SJMSS Section IV: Sports Medicine & Orthopaedics

THE GENETIC ASSOCIATION WITH INJURY RISK IN MALE ACADEMY SOCCER PLAYERS DEPENDS ON MATURITY STATUS

Elliott CR Hall[1], Philipp Baumert[2], Jon Larruskain[3], Susana M Gil[4], Josean A Lekue[3], Edgardo Rienzi[5], Sacha Moreno[5], Marcio Tannure[6], Conall F Murtagh[1,7], Jack D Ade[7], Paul Squires[7], Patrick Orme[8], Liam Anderson[9], Thomas E Brownlee[9], Craig M Whitworth-Turner[1], James P Morton[1], Barry Drust[9], Alun G Williams[10,11] and Robert M Erskine[1,11]

[1]*School of Sport and Exercise Sciences, Liverpool John Moores University, Liverpool, UK;*
[2]*Exercise Biology Group, Faculty of Sport and Health Sciences, Technical University of Munich, Munich, Germany;*
[3]*Medical Services, Athletic Club, Lezama, Spain;*
[4]*Department of Physiology, Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Leioa, Spain;*
[5]*Club Atlético Peñarol, Estadio Campeón del Siglo, Montevideo, Uruguay;*
[6]*Clube de Regatas do Flamengo, Rio de Janiero, Brazil;*
[7]*Liverpool Football Club, Liverpool, UK;*
[8]*Bristol City Football Club, Bristol, UK;*
[9]*School of Sport, Exercise and Rehabilitation Sciences, University of Birmingham, Birmingham, UK;*
[10]*Department of Sport and Exercise Science, Manchester Metropolitan University, Manchester, UK;*
[11]*Institute of Sport, Exercise and Health, University College London, London, UK.*

Address for correspondence:
Dr Rob Erskine, PhD,
School of Sport and Exercise Sciences
Liverpool John Moores University
Liverpool, L3 3AF
United Kingdom.
Email: R.M.Erskine@ljmu.ac.uk;
Tel: +44 151 904 6256;
R.M.Erskine ORCID: 0000-0002-5705-0207

Key words:

Genetics, football, adolescent, DNA, epidemiology

Acknowledgements:

We are extremely grateful to the following individuals for their assistance with data collection: Sam Temple, Dr Mateo Gamarra, Dr Emiliano Vigna, Dr Gustavo Schmitner, Dr Luisina Passarello, Daniel Silva, Diego Morena, Bruno Jotta Costa and Dr John Chaffe. This study was funded as part of a Liverpool John Moores University PhD Scholarship.

Abstract

It is currently unknown if injury risk is associated with genetic variation in academy soccer players (ASP). We investigated whether nine candidate single nucleotide polymorphisms were associated (individually and in combination) with injury in ASP at different stages of maturation. Saliva samples and one season's injury records were collected from 402 Caucasian male ASP from England, Spain, Uruguay and Brazil, whose maturity status was defined as pre- or post-peak height velocity (PHV). Pre-PHV *COL5A1* rs12722 CC homozygotes had relatively higher prevalence of any musculoskeletal soft-tissue (22.4% *vs.* 3.0%, *P*=0.018) and ligament (18.8% *vs.* 11.8%, *P*=0.029) injury than T-allele carriers, while *VEGFA* rs2010963 CC homozygotes had greater risk of ligament/tendon injury than G-allele carriers. Post-PHV *IL6* rs1800795 CC homozygotes had a relatively higher prevalence of any (67.6% *vs.* 40.6%, *P*=0.003) and muscle (38.2% *vs.* 19.2%, *P*=0.013) injuries than G-allele carriers. Relatively more post-PHV *EMILIN1* rs2289360 CC homozygotes suffered any injury than CT and TT genotypes (56.4% *vs.* 40.3% and 32.8%, *P*=0.007), while the 'protective' *EMILIN1* TT genotype was more frequent in post- than pre-PHV ASP (22.3 *vs.* 10.0%, *P*=0.008). Regardless of maturity status, T-alleles of *ACTN3* rs1815739 and *EMILIN1* rs2289360 were associated with greater absence following ankle injury, while the *MMP3* rs679620 T-allele and *MYLK* rs28497577 GT genotype were associated with greater absence following knee injury. The combination of injury-associated genotypes was greater in injured *vs.* non-injured ASP. This study is the first to demonstrate that a genetic association exists with injury prevalence in ASP, which differs according to maturity status.

Key words: Genetics, football, adolescent, DNA, epidemiology

Introduction

Soccer injuries negatively influence player availability and team success[1] and may affect the development of academy soccer players (ASP)[2]. Previous injury represents a major risk factor, meaning that players who suffer injuries at younger ages increase their risk of subsequent injury[1]. Accordingly, injury prevention is an important consideration for ASP but before prevention strategies are considered, the factors that influence injury risk must be identified[2]. Despite several known risk factors[1], inter-individual variation in the frequency and severity of soccer injuries suggests not all factors are fully understood.

Genetic variation is an intrinsic factor associated with injury risk in professional (senior) soccer players[5,6,7,8,9] but it is not known if it can explain the variability in injury risk in ASP. Common genetic variants have the potential to influence the structure, function or expression of proteins within tissues, which can affect tissue phenotypes[10]. It therefore follows that genetic differences between players may alter the mechanical properties of musculoskeletal soft tissues, such as skeletal muscle, ligament and tendon[12], which are commonly injured in ASP[11], thus potentially explaining some of the inter-individual variability in injury frequency and severity. In addition, the polygenic nature of most human phenotypes suggests any genetic influence on injuries in ASP is likely to involve multiple gene variants. However, no study has sought to investigate individual genetic associations, or the combined influence of multiple polymorphisms, on injury risk in ASP. We have recently shown that the genetic profile of ASP differs according to maturity status, with post-PHV players displaying a more power-orientated genetic profile and pre-PHV players a profile more suited to endurance capability[13]. Thus, given that maturity status is also associated with injury risk[3], it is possible that a potential genetic association with injury risk in ASP differs according to maturity status.

The aim of this study was to investigate whether variations in genes encoding key proteins in the structure and/or function of musculoskeletal tissues were associated with injury in high-level ASP. We hypothesised that: (i) the high-risk genotypes of nine single nucleotide polymorphisms (SNPs) previously associated with musculoskeletal injury in other populations would be associated with greater injury prevalence and absence due to injury in ASP; (ii) genetic associations would differ according to maturity status; and (iii) there would be a combined association of multiple gene variants with injury prevalence.

Materials and Methods

Participants and study period

This study recruited 402 Caucasian male ASP aged 9-23 years registered with the academies of eight professional soccer clubs from England (5), Spain (1), Uruguay (1) and Brazil (1). Of the five English academies, two were categorised under the Premier League's Elite Player Performance Plan (EPPP) as Category 1 and two were Category 2. One English academy operated independently of the EPPP but competed regularly with Category 1 academies (Under 23 level). The Uruguayan academy was of the highest national category (Category A). There is no classification system for soccer academies in Spain or Brazil, although the Spanish and Brazilian academies included in this study are recognised as among the most successful in their respective countries. Due to the influence of maturity status on injury in this population[3], players were grouped and analysed according to maturity status using non-invasive procedures[4]. Due to a limited number of mid-PHV players, only pre-PHV (n = 101, age 11.5 ± 1.1 years, height 1.47 ± 0.06 m, mass 37.7 ± 4.5 kg) and post-PHV (n = 301, age 17.5 ± 2.1 years, height 1.78 ± 0.07 m, mass 71.7 ± 8.7 kg) were included for analysis. Written informed consent was obtained from academy officials and players, with parental consent and player assent provided for all participants <16 years of age in England and <18 in other countries. The

study received approval from Liverpool John Moores University Research Ethics Committee and complied with the Declaration of Helsinki.

Injury recording and definitions

Injuries were recorded during seasons 2014-15 (147 players), 2016-17 (11 players), and 2017-18 (244 players), with each player contributing one season only. For each academy, this corresponded to the season with the most individual player records available. Injuries were diagnosed and recorded by medical personnel at each academy using published guidelines[14]. An injury was recorded if it had taken place during soccer-related activity and resulted in a player being unable to participate in training or competition for a minimum of 24 hours following the incidence or onset of injury. Players were considered injured from the date of injury until the date of return to full training and availability for match selection, with absence due to injury defined as the difference between both dates. Previous injury history was not available for this study. The injuries selected for analysis were those occurring most frequently in an audit involving this cohort[11] and are detailed in Table 1. Injuries categorised as muscle rupture/strain/cramps, sprain/ligament injury or tendon injury/rupture/tendinosis/bursitis were investigated as individual categories and considered collectively as 'musculoskeletal soft-tissue' injuries. Thigh injuries included the anterior and posterior thigh, thus inclusive of the quadriceps and hamstrings.

INSERT TABLE 1 NEAR HERE.

Saliva samples and DNA isolation

Saliva samples were collected following abstinence from food or drink for at least 30 minutes. Participants were instructed to add at least 2 mL of saliva to a sterile collection tube containing

2 mL stabilisation buffer (GeneFix, Isohelix, Kent, UK). Samples were stored at -80ºC. Extraction of DNA was performed using a genomic DNA isolation kit (PureLink Genomic DNA Mini Kit, Invitrogen, UK), according to the manufacturer's instructions. All isolated DNA was stored at 4 ºC.

Genotyping

All participants were genotyped for nine SNPs (Table 2) using real-time polymerase chain reaction (PCR) on a Rotor-Gene Q PCR machine (Qiagen, Manchester, UK). Reactions were completed on a 72-well rotor disc, each reaction containing 5 µL Genotyping Master Mix (Applied Biosystems, Foster City, California, USA), 3.5 µL nuclease-free H_2O (Qiagen), 0.5 µL genotyping assay containing SNP-specific TaqMan primers and probes (Applied Biosystems), and 1.0 µL of participant DNA. For negative controls, DNA was replaced by 1 µL nuclease-free H_2O, and positive controls were used to corroborate results. The PCR protocol involved 50 denaturation cycles of incubation at 92ºC for 15 s, followed by annealing and extension at 60ºC for 1 min. Genotype was determined using Rotor-Gene Q Software 2.3.1. All samples and controls were analysed in duplicate with 100% agreement.

INSERT TABLE 2 NEAR HERE.

Total genotype score (TGS) calculation

The SNPs individually associated with injury prevalence and/or absence due to injury (after correction for multiple comparisons) were included in a TGS model[31]. Using the results of this study, genotypes associated with higher injury prevalence for each SNP were given a score of 2, with a linear trend applied to the remaining genotypes so that the genotype with the next highest prevalence was scored as 1 and the genotype with the lowest injury prevalence was

scored as 0. For each player, the TGS was converted to a percentage based on the number of injury-associated alleles. Accordingly, a higher TGS indicates increased injury risk.

Statistical and data analysis

Data are presented as mean ± standard deviation (SD) unless otherwise stated. For each injury category, players were initially grouped according to maturity status and then by whether they had suffered at least one injury, or no injury. The Chi-square (χ^2) test of independence assessed whether injury prevalence (proportion of players injured during the season) for each category was independent of genotype group. All SNPs were analysed in a co-dominant model (AA *vs* Aa *vs* aa) in the first instance (except for SNPs with a minor allele frequency [MAF] <0.35, where a dominant model was used from the outset). Where tendencies existed toward statistical significance (e.g. p = 0.05-0.10), dominant (AA *vs* Aa+aa) or recessive (AA+Aa *vs* aa) models were used, depending on which homozygous genotype tended to associate with higher or lower injury prevalence. Odds ratios (ORs) were calculated where injury prevalence differed significantly between genotype groups. For players with at least one injury for each category, differences in days missed due to injury according to genotype and PHV groups were analysed using two-way between groups ANOVA. Differences in TGS between injured and non-injured players, and between pre- and post-PHV players, were analysed by independent samples t-test. A false discovery rate (FDR) of 0.2 was applied to all χ^2 and ANOVA analyses using the Benjamini-Hochberg (B-H) method to control for multiple comparisons, i.e. the original nine p-values (one per SNP) in each model[32]. Statistical significance was accepted when $p<0.05$. Statistical analyses were performed using IBM SPSS version 25.0 (Armonk, NY, USA).

Results

Hardy-Weinberg equilibrium (HWE) and genotype frequency distribution

Genotype frequencies of all nine SNPs are shown in Table 2. Genotype distributions were in HWE for all nine SNPs amongst pre-PHV ($\chi^2 \leq 3.797$, $p \geq 0.051$) and for all SNPs amongst post-PHV ($\chi^2 \leq 3.179$, $p \geq 0.074$) except *COL5A1* rs12722 ($\chi^2 = 6.887$, $p = 0.008$). *EMILIN1* genotype frequency differed according to maturity status, with relatively more TT homozygotes in post-PHV than pre-PHV (22.3% *vs.* 10.0%, $p = 0.008$; Table 2). No maturity-associated differences in genotype frequency were observed for any other SNP (Table 3).

INSERT TABLE 3 NEAR HERE.

Injury frequency,, prevalence and absence

Two hundred and ninety injuries were recorded over the course of the season, resulting in 7,680 (median = 27.0 per injury) days missed through injury. The majority of injuries suffered were non-contact (66.6%). The most commonly injured tissue was skeletal muscle (65.5%), with the thigh representing the most commonly injured location (24.8%) and the knee injuries leading to the greatest median days missed following injury (25 days). Injury frequency, prevalence and absence are summarised in Table 1.

Injury prevalence

Any injury

In post-PHV alone, *IL6* rs1800795 CC homozygotes were 3.1 times more likely to be injured than G-allele carriers ($\chi^2 = 8.964$, $p = 0.003$; Table 4, Figure 1A), while *EMILIN1* rs2289360 CC homozygotes were 1.9 times more likely to be injured than CT heterozygotes, and 2.7 times more likely to be injured than TT homozygotes ($\chi^2 = 10.019$, $p = 0.007$; Table 4, Figure 2).

INSERT FIGURE 1 NEAR HERE.

INSERT FIGURE 2 NEAR HERE.

Musculoskeletal soft-tissue injury

In pre-PHV alone, *COL5A1* rs12722 C-allele carriers were 9.3 times more likely to be injured than TT homozygotes ($\chi^2 = 6.165$ $p = 0.018$; Table 4, Figure 3).

INSERT FIGURE 3 NEAR HERE.

Muscle injury

In post-PHV alone, *IL6* rs1800795 CC homozygotes were 2.6 times more likely to be injured than G-allele carriers ($\chi^2 = 6.527$, $p = 0.015$; Table 4, Figure 1B). In pre- and post-PHV combined, *IL6* rs1800795 CC homozygotes were 2.4 times more likely to be injured than G-allele carriers ($\chi^2 = 5.930$, $p = 0.019$, Table 4).

Ligament injury

In pre-PHV alone, *COL5A1* rs12722 CC homozygotes were 1.7 times more likely to be injured than CT heterozygotes ($\chi^2 = 6.212$, $p = 0.029$; Table 4). No TT homozygotes suffered ligament injuries. Also in pre-PHV alone, *VEGFA* rs2010963 CC homozygotes were 10.3 times more likely to be injured than GG homozygotes and 11.7 times more likely to be injured than GC heterozygotes ($\chi^2 = 12.871$, $p = 0.010$; Table 4). Further, when combining ligament and tendon injuries, pre-PHV *VEGFA* rs2010963 CC homozygotes were 6.7 times more likely to be injured than GG homozygotes and 11.7 times more likely to be injured than GC heterozygotes ($\chi^2 = 11.269$, $p = 0.011$; Table 4).

Absence (days missed) due to injury

Knee injury

MMP3 rs679620 genotype was associated with days missed following knee injuries [F (1, 57) = 5.17, p = 0.027], where T-allele carriers missed more days than CC homozygotes (median (interquartile range) = 29 (47) *vs* 10 (23)). *MYLK* rs28497577 genotype was also associated with days missed following knee injuries [F (1, 56) = 4.72, p = 0.034], where GT heterozygotes missed more days than GG homozygotes (50 (31) *vs* 16 (29)). Both associations were independent of maturity status.

Ankle injuries

ACTN3 rs1815739 genotype was associated with days missed following ankle injuries [F (1, 35 = 5.10, p = 0.032], with T-allele carriers missing more days than CC homozygotes (27 (45) *vs* 16 (30)). *EMILIN1* rs2289360 genotype was also associated with days missed through ankle injuries [F (1, 35 = 6.05, p = 0.020], with T-allele carriers missing more days than CC homozygotes (27 (51) *vs* 10 (31)). Both associations were independent of maturity status.

INSERT TABLE 4 NEAR HERE.

TGS

Seven SNPs were individually associated with injury prevalence and/or absence due to injury (*ACTN3, COL5A1, EMILIN1, IL6, MMP3, MYLK* and *VEGFA*), and were therefore included in the TGS. In this study, the *EMILIN1* C-allele was associated with greater injury prevalence and the T-allele with greater injury absence. Because we investigated the influence of TGS on injury prevalence only, the C-allele allele was considered the 'risk' allele. Players (regardless of maturity status), who had suffered one or more injury of any description, had a higher TGS than non-injured players (46.5 ± 13.1 *vs.* 43.9 ± 12.6, t(395) = -1.981 , p = 0.048). However,

TGS did not differ between pre- and post-PHV (45.6 ± 12.4 *vs.* 44.8 ± 13.2, $t(396) = 0.551$, $p = 0.582$).

Discussion

This study is the first to investigate the genetic association with injury risk in academy soccer players (ASP). The main findings were the individual associations of *COL5A1*, *VEGFA*, *EMILIN1* and *IL6* SNPs with injury prevalence, and *ACTN3*, *EMILIN1*, *MYLK* and *MMP3* with absence due to injury. Moreover, the types of SNPs associated with injury differed according to maturity status, suggesting pre-PHV ASP injury risk was associated with variation in genes affecting the mechanical and material properties of muscle/tendon/ligament, while the genetic association with injury risk in post-PHV ASP appeared to be linked to inflammation as well as musculoskeletal properties. Further, the combination of high-risk genotypes was higher in those players who had suffered any injury compared to non-injured players. We also report, for the first time, a higher frequency of the *EMILIN1* TT genotype (i.e. the 'protective' genotype with regard to injury prevalence) in post-PHV ASP compared to pre-PHV, suggesting fewer 'genetically predisposed' ASP are retained by academies when they become physically mature. Therefore, this study presents the first evidence for a genetic link with injury risk in ASP, which appears to depend on maturity status.

In our pre-PHV players, *COL5A1* rs12722 C-allele carriers and CC homozygotes suffered relatively more musculoskeletal soft-tissue injuries and ligament injuries, respectively, compared to T-allele carriers. In professional (senior) soccer, however, the T-allele has been associated with musculoskeletal injury severity[7], hamstring injury risk[9] and anterior cruciate ligament (ACL) injury risk[33], with no association in another study[6]. The lack of agreement between our results and previous studies regarding the 'risk' allele could be due to the different populations studied. This is the first time that a genetic association with injury

has been investigated in ASP, and the fact that we found *COL5A1* associations with musculoskeletal soft tissue and ligament injury prevalence in pre- but not post-PHV players suggests that these associations depend on maturity status. These associations could relate to lower tendon stiffness in pre-pubertal boys compared to physically mature men[34]. For example, it is possible that the C-allele negatively influences collagenous tissue properties prior to puberty due to the influence of this SNP on *COL5A1* mRNA stability, where the C-allele is associated with lower type V collagen production than the T-allele, and might affect the tensile strength and stiffness of collagen fibrils in musculoskeletal tissues[12]. The increase in muscle size and strength, and tendon stiffness[34] observed with maturation may counter the unfavourable effects of the C-allele and explain the lack of association observed in post-PHV ASP. The *COL5A1* CC genotype has also been associated with faster sprinting and greater jump height in a similar population[13], and greater speed/power might lead to more eccentric soft tissue damage (e.g. during deceleration/landing), which might explain why our CC homozygotes had a greater risk of musculoskeletal soft-tissue injury than their T-allele counterparts.

Ligament and ligament/tendon injury prevalence was also greater for *VEGFA* rs2010963 CC homozygotes in pre-PHV ASP, which is in accordance with the finding that CC homozygotes suffer more ACL injuries in professional soccer[35]. Whilst limited to pre-PHV, our finding supports the association of the *VEGFA* CC genotype with soccer-related ligament injury, which may be due to enhanced protein expression and plasma VEGF-A concentration in CC homozygotes[30]. Increased VEGF expression upregulates MMP3 expression[36], possibly compromising extracellular matrix (ECM) homeostasis and negatively affecting the biomechanical properties of ligaments and tendons. As with *COL5A1* rs12722, the lack of association between *VEGFA* rs2010963 and injury in post-PHV suggests a genetic influence that is dependent on maturity status, and one that is linked to the structural proteins influencing

the mechanical and material properties of collagenous musculoskeletal soft-tissue in pre-PHV players.

In our post-PHV players, the greater prevalence of any injury among *EMILIN1* rs2289360 CC homozygotes might also suggest a genetic influence on the mechanical properties of collagenous tissues. In professional soccer, the CC genotype has been associated with fewer knee injuries,[5] but also with longer absences after knee injury.[9] Our results appear to be in contrast with these findings, with the CC genotype associated with greater risk of any injury, and the T-allele with longer absence following knee injury. This SNP's intronic location suggests gene or mRNA stability could differ by genotype[10], which might affect the compliance and tensile strength of collagenous tissue, increasing joint laxity and risk of tissue rupture. Elastin is a key protein in the composition of the ECM and within ligament and tendon tissue[19], thus, if this SNP affects elastin protein expression, it could influence the risk of injury to ligament and tendon, and potentially skeletal muscle through variability in muscle-tendon tensile strength. Our findings suggest that the *EMILIN1* T-allele reduces the risk of injury in ASP and could explain the higher proportion of TT homozygotes in post-PHV ASP compared to pre-PHV (e.g. fewer 'genetically predisposed' ASP progressing from pre-PHV to post-PHV). Alternatively, it is possible that the *EMILIN1* TT genotype provides a performance advantage, potentially influencing muscle-tendon unit mechanical properties, which in turn relate to power production in this population[13]. This is supported by our previous work, demonstrating the genotype frequency distributions of SNPs associated with greater power and sprint performance were higher in post- *vs.* pre-PHV ASP[13].

The *IL6* rs1800795 SNP was also associated with post-PHV injury prevalence in our study, with relatively more CC homozygotes suffering any injuries and muscle injuries, which is in accordance with the C-allele being associated with hamstring injury risk in professional soccer[6]. It is possible that more intense and aggressive match-play in physically mature players[1]

exposes post-PHV players to more injury-inciting events, or that this induces greater post-exercise muscle damage and inflammation amongst some post-PHV players. Due to the association of the *IL6* C-allele with an increased creatine kinase response to eccentric contractions[37], we propose that C-allele carriers experience greater muscle damage and inflammation at all ages, but that this may only translate to injury risk in post-PHV due to more intense playing styles and match actions, as ASP progress toward professional soccer[1]. Accordingly, the association of inflammation-related SNPs with injury may only become apparent in post-PHV players through the combination of unfavourable genotype(s), greater match intensity, and more frequent training and/or competition. Whilst further mechanistic research is required to support this hypothesis, it appears post-PHV players' genetic susceptibility to injury might relate to genes involved in the inflammatory response to soccer-specific exercise, as well as the structure of musculoskeletal tissues.

In addition to injury prevalence, we found associations between *ACTN3*, *EMILIN1*, *MMP3* and *MYLK* SNPs and absence due to injury, all independent of maturity status. We observed that *ACTN3* T-allele ('X'-allele) carriers missed more days following ankle injuries, while in professional soccer, the T-allele has been associated with slower recovery from muscle injuries[8], and the TT genotype with musculoskeletal injury incidence[8] but not hamstring injury[6]. The association of the TT genotype with non-contact ankle sprain incidence in non-athletes[38] relates to our finding, and it is possible that the greater muscle size and strength associated with the C-allele[15] might increase lower limb stability and favour return to play after ankle injury. *EMILIN1* TT homozygotes also missed more days following ankle injuries, despite the CC genotype being associated with a greater risk of any injury in post-PHV. If the C-allele does indeed augment collagenous tissue stiffness, a stiffer tendon in CC homozygotes might cause muscle fascicles to undergo greater strain and therefore potentially more damage during eccentric actions. This hypothesis is supported by a stiffer tendon leading to greater

muscle fascicular lengthening and markers of muscle damage in men versus women following eccentric exercise.[39] This might partly explain the greater prevalence of any injury in *EMILIN1* CC homozygotes compared to CT and TT genotypes. Accordingly, it is possible that the T-allele protects against injuries in general through increased (tendon) tissue compliance. However, this increased compliance may be insufficient to withstand more excessive external loads, leading to more severe damage, e.g. tissue rupture. Under these circumstances, an extensive period for healing and rehabilitation is required and could explain the greater number of days missed from ankle injury in TT homozygotes. The association of the TT genotype with longer absence following ankle injuries in ASP in this study, and with medial collateral ligament injury in professional soccer[5,9], suggests the *EMILIN1* rs2289360 SNP may be important in providing a mechanistic explanation for soccer-related ligament injury predisposition and recovery.

Another SNP linked to days absent following injury in this study was *MMP3* rs679620. In this case, the CC genotype was associated with fewer days' absence following knee injuries, and has also been associated with lower risk and severity of hamstring injury in professional soccer[6]. The *MMP3* rs679620 T-allele is in linkage disequilibrium with the *MMP3* rs3025058 5A-allele, which increases *MMP3* expression[40]. Therefore, the rs679620 T-allele might increase RNA expression and intensify the degradation of proteins, such as collagen and elastin, thus weakening the structural integrity of ligament and ECM. The *MYLK* GT genotype was also associated with longer absence following knee injuries. As the rare T-allele has been associated with increased post-exercise creatine kinase and muscle strength loss[28], it is possible that the *MYLK* T-allele facilitates greater muscle damage following exercise and impairs contractile capacity in musculature surrounding the knee, which could contribute to prolonged recovery from injury. However, a low MAF (and therefore low number of TT homozygotes in our study) restricts the ability to generalise our findings, meaning larger sample sizes are

required in future studies to support our *MYLK* rs28497577 T-allele association with injury risk in this population.

Following the individual SNP analyses, we investigated whether the combination of only those associated SNPs was linked to injury prevalence in ASP. We found that players (regardless of maturity status) recording at least one injury of any kind had a higher TGS than non-injured players, suggesting a combination of 'unfavourable' alleles increases injury risk. To our knowledge, this is the first evidence of a polygenic association with soccer injury prevalence, suggesting that possession of enough 'risk' alleles can predispose some ASP to injury, while sufficient 'protective' alleles have the opposite effect. Importantly, there was no difference in TGS between pre- and post-PHV, suggesting that ASP progression is not limited by a polygenic susceptibility to injury. If this were the case, post-PHV might be expected to have a lower TGS than pre-PHV, according to the hypothesis that carrying more 'protective' alleles would increase the chance of remaining injury-free and assist players' progression through academy soccer. However, it should be noted that post-PHV ASP did have a lower frequency of the *EMILIN1* 'risk' genotype than pre-PHV, although this may be related to the T-allele having a potentially performance enhancing benefit, as discussed above.

While our study presents extremely novel findings concerning the genetic association with injury in ASP, it has some limitations. Firstly, mid-PHV players were excluded due to being relatively few, and it is possible that the investigated SNPs might affect injury risk during this period of rapid growth differently to pre- and post-PHV ASP. Secondly, genetic variability is one of numerous risk factors, and quantifying other factors (in addition to maturity status) here might provide greater context to our findings. Nevertheless, the current sample size is greater than any previous studies of genotype-injury associations in soccer (with only one exception being a study on 710 elite (professional) soccer players from Italy and Japan[41]), and our findings are the first of their kind in academy (youth) soccer. It is pertinent to note that

training and match hours were not recorded in the current study, meaning we could not calculate injury incidence rates or account for the possibility of different training protocols between academies. Finally, whilst these results offer greater external validity than previous studies due to the inclusion of players from several clubs and countries, replication by independent groups is required to support our findings.

Perspective

This study is the first to demonstrate that injury risk in ASP is associated with genetic variability, and that this association is dependent on maturity status. We also demonstrate that the *EMILIN1* TT ('protective') genotype is overrepresented in post-PHV ASP compared to pre-PHV, suggesting fewer 'genetically predisposed' pre-PHV ASP progress when they become mature. Further, the combination of high-risk genotypes was greater in those players who had suffered at least one injury during the season than non-injured players, indicating that injury risk in this under-researched population is polygenic. Future studies may wish to adopt a longitudinal study design (thus following the same ASP from pre- through to post-PHV) to compare with our cross-sectional data. We believe our novel findings have important implications for the aetiology of injuries in ASP and demonstrate the potential for genetic data to help inform traditional injury prevention and rehabilitation strategies in future practice.

Conflict of Interest

The authors have no conflicts of interests to declare, financial or otherwise.

References

1. Arnason A, Sigurdsson SB, Gudmundsson A, Holme I, Engebretsen L, Bahr R. Physical fitness, injuries, and team performance in soccer. Med Sci Sports Exerc. 2004;36(2):278-285.
2. Price RJ, Hawkins RD, Hulse MA, Hodson A. The Football Association medical research programme: an audit of injuries in academy youth football. Br J Sports Med. 2004;38(4):466-471.
3. Bult HJ, Barendrecht M, Tak IJR. Injury risk and injury burden are related to age group and peak height velocity among talented male youth soccer players. Orthop J Sports Med. 2018;6(12): 2325967118811042.
4. Mirwald RL, Baxter-Jones AD, Bailey DA, Beunen GP. An assessment of maturity from anthropometric measurements. Med Sci Sports Exerc. 2002;34(4):689-694.
5. Artells R, Pruna R, Dellal A, Maffulli N. Elastin: a possible genetic biomarker for more severe ligament injuries in elite soccer. A pilot study. Muscles Ligaments Tendons J. 2016;6(2):188-192.
6. Larruskain J, Celorrio D, Barrio I, et al. Genetic Variants and Hamstring Injury in Soccer: An Association and Validation Study. Med Sci Sports Exerc. 2018;50(2):361-368.
7. Massidda M, Bachis V, Corrias L, Piras F, Scorcu M, Calò CM. Influence of the COL5A1 rs12722 on musculoskeletal injuries in professional soccer players. J Sports Med Phys Fitness. 2015;55(11):1348-1353.
8. Massidda M, Voisin S, Culigioni C, et al. ACTN3 R577X Polymorphism Is Associated With the Incidence and Severity of Injuries in Professional Football Players. Clin J Sport Med. 2019;29(1):57-61.

9. Pruna R, Artells R, Ribas J, et al. Single nucleotide polymorphisms associated with non-contact soft tissue injuries in elite professional soccer players: influence on degree of injury and recovery time. BMC Musculoskelet Disord. 2013;14:221.
10. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. In: Nat Rev Genet. Vol 3. England, 2002:391-397.
11. Hall ECR, Larruskain J, Gil SM, et al. An injury audit in high-level male youth soccer players from English, Spanish, Uruguayan and Brazilian academies. Physical Therapy in Sport. 2020;44:53-60.
12. Collins M, Posthumus M. Type V collagen genotype and exercise-related phenotype relationships: a novel hypothesis. Exerc Sport Sci Rev. 2011;39(4):191-198.
13. Murtagh CF, Brownlee TE, Rienzi E, et al. The genetic profile of elite youth soccer players and its association with power and speed depends on maturity status. PLos ONE. 2020;15(6):e0234458.
14. Fuller CW, Ekstrand J, Junge A, et al. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. Br J Sports Med. 2006;40(3):193-201.
15. Seto JT, Quinlan KG, Lek M, et al. ACTN3 genotype influences muscle performance through the regulation of calcineurin signaling. J Clin Invest. 2013;123(10):4255-4263.
16. Hubal MJ, Devaney JM, Hoffman EP, et al. CCL2 and CCR2 polymorphisms are associated with markers of exercise-induced skeletal muscle damage. J Appl Physiol. 2010;108(6):1651-1658.
17. Harmon BT, Orkunoglu-Suer EF, Adham K, et al. CCL2 and CCR2 variants are associated with skeletal muscle strength and change in strength with resistance training. J Appl Physiol. 2010;109(6):1779-1785.

18. McKee TJ, Perlman G, Morris M, Komarova SV. Extracellular matrix composition of connective tissues: a systematic review and meta-analysis. Sci Rep. 2019;9(1):10542.
19. Mann V, Hobson EE, Li B, et al. A COL1A1 Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. J Clin Invest. 2001;107(7):899-907.
20. Birk DE, Fitch JM, Babiarz JP, Doane KJ, Linsenmayer TF. Collagen fibrillogenesis in vitro: interaction of types I and V collagen regulates fibril diameter. J Cell Sci. 1990;95 (4):649-657.
21. Bressan GM, Daga-Gordini D, Colombatti A, Castellani I, Marigo V, Volpin D. Emilin, a component of elastic fibers preferentially located at the elastin-microfibrils interface. J Cell Biol. 1993;121(1):201-212.
22. Oh VM, Chua BM, Heng CK, Yeo SB, Yim OS, Yap EP. Association of intronic single-nucleotide polymorphisms in the EMILIN1 gene with essential hypertension in a Chinese population. J Hum Hypertens. 2012;26(9):553-561.
23. Andersen MB, Pingel J, Kjær M, Langberg H. Interleukin-6: a growth factor stimulating collagen synthesis in human tendon. J Appl Physiol (1985). 2011;110(6):1549-1554.
24. Fishman D, Faulds G, Jeffery R, et al. The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis. J Clin Invest. 1998;102(7):1369-1376.
25. Raleigh SM, van der Merwe L, Ribbans WJ, Smith RK, Schwellnus MP, Collins M. Variants within the MMP3 gene are associated with Achilles tendinopathy: possible interaction with the COL5A1 gene. Br J Sports Med. 2009;43(7):514-520.

26. Nie G, Wen X, Liang X, Zhao H, Li Y, Lu J. Additional evidence supports association of common genetic variants in MMP3 and TIMP2 with increased risk of chronic Achilles tendinopathy susceptibility. J Sci Med Sport. 2019;22(10):1074-1078.
27. Sweeney HL, Bowman BF, Stull JT. Myosin light chain phosphorylation in vertebrate striated muscle: regulation and function. Am J Physiol. 1993;264(5 Pt 1):C1085-1095.
28. Clarkson PM, Hoffman EP, Zambraski E, et al. ACTN3 and MLCK genotype associations with exertional muscle damage. J Appl Physiol (1985). 2005;99(2):564-569.
29. Matsumoto K, Ema M. Roles of VEGF-A signalling in development, regeneration, and tumours. J Biochem. 2014;156(1):1-10.
30. Schneider BP, Radovich M, Sledge GW, et al. Association of polymorphisms of angiogenesis genes with breast cancer. Breast Cancer Res Treat. 2008;111(1):157-163.
31. Williams AG, Folland JP. Similarity of polygenic profiles limits the potential for elite human physical performance. J Physiol. 2008;586(1):113-121.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc. 1995;57(1):289-300.
33. Lulińska-Kuklik E, Rahim M, Domańska-Senderowska D, et al. Interactions between COL5A1 Gene and Risk of the Anterior Cruciate Ligament Rupture. J Hum Kinet. 2018;62:65-71.
34. Charcharis G, Mersmann F, Bohm S, Arampatzis A. Morphological and mechanical properties of the quadriceps femoris muscle-tendon unit from adolescence to adulthood: effects of age and athletic training. Front Physiol. 2019;10:1082.
35. Lulińska-Kuklik E, Leźnicka K, Humińska-Lisowska K, et al. The VEGFA gene and anterior cruciate ligament rupture risk in the Caucasian population. Biol Sport. 2019;36(1):3-8.

36. Wang H, Keiser JA. Vascular endothelial growth factor upregulates the expression of matrix metalloproteinases in vascular smooth muscle cells: role of flt-1. Circ Res. 1998;83(8);832-840.
37. Yamin C, Duarte JA, Oliveira JM, et al. IL6 (-174) and TNFA (-308) promoter polymorphisms are associated with systemic creatine kinase response to eccentric exercise. Eur J Appl Physiol. 2008;104(3):579-586.
38. Shang X, Li Z, Cao X, et al. The association between the ACTN3 R577X polymorphism and noncontact acute ankle sprains. J Sports Sci. 2015;33(17):1775-1779.
39. Hicks KM, Onambele-Pearson GL, Winwood K, Morse CI. Gender differences in fascicular lengthening during eccentric contractions: the role of patella tendon stiffness. Acta Physiol (Oxf). 2013;209(3):234-244.
40. Medley TL, Kingwell BA, Gatzka CD, Pillay P, Cole TJ. Matrix metalloproteinase-3 genotype contributes to age-related aortic stiffening through modulation of gene and protein expression. Circ Res. 2003;92(11):1254-61.
41. Massidda M, Miyamoto-Mikami E, Kumagai H, Ikeda H, et al. Association between the ACE I/D polymorphism and muscle injuries in Italian and Japanese elite football players. J Sports Sci. 2020;38(21):2423-2429.

Table 1. Definitions and details of recorded injuries for each injury category according to maturity status (pre- and post-peak height velocity, PHV).

Injury category *Definition*	**Injury Prevalence (%)**			**Number of Injuries**			**Total days missed**			**Median days missed**		
	Pre-PHV	**Post-PHV**	**All ASP**	**Pre-PHV**	**Post-PHV**	**All ASP**	**Pre-PHV**	**Post-PHV**	**All ASP**	**Pre-PHV**	**Post-PHV**	**All ASP**
Any injury *Any injury recorded in the study*	34.7	46.5	41.3	55	235	290	1,112	6,568	7,680	19	28	27
Non-contact injury *Injuries occurring without physical contact*	29.7	30.2	30.1	45	148	193	787	4,224	5,011	19	24	23
Musculoskeletal soft-tissue injury *Injuries occurring in muscle, ligament or tendon*	16.8	36.2	31.3	45	145	190	508	4,245	4,753	10	22	19
Muscle injury *Injuries recorded as muscle rupture/tear/strain/cramps*	6.0	21.3	17.4	6	86	92	123	1,373	1,496	12	14	14
Ligament injury *Injuries recorded as sprain/ligament injury*	8.9	16.3	14.4	9	57	66	173	2,502	2,675	13	25	23
Ligament/tendon injury *Injuries recorded as sprain/ligament injury OR tendon injury/rupture/tendinosis/bursitis*	10.9	19.9	17.7	11	70	81	189	2,989	3,178	10	28	24
Non-contact musculoskeletal soft-tissue injury *Injury to muscle, ligament or tendon occurring without physical contact*	29.7	30.2	30.1	17	151	168	312	4,059	4,371	14	24	23
Growth-related injury *Injuries linked to somatic growth*	17.8	1.3	5.5	24	4	28	539	41	580	17	10	15
Low back/sacrum/pelvis injury *Injuries recorded as low back/sacrum/pelvis injury*	10.9	6.3	7.5	13	25	38	244	764	1,008	19	21	20
Knee injury *Injuries located at the knee*	13.9	14.3	14.2	15	46	61	374	2,354	2,728	14	27	25
Ankle injury *Injuries located at the ankle*	4.0	10.3	8.7	5	38	43	54	1,202	1,256	13	27	24
Thigh injury *Injuries located in any muscle of the thigh*	5.0	15.9	13.7	5	67	72	129	1,147	1,276	27	17	18
Hamstring injury *Injuries located within the hamstring muscles*	5.0	8.3	7.5	5	34	39	129	600	729	27	14	17

Table 2. Information pertaining to the nine single nucleotide polymorphisms (SNPs) investigated in the current study. Locations of each SNP are described according to the Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13). Alleles for each SNP are reported according to the forward DNA strand.

Gene	Encoded protein	Protein function	Alleles and rs number	Location	MAF	SNP Type	SNP function
ACTN3	α-actinin-3	Binds actin to Z-line in type II skeletal muscle fibres [15]; blocks calcineurin and inhibits slow myogenic programme [15]	C>T rs1815739	Chr.11:66560624	T: 0.43	Transition substitution, nonsense mutation, intragenic	T-allele produces stop codon at amino acid 577, which prevents protein production [15]
CCL2	Chemokine (C-C motif) ligand-2	Mediates systemic changes induced by chronic exercise; expressed within interstitial space following muscle damage [16]	G>C rs2857656	Chr.17:34254988	C: 0.42	Intron, transversion substitution, intragenic	Purportedly involved in transcription regulation. C allele associated with lower pre-exercise muscle strength [17]
COL1A1	Pro-α1 (I) chain	Major component of type I collagen, a structural protein in connective tissues such as ligament, tendon [18]	A>C rs1800012	Chr.17:50200388	A: 0.09	Intron, Transversion substitution, intragenic	Affects binding site for the *COL1A1* Sp1 transcription factor. A-allele associated with greater α1 (I) chain mRNA and protein [19]
COL5A1	Pro-α1 (V) chain	Major component of type V collagen, a regulator of collagen fibril diameter [20]	C>T rs12722	Chr.9:134842570	T: 0.35	Transition substitution, UTR 3, intron, intragenic	3'-UTR SNPs may alter level, location or timing of gene expression [10]. TT genotype could reduce tensile strength and increase stiffness of ligament/tendon [12]
EMILIN1	Elastin microfibril interfacer-1	ECM glycoprotein involved in the microfibrillar structure of elastic fibres. Regulates elastogenesis [21]	C>T rs2289360	Chr.2:27079297	T: 0.37	Intron, Transition substitution, intragenic	Currently unknown. CC genotype associated with hypertension, possibly by reducing arterial compliance [22]
IL6	Interleukin-6	Modulates cytokine release and involved in collagen synthesis [23]. Potential as signalling molecule associated with post-exercise satellite cell proliferation	C>G rs1800795	Chr.7:22727026	C: 0.14	Intron, transversion substitution, intragenic	Possibly affects glucocorticoid receptor and transcription. G-allele associated with increased plasma IL6 [24]
MMP3	Matrix metalloproteinase-3	One of >20 MMPs that can catalytically degrade collagens and ECM substrates [25]. Also activates other MMPs	C>T rs679620	Chr.11:102842889	C: 0.48	Mis-sense mutation, transition substitution, intragenic	Currently unknown. C>T substitution replaces glutamate residue with lysin residue but does not affect MMP3 protein function [26]
MYLK	Myosin light-chain kinase	Activated by contraction-induced Ca^{2+} influx and phosphorylates the regulatory light chains that connect myosin heads [27]	G>T rs28497577	Chr.3:123793780	T: 0.20	Transversion substitution, mis-sense mutation, UTR 5, intron, intragenic	Potentially alters regulatory light chain phosphorylation. TT genotype associated with elevated muscle damage biomarkers [28]
VEGFA	Vascular endothelial growth factor-A	Considered the dominant inducer of angiogenesis. Production increased in oxygen-deprived cells [29]	C>G rs2010963	Chr.6:43770613	C: 0.30	Intron, transversion substitution, UTR 5, intragenic	Influences *VEGFA* expression and protein production. CC genotype associated with enhanced gene expression and higher plasma VEGF-A concentration [30]

SNP, single nucleotide polymorphism; *MAF*, minor allele frequency; *Chr*, chromosome; *ECM*, extracellular matrix; *MMP*, matrix metalloproteinas

Table 3. Genotype frequency distribution of the nine single nucleotide polymorphisms (SNPs) analysed in pre- and post-peak height velocity (PHV) ASP, and in all ASP (pre- and post-PHV combined)

SNP	Genotype	Pre-PHV *n* (%)	Post-PHV *n* (%)	All ASP *n* (%)	MAF	
					Pre-PHV	Post-PHV
ACTN3	CC	33 (33.6)	94 (31.3)	127 (31.9)	T: 0.42	T: 0.43
rs1817539	CT	47 (48.0)	149 (49.7)	196 (49.3)		
	TT	18 (18.4)	57 (19.0)	75 (18.8)		
CCL2	GG	47 (46.5)	138 (45.8)	185 (46.0)	C: 0.33	C: 0.32
rs2857656	GC	49 (48.5)	128 (42.5)	177 (44.0)		
	CC	5 (5.0)	35 (11.7)	40 (10.0)		
COL1A1	CC	63 (63.0)	199 (66.1)	262 (65.3)	A: 0.21	A: 0.19
rs1800012	CA	33 (33.0)	89 (29.6)	122 (30.4)		
	AA	4 (4.0)	13 (4.3)	17 (4.3)		
COL5A1	CC	16 (16.0)	72 (24.0)	88 (22.0)	T: 0.42	T: 0.45
rs12722	CT	51 (51.0)	126 (42.0)	177 (44.3)		
	TT	33 (33.0)	102 (34.0)	135 (33.7)		
EMILIN1	CC	36 (36.0)	95 (31.6)	131 (32.7)	T: 0.37	T: 0.45
rs2289360	CT	54 (54.0)	139 (46.2)	193 (48.1)		
	TT	**10 (10.0)**	**67 (22.2)***	77 (19.2)		
IL6	GG	36 (36.0)	134 (44.5)	170 (42.4)	C: 0.37	C: 0.33
rs1800795	GC	55 (55.0)	133 (44.2)	188 (46.9)		
	CC	9 (9.0)	34 (11.3)	43 (10.7)		
MMP3	CC	28 (28.0)	81 (27.0)	109 (27.3)	T: 0.47	T: 0.47
rs679620	CT	50 (50.0)	154 (51.3)	204 (51.0)		
	TT	22 (22.0)	65 (21.7)	87 (21.7)		
MYLK	GG	85 (85.9)	244 (81.3)	329 (82.5)	T: 0.10	T: 0.09
rs28497577	GT	22 (14.1)	56 (18.7)	70 (17.5)		
	TT	0 (0.0)	0 (0.0)	0 (0.0)		
VEGFA	GG	33 (33.3)	135 (45.0)	168 (42.1)	C: 0.38	C: 0.33
rs2010963	GC	56 (56.6)	133 (44.3)	189 (47.4)		
	CC	10 (10.1)	32 (10.7)	42 (10.5)		

**greater than pre-PHV (p = 0.008); MAF, minor allele frequency.*

Table 4. Single nucleotide polymorphisms (SNPs) associated with injury prevalence in pre- and post-PHV players.

Group	Injury Type	Gene	SNP	Genotype	Prevalence (%)	*p* value
Pre-PHV	Musculoskeletal soft-tissue	*COL5A1*	rs12722	**CC/CT***	22.4	0.018
				TT	3.0	
	Ligament	*COL5A1*	rs12722	**CC***	18.8	0.029
				CT	11.8	
				TT	0.0	
	Ligament	*VEGFA*	rs2010963	GG	6.1	0.010
				GC	5.4	
				CC*	40.0	
	Ligament/tendon	*VEGFA*	rs2010963	GG	9.1	0.011
				GC	5.4	
				CC*	40.0	
Post-PHV	Any injury	*IL6*	rs1800795	GG/GC	40.6	0.003
				CC*	67.6	
	Muscle injury	*IL6*	rs1800795	GG/GC	19.2	0.013
				CC*	38.2	
	Any injury	*EMILIN1*	rs2289360	**CC***	56.4	0.007
				CT	40.3	
				TT	32.8	
All ASP	Muscle	*IL6*	rs1800795	GG/GC	15.9	0.014
				CC*	31.0	

*Genotype with greater injury prevalence compared to the other genotype(s)

Figure legends

Figure 1. Relative proportions of injured and non-injured post-PHV players according to *IL6* rs1800795 GG/GC (white bars) and CC (black bars) genotype regarding any injury (A); and muscle injuries (B); *greater relative proportion of CC homozygotes injured compared to GG/GC genotypes ($p < 0.02$).

Figure 2. Relative proportions of injured (any injury) and non-injured post-PHV players according to *EMILIN1* rs2289360 CC (black bars), CT (grey bars) and TT (white bars) genotype; *greater relative proportion of CC homozygotes injured compared to CT and TT genotypes ($p < 0.05$).

Figure 3. Relative proportions of pre-PHV players suffering musculoskeletal soft-tissue injuries according to *COL5A1* rs12722 CC/CT (black bars) and TT (white bars) genotype. *greater relative proportion of CC/CT genotypes injured compared to TT homozygotes ($p = 0.018$.

Figure 1

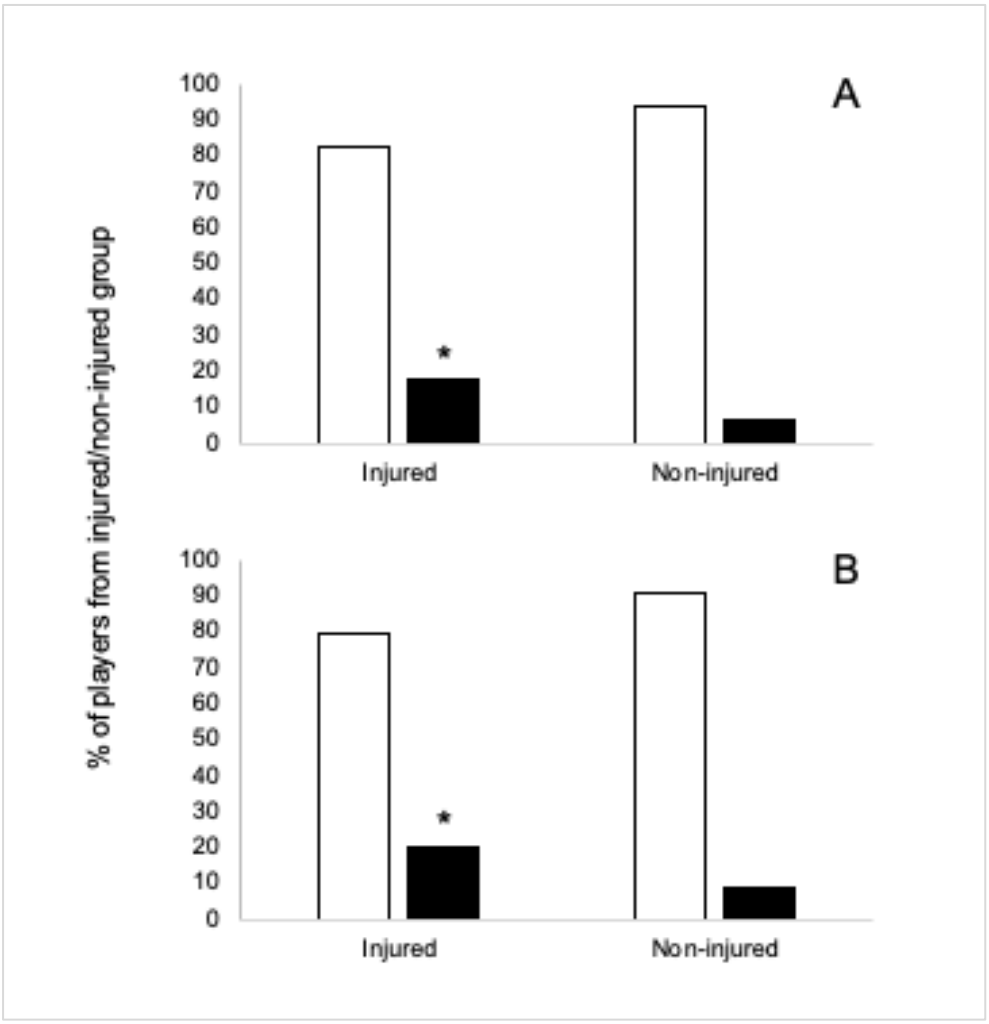

% of players from injured/non-injured group
100
90
80
70
60
50
40
30
20
10
0
A
*
Injured
Non-injured
B
*
Injured
Non-injured

Figure 2

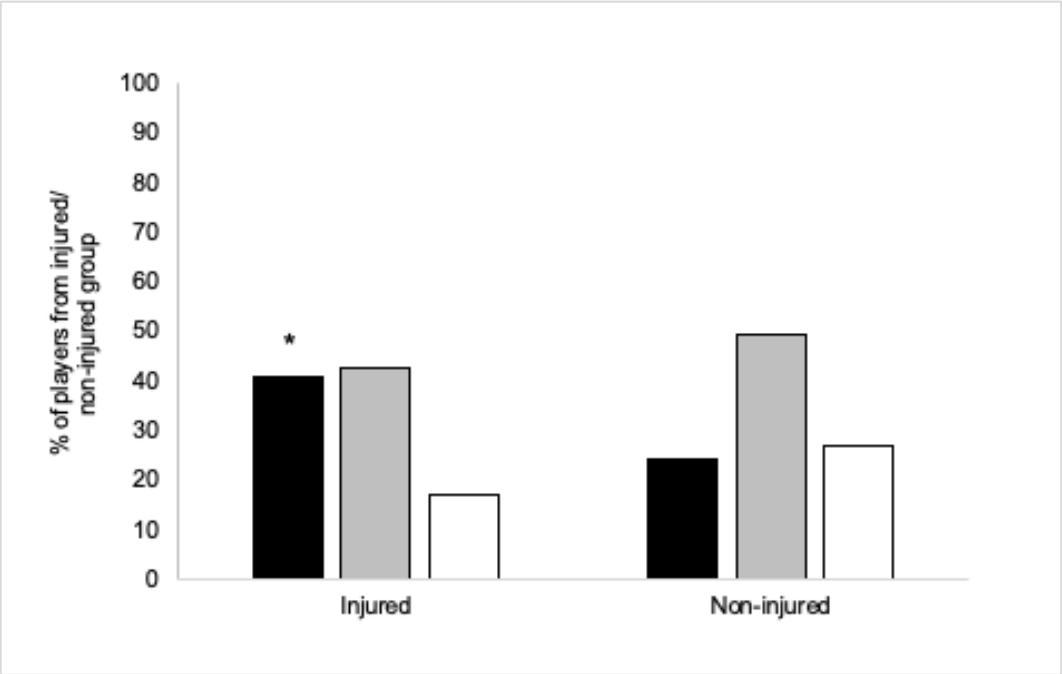
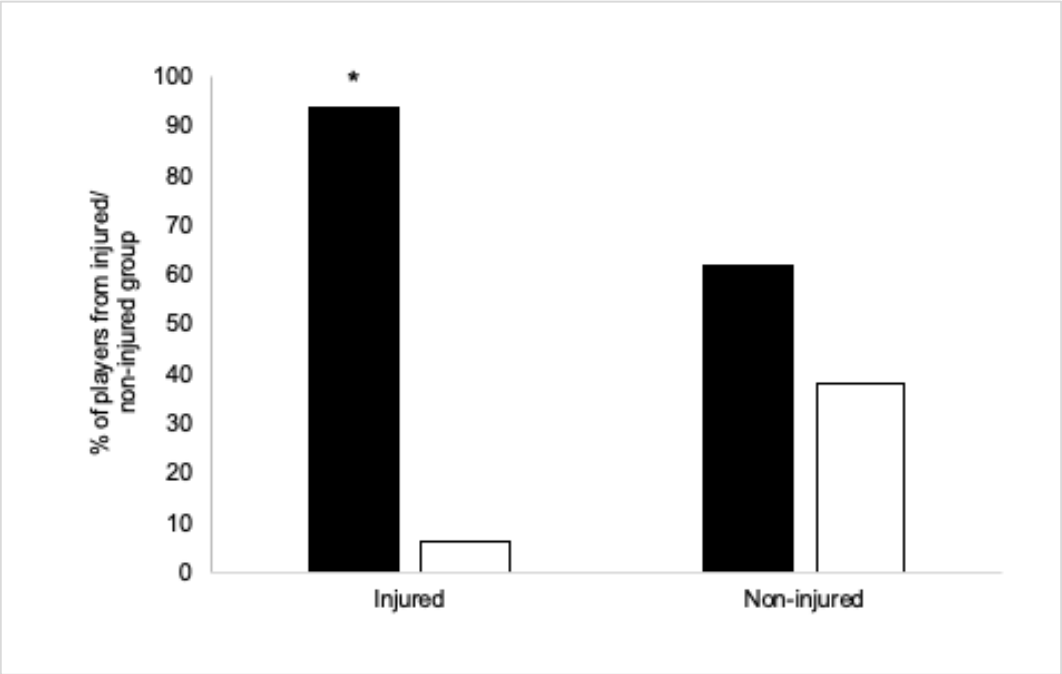

100
90
80
70
60
50
40
30
20
10
0
% of players from injured/ non-injured group
*
Injured
Non-injured

Figure 3


100
90
80
70
60
50
40
30
20
10
0
% of players from injured/ non-injured group
*
Injured
Non-injured