




## Application of an *in silico* approach identifies a genetic locus within *ITGB2*, and its interactions with *HSPG2* and *FGF9*, to be associated with anterior cruciate ligament rupture risk

Senanile B. Dlamini, Colleen J. Saunders, Mary-Jessica N. Laguette, Andrea Gibbon, Junaid Gamiieldien, Malcolm Collins & Alison V. September

To cite this article: Senanile B. Dlamini, Colleen J. Saunders, Mary-Jessica N. Laguette, Andrea Gibbon, Junaid Gamiieldien, Malcolm Collins & Alison V. September (2023): Application of an *in silico* approach identifies a genetic locus within *ITGB2*, and its interactions with *HSPG2* and *FGF9*, to be associated with anterior cruciate ligament rupture risk, European Journal of Sport Science, DOI: [10.1080/17461391.2023.2171906](https://doi.org/10.1080/17461391.2023.2171906)

To link to this article: <https://doi.org/10.1080/17461391.2023.2171906>

 View supplementary material 

 Published online: 23 Feb 2023.

 Submit your article to this journal 

 Article views: 89

 View related articles 

 View Crossmark data 

ORIGINAL INVESTIGATION



# Application of an *in silico* approach identifies a genetic locus within *ITGB2*, and its interactions with *HSPG2* and *FGF9*, to be associated with anterior cruciate ligament rupture risk

Senanile B. Dlamini<sup>a,c</sup>, Colleen J. Saunders<sup>b,d</sup>, Mary-Jessica N. Laguette<sup>a,c</sup>, Andrea Gibbon<sup>a</sup>, Junaid Gamielien<sup>d</sup>, Malcolm Collins<sup>b,a,c,e</sup> and Alison V. September<sup>a,c,e</sup>

<sup>a</sup>Division of Physiological Sciences, Department of Human Biology, University of Cape Town, Cape Town, South Africa; <sup>b</sup>Division of Emergency Medicine, Department of Surgery, University of Cape Town, Cape Town, South Africa; <sup>c</sup>Department of Human Biology, Health through Physical Activity Lifestyle and Sport Research Centre (HPALS), Newlands, South Africa; <sup>d</sup>South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa; <sup>e</sup>Department of Human Biology, International Federation of Sports Medicine (FIMS) Collaborative Centre of Sports Medicine, University of Cape Town, Newlands, South Africa

## ABSTRACT

We developed a Biomedical Knowledge Graph model that is phenotype and biological function-aware through integrating knowledge from multiple domains in a Neo4j, graph database. All known human genes were assessed through the model to identify potential new risk genes for anterior cruciate ligament (ACL) ruptures and Achilles tendinopathy (AT). Genes were prioritised and explored in a case-control study comparing participants with ACL ruptures (ACL-R), including a sub-group with non-contact mechanism injuries (ACL-NON), to uninjured control individuals (CON). After gene filtering, 3376 genes, including 411 genes identified through previous whole exome sequencing, were found to be potentially linked to AT and ACL ruptures. Four variants were prioritised: *HSPG2*:rs2291826A/G, *HSPG2*:rs2291827G/A, *ITGB2*:rs2230528C/T and *FGF9*:rs2274296C/T. The rs2230528 CC genotype was over-represented in the CON group compared to ACL-R ( $p < 0.001$ ) and ACL-NON ( $p < 0.001$ ) and the TT genotype and T allele were over-represented in the ACL-R group and ACL-NON compared to CON ( $p < 0.001$ ) group. Several significant differences in distributions were noted for the gene-gene interactions: (*HSPG2*:rs2291826, rs2291827 and *ITGB2*:rs2230528) and (*ITGB2*:rs2230528 and *FGF9*:rs2297429). This study substantiates the efficiency of using a prior knowledge-driven *in silico* approach to identify candidate genes linked to tendon and ACL injuries. Our biomedical knowledge graph identified and, with further testing, highlighted novel associations of the *ITGB2* gene which has not been explored in a genetic case control association study, with ACL rupture risk. We thus recommend a multistep approach including bioinformatics in conjunction with next generation sequencing technology to improve the discovery potential of genomics technologies in musculoskeletal soft tissue injuries.

## KEYWORDS

Semantic modelling; anterior cruciate ligament; whole exome sequencing; *Heparan sulfate proteoglycan*; *Integrin  $\beta 2$  subunit*; *Fibroblast Growth Factor 9*

## Highlights


- A biomedical knowledge graph was modelled for musculoskeletal soft tissue injuries to efficiently identify candidate genes for genetic susceptibility analyses.
- The biomedical knowledge graph and sequencing data identified potential biologically relevant variants to explore susceptibility to common tendon and ligament injuries. Specifically genetic variants within the *ITGB2* and *FGF9* genes were associated with ACL risk.
- Novel allele combinations (*HSPG2-ITGB2* and *ITGB2-FGF9*) showcase the potential effect of *ITGB2* in influencing risk of ACL rupture.

## Introduction

Ligaments and tendons, although functionally and anatomically distinct, are both dense connective tissues with similarities in molecular components (Benjamin & Ralphs, 1997).

They are common sites of acute and overuse injuries during certain occupational or physical activities (Meeuwisse, 1994). Although multiple risk factors have been implicated in the aetiology of these injuries (Meeuwisse, 1994) there is a

**CONTACT** Alison V. September  alison.september@uct.ac.za  Division of Physiological Sciences, Department of Human Biology, University of Cape Town, Anatomy building, Level 5, Anzio Road, Observatory, Cape Town 7925, South Africa; Department of Human Biology, Health through Physical Activity Lifestyle and Sport Research Centre (HPALS), 3rd Floor, Sports Science Institute Building, Boundary Road, Newlands 7701, South Africa; Department of Human Biology, International Federation of Sports Medicine (FIMS) Collaborative Centre of Sports Medicine, University of Cape Town, 3rd Floor, Sports Science Institute Building, Boundary Road, Newlands 7701, South Africa

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/17461391.2023.2171906>.

growing body of evidence suggesting that genetic factors are also important modulators of injury risk (Ljungqvist et al., 2008; Rahim et al., 2016).

To date, candidate gene association studies based on our understanding of connective tissue biology have successfully identified more than 80 genetic loci, highlighting at least 46 genes, to be implicated in the predisposition to connective tissue injuries including those affecting ligaments and tendons (Rahim et al., 2016). Many more genes require identification and validation as each effective risk allele is most likely making a small contribution to the overall genetic risk. The associated genes implicated thus far encode structural components, as well as regulators of biological processes within these connective tissues (Rahim et al., 2016). A recognisable limitation of the candidate gene approach is that it disregards potentially important genes that are not obvious candidates. It is therefore important that researchers use a more comprehensive approach to improve our understanding of the genetic complexity contributing to musculoskeletal soft tissue injury risks. There is a vast array of research information stored in public domains and, as researchers, we are not always able to keep up with the speed of these publications and how they link and integrate with clinical conditions. It is therefore not surprising that we see an increased creation of knowledge graphs for many multifactorial phenotypes (Hassani-Pak & Rawlings, 2017; Mohamed et al., 2021; Nicholson & Greene, 2020). These knowledge graphs are created by integrating and extracting knowledge from publicly available biomedical literature and linking it with the relevant information from curated biological databases (Hassani-Pak & Rawlings, 2017; Mohamed et al., 2021; Nicholson & Greene, 2020). Biomedical knowledge graphs can be used to effectively gain more comprehensive insight into identifying biological relationships between genes, proteins and molecular networks linked to a clinical condition. Thereby, it can assist in exploring the pathobiology of complex phenotypes such as ligament and tendon injuries towards improved diagnoses, treatment design and drug therapy.

Next generation sequencing is another advanced approach that can be utilised to discover promising candidate genetic loci (Gibbon et al., 2018). Gibbon et al. (2018) used whole exome sequencing (WES) and a biomedical knowledge graph to filter genes and prioritise variants mapping to 35 plausible candidate genes for investigation in a candidate gene association study for Achilles tendinopathy (AT) risk (Gibbon et al., 2018).

Ligament injuries, such as anterior cruciate ligament (ACL) ruptures, are also one of the common lower limb

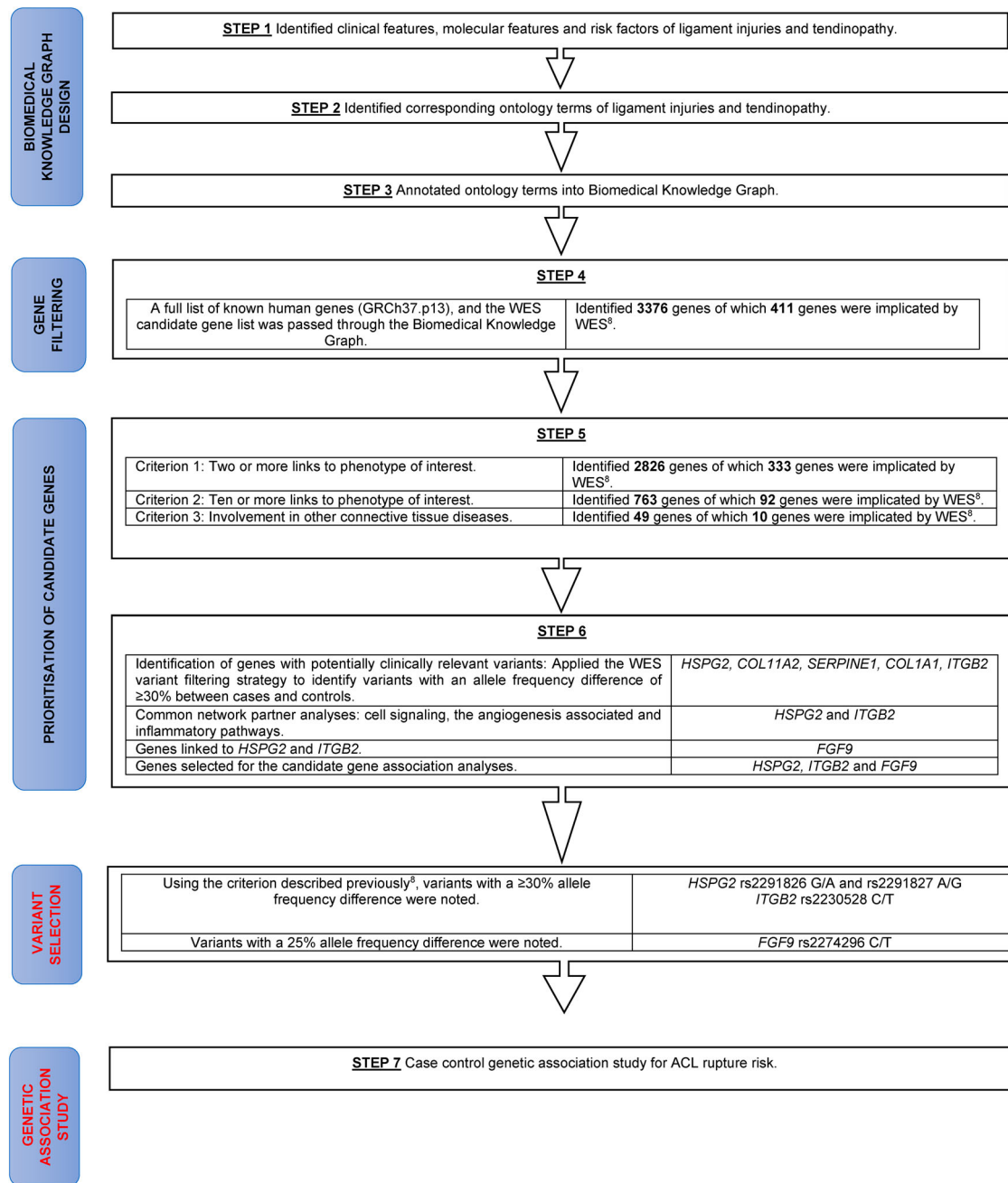
injuries and the identification of additional genetic factors that modulate injury risk is an important step in elucidating the biological mechanism of injury risk. The aim of the current study was to identify candidate genes linked to both ACL ruptures and Achilles tendinopathy by using a biomedical knowledge discovery approach and to test these candidate genes in a genetic association study using a case–control study design for ACL rupture. The objectives included incorporating knowledge related to connective tissue and this involved the integration (i) of phenotypic features of ligament injuries into the biomedical knowledge graph previously developed specifically for tendon features by Saunders et al. (2016); (ii) relevant protein network pathways, known protein–protein interactions, specifically inflammatory signalling pathways.

## Methodology

### *Biomedical knowledge graph design*

A previously described biomedical knowledge graph implemented in a Neo4j. graph database management system (<https://neo4j.com/>)<sup>9</sup> was modified to include both ligament and tendon features. We also integrated multiple bioinformatic data sources to identify potential genetic loci for risk susceptibility to both ligament and tendon injuries, specifically ACL ruptures and Achilles tendinopathy. This was achieved in three steps (Figure 1):

- (1) Reviewed published data from PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) to identify clinical features, molecular features, and risk factors for both ligament injuries and tendinopathy.
- (2) The ontology look-up service was used to identify corresponding ontology terms for clinical features, molecular features and risk factors of ligament injuries and tendinopathy.
- (3) New relations between the prioritised ontology terms and the relevant Disease Ontology term were created in the Neo4j. graph database, thereby modifying the stored semantic network to expand the representation of the phenotype of interest through links to (i) phenotypic features of ACL ruptures, (ii) known connective tissue associated network pathways, (iii) known protein–protein interactions and (iv) inflammatory signalling pathways (Table S1, Figure S1). Neo4j path-based queries were used to perform directed walks on the semantic network to identify biologically plausible direct and transitive links that may explain a gene's potential relationship to the phenotype of interest.



**Figure 1.** Flow diagram showing the Biomedical Knowledge Graph development and candidate gene prioritisation for a case-control genetic association study in ACL ruptures.

### Gene filtering

A full list of known human genes (GRCh37.p13), as well as the WES candidate gene list previously identified by Gibbon et al. (2018), was then filtered through the biomedical knowledge graph developed to identify candidates supported by prior knowledge as having potential roles in our phenotype of interest.

### Prioritisation of candidate genes

The genes highlighted by the biomedical knowledge graph were prioritised in a stepwise approach (Figure 1):

- Genes with links to the phenotype of interest through at least two independent pathways were identified.
- Genes with multiple (ten or more) links to the phenotype of interest were prioritised.
- Genes previously implicated in other connective tissue diseases were further prioritised.
- Finally, genes containing variants which were highlighted in the WES study by Gibbon et al. (2018) to have an allele frequency difference of  $\geq 30\%$  between tendinopathy cases and controls were identified (Gibbon et al., 2018).

- Several bioinformatic tools were used to characterise the functional partners of prioritised candidate genes. These tools included Enrichr, which is an enrichment analysis tool (<https://maayanlab.cloud/Enrichr/>) (Chen et al., 2013, Kuleshov et al., 2016, Xie et al., 2021) and GeneMANIA (<https://genemania.org/>) (Warde-Farley et al., 2010) a predictive gene interaction tool used to explore evidence for the shared interactions between genes and partner sets.

### **Variant selection and their potential functional effects**

*HSPG2*:rs2291826 A/G, rs2291827 G/A and *ITGB2*:rs2230528 C/T variants were selected because, in our previous WES study, it was shown that they had a  $\geq 30\%$  allele frequency distribution difference between the tendinopathy cases and controls (Gibbon et al., 2018). The *FGF9*: rs2274296 C/T variant was selected even though it had a 25% allele frequency difference between the tendinopathy cases and controls.

Very little is known about the functional effects of these four variants and for this reason their potential functional consequences were characterised *in silico* using the online software programmes (Fathmm, SIFT and PolyPhen):

- (1) Fathmm is a species-independent method used to predict functional effects of various changes within coding and non-coding regions of a gene by applying a model incorporating sequence conservation, nucleotide sequence characteristics, genomic features (codons or splice sites), amino acid features and if there is an association based on previous expression level data in different tissues (Shihab et al., 2013).
- (2) SIFT is a multistep algorithm that predicts whether an amino acid substitution is deleterious and affects protein function, by taking into account protein sequence conservation (Vaser et al., 2016).
- (3) PolyPhen, is a tool which predicts the possible impact of amino acid substitutions, annotates them to the gene transcripts and predicts both the stability and function of the encoded proteins using structural and comparative evolutionary considerations (Adzhubei et al., 2010).

### **Case control genetic association study**

#### **Participants**

The study included 237 South African Caucasian participants with ACL ruptures (ACL-R) and 232 healthy

participants with no history of ACL ruptures (CON) which were previously recruited and described (Mannion et al., 2014; Rahim et al., 2014). Participants with ACL ruptures were included regardless of the mechanism of injury and included a subgroup of 149 participants with a non-contact mechanism of ACL ruptures (ACL-NON) (Mannion et al., 2014; Rahim et al., 2014) All participants gave written informed consent and completed personal and medical questionnaires. This study was approved by the University of Cape Town Human Research Ethics Committee (HREC 164/2006). ACL ruptures were diagnosed using clinical criteria and confirmed by ultrasound, magnetic resonance imaging, arthroscopy, or during surgery.

#### **Genotyping**

The variants were genotyped using catalogued TaqMan® Genotyping Assays (Applied Biosystems, Foster City, CA, USA). PCR was conducted and analysed using the Applied Biosystems QuantaStudio3™ Real-Time PCR system (Applied Biosystems). For quality control purposes, DNA free samples and a number of positive (known genotypes) were randomly included on every 96-welled PCR plate. The genotypes were called using the Thermo Fisher Cloud suite (Thermo Fisher Scientific, Foster City, CA, USA) with an average call rate of 98%. All laboratory work was conducted at the Health through Physical Activity, Lifestyle and Sport Research (HPALS), Division of Physiological Sciences, University of Cape Town.

#### **Statistical analysis**

Power calculations were performed using QUANTO v1.2.4 (<http://biostats.usc.edu/software>). Assuming minor allele of between 0.15 and 3.5 the sample size was adequate to detect an allelic odds ratio of  $\geq 2.0$  with 80% power and an alpha value of 0.05. The programming environment R (R Development Core Team, 2010), STATISTICA v11 (StatSoft Inc., Tulsa, OK) and GraphPad Prism v5.02 (GraphPad Software Inc., San Diego, CA) were used. One-way analysis of variance (ANOVA) was used to determine significant differences in participant characteristics between the CON and ACL-R groups, and between the CON group and ACL-NON subgroup. The R packages genetics (González et al., 2007; Warnes, 2011) and SNPAssoc (Sinnwell & Schaid, 2011) were used to identify differences in genotype and allele frequencies between the groups and to calculate Hardy-Weinberg equilibrium (HWE) probabilities. Analyses were corrected for two confounders, age and sex. All genetic models were investigated, and AIC

was used to identify the most significant model (Burnham & Anderson, 2004). Haplotypes were inferred using the R package haplo.stats (Sinnwell & Schaid, 2011) and inferred allele combinations were constructed as a proxy for potential gene-gene interactions using genotype data between *HSPG2* (rs2291826, rs2291827) and *ITGB2* (rs2230528); *ITGB2* (rs2230528) and *FGF9* (rs2274296); *HSPG2* (rs2291826, rs2291827) and *FGF9* (rs2274296). The most common allele combination was automatically used as a reference. For all analyses, statistical significance was accepted when  $p < 0.05$ . The False Discovery Rate (FDR) procedure was used to adjust for multiple comparisons using the method applied for multiple testing under dependency (Benjamini & Hochberg, 1995). FDR correction was performed for significant  $p$  values and included correction for all tests and models for all four SNPs.

## Results

### Biomedical knowledge graph and gene filtering

The modified biomedical knowledge graph together with gene filtering identified 3376 genes from all known human genes, and 411 genes from the WES genes, with links to one or more ontology terms describing features of ligament injuries and tendinopathy (Figure 1).

### Prioritisation of candidate genes and variant selection

Prioritisation revealed that (i) 2826 genes (including 333 implicated by WES) were linked to ligament injuries and tendinopathy by more than one path, (ii) 763 known genes (including 92 also implicated by WES) were linked to ligament injuries and tendinopathy by more than ten paths and (iii) 49 known genes (including 10 implicated by WES) were previously implicated in other connective tissue diseases.

Applying the WES variant filtering strategy revealed five genes (*HSPG2*, *COL11A2*, *SERPINE1*, *COL1A1*, *ITGB2*) (Figure 1) containing variants with an allele frequency difference of  $\geq 30\%$  between tendinopathy cases and controls (Gibbon et al., 2018) (Table S2). These prioritised genes were further explored using the bioinformatic tools, Enrichr and GeneMANIA. The interacting genes and implicated pathways are summarised in supplementary table 3. Of the five prioritised genes, *HSPG2* and *ITGB2* were found to be most often implicated in common pathways compared to *COL11A2*, *SERPINE1* and *COL1A1*. For this reason, *HSPG2* and *ITGB2* were prioritised. The pathways implicated included cell signalling, the angiogenesis associated and inflammatory

pathways. It was interesting to note that the network analyses highlighted that *HSPG2* and *ITGB2*, together with *FGF9*, played an integral role in ECM regulation (Table S3). The Enrichr tool specifically showed interactions between (i) *ITGB2* and *FGF9* through functional pathways such as the *rap1* signalling pathway, which is an essential regulator of basic cell functions (formation, cell adhesions, cellular migration, and polarisation) (Zhang et al., 2017) and regulation of actin cytoskeleton, which is a network of actin binding proteins responsible for essential cellular processes like cell migration, organelle transport, axonal growth, cytoplasmic streaming, and phagocytosis. (ii) *ITGB2*, *FGF9* and *HSPG2* genes are all expressed in the extracellular vesicle, exosome and organelle of cells. The GeneMANIA tool showed that *ITGB2* and *HSPG2* genes shared functional pathways, e.g. extracellular matrix organisation and integrin cell surface interactions (Table S3). For these collective reasons *HSPG2*, *ITGB2* and *FGF9* were prioritised and the variants in each gene were selected as described in the methods (rs2291826, rs2291827, rs2230528 and rs2274296) for candidate gene association analyses.

The *in silico* functional effects of the specific variants were further explored and presented in Table 1. Fathmm predicted *HSPG2*: rs2291826 A/G to be functionally deleterious, although SIFT predicted it to be tolerated. Polymorphism rs2291827 G/A was predicted to be deleterious by SIFT and benign by Polyphen. *ITGB2* rs2230528 C/T was predicted to be deleterious by Fathmm and tolerable by SIFT. *FGF9* rs2274296 C/T was predicted to potentially be deleterious by Fathmm.

### Case control genetic association study

#### Participant characteristics

Participant groups were not significantly different in body mass index (BMI) nor in distribution by country of birth. The CON group was significantly older than both the ACL-R group and the ACL-NON subgroup. There were significantly less males in the CON group compared to the ACL-R group ( $p < 0.001$ ) and the ACL-NON subgroup ( $p < 0.001$ ). When adjusted for sex, there were no significant differences in height between the groups. However, when adjusted for age and sex, the CON group weighed significantly less than the ACL-R and the ACL-NON subgroup (Table S4). No significant genotype effects were noted on age, height, weight, BMI, sex nor country of birth for any of the variants investigated (Table S5).

#### Genotype and Allele frequencies

There were no significant differences in genotype or allele frequencies for *HSPG2* rs2291826 and rs2291827

**Table 1.** Summary of the predicted functional effects of the *HSPG2*, *ITGB2* and *FGF9* variants.

Gene	Variant	Chromosomal location	Variant location	Fathmm	SIFT	Polyphen	Coding effect
<i>HSPG2</i>	rs2291826 A > G	1: 21839408	Exon 6	Deleterious	Tolerated	Non-Predicted	synonymous
<i>HSPG2</i>	rs2291827 G > A	1: 21839494	Exon 6	Non-Predicted	Deleterious	Benign	non-synonymous
<i>ITGB2</i>	rs2230528 C > T	21: 44900398	Exon 7	Deleterious	Tolerated	Non-Predicted	synonymous
<i>FGF9</i>	rs2274296 C > T	13: 21681162	Intron 3	Deleterious	Non-Predicted	Non-Predicted	Unknown

between the groups (Table 2). Significant deviations from HWE were noted in the ACL-R group (0.040) and ACL-NON subgroup (0.025) for rs2291826. The *ITGB2* rs2230528 CC genotype, dominant model, was significantly over-represented in CON compared to ACL-R ( $p < 0.001$ ; OR:2.59; 95% CI:1.74-3.86) group and ACL-NON subgroup ( $p < 0.001$ ; OR:2.51; 95% CI:1.60-3.94), while the recessive model, the TT genotype was significantly over-represented in the ACL-R ( $p < 0.001$ ; OR:6.54; 95% CI:3.57-11.99) group and ACL-NON ( $p < 0.001$ ; OR:7.02; 95% CI:3.64-13.53) subgroup compared to the CON group. The over-dominant model was not significant [ $p = 0.585$  (CON vs ACL-R),  $p = 0.451$  (CON vs ACL-NON)]. The AIC score indicated the recessive model to be the most significant model for *ITGB2* rs2230528 (Table 2). Furthermore, the T allele was significantly over-represented in the ACL-R ( $p < 0.001$ ; OR:2.85; 95% CI:2.10-3.87) group and ACL-NON ( $p < 0.001$ ; OR:2.84 95% CI:2.02-3.99) subgroup compared to the CON group. Significant deviations from HWE were noted in the ACL-R group (0.000) and ACL-NON subgroup (0.000). The *FGF9* rs2274296 T allele was significantly over-represented in the ACL-R ( $p = 0.029$ ; OR:1.37; 95% CI:1.03-1.83) group compared to the CON group (Table 2).

#### Inferred haplotypes and gene-gene interactions

There were no significant differences in the distribution of inferred haplotypes constructed from *HSPG2*: rs2291826 and rs2291827 between groups (Figure S2). Allele combinations were constructed between *HSPG2*-*ITGB2*, *ITGB2*-*FGF9* and *HSPG2*-*FGF9*. Significant differences in frequencies were noted where *ITGB2*: rs2230528 was included. For *HSPG2*: rs2291826, rs2291827 and *ITGB2*: rs2230528 specifically, the (G-A-C) combination was under-represented in the ACL-R (ACL-R: 5%,  $p < 0.001$ ; OR:0.45; 95% CI:0.25-0.83) group and ACL-NON (ACL-NON: 5%,  $p < 0.001$ ; OR:0.51; 95% CI:0.25-1.03) subgroup compared to the CON group (14%) (Figure 2A); while the (A-G-T) combination was over-represented in the ACL-R (CON:19%, ACL-R:34%,  $p < 0.001$ ; OR:1.84; 95% CI:1.32-2.58) group and ACL-NON (ACL-NON:35%,  $p < 0.001$ ; OR:1.91; 95% CI:1.32-2.77) subgroup compared to the CON group (Figure 2A). Significant differences in the frequency distribution of the inferred allele combinations for *ITGB2*: rs2230528

and *FGF9*: rs2297429 were noted where the (T-T) combination was over-represented in ACL-R (CON:8%, ACL-R:17%,  $p < 0.001$ ; OR:3.19; 95% CI:1.99-5.13) group and ACL-NON (ACL-NON:17%,  $p < 0.001$ ; OR:3.09; 95% CI:1.80-5.29) subgroup compared to the CON group (Figure 2C). The (T-C) combination was also found to be over-represented in ACL-R (CON:16%, ACL-R:32%,  $p < 0.001$ ; OR:2.59; 95% CI:1.80-3.73) group and ACL-NON (ACL-NON:31%,  $p < 0.001$ ; OR:2.47; 95% CI:1.62-3.79) subgroup compared to the CON group (Figure 2B). Inferred allele combinations for *HSPG2*: rs2291826, rs2291827 and *FGF9*: rs2297429 showed no significant differences in frequency distributions between groups (Figure 2C).

#### Discussion

The biomedical knowledge graph developed for ligament injuries and tendinopathy incorporated phenotypic features of ACL ruptures and Achilles tendon (the existing already had some Achilles tendon clinical features). In addition, we incorporated knowledge related to connective tissue pathways such as signalling pathways, known protein-protein network partners and also inflammatory signalling pathways. A total of 3376 candidate genes were prioritised as having potential roles in these conditions. We propose therefore, that this tool provided a more comprehensive approach for identifying biologically-plausible candidate genes based on multiple levels of prior evidence for connective tissue injuries such as ACL ruptures and Achilles tendinopathy, in comparison to the 46 prioritised genes previously highlighted in a review (Rahim et al., 2016). Furthermore, the tool was tested using a genetic association study in which four prioritised variants in three genes: *perlecan* (*HSPG2* rs2291826 A/G and rs2291827 G/A), *integrin  $\beta 2$  subunit* (*ITGB2* rs2230528 C/T) and *fibroblast growth factor 9* (*FGF9* rs2274296 C/T) were explored in ACL rupture risk.

The *ITGB2* CC and TT genotypes were associated with a decreased and increased risk for ACL and non-contact ACL ruptures respectively, while the T allele was associated with increased risk for rupture. The *FGF9* T allele was associated with an increased risk for ACL rupture. In addition, several novel inferred allele-allele combinations, a proxy for gene-gene interactions,

**Table 2.** Genotype, minor allele frequencies and Hardy Weinberg equilibrium distributions for all participants (males and females) in the control (CON) group, anterior cruciate ligament rupture (ACL-R) group and subgroup of individuals reporting a noncontact mechanism of injury (ACL-NON) for the investigated variants *HSPG2* (rs2291826 A/G, rs2291827 G/A), *ITGB2* (rs2230528 C/T) and *FGF9* (rs2274296 C/T).

	CON % (n)	ACL-R % (n)	<i>p</i> -values <sup>a</sup>	AIC	ACL-NON % (n)	<i>p</i> -values <sup>b</sup>	AIC
<i>HSPG2</i>	218	212			133		
rs2291826							
AA	53 (115)	61 (130)	0.115		63 (84)	0.094	
AG	40 (88)	30 (63)			28 (37)		
GG	7 (15)	9 (19)			9 (12)		
G allele	27 (118)	24 (101)	0.309		23 (61)	0.246	
HWE	0.870	0.040			0.025		
<i>HSPG2</i>	218	212			133		
rs2291827							
GG	70 (152)	74 (157)	0.402		74 (99)	0.392	
GA	26 (58)	24 (51)			24 (32)		
AA	4 (8)	2 (4)			2 (2)		
A allele	17 (74)	14 (59)	0.222		14 (36)	0.240	
HWE	0.478	1.000			1.000		
<i>ITGB2</i>	212	209			135		
rs2230528							
CC	60 (127)	38 (79)	<0.001 (<0.001) <sup>c</sup>	533.9	38 (52)	<0.001 (<0.001) <sup>d</sup>	422.8
			D = <0.001	558.8		D = <0.001	445.8
			R = <0.001	535.1		R = <0.001	422.5
			O = 0.585	581.0		O = 0.451	461.8
CT	32 (69)	30 (63)			29 (39)		
TT	8 (16)	32 (67)			33 (44)		
T allele	24 (101)	47 (197)	<0.001 (<0.001) <sup>e</sup>		47 (127)	<0.001 (<0.001) <sup>f</sup>	
HWE	0.063	0.000			0.000		
<i>FGF9</i> rs2274296	220	216			138		
CC	46 (100)	38 (82)	0.057		37 (51)	0.170	
CT	43 (95)	43 (94)			47 (65)		
TT	11 (25)	19 (40)			16 (22)		
T allele	33 (145)	40 (174)	0.029 (0.034)		39 (109)	0.065	
HWE	0.880	0.180			0.863		

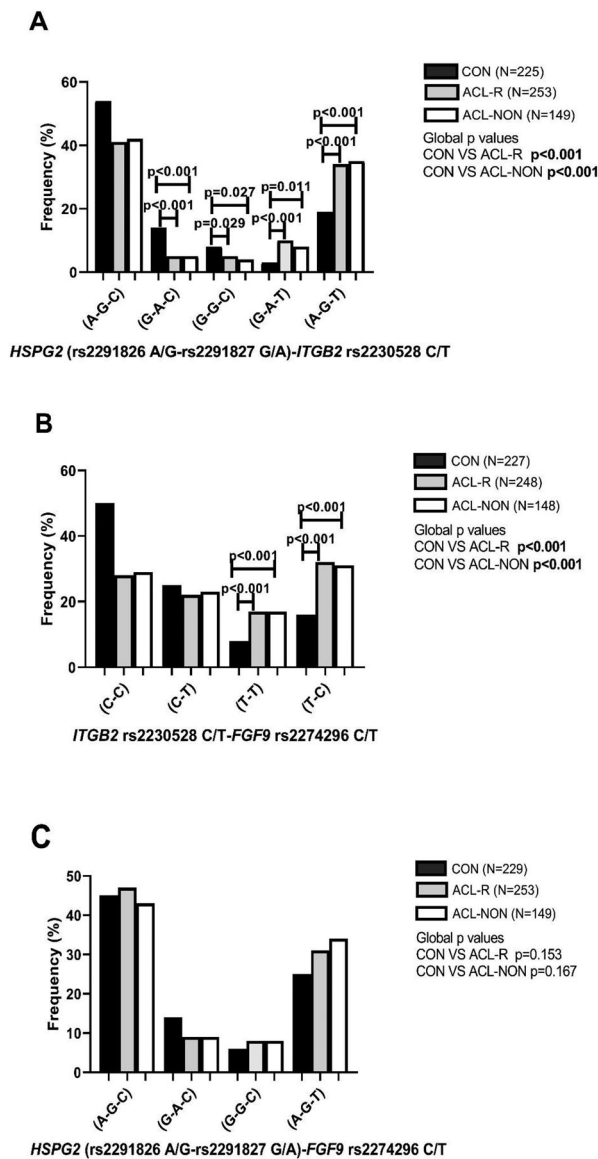
Genotype and allele frequencies are expressed as a percentage with the number of participants (n) in parentheses. CON vs. ACL-R<sup>a</sup> (adjusted *P*-values). CON vs. ACL-NON<sup>b</sup> (adjusted *P*-values). *P*-values in bold typeset indicate significance (*P* < 0.05). *P*-values corrected for multiple testing are in parenthesis;  $3.66 \times 10^{-10}$ <sup>c</sup>,  $1.24 \times 10^{-9}$ <sup>d</sup>,  $5.12 \times 10^{-11}$ <sup>e</sup>,  $9.95 \times 10^{-9}$ <sup>f</sup>. D indicates the dominant model; R indicates the recessive model and O indicates the over-dominant model. AIC indicates Akaike information criterion. *ITGB2* significant *p*-values remained significance after FDR correction while the *FGF9* T allele frequency lost significance (*p* = 0.118) after all *p* values were corrected for all models and tests for all SNPs.

were noted where *ITGB2* was included. These novel independent associations highlight (i) the successful application of the knowledge graph and (ii) the value of using the sequencing data from the previous WES project in identifying potential biologically relevant variants for exploration. The novel allele combinations showcase the potential effect of *ITGB2* in influencing risk of ACL rupture. The bioinformatic analyses does showcase evidence of key protein–protein interactions between these genes (*HSPG2-ITGB2* and *ITGB2-FGF9*) and therefore it is plausible that these interactions may potentially signal new therapeutic targets requiring exploration.

No independent associations were identified for *HSPG2*. *HSPG2* encodes for perlecan, a major heparin sulfate proteoglycan found in basement membranes which has been implicated to play a key role in binding and delivering growth factors such as platelet derived and fibroblast growth factors to the extracellular space (Melrose et al., 2006). Evidence suggests it plays a role in facilitating (i) several cellular environmental

interactions and (ii) interactions between several extracellular matrix (ECM) components and thereby it may contribute to regulating tissue homeostasis (Melrose et al., 2006; Whitelock & Iozzo, 2005). *Perlecan* is composed of five domains each with specific functions, the rs2291826 and rs2291827 variants localise within the third domain responsible for cell surface binding and secretion into the extracellular space (Colognato & Yurchenco, 2000; Melrose et al., 2006). The function of these SNPs has not been investigated in ACL ruptures. It is interesting to note that perlecan and elastin were found to colocalise within connective tissues such as the ACL attachment regions to bone, paraspinal stromal tissues and synovial tissues of the bovine knee joint (Hayes et al., 2011). These findings suggest the potential significant interactive properties between the two proteins and importance of perlecan in contributing to ligament elasticity. Elastin is regarded as one of the major structural components of elastic fibres that give tendons and ligaments elastic recoil properties (Hayes et al., 2011).





**Figure 2.** Distribution of inferred allele-allele combinations for (A) *HSPG2* (rs2291826 A/G, rs2291827 G/A) and *ITGB2* (rs2230528 C/T), (B) *ITGB2* (rs2230528 C/T) and *FGF9* (rs2297429 C/T), and (C) *HSPG2* (rs2291826 A/G, rs2291827 G/A) and *FGF9* (rs2274296 C/T). CON: Control group (black bars), ACL-R: anterior cruciate ligament rupture group (grey bars), ACL-NON: subgroup of participants with a non-contact mechanism of injury (white bars). The number of participants (n) in each group is in parentheses. Statistically significant differences in frequency between groups are indicated, with  $p$ -values adjusted for age and sex.  $P$ -values in bold typeset indicate significance ( $p < 0.05$ ).

The third domain of perlecan also contains laminin binding sites and laminins have the ability to bind to ECM macromolecules and thereby facilitating cell-matrix and cell-cell interactions (Cognato & Yurchenco, 2000). There is evidence to suggest that interactions between laminins and perlecan contribute to tissue survival, maintenance of tissue phenotype, cell

shape and movement, and cell attachment and differentiation (Cognato & Yurchenco, 2000).

Investigation of *ITGB2* highlighted that the rs2230528 CC genotype was associated with a 3-fold decreased risk for ACL and non-contact ACL injuries, while the TT genotype was associated with a 7-fold increased risk for sustaining ACL ruptures and non-contact ACL injuries. The T allele was associated with a 3-fold increased risk for sustaining an ACL ruptures and non-contact ACL injuries. The  $\beta 2$  integrins are a subgroup of the heterodimeric transmembrane receptor family which are expressed in leukocytes of all nucleated cells and share a common  $\beta 2$  chain and ligand binding sites (Fagerholm et al., 2019). Integrins play an important role in several signalling cascades through ligand binding and in cell migration, proliferation, angiogenesis, differentiation and apoptosis (Tan, 2012). and some of these processes have been implicated in maintaining ligament homeostasis. Interestingly, integrins also facilitate adhesion to the extracellular matrix by binding to the tripeptide Arg-Gly-Asp of collagens, fibronectin, laminins to name only a few ECM components of ligaments (Nikonenko et al., 2003). More recently, a WES study in twin siblings with ACL ruptures identified *ITGB2* as a predictor for non-contact ACL rupture (Caso et al., 2016). Another study which aimed to identify mechanoresponsive genes within the ECM of human periodontal ligament cells, found *ITGB2* to be one of the differentially expressed genes (Ma et al., 2015). Mutations in *ITGB2* have also been identified to cause leukocyte adhesion deficiency, a rare immunodeficiency syndrome for which a reduction or loss of expression of  $\beta 2$  integrins have been implicated and proposed to lead to compromise cytokine responses (Fagerholm et al., 2019). There is currently no empirical evidence describing the biological role of *ITGB2* rs2230528 and this would be interesting to resolve taking into account the finding of this study.

Evaluation of *FGF9* showcased the novel association of the T-allele with a 1-fold increased susceptibility for ACL ruptures. *Fibroblast growth Factor 9* is a member of the FGF ligand family which is important for many biological processes including cell migration, proliferation, and differentiation (Behr et al., 2010; Hung et al., 2007). A previous study showed that *Fgf9* +/- mice had impaired angiogenesis implicating *FGF9* in blood vessel formation. Furthermore, when a separate *Fgf9* -/- group of mice were treated with *FGF9*, angiogenesis was stimulated, and bone healing was improved (Hung et al., 2007), however, the angiogenesis signalling pathway has been implicated with ligament rupture susceptibility (Rahim et al., 2014) and taken together, these findings highlight the potential hypothesis that *FGF9* may play a significant biological role in the ACL injury

pathways. Therefore, it would be important that the functional significance of this variant be characterised to further understand the underlying molecular mechanisms related to ligament biology.

The gene-gene interactions showcase a potential dominant effect from *ITGB2*. The use of Enrichr and GeneMANIA bioinformatic tools highlighted that *HSPG2*, *ITGB2* and *FGF9* share overlapping functions in cell signalling pathways regulating extracellular matrix function, cell-cell and cell-matrix interactions. *HSPG2* has been found to enhance *FGF9* signalling through capturing and accumulating diffused growth factors such as *FGF9* further supplying them to target cells when in need as a way to prevent uncontrolled diffusion of growth factors into the extracellular space of produced cells (Matsuo & Kimura-Yoshida, 2014). Integrins and growth factors have also been found to be essential in the formation of new blood vessels through signalling cascades (Eslava-Alcon et al., 2020) and the identification of *HSPG2* as an *FGF9* regulator can mean that they all have important roles in cell repair and healing various damaged tissues including ACL and AT. The collective findings add to the growing evidence suggesting that ECM integrity is in part dependent on the proteins responsible for the regulation of cell signalling, cell-cell communication, cell proliferation, cell metabolism, inflammation and blood vessel formation (Broughton et al., 2006).

The biomedical knowledge graph described in the study was developed incorporating both ligament and tendon linked descriptors and for this reason, it is not specific to identify potential candidate genes exclusive to ligament or tendon related injuries. Furthermore, the sample size was limited and thus we were only able to detect large effect sizes. *FGF9* allele frequency significance was lost after FDR correction for all SNPs which further supports that a larger sample size is required to investigate the collective biological significance of these genes in connective tissue phenotypes. In addition, the cases and controls were not matched for sex and weight, which are confounding variables for ACL ruptures. Sports participation data was self-reported, and the participants were not matched for participation in contact and noncontact non-jumping sports (Mannion et al., 2014).

This study demonstrated the capability of a biomedical knowledge graph in conjunction with WES in identifying candidate genes for ligament and tendon injuries, as well as the importance of integrating new salient facts into an adaptable framework to facilitate further biomedical discoveries. We were able to apply this approach to ACL ruptures and found novel associations for *ITGB2*: rs2230528 and *FGF9*: rs2274296 and from the gene

interactions, *ITGB2* was further implicated in risk of ACL ruptures. These genes should therefore be prioritised for investigation into the genetic susceptibility of tendinopathy. This data contributes to a growing body of research characterising the functional and genetic signature underpinning the aetiology of musculoskeletal soft tissue injuries.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research was supported in part by funds received from the National Research Foundation (NRF) [grant number NRF 90552 and NRF 88177], the University of Cape Town (UCT) Research Council and the Department of Science and Technology [grant number NRF 92550]. SBD and AG were financially supported by the European Union funded project RUBICON H2020-MSCA-RISE-2015-690850. M-JNL is a HPALS research fellow.

## ORCID

Malcolm Collins  <http://orcid.org/0000-0002-2564-0480>

## References

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), 248–249. <https://doi.org/10.1038/nmeth0410-248>
- Behr, B., Leucht, P., Longaker, M. T., & Quarto, N. (2010). Fgf-9 is required for angiogenesis and osteogenesis in long bone repair. *Proceedings of the National Academy of Sciences*, 107(26), 11853–11858. <https://doi.org/10.1073/pnas.1003317107>
- Benjamin, M., & Ralphs, J. R. (1997). Tendons and ligaments—an overview. *Histol. Histopathol*, 12(4), 1135–1144. <https://doi.org/10.14670/HH-12.1135>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B*, 57(1), 289–300. <https://doi.org/10.2307/2346101>
- Broughton IIG., Janis, J. E., Attinger, C. E., (2006). The basic science of wound healing. *Plastic and Reconstructive Surgery*, 117(SUPPLEMENT), 12S–34S. <https://doi.org/10.1097/01.prs.0000225430.42531.c2>
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Caso, E., Maestro, A., Sabiers, C. C., Godino, M., Caracuel, Z., Pons, J., Gonzalez, F. J., Bautista, R., Claros, M. G., Caso-Onzain, J., Viejo-Allende, E., Giannoudis, P. V., Alvarez, S., Maietta, P., Guerado, E., et al. (2016). Whole-exome sequencing analysis in twin sibling males with an anterior cruciate

- ligament rupture. *Injury*, 47(3), S41–S50. [https://doi.org/10.1016/S0020-1383\(16\)30605-2](https://doi.org/10.1016/S0020-1383(16)30605-2)
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1), 1. <https://doi.org/10.1186/1471-2105-14-128>
- Colognato, H., & Yurchenco, P. D. (2000). Form and function: The laminin family of heterotrimers. *Developmental Dynamics*, 218(2), 213–234. [https://doi.org/10.1002/\(SICI\)1097-0177\(200006\)218:2<213::AID-DVDY1>3.0.CO;2-R](https://doi.org/10.1002/(SICI)1097-0177(200006)218:2<213::AID-DVDY1>3.0.CO;2-R)
- Eslava-Alcon, S., Extremera-García, M. J., Sanchez-Gomar, I., Beltrán-Camacho, L., Rosal-Vela, A., Muñoz, J., Ibarz, N., Alonso-Piñero, J. A., Rojas-Torres, M., Jiménez-Palomares, M., González-Rovira, A., Conejero, R., Doiz, E., Rodríguez-Piñero, M., Moreno-Luna, R., & Durán-Ruiz, M. C. (2020). Atherosclerotic Pre-conditioning affects the paracrine role of circulating angiogenic cells Ex-vivo. *International Journal of Molecular Sciences*, 21(15), 5256. <https://doi.org/10.3390/ijms21155256>. PMID: 32722151; PMCID: PMC7432497.
- Fagerholm, S. C., Guenther, C., Llorc Asens, M., et al. (2019). Beta2-Integrins and interacting proteins in leukocyte trafficking, immune suppression, and immunodeficiency disease. *Front. Immunol*, 10(254), 1–10. <https://doi.org/10.3389/fimmu.2019.00254>
- Gibbon, A., Saunders, C. J., Collins, M., Gamielien, J., & September, A. V. (2018). Defining the molecular signatures of achilles tendinopathy and anterior cruciate ligament ruptures: A whole-exome sequencing approach. *PLoS One*, 13(10), 1–20. <https://doi.org/10.1371/journal.pone.0205860>
- González, J. R., Armengol, L., Solé, X., Guino, E., Mercader, J. M., Estivill, X., & Moreno, V. (2007). SNPAssoc: An R package to perform whole genome association studies. *Bioinformatics (oxford, England)*, 23(5), 654–655. <https://doi.org/10.1093/bioinformatics/btm025>
- Hassani-Pak, K., & Rawlings, C. (2017). Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. *Journal Of Integrative Bioinformatics*, 14(1), 1–9. <https://doi.org/10.1515/jib-2016-0002>
- Hayes, A. J., Lord, M. S., Smith, S. M., Smith, M. M., Whitelock, J. M., Weiss, A. S., & Melrose, J. (2011). Colocalization in vivo and association in vitro of perlecan and elastin. *Histochemistry and Cell Biology*, 136(4), 437–454. <https://doi.org/10.1007/s00418-011-0854-7>
- Hung, I. H., Yu, K., Lavine, K. J., Ornitz, D. M., et al. (2007). FGF9 regulates early hypertrophic chondrocyte differentiation and skeletal vascularization in the developing stylopod. *Developmental Biology*, 307(2), 300–313. <https://doi.org/10.1016/j.ydbio.2007.04.048>
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1), W90–W97. <http://dx.doi.org/10.1093/nar/gkw377>
- Ljungqvist, A., Schwellnus, M. P., Bachl, N., Collins, M., Cook, J., Khan, K. M., Maffulli, N., Pitsiladis, Y., Riley, G., Golspink, G., Venter, D., Derman, E. W., Engebretsen, L., & Volpi, P. (2008). International Olympic committee consensus statement: Molecular basis of connective tissue and muscle injuries in sport. *Clinics in Sports Medicine*, 27(1), 231–239. <https://doi.org/10.1016/j.csm.2007.10.007>
- Ma, J., Zhao, D., Wu, Y., Xu, C., Zhang, F., et al. (2015). Cyclic stretch induced gene expression of extracellular matrix and adhesion molecules in human periodontal ligament cells. *Archives of Oral Biology*, 60(3), 447–455. <https://doi.org/10.1016/j.archoralbio.2014.11.019>
- Mannion, S., Mtintsilana, A., Posthumus, M., van der Merwe, W., Hobbs, H., Collins, M., September, A. V., et al. (2014). Genes encoding proteoglycans are associated with the risk of anterior cruciate ligament ruptures. *British Journal of Sports Medicine*, 48(22), 1640–1646. <https://doi.org/10.1136/bjsports-2013-093201>
- Matsuo, I., & Kimura-Yoshida, C. (2014). Extracellular distribution of diffusible growth factors controlled by heparan sulfate proteoglycans during mammalian embryogenesis. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 369(1657), 1657. <https://doi.org/10.1098/rstb.2013.0545>
- Meeuwisse, W. H. (1994). Assessing causation in sport injury: A multifactorial model. *Clinical Journal of Sport Medicine*, 4(3), 166–170. <https://doi.org/10.1097/00042752-199407000-00004>
- Melrose, J., Roughley, P., Knox, S., Smith, S., Lord, M., Whitelock, J., et al. (2006). The structure, location, and function of perlecan, a prominent pericellular proteoglycan of fetal, post-natal, and mature hyaline cartilages. *Journal of Biological Chemistry*, 281(48), 36905–36914. <https://doi.org/10.1074/jbc.M608462200>
- Mohamed, S. K., Nounu, A., & Nováček, V. (2021). Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics*, 22(2), 1679–1693. <https://doi.org/10.1093/bib/bbaa012>
- Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 2(18), 1414–1428. <https://doi.org/10.1016/j.csbj.2020.05.017>
- Nikonenko, I., Toni, N., Moosmayer, M., Shigeri, Y., Muller, D., Sargent Jones, L., et al. (2003). Integrins are involved in synaptogenesis, cell spreading, and adhesion in the post-natal brain. *Developmental Brain Research*, 140(2), 185–194. [https://doi.org/10.1016/S0165-3806\(02\)00590-4](https://doi.org/10.1016/S0165-3806(02)00590-4)
- Rahim, M., Collins, M., & September, A. (2016). Genes and musculoskeletal soft-tissue injuries. *J. Sports Sci. Med*, 61, 68–91. <https://doi.org/10.1159/000445243>
- Rahim, M., Gibbon, A., Hobbs, H., van der Merwe, W., Posthumus, M., Collins, M., & September, A. V. (2014). The association of genes involved in the angiogenesis-associated signalling pathway with risk of anterior cruciate ligament rupture. *Journal of Orthopaedic Research*, 32(12), 1612–1618. <https://doi.org/10.1002/jor.22705>
- R Development Core Team. (2010). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. [www.r-project.org](http://www.r-project.org).
- Saunders, C. J., Dashti, M. J. S., Gamielien, J., (2016). Semantic interrogation of a multi knowledge domain ontological model of tendinopathy identifies four strong candidate risk genes. *Scientific Reports*, 6(19820), 1–10. <https://doi.org/10.1038/srep19820>
- Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., Gaunt, T. R., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, 34(1), 57–65. <https://doi.org/10.1002/humu.22225>

- Sinnwell, J., & Schaid, D. (2011). Haplo.stats: A package for statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. *R package (version 1.5.4)*. <http://cran.rproject.org/package=haplo.stats>.
- Tan, S. M. (2012). The leucocyte beta2 (CD18) integrins: The structure, functional regulation and signalling properties. *Bioscience Reports*, 32(3), 241–269. <https://doi.org/10.1042/BSR20110101>
- Vaser, R., Adusumalli, S., Ngak Leng, S., Sikic, M., Ng, P. C., et al. (2016). SIFT missense predictions for genomes. *Nature Protocols*, 11(1), 1–9. <https://doi.org/10.1038/nprot.2015.123>
- Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., ... Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38 (suppl\_2), W214–W220. <https://doi.org/10.1093/nar/gkq537>
- Warnes, G. (2011). Genetics: A package for population genetics. *R Package (version 1.3.6)*. <http://cran.r-project.org/package=genetics>.
- Whitelock, J. M., & Iozzo, R. V. (2005). Heparan sulfate: A complex polymer charged with biological activity. *Chemical Reviews*, 105(7), 2745–2764. <https://doi.org/10.1021/cr010213m>
- Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., ... Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, 1(3), e90. <https://doi.org/10.1002/cpz1.v1.3>.
- Zhang, Y. L., Wang, R. C., Cheng, K., et al. (2017 Feb). Roles of Rap1 signalling in tumor cell migration and invasion. *Cancer Biology & Medicine*, 14(1), 90–99. <https://doi.org/10.20892/j.issn.20953941.2016.0086>. PMID: 28443208; PMCID: PMC5365179.