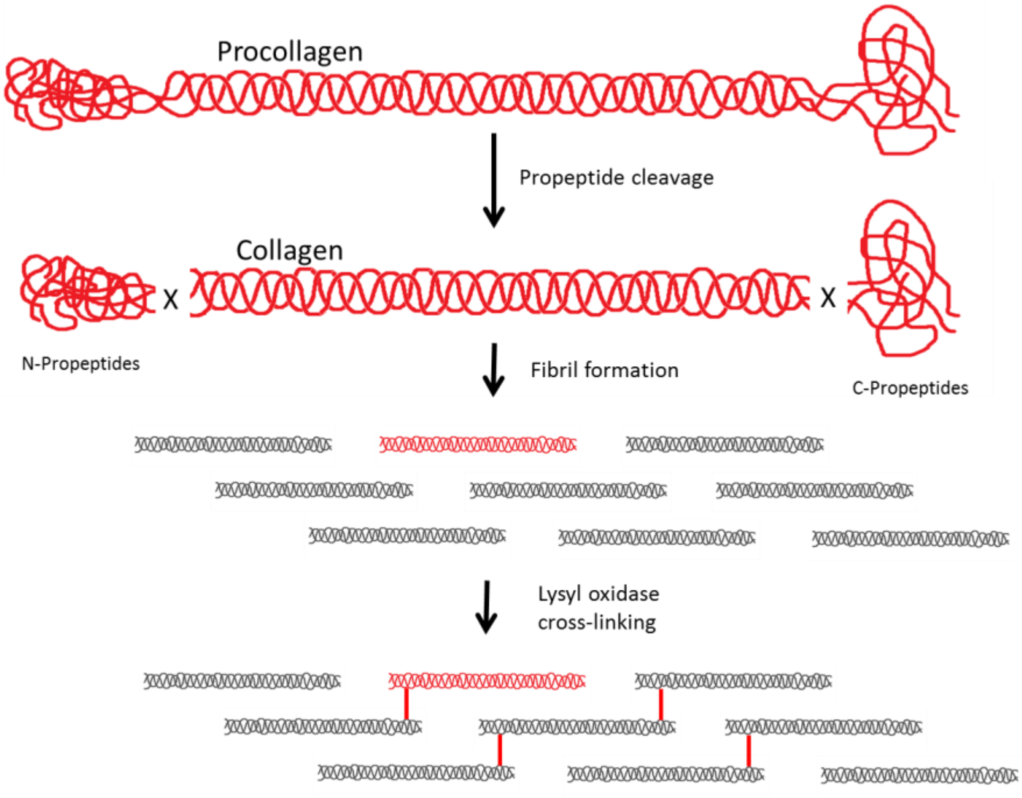


Figure 1.6 Collagen fibrillogenesis. Procollagen trimers are synthesised and excreated from the cell where they undergo propeptide cleavage. The trimers are then arranged into long fibrils where they a covalently bonded together at their terminus between lysine and hydroylysine residues. Figure adapted from (117).

1.20 Collagen, type XI, alpha 1

When stage 1 of the arcOGEN GWAS was published (82), and following replication analysis, one of the associations uncovered was to the SNP rs2615977. The T-allele of this T/G SNP associated with OA with a p-value of $1.2x10^{-5}$ and an odds ratio of 1.10 in hip OA of both males and females. rs2615977 resides on chromosome 1p21 within intron 31 of *COL11A1*, a 67 exon gene which codes for the α1 polypeptide chain of type XI collagen and is not in LD with any amino acid substitutions within the gene. Type XI collagen is a heterotrimer comprised of an α1 type XI chain (*COL11A1*), an α2 type XI chain (*COL11A2*) and a post-translationally modified α1 type II chain (*COL2A1*) (85). Although found in low abundance within the cartilage extracellular matrix, type XI collagen is a structurally important component. It is believed that the globular nature of the amino terminal domain of the α1(XI) chain, preventing its incorporation into the collagen fibrils along with the rest of the type XI molecule, is responsible for controlling the diameter of the overall collagen fibrils within the ECM (118). Chondrodysplasia in mouse (*cho/+*) is caused by a frame-shift mutation within *Col11a1*, producing a truncated α1(XI) chain, with an inability to regulate correct collagen fibril thickness, and as a result *cho/+* mice have thicker fibrils than wild-type mice (119). This highlights the critical role type XI collagen plays in the organisation of the supramolecular architecture of cartilage collagen (83). These haploinsufficient *cho*/+ mice develop OA-like changes in both their knee and temporomandibular joints, yet show no other skeletal abnormalities (120). The result of the reduced expression of type XI collagen within the *cho*/+ mouse is that the type II collagen network is altered whereby there is a higher proportion of pericellular type II collagen than in wild type cartilage. This in turn increases the levels of both discoidin domain receptor 2 (*Ddr2*) and matrix metalloproteinase 13 (*Mmp13*), which lead to cartilage degradation (120).

Type XI mutations, including those in *COL11A1*, typically affecting highly conserved glycine residues preventing the correct formation of the heterotrimer, are found in certain forms of Stickler syndrome, a rare and Mendelian disorder of severe OA that presents during adolescence (84, 85). Patients with Stickler syndrome also have abnormalities with their spine, including a narrowing of the intervertebral disc, further highlighting type XI collagen's critical role (83).

In this thesis I report on genetic and functional analyses that I have performed on *COL11A1* and *GDF5*.

Chapter 2: Allelic expression of *COL11A1* in human articular cartilage

2.1 Introduction

In late 2007, Mio *et al.* reported that a common non-synonymous SNP, rs1676486, found in exon 62 of *COL11A1*, was associated with lumbar disc herniation (LDH) in a Japanese population (83). LDH and OA are both characterised by age-associated, loss of function and degradation of cartilage. However, OA is categorised as degradation of hyaline, articular cartilage whereas LDH is categorised as degradation of type I collagen containing fibrocartilage. rs1676486 is a C-T transition which codes for a proline (hydrophobic, non-polar) to serine (polar) substitution. The T-allele of rs1676486, with an allele frequency of 25% in the Japanese population, was shown to be elevated within LDH cases, and correlated with decreased stability of the *COL11A1* transcript *in vitro* when compared to that of the more common C-allele. A difference in allelic output or transcript stability can impact upon the total amount of transcript of a gene and is known as an allelic expression imbalance (AEI). AEI can be caused by mRNA differences in transcription efficiencies or by differences in mRNA stability, with the end result being less mRNA able to be translated into protein. Mio *et al.* concluded that rs1676486 has a quantitative effect upon *COL11A1* transcript abundance, implying that a quantitative deficiency also exists of the α1 polypeptide of type XI collagen and it is this deficiency which is contributing towards LDH susceptibility.

As more and more susceptibility loci are being uncovered by GWASs it is becoming apparent that common disease risk is increasingly being linked to alleles which influence gene expression via the modulation of transcription levels or transcript stability (121, 122). A good example of such a disease susceptibility locus is rs1676486 for LDH, affecting transcript levels in a tissue relevant to that disease. It is key that investigations into the effects of susceptibility alleles are carried out using RNA extracted from not just organs that exhibit the disease phenotype but specifically from tissues exhibiting the phenotype and ideally at the time of disease development, as AEI may not only be spatial but also temporal (122, 123).

With this knowledge, I hypothesised that the OA association to *COL11A1* reported by the arcOGEN study, marked by rs2615977, may be mediated by the modulation of *COL11A1* expression in articular cartilage. Additionally, I hypothesised that the AEI reported by Mio et al. in LDH, marked by rs1676486, may also be observed in OA, again in articular cartilage.

2.2 Aims of this study

The aims of this study were to firstly test whether the OA association to *COL11A1* reported by the arcOGEN study, marked by rs2615977, is mediated by the modulation of *COL11A1* expression in articular cartilage, and secondly to test whether the AEI reported by Mio et al. in LDH, marked by rs1676486, may also be observed in OA, again in articular cartilage. In order to test these hypotheses I used two approaches; firstly to quantitatively measure overall *COL11A1* expression in articular cartilage and stratify by genotype at the two SNPs, and secondly to test for AEI within *COLL11A1* using assays which can both discriminate and quantify the mRNA levels of each allele of a SNP within the transcript of the target gene.

2.3 Materials and methods

2.3.1 Patients, tissues and nucleic acid extraction

Macroscopically normal articular cartilage tissue located distal to the OA lesion was dissected from individuals undergoing elective joint replacement for OA of the hip (total hip replacement, THR) or of the knee (total knee replacement, TKR), as described extensively previously (53, 93). The Newcastle and North Tyneside research ethics committee granted ethical approval for the collection (REC reference number 09/H0906/72) and informed consent was obtained from each donor. On the day of operation, tissue samples were snap frozen and ground to a powder using a Resch mixermill 200 (Retsch Limited, UK) under liquid nitrogen to prevent RNA degradation. Genomic DNA and RNA were extracted from ground tissue using an EZNA DNA/RNA Isolation Kit (Omega bio-tek, R6731-02). Nucleic acids were then quantified using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, USA). Tissue grinding and nucleic acid extraction for all tissue samples was carried out exclusively by Emma Raine, a colleague working within my supervisor, Professor Loughlin's, research group.

2.3.2 DNA genotyping COL11A1 SNPs

Genomic DNA isolated during the nucleic extraction from articular cartilage was used to genotype three SNPs within *COL11A1*: the OA associated non-transcript SNP rs2615977, located within intron 31; the LDH associated SNP rs1676486, located within exon 62; and SNP rs9659030, located within the 3' UTR. rs9659030 was tested for AEI due to the absence of any transcript SNPs in high LD ($r^2 \geq 0.8$) with the OA associated SNP, rs2615977. Heterozygosity was determined by PCR-restriction fragment length polymorphism (RFLP) analysis. In order to genotype both rs2615977 and rs9659030 PCR was performed using a forward primer containing a single base change allowing the creation of a restriction enzyme site in the presence of one allele but not in the presence of the other allele, along with a reverse primer which was also specific to the genomic sequence. In order to genotype rs1676486 no base changes were needed to be made to the primers as the SNP naturally altered a restriction enzyme site. A 15 µl PCR contained 50 ng DNA, along with 7.5 µM of forward and 7.5 µM of reverse primer, 1.5 µl of 10X PCR buffer (50 mM KCl, 10 mM Tris-HCl pH 8.3), 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.08 units of AmpliTaq Gold DNA polymerase (Applied Biosystems) and the total volume was brought up to 15 µl using H_2O. Thermocycling conditions to genotype both rs2615977 and rs1676486 were an initial denaturation at 94^{o}C for 10 minutes, followed by 35 cycles of 94^{o}C for 30 seconds, annealing at 64^{o}C for 30 seconds, and a polymerase extension at 72^{o}C for 30 seconds, with a final extension step at 72^{o}C for 10 minutes. Thermocycling conditions to genotype rs9659030 were an initial denaturation at 94^{o}C for 10 minutes, followed by 35 cycles of 94^{o}C for 30 seconds, annealing at 55^{o}C for 30 seconds, and a polymerase extension at 72^{o}C for 30 seconds, with a final extension step at 72^{o}C for 10 minutes. A 5 µl digestion mixture was added to each PCR to create a 20 µl reaction mixture. Five units of restriction enzyme were added along with 2 µl of the appropriate 10X NEB Buffer and 2.5 µl of H_2O. The PCR and digestion mixture was then incubated at 37^{o}C for 3 hours followed by a 20 minute heat inactivation at 60^{o}C. Primer sequences, PCR conditions and the enzymes used are listed in Table 2.1. Digestion products were electrophoresed through a 3% (weight/volume) agarose gel containing ethidium bromide, and genotypes were scored after UV visualisation.

2.3.3 Reverse transcriptase (RT) polymerase chain reaction (RT-PCR)

RNA was stored in diethylpyrocarbonate (DEPC)-treated H_2O (Invitrogen) at a concentration of 250 ng/µl. RT-PCR was carried out using the SuperScript First-Strand Synthesis System (Invitrogen). Initially 4ul of RNA in DEPC-treated H_2O was mixed with 1 µl of random primers (50 ng/µl), 1 µl of 10 mM dNTP mix (10 mM each: dATP; dCTP; dGTP; dTTP), and 2 µl of DEPC-treated H_2O. This reaction mix was then incubated at 65^{o}C for 5 minutes followed by incubation on ice for at least 1 minute. Following this, 4 µl of 5X Reaction buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 4 µl of 25 mM $MgCl_2$, 2 µl of 0.1 M DTT and 1 µl of RNaseOUT inhibitor (40 units/µl) were added and incubated at 25°C for 1 minute. Then 50 units of SuperScript II RT enzyme was added to the reaction mixture and mixed via pipetting. The 20 µl total reaction mix was incubated at 25°C for 10 minutes, 42°C for 50 minutes and finally 70°C for 15 minutes. Following the incubation cycle, 2 units of RNase H (New England Biolabs, 5 units/µl) were added and the reaction mixture was incubated at 37°C for 20 minutes. The newly synthesised cDNA was then stored at -20°C until required.

2.3.4 Quantitative real-time PCR

PrimeTime Mini qPCR Assays (Table 2.2) were purchased from Integrated DNA Technologies (IDT; Iowa, USA) and overall expression of *COL11A1* was measured relative to the housekeeping genes *HPRT1*, *GAPDH* and *18S* using an ABI PRISM 7900HT Sequence Detection System following the manufacturer's protocol (Applied Biosystems). Three replicates for each patient cDNA sample were carried out per gene, creating a total of 9 housekeeper cycle threshold (Ct) values. A mean housekeeper Ct value was calculated and used as a control against *COL11A1* expression using the $2^{-\Delta Ct}$ method. Mann-Whitney U and Kruskal-Wallis tests were performed to test whether genotype at rs2615977 and genotype at rs1676486 correlated with *COL11A1* expression.

2.3.3 Allelic expression imbalance analysis

Allelic expression imbalance was assessed using the transcript SNPs rs1676486 and rs9659030 and quantitative real time PCR genotyping assays. These assays are standard real time assays except that they employ a probe (FAM or VIC labelled) specific to each of the two alleles of a SNP. Readymade TaqMan genotyping assays for rs9659030 and rs1676486 for use with genomic DNA were purchased from Applied Biosystems. A custom made assay was required to PCR cDNA for rs1676486 as the SNP is only 5 bp from the exon 62/intron 62 boundary. Real time PCR was carried out according to the manufacturer's instructions. In brief, 4 µl of cDNA (corresponding to 200 ng of patient tissue RNA) or 20 ng of DNA (also extracted from the tissue) was added to 5 µl of TaqMan Universal Master Mix II no UNG (Applied Biosystems) and 0.25 µl of 40 x TaqMan assay in a 10 µl reaction. The samples were then amplified on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems), under the following cycle conditions: 50°C for 2 minutes, 95°C for 10 minutes and 40 cycles of 92°C for 15 seconds and 60°C for 1 minute. The reactions were performed in five replicates. The allelic ratios were calculated using the formula $(2^{\wedge\text{-FAM Ct}})/ (2^{\wedge\text{-VIC Ct}})$.

For each assay, the ratios between the amounts of each allele in every sample were calculated for genomic DNA and cDNA. For each sample the average allelic ratio for genomic DNA, which represents the 1:1 ratio, was then used to normalise the cDNA ratio to generate a corrected allelic ratio. To determine if there was an overall difference in expression between alleles for a particular tissue across all patients the mean allelic ratios for the patient cDNAs were compared to the mean allelic ratios for the patient genomic DNAs using a 2-tailed Mann-Whitney exact test.

Table 2.1 Table of primers and PCR conditions along with restriction enzymes used for genotyping SNPs by RFLP analysis.

SNP		Sequence (5'-3')	PCR Conditions	Enzyme
rs2615977	Forward primer	GTTTTAAAACATCCCCAGATAATTATCATATTCACTG	2 mM $MgCl_2$	BsrI
	Reverse primer	GTAACTAACCTGCACGTTGT	64^{o}C annealing temp.	
rs1676486	Forward primer	GCATTTTGTAGGGTCCTCAAGGC	2 mM $MgCl_2$	BsaI
	Reverse primer	ACTTTGCTATGGTAGCCCTAGC	64^{o}C annealing temp.	
rs9659030	Forward primer	GACCTACCTAATTGCTAAATGAATAACATATGGTGGACTGTTATTAACAG	2 mM $MgCl_2$	TspRI
	Reverse primer	ACTGTTCCAGTGAAATCTGG	55^{o}C annealing temp.	

Table 2.2 Table of the PrimeTime Mini qPCR Assays used for quantitative PCR analysis of *COL11A1*

Gene	Primer 1 (5'-3')	Primer 2 (5'-3')	Probe
COL11A1	TTCTCCACGCTGATTGCTAC	TTGGTGTTGAGGTTGGGAG	56-FAM/TTAACATCGCTGACGGGAAGTGGC/36-TAMSp
HPRT1	TGCTGAGGATTTGGAAAGGG	ACAGAGGGCTACAATGTGATG	56-FAM/AGGACTGAACGTCTTGCTCGAGATG/36-TAMSp
GAPDH	GGCCATCCACAGTCTTCTG	CAGCCTCAAGATCATCAGCAA	56-FAM/ATGACCACA/ZEN/GTCCATGCCATCACT/31ABkFQ
18S	CGAATGGCTCATTAAATCAGTTATGG	TATTAGCTCTAGAATTACCACAGTTATCC	56-FAM/TCCTTTGGTCGCTCGCTCCTCTCCC/TAMRA

2.4 Results

2.4.1 Genotyping of OA patients at rs1676486, rs2615977 and rs9659030

DNA extracted from patient articular cartilage was used to genotype the three SNPs by RFLP analysis. Genotyping of rs1676486 resulted in an uncut 753 bp C-allele fragment and two 564 bp and 189 bp cut T-allele fragments (Figure 2.1, A). Of the 78 patients tested, 48 were CC, 29 were CT and one was TT at rs1676486 (Table 2.3). Genotyping of rs2615977 resulted in an uncut 360 bp T-allele and two 324 bp and 36 bp cut G-allele fragments (Figure 2.1, B). Of the 78 patients tested, 40 were TT, 32 were TG and six were GG at rs1676486 (Table 2.3). Genotyping of rs9659030 resulted in an uncut 256 bp A-allele and two 206 bp and 50 bp cut G-allele fragments (Figure 2.1, C). Of the 78 patients tested, 34 were AA, 39 were TG and five were GG at rs1676486 (Table 2.3).

2.4.2 COL11A1 expression in articular cartilage stratified by genotype at associated SNPs

Using cDNA synthesised from articular cartilage RNA I was able to measure the level of *COL11A1* expression using quantitative real-time PCR. This data was then stratified by genotype at the OA associated locus, rs2615977 and by genotype at the LDH associated locus, rs1676486 (Figure 2.2). Due to rs1676486 having a MAF of only 20% (T-allele), and homozygous TT individuals being limited to only 1.8% of the HapMap CEU population, I only encountered one homozygous TT individual. Due to the low incidence of these individuals I decided to stratify by carriers (TT and CT individuals) and non-carriers (CC individuals) of the LDH associated T-allele. Stratification yielded no significant correlation ($p<0.05$) between the level of *COL11A1* expression and genotype at either rs2615977 or rs1676486. Additionally I stratified these data further by those patients who had undergone TKR and those who had undergone THR, but again there was no correlation between *COL11A1* expression levels and genotype at either rs2615977 or rs1676486 (Figure 2.3).

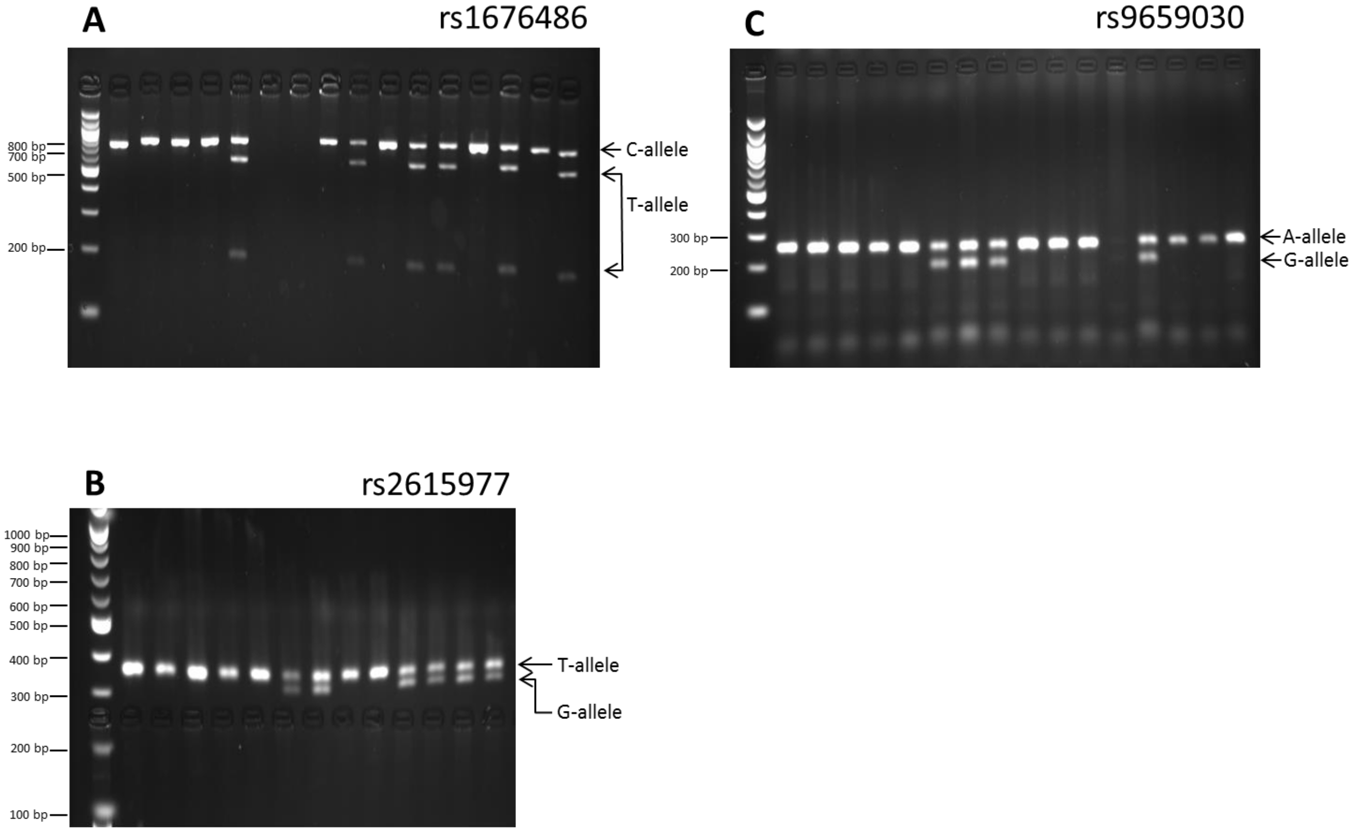


Figure 2.1 Restriction fragment length polymorphism analysis of SNPs within *COL11A1.* Agarose gel images using 100 bp ladder of: (A) rs1676486 genotyping, *Bsa*I digestion produces an uncut C-allele band of 753 bp and two cut T-allele bands of 564 bp and 189 bp in length. (B) rs2615977 genotyping, *Bsr*I digestion produces an uncut T-allele band of 360 bp and two cut G-allele bands of 324 bp and 36 bp in length. (C) rs9659030 genotyping, *TspR*I digestion produces an uncut A-allele band of 256 bp and two cut G-allele bands of 206 bp and 50 bp in length.

Table 2.3 Table of patient characteristics and their genotype at rs2615977, rs1676486 and rs9659030. F, female; M, male, K, knee, H, hip.

Patient	Sex	Age at surgery (years)	Joint replaced at surgery	Genotype		
				rs2615977	rs1676486	rs9659030
1	F	75	H	TT	CT	AG
2	F	67	K	TT	CC	AA
3	M	69	K	TT	CT	AG
4	F	77	H	TG	CC	AG
5	F	73	K	TT	CC	AA
6	M	71	K	TT	CT	AG
7	F	73	K	GG	CC	AA
8	F	51	H	TT	CT	AG
9	F	60	K	GG	CC	AG
10	F	67	K	TT	CT	AG
11	M	62	K	TT	CT	AG
12	M	50	K	TG	CC	AA
13	M	67	K	TT	CT	AG
14	F	76	H	TT	CC	AA
15	F	71	K	TG	CC	AG
16	F	67	H	TT	CC	AA
17	F	70	H	TT	CC	AA
18	F	60	H	TG	CC	AA
19	M	80	K	TG	CC	AG
20	F	81	K	TG	CC	AG
21	M	76	K	TG	CT	AA
22	F	55	K	TT	CT	AA
23	M	57	K	TT	CT	AA
24	M	57	H	TT	CT	AG
25	M	69	K	TG	CT	GG
26	F	67	K	TT	CC	AA
27	F	69	K	GG	CC	AA
28	M	57	K	GG	CC	AG
29	F	60	K	TG	CC	AA
30	F	66	K	TT	CC	AA
31	M	63	K	TT	CT	AG
32	M	77	K	TT	CC	AA

33	M	82	K	TT	CT	AA
34	F	78	K	TT	CC	AA
35	M	82	K	TG	CT	AG
36	M	46	K	TT	CC	AA
37	M	56	K	TT	CT	AG
38	F	54	K	TT	TT	GG
39	M	71	K	TT	CT	AG
40	F	58	H	TT	CC	AA
41	F	69	H	TG	CT	GG
42	F	64	K	TT	CT	AA
43	M	63	K	TG	CC	AA
44	F	71	H	TG	CC	AG
45	M	70	K	TG	CT	AG
46	M	67	K	TT	CT	AG
47	M	86	K	TT	CT	AA
48	F	67	K	TT	CC	AA
49	M	71	K	TG	CC	AA
50	F	46	K	TT	CC	AA
51	F	62	K	TT	CC	AA
52	F	58	K	TG	CC	AA
53	M	69	K	TG	CC	AG
54	M	59	K	TT	CT	AG
55	M	64	K	TT	CT	GG
56	F	81	K	GG	CC	AG
57	F	80	K	TT	CT	AG
58	F	64	K	TG	CC	AA
59	F	78	K	TG	CC	AG
60	F	61	K	TT	CT	AG
61	F	80	K	TG	CT	GG
62	F	80	K	TT	CC	AA
63	F	59	K	TG	CT	AG
64	F	71	H	TT	CT	AG
65	M	74	K	GG	CC	AG
66	F	74	H	TT	CC	AG
67	M	72	K	TG	CC	AG
68	M	72	K	TT	CC	AA
69	M	68	H	TG	CC	AA

70	F	72	H	TG	CC	AA
71	M	68	K	TG	CC	AA
72	M	75	K	TG	CC	AG
73	F	82	H	TG	CC	AG
74	F	78	K	TG	CC	AG
75	M	72	K	TG	CC	AG
76	F	59	K	TG	CC	AG
77	F	71	H	TG	CC	AG
78	M	69	K	TG	CC	AG

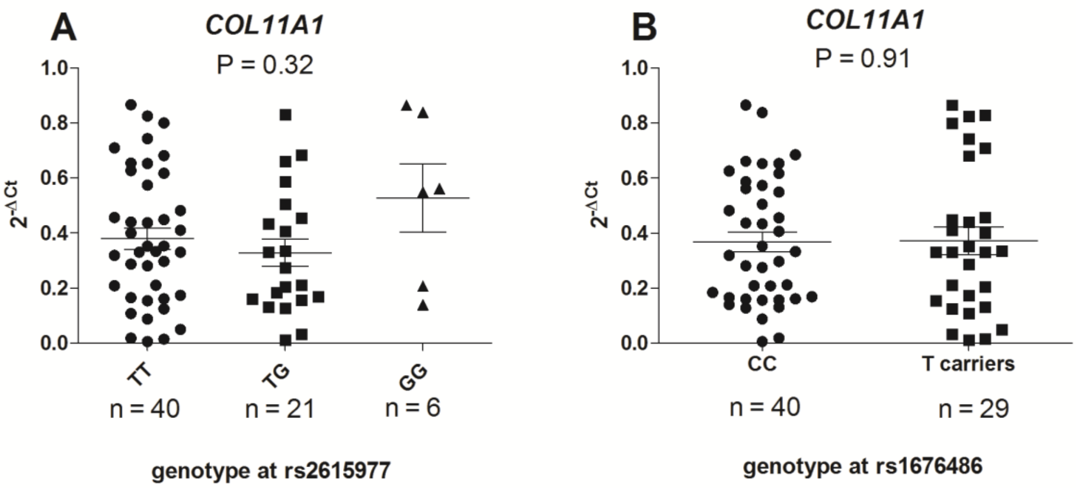


Figure 2.2 Columnar scatter plots of the quantitative expression of *COL11A1* in cartilage cDNA stratified by genotype at A) the OA associated SNP rs2615977 and B) the LDH associated SNP rs1676486. Due to the low frequency of TT homozygotes at rs1676486 the analysis in B) was between CC homozygotes and T-allele carriers (CT and TT combined) at this SNP. n is the number of patients studied for each group. The horizontal lines in each plot represent the mean and the standard error of the mean. P-values were calculated using a Kruskal-Wallis test for A) and a Mann-Whitney U test for B).

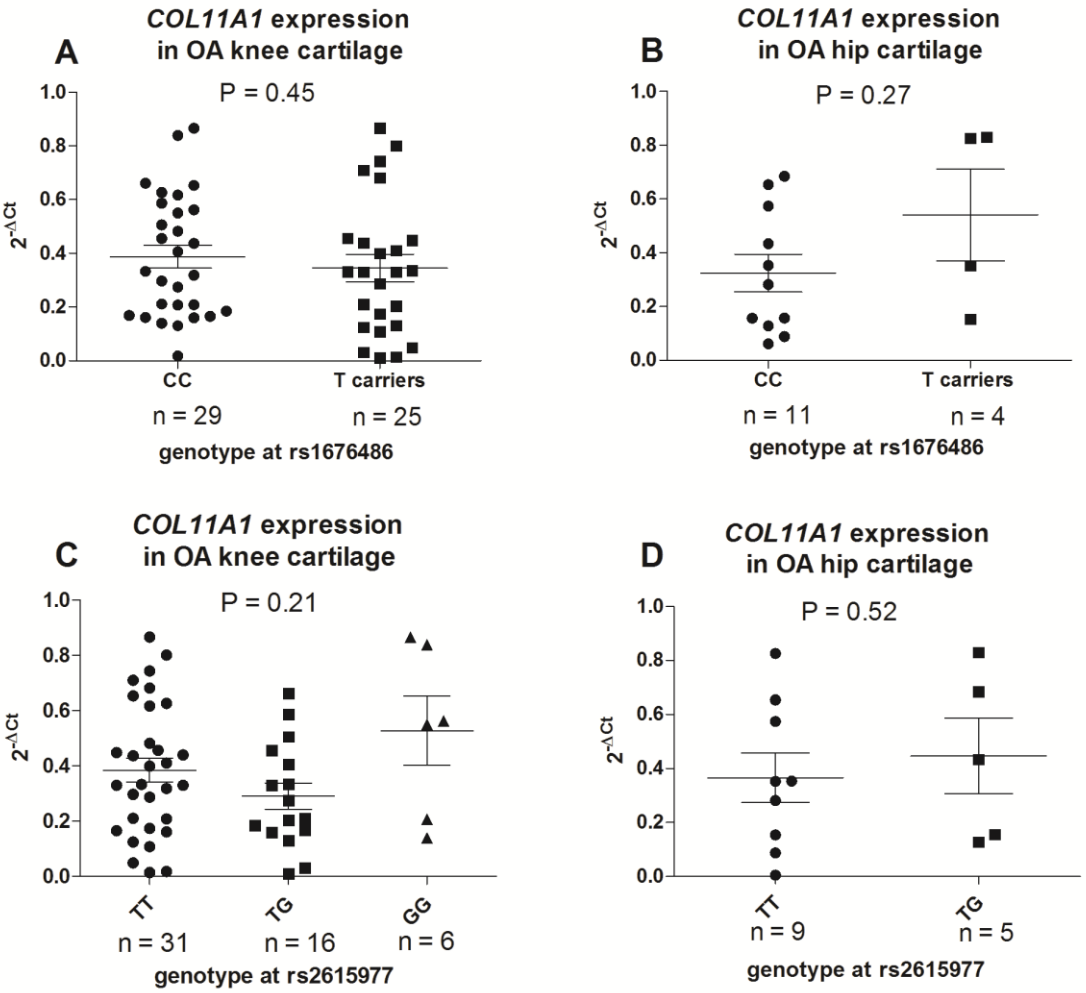


Figure 2.3 Columnar scatter plots of the quantitative expression of COL11A1 in OA cartilage cDNA stratified by A) genotype at the LDH associated SNP rs1676486 in knee cases, B) genotype at the LDH associated SNP rs1676486 in hip cases, C) genotype at the OA associated SNP rs2615977 in knee cases, and D) genotype at the OA associated SNP rs2615977 in hip cases. Due to the low frequency of TT homozygotes at rs1676486 the analysis in A) and B) was between CC homozygotes and T-allele carriers (CT and TT combined) at this SNP. Due to the absence in the hip strata of GG homozygotes at rs2615977 the analysis in D) was between TT homozygotes and TG heterozygotes at this SNP. n is the number of patients studied for each group. The horizontal lines in each plot represent the mean and the standard error of the mean. P-values were calculated using a Mann-Whitney U test for A), B) and D) and a Kruskal-Wallis test for C).

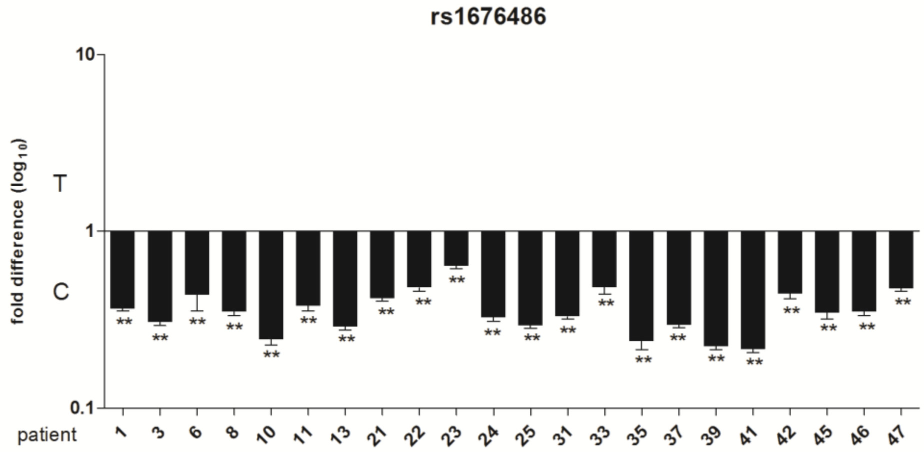
2.4.3 Allelic Expression Imbalance in COL11A1

The technique used in section 3.4.1 did not yield a correlation between *COL11A1* expression and genotype at either SNP. However, as both Figures 2.2 and 2.3 show, there is a large variation in expression between individuals of the same genotype. For example, the expression of *COL11A1* can vary by more than 200 fold between individuals of the same genotype at the targeted SNPs. This indicates that the technique used is vulnerable to natural fluctuation in gene expression which impinges upon the sensitivity and accuracy of the assay (124), potentially creating false negatives which I believe is happening in this instance and measuring gene expression and stratifying by genotype is therefore unable to accurately reflect any imbalances in expression at the allelic level. Instead, one can test directly for AEI by using transcript SNPs to measure the cDNA output from each allele within each heterozygous patient. By only using heterozygotes and testing each allele's output against the other, one can overcome inter-individual fluctuations in overall gene expression. Additionally, by using genomic DNA as a control to provide a 1:1 ratio one is able to normalise the cDNA allelic ratio and overcome any allele specific assay amplification differences.

The LDH associated SNP, rs1676486, is a transcript SNP and therefore can be subjected to direct AEI analysis. However, rs26715977 is within intron 31 of COL11A1 and therefore cannot be used to detect an AEI. Using public databases I was unable to identify any SNPs with an LD of 0.8 or greater with rs26715977, and so opted to use a transcript SNP with a high heterozygosity. By doing this I ensured that I could utilise as many of our patients for the analysis as possible, focusing on individuals who were compound heterozygotes at rs26715977 and the transcript SNP that was chosen, rs9659030. The heterozygosity of rs9659030 is 31% in the European (HapMap CEU) population. rs9659030 is not in LD with rs2615977 (pair wise r^2=0.08), but by testing compound heterozygotes for AEI I was able to use the two alleles of rs9659030 to differentiate between the two alleles of rs2615977. Due to there being no LD between the two SNPs it is not possible to determine the phase between the two SNPs, but should rs2615977 drive, or correlate with a SNP that drives AEI, then one would expect a high proportion of the compound heterozygotes to show either a positive or negative deviation from an allelic expression ratio of 1:1.

2.4.3.1 rs1676486

I tested for AEI on a total of 22 OA patients who were heterozygous at the LDH associated SNP rs1676486. All 22 patients exhibited a significant (p<0.05) decrease in expression of the disease associated T-allele relative to that of the C-allele. The mean allelic ratio of these 22 patients was 0.36 (p<0.0001) (Figure 2.4 and Table 2.4), meaning that there is only 0.36 molecules of T-allele cDNA for every one molecule of C-allele cDNA. This AEI in articular cartilage echoes the findings of the original LDH study where the T-allele was shown to exhibit reduced stability (83).



patient	1	3	6	8	10	11	13	21	22	23	24	25	31	33	35	37	39	41	42	45	46	47
LDH associated SNP rs1676486	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
OA associated SNP rs2615977	TT	TT	TT	TT	TT	TT	TT	TG	TT	TT	TT	TG	TT	TT	TG	TT	TT	TG	TT	TG	TT	TG
3'UTR SNP rs9659030	AG	AG	AG	AG	AG	AG	AG	AA	AA	AA	AG	GG	AG	AA	AG	AG	AG	GG	AA	AG	AG	AA

Figure 2.4 Allelic expression analysis in cartilage samples from OA patients assessed using the LDH associated SNP rs1676486. The genotypes of the patients at rs1676486, at the OA associated SNP rs2615977 and at the 3'UTR SNP rs9659030 are shown. P-values were calculated by comparing the cDNA allelic ratio to the DNA allelic ratio using a Mann-Whitney U test. ** denotes p < 0.01.

Table 2.4 Table of the allelic ratios for the patients analysed for AEI at rs9659030 and rs1676486. Allelic ratios are shown ± the standard error of the mean. P-values were calculated using a Mann Whitney U test. P-values < 0.01 are highlighted in bold. Experiments to assess AEI on patient samples at rs9659030 denoted with a * were carried out by Emma Raine.

rs9659030			rs1676486		
Patient	Allelic ratio (A/G)	P value	Patient	Allelic ratio (C/T)	P value
1*	1.16 ± 0.16	0.42	**1**	**0.36 ± 0.01**	**<0.01**
3*	0.98 ± 0.2	1.00	**3**	**0.31 ± 0.01**	**<0.01**
4	1.08 ± 0.08	0.84	**6**	**0.43 ± 0.08**	**<0.01**
8*	1.09 ± 0.11	0.55	**8**	**0.35 ± 0.02**	**<0.01**
9*	0.94 ± 0.1	1.00	**10**	**0.24 ± 0.02**	**<0.01**
10*	0.8 ± 0.06	0.1	**11**	**0.38 ± 0.02**	**<0.01**
11*	**0.66 ± 0.03**	**<0.01**	**13**	**0.29 ± 0.01**	**<0.01**
13*	0.81 ± 0.07	0.06	**21**	**0.42 ± 0.01**	**<0.01**
15	0.94 ± 0.12	0.69	**22**	**0.48 ± 0.02**	**<0.01**
19	1.16 ± 0.05	0.06	**23**	**0.64 ± 0.02**	**<0.01**
20	**0.81 ± 0.02**	**<0.01**	**24**	**0.33 ± 0.02**	**<0.01**
24*	0.83 ± 0.08	0.22	**25**	**0.29 ± 0.01**	**<0.01**
28*	0.92 ± 0.05	0.55	**31**	**0.33 ± 0.01**	**<0.01**
31*	0.75 ± 0.05	0.15	**33**	**0.48 ± 0.04**	**<0.01**
35	1.35 ± 0.6	0.33	**35**	**0.24 ± 0.03**	**<0.01**
37*	**0.69 ± 0.03**	**<0.01**	**37**	**0.29 ± 0.01**	**<0.01**
39*	**0.66 ± 0.04**	**<0.01**	**39**	**0.22 ± 0.01**	**<0.01**
45	1.2 ± 0.16	0.69	**41**	**0.22 ± 0.01**	**<0.01**
46*	1.06 ± 0.1	0.69	**42**	**0.44 ± 0.03**	**<0.01**
72	1.06 ± 0.12	0.73	**45**	**0.34 ± 0.03**	**<0.01**
73	0.96 ± 0.18	0.57	**46**	**0.35 ± 0.02**	**<0.01**
74*	0.90 ± 0.04	1.00	**47**	**0.47 ± 0.02**	**<0.01**
75*	0.98 ± 0.04	1.00			
76*	0.99 ± 0.21	0.79			
77*	0.93 ± 0.14	0.69			
78*	0.72 ± 0.08	0.55			

2.4.3.2 rs9659030

I tested for AEI on a total of 26 OA patients who were heterozygous at rs9659030 (Figure 2.5 and Table 2.4). Of these 26 patients, 13 were compound heterozygotes at rs2615977 and rs9659030 but only one of these 13 patients exhibited a significant imbalance in allelic expression (Patient 20), demonstrating that the two alleles of the OA associated SNP rs2615977 do not show any correlation with an expression imbalance of COL11A1. Three additional patients exhibited a significant AEI (patients 13, 37 and 39 in Figure 2.5) who, as mentioned, were not heterozygous at rs2615977 but they were all heterozygous at rs1676486 – the LDH associated SNP – and it is most likely this which is driving the imbalance observed.

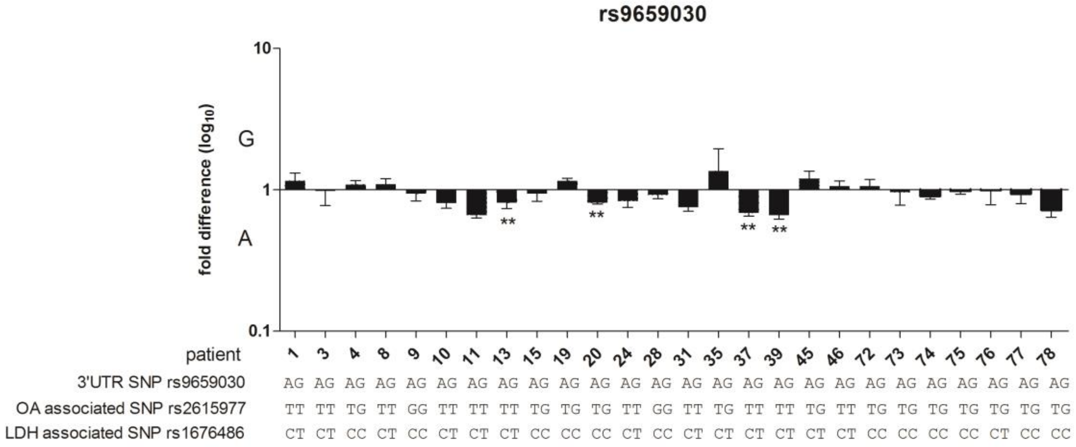


Figure 2.5 Allelic expression analysis in cartilage samples from OA patients assessed using the 3'UTR SNP rs9659030. The genotypes of the patients at rs9659030, at the OA associated SNP rs2615977 and at the LDH associated SNP rs1676486 are shown. P-values were calculated by comparing the cDNA allelic ratio to the DNA allelic ratio using a Mann-Whitney U test. ** denotes $p < 0.01$.

However, it is to be noted that nine patients who are compound heterozygous for rs1676486 and rs9659030 did not exhibit a significant AEI when allelic expression was measured with rs9659030 (patients 1, 3, 8, 10, 11, 24, 31, 46 and 76 in Figure 2.5). Based on the AEI data that I had obtained for rs1676486 (*Section 2.4.3.1*) I had expected all patients who were heterozygous for this SNP to demonstrate AEI, even when the AEI was measured using a SNP other than rs1676486. Eight of these nine patients had undergone AEI testing at rs1676486 in *Section 2.4.3.1*, and exhibited a significant AEI (patients 1, 3, 8, 10, 11, 24, 31, and 46 in Figure 2.4). Mio *et al.* 2007 postulated that the COL11A1 transcript is subjected to differential degradation, and hence a variable stability, dictated by genotype (83). A differential allelic degradation could affect the AEI observed depending upon where on the transcript the AEI was being measured. In order to investigate this further I used the AEI data from the 22 patients tested using rs1676486 and plotted their allelic ratios stratified by their genotype at the 3' UTR SNP rs9659030 (Figure 2.6). When this data was assessed using a Kruskal–Wallis one-way analysis of variance test, it was apparent that genotype at the rs9659030 correlated with the degree of AEI observed at rs1676486, with the G-allele of rs9659030 correlating with a greater magnitude of AEI at rs1676486 (p=0.001).

2.4.4 Genetic association of rs1676486 and OA

Having confirmed rs1676486 as exhibiting AEI in articular OA cartilage I tested whether this SNP was associated with OA. Using the arcOGEN GWAS dataset (80) it was revealed that the SNP showed minimal evidence of an association with OA (Table 2.5), with a comparison of 7 410 cases and 11 009 controls producing a p-value of 0.012. As in the LDH study, the T-allele of rs1676486 showed an elevation in cases versus controls (frequency of 0.201 vs 0.1905) with an odd ratio of 1.069 (95% confidence interval of 1.015-1.127).

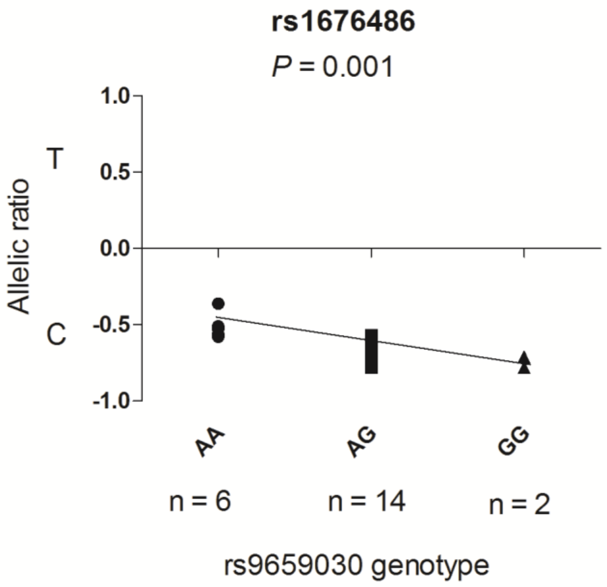


Figure 2.6 Columnar scatter plot of the allelic ratios for AEI analysis of rs1676486 stratified by genotype at the 3'UTR SNP rs9659030. An allelic ratio of 0 represents no difference between the two alleles. A negative allelic ratio value represents increased expression of the C-allele. n is the number of patients analysed for each group. The p-value was calculated using a Kruskal-Wallis test.

Table 2.5 Association analysis of rs1676486 with OA using the arcOGEN data.

Stratum	Number of cases	Number of controls	T-allele frequency		P-value	Odds ratio
			Cases	Controls		(95% confidence intervals)
All cases	7410	11009	0.201	0.1905	0.012	1.069 (1.015-1.127)
Hip cases	3912	11009	0.201	0.1905	0.043	1.069 (1.002-1.141)
Knee cases	4144	11009	0.2035	0.1905	0.010	1.086 (1.02-1.157)

2.5 Discussion

Decreased expression of type XI collagen in *Col11A1* haploinsufficient mice (*cho*/+) can lead to an OA-like phenotype in the joints of these mice due to the altered collagen structure of the articular cartilage and subsequent increase in *Mmp13* expression (120). A key mechanism of OA susceptibility in humans has been shown to be that of allelic expression imbalance, with the 5′ UTR SNP rs143383 of *GDF5* being a prime example whereby the disease associated T-allele confers a reduced level of expression relative to that of the C-allele in articular cartilage and all other tissues of the synovial joint (54). Following investigations such as this, I hypothesised that the OA association to *COL11A1* marked by rs2615977, as reported by the arcOGEN GWAS (82), was due to AEI acting upon *COL11A1* in OA articular cartilage. Additionally, following an investigation into LDH, it was apparent that *COL11A1* expression can be subjected to AEI in intervertebral disc (83) marked by rs1676486 and I was keen to test whether this imbalance is present within OA cartilage. Should this prove to be the case, I also wanted to see if genotype at this SNP associates with OA as it does with LDH. Following our AEI analysis I did not see a correlation between the arcOGEN GWAS OA associated SNP rs2615977 and AEI within *COL11A1*. However, my analysis of the LDH associated SNP rs1676486 did yield a definite correlation between genotype and AEI within *COL11A1*. Like in LDH, it is the T-allele of rs1676486 which corresponds to reduced expression of the gene, however when tested for an association with OA the SNP yielded only a very weak association with the disease. This result stems from a GWAS of over 480 000 SNPs and as a result the very modest p-value does not support rs1676486 as a locus conferring an OA susceptibility, despite it conferring an imbalance in allelic output.

From this study it can be concluded that while *COL11A1* exhibits AEI in OA cartilage it is not a risk factor for the progression of OA, unlike LDH. There are clearly differing mechanistic effects within human articular cartilage and human disc cartilage, whereby an AEI in *COL11A1* can be tolerated in the former but not in the latter. The fact that we see AEI in

COL11A1 mediated by rs1676486 in OA cartilage as it occurs in LDH cartilage implies that the same *trans*-acting factors are at work in both tissue types. When Mio and colleagues investigated the effect of rs1676486 on LDH, they quantified the level of *COL11A1* expression in both intervertebral disc and articular cartilage. They found *COL11A1* expression to be almost 5 times greater in intervertebral disc than in articular cartilage, which could provide an explanation as to why AEI mediated by rs1676486 is not a risk factor for OA, if the ratio of imbalance is the same in both tissues then the actual disparity in the number of allelic molecules would be greater within intervertebral disc.

The association of OA marked by rs2615977 is obviously acting in a different manner than affecting the level of *COL11A1* expression in mature cartilage, as our AEI analysis of the SNP suggests. It is possible that rs2615977 marks an association to a polymorphism which impinges upon *COL11A1* expression at a time point not investigated here. Our OA cartilage samples were from patients with end stage OA electing for total joint replacement. For this reason one cannot rule out the possibility that rs2615977 may induce AEI of *COL11A1* during skeletogenesis or in adolescent tissues before OA is present, affecting the structural integrity of the cartilage as it is laid down or early on in life, initiating the long term progression of OA. However, testing this theory by examining cartilage at these time points is obviously very challenging due to their unavailability. One avenue of investigation would be to take mesenchymal stem cells from the bone marrow of patients undergoing total joint replacement and culturing these cells while differentiating them into chondrocytes. During this differentiation, at various time points RNA could be extracted and tested for AEI.

It should be noted that in nine patients that were compound heterozygotes for rs1676486 and rs9659030, to detect no AEI at rs9659030 when AEI was observed when measuring at rs1676486 was surprising. It is fair to assume that should AEI exist then it would be detectable using any SNP as the marker. However, as Mio *et al.* postulated, rs1676486 alters the stability of the mRNA transcript (83). The data suggests that there is variation in the rate of RNA degradation across the transcript length and that the AEI values obtained could be in part dependent upon the physical position of the SNP used to measure AEI within the transcript, and rs9659030 was located in the 3′ UTR only approximately 360 nucleotides from the end of the transcript.

The mRNA molecules being translated within the cytoplasm of a cell are protected from degradation by a 5′ cap and a polyadenylated 3′ tail (125). The polyadenylated tail is reduced with each cycle of translation, upon the removal of the poly(A) tail the mRNA can be decapped by the DCP1-DCP2 complex. This process then allows for the mRNA to be degraded 5′-3′ by XRN1/XRN2 or by the exosome complex which degrades in a 3′-5′ direction. The presence of AU-rich elements within mRNAs bind proteins which aid the removal of the poly(A) tail, which in turn initiates decapping and the degradation by the exosome complex. As rs9659030 is located closer to the end of the transcript than rs1676486, if mRNA degradation is causing AEI at rs1676486 then it suggests that decapping and subsequent 5′-3′ degrading of the mRNA transcript is affecting the AEI measured. With rs9659030 being closer to the 3′ end of the transcript it would remain longer before being subjected to 5′-3′ degradation.

Chapter 3: Deep Sequencing of GDF5 reveals the absence of rare variants

3.1 Introduction

The hypothesis that common diseases are caused by common variants (126–129) has been at the forefront of thinking in case-control association studies for over a decade. This hypothesis originally dictated that a common polymorphism had a minor allele frequency (MAF) of ≥1%, however in practice researchers often omit SNPs with MAFs of <5% (98). The International HapMap Project (68) is a key resource for improving case-control studies, but currently concentrates on SNPs with MAFs of ≥5%. Rare SNPs are often overlooked during case-control studies due to the reduced statistical power when cohorts are of limited size and also because common SNPs can significantly contribute to disease prevalence even if their effect on disease risk is modest. While case-control association studies have yielded many polymorphisms that affect a person's risk to common diseases, this dominance of common variants in research has deterred genotyping companies from including rare SNPs in coding and regulatory elements into their genotyping panels (98).

As mentioned previously, there are many factors that contribute to the development of OA including BMI, age, sex and genetic susceptibility. The genetic heritability of OA is estimated to be between 39-65% for the hand or knee, 58% for the hip and 63-79% for the spine (9). However, few loci have achieved a genome-wide significance ($p \leq 5x10^{-8}$) of association in Europeans, *GDF5* being one such loci. These loci have small effect sizes and by no means account for the overall heritability attributed to OA.

Gorlov et al. (2008) suggests that within the rare SNPs category lie slightly deleterious SNPs, SNPs which are subjected to weak purifying selection. They also hypothesised that slightly deleterious SNPs in genes implicated in disease pathways are mainly responsible for the inter-individual variation in susceptibility to complex disease. This may be due “to the summation of the effects of a series of low frequency dominantly and independently acting variants of a variety of different genes, each conferring a moderate but readily detectable increase in relative risk” (99). These SNPs have lower MAFs than common variants (which have MAFs of 1% or more), but they have higher MAFs than clearly deleterious SNPs (which have MAFs of 0.1% or less). Slightly deleterious SNPs are in large numbers in the human genome, with the rare allele disrupting gene function and dramatically increasing susceptibility. However, their actions are not strong enough to be eradicated from the population (98, 99). The same can be said of course for slightly advantageous SNPs, where their actions are not strong enough to become fixed within a population. OA is of course characteristically a late onset disease, when reproduction has been completed and so risk alleles are not affected by natural selection purely on their basis of disease risk. It has been estimated that within the global population there should be on average at least two to three slightly deleterious SNPs per gene in the human genome (98), which leads to the idea that the discovery of these rare variants can radically improve the understanding of common complex disease susceptibility. Such SNPs can be retained in the population by random genetic drift or population bottlenecks, however, as such SNPs are likely to be rather novel, evolutionarily speaking, then they are likely to be highly population specific (98, 99). A strong criticism of the notion that slightly deleterious rare SNPs are responsible for common diseases is that their prevalence does not in any way reflect that of the disease in question (98). This argument can be countered in that the slightly deleterious SNPs may be rare, but on the whole such SNPs are very common throughout the genome, particularly if the Gorlov *et al.'*s prediction of there being two to three rare variants within each gene of the genome is correct. Conversely to this, rare variants may be slightly protective rather than deleterious, and such variants would have MAFs that were elevated in controls rather than cases (130).

The detection of rare causal variants is not possible using common tagSNPs in a GWAS as the weak correlations between the rare and common SNPs leave the study low-powered (97). Instead the discovery of rare variants must be achieved by the re-sequencing of candidate genes in a cohort of both cases and controls. If variants are slightly deleterious then they should be elevated in cases, where as if they are protective they should be elevated in controls. For this reason it is important to re-sequence both groups, and not re-sequence cases and then merely genotype controls (97).

As mentioned previously *GDF5* harbours a common SNP, rs143383, who's T-allele has been shown to associate with OA. This association signal has been replicated in both Japanese (52) and European (55) populations. This makes *GDF5* a good candidate for re-sequencing to look for rare variants which could account for some of the missing heritability in OA. One such example of this approach in another disease is that of the resequencing of interferon induced with helicase C domain 1 (*IFIH1*), a gene located within a genetic region shown by GWAS to be associated with type 1 diabetes (130). In this study, four rare variants were discovered with MAFs ranging from 0.46-1.1% in patients with type 1 diabetes. These MAFs were all elevated in controls. The variants all associated with disease; with the rare alleles consistently protecting against type 1 diabetes, and the common allele carried by the majority of the population predisposing the carrier to type 1 diabetes.

3.2 Aims of this study

The aim of this study is to interrogate the genetic architecture of this important OA locus in both OA cases and symptom free controls. This was to be achieved by resequencing *GDF5*, including its promoter, UTRs and open reading frame in 502 OA and 460 controls, looking for novel variants and cataloguing the allele frequencies of common known polymorphisms.

3.3 Materials and methods

3.3.1 Cases and controls

There were a total of 502 cases (383 females and 119 males; 220 knee cases and 282 hip) and 460 controls (184 females and 276 males). These cases were ascertained using the criteria of signs and symptoms of OA sufficiently severe to require joint replacement surgery. The radiological stage of the disease was a Kellgren and Lawrence grade of 2 or more in all cases with over 90% beng grade 3 or 4. Inflammatory arthritis (rheumatoid, polyarthritic or autoimmune disease) was excluded, as was post-traumatic or post-septic arthritis. The cases had an age range of 56-85 years. The controls had no signs or symptoms of arthritis or joint disease (pain, swelling, tenderness or restriction of movement) and had an age range of 55-89 years. Ethical approvals for the use of the DNAs in OA genetic studies was obtained from the Oxfordshire Clinical Research Ethics Committee (COREC reference 541).

3.3.2 Sequence analysis of GDF5

DNA was extracted from peripheral blood samples using guanidine hydrochloride by members of my supervisor's research group prior to my undertaking of this study. Deep sequencing of *GDF5* was carried out by resequencing the 962 cases and controls (1 924 chromosomes) using six PCR amplicons. Genomic DNA was first amplified by PCR in a 15 µl reaction containing 50 ng DNA, along with 7.5 µM of forward and 7.5 µM of reverse primer, 1.5 µl of 10X PCR buffer (50 mM KCl, 10 mM Tris-HCl pH 8.3), 1 or 2 mM $MgCl_2$ (depending upon primer pairs used), 0.2 mM dNTPs, 0.08 units of AmpliTaq Gold DNA polymerase (Applied Biosystems) and the total volume was brought up to 15 µl using H_2O. For amplicons 1, 2 and 4-6 specific M13 phage DNA sequence was included to the 5′ end of the forward and reverse primers. Incorporating these M13 sequences into the amplicons allowed for the same sequencing primers and the same sequencing reaction conditions to be used when amplicons 1, 2 and 4-6 were sequenced. This was found not to be of benefit for amplicon 3, which was sequenced without the addition of the M13 sequences to the PCR primers. For amplicon 4 the use of PCR primers at a concentration of 0.75 µM together with the addition of 25% betaine (Sigma Aldrich) to the PCR reaction was needed to generate an intense PCR product. Thermocycling conditions were 94°C for 4 minutes, followed by a denaturation step of 94°C for 30 seconds, a specific annealing temperature step for 30 seconds, an extension step of 72°C for 1 minute, cycled 35-40 times. Table 3.1 lists the sequences of the PCR primers, the annealing temperatures and $MgCl_2$ concentrations used in the PCRs, and the length of the amplified products. After amplification and prior to the sequencing chemistry reaction, 10 µl of PCR product was incubated with 2.4 units of Shrimp Alkaline Phosphatase (SAP) (Fisher Scientific UK), 12 units of Exonuclease I (EXO) (New England Biolabs) and 1.4 µl of 10X Exonuclease I buffer (New England Biolabs) for 15 minutes at 37°C followed by another 15 minute incubation at 80°C. The sequencing chemistry reaction was performed with the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). 6 µl of the SAP/EXO treated PCR product was added to 0.5 µl of BigDye premix v3.1, 2 µl of 5X BigDye v3.1 Sequencing Buffer, 1.2 µl of H_2O and 30 µM of M13 reverse primer (amplicons 1,2, 5 and 6), or 30 µM of Amp 3 7F forward primer for amplicon 3 (Table 3.1). Amplicon 4 contained 240 µM M13 reverse primer. This reaction mixture was incubated for 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds, and 60°C for 4 minutes with temperature ramping at 1°C/second.

To precipitate the 10 µl of sequencing products, 10 µl of H_2O were added along with 60 µl of 100% ethanol and 0.1 µl of Pellet Paint Non-Fluorescent Co-Precipitant (Novagen) with the mixture then incubated for 30 minutes at room temperature and in the dark. After the incubation, the products were centrifuged at room temperature at 1 600 *g* for 45 minutes followed by the removal of the supernatant via an inversion of the sample and brief centrifuge ramping up to 50 *g*. To each pellet, 70 µl of 70% (volume/volume) ethanol was added and this mixture was again centrifuged at 1 600 *g*, this time for 15 minutes. After removal of the supernatant, the pellet was resuspended in 10 µl of Hi-Di formamide (Applied Biosystems) and

10 µl of H_2O. These products were then separated by electrophoresis through a 36 cm capillary array with POP-7 polymer (Applied Biosystems) at 55°C using an Applied Biosystems 3130xl Genetic Analyzer. These data were then analysed with the Sequencing Analysis and Seqscape software (Applied Biosystems).

3.3.3 Verification of novel variants

In order to verify the novel variants discovered, new sequencing primers were designed and each relevant patient was re-sequenced in the opposite strand to the strand that was sequenced when the variant was discovered. The primers used along with their PCR conditions are listed in Table 3.2. In addition to re-sequencing, a restriction fragment length polymorphism (RFLP) assay was designed to further confirm the existence of a variant discovered in amplicon 1. An elongated forward primer with a single base change (underlined) was synthesised in order to force a *Hinf*I restriction site (5′-AAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGCGATT-3′). This primer was used in conjunction with a standard reverse primer (5′-CCGGGTGTGTGTTTGTATCCAG-3′). PCR was carried out as described in *Section 3.3.2*, using an annealing temperature of 64°C and a $MgCl_2$ concentration of 2 mM. The *Hinf*I restriction site of GANTC allows *Hinf*I to cut the wildtype C-allele producing two fragments of 190 bp and 50 bp in length, where as a mutant allele would remain uncut with a single 240 bp undigested band (Figure 3.1). Five units of *Hinf*I were added to a 15 µl PCR in addition to 2 µl of 10X Buffer 2 (New England Biolabs) and the reaction made up to 20 µl with H_2O. This digestion mix was then incubated at 37°C for 3 hours and then electrophoresed on a 3% (weight/volume) agarose gel.

```
TGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAAGTATTTTC
ACTGGAAAGGATTCAAAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGC
GATTCTTCCTTCTTGGAAAAATCCCTCAGCCTTATACAAGCCTCCTTCAAGCCCTCAGTC
AGTTGTGCAGGAGAAAGGGGGCGGTTGGCTTTCTCCTTTCAAGAACGAGTTATTTTCAGC
TGCTGACTGGAGACGGTGCACGTCTGGATACGAGAGCATTTCCACTATGGGACTGGATAC
AAACACACACCCGGCAGACTTCAAGAGTCTCAGACTGAGGAGAAAGCCTTTCCTTCTGCT
GCTACTGCTGCTGCCGCTGCTTTTGAAAGTCCACTCCTTTCATGGTTTTTCCTGCCAAAC
CAGAGGCACCTTTGCTGCTGCCGCTGTTCTCTTTGGTGTCATTCAGCGGCTGGCCAGAGG
ATGAGACTCCCCAAACTCCTCACTTTCTTGCTTTGGTACCTGGCTTGGCTGGACCTGGAA
```

```
TGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAAGTATTTTC
ACTGGAAAGGATTCAAAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGC
GATTATTCCTTCTTGGAAAAATCCCTCAGCCTTATACAAGCCTCCTTCAAGCCCTCAGTC
AGTTGTGCAGGAGAAAGGGGGCGGTTGGCTTTCTCCTTTCAAGAACGAGTTATTTTCAGC
TGCTGACTGGAGACGGTGCACGTCTGGATACGAGAGCATTTCCACTATGGGACTGGATAC
AAACACACACCCGGCAGACTTCAAGAGTCTCAGACTGAGGAGAAAGCCTTTCCTTCTGCT
GCTACTGCTGCTGCCGCTGCTTTTGAAAGTCCACTCCTTTCATGGTTTTTCCTGCCAAAC
CAGAGGCACCTTTGCTGCTGCCGCTGTTCTCTTTGGTGTCATTCAGCGGCTGGCCAGAGG
ATGAGACTCCCCAAACTCCTCACTTTCTTGCTTTGGTACCTGGCTTGGCTGGACCTGGAA
```

Figure 3.1 *HinfI* RFLP assay. Top) GDF5 proximal promoter containing the wildtype C-allele of the -41 bp variant (C – large bolded) and immediate 5′ sequence (highlighted orange). Forward RFLP primer (highlighted green) with forced base change (red and underlined G) and reverse primer (highlighted blue). The black box indicates a positive site for *HinfI* digestion. Below) As above, but variant A-allele of the -41 bp variant (A- large bolded). No *HinfI* digestion site is present.

Table 3.1 Primer sequences and PCR conditions used . Underlined bases denote *GDF5* specific sections of each primer. Also listed are the sequences of the M13 primers used in the sequencing reactions of amplicons 1, 2 and 4-6.

Amplicon	Primer name	Primer Sequence (5' - 3')	$MgCl_2$ Concentration	Annealing Temperature	Number of Cycles	Amplicon Length
1	Amp 1 1F	TGTAAAACGACGGCCAGTATTTTCACTGGAAAGGATTC				
	Amp 1 1R	CAGGAAACAGCTATGACCAGGGGCACCCAACACAGTGC	1 mM	66^{o}C	35	513
2	Amp 2 3F	TGTAAAACGACGGCCAGTACACCCGGCAGACTTCAAGAG				
	Amp 2 3R	CAGGAAACAGCTATGACCGACAGATCCTGCTTTTGGGGG	1 mM	70^{o}C	35	540
3	Amp 3 7F	GCCCGGAACGTCTTCAGGCCAG				
	Amp 3 7R	TGCCAGGGCTTTGAAAGCCCCT	1 mM	64^{o}C	35	551
4	Amp 4 2F	TGTAAAACGACGGCCAGTTCGAAGTGACTGGCTCCCTTG				
	Amp 4 3R	CAGGAAACAGCTATGACCTGGAAAGCCTCGTACTCAAGG	2 mM	56^{o}C	40	712
5	Amp 5 1F	TGTAAAACGACGGCCAGTCTGGCCAGGACGATAAGACC				
	Amp 5 2R	CAGGAAACAGCTATGACCGAGTCTGTCTCCCTGGACCTG	2 mM	68^{o}C	35	679
6	Amp 6 2F	TGTAAAACGACGGCCAGTTTCCTGCACTCCTGGAATCAC				
	Amp 6 1R	CAGGAAACAGCTATGACCGATAGGGCTCTGAAAACTGAG	2 mM	66^{o}C	35	630
	M13-F	TGTAAAACGACGGCCAGT	N/A	50^{o}C	25	
	M13-R	CAGGAAACAGCTATGACC	N/A	50^{o}C	25	

Table 3.2 Primer sequences and PCR conditions used for variant verification. Underlined bases denote *GDF5* specific sections of each primer.

		$MgCl_2$	Annealing	Number of	
Variant	Primer Sequence (5' - 3')	Concentration	Temperature	Cycles	
1	TGTAAAACGACGGCCAGTGATTTTTCTGAGCACCTGCAG				
	CAGGAAACAGCTATGACCGCAGTAGCAGCAGAAGGAAAG	1 mM	62°C	35	
2, 3	TGTAAAACGACGGCCAGTCACTCCTTTCATGGTTTTTCC				
	TGTAGCCTGCCTTGTTTGGG	2mM	60°C	35	
4, 5	TGTAAAACGACGGCCAGTTCGAAGTGACTGGCTCCCTTG				25%
	CAGGAAACAGCTATGACCAGAACAGGTCCCGTTTCTTGGTG	3 mM	59°C	35	Betaine
6	TGTAAAACGACGGCCAGTAAGGCACTGCATGTCAACTTC				
	CAGGAAACAGCTATGACCGAGTCTGTCTCCCTGGACCTG	2 mM	66°C	35	

3.3.3 Online prediction databases

In order to assess the implications of a variant discovered within the promoter of *GDF5,* two online prediction databases were used; Promo 3.0 (http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3) and TESS: Transcription Element Search System (http://www.cbil.upenn.edu/cgi-bin/tess/tess). To assess the implications of a non-synonymous variant, PolyPhen (http://genetics.bwh.harvard.edu/pph2/) was used, which gave a prediction on the effect of the variant on protein function ranging from benign to seriously damaging.

3.3.4 Protein modelling

To model and mutate a non-synonymous variant I used a program called Coot (131) in collaboration with Professor Rick Lewis of the Institute of Cell and Molecular Biosciences at Newcastle University. The images from this modelling were created using the program PyMOL (pymol.sourceforge.net/).

3.4 Results

In order to cover both exons of GDF5, as well as the proximal promoter and the exon/intron boundaries, six PCR amplicons were designed. These six PCR amplicons covered 111 bp of the proximal promoter, the whole 2 383 bp transcript including both 5' and 3' UTRs as well as 76 bp of intronic sequence following exon one, 60 bp of intronic sequence before exon two and 114 bp of intergenic sequence following exon 2. These amplicons varied both in length and ease of amplification, with the shortest being 513 bp and the largest being 712 bp (Figures 3.2 and 3.3).

```
            ATTCCGTTTCCAATTCCTGAGTTCAGGTTTGTAAAAGATTTTTCTGAGCACCTGCAGGCC
            TGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAAGTATTTTC
Amp1 1F     ACTGGAAAGGATTCAAAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGC
            CATTCTTCCTTCTTGGAAAAATCCCTCAGCCTTATACAAGCCTCCTTCAAGCCCTCAGTC
            AGTTGTGCAGGAGAAAGGGGGCGGTTGGCTTTCTCCTTTCAAGAACGAGTTATTTTCAGC
            TGCTGACTGGAGACGGTGCACGTCTGGATACGAGAGCATTTCCACTATGGGACTGGATAC
            AAACACACACCCGGCAGACTTCAAGAGTCTCAGACTGAGGAGAAAGCCTTTCCTTCTGCT  Amp2 3F
            GCTACTGCTGCTGCCGCTGCTTTTGAAAGTCCACTCCTTTCATGGTTTTTCCTGCCAAAC
            CAGAGGCACCTTTGCTGCTGCCGCTGTTCTCTTTGGTGTCATTCAGCGGCTGGCCAGAGG
            ATGAGACTCCCCAAACTCCTCACTTTCTTGCTTTGGTACCTGGCTTGGCTGGACCTGGAA
EXON 1      TTCATCTGCACTGTGTTGGGTGCCCCTGACTTGGGCCAGAGACCCCAGGGGACCAGGCCA  Amp1 1R
            GGATTGGCCAAAGCAGAGGCCAAGGAGAGGCCCCCCCTGGCCCGGAACGTCTTCAGGCCA  Amp3 7F
            GGGGGTCACAGCTATGGTGGGGGGGCCACCAATGCCAATGCCAGGGCAAAGGGAGGCACC
            GGGCAGACAGGAGGCCTGACACAGCCCAAGAAGGATGAACCCAAAAAGCTGCCCCCCAGA
            CCGGGCGGCCCTGAACCCAAGCCAGGACACCCTCCCCAAACAAGGCAGGCTACAGCCCGG
            ACTGTGACCCCAAAAGGACAGCTTCCCGGAGGCAAGGCACCCCCAAAAGCAGGATCTGTC  Amp2 3R
            CCCAGCTCCTTCCTGCTGAAGAAGGCCAGGGAGCCCGGGCCCCCACGAGAGCCCAAGGAG
            CCGTTTCGCCCACCCCCCATCACACCCCACGAGTACATGCTCTCGCTGTACAGGACGCTG
            TCCGATGCTGACAGAAAGGGAGGCAACAGCAGCGTGAAGTTGGAGGCTGGCCTGGCCAAC
            ACCATCACCAGCTTTATTGACAAAGGGCAAGGTGAGGGGGCGGGGTGGCAGGGGCACGGC
            TCAGAGGGAGGGGCATCTGCATGAATGGAGGGGCTTTCAAAGCCCTGGCACTGCCCTGGT  Amp3 7R
            GGGAGACACTGTGTGCATCTGGCCCGGGGTGGGTGTCTGGGGACACTGACACATATCTCA
            CACATAGCAGATCCTAGGCTGCCCACACATCAGGGCTGGAGGGCACCTAGGAAGCACCAG

            CAGAATGGGGCAGAGGTGAAAGAAAGCTCTCTGGACGGGGAGAGAGCTGGGGCCTTCCCC
            CGCAGCCTCGAAGTGACTGGCTCCCTTGGTGAGGTTGCAGGGAATGACTTCTGGGTGTTC  Amp4 2F
            TCTCTAGATGACCGAGGTCCCGTGGTCAGGAAGCAGAGGTACGTGTTTGACATTAGTGCC
            CTGGAGAAGGATGGGCTGCTGGGGGCCGAGCTGCGGATCTTGCGGAAGAAGCCCTCGGAC
            ACGGCCAAGCCAGCGGCCCCCGGAGGCGGGCGGGCTGCCCAGCTGAAGCTGTCCAGCTGC
            CCCAGCGGCCGGCAGCCGGCCTCCTTGCTGGATGTGCGCTCCGTGCCAGGCCTGGACGGA
            TCTGGCTGGGAGGTGTTCGACATCTGGAAGCTCTTCCGAAACTTTAAGAACTCGGCCCAG
            CTGTGCCTGGAGCTGGAGGCCTGGGAACGGGGCAGGGCCGTGGACCTCCGTGGCCTGGGC
            TTCGACCGCGCCGCCCGGCAGGTCCACGAGAAGGCCCTGTTCCTGGTGTTTGGCCGCACC
            AAGAAACGGGACCTGTTCTTTAATGAGATTAAGGCCCGCTCTGGCCAGGACGATAAGACC  Amp5 1F
            GTGTATGAGTACCTGTTCAGCCAGCGGCGAAAACGGCGGGCCCCACTGGCCACTCGCCAG
            GGCAAGCGACCCAGCAAGAACCTTAAGGCTCGCTGCAGTCGGAAGGCACTGCATGTCAAC
            TTCAAGGACATGGGCTGGGACGACTGGATCATCGCACCCCTTGAGTACGAGGCTTTCCAC  Amp4 3R
EXON 2      TGCGAGGGGCTGTGCGAGTTCCCATTGCGCTCCCACCTGGAGCCCACGAATCATGCAGTC
            ATCCAGACCCTGATGAACTCCATGGACCCCGAGTCCACACCACCCACCTGCTGTGTGCCC
            ACGCGGCTGAGTCCCATCAGCATCCTCTTCATTGACTCTGCCAACAACGTGGTGTATAAG
            CAGTATGAGGACATGGTCGTGGAGTCGTGTGGCTGCAGGTAGCAGCACTGGCCCTCTGTC
            TTCCTGGGTGGCACATCCCAAGAGCCCCTTCCTGCACTCCTGGAATCACAGAGGGGTCAG  Amp6 2F
            GAAGCTGTGGCAGGAGCATCTACACAGCTTGGGTGAAAGGGGATTCCAATAAGCTTGCTC
            GCTCTCTGAGTGTGACTTGGGCTAAAGGCCCCCTTTTATCCACAAGTTCCCCTGGCTGAG
            GATTGCTGCCCGTCTGCTGATGTGACCAGTGGCAGGCACAGGTCCAGGGAGACAGACTCT  Amp5 2R
            GAATGGGACTGAGTCCCAGGAAACAGTGCTTTCCGATGAGACTCAGCCCACCATTTCTCC
            TCACCTGGGCCTTCTCAGCCTCTGGACTCTCCTAAGCACCTCTCAGGAGAGCCACAGGTG
            CCACTGCCTCCTCAAATCACATTTGTGCCTGGTGACTTCCTGTCCCTGGGACAGTTGAGA
            AGCTGACTGGGCAAGAGTGGGAGAGAAGAGGAGAGGGCTTGGATAGAGTTGAGGAGTGTG
            AGGCTGTTAGACTGTTAGATTTAAATGTATATTGATGAGATAAAAAGCAAAACTGTGCCT
            AAAACTGTGGCAAATTTCTTGTTTACTCCAGGGGACCCAGTGACCCCTTGAAGAATTACT
            GACCCAGGGCAGGGAGGTGGGTTGCTTCATGTATAAACTCAGTTTTCAGAGCCCTATCCC  Amp6 1R
```

Figure 3.2 *GDF5* Sequencing amplicons. All six sequencing amplicons are shown with each colour depicting a primer pair. Green shows amplicon 1 primers, blue shows amplicon 2 primers, violet shows amplicon 3 primers all of which cover the 5’ UTR and exon 1 (upper highlighted section) and its flanking sequence. The ATG start codon is shown in black. In exon 2 and the 3’ UTR (lower highlighted section), pink shows amplicon 4 primers, yellow shows amplicon 5 primers and red shows amplicon 6 primers. The TAG stop codon is shown in black.

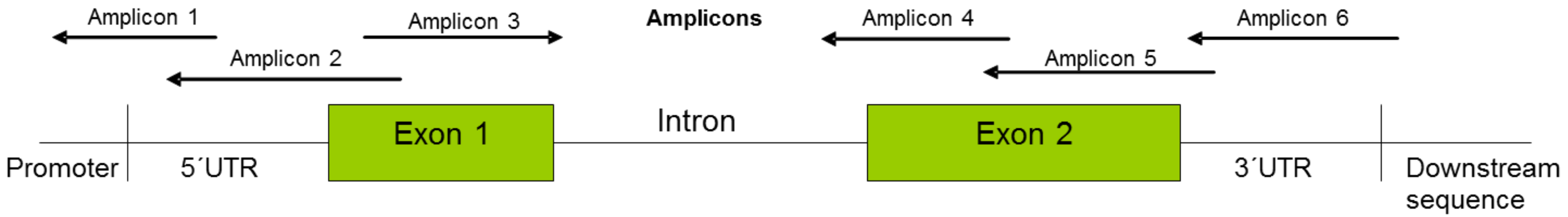


Figure 3.3 Sequenced areas of *GDF5* and amplicon locations. Schematic showing the areas of GDF5 covered by each amplicon. Arrows depict the direction the sequence was generated, relative to transcription.

3.4.1 Amplicon 1

Amplicon 1, once amplified and sequenced, produced an average read length of 354 bp. Successful sequencing data was generated from 458 OA cases and 418 controls (1 752 chromosomes). In addition to common catalogued SNPs, a novel variant was discovered in a female control individual from this cohort (Figure 3.4, Panel A). This mutation, a C to A transversion, resides 41 bp upstream of the transcription start site of *GDF5*. According to the online database PROMO, this -41 bp variant lies within several possible *trans*-acting factor binding sites, including YY1, UCRBP, SOX9, SEF-4, VDR and abaA (Figure 3.5). A RFLP assay (Figure 3.4, Panel D) and resequencing in the reverse direction using new primers that reside distal to the original amplicon 1 PCR primers, were used to confirm the mutation (Figure 3.4, Panel B). An unrelated control individual was also sequenced to demonstrate the wildtype sequence (Figure 3.4, Panel C).

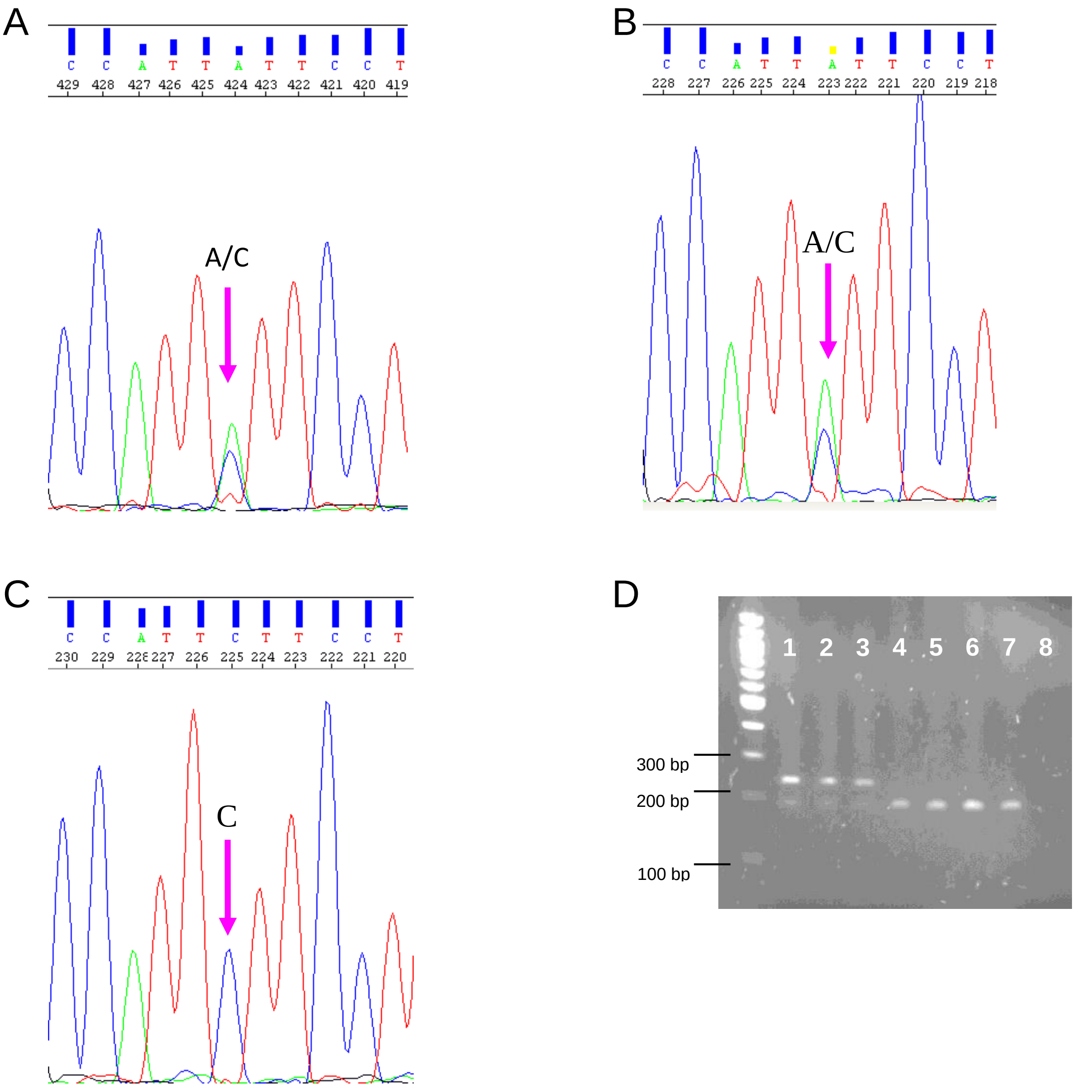


Figure 3.4 The -41 bp mutation identified in a female control individual (variant 1). (A) Electropherogram showing two alleles, A and C, following sequencing of amplicon 1 in the female control. Sequence is shown along the top along with a bar height corresponding to the confidence in which the Sequencing Analysis program calls each base. Blue corresponds to the greatest confidence, yellow is low confidence and red is zero confidence. (B) Electropherogram showing the same double peak when the individual was resequenced using a new primer pair. (C) Electropherogram showing only a single allele present in an unrelated individual when sequenced using the same conditions as in B. (D) Agarose gel image of an RFLP assay along with a 100 bp ladder, carried out when DNA from the individual carrying the variant (three replicates, lanes 1-3) and four unrelated individuals (lanes 4-7) were PCR amplified and then digested with *HinfI* restriction enzyme, which cuts only when the C allele is present (lane 8 is a PCR blank).

```
TGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAAGTATTTTC
ACTGGAAAGGATTCAAAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGC
CATTCTTCCTTCTTGGAAAAATCCCTCAGCCTTATACAAGCCTCCTTCAAGCCCTCAGTC
AGTTGTGCAGGAGAAAGGGGGCGGTTGGCTTTCTCCTTTCAAGAACGAGTTATTTTCAGC
TGCTGACTGGAGACGGTGCACGTCTGGATACGAGAGCATTTCCACTATGGGACTGGATAC
AAACACACACCCGGCAGACTTCAAGAGTCTCAGACTGAGGAGAAAGCCTTTCCTTCTGCT
```

```
CGCCATTCTTCCTT           CGCCATTCTTCCTT
  CCATTTT - YY1            GGATCCTGTCG - SOX9
CGCCATTATTCCTT           CGCCATTATTCCTT

CGCCATTCTTCCTT           CGCCATTCTTCCTT
CGCCATTCT - UCRBP            CATTCTT - SEF-4
CGCCATTATTCCTT           CGCCATTATTCCTT

CGCCATTCTTCCTT           CGCCATTCTTCCTT
   CATTCTTCC - VDR           CATTCT - abaA
CGCCATTATTCCTT           CGCCATTATTCCTT
```

Figure 3.5 Predicted binding *trans*-factors. Top) *GDF5* proximal promoter sequence (non-highlighted) and initial 5' UTR sequence (highlighted) with the -41 bp wildtype allele highlighted green. Below) -41 bp wildtype allele (geeen), variant (blue) and immediate sequence surrounding site shown next to predicted binding *trans*-acting factors and their consensus sequences, as given from the PROMO transcription factor binding prediction database.

3.4.2 Amplicon 2

Amplicon 2 produced an average read length of 457 bp, including a degree of overlap with amplicon 1. Successful sequencing data was generated from 491 OA cases and 425 controls (1 832 chromosomes). From these individuals, two mutations were discovered. One was a G to A transition within the coding region of exon 1 in a female OA patient (Figure 3.6, Panel A). This transition is synonymous. This mutation was verified by resequencing of the region with new primers (Figure 3.6, Panel B). The second mutation was also found in the coding region of exon 1, in a male OA patient. This mutation, a G to A transition (Figure 3.7, Panel A), is predicted to cause an amino acid substitution at position 81 of the GDF5 protein. The wild type glycine residue is neutral and non-polar whereas the substitution would be to an arginine residue which is polar and positively charged. The glycine residue is highly conserved in mammals (Figure 3.8). The variant was also verified using the new primers. This non-synonymous mutation was however predicted to be benign by the PolyPhen protein prediction database.

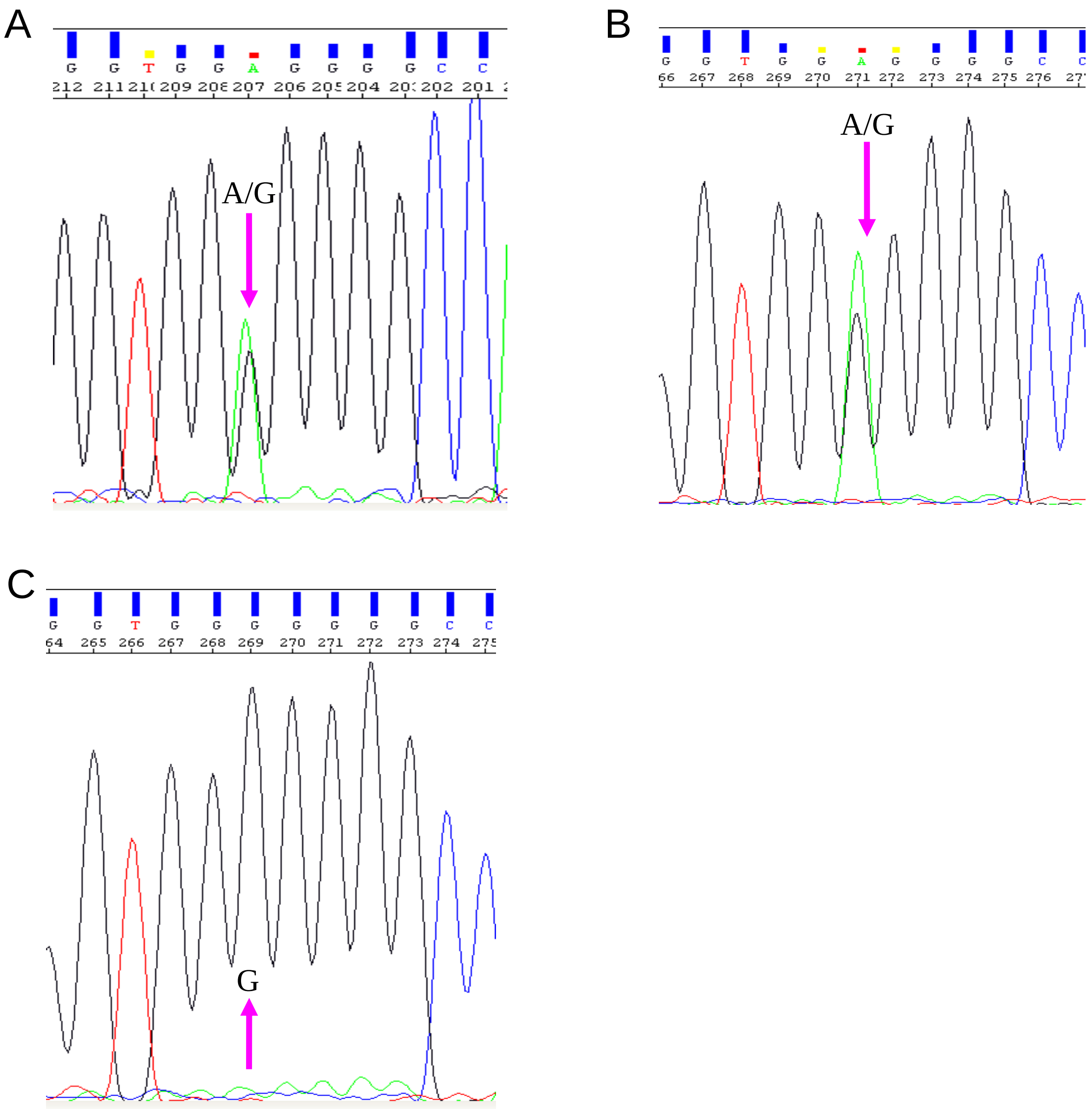


Figure 3.6 The exon 1 synonymous mutation identified in a female OA patient (variant 2). (A) Electropherogram showing two alleles, A and G, following sequencing of amplicon in the female OA patient. Sequence is shown along the top along with a bar height corresponding to the confidence in which the Sequencing Analysis program calls each base. Blue corresponds to the greatest confidence, yellow is low confidence and red is zero confidence. (B) Electropherogram showing the same double peak when the individual was resequenced using a new primer pair. (C) Electropherogram showing only a single wildtype allele, G, present in an unrelated individual when sequenced using the same conditions as in B.

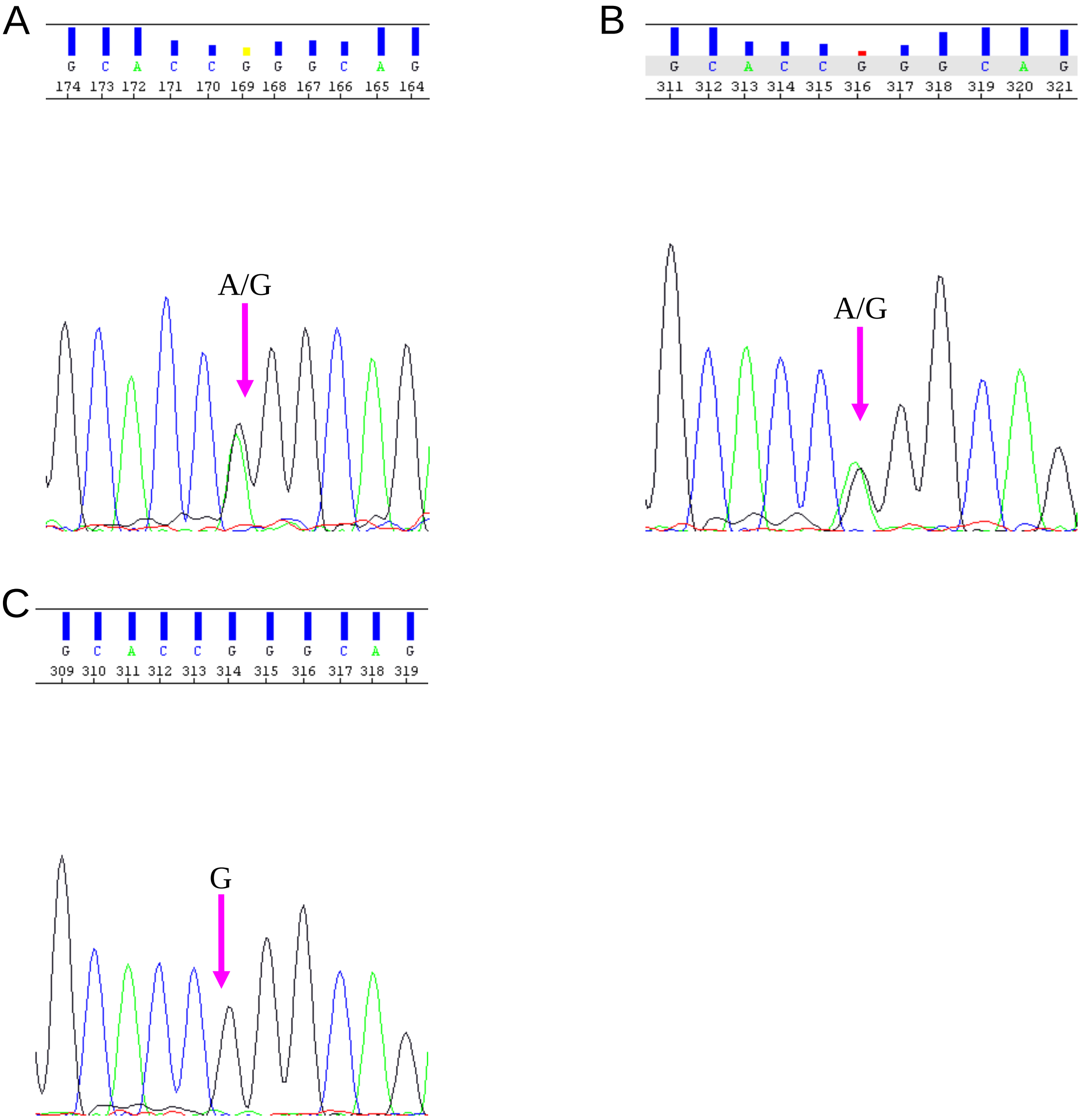


Figure 3.7 The exon 1 non-synonymous mutation identified in a male OA patient (variant 3). (A) Electropherogram showing two the alleles, A and G, following sequencing of amplicon 2 in the male OA patient. Sequence is shown along the top along with a bar height corresponding to the confidence in which the Sequencing Analysis program calls each base. Blue corresponds to the greatest confidence, yellow is low confidence and red is zero confidence. (B) Electropherogram showing the same double peak when the individual was resequenced using a new primer pair. (C) Electropherogram showing only the single wildtype allele, G, present in unrelated individual when sequenced using the same conditions as in B.

```
Human           GGHSYGGGATNANARAKGGTGQTGGLTQPKKDEPKKLPPRPGGPEPKPGHPPQTRQATAR 120
Gorilla         GGHSYGGGATNANARAKGGTGQTGGLTQPRKDEPKKLPPRPGGPEPKPGHPPQTRQATAR 120
Chimpanzee      GGHSYGGGATNANARAKGGTGQKGGLTQPKKDEPKKLPPRPGGPEPKPGHPPQTRQTTAR 120
Orangutan       GGHSYGGGAANANARAKGGTGQTGGLTQPKKDEPKKLSPRPGGPEPKPGHPPQTRQATAR 120
Macaque         GGHSYGGGA--ANARAKGGTGHTGGLTQPKKDEPKKLPPRPGGPEPKPGHPPQTRQATAR 118
Dog             GGHSYGGGA--ANARAKGGTGQTGGLTQPKKDEPKKLPPRPGSPEPKPGHPTQTRQAAPR 118
Mouse           GGHIYGVGA--TNARAKGSSGQ----TQAKKDEPRKMPPRSGGPETKPGPSSQTRQAAAR 114
Zebra fish      ------------------------------------------------------------
Xenopus         RTGILGHGVGLQKGRSKVPLVQSRIFLSKNEDIKKQAASRAN-PHVKTGN-AENRQSGEK 115
                *********:**:*:*:**:**********:**********.:****************
```

Figure 3.8 Amino acid sequence alignments. (A) The amino acid sequence alignment proximal to the non-synonymous mutation identified in exon 1 of an OA patient (Variant 3). The wildtype amino acid residue is highlighted in blue, and shows that in mammals the glycine residue is highly conserved. This mutation was predicted to be benign.

3.4.3 Amplicon 3

Amplicon 3, produced an average read length of 441 bp, including a 24 bp overlap with amplicon 2. Successful sequencing data was generated for a total of 442 OA cases and 332 controls (1 548 chromosomes). No novel mutations were discovered.

3.4.4 Amplicon 4

Amplicon 4, which covered the 5′ end of the second exon of *GDF5*, produced an average read length of 546 bp. Successful sequencing data was generated for 474 OA cases and 389 controls (1 726 chromosomes). From the sequencing of these individuals two novel mutations were discovered (Figures 3.9 and 3.10, Panels A), both of which are synonymous C to T transitions with the first found in a male control individual and the second in a female control individual. Both mutations were verified by resequencing each using new primers.

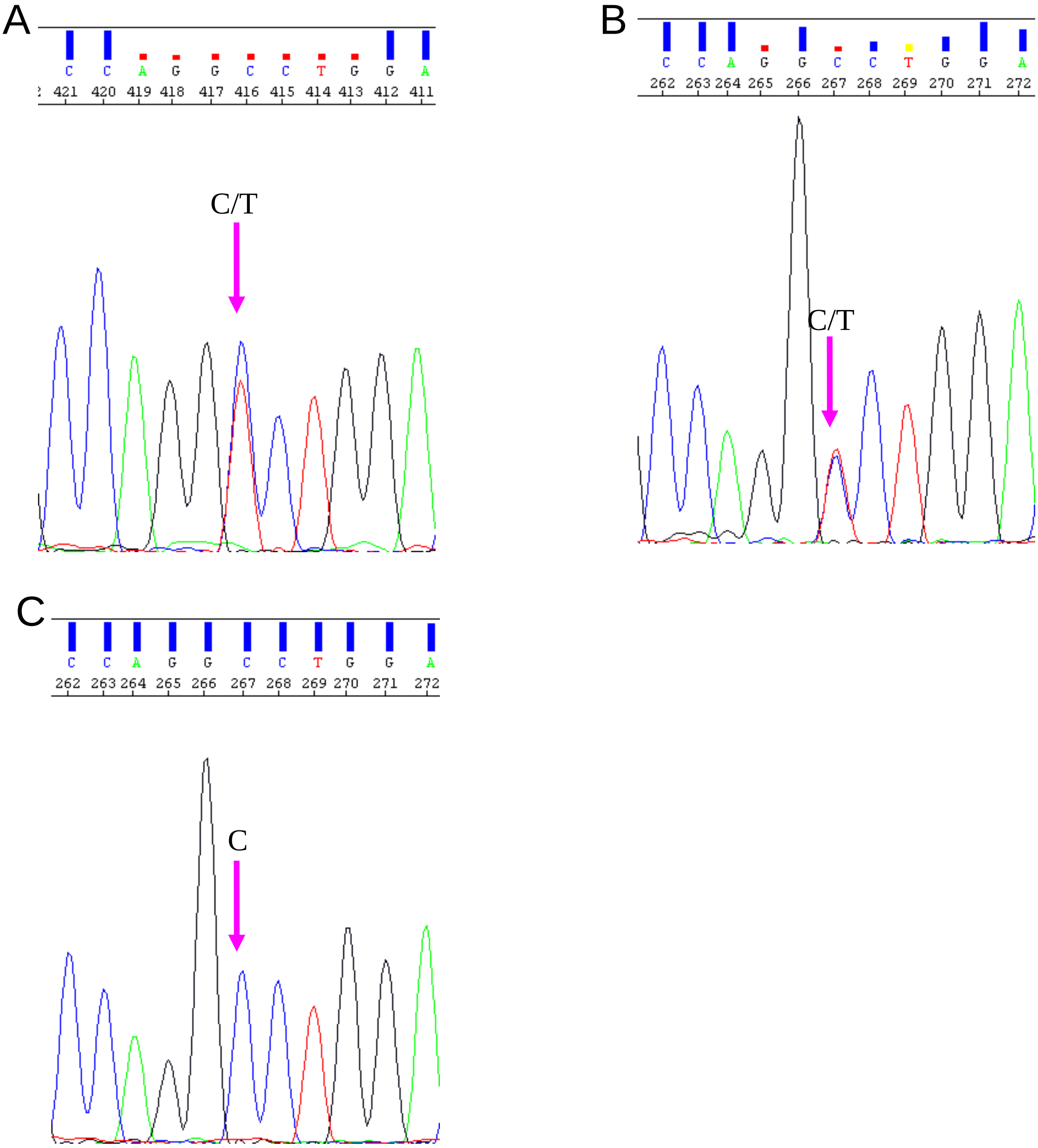


Figure 3.9 The exon 2 synonymous mutation identified in a male control individual (variant 4). (A) Electropherogram showing two alleles, C and T, following sequencing of amplicon 4 in the male control. Sequence is shown along the top along with a bar height corresponding to the confidence in which the Sequencing Analysis program calls each base. Blue corresponds to the greatest confidence, yellow is low confidence and red is zero confidence. (B) Electropherogram showing the same double peak when the individual was resequenced using a new primer pair. (C) Electropherogram showing only the single wildtype allele, C, present in an unrelated individual when sequenced using the same conditions as in B.

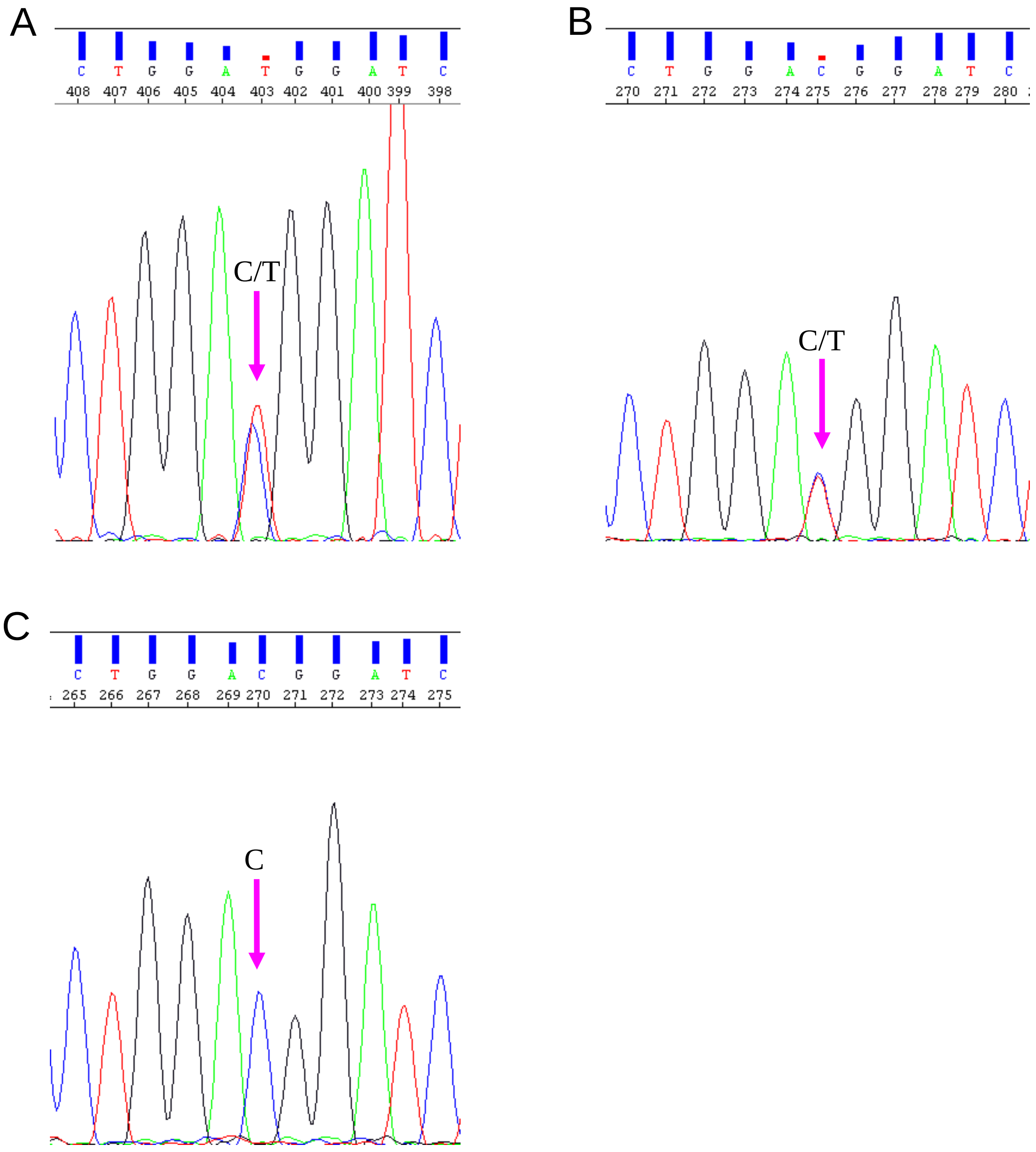


Figure 3.10 The exon 2 synonymous mutation identified in a female control individual (variant 5). (A) Electropherogram showing two alleles, C and T, following sequencing of amplicon 4 in the female control. Sequence is shown along the top along with a bar height corresponding to the confidence in which the Sequencing Analysis program calls each base. Blue corresponds to the greatest confidence, yellow is low confidence and red is zero confidence. (B) Electropherogram showing the same double peak when the individual was resequenced using a new primer pair. (C) Electropherogram showing only the single wildtype allele, C, present in an independent control individual when sequenced using the same conditions as in B.

3.4.5 Amplicon 5

Amplicon 5 covered the intermediate section of exon 2 producing an average sequencing read length of 454 bp which overlapped with amplicons 4 and 6. Successful sequencing data was generated for 495 OA cases and 436 controls (1 862 chromosomes). From these individuals, one mutation was detected in a control individual (Figure 3.11, Panel A). This mutation is predicted to result in a non-synonymous amino acid substitution and was validated by resequencing it using new primers (Figure 3.11, Panel B). The C to G transversion results in a non-polar, neutral, highly conserved (Figure 3.12) threonine residue being substituted by a non-polar but positively charged arginine residue at position 469. This amino acid substitution is predicted to be “possibly damaging” by PolyPhen. When GDF5 was modelled using the program Coot, Thr469 was shown to be interacting with a serine residue at position 115. It was also apparent that the threonine rotamer is most likely to be the standard rotamer for this side chain as seen in 49% of all threonine conformations (131). After *in silico* mutating of Thr469 to an arginine, which has a much longer side chain, it was apparent that the serine interaction is lost and the integration of this amino acid side chain becomes much more complex, with the suggested rotamer only observed in 9% of all arginine rotamers, and with another 32 possible rotamers suggested all with less frequency. The mutant Arg469 side chain becomes too close to Leu441 and Arg501 to allow for a wildtype protein folding due to the large collection of positive charge (Figure 3.13). In the GDF5 dimer the threonine 469s are only 10 Å apart, this is too small a distance to accommodate the arginine side chain coupled with its positive charge along with that of arginine 501.

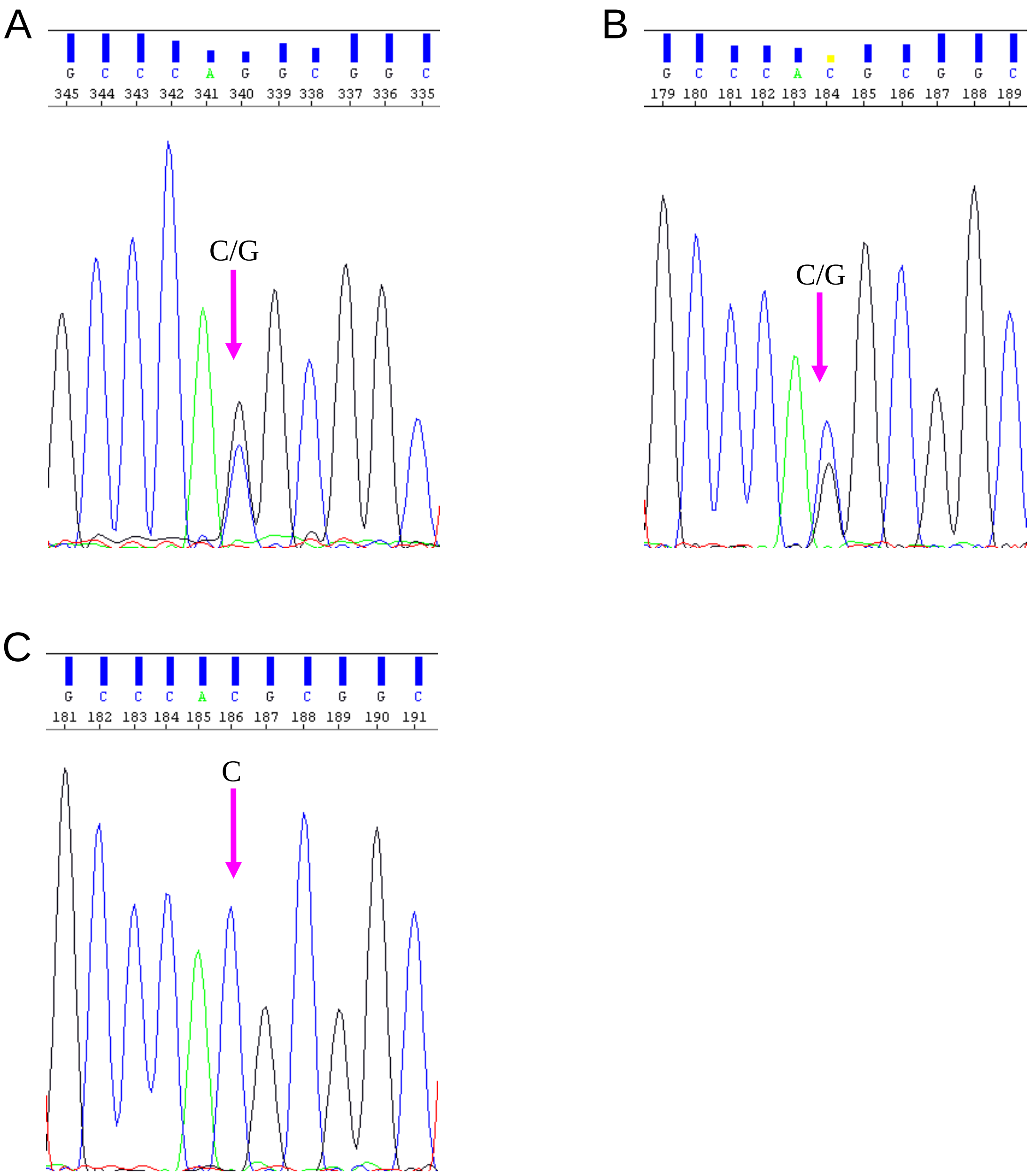


Figure 3.11 The exon 2 non-synonymous mutation identified in a female control individual (variant 6). (A) Electropherogram showing two alleles, C and G, following sequencing of amplicon 5 in the female control. Sequence is shown along the top along with a bar height corresponding to the confidence in which the Sequencing Analysis program calls each base. Blue corresponds to the greatest confidence, yellow is low confidence and red is zero confidence. (B) Electropherogram showing the same double peak when the individual was resequenced using a new primer pair. (C) Electropherogram showing only the single wildtype allele, C, present in an independent control individual when sequenced using the same conditions as in B.

```
Human           IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 477
Gorilla         IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 477
Chimpanzee      IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 477
Orangutan       IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 477
Macaque         IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 475
Dog             IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 475
Mouse           IIAPLEYEAFHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPESTPPTCCVPTRLSPISIL 471
Zebra fish      IIAPLEYEAFHCDGVCDFPIRSHLEPTNHAIIQTLMNSMDPRSTPPTCCVPTRLSPISIL 233
Xenopus         IIAPLEYEAYHCEGLCEFPLRSHLEPTNHAVIQTLMNSMDPETTPPTCCVPTRLSPISIL 472
                *********:**:*:*:**:**********:**********.:*****************
```

Figure 3.12 Amino acid sequence alignments. The amino acid sequence alignment proximal to the non-synonymous mutation identified in exon 2 of a female control individual (variant 6). This threonine residue (highlighted in red) is highly conserved across many species, not just mammals. The mutation to an arginine residue is predicted to be possibly damaging.

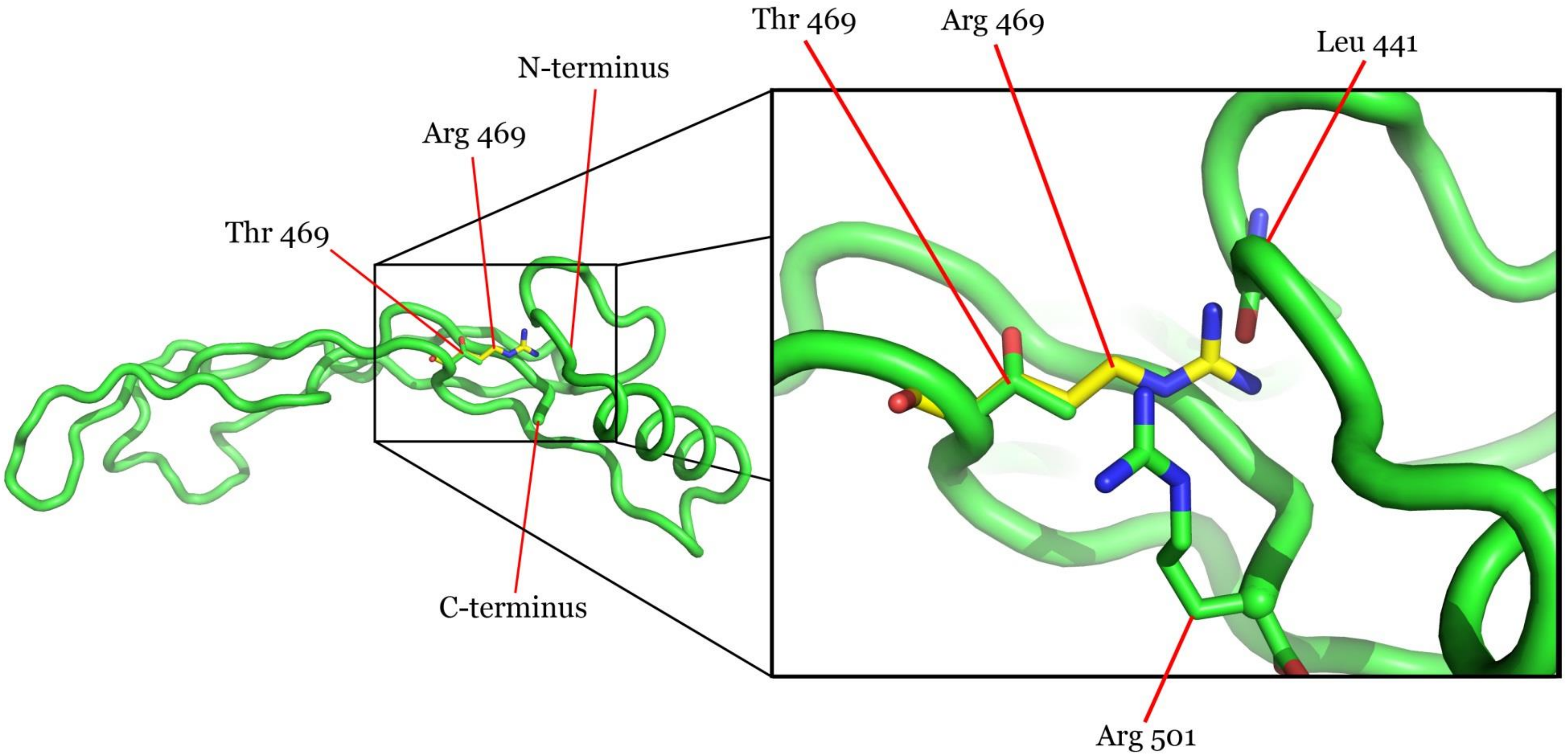


Figure 3.13 Protein model of a single GDF5 molecule. *Left*, shows the 3D structure of a single GDF5 molecule with both threonine 469 (shown in green) and arginine 469 (shown in yellow) as well as the C- and N-termini. *Right*, shows the area surrounding position 469 more clearly. Wildtype amino acids are shown with green carbons, whereas the variant Arg469's carbons are coloured yellow. Blue depicts nitrogen atoms and red depicts oxygen atoms. As shown the extended sidechain of the variant Arg469 brings it into close proximity with both Arg501 and Leu441, both of which carry a positive charge.

3.4.6 Amplicon 6

Amplicon 6 covered the 3’ end of exon 2 producing an average read length of 528 bp with some overlap with amplicon 5. Successful sequencing data was generated for 489 OA cases and 410 controls (1 798 chromosomes). No novel mutations were discovered.

3.4.7 Novel variants

Overall, six novel variants were identified, all extremely rare (Figure 3.14, Figure 3.15 and Table 3.3). Each variant was detected in a single individual, two in cases and four in controls, with each individual being heterozygote for the variant allele that they carry. Variant 1, discovered in a female control had a MAF of 0.0006 (MAF = number of variant alleles dived by the total number of chromosomes, ie. 1/1 752 for variant 1). Variant 2, discovered in a female case with knee OA and variant 3, discovered in a male case with knee OA both had MAFs of 0.0005. Variants 4, discovered in a male control, and 5, discovered in a female control, both had MAFs of 0.0006. Variant 6, discovered in a female control had a MAF of 0.0005.

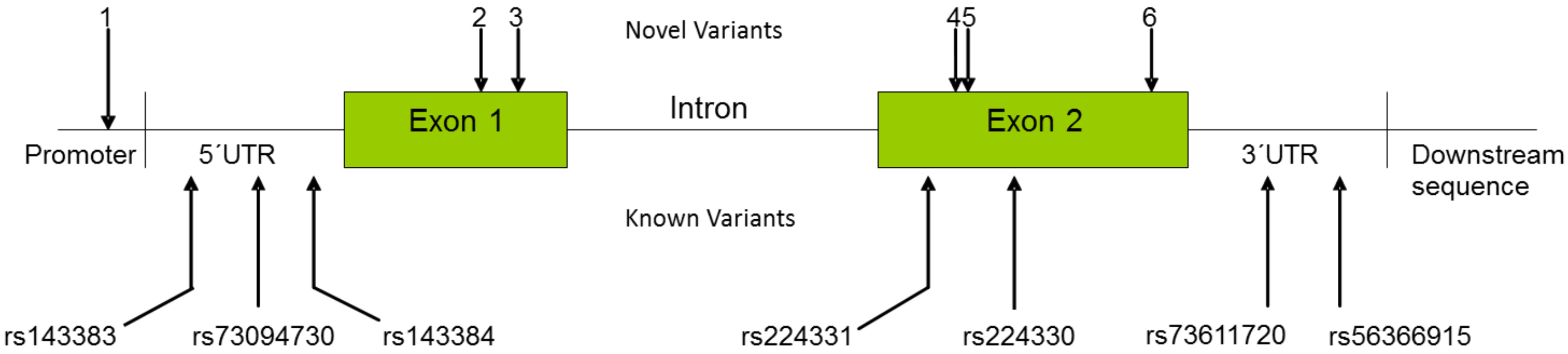


Figure 3.14 Locations of novel and known variants within *GDF5.* Schematic diagram depicting the locations of the six novel variants (top) found by deep sequencing and the seven previously known variants (bottom) within *GDF5*.

Table 3.3 Characteristics of each of the six novel mutations discovered within *GDF5* and its proximal promoter. Numbers correspond with mutation location in relation to Figure 3.14

Variant	Individual	Wild-type allele	Mutant allele	Location	Predicted effect
1	Control	C	A	Promoter	Alteration of a possible *trans*-factor binding site
2	Case	G	A	Exon 1	Synonymous, Gly67Gly
3	Case	G	A	Exon 1	Non-synonymous, Gly81Arg
4	Control	C	T	Exon 2	Synonymous, Gly285Gly
5	Control	C	T	Exon 2	Synonymous, Asp287Asp
6	Control	C	G	Exon 2	Non-synonymous, Thr469Arg

```
ATTCCGTTTCCAATTCCTGAGTTCAGGTTTGTAAAAGATTTTTCTGAGCACCTGCAGGCC
TGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAAGTATTTTC
ACTGGAAAGGATTCAAAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGC
CATTMTTCCTTCTTGGAAAAATCCCTCAGCCTTATACAAGCCTCCTTCAAGCCCTCAGTC
AGTTGTGCAGGAGAAAGGGGGCGGTYGGCTTTCTCCTTTCAAGAACGAGTTATTTTCAGC
TGCTGACTGGAGACGGTGCACGTCTGGATACGAGAGCATTTCCACTATGGGACTGGATAC
AAACACACACCCGGCAGACTTCAAGAGTCTCAGACTGAGGAGAAARCCTTTCCTTCTGCT
GCTACTGCTGCTGCCGCTGCTTTTGAAAGTCCACTCCTTTCATGGTTTTTCCTGCCAAAC
CAGAGGCACCTTYGCTGCTGCCGCTGTTCTCTTTGGTGTCATTCAGCGGCTGGCCAGAGG
ATGAGACTCCCCAAACTCCTCACTTTCTTGCTTTGGTACCTGGCTTGGCTGGACCTGGAA
TTCATCTGCACTGTGTTGGGTGCCCCTGACTTGGGCCAGAGACCCCAGGGGACCAGGCCA
GGATTGGCCAAAGCAGAGGCCAAGGAGAGGCCCCCCCTGGCCCGGAACGTCTTCAGGCCA Bd-C
GGGGGTCACAGCTATGGTGGRGGGGCCACCAATGCCAATGCCAGGGCAAAGGGAGGCACC
RGGCAGACAGGAGGCCTGACACAGCCCAAGAAGGATGAACCCAAAAAGCTGCCCCCCAGA
CCGGGCGGCCCTGAACCCAAGCCAGGACACCCTCCCCAAACAAGGCAGGCTACAGCCCGG
ACTGTGACCCCAAAAGGACAGCTTCCCGGAGGCAAGGCACCCCCAAAAGCAGGATCTGTC
CCCAGCTCCTTCCTGCTGAAGAAGGCCAGGGAGCCCGGGCCCCCACGAGAGCCCAAGGAG
CCGTTTCGCCCACCCCCCATCACACCCCACGAGTACYTGCTCTCGCTGTACAGGACGCTG Bd-C
TCCGATGCTGACAGAAAGGGAGGCAACAGCAGCGTGAAGTTGGAGGCTGGCCTGGCCAAC
ACCATCACCAGCTTTATTGACAAAGGGCAAGGTGAGGGGGCGGGGTGGCAGGGGCACGGC
TCAGAGGGAGGGGCATCTGCATGAATGGAGGGGCTTTCAAAGCCCTGGCACTGCCCTGGT
GGGAGACACTGTGTGCATCTGGCCCGGGGTGGGTGTCTGGGGACACTGACACATATCTCA

CAGAATGGGGCAGAGGTGAAAGAAAGCTCTCTGGACGGGGAGAGAGCTGGGGCCTTCCCC
CGCAGCCTCGAAGTGACTGGCTCCCTTGGTGAGGTTGCAGGGAATGACTTCTGGGTGTTC
TCTCTAGATGACCGAGGTCCCGTGGTCAGGAAGCAGAGGTACGTGTTTGACATTAGTGCC
CTGGAGAAGGATGGGCTGCTGGGGGCCGAGCTGCGGATCTTGCGGAAGAAGCCCTCGGAC
ACGGCCAAGCCAGCGGCCCCCGGAGGCGGGCGGGCTGCCCAGCTGAAGCTGTCCAGCTGC
CCCAGCGGCCGGCAGCCGGCCTCCTTGCTGGATGTGCGCTCCGTGCCAGGYCTGGAYGGA
TCTGGCTGGGAGGTGTTCGACATCTGGAAGCTCTTCRGAAACTTTAAGAACTCGGCCCAG CGT
CTGTGCCTGGAGCTGGAGGCCTGGGAACGGGGCAGGGCCGTGGACCTCCGTGGCCTGGGC
TTCGACCGCGCCGCCCGGCAGGTCCACGAGAAGGCCCTGTTCCTGGTGTTTGGCCGCACC
AAGAAACGGGACCTGTTCTTTAATGAGATTAAGGCCCGCTCTGGCCAGGACGATAAGACC
GTGTATGAGTACCTGTTCAGCCAGCGGCGAAAACGGCGGGCCCCACTGGCCACTCGCCAG
GGCAAGCGACCCAGCAAGAACCTTAAGGCTCGCTYCAGTCGGAAGGCACTGCATGTCAAC CGT
TTCAAGGACATGGGCTGGGACGACTGGATCATCGCACCCCTTGAGTACGAGGCTTTCCAC
TGCGAGGGGCTGTGCGAGTTCCCATTGCGCTCCCACCRGGAGCCCACGAATCATGCAGTC Bd-A2
ATCCAGACCCTGATGAACTCCATGGACCCCGAGTCCACACCACCCACCTGCTGTGTGCCC
ASGCGGCTGAGTCCCATCAGCATCCTCTTCATTGACTCTGCCAACAACGTGGTGTAMAAG Bd-C
CAGTATGAGGACATGGTCGTGGAGTCGTGTGGCTGCAGGTAGCAGCACTGGCCCTCTGTC
TTCCTGGGTGGCACATCCCAAGAGCCCCTTCCTGCACTCCTGGAATCACAGAGGGGTCAG
GAAGCTGTGGCAGGAGCATCTACACAGCTTGGGTGAAAGGGGATTCCAATAAGCTTGCTC
GCTCTCTGAGTGTGACTTGGGCTAAAGGCCCCCTTTTATCCACAAGTTCCCCTGGCTGAG
GATTGCTGCCCGTCTGCTGATGTGACCAGTGGCAGGCACAGGTCCAGGGAGACAGACTCT
GAATGGGACTGAGTCCCAGGAAACAGTGCTTTCCGATGAGACTCAGCCCACCATTTCTCC
TCACCTGGGCCTTCTCKGCCTCTGGACTCTCCTAAGCACCTCTCAGGAGAGCCACAGGTG
CCACTGCCTCCTCAAATCACATTTGTGCCTGGTGACTTCCTGTCCCRGGGACAGTTGAGA
AGCTGACTGGGCAAGAGTGGGAGAGAAGAGGAGAGGGCTTGGATAGAGTTGAGGAGTGTG
AGGCTGTTAGACTGTTAGATTTAAATGTATATTGATGAGATAAAAAGCAAAACTGTGCCT
AAAACTGTGGCAAATTTCTTGTTTACTCCAGGGGACCCAGTGACCCCTTGAAGAATTACT
GACCCAGGGCAGGGAGGTGGGTTGCTTCATGTATAAACTCAGTTTTCAGAGCCCTATCCC
```

R= A/G Y= C/T M= A/C K= G/T S= G/C W= A/T

Figure 3.15 Variants within *GDF5*. GDF5 sequence with variants recorded (novel variants in blue and common, known variants in green). Variants in red are disease causing; Brachytactyly type-C (Bd-C), chondrodysplasia Grebe Type (CGT) and Brachytactyly type A2 (Bd-A2) were not found. ATG translation start site and TAG stop codon are underlined. The yellow bases mark the location of cleavage site.

3.4.8 Known variants

GDF5 is 4.9 kb in length from its transcription start site to termination site and contains one intron between two exons coding for a 510 amino acid protein. In addition to the six novel variants that I have discovered here, I also identified seven common variants that were previously known to exist in the gene: rs143383, rs73094730, rs143384, rs224331, rs224330, rs73611720 and rs56366915 (Figures 3.14 and 3.15). The resequencing data gave rise to over 15 000 genotypes in the OA cases and controls. I therefore tested each SNP for association to OA by case-control analysis, using a χ^2 test. I examined both genotpic and allelic association. None of the known variants identified demonstrated association to OA ($P < 0.05$), with association analysis performed unstratified and stratified by sex and by joint (Tables 3.4-3.10). This is not particularly surprising since it is known that large sample sizes are required to generate robust association to common SNPs, such as rs143383 (55–57), and by comparison our sample size is small and underpowered.

There are several other *GDF5* SNPs listed in the public SNP database dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) but these either lack frequency data or have so far been shown to be polymorphic only in non-European samples. These are rs34534075, rs61754581, rs28936397, rs76603468, rs61754580, rs28936683, rs17853055, rs5841206, rs6120942 and rs79051206. I did not detect these variants despite the coverage in my sequence analysis of the relevant parts of *GDF5* that are purported to harbour these SNPs and I can conclude therefore that they are not polymorphic in Europeans, at least not in the large sample that I have studied.

Table 3.4 Genotype and allele frequencies of rs143383, along with χ^2 *p values.*

			Genotype				Allele	
					P value			*P* value
Group		(CC)	(CT)	(TT)		(C)	(T)	
All cases	Count	64	199	195	0.53	327	589	0.42
(n =458)	Frequency (%)	14.0	43.4	42.6		35.7	64.3	
All controls	Count	59	196	163		314	522	
(n = 418)	Frequency (%)	15.3	45.3	39.4		38.0	62.0	
Female cases	Count	47	148	144	0.675	242	436	0.465
(n=339)	Frequency (%)	13.9	43.6	42.5		35.7	64.3	
Female controls	Count	25	81	67		131	215	
(n =173)	Frequency (%)	14.5	46.8	38.7		37.9	62.1	
Male cases	Count	17	51	51	0.752	85	153	0.668
(n =119)	Frequency (%)	14.2	42.9	42.9		35.7	64.3	
Male controls	Count	34	115	96		183	307	
(n =245)	Frequency (%)	13.9	46.9	39.2		37.3	62.7	
All knees	Count	22	90	78		134	246	
(n = 190)	Frequency (%)	11.6	47.4	41.0	0.678	35.3	64.7	0.442
All hips	Count	43	116	123		202	362	
(n = 282)	Frequency (%)	15.3	41.1	43.6	0.32	35.8	64.2	0.507

Table 3.5 Genotype and allele frequencies of SNP 224, along with χ^2 *p values.*

			Genotype				Allele	
					P value			*P* value
Group		(AA)	(GA)	(GG)		(A)	(G)	
All cases	Count	0	20	486	0.43	20	992	0.43
(n = 506)	Frequency (%)	0	4.0	96.0		2.0	98.0	
All controls	Count	0	21	396		21	813	
(n = 417)	Frequency (%)	0	5.0	95.0		3.0	97.0	
Female cases	Count	0	17	370	0.55	17	757	0.56
(n= 387)	Frequency (%)	0	4.0	96.0		2.0	98.0	
Female controls	Count	0	10	171		10	352	
(n = 181)	Frequency (%)	0	5.5	94.5		2.8	97.2	
Male cases	Count	0	3	116	0.33	3	235	0.33
(n = 119)	Frequency (%)	0	2.5	97.5		1.3	98.7	
Male controls	Count	0	11	225		11	461	
(n = 236)	Frequency (%)	0	4.7	95.3		2.3	97.7	
All knees	Count	0	11	209		11	429	
(n = 220)	Frequency (%)	0	5.0	95.0	1	2.5	97.5	1
All hips	Count	0	13	291		13	595	
(n = 304)	Frequency (%)	0	4.3	95.7	0.63	2.1	97.9	0.64

Table 3.6 Genotype and allele frequencies of rs143384, along with χ^2 *p values.*

			Genotype			Allele		
					P value			*P* value
Group		(CC)	(CT)	(TT)		(C)	(T)	
All cases	Count	78	224	189	0.6	380	602	0.53
(n = 491)	Frequency (%)	15.9	45.6	38.5		38.7	61.3	
All controls	Count	67	207	151		341	509	
(n = 425)	Frequency (%)	14.8	45.8	33.4		40.1	59.9	
Female cases	Count	62	174	150	0.54	298	474	0.473
(n= 386)	Frequency (%)	16.1	45.0	38.7		38.6	61.4	
Female controls	Count	28	87	60		143	207	
(n= 175)	Frequency (%)	16.0	49.7	34.3		40.9	59.1	
Male cases	Count	16	50	39	0.99	82	128	0.89
(n = 105)	Frequency (%)	15.2	47.6	37.1		39.0	61.0	
Male controls	Count	39	120	91		198	302	
(n = 250)	Frequency (%)	15.6	48.0	36.4		39.6	60.4	
All knees	Count	32	104	85		168	274	
(n = 221)	Frequency (%)	14.5	47.1	38.4	0.75	38.0	62.0	0.46
All hips	Count	47	130	111		224	352	
(n = 288)	Frequency (%)	16.3	45.1	38.6	0.63	38.9	61.1	0.64

Table 3.7 Genotype and allele frequencies of rs224331, along with χ^2 *p values.*

			Genotype			Allele		
					P value			*P* value
Group		(TT)	(TG)	(GG)		(T)	(G)	
All cases	Count	218	191	61	0.67	627	313	0.77
(n = 470)	Frequency (%)	46.4	40.6	13.0		66.7	33.3	
All controls	Count	171	169	47		511	263	
(n = 387)	Frequency (%)	44.2	43.7	12.1		66.0	34.0	
Female cases	Count	163	145	45	0.56	471	235	0.41
(n= 353)	Frequency (%)	40.2	41.1	12.7		66.7	33.3	
Female controls	Count	63	70	20		196	110	
(n = 153)	Frequency (%)	41.2	45.8	13.0		64.1	35.9	
Male cases	Count	55	46	16	0.79	156	78	0.86
(n = 117)	Frequency (%)	47.0	39.3	13.7		66.7	33.3	
Male controls	Count	108	99	27		315	153	
(n = 234)	Frequency (%)	46.2	42.3	11.5		67.3	32.7	
All knees	Count	91	77	21		259	119	
(n = 189)	Frequency (%)	48.2	40.7	11.1	0.67	68.5	31.5	0.4
All hips	Count	134	121	41		389	203	
(n = 296)	Frequency (%)	45.2	40.9	13.9	0.69	65.7	34.3	0.77

Table 3.8 Genotype and allele frequencies of rs224330, along with χ^2 *p values.*

			Genotype			Allele		
					P value			*P* value
Group		(AA)	(AG)	(GG)		(A)	(G)	
All cases	Count	2	39	433	0.63	43	905	0.77
(n = 474)	Frequency (%)	0.4	8.2	91.4		4.5	95.5	
All controls	Count	3	27	359		33	745	
(n = 389)	Frequency (%)	0.8	6.9	92.3		4.2	95.8	
Female cases	Count	1	30	325	0.24	32	680	0.85
(n= 356)	Frequency (%)	0.3	8.4	91.3		4.5	95.5	
Female controls	Count	2	9	143		13	295	
(n = 154)	Frequency (%)	1.3	5.8	92.9		4.2	95.8	
Male cases	Count	1	9	108	0.88	11	225	0.8
(n = 118)	Frequency (%)	0.8	7.6	91.6		4.7	95.3	
Male controls	Count	1	18	216		20	450	
(n = 235)	Frequency (%)	0.4	7.7	91.9		4.3	95.7	
All knees	Count	2	17	172		21	361	
(n = 191)	Frequency (%)	1.1	8.9	90.0	0.66	5.5	94.5	0.34
All hips	Count	0	24	274		24	572	
(n = 298)	Frequency (%)	0	8.1	91.9	0.27	4.0	96.0	0.84

Table 3.9 Genotype and allele frequencies of SNP 4715, along with χ^2 *p values.*

			Genotype			Allele		
					P value			*P* value
Group		(CC)	(CA)	(AA)		(C)	(A)	
All cases	Count	2	44	443	0.95	48	930	0.84
(n = 489)	Frequency (%)	0.4	9.0	90.6		4.9	95.1	
All controls	Count	2	38	370		42	778	
(n = 410)	Frequency (%)	0.5	9.3	90.2		5.1	94.9	
Female cases	Count	1	32	339	0.86	34	710	0.82
(n= 372)	Frequency (%)	0.3	8.6	91.1		4.4	95.6	
Female controls	Count	1	15	158		17	331	
(n = 174)	Frequency (%)	0.6	8.6	90.8		4.9	95.1	
Male cases	Count	1	12	104	0.87	14	220	0.71
(n = 117)	Frequency (%)	0.9	10.2	88.9		6.0	94.0	
Male controls	Count	1	23	212		25	447	
(n = 236)	Frequency (%)	0.5	9.7	89.8		5.3	94.7	
All knees	Count	2	21	201		25	423	
(n = 224)	Frequency (%)	0.9	9.4	89.7	0.99	5.6	94.4	0.73
All hips	Count	0	24	257		24	538	
(n = 281)	Frequency (%)	0	8.5	91.5	0.88	4.3	95.7	0.47

Table 3.10 Genotype and allele frequencies of rs56366915, along with χ^2 *p values.*

Group		Genotype				Allele		
					P value			*P* value
Group		(CC)	(CT)	(TT)		(C)	(T)	
All cases	Count	1	42	442	0.5	44	926	0.56
(n = 485)	Frequency (%)	0.2	8.7	91.1		4.5	95.5	
All controls	Count	3	36	370		42	776	
(n = 409)	Frequency (%)	0.7	8.8	90.5		5.1	94.9	
Female cases	Count	0	30	340	0.12	30	710	0.3
(n= 370)	Frequency (%)	0	8.1	91.9		4.1	95.9	
Female controls	Count	2	15	157		19	329	
(n = 174)	Frequency (%)	1.1	8.7	90.2		5.6	94.4	
Male cases	Count	1	12	102	0.79	14	216	0.51
(n = 115)	Frequency (%)	0.9	10.4	88.7		6.1	93.9	
Male controls	Count	1	21	213		23	447	
(n = 235)	Frequency (%)	0.4	8.9	90.7		4.9	95.1	
All knees	Count	1	20	200		22	420	
(n = 221)	Frequency (%)	0.5	9.0	90.5	0.91	5.0	95.0	0.9
All hips	Count	0	23	257		23	537	
(n = 280)	Frequency (%)	0	8.2	91.8	0.69	4.1	95.9	0.38

3.4.9 Absence of any rare variants

Overall, my sequencing of *GDF5* did not detect any rare variants, that is variants with a MAF in the region of 0.001–0.025 (i.e., 0.1–2.5%). Instead, variants were either extremely rare (MAFs ≤ 0.0006, i.e., ≤0.06%, *n* = 6) or common (MAFs > 0.025, i.e., 2.5%, *n* = 7), as represented in Figure 3.16.

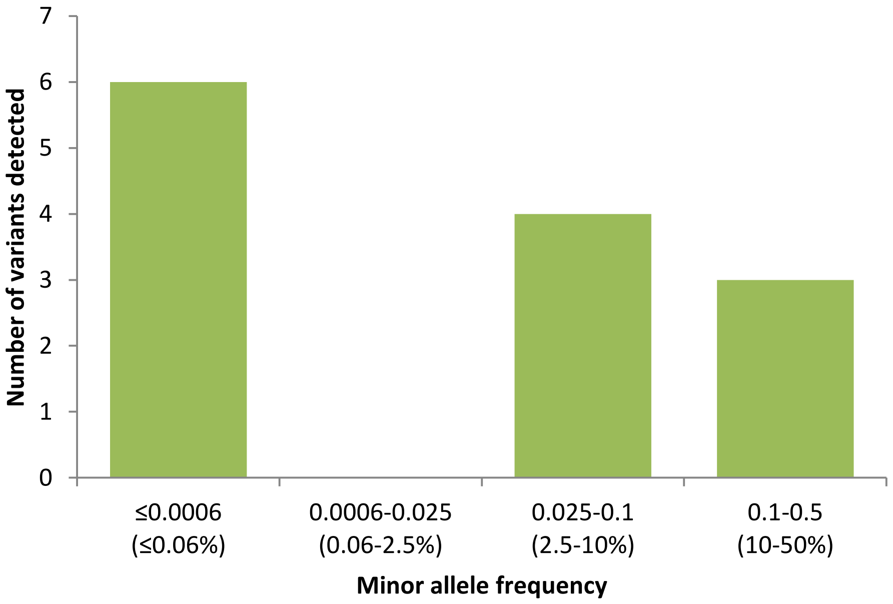


Figure 3.16 The number of variants within *GDF5* displayed according to their MAFs. Graph showing that variants within *GDF5* are either very rare (MAFs ≤0.0006) or common (MAFs ≥0.025).

3.5 Discussion

Current literature is increasingly suggesting that loci harbouring alleles with major disease susceptibility risk to a certain trait are also likely to contain other alleles with varying influence upon that same trait (132). *GDF5* contains a common functional 5′ UTR SNP, rs143383, whereby the under expressing T-allele has robustly been associated to OA in both European and Asian populations (52, 53, 55–57). Furthermore, the allelic imbalance observed at rs143383 has been shown to be modulated by genotype at a neighbouring SNP, rs143384 (54). Additionally, Egli *et al*. also demonstrated that another common SNP in the 3′ UTR, rs56366915, is able to influence the allelic output of *GDF5* but completely independent of genotype at rs143383 (54). *GDF5* therefore exhibits its potential to be influenced by genetic polymorphism within it, by both an interacting polymorphism and independent polymorphism. These three mentioned polymorphisms are all common, with high MAFs. In this study I set out to assess the genetic architecture of *GDF5* to evaluate whether this gene harbours any other variants which may also be functional and that could impact upon OA aetiology. At present this is the first example of a deep-sequencing analysis of an OA susceptibility locus.

I set out to sequence *GDF5* in both cases and controls from the UK population with the hypothesis that by focusing on cases and controls rather than random individuals from the general population I would increase the chances of discovering penetrant risk or protective mutations. I demonstrated the accuracy of my sequencing by detecting all of the known variants from dbSNP that reside within the sections of the gene that I covered. By virtue of detecting six extremely rare novel variants present in only single individuals from over 1 800 chromosomes, I also validated the sensitivity of my study.

Of these six extremely rare variants, three are potentially functional; variant 1, which is located in the proximal promoter of *GDF5* and is within predicted transcription factor binding sites, variant 3, which is located in exon 1 and is predicted to code for the substitution of a highly conserved glycine residue and variant 6, which is located in exon 2 and is predicted to code for the substitution of a highly conserved threonine residue. However, as these variants are extremely rare, occurring each in single individuals only, they can't be having any effect at a population level on OA susceptibility. Should these variants be functional then it is possible that they have an effect at the individual or family level. It is known that penetrant deleterious mutations of *GDF5* can cause Mendelian phenotypes. For example, brachydactyly types A2 and C, characterized by malformation of the phalanges can be caused by mutation within *GDF5*, although these mutations typically involve or are proximal to highly conserved cysteine residues which are involved in both monomer formation and dimerization of GDF5 (133). The ethics used to collect the OA cases and controls used in this thesis does not allow us to retrospectively contact the individuals. It is therefore not possible for us to assess whether the two individuals harbouring the amino-acid substitution mutations have any overt skeletal abnormalities that could be the result of the mutations. Variant 3, Gly81Arg, occurs within the immature form of GDF5 and this makes it difficult to assess the effect this mutation would have on GDF5 through modelling and functional protein studies as the crystal structure of the immature form of GDF5 has not been determined, although the PolyPhen database predicts this amino acid substitution to be benign. This variant does not reside anywhere near to any previously known disease causing variants within the gene. Polyphen predicts that variant 6, Thr469Arg, could be potentially damaging. As this mutation resides within the mature, cleaved form of GDF5, more scope is offered up for potential investigation. The crystal structure is known for this protein form making *in silico* modelling possible. It was found that the introduction of an arginine side chain to this area of GDF5 would probably impact on the folding of the whole protein due to the size and charge of the arginine side chain. Also assays assessing GDF5 activity have been performed in chicken micromass cultures (100) providing a potential further avenue for interrogation of Thr469Arg and its effect upon protein activity.

The absence of rare alleles could be due to the biological importance of *GDF5*, whereby new mutation within the gene cannot be tolerated unless it is benign. One other plausible explanation for this situation is that the study population has undergone some form of evolutionary or population bottleneck. A population bottleneck is a sudden reduction in population size, and with it reduced heterozygosity within that population (134, 135). During a population bottleneck, MAFs and genetic diversity are severely affected by random genetic drift (136). Rare alleles can be wiped out altogether and common alleles can be reduced to very rare alleles, all depending upon which individuals survive the bottleneck. It is possible that the current genetic architecture of *GDF5* is reflective of a bottleneck imparted upon the population, whereby the seven common polymorphisms that were confirmed in my study made it through the bottleneck with all other potential *GDF5* alleles lost. The six very rare novel variants that I discovered may be the result of mutation that has occurred after the bottleneck, as the population is expanding once again and greater heterogeneity is developing. It is believed that, in the last 10 000 years, the global population has exploded from a few million to the estimated 7 billion that it is today. Such an explosion allows for rare variants within sub-populations due to mutation, which can exact upon complex disease (137). The bottleneck hypothesis may be supported by recent findings that both rs143383 and rs143384 have undergone positive selection within the East Asian population (138). In this study it was shown that the T-allele of rs143383 – which is associated with OA – along with the T-allele of rs143384 – which has been shown to modulate the effect of rs143383 of GDF5 expression (54) – are both associated with height and that there has been a positive selection on these alleles. The ages of these alleles are around 12 000 years old, a time which saw an increase in early agriculture, a decline in the average body mass of humans, as well as the spread of infectious disease. By growing and rearing one's own food rather than hunting and foraging, the necessity for a large and high energy dependant body mass becomes reduced (138). Lower expression of *GDF5* confers with shorter limb growth (139), aiding the reduction in body mass.

It is likely that the onset of a typically high-age related disease like OA would rarely come into play as a negative selection pressure during this time period.

In summary, the deep sequencing of the transcript and promoter sequence of *GDF5* has revealed that this gene harbours both extremely rare variants and also common polymorphism but there is an absence of mutation of intermediate frequency. This study was further expanded by the collaboration of another research group from Santiago de Compostela, Spain who sequenced the open reading frame of *GDF5* in a cohort of Spanish and Greek individuals, and thus allowing the study to incorporate both Northern and Southern Europeans (140). The research group in Spain only detected common variants, finding no novel variants, either unique to that cohort or the same as the extremely rare variants found in the UK population.

Of the six extremely rare variants found, the promoter variant (variant 1) and the non-synonymous variant (variant 6) appear to have the greatest potential to impart a functional effect upon *GDF5* expression or GDF5 activity, respectively. However, none of those six variants can be having any effect on OA susceptibility at a population level. Due to the size of and coverage of this study it is unlikely that there is any other mutation within *GDF5* within the population tested.

Chapter 4: A rare variant in the promoter of the OA-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate gene expression

4.1 Introduction

As previously mentioned, rs143383 is a common polymorphism in Asians and Europeans, with a minor allele frequency (MAF) >20%. During the recent frenetic activity into the genetic dissection of polygenic traits it has become apparent that common alleles individually contribute only modestly to trait variance (132) and that rare variants of greater individual impact could account for some of the missing heritability that is a current conundrum in the field (99). As such I was keen to assess whether *GDF5* harbored rare variants that could influence OA susceptibility by either acting as independent risk factors or by potentially modulating the influence of rs143383, akin to rs143384. I therefore performed a deep sequence analysis of *GDF5*, on >1900 European individuals, and focused the sequencing on the proximal regulatory elements of the gene (promoter, UTRs, intron/exon boundaries) and on the protein coding sequence. This revealed an absence of rare *GDF5* variants, with the gene harboring either known common variants, with MAFs ≥2.5%, or very rare novel variants, with MAFs ≤0.006% as described in *Chapter 3* (140). In total, the study uncovered six novel mutations; five within the protein coding sequence of *GDF5* and one in the proximal promoter, 41 base pairs upstream of the transcription start site. I was particularly intrigued by the -41 bp variant, which is a C/A transversion that is located in a conserved sequence of *GDF5*; as mentioned in *Chapter 3*, a search of databases of transcription factor binding sites predicted that the novel A-allele is likely to impact on the binding of *trans*-acting regulators of gene transcription. If the -41 bp variant does highlight a genuine regulatory domain of *GDF5* then this could offer us with a DNA site that could be exploited to manipulate the expression of the gene and therefore potentially alleviate the detrimental effect that the T-allele of rs143383 has on *GDF5* expression and therefore on OA risk.

4.2 Aims of this study

The aim of this study is to assess whether the novel -41 bp variant which I discovered in the promoter of *GDF5* is functional. If it is, I plan to use EMSAs in order to identify possible *trans*-acting factors which bind to this locus.

4.3 Materials and methods

4.3.1 Molecular haplotyping and construction of GDF5-pGL3 basic luciferase reporter plasmids

We investigated the -41 bp variant in combination with rs143383 and rs143384. The physical order of the three polymorphisms is -41 bp-rs143383-rs143384 (Figure 4.1). To generate constructs for use in the *GDF5* promoter/5′ UTR luciferase reporter gene assays, a 403bp fragment of genomic DNA corresponding to the *GDF5* proximal promoter and part of the 5′UTR (-97 to +305 with respect to the transcriptional start site of *GDF5*) and encompassing the three polymorphisms was PCR amplified using gene specific primers (underlined) containing either a *Mlu*I (forward primer 5′-ACTCA**ACGCGT**GGATTCAAAACTAGGGG-3′; *Mlu*I site marked in bold) or a *Bgl*II (reverse primer 5′-GCACA**AGATCT**AGCCGCTGAATGACACCAAAG-3′; *Bgl*II site marked in bold) restriction site at the 5′ end of the primer. PCR was carried out as described in *Chapter 2.3.2* with thermocycling conditions of an initial denaturation at 94^{o}C for 14 minutes, followed by 37 cycles of 94^{o}C for 30 seconds, annealing at 70^{o}C for 30 seconds, and extension at 72^{o}C for 45 seconds, with a final extension at 72^{o}C for 5 minutes. After amplification, the PCR product was purified and eluted in 30 µl of water using the QIAquick

PCR purification kit following the manufacturer's instructions (Qiagen). After PCR purification, double digestion of the PCR products was achieved with 15 units of *Mlu*I and 15 units of *Bgl*II restriction enzymes (NEB), in the presence of 10X NEB Buffer 3 (50 mM Tris-HCl, 10 mM $MgCl_2$, 100 mM NaCl and 1 mM DTT, pH 7.9) incubated for 3 hours at 37°C, followed by heat inactivation at 80°C for 20 minutes. Following digestion the product was again purified using the QIAquick PCR purification kit (QIAGEN).

```
-165 TGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAAGTATTTTC
-105 ACTGGAAAGGATTCAAAACTAGGGGGAAAAAAAAACTGGAGCACACAGGCAGCATTACGC
-45  CATTMTTCCTTCTTGGAAAAATCCCTCAGCCTTATACAAGCCTCCTTCAAGCCCTCAGTC  -41 bp  C/A
  16 AGTTGTGCAGGAGAAAGGGGGCGGTYGGCTTTCTCCTTTCAAGAACGAGTTATTTTCAGC rs143383 C/T
  76 TGCTGACTGGAGACGGTGCACGTCTGGATACGAGAGCATTTCCACTATGGGACTGGATAC
 136 AAACACACACCCGGCAGACTTCAAGAGTCTCAGACTGAGGAGAAAGCCTTTCCTTCTGCT
 196 GCTACTGCTGCTGCCGCTGCTTTTGAAAGTCCACTCCTTTCATGGTTTTTCCTGCCAAAC
 256 CAGAGGCACCTTYGCTGCTGCCGCTGTTCTCTTTGGTGTCATTCAGCGGCTGGCCAGAGG rs143384 C/T
 316 ATGAGACTCCCCAAACTCCTCACTTTCTTGCTTTGGTACCTGGCTTGGCTGGACCTGGAA
 376 TTCATCTGCACTGTGTTGGGTGCCCCTGACTTGGGCCAGAGACCCCAGGGGACCAGGCCA
 436 GGATTGGCCAAAGCAGAGGCCAAGGAGAGGCCCCCCCTGGCCCGGAACGTCTTCAGGCCA
 496 GGGGGTCACAGCTATGGTGGGGGGGCCACCAATGCCAATGCCAGGGCAAAGGGAGGCACC
 556 GGGCAGACAGGAGGCCTGACACAGCCCAAGAAGGATGAACCCAAAAAGCTGCCCCCCAGA
 616 CCGGGCGGCCCTGAACCCAAGCCAGGACACCCTCCCCAAACAAGGCAGGCTACAGCCCGG
 676 ACTGTGACCCCAAAAGGACAGCTTCCCGGAGGCAAGGCACCCCCAAAAGCAGGATCTGTC
 736 CCCAGCTCCTTCCTGCTGAAGAAGGCCAGGGAGCCCGGGCCCCCACGAGAGCCCAAGGAG
 796 CCGTTTCGCCCACCCCCCATCACACCCCACGAGTACATGCTCTCGCTGTACAGGACGCTG
 856 TCCGATGCTGACAGAAAGGGAGGCAACAGCAGCGTGAAGTTGGAGGCTGGCCTGGCCAAC
 916 ACCATCACCAGCTTTATTGACAAAGGGCAAGGTGAGGGGGCGGGGTGGCAGGGGCACGGC
```

Figure 4.1 Exon 1 (highlighted) and flanking intronic sequence of *GDF5*. -41 bp, rs143383 and rs143384 are shown in blue. Numbering is relative to the transcription start site.

In preparation for cloning, 10 µg of pGL3-Basic luciferase reporter vector (Promega) was double digested with *Bgl*II and *Mlu*I restriction enzymes as described above. The digested vector was treated with 5 units of Antarctic Phosphatase (New England Biolabs) and then purified using the QIAquick gel extraction kit (Qiagen) following the manufacturer's instructions and eluted in water. The digested and purified PCR product was then cloned into the *Mlu*I/*Bgl*II sites of the purified pGL3-Basic vector using a 5:1 ratio of PCR fragment to plasmid and 400 units of T4 ligase (New England Biolabs), incubated at 16^{o}C overnight to create the *GDF5* promoter/5' UTR-Luciferase construct (Figure 4.2). Plasmid DNA was transformed into MACH 1 E. *coli* bacterial cells (Invitrogen) following manufacturers guidelines. Following incubation, 50-100 µl of transformed cells were spread out onto agar plates containing 100 µg/ml of ampicillin (Sigma Aldrich). These plates were then incubated upside down at 37^{o}C overnight to allow colonies to grow.

Colonies were picked from the bacterial culture plates and grown overnight at 37^{o}C at 200 rpm in 3.5 ml of LB broth (Sigma Aldrich) containing 100 µg/ml of ampicillin (Sigma Aldrich). Glycerol stocks were made by adding 200 µl of bacterial culture to 200 µl of glycerol (Sigma Aldrich), stocks were then stored at -80^{o}C. Each culture was mini-prepped by centrifuging 1.4 ml of bacterial culture in a 1.5 ml centrifuge tube (Griener Bio) at 8 200 rpm for 3 minutes to pellet the bacteria. Following the aspiration of the supernatant, the bacterial pellet was re-suspended in 100 µl of P1 + RNaseA solution (Qiagen). Following re-suspension, 200 µl of P2 solution (Qiagen) was added and the tubes were inverted to mix and left to incubate at room temperature for 2 minutes. After incubation, 150 µl of P3 solution (Qiagen) was added, mixed and then centrifuged for 5 minutes at 13 000 rpm. The supernatant was then removed and added to 1 ml of 100% ethanol. The ethanol/supernatant mix was then stored at -80^{o}C for at least 15 minutes. Following storage at -80^{o}C the tubes were centrifuged for 10 minutes at 13 000 rpm at room temperature to pellet the DNA. All of the supernatant was then removed and the DNA pellet was air-dried for 10-15 minutes until all ethanol had evaporated. Each pellet was then re-suspended in 30 µl of DEPC water (Invitrogen) and stored at -20°C. These individual plasmid clones were sequenced via Sanger capillary sequencing as described in *Chapter 3.3.2*, by Genome Enterprises, Norwich, UK using the primer *pGL3basic R* 5'-ACCAGGGCGTATCTCTTCAT-3'.

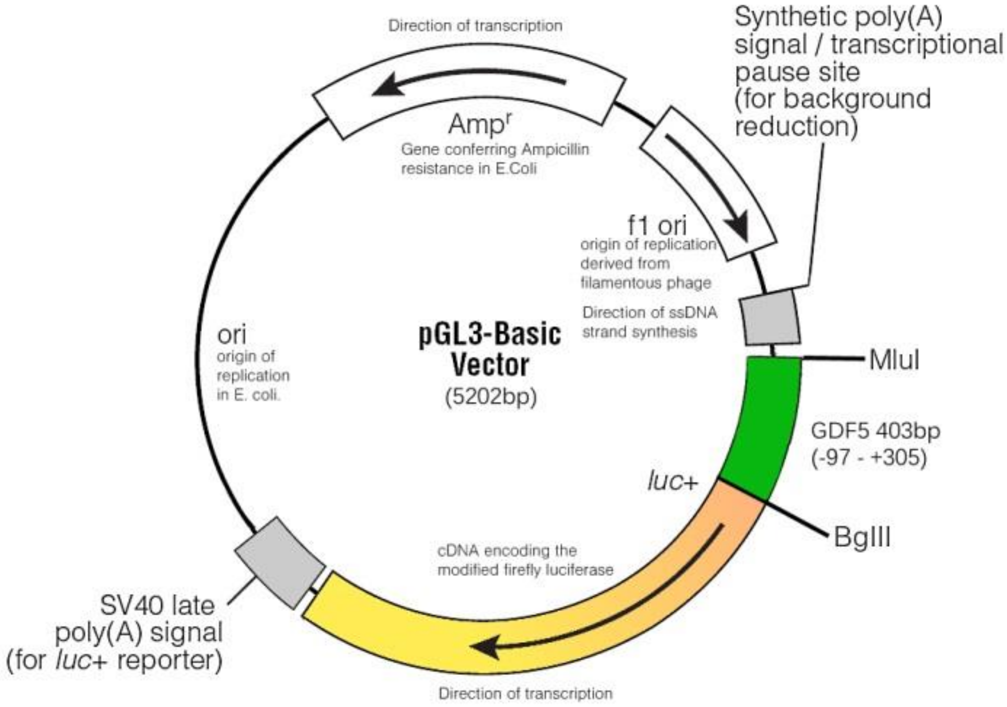

Direction of transcription
Synthetic poly(A) signal / transcriptional pause site (for background reduction)
Amp^r
Gene conferring Ampicillin resistance in E.Coli
f1 ori
origin of replication derived from filamentous phage
Direction of ssDNA strand synthesis
ori
origin of replication in E. coli.
pGL3-Basic Vector
(5202bp)
MluI
GDF5 403bp
(-97 - +305)
luc+
BglII
cDNA encoding the modified firefly luciferase
SV40 late poly(A) signal (for luc+ reporter)
Direction of transcription

Figure 4.2 Vector map of the pGL3-Basic Vector with the 403 bp *GDF5* insert, containing the proximal promoter and 305 bp of the 5' UTR. The insert was cloned into the vector's multiple cloning site using the enzymes *Mlu*I and *Bgl*II. Image adapted from *http://vesuvias.files.wordpress.com/2009/06/plasmidvector.png*

The starting input DNA for this cloning strategy came from the individual in whom the -41 bp variant had been originally detected (140), as described in *Chapter 3*. This individual was heterozygous for -41 bp and heterozygous for both rs143383 and rs143384. Following successful cloning and sequencing, it was discovered that this individual carried a -41 bp (A) - rs143383 (C) - rs143384 (C) haplotype (in future known as the A-C-C haplotype) and a -41 bp

(C) - rs143383 (T) - rs143384 (T) haplotype (in future known as the C-T-T haplotype). The non-naturally occurring A-T-T and C-C-C haplotypes were generated using the Agilent Quickchange II site directed mutagenesis kit (Agilent). To generate A-T-T I used the C-T-T clone and the forward primer 5′-CAGCATTACGCCATTATTCCTTCTTGGAAA and the reverse primer 5′-TTTCCAAGAAGGAATAATGGCGTAATGCTG-3′. To generate C-C-C I used the A-C-C clone and the forward primer 5′-CAGCATTACGCCATTCTTCCTTCTTGGAAA-3′ and the reverse primer 5′-TTTCCAAGAAGGAAGAATGGCGTAATGCTG-3′. The reaction mix contained 50 ng of plasmid, 125 ng of forward primer, 125 ng of reverse primer, 0.01 mM dNTP mix, 10X reaction buffer, 2.5 units of *Pfu* Ultra high fidelity DNA polymerase (Agilent) in a total volume of 50 µl. Thermocycling conditions were denaturing at 95^{o}C for 30 seconds followed by 18 cycles of denaturing at 95^{o}C for 30 seconds, annealing at 50^{o}C for 1 minute and extension at 68^{o}C for 7 minutes. Following thermocycling the reaction was placed on ice for 2 minutes to cool to below 37^{o}C.

Following mutagenesis, methylated and hemimethylated DNA was digested by the addition of 10 units of *Dpn*I (Agilent). Reactions were mixed and briefly centrifuged for 1 minute before incubation at 37^{o}C for 1 hour. The transformation of these mutated plasmids, along with the control plasmid, was carried out following the manufacturer's instructions (Agilent). Colonies were picked and cultured as described above, and miniprepped in the same manner along with glycerol stocks being made. Clones were again sent for sequencing by Genome Enterprises Ltd, Norwich, UK.

4.3.2 Maxiprep of the four haplotypes of the GDF5 promoter/5' UTR Luciferase reporter plasmids

In order to obtain large quantities of pure plasmid DNA, 10 µl of glycerol stock was thawed and initially cultured in 5 ml of LB containing 100 ng/µl of ampicillin at 37^{o}C and 200 rpm for 3 hours. This culture was then transferred to a conical flask containing 250 ml of LB containing 100 ng/µl of ampicillin, and incubated overnight at 37^{o}C and 200 rpm. Cultures were then centrifuged at 6000 *g* for 15 minutes at 4^{o}C in a 50 ml centrifuge tube (Nalgene). Bacterial pellets were then re-suspended in 10 ml of P1 + RNaseA solution (Qiagen). Following re-suspension, 10 ml of P2 solution (Qiagen) was added and the solution was vigorously mixed manually and incubated at room temperature for 5 minutes. After incubation, 10 ml of ice cold P3 solution (Qiagen) was added and again the solution was vigorously mixed manually before being incubated on ice for 20 minutes. Lysates were then centrifuged at 20 000 *g* for 30 minutes at 4^{o}C. The supernatant was then transferred to a fresh tube and centrifuged under the same conditions for a further 15 minutes. Following the equilibration of a QIAGEN-tip500 gravity column with 10 ml of Buffer QBT (Qiagen), the supernatant was applied to the column and allowed to run through by gravity into a waste receptacle. The column was then washed with two 30 ml volumes of Buffer QC, and then eluted with 15 ml of Buffer QF into a fresh centrifuge tube. DNA was precipitated by adding 10.5 ml of isopropanol which was mixed and centrifuged at 15 000 *g* for 30 minutes at 4^{o}C. The supernatant was aspirated and the pellet air dried. The DNA pellet was then re-suspended in 500 µl of de-ionized H_2O. Two and half times volume of 100% ethanol was then added to the re-suspended DNA, and after mixing was stored at -80^{o}C for at least 30 minutes. This precipitant was centrifuged at 13 000 rpm in a desktop centrifuge for 10 minutes. The supernatant was aspirated and pellets were air dried for 5-10 minutes before being re-suspended in 300 µl of H_2O and quantified using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies).

4.3.3 Cell Culture

Three human cell lines were cultured for transfection experiments; the SW1353 chondrosarcoma cell line, the MG63 osteosarcoma cell line and the CH8 articular chondrocyte cell line. The SW1353 cell line was cultured in a T75 or T162 cell culture flask (Corstar) in Dulbecco's Modified Eagle’s Medium: Nutrient Mixture F-12 (DMEM/F12) (GIBCO), containing 5% foetal bovine serum (FBS), 10 ml of 2X Penicillin:Streptomycin mix (Sigma Aldrich) and 2 mM of glutamine (Sigma Aldrich). The MG63 cell line was cultured in a T75 or T162 cell culture flask in Dulbecco's Modified Eagle's Medium (DMEM) (GIBCO), containing 5% foetal bovine serum (FBS), 1X Penicillin:Streptomycin mix (Sigma Aldrich) and 2 mM of glutamine (Sigma Aldrich). The CH8 cell line was cultured in a T75 flask coated with 0.4 mg/ml of rat tail Collagen Type I (BD Biosciences). Collagen coating was achieved by making 100 mg/ml of rat tail Collagen Type I in Dulbecco's Phosphate Buffered Solution (Lonza) and adding this to a T75 flask until the flask surface was completely covered. The coated flask was then incubated at 37°C for one hour, and subsequently washed twice with PBS. The CH8 cell line was cultured on this collagen surface in DMEM containing 10% FBS and 1X Penicillin:Streptomycin mix. Cells were allowed to grow until 70% confluencey before being passaged and reseeded.

4.3.4 Transfection of cell lines

Cells were seeded in 300 µl of media at a density of 17 500 cells per well in a 48 well cell culture plate (Costar). Cells were treated with trypsin and placed in fresh media and then pelleted by centrifugation at room temperature at 400 *g* for 5 minutes. They were then resuspended in media and counted using a Rosenthal haemocytometer. The seeded cells were then cultured for 24 hours with respect to MG63 cells and 48 hours with respect to SW1353 and CH8 cells. After this period of culture the cells in each well were transfected using a transfection mixture consisting of 500 ng of pGL3 plasmid DNA and 15 ng of pTK-RL Renilla plasmid in 28.35 µl of a 0.9 M NaCl solution. This mixture was then mixed and pelleted and then 1.65 µl of Exgen 500 Transfection reagent (Fermentas) was added before being mixed again for 10 seconds. This mixture was then incubated at room temperature for 10 minutes. During this incubation, cells were aspirated prior to the 30 µl of transfection mixture being added directly to the cells. Following this, 270 µl of appropriate fresh media was added per well and cultured for a further 24 hours prior to cell lysis.

4.3.5 Cell Lysis and Luciferase Activity Reading

Twenty-four hours after transfection cells were lysed and their protein extracted. The 48 well cell culture plates had their media aspirated and the wells were then washed twice in 37°C PBS. Following this washing process, 65 µl of 1X Passive Lysis Buffer (Promega) was added to each well. The plate was then incubated for 20 minutes on a platform rocker, and then frozen to -20°C. In order to read the luciferase activity from the cell lysate, the plate was thawed again on a platform rocker. Once thawed, 20 µl of lysate was added to a luminometer plate, which was loaded onto a Microlumat Plus LB96V luminometer. To each 20 µl of lysate sample, 50 µl of Luciferase Activating Reagent II (LAR II, Promega) was added. The plate was then read, using a 1 second measurement and the WinGlow software. Following this reading, 50 µl of Stop & Glo (Promega) was added to each sample to measure the Renilla activity; the measurement was carried out using the same exposure time. This measurement of Renilla activity is used as a transfection efficiency control. Six biological repeats were measured once each on the machine. The whole experiment was repeated three times, producing a total of 18 data readings for each condition. Statistical analysis was then carried out by performing a student's t-test.

4.3.6 Identification of candidate trans-acting factors

I used two in *silico* prediction programs to identify candidate *trans*-acting factors which could be binding at the -41 bp site. The two programs I used were both internet based; Promo 3.0 (http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3) (141) and TESS: Transcription Element Search System (http://www.cbil.upenn.edu/cgi-bin/tess/tess).

4.3.7 Nuclear protein extraction

Twenty million SW1353 or MG63 cells were plated on to 500cm^2 cell culture dishes (Corstar). After 24 hours, culture medium was removed and cells were washed in ice-cold PBS.

The PBS was then removed and cells were scraped, using an 18 cm cell lifter, into 5 ml of fresh PBS and centrifuged for 30 seconds at 10 000 *g* at 4°C. Cells were then re-suspended in 1ml of hypotonic buffer (10 mM HEPES, pH 7.6, 1.5 mM $MgCl_2$, 10 mM KCl, 1 mM DTT, 10 mM NaF, 1 mM Na_3VO_4, 0.1% Tergitol (volume/volume), 1x complete protease inhibitor cocktail tablet per 50 ml solution (Roche)) and incubated on ice for 15 minutes. Cells were re-pelleted as described in *Chapter 4.3.4* and the supernatant containing cytosolic proteins was collected, snap-frozen on dry ice, and stored at -80°C. In order to fractionate the nuclei, the pellet was re-suspended in ice-cold hypotonic buffer supplemented with 0.25 M sucrose. Again, cells were centrifuged and the pellet was re-suspended in 1.5 ml high salt buffer (20 mM HEPES, pH 7.9, 420 mM NaCl, 20% glycerol (volume/volume), 1 mM DTT, 10 mM NaF, 1 mM Na_3VO_4, 1x complete protease inhibitor cocktail tablet per 50 ml of buffer). Following a 30 minute incubation on ice, the cells were centrifuged for the final time at 10 000 *g* for 2 minutes at 4°C and the supernatant containing nuclear protein was transferred to a new 1.5 ml tube and stored at -80°C.

4.3.8 Electrophoretic mobility shift assay (EMSA)

Fluorescently labeled (5'DY682) oligonucleotides (Eurofins MWG Operon) were re-suspended in H_20 (Sigma) to a concentration of 100 pmol/µl. To create double-stranded probes a pair of complimentary oligonucleotides were heated to 95^{o}C in an EMSA annealing buffer (100 mM Tris HCl, 500 mM NaCl, 10 mM EDTA). Following this incubation, the reaction was allowed to cool to room temperature for 2-3 hours. A native 2.3% acrylamide gel (weight/volume) (3.2 µl polyacrylamide, 2.6 µl 5X TBE, 19.1 µl diH_2O, 88 µl of ammonium persulfate (APS) and 25 µl of TEMED) was prepared the day before the reaction and left to set at 4°C overnight. The gel was pre-run at 100 V for 30 minutes and at 4^{o}C in 0.5x TBE buffer prior to the loading of samples in order to remove any traces of APS, to equilibriate ions within the running buffer and to ensure constant gel temperature. Following initial optimisation of binding conditions, the Odyssey Infrared EMSA reaction mixture consisted of 2 µl of 10X binding buffer, 2 µl of 25 mM DTT, 1 µg of PolyIC, 1 µl of 1% NP40, 200 fmol of double stranded oligonucleotide, 5 µg of nuclear protein extract brought up to 20 µl with H^2O. An additional 1 µl of unlabeled competitor oligonucleotide was added to the competitor reaction. These reaction mixes were incubated at room temperature for 20 minutes in the dark. When using unlabeled competitor oligonucleotides, these were added prior to the addition of the fluorescently labeled oligonucleotides. 10X Orange Loading dye (Li-Cor Biosciences) was then added (1x final concentration) and samples were loaded onto the gel. Electrophoresis was performed at 100 V at 4°C in the dark for approximately 4 hours, or until the dye front had reached the end of the gel. Visualisation of the EMSA gel was carried out using an Odyssey Infrared Imager (Li-Cor). When carrying out supershift EMSAs, 2 µg of antibody (YY1 antibody YY1(C-20) cat number sc-281X, raised in rabbit) (Transcruz, Santa Cruz) was added in place of the unlabelled competitor oligonucleotides. Table 4.1 lists the nucleotide sequences of the two -41 bp probes and of the competitor oligonucleotides used.

Table 4.1 Nucleotide sequences of the -41 bp probes and of the competitor oligonucleotides used in the EMSA experiments.

Probe/Competitor	Sequence (5'-3')
-41 bp C-allele probe	GCATTACGCCATTCTTCCTTCTTGGAA
-41 bp A-allele probe	GCATTACGCCATTATTCCTTCTTGGAA
abaA Competitor	AATTGGACCCATTCTTTAGTACGTAGCA
GRLF Competitor	AATTGGACCCATTCTTCCTTCTGTAGCA
SOX9 Competitor	AATTGGACGGATCCTGTCGTACGTAGCA
SEF4 Competitor	AATTGGACCCATTCTTTAGTACGTAGCA
UCRBP Competitor	AATTGGCGCCATTCTTTAGTACGTAGCA
VDR Competitor	AATTGGACCCATTCTTCCGTACGTAGCA
YY1 Competitor	AATTGGACCCATTTTTTAGTACGTAGCA

4.3 Results

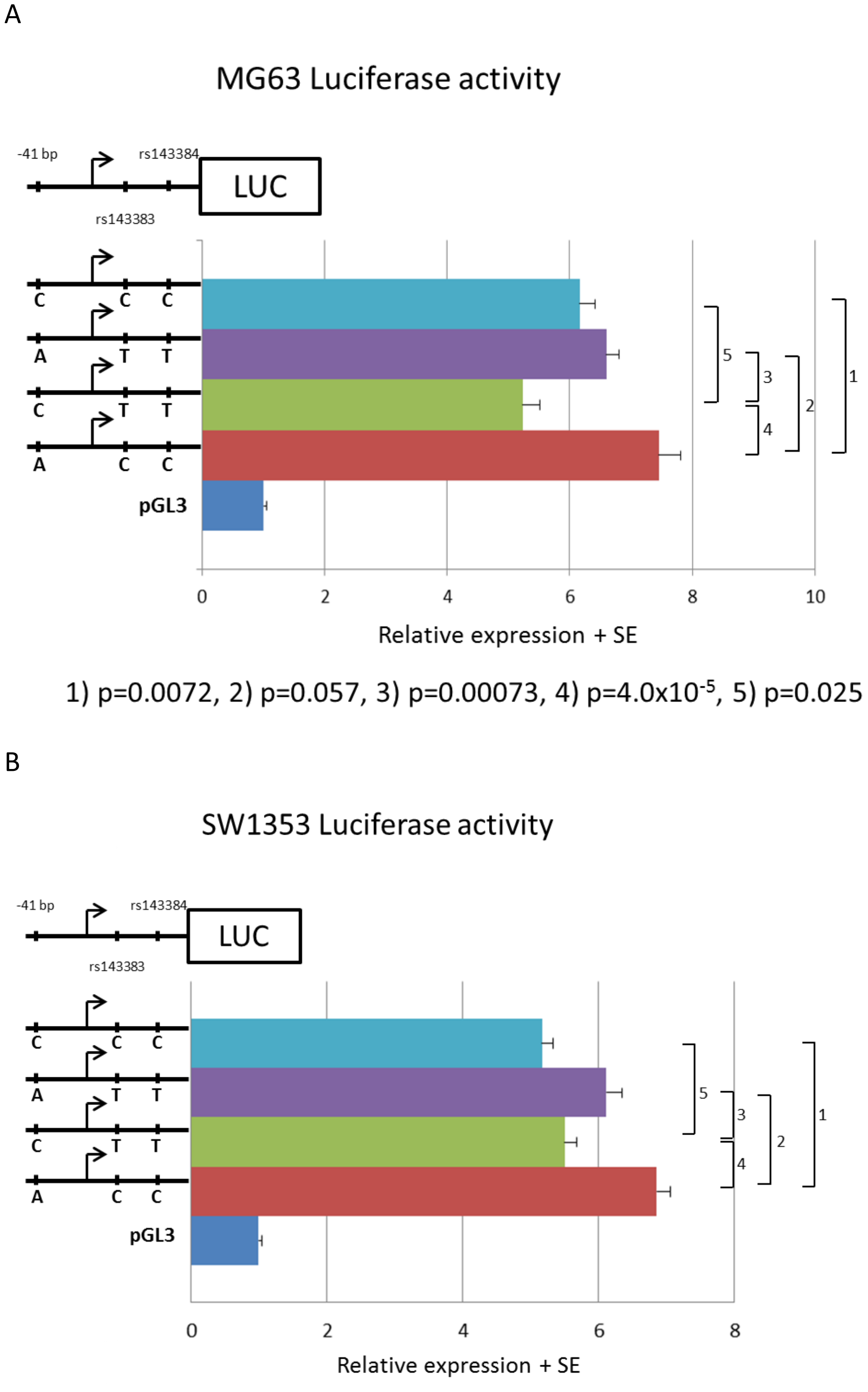
4.3.1 Molecular haplotyping of the -41 bp patient

The -41 bp variant was discovered in an individual who was heterozygous at this site and also at rs143383 and rs143384. Cloning and sequencing of this person's DNA revealed that the haplotypes that the individual carried were A-C-C (-41 bp (A) - rs143383 (C) - rs143384 (C)) and C-T-T (-41 bp (C) - rs143383 (T) - rs143384 (T)). I therefore investigated the effect of the -41 bp variant within the context of these natural haplotypes and in the context of the alternative A-T-T and C-C-C haplotypes.

4.3.2 Luciferase assays

Of the three cell lines which I transfected, one cell line – CH8 – was unable to be efficiently transfected. Luciferase readings from the transfection of this cell line were never over the background reading, indicating transfection had not occurred. This result was consistent for all 5 vectors over 4 repeated experiments, with 6 replicates in each.

In the comparison of A-C-C with C-T-T it was apparent that A-C-C drove greater luciferase expression than C-T-T, with a 30% difference ($p = 4.0 x10^{-5}$) seen in MG63 and a 20% difference ($p = 2.5x10^{-5}$) seen in SW1353 (Figures 4.3 and 4.4). As noted in *Chapter 1.8*, it is already known that the T-allele of rs143383, when in combination with the T-allele of rs143384, results in reduced expression. As such, the greater expression seen here for the A-C-C haplotype could reflect the functionality of rs143383 and rs143384 rather than any functionality of the -41 bp variant. This dilemma could be resolved by studying the alternative haplotypes, A-T-T and C-C-C. When the A-T-T haplotype was compared to the C-T-T haplotype, A-T-T drove greater expression, with a 21% difference (p = 0.00073) seen in MG63 and a 10% difference (p = 0.044) seen in SW1353. Furthermore, when the C-C-C haplotype was compared to the A-C-C haplotype, C-C-C resulted in reduced expression, with a 17% difference (p = 0.0072) observed in MG63 and a 25% difference ($p = 5.0 x 10^{-7}$) observed in SW1353. This data clearly demonstrates that the -41 bp variant is itself functional and that the A-allele discovered in the original deep-sequencing experiments mediates increased gene expression, in the two cell lines that I tested.



1) p=0.0072, 2) p=0.057, 3) p=0.00073, 4) p=$4.0\text{x}10^{-5}$, 5) p=0.025

1) p=$5.0\text{x}10^{-7}$, 2) p=0.022, 3) p=0.044, 4) p=$2.5\text{x}10^{-5}$, 5) p=0.21

Figure 4.3 Results of luciferase (LUC) reporter assays of *GDF5* promoter/5'UTR constructs in the osteogenic MG63 cell line (A) and the chondrogenic SW1353 cell line (B). A schematic drawing of the promoter/5'UTR construct is shown at the top left of each panel and highlights the relative positions of the -41 bp, rs143383 and rs143384 polymorphisms. Data are the fold expression in relation to the control empty pGL3 vector, and are shown as the mean and standard error (SE) from 3 independent experiments, each performed with 6 biological repeats. P-values were calculated using a student's t-test.

	Relative Light Units			
	ACC	**CTT**	**ATT**	**CCC**
MG63	7.45	5.23	6.60	6.16

% difference	CTT	ATT	CCC
ACC	30	11	17
CTT	-	21	15
ATT	-	-	7

	Relative Light Units			
	ACC	**CTT**	**ATT**	**CCC**
SW1353	6.86	5.50	6.12	5.17

% difference	CTT	ATT	CCC
ACC	20	11	25
CTT	-	10	6
ATT	-	-	16

Figure 4.4 The relative mean light unit outputs from each plasmid following renilla correction (top) and the percentage differences in expression between each (bottom), in MG63 (left) and SW1353 (right) cells.

What was particularly striking about the luciferase data was the discovery that introducing a -41 bp A-allele onto a rs143383-rs143384 T-T haplotype reverses the capacity of this T-T haplotype to mediate reduced expression. This is apparent when one compares the relative expressions seen between the C-C-C and the A-T-T haplotypes; in both cell lines the A-T-T expression is greater despite it harboring the T-alleles of both rs143383 and rs143384.

4.3.3 Electrophoretic mobility shift assays

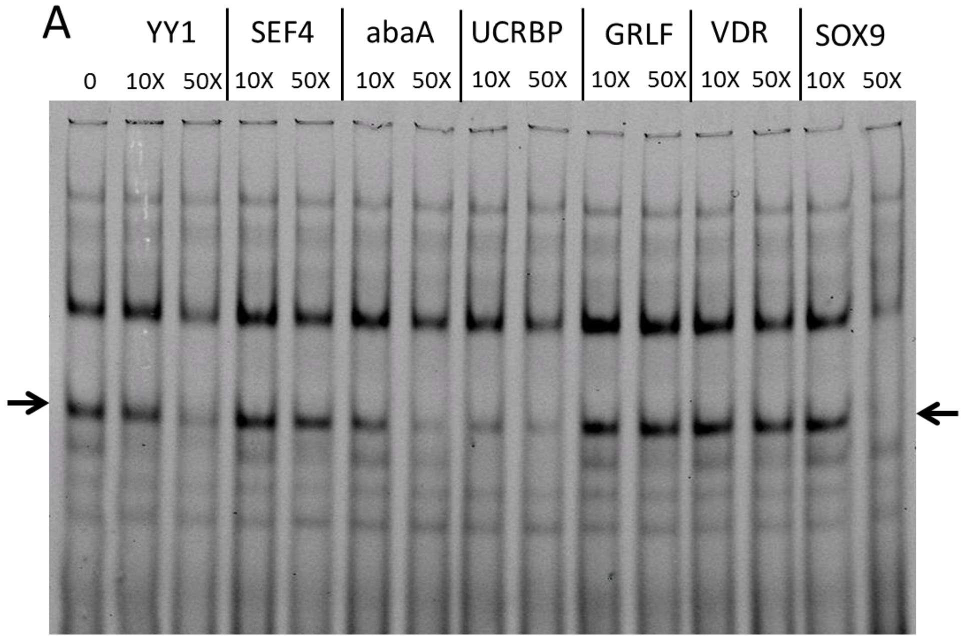
Having established that the -41 bp variant is functional I went on to use EMSAs to characterize trans-acting factors that bind to this site. I investigated both alleles of -41bp and nuclear extracts from MG63 and SW1353 cells. The -41 bp variant resides within a sequence of GDF5 that is conserved amongst mammals (Figure 4.5). Using the online prediction databases TESS and Promo 3.0, I identified several transcription factors whose consensus sequences indicated a potential binding relationship with the sequence encompassing -41 bp and in which this binding is predicted to be influenced by the C/A change (Figure 3.5, Section 3.4.1). I tested the ability of YY1, UCRBP (a synonym of YY1 but with a different registered consensus sequence), SOX9, abaA, SEF4, VDR, and GRLF to out-compete the binding to the C-allele and the A-allele probes, using the competitor oligonucleotides listed in Table 4.1. I firstly tried two concentrations of each competitor – 10x and 50x – and saw possible band intensity changes in

YY1, abaA, UCRBP1 and SOX9 using MG63 nuclear extract (Figure 4.6A). I repeated the negative results using SW1353 nuclear extract and again saw no band intensity changes (Figure 4.6B). By titrating concentrations of the competitor from 5x to 50x probe concentration I was able to determine if the affinity for transcription factor binding differed between the two alleles. I did not detect any consistent band intensity depletion when using abaA, GRLF, SEF4, SOX9 or VDR competitors. I did however detect band intensity depletion for YY1 and its synonym UCRBP in both MG63 and SW1353, and that in each case the site-specific band produced by fluorescent probe binding was diminished by a lower concentration of YY1 consensus sequence probe when testing the C-allele compared to the A-allele (Figures 4.7-4.10). This indicates that YY1 binds to this locus and that binding is stronger to the A-allele than the C-allele.

```
Homo sapiens -41 bp C TGGAGCACACAGGCAGCATTACGCCATTCTTCCTTCTTGGAAAAATCCC 49
Homo sapiens -41 bp A TGGAGCACACAGGCAGCATTACGCCATTATTCCTTCTTGGAAAAATCCC 49
Pan troglodytes       TGGAGCACACAGGCAGCATTACGCCATTCTTCCTTCTTGGAAAAATCCC 49
Equus caballus        TGGAGCACACAGGCAGCATTACGCCATTCTTCCTTCTTGGAAAAATCCC 49
Mus musculus          TGGAGCGCACAGGCAGCATTACGCCATTCTTCCTTCTTGGAAAAATCCC 49
```

Figure 4.5 Nucleotide alignment of the -41bp variant site. The wild type C-allele is shown in red and the mutant A-allele is shown in blue. The homologous sequences of Pan troglodytes (common chimpanzee), Equus caballus (horse) and Mus musculus (house mouse) are also shown.

I next carried out an antibody supershift EMSA for both MG63 and SW1353 and this confirmed that YY1 does bind to the -41 bp locus (Figure 4.11). Although I didn’t observe a supershift of the YY1 specific band, I did observe that the YY1 specific band disappeared when antibody was added to the C-allele, and certainly diminished when added to the stronger affinity A-allele. It is possible that the antibody is binding directly over the site of the fluorescently tagged oligonucleotide and that is why I see a band disappearance/diminishment rather than a shift.



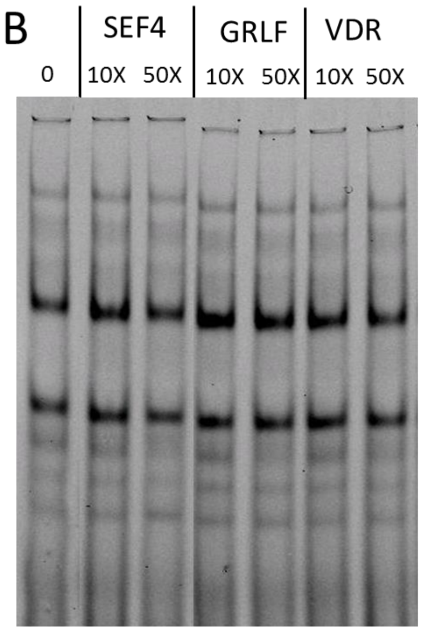


Figure 4.6 EMSA performed on MG63 nuclear extract (A) and SW1353 nuclear extract (B), using increasing concentrations of transcription factor binding consensus sequence unlabeled probes. Initial two concentration (10x and 50x) EMSAs carried out using unlabeled probes to test for their suitability to out compete the site specific band (denoted by arrows). A) shows that YY1, abaA, UCRBP and SOX9 may be possible candidates when using MG63 nuclear extract. B) shows those negative results repeated with SW1353 nuclear extract but again they still show no out competition of any bands.

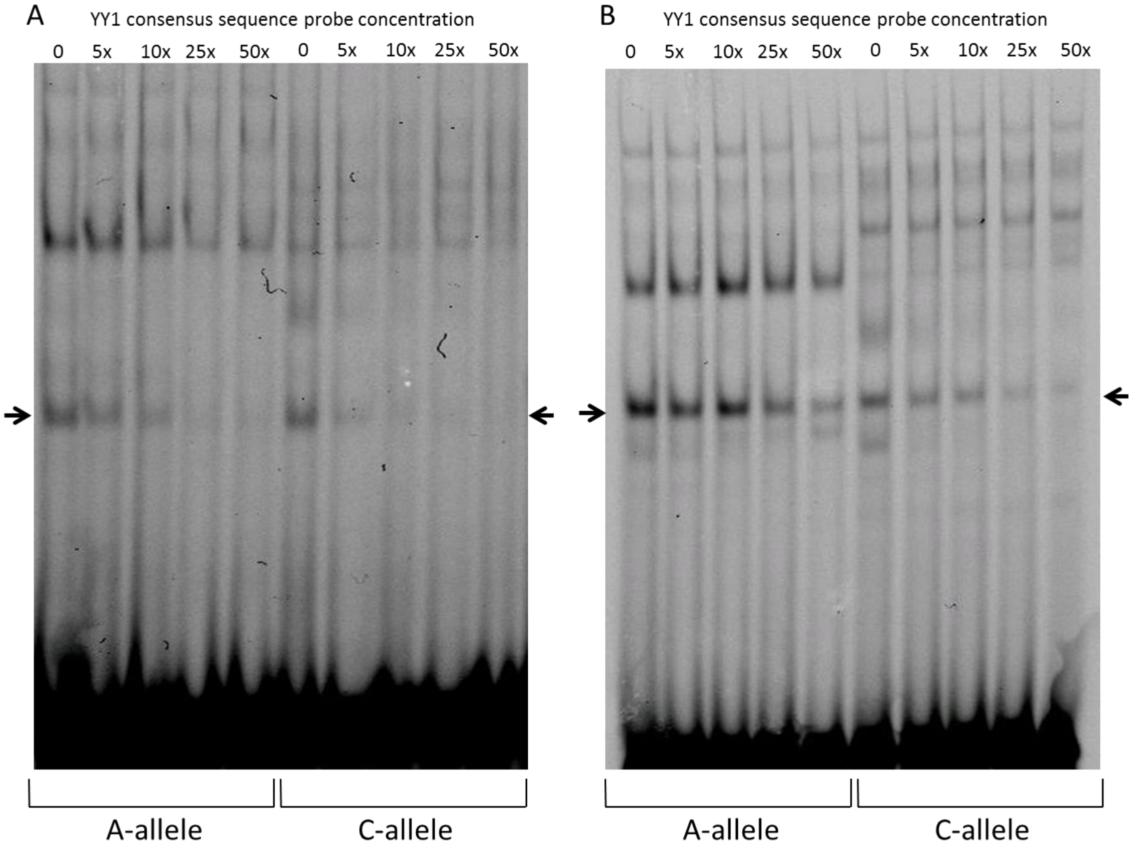

A
YY1 consensus sequence probe concentration
0
5x
10x
25x
50x
0
5x
10x
25x
50x
A-allele
C-allele
B
YY1 consensus sequence probe concentration
0
5x
10x
25x
50x
0
5x
10x
25x
50x
A-allele
C-allele

Figure 4.7 EMSA performed on MG63 nuclear extract (A) and SW1353 nuclear extract (B), using increasing concentrations of YY1 consensus sequence unlabeled probe. Both EMSAs show that the site specific band (denoted by the arrows) produced by nuclear protein/fluorescent probe binding can be diminished by a lower concentration of YY1 consensus sequence probe when testing the wildtype C-allele compared to the mutant A-allele. This indicates that YY1 is a strong candidate for binding to this locus and that binding is stronger to the A-allele than the C-allele.

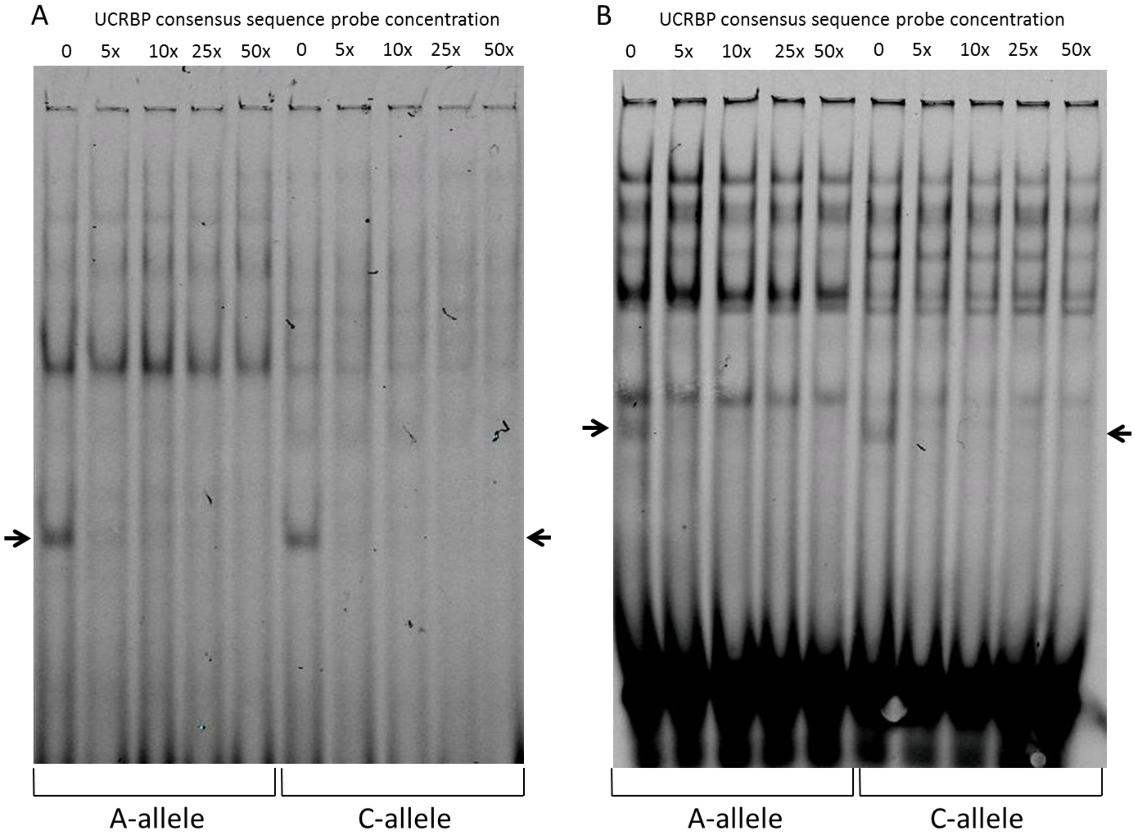

A
UCRBP consensus sequence probe concentration
0 5x 10x 25x 50x 0 5x 10x 25x 50x
A-allele
C-allele
B
UCRBP consensus sequence probe concentration
0 5x 10x 25x 50x 0 5x 10x 25x 50x
A-allele
C-allele

Figure 4.8 EMSA performed on MG63 nuclear extract (A) and SW1353 nuclear extract (B), using increasing concentrations of UCRBP consensus sequence unlabeled probe. Both EMSAs show that the site specific band (denoted by the arrows) produced by nuclear protein/fluorescent probe binding can be diminished by a lower concentration of UCRBP consensus sequence probe when testing the wildtype C-allele compared to the mutant A-allele. URCBP is an synonym of YY1 but with a different registered consensus sequence. This verifies YY1 as a strong candidate for binding to this locus and that again binding is stronger to the A-allele than the C-allele.

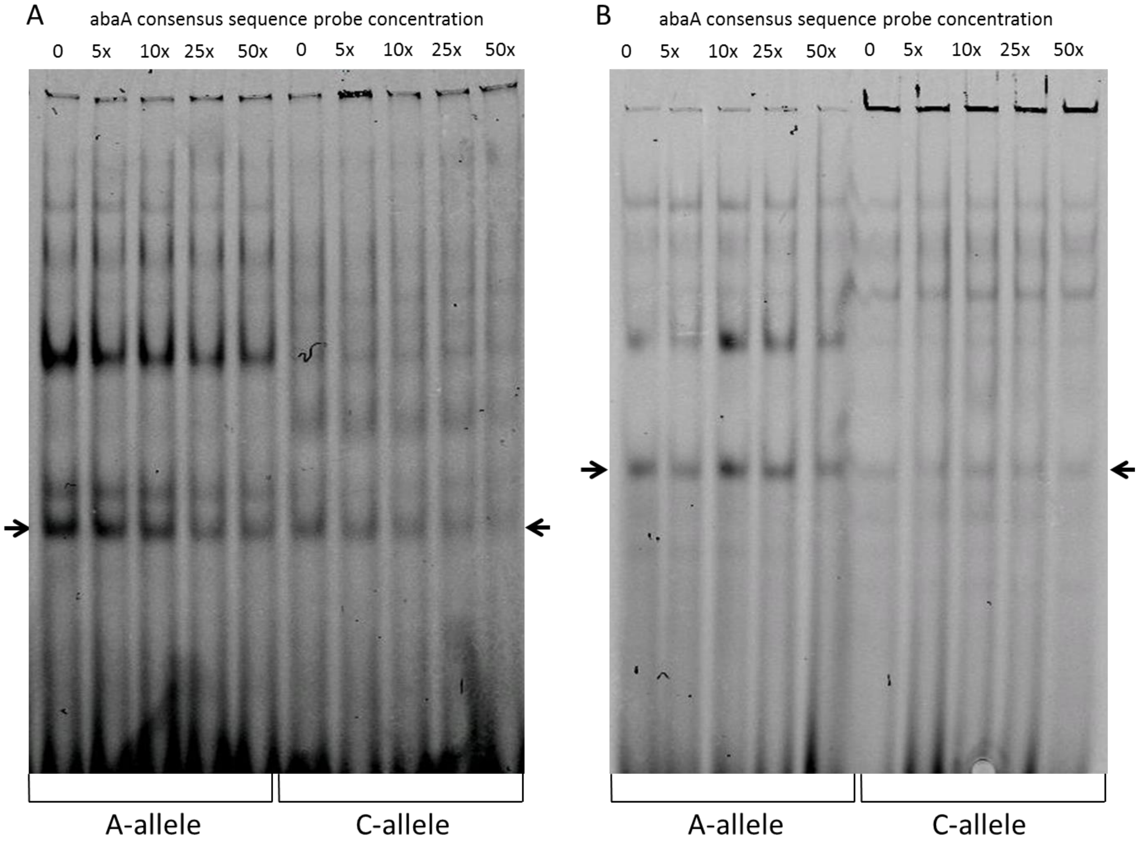

A
abaA consensus sequence probe concentration
0 5x 10x 25x 50x 0 5x 10x 25x 50x
A-allele
C-allele
B
abaA consensus sequence probe concentration
0 5x 10x 25x 50x 0 5x 10x 25x 50x
A-allele
C-allele

Figure 4.9 EMSA performed on MG63 nuclear extract (A) and SW1353 nuclear extract (B), using increasing concentrations of abaA consensus sequence unlabeled probe. Both EMSAs show that the site specific band (denoted by the arrows) produced by nuclear protein/fluorescent probe binding cannot be diminished by varying the concentration of a abaA unlabeled probe.

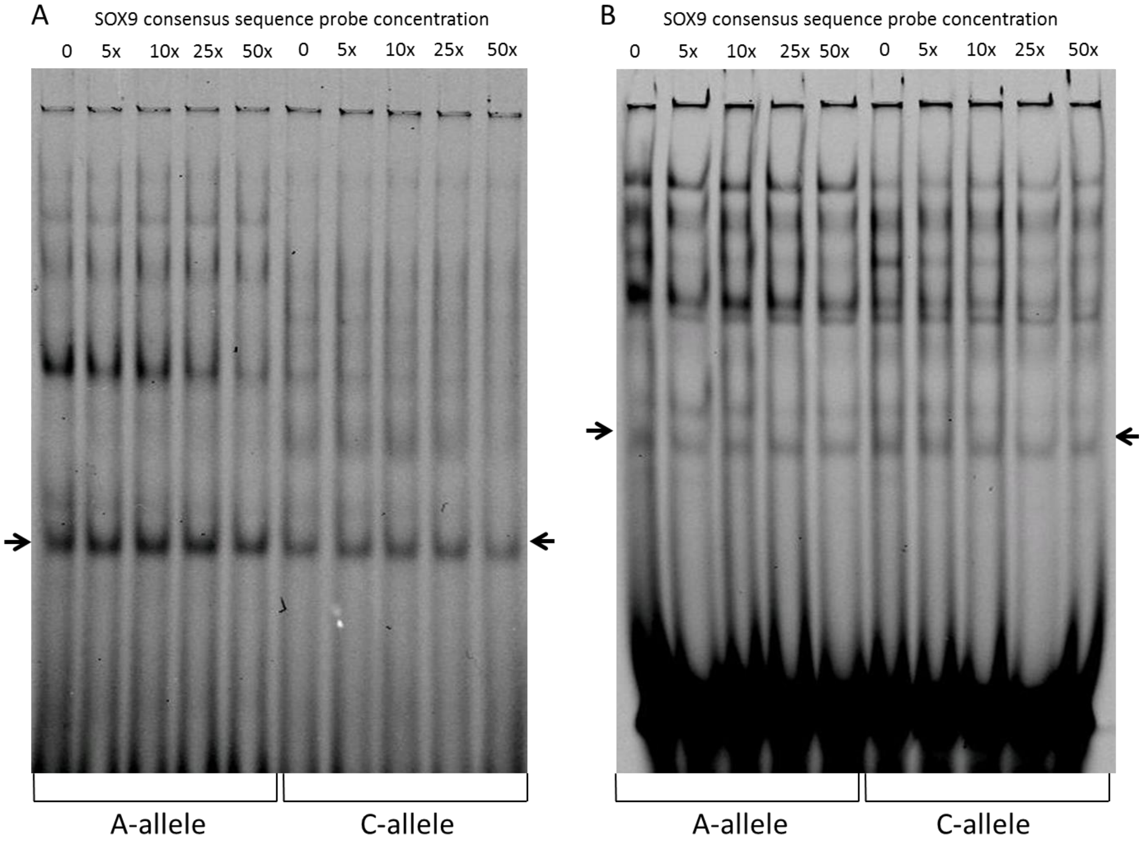

A
SOX9 consensus sequence probe concentration
0
5x
10x
25x
50x
0
5x
10x
25x
50x
A-allele
C-allele
B
SOX9 consensus sequence probe concentration
0
5x
10x
25x
50x
0
5x
10x
25x
50x
A-allele
C-allele

Figure 4.10 EMSA performed on MG63 nuclear extract (A) and SW1353 nuclear extract (B), using increasing concentrations of SOX9 consensus sequence unlabeled probe. Both EMSAs show that the site specific band (denoted by the arrows) produced by nuclear protein/fluorescent probe binding cannot be diminished by varying the concentration of a SOX9 unlabeled probe.

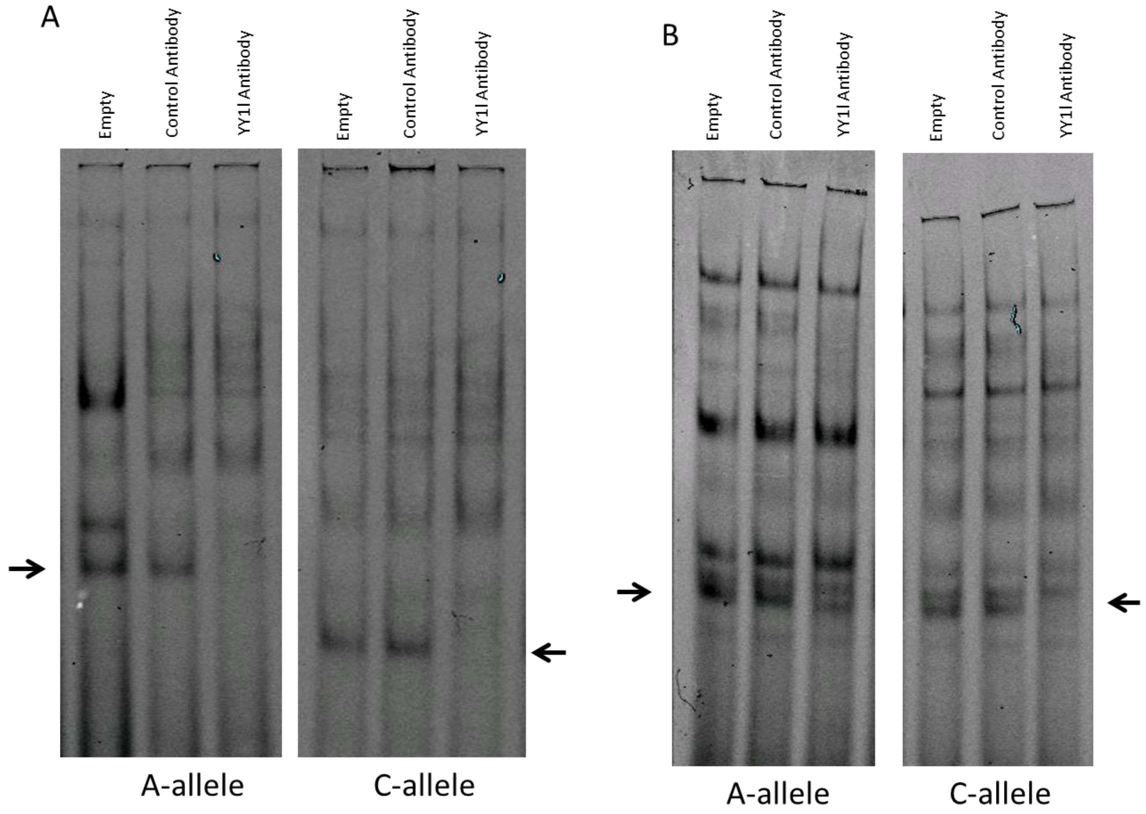


Figure 4.11 Antibody supershift EMSA of (A) MG63 nuclear extract and (B) SW1353 nuclear extract. The addition of the YY1 antibody clearly shows a reduction in the intensity of the site specific band (denoted by the arrows) when compared to when either no antibody (empty) or a non-targeting (control) antibody was added. In the case of the C allele the band is completely diminished. A shift in the band is not observed as I hypothesize the antibody is binding directly over the fluorescently tagged oligonucleotide.

4.4 Discussion

Using a range of techniques I have demonstrated that the -41 bp variant discovered by my deep-sequencing analysis of *GDF5* (Chapter 3) is functional, with the novel A-allele mediating increased gene expression. One of the particularly exciting outcomes of my luciferase assays was the discovery that this allele is able to compensate for the reduced expression that is mediated by the OA-associated T-allele of rs143383. This result clearly demonstrates that the activity of an OA susceptibility allele is highly context specific and that negative effects on gene expression mediated by *cis*-acting regulatory sites are potentially modifiable by targeting other *cis* sites within the gene.

Although previously implicated in binding to the promoter of GDF5 (142), I have identified YY1 as a *trans*-acting factor that differentially binds to the C and A alleles of the -41 bp variant, with more avid binding to the A allele. YY1 (Yin Yang 1) is a widely expressed zinc-finger nuclear protein that is conserved between species and which can act as a transcriptional repressor or activator, depending on cellular context (143). It has roles in cellular proliferation, differentiation and apoptosis, with *yy1* knockout mice having *in utero* lethality (144). YY1 expression has been associated with the development of malignant phenotypes of human cancers, along with tumour progression, including metastasis and also in tumour survival (145). However, due to its ability to act as a repressor and as an activator, its role in tumor progression is controversial as YY1 plays a key role in both cell proliferation and apoptosis. As far as I am aware the protein has not so far been implicated in OA pathogenesis. Its discovery as a differential regulator of *GDF5* expression now opens up the possibility that it can be exploited to modulate the expression of this important OA susceptibility locus.

When I carried out the deep-sequencing of *GDF5* which I described in Chapter 3, my primary goal was to assess whether this gene harbored other, novel, polymorphisms that could directly contribute to OA genetic susceptibility at a population level; by definition, such polymorphisms would need to be common in the population. My experiment however only identified extremely rare variants, with MAFs ≤0.006% (Table 3.3). My functional analysis of the -41 bp variant has demonstrated that despite its low frequency this mutation offers insight into the regulation of expression of *GDF5* and provides us with a protein, YY1, that can be exploited to alleviate the OA risk coded for by this gene. Deep sequencing therefore offers not only a means to identify novel susceptibility alleles but also a means to identify a mechanism to counteract the effects of those that are already known.

In order to further understand YY1's role in the regulation of *GDF5*, I collaborated with another member of my research group, Catherine Syddall, who demonstrated that YY1 was expressed (mRNA and protein) in cartilage and also performed a knockdown of *YY1* to assess what effect this had on *GDF5* expression. It was found that knock down of *YY1* reduces the expression of *GDF5*, supporting my data. This research has now been published under the title *"A rare variant in the promoter of the OA-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate gene expression"* within the European Journal of Human Genetics (146).

Chapter 5: General Discussion

Osteoarthritis (OA) is the most common musculoskeletal disease, and in some populations up to 90% of the over 65s suffer from OA in one or more of their joints (2). It is characterised by focal areas of loss of articular cartilage resulting in the deterioration of joint form and function. In Europe an osteoarthritic joint is replaced once every 90 seconds, placing an enormous economic burden upon governments and heath care authorities (1). Genetic analysis is a useful tool for uncovering biological factors important in healthy body function and in the initiation and progression of common diseases. At present, OA research has yielded fewer robust genetic associations when compared to other common complex diseases such as type 1 diabetes or Crohn's disease (96, 113). The heterogeneity of OA has a large responsibility for situation. OA is an end stage phenotype achieved from many different single starting points or indeed multiple starting points, stemming from multiple possible triggers (147). Heterogeneity exists amongst genders, ethnicities, anatomical sites and this manifests observed differences in joint pathology even at end stage disease.

The strategies employed in furthering genetic research into OA can be classified into two categories: the candidate gene approach and the genome-wide approach. The former relies on a hypothesis leading to the candidate gene whereas the latter is hypothesis-free and agnostic of gene function. Both strategies have yielded significant associations to disease and this thesis demonstrates research following each strategy.

Genome-wide association scans (GWASs) compare the frequencies of SNPs within a cohort of affected individuals and a cohort of people either free of disease or representative of the normal population as a whole. By exploiting linkage disequilibrium (LD) we are able to cover the whole genome by genotyping only a fraction of the total number of SNPs present within the human genome. The consequence of this is that any SNP providing an association signal is not necessarily the SNP that is causal with regard to disease. A GWAS carried out in

Rotterdam reported a locus on chromosome 7q22 with an association to knee and/or hand OA (75). This locus contains a region of high LD which encapsulates six genes, none of which are obvious candidates for OA susceptibility: *BCAP29* (encoding B cell receptor-associated protein 29), *COG5* (encoding component of oligomeric golgi complex 5), *DUS4L* (encoding dihydrouridine synthase four-like), *GPR22* (encoding G protein-coupled receptor 22), *HBP1* (encoding HMG-box transcription factor 1) and *PRKAR2B* (encoding protein kinase-cAMP-dependant-regulatory-type II-β). A further investigation failed to find *GPR22* expressed anywhere within the joint and went on to single out *HBP1* as possessing an expression quantitative trait locus in carriers of the associated allele (77). This demonstrates that functional studies are a crucial in order to understand the results from GWASs

Following the inception of the arcOGEN Consortium and the subsequent publication of Stage 1 of its GWAS findings, one association signal was that of the T-allele of rs2615977, a G/T transversion, located within intron 31 of the type XI collagen α1 polypeptide encoding gene *COL11A1* (82). Unlike the Rotterdam GWAS and the 7q22 locus, an association to a collagen gene is a much more compelling candidate due to type XI collagen being such key components of the articular cartilage. The suitable candidacy of *COL11A1* is further reinforced due to the fact a Japanese group found a common non-synonymous SNP, rs1676486, within exon 62 of *COL11A1*, was associated with lumbar disc herniation (LDH) in a Japanese population (83). Both LDH and OA are characterised by age-associated, loss of function and degradation of cartilage.

Chondrodysplasia in the mouse model (*cho/+*) is caused by a frame-shift mutation within *Col11a1*, producing a truncated α1(XI) chain, with an inability to regulate correct collagen fibril thickness, and as a result *cho/+* mice have thicker fibrils than wild-type mice (119). These haploinsufficient *cho*/+ mice develop OA-like changes in both their knee joints, yet show no other skeletal abnormalities (120). The type II collagen network is altered whereby there is a higher proportion of pericellular type II collagen than in wild type cartilage.

This is a result of the reduced expression of type XI collagen within the *cho*/+ mouse which in turn increases the levels of both discoidin domain receptor 2 (*Ddr2*) and matrix metalloproteinase 13 (*Mmp13*), which lead to cartilage degradation (120). In humans, type IX collagen mutations are found in certain forms of Stickler syndrome, a rare and Mendelian disorder of severe OA that presents during adolescence (84, 85).

A key mechanism of OA susceptibility in humans has been shown to be that of allelic expression imbalance (AEI). The T-allele of the 5' UTR SNP rs143383 of *GDF5* exhibits AEI and it is one of the most robust associations to OA we currently know of. I therefore hypothesised that the OA association to *COL11A1* marked by rs2615977, as reported by the arcOGEN GWAS (82), was due to AEI acting upon *COL11A1* in OA articular cartilage. Additionally, I was keen to test whether imbalance seen in LDH intervertebral disc cartilage is present within OA cartilage and should this prove to be the case, I also wanted to see if genotype at this SNP associates with OA as it does with LDH.

Upon testing for AEI within *COL11A1*, I did not detect a correlation between genotype at the arcOGEN GWAS OA associated SNP rs2615977, and the level of expression of *COL11A1*. It should be pointed out that this study was carried out on OA cartilage samples from patients with end stage OA who elected for total joint replacement. It is possible that rs2615977 marks an association to a polymorphism which impinges upon *COL11A1* expression at a time-point during skeletogenesis or in adolescent tissues before OA is present, affecting the structural integrity of the cartilage as it is laid down or in early life, initiating the long term progression of OA as AEI may not only be spatial but also temporal (122, 123). To further investigate this association will be complex as no transcript SNP was found within the gene that reflected a similar frequency or a high degree of LD. If the association is marking a polymorphism which is exhibiting AEI during development or adolescence then it is obviously very difficult to test for this *in vivo* due to the inability to acquire tissue from such a time-point. A possible direction to take this investigation further would be to extract mesenchymal stem cells from the bone marrow of patients undergoing total joint replacement and culture these cells *in vitro* while differentiating them down the chondrocyte lineage. During this differentiation, as the cells undergo phenotypic change, RNA could be extracted and tested for AEI. It is also possible that rs2615977 is marking a specific haplotype rather than an individual polymorphism, assessing *COL11A1* expression stratified by haplotypes could yield further results.

I then wanted to test if AEI within *COL11A1* marked by rs1676486 in the intervertebral disc of LDH patients by Mio *et al.* is also true of articular cartilage in OA patients. I found this to be the case; I observed a reduction in the expression of the T-allele in articular cartilage, just as was seen in the intervertebral disc of LDH patients. However upon testing the T-allele of rs1676486 for an association to OA susceptibility only a very modest significance was observed. It can therefore be concluded that despite carriers of the T-allele of rs1676486 exhibiting reduced expression of *COL11A1*, this does not make them susceptible to acquiring OA, unlike LDH. There are clearly differing mechanistic effects between articular and intervertebral disc cartilages whereby a reduction in *COL11A1* expression can be tolerated in articular cartilage but it leads to disease progression within intervertebral disc cartilage. However, due to seeing AEI in *COL11A1* mediated by rs1676486 in both tissues, it implies that the same *trans*-acting factors are at work in both types of cartilages.

Following my research into a locus highlighted by a GWAS, I then investigated a candidate gene, *GDF5*, which is a member of the TGFβ superfamily (110). The TGFβ signalling pathway is involved in the homeostasis of all articular joint tissues and GDF5 itself participates in development, maintenance and repair of bone, cartilage, ligament and tendon via extracellular signalling (16). This led Miyamoto and colleagues to genotype common polymorphisms in *GDF5* in an Asian cohort of OA cases and controls. This resulted in rs143383, a SNP in the 5′ UTR of the gene, to be associated with OA (52). Replication in European cohorts have shown that this association has global significance to OA susceptibility (55–57), with functional studies indicating that rs143383 itself is influencing the susceptibility risk of OA due to the disease associated T-allele mediating a reduction in *GDF5* transcription, relative to the C-allele (52–54).

Further investigations into the effects of common polymorphisms indicated that genotype at another 5’ UTR SNP within *GDF5*, rs143384, is able to modulate the AEI observed at rs143383, and it was also demonstrated that another SNP in the 3’ UTR, rs56366915, is able to subject *GDF5* to further AEI, which is independent of genotype at both rs143383 and rs143384 (54). These findings display *GDF5*’s potential to be influenced by genetic polymorphism within it, by both an interacting polymorphisms (in the case of rs143383 and rs143384) and independent polymorphism (in the case of rs56366915). These three discussed polymorphisms are all common, with high minor allele frequencies (MAFs).

I set out to further assess the genetic architecture of *GDF5* to evaluate whether this important OA susceptibility locus harbours any more variants that may have lower MAFs than 0.01, which could impact upon OA aetiology. It has been demonstrated that loci harbouring alleles that confer disease susceptibility risk are also likely to contain other alleles able to influence the same trait (132). Furthermore, studies have shown that rapid expansion of the human population, coupled with weak purifying selection has allowed for a plethora of rare, and population-specific variants, many of which are potentially deleterious (148, 149). This is the first example of a deep-sequencing analysis of an OA susceptibility locus looking for the missing heritability of the disease.

Following the sequencing of *GDF5*, from a total of 1 800 chromosomes, only six novel variants were detected, all with very low MAFs (≤0.0006). Of these six rare variants, three suggested that they had the potential to be functional. One variant resides within the proximal promoter of the gene and two are non-synonymous base changes. The first of the non-synonymous base changes lies within the immature form, or ‘pro-GDF5’, and as such this makes further studies such as modelling difficult as the crystal structure for GDF5 is currently only available for the cleaved and therefore active, mature GDF5. The PolyPhen database predicted that this glycine to arginine variant at position 81 would have a benign effect. The second non-synonymous mutation, a threonine to arginine substitution at position 469 lies within the mature form of GDF5 and as such allowed me to model the change *in silico*. The introduction of an arginine, with its positively charged and large side chain, would impact upon the folding of the protein, partly due to no common rotomer being able to be accommodated in to the wildtype polypeptide conformation of GDF5. PolyPhen also predicted this amino acid change to be possibly damaging. A group in Germany have a chicken micromass culture assay in which they are able to assess the activity of GDF5 as an extracellular signalling molecule (100), and I set up a collaboration with them to further investigate Thr469Arg but this experiment requires further optimisation due to measurement sensitivity issues. The ethics used to collect the OA and control samples used in this thesis does not allow us to retrospectively contact the individuals. It is therefore not possible for us to assess whether the two individuals harbouring the amino-acid substitutions have any overt skeletal abnormalities that could be the result of the mutations.

While I found six novel variants, they each existed in one individual only, meaning while they could be having an effect upon familial OA susceptibility, they are not acting at the population level. It can therefore be concluded, in terms of population genetics, that I observed an absence of rare variants within *GDF5*. This could be due to *GDF5* being subjected to a population bottleneck in the past. A population bottleneck is a sudden reduction in population size, and with it reduced heterozygosity within that population (134, 135) (Figure 5.1). During a population bottleneck, MAFs and genetic diversity are severely affected by random genetic drift (136). It is possible that the current genetic architecture of *GDF5* is reflective of a bottleneck imparted upon the population, whereby the seven common polymorphisms that were confirmed in my study made it through the bottleneck with all other potential *GDF5* alleles lost. The six very rare novel variants that I discovered may be the result of mutation that has occurred after the bottleneck, as the population is expanding once again and greater heterogeneity is developing. It is believed that, in the last 10 000 years, the global population has exploded from a few million to the estimated 7 billion that it is today. Such an explosion allows for rare variants within sub-populations due to mutation, which can exact upon complex disease (137). This hypothesis that *GDF5* has been subjected to a bottleneck is supported by a recent study which found that T-alleles of both rs143383 and rs143384 are both associated with height and have undergone positive selection within the East Asian population (138). As mentioned, the T-allele of rs143383 has been robustly associated with OA across both Asian and European populations and also shown to correlate with reduced expression of *GDF5*; the T-allele of rs143384 has been shown to modulate the further reduction in *GDF5* expression. Wu and colleagues postulate that the age of these two T-alleles correlates with a period in human history which saw an increase in early agriculture and a reduction of body mass as the need for a large hunter-gatherer, highly energy dependant body-type declined (138). Further evidence for this reduction in body mass during this period is observed when looking at human brain size. One of the best known and major characteristics of human evolution is the increase in brain size, and around 2.5 million years ago the human brain was approximately 750 ml (150). If we jump to only 30 000 years ago this value had doubled with most of the brain volume expansion coming latterly at a rate of around 70 ml per 100 000 years up until this 30 000 years ago mark (151). However, it is less well known that in the remaining 30 000 years up until present day, the human brain has decreased, and at a substantial rate (152, 153). In the last 10 000 years the brain size has dropped from 1 500 ml to 1 240 ml in a European female, this rate of decrease in 10 000 years is 36 times greater than the rate of increase over the previous 800 000 years, suggesting the presence of a strong selection pressure. Brain size and body mass are closely linked in humans and the time when humans began hunting using projectile weapons and developing primitive forms of agriculture coincides with reduced body mass (154) and the emergence of the rs143383 and rs143384 T-alleles.

We know that reduced *GDF5* expression confers shorter limb growth (139), which would produce a shorter stature and lower body mass, so the possession of these two lower *GDF5*-expressing T-alleles could help produce this advantageous phenotype. It is reasonable to suggest that a characteristically high-age related disease like OA would be unlikely to exert a negative selection pressure during this time period and so any OA susceptibility would not manifest itself.

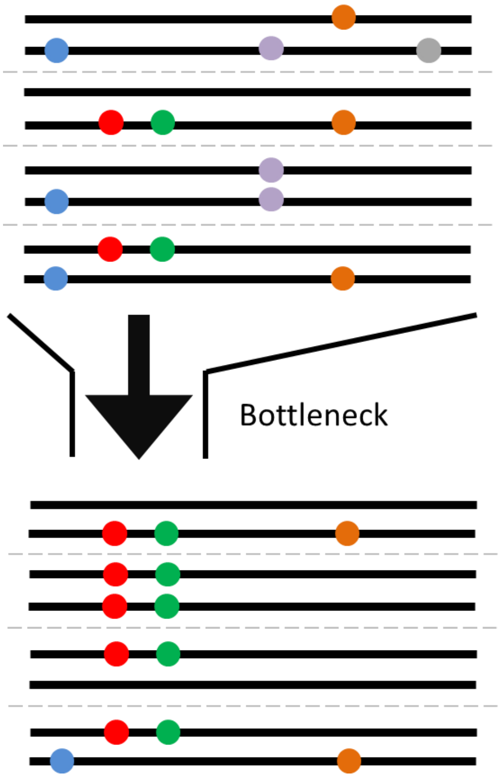


Figure 5.1 Reduced heterogeneity due to a bottleneck. Top) Four diploid copies of a gene, all carrying various mutant alleles of multiple loci (coloured circles). Bottom) A selection pressure is applied whereby the carrying of the red and green mutant alleles is advantageous. The frequencies of these two mutant alleles is greatly altered while some other mutant alleles are lost from the population all together (purple and grey) and others' frequencies are reduced. Overall the population is less heterogeneous following the bottleneck.

Fossil records indicate that a reduction in stature occurred in humans at around the same time as the arrival of the T-alleles of rs143383 and rs143384, which helps support the hypothesis that *GDF5* has passed through a bottleneck. However, it can't be said for certain that a positive selection pressure on these alleles caused a reduction in heterogeneity within *GDF5*. In order to add more weight to this hypothesis it may be prudent to look at the ages of other alleles which have been associated with height. Human height is a complex and polygenic trait, therefore one should see other alleles in other genes undergoing positive selection also at the same point in history.

Collaborating with another research group from Santiago de Compostela, Spain – who sequenced the open reading frame of *GDF5* in a cohort of Spanish and Greek individuals – allowed expansion of the study to incorporate both Northern and Southern Europeans. This research was published in *Osteoarthritis and Cartilage* in 2011 (140). The Spanish group detected the previously reported common *GDF5* polymorphisms but detected no novel variants. Of the six extremely rare variants I found in the UK population, the promoter variant (variant 1, Chapter 3.4.1) and the non-synonymous variant (variant 6, Chapter 3.4.5) appear to have the greatest potential to impart a functional effect upon *GDF5* expression or GDF5 activity, respectively. It can be concluded that *GDF5* possesses no rare variants that can impact at a population level; of course I cannot discount the possibility of rare variants existing within distal *cis*-acting sites such as enhancers or inducers, but there are currently no such sites known to be controlling *GDF5*. It is reasonable to say that any additional OA susceptibility within *GDF5* is due to either common polymorphism or epigenetics.

I investigated the effect that the -41 bp (C/A) variant could have upon *GDF5* expression by firstly creating luciferase constructs of the two naturally occurring haplotypes when coupled with the alleles of rs143383 and rs143384. I also created constructs containing two non-naturally occurring haplotypes, so that I could circumvent the known functional effect of these two common polymorphisms (54). My luciferase assays on both chondrogenic and oseotgenic human cell lines demonstrated that the novel A-allele of the -41 bp variant is functional and not only able to mediate increased expression of *GDF5*, but the increase is of such a degree that it is able to reverse the capacity that the T-T rs143383-rs143384 haplotype has to mediate a reduction in expression of *GDF5*. In both cell lines I observed that the A-T-T (-41 bp – rs143383 – rs143384) haplotype was able to drive expression to a greater degree than the C-C-C haplotype despite the latter containing the two higher expressing C-alleles of both rs143383 and rs143384. This result clearly demonstrates that the activity of an OA susceptibility allele is highly context specific and that negative effects on gene expression mediated by *cis*-acting regulatory sites are potentially modifiable by targeting other *cis* sites within the gene.

I identified YY1 as a *trans*-acting factor which is able to bind differentially to the C- and A-alleles of the -41 bp variant, with an increased binding affinity for the novel A-allele. YY1 is a widely expressed zinc-finger nuclear protein that is conserved between species and which can act as a transcriptional repressor or activator, depending on cellular context (143). It has roles in cellular proliferation, differentiation and apoptosis, with *Yy1* knockout mice having *in utero* lethality (144). As its name suggests, Yin Yang 1 is able to perform conflicting functions and its expression has been associated with the development of malignant phenotypes of human cancers, along with tumour progression, including metastasis but also, conversely, in tumour survival (145). To my knowledge, this protein has not been implicated in OA pathogenesis, and as a regulator of *GDF5* expression there is a possibility that it can be exploited to modulate the expression of this OA susceptibility locus. Due to *YY1* being conserved between species it allows for the possibility of mouse models to be utilised in taking these findings further. For instance, it would be interesting to place an additional *Yy1* gene under the control of a chondrogenic specific promoter such as *Sox9,* so that there is an elevated level of Yy1 protein during chondrogensis. This would be especially interesting in the haploinsufficient brachypodism (bp) mouse (*Gdf5*$^{Bp-J/+}$*)*, to see if an increase in Yy1 during chondrogenesis can protect the *bp* mouse from developing an OA phenotype when challenged.

In summary, this thesis has pursued research into both common variants in GWAS associated loci in the case of *COL11A1*, and rare variants in candidate gene studies in the case of *GDF5* to elucidate genetic factors which play a role in to the common, complex disease of OA. I established that the arcOGEN GWAS association signal to rs2615977 is not due to AEI within *COL11A1* within articular cartilage from end stage OA patients. It is of course possible that the association signal to rs2615977 could have been a false positive from the GWAS. I also found that the AEI within *COL11A1* marked by rs1676486 in the intervertebral disc of LDH patients detected by Mio *et al.* (83) also exists in the articular cartilage in OA patients. However, this AEI does not prove to be a risk factor for OA. By resequencing *GDF5* in 962 individuals, I carried out the first deep-sequencing of an OA susceptibility locus. I discovered six novel variants but also discovered that there is an absence of rare variants with MAFs between 0.06% and 2.5%, suggesting that *GDF5* has been subjected to a bottleneck in the past. I then carried out functional studies on a rare variant (-41 bp) I found in the proximal promoter of *GDF5*. I found that the variant A-allele of this -41 bp locus is able to increase the expression of *GDF5* to such an extent that it counteracts the effect of the OA associated T-allele of rs143383. I established that this is due to the trans-acting factor YY1 having a greater binding affinity to this variant allele than that of the wildtype C-allele. These findings expose a site for intervention and modulation of *GDF5* expression to overcome a disease susceptibility allele. In terms of the missing heritability of OA, this thesis has been unable to identify any loci that are contributing to the heritability of the disease on *the population level*. The discovery of the -41 bp variant, and the identification of a *trans*-acting factor which binds differentially to these two alleles may well be having an effect on the heritability of OA on a *personal and familial level*.

In terms of genetic signals coming from GWASs and other genetic studies, with regards to OA there has been small effect sizes yielded. Some argue that this could be due to low statistical power of these studies, and by greatly increasing the power then whole pathways could be identified. In age-related macular degeneration five loci have been identified which accounts for 50% of the heritability of the disease, whereas research into Crohn’s disease has accounted for approximately 20% of the heritability from 32 loci. Genetic research into OA on the other hand has revealed only a handful of loci – such as GDF5, DIO2 and the 7q22 locus – all with small effect sizes. We see a discordance of 40% for OA between MZ twins, as well as many other diseases and these discordant MZ twins have been shown to have differences in their DNA methylation profiles (155). MMPs – shown to be elevated in expression in OA cartilage – have also found to have hypomethelated promoters which could be the cause of the disease associated alteration in expression (156). Evidence such as this indicates that research into OA should not just be confined to common variants with small effect sizes; we must look for rarer variants which may have larger effect sizes and for epigenetic differences.

Following this research I would suggest deep-sequencing of more OA susceptibility loci for rare variants. It is likely that my failure to record any rare variants with MAFs capable of imparting an effect at a population level is due to the hypothesised bottleneck that *GDF5* has been subjected to due to a positive selection pressure on the decreased height associated alleles of rs143383 and rs143384 (138). Suitable candidates for deep-sequencing would be *DIO2* which has been shown to exhibit AEI (67) or genes implicated in OA susceptibility from the publication of the final arcOGEN GWAS results such as *CHRST11* (80).

Should I carry out any of these mentioned deep-sequencing investigations I would not use the Sanger Sequencing method used in this thesis again. Sequencing technologies are advancing all the time as well as becoming progressively cheaper. From Sanger Sequencing there are now 'next-generation' sequencing methods, and even 'next-next-generation' sequencing methods. Of the next-generation methods three methods have become widely used: Pyrosequencing, Illumina dye sequencing and SOLiD sequencing by Roche, Illumina and Life Technologies (formerly Applied Biosystems) respectively.

Pyrosequencing uses single stranded template DNA along with DNA polymerases and enzymes including luciferase which fluoresce when a specific nucleotide base is incorporated into the synthesising strand, one of the four bases are added in a specific order during each round of elongation. Illumina dye sequencing involves the addition of adapter sequences being added to the ends of double stranded target DNA. These adapters then bind to adapters on a slide and the DNA is replicated to produce local colonies of the same sequence. The dNTPs are added and unincorporated bases washed off. A laser is applied to the slide and the fluorescence given off corresponds to the base. However, unlike Sanger Sequencing the application of the laser removes the 3′ terminator, allowing for the next round of base extensions to occur on the same molecule. This cycles through until the whole molecule is sequenced. The advantage of Illumina over other methods is that this method just uses a DNA polymerase and fluorescent bases, rather than a multitude of enzymes. SOLiD sequencing requires the template DNA to be added to an adapter on an emulsion bead, then a primer of a specific length (*n*) binds to the adapter sequence, a ligase then adds fluorescent interrogation probes which contain two bases complementary to the target sequence plus multi-nonspecific binding sequence. The probe is then cleaved to allow for identification of the first base of the probe. Ligation and then cleavage of a new probe occurs and the sequence read grows until the end of the template is reached. Following this, denaturation occurs from the template and a new primer of length *n*-1 is added and the whole cycle is repeated. This results in each base of the template strand being read twice, increasing read accuracy. Should I repeat the sequencing project on other genes I would use Pyrosequencing by Roche as it would allow long individual read lengths but many, many more reads per run, enabling me to go deeper in re-sequencing, resulting in the ability to detect even rarer variants.

Some large scale sequencing projects include ENCODE, the encyclopaedia of DNA elements which aims to identify and sequence all functional elements within the human genome and the 1 000 genomes project, which aimed to sequence the entire genomes of over 1 000 people from across the world, highlighting population specific differences and giving allele frequencies. All this data has since been added into the HapMap database. While my study was limited to only one gene rather than a whole genome, I sequenced with a greater depth than the 1 000 genome project has done for *GDF5*, as I sequenced over 1 800 chromosomes at this locus. Also my study was solely on Europeans, whereas the 1 000 genomes project covers multiple populations, therefore reducing the population specific depth in which it goes to. However, as I found out there are only very, very rare variants in GDF5 with an absence of rare variants. Had the 1 000 genomes project data for *GDF5* been available at the start of my study, with sufficient depth in the European population, one would not have chosen *GDF5* as the OA susceptibility locus to resequence as it would have been apparent that there were no rare mutations that still had a frequency high enough to have an effect at the population level. Nonetheless, by carrying out my study I identified a site in the promoter of *GDF5* that allows for the modulation of a trans-acting factor's binding affinity, and thus the regulation of the gene's expression.

The HapMap database was invaluable for selecting candidate proxy SNPs to use for my AEI studies. HapMap provides invaluable allele frequencies and LD values between SNPs.

Further research into *GDF5* should be concentrated upon common polymorphisms within the gene. Egli and colleagues found that a SNP in the 3' UTR of *GDF5*, rs56366915, is able to modulate gene expression, with the T-allele exhibiting lower expression than the C-allele, an effect which is independent of the rs143383-rs143384 haplotype (54). At present no cause for this AEI has been found. It is possible that a microRNA is able to bind, or bind more strongly, to the T-allele of rs56366915 than it can to the C-allele, and thus reduces the expression of the T-allele. Another possibility for the AEI observed at rs56366915 is that the SNP is modulating the stability of the GDF5 mRNA transcript. An investigation into the half-life of GDF5 mRNA transcripts containing the two alleles of rs56366915 would be interesting.

As mentioned earlier, in order to elucidate the cause of the arcOGEN association signal at rs2615977 one could extract mesenchymal stem cells from the bone marrow of patients undergoing total joint replacement and culture these cells *in vitro* while differentiating them down the chondrocyte lineage. During this differentiation, as the cells undergo phenotypic change, RNA could be extracted and tested for AEI.

Chapter 6 Appendix

6.1 Conference presentations

Oral presentation at OARSI 2010 World Congress **“ALLELIC EXPRESSION ANALYSIS OF THE GENES WITHIN THE CHROMOSOME 7Q22 OA SUSCEPTIBILITY LOCUS REVEALS EVIDENCE FOR FUNCTIONAL POLYMORPHISM IN COG5”**
A. W. Doddd, N. Wreglesworth, E. V. Raine, A. Gravani, J. Loughlin;
Newcastle Univ., Newcastle upon Tyne, UNITED KINGDOM.

Poster presentation at OARSI 2010 World Congress **“DEEP SEQUENCING OF GDF5 IN OVER 1900 OSTEOARTHRITIS CASES AND CONTROLS REVEALS NOVEL AND POTENTIALLY FUNCTIONAL RARE VARIANTS IN THE PROTEIN CODING AND PROMOTER REGIONS OF THE GENE”**
A. W. Dodd[1], C. Rodriguez-Fontenla[2], M. Calaza[2], A. Carr[3], J. J.Gomez-Reino[2], A. Tsezou[4], A. Gonzalez[2], J. Loughlin[1]; [1]Newcastle Univ., Newcastle upon Tyne, UNITED KINGDOM, [2]Univ.rio de Santiago, Santiago de Compostela, SPAIN, [3]Univ. of Oxford, Oxford, UNITED KINGDOM, [4]Univ. of Thessaly, Larissa, GREECE.

Oral presentation at OARSI 2012 World Congress **“ALLELIC EXPRESSION ANALYSIS IN PATIENT TISSUES OF THE OSTEOARTHRITIS AND OF THE LUMBAR DISC HERNIATION SUSCEPTIBILITY LOCI THAT MAP TO COL11A1”**
A. W. Dodd, E. V. Raine, L. N. Reynard, J. Loughlin; Newcastle Univ., Newcastle upon Tyne, UNITED KINGDOM.

Poster presentation at OARSI 2012 World Congress **“A FUNCTIONAL INVESTIGATION OF A RARE VARIANT IN THE PROMOTER OF THE OA-ASSOCIATED LOCUS GDF5 AND THE DISCOVERY OF A TRANS-ACTING FACTOR INTERACTING WITH THE VARIANT SITE”**
A. W. Dodd, J. Loughlin; Newcastle Univ., Newcastle upon Tyne, UNITED KINGDOM.

6.2 Publications

Riancho JA, García-Ibarbia C, Gravani A, Raine EV, Rodríguez-Fontenla C, Soto-Hermida A, Rego-Perez I, **Dodd AW**, Gómez-Reino JJ, Zarrabeitia MT, Garcés CM, Carr A, Blanco F, González A, Loughlin J. (2010)

Common variations in estrogen-related genes are associated with severe large-joint osteoarthritis: a multicenter genetic and functional study. *Osteoarthritis Cartilage.* **18** 927-933.

Dodd AW*, Rodriguez-Fontenla C*, Calaza M, Carr A, Gomez-Reino JJ, Tsezou A, Reynard LN, Gonzalez A, Loughlin J. (2011)

Deep sequencing of GDF5 reveals the absence of rare variants at this important osteoarthritis susceptibility locus. *Osteoarthritis Cartilage.* **19** 430-434

Raine EV, Wreglesworth N, **Dodd AW**, Reynard LN, Loughlin J. (2012)

Gene expression analysis reveals HBP1 as a key target for the osteoarthritis susceptibility locus that maps to chromosome 7q22. *Ann Rheum Dis.* **71** 2020-2027

Dodd AW, Syddall CM, Loughlin J. (2012)

A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression. *Eur J Hum Genet.* **5** 517-521

Raine EVA*, **Dodd AW***, Reynard LN, Loughlin J. (2013)

Allelic expression analysis of COL11A1 in human joint tissues and its correlation with osteoarthritis susceptibility. *BMC Musculoskelet Disord.* **14**:85

Chapter 7: Acknowledgements

I would like to acknowledge the support, assistance and friendship of the Osteoarthritis Genetics Group which includes or has included; Miss. Emma Raine, Miss. Catherine Syddall, Miss. Athanasia Gravani, Dr. Louise Reynard, Miss. Madhushika Ratnayake, Dr. Nick Wreglesworth, Dr. Julie Colclough and most importantly my supervisor, Professor John Loughlin, who has guided and developed me throughout my studies and also given me the opportunity to present my research at two international conferences. I would also like to extend my acknowledgements to the Musculoskeletal Research Group as a whole – past and present, scientific and administrative – for all the expertise I have drawn upon throughout my time in the department but would like to specifically acknowledge the help of Dr. Matt Barter for help with countless techniques, Dr. Catherine Bui and Mr. Steven Woods for help with transfections, Mr. Tim Robson, Mr. Tim Williamson, Mrs. Debbie Jones, Miss. Jamie Watson and Miss. Julie Diboll for all their help and advice as laboratory technicians. I wish to thank Dr. Rick Lewis for his help in modelling GDF5. I am grateful to the Newcastle and Oxfordshire NHS trusts and for the collection of patient material. I am also grateful to the William Harker Foundation for their support during the course of my degree.

Chapter 8: References

1. Wieland, H. a, M. Michaelis, B. J. Kirschbaum, and K. a Rudolphi. 2005. Osteoarthritis - an untreatable disease? *Nature reviews. Drug discovery* 4: 331-44.

2. Buckwalter, J. A., C. Saltzman, and T. Brown. 2004. The impact of osteoarthritis: implications for research. *Clinical orthopaedics and related research* S6-15.

3. Hunter, D. J., and D. T. Felson. 2006. Osteoarthritis. *BMJ (Clinical research ed.)* 332: 639-42.

4. Loeser, R. F., S. R. Goldring, C. R. Scanzello, and M. B. Goldring. 2012. Osteoarthritis: a disease of the joint as an organ. *Arthritis and rheumatism* 64: 1697-707.

5. Bennell, K. L., D. J. Hunter, and R. S. Hinman. 2012. Management of osteoarthritis of the knee. *Bmj* 345: e4934-e4934.

6. Kellgren, J. H., and J. S. Lawrence. 1957. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases* 16: 494-502.

7. Archer, C. W., G. P. Dowthwaite, and P. Francis-West. 2003. Development of synovial joints. *Birth defects research. Part C, Embryo today : reviews* 69: 144-55.

8. Salter, D. M. 1998. The tissues we deal with (II) Cartilage. *Current Orthopaedics* 12: 251-257.

9. Loughlin, J. 2005. The genetic epidemiology of human primary osteoarthritis: current status. *Expert reviews in molecular medicine* 7: 1-12.

10. Buckwalter, J. A., and H. J. Mankin. 1998. Articular cartilage: tissue design and chondrocyte-matrix interactions. *Instructional course lectures* 47: 477-86.

11. Peach, C. A., A. J. Carr, and J. Loughlin. 2005. Recent advances in the genetic investigation of osteoarthritis. *Trends in molecular medicine* 11: 186-91.

12. Aigner, T., a Sachse, P. M. Gebhard, and H. I. Roach. 2006. Osteoarthritis: pathobiology-targets and ways for therapeutic intervention. *Advanced drug delivery reviews* 58: 128-49.

13. Jones, A. R. C., and C. R. Flannery. 2007. Bioregulation of lubricin expression by growth factors and cytokines. *European cells & materials* 13: 40-5; discussion 45.

14. Othoteers.org. 2012. Zones of Articular Cartilage. *http://www.orthoteers.org/images/uploaded/Images3/ArtCart.jpg* .

15. Mackie, E. J., L. Tatarczuch, and M. Mirams. 2011. The skeleton: a multi-functional complex organ: the growth plate chondrocyte and endochondral ossification. *The Journal of endocrinology* 211: 109-21.

16. Buxton, P., C. Edwards, C. W. Archer, and P. Francis-West. 2001. Growth/differentiation factor-5 (GDF-5) and skeletal development. *The Journal of bone and joint surgery. American volume* 83-A Suppl: S23-30.

17. Storm, E. E., and D. M. Kingsley. 1999. GDF5 coordinates bone and joint formation during digit development. *Developmental biology* 209: 11-27.

18. Pacifici, M., E. Koyama, and M. Iwamoto. 2005. Mechanisms of synovial joint and articular cartilage formation: recent advances, but many lingering mysteries. *Birth defects research. Part C, Embryo today : reviews* 75: 237-48.

19. Lee, Y. C., B. Lu, J. M. Bathon, J. A. Haythornthwaite, M. T. Smith, G. G. Page, and R. R. Edwards. 2011. Pain sensitivity and pain reactivity in osteoarthritis. *Arthritis care & research* 63: 320-7.

20. Lindblad, S., and E. Hedfors. 1987. Arthroscopic and immunohistologic characterization of knee joint synovitis in osteoarthritis. *Arthritis and rheumatism* 30: 1081-8.

21. Oehler, S., D. Neureiter, C. Meyer-Scholten, and T. Aigner. 2002. Subtyping of osteoarthritic synoviopathy. *Clinical and experimental rheumatology* 20: 633-40.

22. Felson, D. T., and T. Neogi. 2004. Osteoarthritis: is it a disease of cartilage or of bone? *Arthritis and rheumatism* 50: 341-4.

23. Mankin, H. J., and L. Lippiello. 1970. Biochemical and metabolic abnormalities in articular cartilage from osteo-arthritic human hips. *The Journal of bone and joint surgery. American volume* 52: 424-34.

24. Vornehm, S. I., J. Dudhia, K. Von der Mark, and T. Aigner. 1996. Expression of collagen types IX and XI and other major cartilage matrix components by human fetal chondrocytes in vivo. *Matrix biology : journal of the International Society for Matrix Biology* 15: 91-8.

25. Aigner, T., S. I. Vornehm, G. Zeiler, J. Dudhia, K. von der Mark, and M. T. Bayliss. 1997. Suppression of cartilage matrix gene expression in upper zone chondrocytes of osteoarthritic cartilage. *Arthritis and rheumatism* 40: 562-9.

26. Roach, H. I., N. Yamada, K. S. C. Cheung, S. Tilley, N. M. P. Clarke, R. O. C. Oreffo, S. Kokubun, and F. Bronner. 2005. Association between the abnormal expression of matrix-degrading enzymes by human osteoarthritic chondrocytes and demethylation of specific CpG sites in the promoter regions. *Arthritis and rheumatism* 52: 3110-24.

27. Martin, J. A., and J. A. Buckwalter. 2001. Telomere erosion and senescence in human articular cartilage chondrocytes. *The journals of gerontology. Series A, Biological sciences and medical sciences* 56: B172-9.

28. Loeser, R. 2010. Age-related changes in the musculoskeletal system and the development of osteoarthritis. *Clinics in geriatric medicine* 26: 371-386.

29. Dai, S.-M., Z.-Z. Shan, H. Nakamura, K. Masuko-Hongo, T. Kato, K. Nishioka, and K. Yudoh. 2006. Catabolic stress induces features of chondrocyte senescence through overexpression of caveolin 1: possible involvement of caveolin 1-induced down-regulation of articular chondrocytes in the pathogenesis of osteoarthritis. *Arthritis and rheumatism* 54: 818-31.

30. Itahana, K., J. Campisi, and G. P. Dimri. 2004. Mechanisms of cellular senescence in human and mouse cells. *Biogerontology* 5: 1-10.

31. Campisi, J., and F. d' Adda di Fagagna. 2007. Cellular senescence: when bad things happen to good cells. *Nature reviews. Molecular cell biology* 8: 729-40.

32. Hiran, T. S., P. J. Moulton, and J. T. Hancock. 1997. Detection of superoxide and NADPH oxidase in porcine articular chondrocytes. *Free radical biology & medicine* 23: 736-43.

33. Tiku, M. L., R. Shah, and G. T. Allison. 2000. Evidence linking chondrocyte lipid peroxidation to cartilage matrix protein degradation. Possible role in cartilage aging and the pathogenesis of osteoarthritis. *The Journal of biological chemistry* 275: 20069-76.

34. Jallali, N., H. Ridha, C. Thrasivoulou, C. Underwood, P. E. M. Butler, and T. Cowen. 2005. Vulnerability to ROS-induced cell death in ageing articular cartilage: the role of antioxidant enzyme activity. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 13: 614-22.

35. Aigner, T., K. Fundel, J. Saas, P. M. Gebhard, J. Haag, T. Weiss, A. Zien, F. Obermayr, R. Zimmer, and E. Bartnik. 2006. Large-scale gene expression profiling reveals major pathogenetic pathways of cartilage degeneration in osteoarthritis. *Arthritis and rheumatism* 54: 3533-44.

36. Gavriilidis, C., S. Miwa, T. von Zglinicki, R. W. Taylor, and D. A. Young. 2013. Mitochondrial dysfunction in osteoarthritis is associated with down-regulation of superoxide dismutase 2. *Arthritis and rheumatism* 65: 378-87.

37. Yudoh, K., van T. Nguyen, H. Nakamura, K. Hongo-Masuko, T. Kato, and K. Nishioka. 2005. Potential involvement of oxidative stress in cartilage senescence and development of osteoarthritis: oxidative stress induces chondrocyte telomere instability and downregulation of chondrocyte function. *Arthritis research & therapy* 7: R380-91.

38. Dieppe, P. A., and L. S. Lohmander. 2005. Pathogenesis and management of pain in osteoarthritis. *Lancet* 365: 965-73.

39. Stecher, R. M., and A. H. Hersh. 1944. HEBERDEN'S NODES: THE MECHANISM OF INHERITANCE IN HYPERTROPHIC ARTHRITIS OF THE FINGERS. *The Journal of clinical investigation* 23: 699-704.

40. Spector, T. D., and A. J. MacGregor. 2004. Risk factors for osteoarthritis: genetics. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 12 Suppl A: S39-44.

41. Spector, T. D., F. Cicuttini, J. Baker, J. Loughlin, and D. Hart. 1996. Genetic influences on osteoarthritis in women: a twin study. *BMJ (Clinical research ed.)* 312: 940-3.

42. MacGregor, A. J., L. Antoniades, M. Matson, T. Andrew, and T. D. Spector. 2000. The genetic contribution to radiographic hip osteoarthritis in women: results of a classic twin study. *Arthritis and rheumatism* 43: 2410-6.

43. Sambrook, P. N., A. J. MacGregor, and T. D. Spector. 1999. Genetic influences on cervical and lumbar disc degeneration: a magnetic resonance imaging study in twins. *Arthritis and rheumatism* 42: 366-72.

44. MacGregor, A. J., Q. Li, T. D. Spector, and F. M. K. Williams. 2009. The genetic influence on radiographic osteoarthritis is site specific at the hand, hip and knee. *Rheumatology (Oxford, England)* 48: 277-80.

45. Ballantyne, P. J., M. a M. Gignac, and G. a Hawker. 2007. A patient-centered perspective on surgery avoidance for hip or knee arthritis: lessons for the future. *Arthritis and rheumatism* 57: 27-34.

46. Kerkhof, H. J. M., I. Meulenbelt, T. Akune, N. K. Arden, a Aromaa, S. M. a Bierma-Zeinstra, a Carr, C. Cooper, J. Dai, M. Doherty, S. a Doherty, D. Felson, a Gonzalez, a Gordon, a Harilainen, D. J. Hart, V. B. Hauksson, M. Heliovaara, a Hofman, S. Ikegawa, T. Ingvarsson, Q. Jiang, H. Jonsson, I. Jonsdottir, H. Kawaguchi, M. Kloppenburg, U. M. Kujala, N. E. Lane, P. Leino-Arjas, L. S. Lohmander, F. P. Luyten, K. N. Malizos, M. Nakajima, M. C. Nevitt, H. a P. Pols, F. Rivadeneira, D. Shi, E. Slagboom, T. D. Spector, K. Stefansson, a Sudo, a Tamm, a E. Tamm, a Tsezou, a Uchida, a G. Uitterlinden, J. M. Wilkinson, N. Yoshimura, a M. Valdes, and J. B. J. van Meurs. 2011. Recommendations for standardization and phenotype definitions in genetic studies of osteoarthritis: the TREAT-OA consortium. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 19: 254-64.

47. Bos, S. D., P. E. Slagboom, and I. Meulenbelt. 2008. New insights into osteoarthritis: early developmental features of an ageing-related disease. *Current opinion in rheumatology* 20: 553-9.

48. Meulenbelt, I. 2012. Osteoarthritis year 2011 in review: genetics. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 20: 218-22.

49. Loughlin, J. 2003. Genetics of osteoarthritis and potential for drug development. *Current opinion in pharmacology* 3: 295-9.

50. Reginato, A. M., and B. R. Olsen. 2002. The role of structural genes in the pathogenesis of osteoarthritic disorders. *Arthritis research* 4: 337-45.

51. Kizawa, H., I. Kou, A. Iida, A. Sudo, Y. Miyamoto, A. Fukuda, A. Mabuchi, A. Kotani, A. Kawakami, S. Yamamoto, A. Uchida, K. Nakamura, K. Notoya, Y. Nakamura, and S. Ikegawa. 2005. An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nature genetics* 37: 138-44.

52. Miyamoto, Y., A. Mabuchi, D. Shi, T. Kubo, Y. Takatori, S. Saito, M. Fujioka, A. Sudo, A. Uchida, S. Yamamoto, K. Ozaki, M. Takigawa, T. Tanaka, Y. Nakamura, Q. Jiang, and S. Ikegawa. 2007. A functional polymorphism in the 5′ UTR of GDF5 is associated with susceptibility to osteoarthritis. *Nature genetics* 39: 529-33.

53. Southam, L., J. Rodriguez-Lopez, J. M. Wilkins, M. Pombo-Suarez, S. Snelling, J. J. Gomez-Reino, K. Chapman, A. Gonzalez, and J. Loughlin. 2007. An SNP in the 5′-UTR of GDF5 is associated with osteoarthritis susceptibility in Europeans and with in vivo differences in allelic expression in articular cartilage. *Human molecular genetics* 16: 2226-32.

54. Egli, R. J., L. Southam, J. M. Wilkins, I. Lorenzen, M. Pombo-Suarez, A. Gonzalez, A. Carr, K. Chapman, and J. Loughlin. 2009. Functional analysis of the osteoarthritis susceptibility-associated GDF5 regulatory polymorphism. *Arthritis and rheumatism* 60: 2055-64.

55. Chapman, K., A. Takahashi, I. Meulenbelt, C. Watson, J. Rodriguez-Lopez, R. Egli, A. Tsezou, K. N. Malizos, M. Kloppenburg, D. Shi, L. Southam, R. van der Breggen, R. Donn, J. Qin, M. Doherty, P. E. Slagboom, G. Wallis, N. Kamatani, Q. Jiang, A. Gonzalez, J. Loughlin, and S. Ikegawa. 2008. A meta-analysis of European and Asian cohorts reveals a global role of a

functional SNP in the 5' UTR of GDF5 with osteoarthritis susceptibility. *Human molecular genetics* 17: 1497-504.

56. Evangelou, E., K. Chapman, I. Meulenbelt, F. B. Karassa, J. Loughlin, A. Carr, M. Doherty, S. Doherty, J. J. Gómez-Reino, A. Gonzalez, B. V. Halldorsson, V. B. Hauksson, A. Hofman, D. J. Hart, S. Ikegawa, T. Ingvarsson, Q. Jiang, I. Jonsdottir, H. Jonsson, H. J. M. Kerkhof, M. Kloppenburg, N. E. Lane, J. Li, R. J. Lories, J. B. J. van Meurs, A. Näkki, M. C. Nevitt, J. Rodriguez-Lopez, D. Shi, P. E. Slagboom, K. Stefansson, A. Tsezou, G. A. Wallis, C. M. Watson, T. D. Spector, A. G. Uitterlinden, A. M. Valdes, and J. P. A. Ioannidis. 2009. Large-scale analysis of association between GDF5 and FRZB variants and osteoarthritis of the hip, knee, and hand. *Arthritis and rheumatism* 60: 1710-21.

57. Valdes, A. M., E. Evangelou, H. J. M. Kerkhof, A. Tamm, S. A. Doherty, K. Kisand, A. Tamm, I. Kerna, A. Uitterlinden, A. Hofman, F. Rivadeneira, C. Cooper, E. M. Dennison, W. Zhang, K. R. Muir, J. P. A. Ioannidis, M. Wheeler, R. A. Maciewicz, J. B. van Meurs, N. K. Arden, T. D. Spector, and M. Doherty. 2011. The GDF5 rs143383 polymorphism is associated with osteoarthritis of the knee with genome-wide statistical significance. *Annals of the rheumatic diseases* 70: 873-5.

58. Simonet, W. S. 2002. Genetics of primary generalized osteoarthritis. *Molecular genetics and metabolism* 77: 31-4.

59. Chapman, K., Z. Mustafa, C. Irven, A. J. Carr, K. Clipsham, A. Smith, J. Chitnavis, J. S. Sinsheimer, V. A. Bloomfield, M. McCartney, O. Cox, L. R. Cardon, B. Sykes, and J. Loughlin. 1999. Osteoarthritis-susceptibility locus on chromosome 11q, detected by linkage. *American journal of human genetics* 65: 167-74.

60. Loughlin, J., B. Dowling, K. Chapman, L. Marcelline, Z. Mustafa, L. Southam, A. Ferreira, C. Ciesielski, D. A. Carson, and M. Corr. 2004. Functional variants within the secreted frizzled-related protein 3 gene are associated with hip osteoarthritis in females. *Proceedings of the National Academy of Sciences of the United States of America* 101: 9757-62.

61. Hoang, B., M. Moos, S. Vukicevic, and F. P. Luyten. 1996. Primary structure and tissue distribution of FRZB, a novel protein related to Drosophila frizzled, suggest a role in skeletal morphogenesis. *The Journal of biological chemistry* 271: 26131-7.

62. Valdes, A. M., J. Loughlin, M. V. Oene, K. Chapman, G. L. Surdulescu, M. Doherty, and T. D. Spector. 2007. Sex and ethnic differences in the association of ASPN, CALM1, COL2A1, COMP, and FRZB with genetic susceptibility to osteoarthritis of the knee. *Arthritis and rheumatism* 56: 137-46.

63. Snelling, S., a Ferreira, and J. Loughlin. 2007. Allelic expression analysis suggests that cis-acting polymorphism of FRZB expression does not contribute to osteoarthritis susceptibility. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 15: 90-2.

64. Meulenbelt, I., A. B. Seymour, M. Nieuwland, T. W. J. Huizinga, C. M. van Duijn, and P. E. Slagboom. 2004. Association of the interleukin-1 gene cluster with radiographic signs of osteoarthritis of the hip. *Arthritis and rheumatism* 50: 1179-86.

65. Baker-Lepain, J. C., J. A. Lynch, N. Parimi, C. E. McCulloch, M. C. Nevitt, M. Corr, and N. E. Lane. 2012. Variant alleles of the Wnt antagonist FRZB are determinants of hip shape and modify the relationship between hip shape and osteoarthritis. *Arthritis and rheumatism* 64: 1457-65.

66. Meulenbelt, I., J. L. Min, S. Bos, N. Riyazi, J. J. Houwing-Duistermaat, H.-J. van der Wijk, H. M. Kroon, M. Nakajima, S. Ikegawa, A. G. Uitterlinden, J. B. J. van Meurs, W. M. van der Deure, T. J. Visser, A. B. Seymour, N. Lakenberg, R. van der Breggen, D. Kremer, C. M. van Duijn, M. Kloppenburg, J. Loughlin, and P. E. Slagboom. 2008. Identification of DIO2 as a new susceptibility locus for symptomatic osteoarthritis. *Human molecular genetics* 17: 1867-75.

67. Bos, S. D., J. V. M. G. Bovée, B. J. Duijnisveld, E. V. A. Raine, W. J. van Dalen, Y. F. M. Ramos, R. van der Breggen, R. G. H. H. Nelissen, P. E. Slagboom, J. Loughlin, and I. Meulenbelt. 2012. Increased type II deiodinase protein in OA-affected cartilage and allelic imbalance of OA risk polymorphism rs225014 at DIO2 in human OA joint tissues. *Annals of the rheumatic diseases* 71: 1254-8.

68. The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437: 1299-320.

69. Polychronakos, C., and Q. Li. 2011. Understanding type 1 diabetes through genetics: advances and prospects. *Nature reviews. Genetics* 12: 781-92.

70. The Wellcome Trust Case Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678.

71. Spector, T. D., R. H. Reneland, S. Mah, A. M. Valdes, D. J. Hart, S. Kammerer, M. Langdown, C. R. Hoyal, J. Atienza, M. Doherty, P. Rahman, M. R. Nelson, and A. Braun. 2006. Association between a variation in LRCH1 and knee osteoarthritis: a genome-wide single-nucleotide polymorphism association study using DNA pooling. *Arthritis and rheumatism* 54: 524-32.

72. Snelling, S., J. S. Sinsheimer, A. Carr, and J. Loughlin. 2007. Genetic association analysis of LRCH1 as an osteoarthritis susceptibility locus. *Rheumatology (Oxford, England)* 46: 250-2.

73. Jiang, Q., D. Shi, M. Nakajima, J. Dai, J. Wei, K. N. Malizos, J. Qin, Y. Miyamoto, N. Kamatani, B. Liu, A. Tsezou, T. Nakamura, and S. Ikegawa. 2008. Lack of association of single nucleotide polymorphism in LRCH1 with knee osteoarthritis susceptibility. *Journal of human genetics* 53: 42-7.

74. Valdes, A. M., J. Loughlin, K. M. Timms, J. J. B. van Meurs, L. Southam, S. G. Wilson, S. Doherty, R. J. Lories, F. P. Luyten, A. Gutin, V. Abkevich, D. Ge, A. Hofman, A. G. Uitterlinden, D. J. Hart, F. Zhang, G. Zhai, R. J. Egli, M. Doherty, J. Lanchbury, and T. D. Spector. 2008. Genome-wide association scan identifies a prostaglandin-endoperoxide synthase 2 variant involved in risk of knee osteoarthritis. *American journal of human genetics* 82: 1231-40.

75. Kerkhof, H. J. M., R. J. Lories, I. Meulenbelt, I. Jonsdottir, A. M. Valdes, P. Arp, T. Ingvarsson, M. Jhamai, H. Jonsson, L. Stolk, G. Thorleifsson, G. Zhai, F. Zhang, Y. Zhu, R. van der Breggen, A. Carr, M. Doherty, S. Doherty, D. T. Felson, A. Gonzalez, B. V. Halldorsson, D. J. Hart, V. B. Hauksson, A. Hofman, J. P. A. Ioannidis, M. Kloppenburg, N. E. Lane, J. Loughlin, F. P. Luyten, M. C. Nevitt, N. Parimi, H. A. P. Pols, F. Rivadeneira, E. P. Slagboom, U. Styrkársdóttir, A. Tsezou, T. van de Putte, J. Zmuda, T. D. Spector, K. Stefansson, A. G. Uitterlinden, and J. B. J. van Meurs. 2010. A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22. *Arthritis and rheumatism* 62: 499-510.

76. Evangelou, E., A. M. Valdes, H. J. M. Kerkhof, U. Styrkarsdottir, Y. Zhu, I. Meulenbelt, R. J. Lories, F. B. Karassa, P. Tylzanowski, S. D. Bos, T. Akune, N. K. Arden, A. Carr, K. Chapman, L. A. Cupples, J. Dai, P. Deloukas, M. Doherty, S. Doherty, G. Engstrom, A. Gonzalez, B. V.

Halldorsson, C. L. Hammond, D. J. Hart, H. Helgadottir, A. Hofman, S. Ikegawa, T. Ingvarsson, Q. Jiang, H. Jonsson, J. Kaprio, H. Kawaguchi, K. Kisand, M. Kloppenburg, U. M. Kujala, L. S. Lohmander, J. Loughlin, F. P. Luyten, A. Mabuchi, A. McCaskie, M. Nakajima, P. M. Nilsson, N. Nishida, W. E. R. Ollier, K. Panoutsopoulou, T. van de Putte, S. H. Ralston, F. Rivadeneira, J. Saarela, S. Schulte-Merker, D. Shi, P. E. Slagboom, A. Sudo, A. Tamm, A. Tamm, G. Thorleifsson, U. Thorsteinsdottir, A. Tsezou, G. a Wallis, J. M. Wilkinson, N. Yoshimura, E. Zeggini, G. Zhai, F. Zhang, I. Jonsdottir, A. G. Uitterlinden, D. T. Felson, J. B. van Meurs, K. Stefansson, J. P. a Ioannidis, and T. D. Spector. 2011. Meta-analysis of genome-wide association studies confirms a susceptibility locus for knee osteoarthritis on chromosome 7q22. *Annals of the rheumatic diseases* 70: 349-55.

77. Raine, E. V. A., N. Wreglesworth, A. W. Dodd, L. N. Reynard, and J. Loughlin. 2012. Gene expression analysis reveals HBP1 as a key target for the osteoarthritis susceptibility locus that maps to chromosome 7q22. *Annals of the rheumatic diseases* 6-15.

78. Sampson, E. M., Z. K. Haque, M. C. Ku, S. G. Tevosian, C. Albanese, R. G. Pestell, K. E. Paulson, and a S. Yee. 2001. Negative regulation of the Wnt-beta-catenin pathway by the transcriptional repressor HBP1. *The EMBO journal* 20: 4500-11.

79. Nakajima, M., A. Takahashi, I. Kou, C. Rodriguez-Fontenla, J. J. Gomez-Reino, T. Furuichi, J. Dai, A. Sudo, A. Uchida, N. Fukui, M. Kubo, N. Kamatani, T. Tsunoda, K. N. Malizos, A. Tsezou, A. Gonzalez, Y. Nakamura, and S. Ikegawa. 2010. New sequence variants in HLA class II/III region associated with susceptibility to knee osteoarthritis identified by genome-wide association study. *PloS one* 5: e9723.

80. arcOGEN Consortium and arcOGEN Collaborators. 2012. Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet* 380: 815-23.

81. TwinsUK. 2012. TwinsUK. *http://www.twinsuk.ac.uk/* .

82. Panoutsopoulou, K., L. Southam, K. S. Elliott, N. Wrayner, G. Zhai, C. Beazley, G. Thorleifsson, N. K. Arden, A. Carr, K. Chapman, P. Deloukas, M. Doherty, A. McCaskie, W. E. R. Ollier, S. H. Ralston, T. D. Spector, A. M. Valdes, G. A. Wallis, J. M. Wilkinson, E. Arden, K. Battley, H. Blackburn, F. J. Blanco, S. Bumpstead, L. A. Cupples, A. G. Day-Williams, K. Dixon, S. A. Doherty, T. Esko, E. Evangelou, D. Felson, J. J. Gomez-Reino, A. Gonzalez, A. Gordon, R. Gwilliam, B. V. Halldorsson, V. B. Hauksson, A. Hofman, S. E. Hunt, J. P. A. Ioannidis, T. Ingvarsson, I. Jonsdottir, H. Jonsson, R. Keen, H. J. M. Kerkhof, M. G. Kloppenburg, N. Koller, N. Lakenberg, N. E. Lane, A. T. Lee, A. Metspalu, I. Meulenbelt, M. C. Nevitt, F. O'Neill, N. Parimi, S. C. Potter, I. Rego-Perez, J. A. Riancho, K. Sherburn, P. E. Slagboom, K. Stefansson, U. Styrkarsdottir, M. Sumillera, D. Swift, U. Thorsteinsdottir, A. Tsezou, A. G. Uitterlinden, J. B. J. van Meurs, B. Watkins, M. Wheeler, S. Mitchell, Y. Zhu, J. M. Zmuda, E. Zeggini, and J. Loughlin. 2011. Insights into the genetic architecture of osteoarthritis from stage 1 of the arcOGEN study. *Annals of the rheumatic diseases* 70: 864-7.

83. Mio, F., K. Chiba, Y. Hirose, Y. Kawaguchi, Y. Mikami, T. Oya, M. Mori, M. Kamata, M. Matsumoto, K. Ozaki, T. Tanaka, A. Takahashi, T. Kubo, T. Kimura, Y. Toyama, and S. Ikegawa. 2007. A functional polymorphism in COL11A1, which encodes the alpha 1 chain of type XI collagen, is associated with susceptibility to lumbar disc herniation. *American journal of human genetics* 81: 1271-7.

84. Kannu, P., J. F. Bateman, D. Belluoccio, A. J. Fosang, and R. Savarirayan. 2009. Employing molecular genetics of chondrodysplasias to inform the study of osteoarthritis. *Arthritis and rheumatism* 60: 325-34.

85. Snead, M. P., and J. R. Yates. 1999. Clinical and Molecular genetics of Stickler syndrome. *Journal of medical genetics* 36: 353-9.

86. Wang, K., H. Zhang, S. Kugathasan, V. Annese, J. P. Bradfield, R. K. Russell, P. M. A. Sleiman, M. Imielinski, J. Glessner, C. Hou, D. C. Wilson, T. Walters, C. Kim, E. C. Frackelton, P. Lionetti, A. Barabino, J. Van Limbergen, S. Guthery, L. Denson, D. Piccoli, M. Li, M. Dubinsky, M. Silverberg, A. Griffiths, S. F. A. Grant, J. Satsangi, R. Baldassano, and H. Hakonarson. 2009. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *American journal of human genetics* 84: 399-405.

87. Fuss, I. J., C. Becker, Z. Yang, C. Groden, R. L. Hornung, F. Heller, M. F. Neurath, W. Strober, and P. J. Mannon. 2006. Both IL-12p70 and IL-23 are synthesized during active Crohn's disease and are down-regulated by treatment with anti-IL-12 p40 monoclonal antibody. *Inflammatory Bowel Diseases* 12: 9-15.

88. Rockman, M. V., and G. A. Wray. 2002. Abundant Raw Material for Cis-Regulatory Evolution in Humans. *Molecular Biology and Evolution* 19: 1991-2004.

89. Kleinjan, D. A., and V. van Heyningen. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *American journal of human genetics* 76: 8-32.

90. Borel, P. 2012. Genetic variations involved in interindividual variability in carotenoid status. *Molecular nutrition & food research* 56: 228-40.

91. Fu, J., M. G. M. Wolfs, P. Deelen, H.-J. Westra, R. S. N. Fehrmann, G. J. Te Meerman, W. A. Buurman, S. S. M. Rensen, H. J. M. Groen, R. K. Weersma, L. H. van den Berg, J. Veldink, R. A. Ophoff, H. Snieder, D. van Heel, R. C. Jansen, M. H. Hofker, C. Wijmenga, and L. Franke. 2012. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS genetics* 8: e1002431.

92. Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature genetics* 44: 502-10.

93. Wilkins, J. M., L. Southam, A. J. Price, Z. Mustafa, A. Carr, and J. Loughlin. 2007. Extreme context specificity in differential allelic expression. *Human molecular genetics* 16: 537-46.

94. Valdes, A. M., and T. D. Spector. 2008. The contribution of genes to osteoarthritis. *Rheumatic diseases clinics of North America* 34: 581-603.

95. Visscher, P. M. 2008. Sizing up human height variation. *Nature genetics* 40: 489-90.

96. Manolio, T., F. Collins, and N. Cox. 2009. Finding the missing heritability of complex diseases. *Nature* 461: 747-753.

97. Li, B., and S. M. Leal. 2009. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS genetics* 5: e1000481.

98. Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. 2008. Shifting Paradigm of Association Studies : Value of Rare Single-Nucleotide Polymorphisms. *Journal of Human Genetics* 100-112.

99. Bodmer, W., and C. Bonilla. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics* 40: 695-701.

100. Plöger, F., P. Seemann, M. Schmidt-von Kegler, K. Lehmann, J. Seidel, K. W. Kjaer, J. Pohl, and S. Mundlos. 2008. Brachydactyly type A2 associated with a defect in proGDF5 processing. *Human molecular genetics* 17: 1222-33.

101. Wolfman, N. M., G. Hattersley, K. Cox, A. J. Celeste, R. Nelson, N. Yamaji, J. L. Dube, E. DiBlasio-Smith, J. Nove, J. J. Song, J. M. Wozney, and V. Rosen. 1997. Ectopic induction of tendon and ligament in rats by growth and differentiation factors 5, 6, and 7, members of the TGF-beta gene family. *The Journal of clinical investigation* 100: 321-30.

102. Edwards, C. J., and P. H. Francis-West. 2001. Bone morphogenetic proteins in the development and healing of synovial joints. *Seminars in arthritis and rheumatism* 31: 33-42.

103. Kotzsch, A., J. Nickel, A. Seher, W. Sebald, and T. D. Müller. 2009. Crystal structure analysis reveals a spring-loaded latch as molecular mechanism for GDF-5-type I receptor specificity. *The EMBO journal* 28: 937-47.

104. Rountree, R. B., M. Schoor, H. Chen, M. E. Marks, V. Harley, Y. Mishina, and D. M. Kingsley. 2004. BMP receptor signaling is required for postnatal maintenance of articular cartilage. *PLoS biology* 2: e355.

105. Storm, E. E., and D. M. Kingsley. 1996. Joint patterning defects caused by single and double mutations in members of the bone morphogenetic protein (BMP) family. *Development* 122: 3969-79.

106. Mishina, Y., A. Suzuki, N. Ueno, and R. R. Behringer. 1995. Bmpr encodes a type I bone morphogenetic protein receptor that is essential for gastrulation during mouse embryogenesis. *Genes & development* 9: 3027-37.

107. Nickel, J., A. Kotzsch, W. Sebald, and T. D. Mueller. 2005. A single residue of GDF-5 defines binding specificity to BMP receptor IB. *Journal of molecular biology* 349: 933-47.

108. Tina V. Hellmann, J. N. and T. D. M. 2012. *Mutations in Human Genetic Disease*, (D. Cooper, ed). InTech.

109. Cornelis, F. M. F., F. P. Luyten, and R. J. Lories. 2011. Functional effects of susceptibility genes in osteoarthritis. *Discovery medicine* 12: 129-39.

110. Storm, E. E., T. V. Huynh, N. G. Copeland, N. A. Jenkins, D. M. Kingsley, and S. J. Lee. 1994. Limb alterations in brachypodism mice due to mutations in a new member of the TGF beta-superfamily. *Nature* 368: 639-43.

111. Daans, M., F. P. Luyten, and R. J. U. Lories. 2011. GDF5 deficiency in mice is associated with instability-driven joint damage, gait and subchondral bone changes. *Annals of the rheumatic diseases* 70: 208-13.

112. Mikic, B., B. J. Schalet, R. T. Clark, V. Gaschen, and E. B. Hunziker. 2001. GDF-5 deficiency in mice alters the ultrastructure, mechanical properties and composition of the Achilles tendon. *Journal of orthopaedic research : official publication of the Orthopaedic Research Society* 19: 365-71.

113. Loughlin, J. 2011. Genetic indicators and susceptibility to osteoarthritis. *British journal of sports medicine* 45: 278-82.

114. Gentili, C., and R. Cancedda. 2009. Cartilage and bone extracellular matrix. *Current pharmaceutical design* 15: 1334-48.

115. Kolácná, L., J. Bakesová, F. Varga, E. Kostáková, L. Plánka, A. Necas, D. Lukás, E. Amler, and V. Pelouch. 2007. Biochemical and biophysical aspects of collagen nanostructure in the extracellular matrix. *Physiological research / Academia Scientiarum Bohemoslovaca* 56 Suppl 1: S51-60.

116. Khoshnoodi, J., V. Pedchenko, and B. G. Hudson. 2008. Mammalian collagen IV. *Microscopy research and technique* 71: 357-70.

117. Kadler, K. E., D. F. Holmes, J. A. Trotter, and J. A. Chapman. 1996. Collagen fibril formation. *The Biochemical journal* 316 (Pt 1: 1-11.

118. Brown, R. J., C. Mallory, O. M. McDougal, and J. T. Oxford. 2011. Proteomic analysis of Col11a1-associated protein complexes. *Proteomics* 11: 4660-76.

119. Li, Y., D. A. Lacerda, M. L. Warman, D. R. Beier, H. Yoshioka, Y. Ninomiya, J. T. Oxford, N. P. Morris, K. Andrikopoulos, and F. Ramirez. 1995. A fibrillar collagen gene, Col11a1, is essential for skeletal morphogenesis. *Cell* 80: 423-30.

120. Xu, L., H. Peng, D. Wu, K. Hu, M. B. Goldring, B. R. Olsen, and Y. Li. 2005. Activation of the discoidin domain receptor 2 induces expression of matrix metalloproteinase 13 associated with osteoarthritis in mice. *The Journal of biological chemistry* 280: 548-55.

121. Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. 2009. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics* 10: 184-94.

122. Montgomery, S. B., and E. T. Dermitzakis. 2011. From expression QTLs to personalized transcriptomics. *Nature reviews. Genetics* 12: 277-82.

123. Musunuru, K., A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, X. Li, H. Li, N. Kuperwasser, V. M. Ruda, J. P. Pirruccello, B. Muchmore, L. Prokunina-Olsson, J. L. Hall, E. E. Schadt, C. R. Morales, S. Lund-Katz, M. C. Phillips, J. Wong, W. Cantley, T. Racie, K. G. Ejebe, M. Orho-Melander, O. Melander, V. Koteliansky, K. Fitzgerald, R. M. Krauss, C. A. Cowan, S. Kathiresan, and D. J. Rader. 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466: 714-9.

124. Johnson, A. D., Y. Zhang, A. C. Papp, J. K. Pinsonneault, J.-E. Lim, D. Saffen, Z. Dai, D. Wang, and W. Sadée. 2008. Polymorphisms affecting gene transcription and mRNA processing in pharmacogenetic candidate genes: detection through allelic expression imbalance in human target tissues. *Pharmacogenetics and genomics* 18: 781-91.

125. Houseley, J., and D. Tollervey. 2009. The many pathways of RNA degradation. *Cell* 136: 763-76.

126. Long, A. D., M. N. Grote, and C. H. Langley. 1997. Genetic analysis of complex diseases. *Science (New York, N.Y.)* 275: 1328; author reply 1329-30.

127. Scott, W. K., M. A. Pericak-Vance, and J. L. Haines. 1997. Genetic analysis of complex diseases. *Science (New York, N.Y.)* 275: 1327; author reply 1329-30.

128. Müller-Myhsok, B., and L. Abel. 1997. Genetic analysis of complex diseases. *Science (New York, N.Y.)* 275: 1328-9; author reply 1329-30.

129. Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)* 273: 1516-7.

130. Nejentsev, S., N. Walker, D. Riches, M. Egholm, and J. A. Todd. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.)* 324: 387-9.

131. Emsley, P., and K. Cowtan. 2004. Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography* 60: 2126-32.

132. Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, T. Ferreira, A. R. Wood, R. J. Weyant, A. V. Segrè, E. K. Speliotes, E. Wheeler, N. Soranzo, J.-H. Park, J. Yang, D. Gudbjartsson, N. L. Heard-Costa, J. C. Randall, L. Qi, A. Vernon Smith, R. Mägi, T. Pastinen, L. Liang, I. M. Heid, J. Luan, G. Thorleifsson, T. W. Winkler, M. E. Goddard, K. Sin Lo, C. Palmer, T. Workalemahu, Y. S. Aulchenko, A. Johansson, M. C. Zillikens, M. F. Feitosa, T. Esko, T. Johnson, S. Ketkar, P. Kraft, M. Mangino, I. Prokopenko, D. Absher, E. Albrecht, F. Ernst, N. L. Glazer, C. Hayward, J.-J. Hottenga, K. B. Jacobs, J. W. Knowles, Z. Kutalik, K. L. Monda, O. Polasek, M. Preuss, N. W. Rayner, N. R. Robertson, V. Steinthorsdottir, J. P. Tyrer, B. F. Voight, F. Wiklund, J. Xu, J. H. Zhao, D. R. Nyholt, N. Pellikka, M. Perola, J. R. B. Perry, I. Surakka, M.-L. Tammesoo, E. L. Altmaier, N. Amin, T. Aspelund, T. Bhangale, G. Boucher, D. I. Chasman, C. Chen, L. Coin, M. N. Cooper, A. L. Dixon, Q. Gibson, E. Grundberg, K. Hao, M. Juhani Junttila, L. M. Kaplan, J. Kettunen, I. R. König, T. Kwan, R. W. Lawrence, D. F. Levinson, M. Lorentzon, B. McKnight, A. P. Morris, M. Müller, J. Suh Ngwa, S. Purcell, S. Rafelt, R. M. Salem, E. Salvi, S. Sanna, J. Shi, U. Sovio, J. R. Thompson, M. C. Turchin, L. Vandenput, D. J. Verlaan, V. Vitart, C. C. White, A. Ziegler, P. Almgren, A. J. Balmforth, H. Campbell, L. Citterio, A. De Grandi, A. Dominiczak, J. Duan, P. Elliott, R. Elosua, J. G. Eriksson, N. B. Freimer, E. J. C. Geus, N. Glorioso, S. Haiqing, A.-L. Hartikainen, A. S. Havulinna, A. A. Hicks, J. Hui, W. Igl, T. Illig, A. Jula, E. Kajantie, T. O. Kilpeläinen, M. Koiranen, I. Kolcic, S. Koskinen, P. Kovacs, J. Laitinen, J. Liu, M.-L. Lokki, A. Marusic, A. Maschio, T. Meitinger, A. Mulas, G. Paré, A. N. Parker, J. F. Peden, A. Petersmann, I. Pichler, K. H. Pietiläinen, A. Pouta, M. Ridderstråle, J. I. Rotter, J. G. Sambrook, A. R. Sanders, C. O. Schmidt, J. Sinisalo, J. H. Smit, H. M. Stringham, G. Bragi Walters, E. Widen, S. H. Wild, G. Willemsen, L. Zagato, L. Zgaga, P. Zitting, H. Alavere, M. Farrall, W. L. McArdle, M. Nelis, M. J. Peters, S. Ripatti, J. B. J. van Meurs, K. K. Aben, K. G. Ardlie, J. S. Beckmann, J. P. Beilby, R. N. Bergman, S. Bergmann, F. S. Collins, D. Cusi, M. den Heijer, G. Eiriksdottir, P. V. Gejman, A. S. Hall, A. Hamsten, H. V. Huikuri, C. Iribarren, M. Kähönen, J. Kaprio, S. Kathiresan, L. Kiemeney, T. Kocher, L. J. Launer, T. Lehtimäki, O. Melander, T. H. Mosley, A. W. Musk, M. S. Nieminen, C. J. O’Donnell, C. Ohlsson, B. Oostra, L. J. Palmer, O. Raitakari, P. M. Ridker, J. D. Rioux, A. Rissanen, C. Rivolta, H. Schunkert, A. R. Shuldiner, D. S. Siscovick, M. Stumvoll, A. Tönjes, J. Tuomilehto, G.-J. van Ommen, J. Viikari, A. C. Heath, N. G. Martin, G. W. Montgomery, M. A.

Province, M. Kayser, A. M. Arnold, L. D. Atwood, E. Boerwinkle, S. J. Chanock, P. Deloukas, C. Gieger, H. Grönberg, P. Hall, A. T. Hattersley, C. Hengstenberg, W. Hoffman, G. M. Lathrop, V. Salomaa, S. Schreiber, M. Uda, D. Waterworth, A. F. Wright, T. L. Assimes, I. Barroso, A. Hofman, K. L. Mohlke, D. I. Boomsma, M. J. Caulfield, L. A. Cupples, J. Erdmann, C. S. Fox, V. Gudnason, U. Gyllensten, T. B. Harris, R. B. Hayes, M.-R. Jarvelin, V. Mooser, P. B. Munroe, W. H. Ouwehand, B. W. Penninx, P. P. Pramstaller, T. Quertermous, I. Rudan, N. J. Samani, T. D. Spector, H. Völzke, H. Watkins, J. F. Wilson, L. C. Groop, T. Haritunians, F. B. Hu, R. C. Kaplan, A. Metspalu, K. E. North, D. Schlessinger, N. J. Wareham, D. J. Hunter, J. R. O'Connell, D. P. Strachan, H.-E. Wichmann, I. B. Borecki, C. M. van Duijn, E. E. Schadt, U. Thorsteinsdottir, L. Peltonen, A. G. Uitterlinden, P. M. Visscher, N. Chatterjee, R. J. F. Loos, M. Boehnke, M. I. McCarthy, E. Ingelsson, C. M. Lindgren, G. R. Abecasis, K. Stefansson, T. M. Frayling, and J. N. Hirschhorn. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467: 832-8.

133. Byrnes, A. M., L. Racacho, S. M. Nikkel, F. Xiao, H. MacDonald, T. M. Underhill, and D. E. Bulman. 2010. Mutations in GDF5 presenting as semidominant brachydactyly A1. *Human mutation* 31: 1155-62.

134. Wright, S. 1931. Evolution in Mendelian Populations. *Genetics* 16: 97-159.

135. Nei, M., T. Maruyama, and R. Chakraborty. 1975. The Bottleneck Effect and Genetic Variability in Populations. *Evolution* 29: 1-10.

136. Keller, L. F., K. J. Jeffery, P. Arcese, M. A. Beaumont, W. M. Hochachka, J. N. Smith, and M. W. Bruford. 2001. Immigration and the ephemerality of a natural population bottleneck: evidence from molecular markers. *Proceedings. Biological sciences / The Royal Society* 268: 1387-94.

137. Keinan, A., and A. G. Clark. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science (New York, N.Y.)* 336: 740-3.

138. Wu, D.-D., G.-M. Li, W. Jin, Y. Li, and Y.-P. Zhang. 2012. Positive Selection on the Osteoarthritis-Risk and Decreased-Height Associated Variants at the GDF5 Gene in East Asians. *PloS one* 7: e42553.

139. Francis-West, P. H., a Abdelfattah, P. Chen, C. Allen, J. Parish, R. Ladher, S. Allen, S. MacPherson, F. P. Luyten, and C. W. Archer. 1999. Mechanisms of GDF-5 action during skeletal development. *Development (Cambridge, England)* 126: 1305-15.

140. Dodd, A. W., C. Rodriguez-Fontenla, M. Calaza, A. Carr, J. J. Gomez-Reino, A. Tsezou, L. N. Reynard, A. Gonzalez, and J. Loughlin. 2011. Deep sequencing of GDF5 reveals the absence of rare variants at this important osteoarthritis susceptibility locus. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 19: 430-4.

141. Messeguer, X., R. Escudero, D. Farré, O. Núñez, J. Martínez, and M. M. Albà. 2002. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics (Oxford, England)* 18: 333-4.

142. Sugiura, T., G. Hötten, and S. Kawai. 1999. Minimal promoter components of the human growth/differentiation factor-5 gene. *Biochemical and biophysical research communications* 263: 707-13.

143. He, Y., and P. Casaccia-Bonnefil. 2008. The Yin and Yang of YY1 in the nervous system. *Journal of neurochemistry* 106: 1493-502.

144. Donohoe, M. E., X. Zhang, L. McGinnis, J. Biggers, E. Li, and Y. Shi. 1999. Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality. *Molecular and cellular biology* 19: 7237-44.

145. Nicholson, S., H. Whitehouse, K. Naidoo, and R. J. Byers. 2011. Yin Yang 1 in human cancer. *Critical reviews in oncogenesis* 16: 245-60.

146. Dodd, A. W., C. M. Syddall, and J. Loughlin. 2012. A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression. *European journal of human genetics : EJHG* .

147. Loughlin, J. 2011. Genetics of osteoarthritis. *Current opinion in rheumatology* 23: 479-83.

148. Tennessen, J. a, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. a Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, and J. M. Akey. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)* 337: 64-9.

149. Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, N.Y.)* 337: 100-4.

150. Rightmire, G. P. 2004. Brain size and encephalization in early to Mid-Pleistocene Homo. *American journal of physical anthropology* 124: 109-23.

151. Lee, S.-H., and M. H. Wolpoff. 2003. The pattern of evolution in Pleistocene human brain size. *Paleobiology* 29: 186-196.

152. Hawks, J. 2011. Selection for smaller brains in Holocene human evolution. *arXiv:1102.5604v1 [q-bio.PE]* 1-20.

153. Henneberg, M. 1988. Decrease of human skull size in the Holocene. *Human biology* 60: 395-405.

154. Frayer, D. W. 1981. Body Size, Weapon Use, and Natural Selection in the European Upper Paleolithic and Mesolithic. *American Anthropologist* 83: 57-73.

155. Bell, J. T., and T. D. Spector. 2011. A twin approach to unraveling epigenetics. *Trends in genetics : TIG* 27: 116-25.

156. Barter, M. J., C. Bui, and D. A. Young. 2012. Epigenetic mechanisms in cartilage and osteoarthritis: DNA methylation, histone modifications and microRNAs. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 20: 339-49.