Biological variation in musculoskeletal injuries: current knowledge, future research and practical implications

Malcolm Collins, Alison V September, Michael Posthumus

ABSTRACT

Evidence from familial and genetic association studies have reported that DNA sequence variants play an important role, together with non-genetic factors, in the aetiology of both exercise-associated and occupational-associated acute and chronic musculoskeletal soft tissue injuries. The associated variants, which have been identified to date, may contribute to the interindividual variation in the structure and, by implication, mechanical properties of the collagen fibril and surrounding matrix within musculoskeletal soft tissues, as well as their response to mechanical loading and other stimuli. Future work should focus on the establishment of multidisciplinary international consortia for the identification of biologically relevant variants involved in modulating injury risk. These consortia will improve the limitations of the published hypothesis-driven genetic association studies, since they will allow resources to be pooled in recruiting large well-characterised cohorts required for whole-genome screening. Finally, clinicians and coaches need to be aware that many direct-to-consumer companies are currently marketing genetic tests directly to athletes without it being requested by an appropriately qualified healthcare professional, and without interpretation alongside other clinical indicators or lifestyle factors. These specific genetic tests are premature and are not necessarily required to evaluate susceptibility to musculoskeletal soft tissue injury. Current practice should rather consider susceptibility through known risk factors such as a positive family history of a specific injury, a history of other tendon and/or ligament injuries and participation in activities associated with the specific musculoskeletal injuries. Potential susceptible athletes may then be individually managed to reduce their risk profile.

INTRODUCTION

Acute and chronic musculoskeletal soft tissues are commonly injured as a result of participation in physical activity and specific occupations. Several extrinsic and intrinsic risk factors have been reported to predispose an individual to musculoskeletal soft tissue injuries. Evidence from familial and predominately case-control genetic association studies, and more recently a single genome-wide association study (GWAS) in canines, have reported that DNA sequence variants play an important role in the aetiology of musculoskeletal soft tissue injuries. These include ACL ruptures (reviewed in ref. 1),7–8 Achilles tendon injuries (reviewed in refs. 9 and 10),4,11–12 lateral epicondylitis,12 rotator cuff injuries14–16 and carpal tunnel syndrome17–19 in humans, as well as, equine superficial digital flexor tendinopathy20 and canine cranial cruciate ligament rupture (CCLR).21,22

We have regularly reviewed the involvement of genetic factors in tendon and ligament injuries,1–9,10,23–24 therefore, this review will specifically focus on recent developments and future research directions highlighting the discussions during the recent ‘Genomics, Genetics and Exercise Biology Celebratory Symposium in Santorini, Greece’, as well as the current and future clinical applications of this area of investigation.

GENETIC RISK FACTORS FOR MUSCULOSKELETAL SOFT TISSUE INJURIES IN HUMANS

Although there are important limitations to a case-control candidate-gene design approach,25 all the published studies to date reporting an association of DNA sequence variants with musculoskeletal soft tissue injuries in humans, equines and canines have used this approach (reviewed in ref. 1), with the exception of a single study in canines.22 In this approach, an a priori hypothesis that the protein or non-protein products of the genes are believed to be involved in the aetiology of the injury is used to select the candidates (reviewed in ref. 9). Genes encoding the basic structural components of tendons, ligaments and other connective tissue structures, namely the collagen fibril, have been identified as ideal candidates for evaluation. The collagen fibril will therefore be briefly summarised in the following paragraph.

With the exception of cartilage, the collagen fibril consists predominately of type I collagen. The type I collagen fibril and surrounding matrix also consist of other collagen types, elastin, many non-fibre forming proteins (glycoproteins) and proteoglycans (figure 1) (reviewed in refs. 10 and 24). The cellular content of connective tissue is dispersed between the collagen fibres and is responsible for the turnover and maintenance of the tissue matrix. From an exercise point of view, connective tissue is a dynamic tissue, with its cells able to respond appropriately or inappropriately to tissue loading and other stimuli.26 An appropriate response will allow for repair, healing and/or adaptation of the tissue. An inadequate or inappropriate response will eventually lead to a pathological state with degeneration or even failure of the matrix.

Common DNA sequence variants within genes encoding for structural components or regulators of the collagen fibril and other extracellular matrix proteins, such as collagens, fibrillins, proteoglycans and non-fibre forming proteins (glycoproteins),
have been associated with susceptibility to musculoskeletal soft tissue injuries (Figure 1). Interestingly, the effect sizes of some of these associations are relatively large for multifactorial disorders (OR > 2), suggesting that genetic variants are probably strong modulators of the risk for musculoskeletal soft tissue injuries. \(^{27–29}\) Sex-specific associations have also been reported for a subset of these variants, especially with ACL ruptures. \(^{10–31}\) Since tendons and ligaments are dynamic tissues, associations have not been limited to genes coding for structural proteins. Variants within genes encoding proteins involved in the angiogenesis-associated signalling pathways (such as vascular endothelial growth factor and kinase insert-domain receptor with ACL injuries \(^1\)), the apoptosis cell signalling pathways (such as CASP8 with Achilles tendinopathy), \(^{12–13}\) as well as the remodelling pathways (such as the MMP10, MMP1, MMP3, MMP12 gene locus and TIMP2), \(^{12–13}\) have also been associated with musculoskeletal soft tissue injuries.

Although the association of a limited number of the genetic markers have been repeated in other population groups or investigated in different injury types, the current results appear to indicate that there are similarities and differences with respect to the associated DNA sequence variants with acute and chronic injuries, as well as injuries with males and females. The following section will therefore focus on the reported association of variants within some of the collagen genes where these similarities and differences are illustrated.

The COL1A1 gene encodes for the \(\alpha(1)(I)\) chain of type I collagen, which is a heterotrimer consisting of two \(\alpha(1)(I)\) and one \(\alpha(2)(I)\) chains. However, a heterotrimer consisting of three \(\alpha(1)(I)\) chains may also be produced and incorporated into connective tissues together with the heterotrimer. \(^{16}\) A functional variant located in a transcription factor, known as Sp1, binding site (rs1800012, G/T) within the first intron of COL1A1 has been associated with cruciate ligament, mainly ACL, ruptures in four studies. \(^{7–9,37–39}\) This variant has also been reported to associate with shoulder dislocations \(^{37}\) and although the sample size is small, similar genotype distributions have been reported for Achilles ruptures. \(^{29–30}\) It is reasonable to propose that the COL1A1 variant could possibly predispose individuals to a number of different acute tendon and ligament injuries, but to date it has not been associated with any overuse injuries such as chronic Achilles tendinopathy \(^{40}\) and lateral epicondylitis. \(^{41}\)

With the exception of one study, where the COL1A1 GG genotype was significantly under-represented in the ACL rupture group in recreational skiers, \(^4\) all the other studies have reported similar genotype distributions, with an under-representation of the TT genotype in the acute injury groups (Table 1). \(^{37–39}\) Although a detailed explanation is beyond the
The functional COL1A1 gene is required.

Further research where the mechanism of injury and the importance of accurately documenting and stratifying for (the phantom foot mechanism of injury in skiing) and highlights the association of COL1A1 of injury is clearly documented and considered, as well as the mechanism of the injury. Further research where the mechanism of injury is clearly documented and considered, as well as the association of COL1A1 variants with acute and chronic musculoskeletal soft tissue injuries, is required.

To date, only one study which investigated the association of the functional COL1A1 rs1800012 (G/T) variant with ACL ruptures was reported in undefined populations in Turkey, while all the other publications have been in white populations. In fact, almost all of the ACL genetic risk factors have been identified in white populations. Since the frequency distribution of the COL1A1 rs1800012 (G/T) variant and other variants across the human genome varies between population groups, the reported associations cannot necessarily be extrapolated to other population groups. The major problem with the COL1A1 rs1800012 variant is the reported low frequency of the minor T allele (0.5% in Asian and 4.7% in Nigerian and Kenyan) and the TT genotype in various population groups. Very large data sets will therefore be required to understand the significance of this locus. Investigators should therefore be mindful of this observation and select appropriate informative musculoskeletal soft tissue injury genetic markers for the population group under investigation at this specific COL1A1 locus or in other proposed loci.

Like the COL1A1 variant, the role of the functional 3′-untranslated region (UTR) within the COL5A1 gene in the aetiology of musculoskeletal soft tissue injuries is also of interest. The association of up to five variants, namely rs13946 (C/T), rs12722 (C/T), rs71746744 (+/AGGG), rs16399 (ATCT-), rs1134170 (A/T), rs13946 (C/T), and rs12722 (C/T), has been tested in several acute and chronic injuries (table 1). These variants define two major functional forms of the 3′-UTR, which we have previously termed the C-allelic and T-allelic forms (figure 2B). Although not all five of the variants have been genotyped in the studies, considering all the published investigations to date, the C-allelic form appears nevertheless to be under-represented in white populations in Turkey, while rs1134170 variant is the reported low frequency of the minor T allele (0.5% in Asian and 4.7% in Nigerian and Kenyan) and the TT genotype in various population groups. Very large data sets will therefore be required to understand the significance of this locus.Investigators should therefore be mindful of this observation and select appropriate informative musculoskeletal soft tissue injury genetic markers for the population group under investigation at this specific COL1A1 locus or in other proposed loci.

<table>
<thead>
<tr>
<th>Gene and chromosomal location</th>
<th>Variant</th>
<th>Location</th>
<th>AT</th>
<th>AR</th>
<th>CL/ACL</th>
<th>Sho Dis</th>
<th>TE</th>
<th>CTS</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual genes (single variants and variant-variant interactions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL1A1 (17q21)</td>
<td>rs1800012 (G/T)</td>
<td>SPI binding site within intron 1 at nucleotide 1023</td>
<td>X</td>
<td>↓ TT (†)</td>
<td>↓ TT*</td>
<td>↓ TT</td>
<td>X</td>
<td>–</td>
<td>7 37-40</td>
</tr>
<tr>
<td></td>
<td>rs1107946 (G/T)</td>
<td>Transcription factor binding site within promoter at -1997</td>
<td>–</td>
<td>–</td>
<td>↑ G-T haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>39</td>
</tr>
<tr>
<td>COL5A1 (9q34)</td>
<td>rs13946 (C/T)</td>
<td>3′-UTR</td>
<td>X</td>
<td>X (†)</td>
<td>X</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>13 19 27 31 42 43</td>
</tr>
<tr>
<td></td>
<td>rs12722 (C/T)</td>
<td></td>
<td>↓ CC</td>
<td>X (†)</td>
<td>↓ CC (F)</td>
<td>–</td>
<td>↓ TT</td>
<td>↓ TT</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>rs71746744 (+/AGGG)</td>
<td></td>
<td>↑ AF</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>rs16399 (ATCT- )</td>
<td></td>
<td>↑ AF</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>rs1134170 (A/T)</td>
<td></td>
<td>↑ AF</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>rs13946 (C/T)</td>
<td>3′-UTR</td>
<td>–</td>
<td>–</td>
<td>↓ T-C-T-AGGG</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>rs12722 (C/T)</td>
<td></td>
<td>–</td>
<td>–</td>
<td>↓ T-C-T-AGGG</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44</td>
</tr>
<tr>
<td>COL11A1 (1p21)</td>
<td>rs3753841 (T/C)</td>
<td>Exon 52, Leu1323Pro</td>
<td>↑ T-C-T haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>rs1676486 (C/T)</td>
<td>Exon 62, Pro1535Ser</td>
<td>↑ T-C-T haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>COL12A2 (6q21)</td>
<td>rs1799907 (T/A)</td>
<td>Intron 6</td>
<td>↑ T-C-T haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>COL12A1 (6q12-q13)</td>
<td>rs970547 (A/G)</td>
<td>Exon 65, Ser3058Gly</td>
<td>↓ AA (F)</td>
<td>X (M)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>30 45</td>
</tr>
<tr>
<td>Gene-gene interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL1A1</td>
<td>rs12722 (C/T)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>↑ T-A (F) haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>5</td>
</tr>
<tr>
<td>COL12A1</td>
<td>rs970547 (A/G)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>↑ T-A (F) haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44</td>
</tr>
<tr>
<td>COL11A1</td>
<td>rs3753841 (T/C)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>↑ T-A (F) haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44</td>
</tr>
<tr>
<td>COL12A2</td>
<td>rs1676486 (C/T)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>↑ T-A (F) haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44</td>
</tr>
<tr>
<td>COL5A1</td>
<td>rs1799907 (T/A)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>↑ T-A (F) haplotype</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44</td>
</tr>
</tbody>
</table>

X, not associated; ↓, decreased risk; ↑, increased risk; (†), potentially associated or not associated but needs confirmation in larger cohorts.

*Other mechanisms of injury, except the phantom foot mechanism of injury in skiing.
†The phantom foot mechanism of injury in skiing.
††Two additional informative rare adjacent variants, namely rs14776422 (C/T) and rs55748801 (G/A), which have not been reported in a white population, were identified within the population group investigated in the CTS study. Since the genotyping method was unable to distinguish between these two variants, the CG wild-type alleles of these variants were designated as a W, while the three alternative combinations, CA, TG and TA, were designated as an M. The T-C haplotype in white populations is therefore equivalent to the T-W-C in other population groups, which contain the two additional rare variants.

AR, Achilles tendon rupture; AT, Achilles tendinopathy; CL/ACL, cruciate ligament/anterior cruciate ligament rupture; CTS, carpal tunnel syndrome; F, females; Sho Dis, shoulder dislocation; TE, tennis elbow; UTR, untranslated region.

which is a marker of the functional 3'-UTR C allele, was associated with reduced risk of ACL ruptures in females.\textsuperscript{31}

Additionally, the AA genotype of \textit{COL12A1} rs970547 (A/G) within exon 65, which results in a serine to glycine substitution at amino acid 3058 of the \(\alpha_1(XII)\) chain of type XII collagen (S3058G), was associated with an increased risk of ACL ruptures among female participants.\textsuperscript{30} In agreement, this \textit{COL12A1} polymorphism was not associated with ACL ruptures in Polish male soccer players.\textsuperscript{47} Although some of the sex-specific independent associations were not repeated, a significant interaction was noted between the \textit{COL5A1} rs12722 (T/C) and \textit{COL12A1} rs970547 (A/G) polymorphisms, where the T-A inferred allele pair was associated with increased risk of ACL rupture in females in two separate cohorts (South African and Polish).\textsuperscript{7} These results indicate a possible sex-specific association of genetic markers with muscular skeletal soft tissue injuries and the differences between acute and chronic injuries. These differences need to be considered in future research and any potential clinical applications.

In addition to protein coding RNAs, the human genome also codes for several categories of non-protein coding (non-coding) RNAs. These non-coding RNAs are predominantly involved in cellular processes regulating the flow of information from DNA to proteins.\textsuperscript{10} One of these non-coding RNA categories are small RNA molecules known as microRNAs, miRNA, which in humans play an important role in the post-transcriptional regulation of gene expression, predominately by binding to the 3'-UTRs of protein-coding genes to inhibit translation.\textsuperscript{49,50} The majority of all human protein coding genes are under the control of miRNAs. These miRNAs have been reported to regulate development of tissues and are believed to be involved in the aetiology of many diseases.\textsuperscript{49,50} In support of the important role of miRNAs in pathogenesis, the CC genotype of the polymorphic (rs499510, C/G) \textit{MIR608} gene was implicated with increased risk of chronic Achilles tendinopathy in two populations.\textsuperscript{42} The \textit{MIR608} gene encodes a small miRNA, which binds to recognition sequences within the 3'-UTR of \textit{COL5A1} and other genes to inhibit translation. This is the first non-coding gene shown to be associated with musculoskeletal soft tissue injuries and adds another level of complexity to the poorly understood aetiology of these injuries.

**GENETIC RISK FACTORS IN OTHER SPECIES**

The cranial cruciate ligament in canines and the superficial digital flexor tendon in equines are equivalent to the ACL and Achilles tendon in humans, respectively. Variants within genes encoding for proteins involved in ligament strength and stability, as well as extracellular matrix formation, have been associated with susceptibility to CCLR. These genes include, among others, \textit{COL5A1}, \textit{COL1A1}, \textit{COL3A1}, \textit{COL11A1} and \textit{FBN1}.\textsuperscript{21}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{A schematic representation of the exon (vertical lines) and intron (horizontal lines) boundaries of the human (A) \textit{COL1A1} and (B) \textit{COL5A1} genes. The functional variants within the promoter region (rs107946, G/T) and first intron (rs1800012, G/T) of \textit{COL1A1} are indicated. Both these variants are located in transcription factor binding sites, namely an unknown transcription factor binding site in the promoter and an Sp1 binding site in intron 1, which have been expanded. The variant within each binding site affects the affinity of the binding of the respective transcription factor. Increased binding and transcription have been reported to the alleles in black (ie, G allele of rs107946 and the T allele of rs1800012). An increased \(\alpha_1(I)\) chain is produced from these alleles. Two forms of \textit{COL5A1} exon 66, which encodes for the terminal amino acids of the \(\alpha_1(V)\) chain (black box), the stop codon and the 3'-untranslated region (UTR) (clear and grey boxes), are indicated. The four variants, rs13946 (C/T), rs12722 (C/T), rs71746744 (-/AGGG), rs16399 (A/T) and rs1134170 (A/T), reported to be associated with musculoskeletal soft tissue injuries, are indicated. The variants in the 3'-UTR indicated in grey are generally associated with musculoskeletal soft tissue injuries and determine one of the two major functional forms of the 3'-UTR (T allele) which has been reported to have increased mRNA stability. All the information used to construct this figure was obtained from Ensembl (http://www.ensembl.org).}
\end{figure}
One GWAS has identified three main chromosomal regions associated with CCLR on chromosomes 1, 3 and 33. Interestingly, the results of this study implicate the involvement of neurological pathways in susceptibility to CCLR.22 The involvement of these pathways in ACL injuries requires investigation. Similar to the Achilles tendon, variants within the equine TNC and COLSA1 genes are associated with risk of superficial digital flexor tendinopathy in national hunt thoroughbred racehorses.29 Since the same genes have been identified across species, these results strengthen the hypothesis that genetic variants modulate the risk of musculoskeletal soft tissue injuries in humans.

**FUNCTIONAL SIGNIFICANCE OF GENETIC RISK FACTORS FOR MUSCULOSKELETAL SOFT TISSUE INJURIES**

It is proposed that genetic variants underpinning the genetic loci associated with musculoskeletal soft tissue injuries contribute to the normal interindividual biological variation in the structure and, by implication, mechanical properties of tendons, ligaments and other connective tissues. It is also reasonable to assume that these variants also contribute to the biological variation in the tissue responses to mechanical loading and other stimuli. In addition, many of the intrinsic risk factors for musculoskeletal soft tissue injuries are in their own right multifactorial phenotypes determined, to a lesser or greater extent, by both genetic and environmental factors.9–51 For example, the heritability of flexibility, a common intrinsic risk factor, has been estimated to be between 64% and 70%.52–53 It is therefore unlikely, based on the current evidence, that the associated genetic variants are independent risk factors for musculoskeletal soft tissue injuries. Instead, they most likely modulate risk by contributing to the normal biological variations within many of the anatomical, neuromuscular and other risk factors. The matrix of connective tissues adapts in response to load under normal physiological conditions; however, there are a spectrum of loads, ranging from high forces at low frequencies to low forces at high frequencies. Inappropriate responses to these loads may therefore result in maladaptation, eventually resulting in either an acute or chronic injury. Owing to normal biological variation, we propose that each athlete’s tolerable load will be unique, thus resulting in a wide interindividual variation in the response of musculoskeletal soft tissues to load or other stimuli. A proportion of individuals will therefore intrinsically have an increased risk for injuries while others may have a reduced risk for such injuries. Molecular genetic tools integrated with other scientific disciplines should collectively be applied by researchers to elucidate the biological mechanisms of common recreational and occupational injuries.1 An increase in the correct understanding of the underlying biochemical and mechanical properties of these injuries will assist clinicians and trainers in personalising the optimal training load for the individual, allowing them to reach optimal performance while reducing the risk of injury. The following section will review our current understanding on how the associated genetic variants contribute to the biological variation in connective tissue responses to loading and other stimuli. The function of many of the associated variants is currently unknown. Some of the associated variants are located within exons and are known to potentially alter the amino acid sequence and, therefore, potentially the function of the encoded protein. These amino acid substitutions within the structural, regulatory and/or signalling proteins can contribute to normal biological variation and susceptibility to injury. This hypothesis is supported by the published work on the COL1A1, COS1A1, COL11A1 and COL11A2 associated variants, which are located in regulatory regions of the genes, which has suggested that variations within the amounts of protein produced within the tissue during development, tissue maintenance and in response to load (eg, types I, V and XI collagen) are an important contributing factor for the biological variation in function of connective tissues.

Increased binding of the transcription factor, Sp1, to the rare T allele of COL1A1 intronic polymorphism (rs1800012, G/T) has been reported to produce more α1(I) chains, which was proposed to result in increased production and incorporation of the α1(I) type I collagen homotrimer into connective tissues (figure 2A).46 In addition, an interaction between the T allele of this intronic polymorphism and the G allele of a second polymorphism (rs1107946, G/T) within a regulatory element of the COL1A1 promoter at position 1997 in reducing the risk of ACL ruptures was reported.59 Since both alleles are known to increase the expression of COL1A1,46 this interaction may potentially result in further increased expression of the α1(I) chain. Exactly how the formation of the proposed type I collagen homotrimers affects the mechanical properties of connective tissues is currently unknown.

Type V collagen, a minor fibrillar collagen relative to type I collagen, interacts with types I and III collagen and plays a role in regulating the formation and diameter of the collagen fibril (figure 1).54 Since the 3′-UTR of eukaryotic genes plays an important role in post-transcriptional gene regulation,55 56 we have previously identified two major functional allelic forms of the 3′-UTR of COLSA1, which encodes the α(V) chain of type V collagen, with the c-allelic form having a reduced, mRNA stability compared to the T form (figure 2B).46 42 As reviewed above, versions of the T-allelic form are associated with several musculoskeletal soft tissue injuries (table 1).13 19 27 31 42

On the basis of the genetic association and functional findings of the COLSA1 3′-UTR, we proposed that the relative content of type V collagen in tendons, ligaments and other connective tissues alters the fibril diameters and packing density within these tissues, which in turn will alter the mechanical properties of the tissue and susceptibility to injury.57 Type XI collagen, which is structurally and functionally similar to type V collagen, has been suggested to interact with type V collagen in regulating the formation of the collagen fibril during tendon development.58 Interestingly we have recently reported that gene-gene interactions between COL11A1 and COL11A2, which encode for the α(1)(XI) and α(2)(XI) chains of type XI collagen, respectively, as well as interactions between COL11A1, COL11A2 and COL5A1, modulate the risk of chronic Achilles tendinopathy in two populations (table 1).44 On the basis of the proposed functions of the types V and XI collagen gene variants, it was hypothesised that the gene-gene interactions may further modulate the risk of chronic Achilles tendinopathy by affecting mRNA stability, type V and/or XI collagen production, collagen fibril diameter and packing density, as well as change the biomechanical properties of the collagen fibril.44 Further work is required to test this hypothesis.

Finally, the importance of the collagen fibril in musculoskeletal soft tissue injuries is further highlighted by the fact that lethal and/or severe disease causing mutations, which occur regardless of environmental exposure and independent of other non-genetic factors, have been identified within many of the collagen and non-collagen genes that encode for proteins within or associated with the fibril (reviewed in ref. 10). This suggests that limited biological redundancy is tolerated within the collagen fibril and that common DNA sequence variants within these genes may therefore strategically impact on the biological...
consequences of environmental exposures, such as load, and therefore injury.

**FUTURE WORK**
The establishment of international consortia is an essential ‘next step’ in the identification of genetic risk factors for musculoskeletal soft tissue injuries. This will aim to improve the limitations of the previously published hypothesis-driven genetic association studies and allow resources to be pooled in recruiting large well-characterised independent cohorts required for whole-genome screening methods to identify the landscape of biologically relevant pathways across all populations involved in the aetiology of musculoskeletal soft tissue injuries.

Multidisciplinary teams consisting of clinicians, such as orthopaedic surgeons, sports and exercise medicine physicians, radiologists, as well as applied and basic scientists, should be included in the consortia. The latter group should consist of, among other disciplines, human molecular geneticists, molecular and cellular biologists, exercise physiologists, biostatisticians and bioinformaticians. It is important that bioinformaticians are involved in the study design and planning when next-generation sequencing methods are employed. To this end, an international consortium, which originally included researchers from South Africa and Poland, is currently being expanded to include additional groups from Australia, Qatar, Russia and Italy. The aim of the consortium will be to identify DNA sequence variants that modify the risk of ACL injuries and it is currently the only consortium within the ‘ATHLOME PROJECT’ that will specifically investigate exercise-associated injuries. Initially, the consortium plans to screen the current known loci for ACL injury susceptibility in larger data sets in an attempt to identify if these loci remain susceptibility loci across all populations using the hypothesis-driven candidate gene case–control study design. These replication studies will be followed by the ACL sequencing project and eventually expand to include other next-generation methodologies.

A poorly defined or heterogeneous musculoskeletal soft tissue injury profile is a common reason why genetic associations are often not repeated in independent follow-up studies. The inclusion of only homogeneous clinically well-defined injuries as cases in genetic association studies is therefore an important consideration. The selection of appropriately matched controls is also as important as the selection of the cases. Current results suggest that the control population consists of the survivors who are protected from injury and are more than likely a very select group. Care will therefore be taken by all members of the consortium to use the same criteria to accurately phenotype, with respect to ancestry, sporting and occupational details, the injury profile and mechanism of injury, other injury history and family history, as well as other appropriate medical history and medication use.

Besides the hypothesis-driven or hypothesis independent case–control approach to identify genetic risk factors, researchers also need to investigate the role of these risk factors in the development of musculoskeletal soft tissue injuries in well-designed retrospective and prospective studies. Again, large sample sizes and well-characterised participants need to be included in these studies.

The whole-genome screening method, such as GWAS, whole-genome sequencing (WGS), exon sequencing and targeted sequencing, also needs to be carefully considered. The enriched targeted exome sequencing approach, for example, has better coverage than WGS and is therefore a better technique for identifying rare variants. Since <2% of the human genome contains protein-coding genes, important regulatory elements involved in the aetiology of musculoskeletal soft tissue injuries outside of the exons and non-coding genes, such as miRNAs, will be missed using exome sequencing. For this reason, careful considerations are needed in selecting the most appropriate enriched sequencing platform.

Future research should also consider transcriptomics, which allows for a snapshot of the whole transcriptome within the tissue of interest rather than a predetermined subset of genes extracted from blood or buccal smears. Techniques include total RNA, mRNA, targeted RNA and non-coding RNA sequencing. The role of different isoforms, novel transcripts, gene fusions and small regulatory RNAs in the aetiology of musculoskeletal soft tissue injuries can be explored using these approaches.

There is growing evidence that physiological traits responding to exercise depend on DNA sequence variants as well as on epigenetics. It is reasonable to assume that environmental factors, such as exercise, may also work through epigenetic mechanisms to modulate the risk of musculoskeletal soft tissue injuries. In support of this, an age-dependent increase in the hypermethylation of the COL1A1 promoter in periodontal ligaments has been reported. Future work should therefore consider epigenomic techniques, also within the tissue of interest, such as methylation sequencing and ChIP sequencing, which can be used to identify protein-DNA and protein-RNA interactions.

Finally, the functional variants within or around the genes that modify susceptibility to injury need to be identified and the role of these functional variants and the molecular mechanisms that cause injury needs to be elucidated. This can only be done if human molecular genetics is integrated with the other biological disciplines such as molecular and cellular biology.

**CLINICAL RELEVANCE**
The overarching aim of all research focused on identifying the genetic susceptibility underlying musculoskeletal soft tissue injuries is really to understand the molecular mechanisms underpinning injury susceptibility. Before we can design tests to inform injury susceptibility, we need to understand the cumulative effect of all the risk parameters in the injury model. More important, it is pivotal that we explore the potential biological consequences of these various risk parameters before we can design effective clinical management strategies. For these reasons, research into musculoskeletal soft tissue injury susceptibility requires collaboration from a multidisciplinary network of basic scientists and clinicians.

Diagnostic and prognostic genetic testing has successfully been applied to classical Mendelian disorders, such as osteogenesis imperfecta and Ehlers-Danlos syndrome, where genetic mutations are the major determinants of the conditions. Genetic testing will, however, never be diagnostic or prognostic in nature for multifactorial disorders, such as musculoskeletal soft tissue injuries, which are caused by poorly understood complex interactions between environmental factors (eg, load) and an individual’s genetic background. Clinicians should therefore refrain from using terms such as diagnostic, prognostic or predictive when considering any clinical application of genetic profiling for injury risk. These tests would only be able to inform injury susceptibility in the context of considering all other clinical physical activity as well as current and past injury risk factors.

Although scientists are only starting to understand its contribution, numerous companies are marketing direct-to-consumer (DTC) genetic tests for common sports injuries. These DTC tests are not injury-specific and premature since the genetic data
are incomplete. The genetic data are also not interpreted together with clinical indicators and lifestyle factors to identify an altered risk for injury by an appropriately qualified health-care professional. Although beyond the scope of this review, important medical ethical issues are also not considered by these companies, especially when testing professional athletes.

Further, clinicians should be reminded that specific genetic tests are not required to evaluate susceptibility to musculoskeletal injury. Current practice should rather consider susceptibility through known risk factors such as a positive family history of a specific injury, a history of other tendon and/or ligament ruptures and participation in sports associated with the specific musculoskeletal injuries. Athletes who are potential at higher risk than can then be managed personally by, for example, the prescription of the appropriate prehabilitation exercises to reduce risk.

**CONCLUSION**

Musculoskeletal soft tissue injuries are caused by a complex interaction of a number of biological and non-biological factors. There is evidence suggesting that there is in part a genetic contribution to an individual’s risk of sustaining an injury. To date, several variants at known genetic loci have been implicated in risk. These associated variants may contribute to the interindividual variation in the structure and, by implication, mechanical properties of musculoskeletal soft tissues, as well as their response to mechanical loading and other stimuli.

The establishment of international consortia for the identification of genetic risk factors for these injuries will improve the limitations of the published hypothesis-driven genetic association studies. They will also allow resources to be pooled in recruiting large well-characterised cohorts required for whole-genome screening methods to identify biologically relevant variants involved in modulating injury risk.

Finally, clinicians need to be aware that the genetic tests for sports injuries, which are currently available DTC by several companies, are premature and that specific genetic tests are not necessary or informative to evaluate susceptibility to musculoskeletal soft tissue injury. Current practice should rather consider susceptibility through known risk factors such as a positive family history of a specific injury, a history of other tendon and/or ligament ruptures and participation in sports associated with the specific musculoskeletal injuries. These susceptible athletes may then be individually managed to reduce their risk profile.

**Funding**

Research on the identification of genetic risk factors associated with musculoskeletal soft tissue injuries was supported in part by funds from the National Research Foundation (NRF) of South Africa, University of Cape Town (UCT) and the South African Medical Research Council (MRC). MP was supported by the ThembaNkazi Trust.

**Competing interests**

MC and AVS have filed patents on the application of specific sequence variations related to risk assessment of Achilles tendon and anterior cruciate ligament injuries.

**Provenance and peer review**

Not commissioned; externally peer reviewed.

**REFERENCES**


