

Whole exome sequencing: a customised approach to exploring the genetic basis of musculoskeletal soft tissue injuries

By Andrea Gibbon

This thesis is submitted in fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in the Department of Human Biology, Faculty of Health Sciences

University of Cape Town

June 2018

Supervisors:

Assoc/Prof Alison V September¹, Prof Malcolm Collins¹ and Dr Colleen Saunders^{2,3}

¹ Division of Exercise Science and Sports Medicine, Department of Human Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

² South African National Bioinformatics Institute/SA MRC Bioinformatics Capacity Development Research Unit, University of the Western Cape, Cape Town, South Africa

³ Division of Emergency Medicine, Department of Surgery, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa



The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

I dedicate this thesis to my family for their continued support during my many years of study. To my dad Clive, mom Lesley and sister Lauren, no words are able to convey how grateful I am for your love and patience during this time. Today, I have achieved a life-long dream; something that would not have been possible without you by my side.

TABLE OF CONTENTS

APPROVAL FOR INCLUSION OF PUBLICATIONS IN THESIS	ix
PLAGIARISM DECLARATION	x
ACKNOWLEDGEMENTS.....	xi
ETHICAL APPROVAL AND FUNDING.....	xiii
THESIS-RELATED SCIENTIFIC OUTPUTS	xiv
PUBLICATIONS IN PEER-REVIEWED JOURNALS	xiv
CONFERENCE PRESENTATIONS.....	xiv
International	xiv
Local.....	xv
ABBREVIATIONS	xvi
LIST OF FIGURES	xxi
LIST OF TABLES.....	xxvi
THESIS ABSTRACT	xxxiii
BACKGROUND	xxxiii
PARTICIPANTS AND METHODS	xxxiv
RESULTS.....	xxxvi
CONCLUSION.....	xxxviii
THESIS OUTLINE	xxxix
CHAPTER 1: MUSCULOSKELETAL SOFT TISSUE INJURIES IN REVIEW	1
1.1. INTRODUCTION.....	1
1.2. TENDON AND LIGAMENT ANATOMY.....	3
1.2.1. Achilles tendon	3
1.2.2. Anterior cruciate ligament.....	6
1.3. HIERARCHICAL ORGANISATION OF TENDONS AND LIGAMENTS	10
1.4. PATHOLOGIES OF THE ACHILLES TENDON AND ANTERIOR CRUCIATE LIGAMENT ..	12
1.4.1. Achilles tendon injuries.....	12
1.4.2. Anterior cruciate ligament ruptures	26
1.5. RISK FACTORS.....	30
1.5.1. Genetic predisposition.....	32
1.6. SEQUENCING TECHNOLOGIES	55
1.6.1. First generation sequencing	55
1.6.2. Genome-wide association studies	55

1.6.3.	Next generation sequencing	57
1.7.	LITERATURE REVIEW SUMMARY AND CONCLUSIONS.....	60
1.8.	THESIS AIMS AND OBJECTIVES.....	61
CHAPTER 2: RESEARCH PARTICIPANTS AND STUDY METHODS		63
2.1.	RESEARCH PARTICIPANTS	63
2.1.1.	South African cohorts	65
2.1.2.	Australian cohort	81
2.1.3.	British cohort	83
2.2.	SAMPLE COLLECTION AND DNA EXTRACTION.....	86
2.3.	WHOLE EXOME SEQUENCING	87
2.3.1.	Selection of exemplar cases and controls	87
2.3.2.	Sample preparation	87
2.3.3.	Sequencing and post-sequencing data processing.....	88
2.3.4.	Variant filtering	89
2.3.5.	Targeted gene analysis and variant prioritisation	89
2.4.	GENOTYPING AND ALLELIC DISCRIMINATION ASSAYS	105
2.4.1.	Automated genotyping.....	105
2.4.2.	Restriction fragment length polymorphism analysis.....	105
2.5.	CASE-CONTROL GENETIC ASSOCIATION STUDIES.....	106
2.6.	STATISTICAL ANALYSES	107
2.7.	FUNCTIONAL PREDICTIONS AND ANALYSIS.....	108
CHAPTER 3: PROFILING THE <i>TNC</i> AND <i>COL27A1</i> GENES IN MUSCULOSKELETAL SOFT TISSUE INJURIES.....		111
3.1.	STUDY AIM AND OBJECTIVES.....	111
3.2.	RESULTS.....	115
3.2.1.	General analysis of the whole exome sequencing dataset	115
3.2.2.	Achilles tendon injuries.....	119
3.2.3.	Anterior cruciate ligament ruptures	135
3.3.	DISCUSSION.....	155
3.4.	CONCLUSION.....	162
CHAPTER 4: INVESTIGATION OF VARIANTS WITHIN THE <i>MMP3</i> GENE AND MUSCULOSKELETAL SOFT TISSUE INJURIES		163
4.1.	STUDY AIM AND OBJECTIVES.....	163
4.2.	RESULTS.....	166

4.2.1.	Achilles tendinopathy	166
4.2.2.	Anterior cruciate ligament ruptures	173
4.3.	DISCUSSION.....	178
4.4.	CONCLUSION.....	181
CHAPTER 5: FUNCTIONAL <i>COL1A1</i> VARIANTS AS MODIFIERS OF RISK FOR ACUTE MUSCULOSKELETAL SOFT TISSUE INJURIES		183
5.1.	STUDY AIM AND OBJECTIVES.....	183
5.2.	RESULTS.....	187
5.2.1.	Chronic musculoskeletal soft tissue injuries-Achilles tendinopathy	187
5.2.2.	Acute musculoskeletal soft tissue injuries-Achilles tendon and anterior cruciate ligament ruptures (Self-identified White cohorts)	193
5.2.3.	Acute musculoskeletal soft tissue injuries in other populations-Anterior cruciate ligament ruptures (Self-identified Coloured cohort).....	201
5.3.	DISCUSSION.....	207
5.4.	CONCLUSION.....	216
CHAPTER 6: FUNCTIONAL CHARACTERISATION OF THE GENETIC SIGNATURES UNDERPINNING INJURY SUSCEPTIBILITY		217
6.1.	INTRODUCTION.....	217
6.2.	STUDY AIM AND OBJECTIVES.....	218
6.3.	RESULTS.....	220
6.3.1.	<i>TNC</i> gene variants	220
6.3.2.	<i>COL27A1</i> gene variants	224
6.3.3.	<i>MMP3</i> gene variants.....	228
6.3.4.	<i>COL1A1</i> gene variants	232
6.4.	DISCUSSION.....	236
6.5.	CONCLUSION.....	242
CHAPTER 7: CONCLUDING REMARKS		245
7.1.	THESIS SUMMARY	245
7.2.	BIOLOGICAL SIGNIFICANCE OF FINDINGS.....	253
7.3.	STUDY LIMITATIONS AND STRENGTHS	258
7.4.	FUTURE DIRECTIONS.....	263
7.5.	CONCLUSION.....	267

APPENDICES.....	269
(I) STUDY APPROVAL AND RECRUITMENT FORMS	271
(A) Original study approval letters from the Human Research Ethics Committee.....	271
(B) Study information sheets.....	278
(C) Informed consent forms	285
(D) Medical history and sporting questionnaires	291
(E) Achilles tendinopathy clinical diagnostic criteria	304
(II) MATERIAL AND METHODS.....	305
(A) Protocol for DNA extraction from whole blood.....	305
(B) DNA quality assessment for whole exome sequencing.....	307
(C) Sanger sequencing oligonucleotide primer pair sequences.....	309
(D) Allelic discrimination assays.....	310
(III) SUPPLEMENTARY RESULTS	323
(A) Chapter 3: Profiling the <i>TNC</i> and <i>COL27A1</i> genes in musculoskeletal soft tissue injuries.....	323
(B) Chapter 4: The investigation of variants within the <i>MMP3</i> gene and musculoskeletal soft tissue injuries	330
(C) Chapter 5: Functional <i>COL1A1</i> variants as modifiers of risk for acute musculoskeletal soft tissue injuries	333
REFERENCES.....	337

APPROVAL FOR INCLUSION OF PUBLICATIONS IN THESIS

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publication(s) in my PhD thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publication(s):

- 1.) Gibbon, A, Hobbs H, van der Merwe W, Raleigh SM, Cook J, Handley CJ, et al. The *MMP3* gene in musculoskeletal soft tissue injury risk profiling: a study in two independent sample groups. J Sports Sci. 2017; 35(7):655-662.
- 2.) Gibbon A, Saunders CJ, Collins M, Gamielien J, September AV. Defining the genetic risk profile of musculoskeletal soft tissue injuries: a whole exome sequencing approach (In press).

Signature		Date	20 June 2018
	Signed by candidate		

PLAGIARISM DECLARATION

PhD thesis title:

**Whole exome sequencing: a customised approach to exploring the genetic basis of
musculoskeletal soft tissue injuries**

“This thesis has been submitted to the Turnitin module and I confirm that my supervisors have seen my report and any concerns revealed by such have been resolved with my supervisors.”

Name: Andrea Gibbon

Student number: GBBAND005

Signature		Date	20 June 2018
	Signed by candidate		

ACKNOWLEDGEMENTS

To my supervisors, Assoc/Prof Alison September, Prof Malcolm Collins and Dr Colleen Saunders, for all your support and encouragement. Your investment in my research, in addition to your patience and guidance during the months of writing can never be reciprocated but for which I will always be grateful. Alison- you have constantly encouraged my growth as a young female scientist, teaching me the importance of adherence to the scientific principal but also the importance of developing and exploring my own ideas and aspirations. You provided a secure environment in which any problem, scientific or not, could be discussed. Malcolm- regardless of how busy you were, you always found the time to provide constructive feedback and guidance, especially during the final months of writing. You have helped build my confidence and I am thankful for all the opportunities you have provided. Colleen- for all the assistance with the management and mining of the WES data. You made the “omics” world somewhat less daunting. Furthermore, for the many hours dedicated to reviewing these chapters, thank you.

To our project collaborator at the South African National Bioinformatics Institute, Assoc/Prof Junaid Gamielien, for your guidance and assistance with the more complicated matters concerning the sequencing pipeline. Prof Karl Kadler and his group at the Wellcome Trust Centre for Cell Matrix Research (United Kingdom) for providing the means to develop my skill set in a friendly and encouraging environment during my RUBICON secondment period.

To my friends and colleagues at ESSM, in addition to the administrative staff who provided a supportive and productive environment for research. In particular, Neezaam Kariem and Trevino Larry, for all the assistance in the laboratory and with all IT-related problems. To my closest friends, for all your love and support during this time.

Thank you to all the South African, Australian and British participants who donated their time and DNA samples to these studies. Thank you to Dr George Mokone, Prof Martin Schwellnus, Prof Jill Cook, Prof Christopher Handley, Dr Stuart Raleigh, Dr Mike Posthumus, Dr Masouda

Rahim, Sasha Mannion and Melony Hay, who were involved in the recruitment and/or diagnosis of participants.

I am grateful for the financial assistance received from the National Research Foundation and the University of Cape Town postgraduate funding office (Doctoral Research Scholarship, Marion Beatrice Waddel Bursary, Duncan Baxter Scholarship and UCT conference travel grant).

Finally, thank you to the examiners who kindly agreed to review this thesis. Your input is greatly appreciated and with your guidance we may improve the quality of the research outputs.

ETHICAL APPROVAL AND FUNDING

The work presented in this thesis was approved by the Human Research Ethics Committee of the Faculty of Health Sciences within the University of Cape Town, South Africa (Reference numbers: 224/2013, 176/2015, 159/2011, 086/2005, 289/2004, 370/2013 and 164/2006) and the review boards of the other international collaborating institutions.

This research was funded by the Department of Science and Technology, the National Research Foundation and the University of Cape Town.

THESIS-RELATED SCIENTIFIC OUTPUTS

PUBLICATIONS IN PEER-REVIEWED JOURNALS

1. Gibbon A, Hobbs H, van der Merwe W, Raleigh SM, Cook J, Handley CJ, et al. The *MMP3* gene in musculoskeletal soft tissue injury risk profiling: a study in two independent sample groups. *J Sports Sci.* 2017; 35(7):655-62.
2. Gibbon A, Saunders CJ, Gamielien J, Collins M, September AV. Defining the molecular signatures of Achilles tendinopathy and anterior cruciate ligament ruptures: a whole exome sequencing approach (In press).

CONFERENCE PRESENTATIONS

International

1. September AV, Gibbon A, Saunders CJ, Gamielien J, Collins M. Defining the genetic risk profiles of musculoskeletal soft tissue injuries: a whole exome sequencing approach. The 23rd Annual Congress of the European Collage of Sports Science 2018, Dublin, Ireland (Oral presentation).
2. Gibbon A, Kadler K, Collins M. The role of genetic factors, the circadian clock and signalling pathways in tendinopathies. The 44th European Calcified Tissues Society Congress 2017. Presented at the Rubicon Annual Progress meeting. Marie Curie EU-funded Research and Innovation Staff Exchange Programme, 2017, Salzburg, Austria (Oral presentation).
3. September AV, Gibbon A, Willard K, Dlamini S, Saunders CJ, Gamielien J, Collins M. Decoding musculoskeletal soft tissue injuries. International Congress of Human Genetics 2016, Kyoto, Japan (Poster presentation).

4. Gibbon A, Saunders CJ, Gamielien J, Collins M, September AV. Pieces to the puzzle: identifying variants associated with the risk of musculoskeletal soft tissue injuries using whole exome sequencing. The 4th International Scientific Tendinopathy Symposium 2016, Cape Town, South Africa (Poster presentation).
5. Gibbon A, Hobbs H, van der Merwe W, Posthumus M, Collins M, September AV. Variants in the *MMP3* gene are not associated with the risk of acute ACL ruptures. International Congress of Human Genetics 2015, Glasgow, Scotland (Poster presentation).

Local

1. Gibbon A, Saunders CJ, Gamielien J, Collins M, September AV. Defining gene specific profiles of musculoskeletal soft tissue injuries: a whole exome sequencing approach. Faculty of Health Science Postgraduate Research Open Day 2017, Cape Town, South Africa (Poster presentation).
2. Gibbon A, Willard K, Dlamini S, Saunders CJ, Gamielien J, Collins M, September AV. Decoding musculoskeletal soft tissue injuries. Pharmacogenetics and Personalised Medicine Conference 2016, Cape Town, South Africa (Poster presentation).
3. Gibbon A, Hobbs H, van der Merwe W, Posthumus M, Collins M, September AV. Variants in the *MMP3* gene are not associated with the risk of acute ACL ruptures. Faculty of Health Science Postgraduate Research Open Day 2015, Cape Town, South Africa (Poster presentation).

ABBREVIATIONS

A:	Adenine
<i>ACAN</i> :	Gene encoding aggrecan
ACL:	Anterior cruciate ligament
Ala (A):	Alanine
AM:	Anteromedial bundle of the ACL
ANOVA:	Analysis of variance
Arg (R):	Arginine
Asp (D):	Aspartic acid
AT:	Achilles tendinopathy
ATR:	Achilles tendon rupture
AUS:	Australia
AUS-CON:	Group of Australian control participants
AUS-TEN:	Group Australian participants with Achilles tendinopathy
BAM:	Binary version of sequencing alignment map (SAM) files
bp:	Base pairs
<i>BGN</i> :	Gene encoding biglycan
BMD:	Bone mineral density
BMI:	Body mass index
BORG:	BioOntological Relationship Graph semantic database
C:	Cytosine
<i>CASP8</i> :	Gene encoding caspase 8
<i>CASP9</i> :	Gene encoding caspase 9
Chr:	Chromosome
CI:	Confidence interval
CL:	Cruciate ligament
<i>COL1A1</i> :	Gene encoding α_1 chain (I) collagen
<i>COL1A2</i> :	Gene encoding α_2 chain (I) collagen
<i>COL3A1</i> :	Gene encoding α_1 chain (III) collagen
<i>COL5A1</i> :	Gene encoding α_1 chain (V) collagen
<i>COL11A1</i> :	Gene encoding α_1 chain (XI) collagen
<i>COL11A2</i> :	Gene encoding α_2 chain (XI) collagen
<i>COL12A1</i> :	Gene encoding α_1 chain (XII) collagen
<i>COL27A1</i> :	Gene encoding α_1 chain (XXVII) collagen
CNV:	Copy number variation
COOH:	Carboxyl terminal
D':	Measure of linkage disequilibrium
3D:	Three dimensional
dATP:	Deoxyadenosine triphosphate
dCTP:	Deoxycytidine triphosphate
dGTP:	Deoxyguanosine triphosphate
<i>DCN</i> :	Gene encoding decorin

DNA:	Deoxyribonucleic acid
dTTP:	Deoxythymidine triphosphate
ECM:	Extracellular matrix
EDTA:	Ethylenediaminetetraacetic acid
EGFL:	Epidermal growth factor-like repeats
eQTL:	Expression quantitative trait locus
<i>FBN2</i> :	Gene encoding fibrillin 2
FNIII:	Fibronectin type III domains
<i>FAS</i> :	Gene encoding Fas cell surface death receptor
<i>FASLG</i> :	Gene encoding Fas ligand
G:	Guanine
GAG:	Glycosaminoglycan
<i>GDF5</i> :	Gene encoding growth differentiation factor 5
Gln (Q):	Glutamine
Glu (E):	Glutamic acid
GT:	Guanine-thymine dinucleotide repeat
GWAS:	Genome-wide association studies
His (H):	Histidine
HSPG:	Heparan sulfate proteoglycan
<i>HSPG2</i> :	Gene encoding heparan sulfate proteoglycan 2 (Perlecan)
HWE:	Hardy-Weinberg equilibrium
ICD:	International Classification of Diseases
Ile (I):	Isoleucine
ILK:	Integrin-linked kinase
IL1B:	Interleukin 1 beta
<i>IL1B</i> :	Gene encoding interleukin 1 beta
<i>IL1RN</i> :	Gene encoding interleukin 1 receptor antagonist
IL6:	Interleukin 6
<i>IL6</i> :	Gene encoding interleukin 6
IL8:	Interleukin 8
Indel:	Insertion/deletion polymorphism
kb:	Kilobase (1000 base pairs)
<i>KDR</i> :	Gene encoding the kinase insert domain receptor
LCL:	Lateral collateral ligament
LD:	Linkage disequilibrium
Leu (L):	Leucine
<i>LUM</i> :	Gene encoding lumican
Lys (K):	Lysine
MKL1/MAL:	Megakaryoblastic leukemia-1/Megakaryocytic acute leukemia
MAPK:	Mitogen-activated protein kinase

MCL:	Medial collateral ligament
MGA:	Middle genicular artery
miRNA:	Micro ribonucleic acid
<i>MIR455</i> :	Gene encoding microRNA 455
<i>MIR608</i> :	Gene encoding microRNA 608
MMP:	Matrix metalloproteinase
<i>MMP</i> :	Gene encoding a matrix metalloproteinase
<i>MMP1</i> :	Gene encoding matrix metalloproteinase 1
MMP1:	Matrix metalloproteinase 1 (Collagenase 1)
<i>MMP3</i> :	Gene encoding matrix metalloproteinase 3
MMP3:	Matrix metalloproteinase 3 (Stromelysin 1)
<i>MMP10</i> :	Gene encoding matrix metalloproteinase 10
<i>MMP12</i> :	Gene encoding matrix metalloproteinase 12
MMP13:	Matrix metalloproteinase 13 (Collagenase 3)
MRI:	Magnetic resonance imaging
mRNA:	Messenger ribonucleic acid
MTJ:	Myotendinous junction
MUSCLE_INIT_B:	Muscle initiator nuclear protein
NCBI:	National Centre for Biotechnology Information
NFKB:	Nuclear factor kappa B
NGS:	Next generation sequencing
NH ₂ :	Amino terminal
NON-SYN:	Non-synonymous variant
<i>NOS</i> :	Gene encoding nitric oxide synthase
OR:	Odds ratio
OTJ:	Osteotendinous junction
P:	Probability
p:	Short arm of chromosome
PCL:	Posterior cruciate ligament
PCR:	Polymerase chain reaction
PDGF:	Platelet derived growth factor
Phe (F):	Phenylalanine
PL:	Posterolateral bundle of the ACL
POLR2A:	RNA polymerase II subunit A
pQTL:	Protein quantitative trait locus
q:	Long arm of chromosome
r ² :	Correlation coefficient estimate
RCT:	Rotator cuff tear
RFLP:	Restriction fragment length polymorphism
RhoA:	RhoA-actin signalling pathway
ROCK:	Rho-dependent protein kinase

SA:	South Africa
SAC:	South African Coloured
SAC-ACL:	Group of South African Coloured participants with ACL ruptures
SAC-CON:	Group of South African Coloured control participants
SAC-NON:	Group South African Coloured participants with non-contact ACL ruptures
SAM:	Sequencing alignment map file
<i>SAP30BP</i> :	Gene encoding the SAP30 binding protein
<i>SASH1</i> :	Gene encoding SAM and SH3 domain containing 1
SAW:	South African White
SAW-ACL:	Group of South African White participants with ACL ruptures
SAW-CON ^{ACL} :	Group of South African White control participants in the ACL rupture cohort
SAW-CON ^{AT} :	South African White control participants in the Achilles tendon injury cohort
SAW-CON ^{SEQ} :	Sample of SAW-CON ^{AT} participants selected for whole exome sequencing
SAW-NON:	South African White participants with non-contact ACL ruptures
SAW-RUP:	Group of South African White participants with acute Achilles tendon ruptures
SAW-TEN:	Group of South African White participants with Achilles tendinopathy
SAW-TEN ^{SEQ} :	Sample of SAW-TEN participants selected for whole exome sequencing
<i>SEMA5B</i> :	Gene encoding semaphorin 5B
SIRE:	Stromelysin interleukin 1 responsive element
SNP:	Single nucleotide polymorphism
<i>SORCS2</i> :	Gene encoding sortilin related VPS10 domain containing receptor 2
Sp1:	Specificity protein 1 transcription factor
SRF:	Serum response factor
STREGA:	STrengthening the REporting of Genetic Association studies
STROBE:	STrengthening the REporting of OBservational Studies in Epidemiology
SYN:	Synonymous variant
T:	Thymine
TA:	Tenascin assembly domain
TATA:	TATA box
TGFB:	Transforming growth factor beta
Thr (T):	Threonine
TIMP:	Tissue inhibitor of metalloproteinase
<i>TIMP2</i> :	Gene encoding TIMP metalloproteinase inhibitor 2
<i>TNC</i> :	Gene encoding tenascin-C
TNF:	Tumor necrosis factor
<i>TNFRSF1A</i> :	Gene encoding TNF receptor superfamily member 1A
UK:	United Kingdom/British
UK-CON:	Group of British control participants
UK-RUP:	Group of British participants with acute Achilles tendon ruptures
UK-TEN:	Group of British participants with Achilles tendinopathy

USA:	United States of America
UTR:	Untranslated region
Val (V):	Valine
vcf:	Variant call format
<i>VEGFA</i> :	Gene encoding vascular endothelial growth factor A
WES:	Whole exome sequencing
WGS:	Whole genome sequencing

LIST OF FIGURES

Figure 1.1: Basic anatomy of the lower limb	4
Figure 1.2: Basic anatomy of the ankle joint	5
Figure 1.3: Basic anatomy of the knee	7
Figure 1.4: Orientation of the anterior cruciate ligament bundles	8
Figure 1.5: The hierarchical structure of a tendon	11
Figure 1.6: The stress-strain relationship of tendons	18
Figure 1.7: The 'Iceberg theory' of the pathogenesis of tendinopathy	23
Figure 1.8: The pathological continuum model of tendinopathy as proposed by Cook and Purdam et al. (2009)	25
Figure 1.9: Proposed scheme for the classification of ACL injuries by type of contact, as devised by the American Orthopaedic Society for Sports Medicine. Figure modified from Marshall et al. (2007)	29
Figure 1.10: Intrinsic and extrinsic factors proposed to influence the risk of chronic Achilles tendinopathy and anterior cruciate ligament ruptures	30
Figure 1.11: Multifactorial model of the aetiology of sports-related injuries as proposed by Meeuwisse (1994)	31
Figure 1.12: Gene encoding the tenascin-C glycoprotein	38
Figure 1.13: Structure of the tenascin-C hexamer	40
Figure 1.14: Proposed mechanism by which the mechano-induction of tenascin-C inhibits fibronectin-mediated cell spreading	43
Figure 1.15: Gene encoding the α_1 chain of type XXVII collagen	45
Figure 1.16: Matrix metallopeptidase gene cluster on chromosome 11q22	47
Figure 1.17: Gene encoding the MMP3 (Stromelysin 1) endopeptidase	48
Figure 1.18: Gene encoding the α_1 chain of type I collagen	53
Figure 2.1: The customised, tiered filtering strategy used to prioritise loci for interrogation from the whole exome sequencing dataset	91

Figure 3.1: Map of <i>TNC</i> gene displaying candidate variants prioritised for interrogation from the whole exome sequencing dataset in relation to variants previously associated with the risk of musculoskeletal soft tissue injuries	113
Figure 3.2: Map of <i>COL27A1</i> gene displaying candidate variants prioritised for interrogation from the whole exome sequencing dataset in relation to variants previously investigated for their potential association with the risk of musculoskeletal soft tissue injuries.....	114
Figure 3.3: Distribution of variants identified by whole exome sequencing within genes previously associated with the risk of Achilles tendinopathy and other musculoskeletal soft tissue injuries	117
Figure 3.4: Distribution of variants identified by whole exome sequencing within genes proposed as plausible candidates for Achilles tendinopathy using the BioOntological Relationship Graph (BORG) semantic database	118
Figure 3.5: Inferred <i>TNC</i> haplotypes in the South African White (SAW)-, British (UK)- and Combined-Achilles tendon injury cohorts.....	130
Figure 3.6: Inferred <i>COL27A1</i> haplotypes in the South African White (SAW)-, British (UK)- and combined-Achilles tendon injury cohorts	131
Figure 3.7: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>TNC</i> and <i>COL27A1</i> variants in the South African White (SAW)-Achilles tendon injury cohort.....	133
Figure 3.8: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>TNC</i> and <i>COL27A1</i> variants in the British (UK)-Achilles tendon injury cohort	134
Figure 3.9: Genotype frequency distributions of the <i>TNC</i> rs2104772 (T>A) variant in the South African White (SAW)-ACL rupture cohort	139
Figure 3.10: Genotype frequency distributions of the <i>COL27A1</i> rs2241671 (G>A) variant in the South African Coloured (SAC)-ACL rupture cohort	144
Figure 3.11: Inferred <i>TNC</i> haplotypes in the South African White (SAW)-ACL rupture cohort	146
Figure 3.12: Inferred <i>COL27A1</i> haplotypes in the South African White (SAW)-ACL rupture cohort	147
Figure 3.13: Inferred <i>TNC</i> haplotypes in the South African Coloured (SAC)-ACL rupture cohort	149
Figure 3.14: Inferred <i>COL27A1</i> haplotypes in the South African Coloured (SAC)-ACL cohort	151

Figure 3.15: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>TNC</i> and <i>COL27A1</i> variants in the South African White (SAW)-ACL rupture cohort.....	153
Figure 3.16: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the four genotyped <i>TNC</i> and three <i>COL27A1</i> variants in the South African Coloured (SAC)-ACL rupture cohort	154
Figure 3.17: Inferred risk-associated haplotypes mapping to the <i>TNC</i> genomic interval.....	157
Figure 4.1: Map of <i>MMP3</i> gene displaying candidate variants identified by whole exome sequencing in relation to variants previously associated with the risk of musculoskeletal soft tissue injuries	165
Figure 4.2: Inferred <i>MMP3</i> haplotypes in the Australian (AUS)-, South African White (SAW)- and combined-Achilles tendon injury cohorts	170
Figure 4.3: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>MMP3</i> variants in the Australian (AUS)- and South African White (SAW)-Achilles tendon injury cohorts	172
Figure 4.4: Inferred <i>MMP3</i> haplotypes in the South African White (SAW)-ACL rupture cohort	176
Figure 4.5: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>MMP3</i> variants in the South African White (SAW)-ACL rupture cohort.....	177
Figure 5.1: Map of <i>COL1A1</i> gene displaying variants previously associated with the risk of acute musculoskeletal soft tissue injuries and prioritised for further interrogation in the current study	186
Figure 5.2: Inferred <i>COL1A1</i> haplotypes in the British (UK)-, South African White (SAW)- and combined-Achilles tendon injury cohorts	191
Figure 5.3: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>COL1A1</i> variants in the British (UK)- and South African White (SAW)-Achilles tendon injury cohorts.....	192
Figure 5.4: Inferred <i>COL1A1</i> haplotypes in the combined-Achilles tendon injury cohort	198
Figure 5.5: Inferred <i>COL1A1</i> haplotypes in the South African White (SAW)-ACL rupture cohort	199
Figure 5.6: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>COL1A1</i> variants in the South African White (SAW)- ACL rupture cohort.....	200
Figure 5.7: Inferred <i>COL1A1</i> haplotypes in the South African Coloured (SAC)-ACL rupture cohort	205

Figure 5.8: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped <i>COL1A1</i> variants in the South African Coloured (SAC)- ACL rupture cohort	206
Figure 5.9: Genotype frequency distributions of the <i>COL1A1</i> (A) rs1107946 (G>T) and (B) rs1800012 (G>T) variants in the Asian, African, European and South African Coloured populations	213
Figure 6.1: Overview of <i>TNC</i> variants genotyped in the study described in chapter 3	220
Figure 6.2: Functional annotation of the risk-associated <i>TNC</i> (A) rs1061494 (T>C) and (B) rs2104772 (T>A) variants using the 3DSNP annotation tool	222
Figure 6.3: Sfold predicted pre-mRNA secondary structures immediately surrounding the <i>TNC</i> (A) rs1061494 (T>C) and (B) rs2104772 (T>A) variants	223
Figure 6.4: Overview of <i>COL27A1</i> variants genotyped in the study described in chapter 3	224
Figure 6.5: Functional annotation of the risk-associated <i>COL27A1</i> (A) rs2567705 (A>G), (B) rs2241671 (G>A) and (C) rs2567706 (A>T) variants using the 3DSNP annotation tool	226
Figure 6.6: Sfold predicted pre-mRNA secondary structures immediately surrounding the <i>COL27A1</i> (A) rs2567706 (A>G), (B) rs2241671 (G>A) and (C) rs2567705 (A>T) variants	227
Figure 6.7: Overview of <i>MMP3</i> variants genotyped in the study described in chapter 4	229
Figure 6.8: Functional annotation of the risk-associated <i>MMP3</i> (A) rs3025058 (5A>6A), (B) rs679620 (A>G), (C) rs591058 (T>C) and (D) rs650108 (G>A) variants using the 3DSNP annotation tool	230
Figure 6.9: Sfold predicted pre-mRNA secondary structures immediately surrounding the <i>MMP3</i> (A) rs3025058 (5A>6A), (B) rs679620 (A>G), (C) rs591058 (T>C) and (D) rs650108 (G>A) variants	231
Figure 6.10: Overview of <i>COL1A1</i> variants genotyped in the study described in chapter 5	232
Figure 6.11: Functional annotation of the risk-associated <i>COL1A1</i> (A) rs1107946 (G>T) and (B) rs1800012 (G>T) variants using the 3DSNP annotation tool	234
Figure 6.12: Sfold predicted pre-mRNA secondary structures immediately surrounding the <i>COL1A1</i> (A) rs1107946 (G>T) and (B) rs1800012 (G>T) variants	235
Figure 6.13: Proposed mechanism of DNA-protein binding and the interaction of variants in the 5' flank of the <i>COL1A1</i> gene regulated by chromatin looping	241
Figure 7.1: Proposed interactions between proteins encoded by the genes investigated in the studies presented in this thesis	256

Figure IIB.1: Quality assessment of DNA samples obtained from the 20 participants selected for whole exome sequencing	308
Figure IID.1: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs1061494) on the StepOnePlus [™] Real-Time PCR System	310
Figure IID.2: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs1138545) on the QuantStudio [™] 3 Real-Time PCR system.....	311
Figure IID.3: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs2104772) on the StepOnePlus [™] Real-Time PCR System	312
Figure IID.4: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs1061495) on the StepOnePlus [™] Real-Time PCR System.....	313
Figure IID.5: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs2567706) on the StepOnePlus [™] Real-Time PCR System	314
Figure IID.6: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs2241671) on the QuantStudio [™] 3 Real-Time PCR system.....	315
Figure IID.7: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs2567705) on the QuantStudio [™] 3 Real-Time PCR system.....	316
Figure IID.8: Typical 3% agarose gel used to discriminate fragments generated after digestion with the Tth111I endonuclease and electrophoresed at 120V for 40 minutes	317
Figure IID.9: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs679620) on the StepOnePlus [™] Real-Time PCR System	318
Figure IID.10: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs591058) on the StepOnePlus [™] Real-Time PCR System	319
Figure IID.11: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs650108) on the StepOnePlus [™] Real-Time PCR System.....	320
Figure IID.12: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs1107946) on the StepOnePlus [™] Real-Time PCR System.....	321
Figure IID.13: Allelic discrimination plot using the Taqman [®] SNP Genotyping Assay (rs1800012) on the StepOnePlus [™] Real-Time PCR System.....	322

LIST OF TABLES

Table 1.1: Country-specific incidence rates for acute Achilles tendon ruptures.....	16
Table 1.2: Variants within genes encoding extracellular matrix proteins and cell-signalling factors previously associated with the risk of chronic Achilles tendinopathy and/or anterior cruciate ligament ruptures using case-control genetic association studies	33
Table 2.1: Overview of the (i) sequencing sample set and (ii) five genotyping cohorts used to investigate the genetic basis of chronic Achilles tendinopathy and anterior cruciate ligament ruptures using a case-control genetic association study design	64
Table 2.2: Descriptive characteristics of the South African White participants in the control (SAW-CON ^{AT}), Achilles tendinopathy (SAW-TEN) and acute Achilles tendon rupture (SAW-RUP) groups.....	69
Table 2.3: Self-reported sports participation by the South African White participants in the control (SAW-CON ^{AT}), Achilles tendinopathy (SAW-TEN) and Achilles tendon rupture (SAW-RUP) groups	70
Table 2.4: Descriptive characteristics of the South African White participants in the control (SAW-CON ^{ACL}) group, anterior cruciate ligament rupture (SAW-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAW-NON)	73
Table 2.5: Self-reported sports participation according to type of activity for the South African White participants in the control (SAW-CON ^{ACL}) and anterior cruciate ligament rupture (SAW-ACL) groups	74
Table 2.6: Self-reported history of musculoskeletal soft tissue injuries for the South African White participants in the control (SAW-CON ^{ACL}) group, anterior cruciate ligament rupture (SAW-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAW-NON).....	76
Table 2.7: Descriptive characteristics of the South African Coloured participants in the control (SAC-CON) group, anterior cruciate ligament rupture (SAC-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAC-NON)	78
Table 2.8: Self-reported sports participation according to type of activity for the South African Coloured participants in the control (SAC-CON) and anterior cruciate ligament (SAC-ACL) rupture groups	79
Table 2.9: Self-reported history of musculoskeletal soft tissue injuries for the South African Coloured participants in the control (SAC-CON) group, anterior cruciate ligament rupture (SAC-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAC-NON)	80

Table 2.10: Descriptive characteristics of the Australian White participants in the control (AUS-CON) and Achilles tendinopathy (AUS-TEN) groups	82
Table 2.11: Descriptive characteristics of the British White participants in the control (UK-CON), Achilles tendinopathy (UK-TEN) and acute Achilles tendon rupture (UK-RUP) groups.....	85
Table 2.12: Overview of variants prioritised for further interrogation in chapters 3 to 5 of this thesis. Candidates included new variants identified from the whole exome sequencing dataset using a tiered filtering strategy or were variants previously proposed as modifiers of risk	92
Table 2.13: Identification of tagged loci from the list of <i>TNC</i> candidates generated by whole exome sequencing.....	94
Table 2.14: The proportion of recruited participants successfully genotyped for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts, in addition to the South African White (SAW)-and South African Coloured (SAC)-ACL rupture cohorts	96
Table 2.15: Identification of tagged loci from the list of <i>COL27A1</i> candidates generated by whole exome sequencing.....	97
Table 2.16: The proportion of recruited participants successfully genotyped for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts, in addition to the South African (SAW)- and South African Coloured (SAC)-ACL rupture cohorts.	98
Table 2.17: Identification of tagged loci from the list of <i>MMP3</i> candidates generated by whole exome sequencing.....	100
Table 2.18: The proportion of recruited participants successfully genotyped for the <i>MMP3</i> rs3025058 (5A>6A) rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the Australian (AUS)- and South African White (SAW)-Achilles tendon injury cohorts, in addition to the South African White (SAW)-ACL rupture cohort.....	102
Table 2.19: The proportion of recruited participants successfully genotyped for the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts, in addition to the South African White (SAW)- and South African Coloured (SAC)-ACL rupture cohorts.....	104
Table 3.1: Overview of variants captured across the twenty sequenced exomes (SAW-CON ^{SEQ} and SAW-TEN ^{SEQ}) using the Illumina HiSeq 2000/2500 system	116

Table 3.2: Genotype associations with the <i>TNC</i> rs1061494 (T>C), rs11384545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in addition to the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants and the physiological characteristics of participants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts.....	121
Table 3.3: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the South African White (SAW)-, British (UK)- and combined-Achilles tendinopathy cohorts	123
Table 3.4: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the South African White (SAW)-, British (UK)- and combined-Achilles tendinopathy cohorts.....	124
Table 3.5: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) and Achilles tendon rupture (RUP) groups in the combined-Achilles tendon injury cohort	126
Table 3.6: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) and Achilles tendon rupture (RUP) groups in the combined-Achilles tendon injury cohort.....	127
Table 3.7: Genotype associations with the <i>TNC</i> rs1061494 (T>C), rs11384545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants, in addition to the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants and the physiological characteristics of recruited participants in the South African White (SAW)- and South African Coloured (SAC)-ACL rupture cohorts.....	136
Table 3.8: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL rupture cohort	138
Table 3.9: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL rupture cohort	140

Table 3.10: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African Coloured (SAC)-ACL rupture cohort.....	142
Table 3.11: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African Coloured (SAC)-ACL rupture cohort	143
Table 4.1: Genotype associations with the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants and the physiological characteristics of participants in the Australian (AUS)- and South African White (SAW)-Achilles tendon injury cohorts.....	167
Table 4.2: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the Australian (AUS)-, South African White (SAW) - and combined-Achilles tendon injury cohorts	169
Table 4.3: Genotype associations with the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants and the physiological characteristics of participants in the South African White (SAW)-ACL rupture cohort	173
Table 4.4: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of participants reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL cohort	175
Table 5.1: Genotype associations with the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants and the physiological characteristics of participants in the British (UK)- and South African White (SAW)-Achilles tendon injury cohorts	188
Table 5.2: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the British (UK)-, South African White (SAW)- and combined-Achilles tendon injury cohorts.....	189
Table 5.3: Genotype associations with the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants and the physiological characteristics of participants in the South African White (SAW)-ACL rupture cohorts	194

Table 5.4: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) and Achilles tendon rupture (RUP) groups in the combined-Achilles tendon injury cohort.....	195
Table 5.5: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T), rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL cohort.....	196
Table 5.6: Genotype associations with the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants and the physiological characteristics of participants in the South African Coloured (SAC)-ACL rupture cohorts.....	202
Table 5.7: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T), rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African Coloured (SAC)-ACL rupture cohort	203
Table 5.8: Comparative and combined analyses of the genotype and allele frequency distributions of the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants from the South African White-ACL rupture (SAW) cohort and the previously published Polish-ACL and Swedish-cruciate ligament (CL) rupture cohorts.....	210
Table 6.1: A brief overview of the <i>in silico</i> tools used to functionally characterise the genetic signatures associated with injury predisposition.....	219
Table 6.2: Predicted functional effects of the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants	223
Table 6.3: Predicted functional effects of the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants	228
Table 7.1: A summary of the associations between the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 3 are presented	248
Table 7.2: A summary of the associations between the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 3 are presented	249
Table 7.3: A summary of the associations between the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 4 are presented	250

Table 7.4: A summary of the associations between the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 5 are presented	251
Table IIB. 1: Concentration and absorbance measurements made on a spectrophotometer used to assess the quantity and purity of DNA samples obtained from the 20 participant selected for whole exome sequencing on the Illumina HiSeq 2000/2500 system.....	307
Table IIC.1: Oligonucleotide primer pair sequences used to validate the new candidate variants identified by whole exome sequencing.....	309
Table IIIA.1: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort.....	324
Table IIIA.2: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort.....	325
Table IIIA.3: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort.....	326
Table IIIA.4: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort.....	327
Table IIIA.5: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>TNC</i> rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African Coloured (SAC)-ACL rupture cohort.....	328

Table IIIA.6: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL27A1</i> rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African Coloured (SAC)-ACL rupture cohort.....	329
Table IIIB.1: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort.....	331
Table IIIB.2: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>MMP3</i> rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort.....	332
Table IIIC.1: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort.....	334
Table IIIC.2: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort.....	335
Table IIIC.3: Genotype, minor allele frequencies and Hardy-Weinberg equilibrium (HWE) probabilities for the <i>COL1A1</i> rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African Coloured (SAC)-ACL rupture cohort.....	336

THESIS ABSTRACT

BACKGROUND

Several DNA sequence variants have previously been associated with the risk of musculoskeletal soft tissue injuries, suggesting a role for genetics in the aetiology of common sporting injuries such as chronic Achilles tendinopathy (AT) and anterior cruciate ligament (ACL) ruptures. Genetic risk modifiers have primarily been identified using a hypothesis driven candidate gene approach. However, the ability to identify all risk-conferring variants using this approach alone is limited. Therefore, the primary aim of this thesis was to further define the molecular signatures of musculoskeletal soft tissue injuries mapping to specific genomic intervals encoding several structural and regulatory components of the extracellular matrix (ECM). Genes encoding the tenascin-C (*TNC*) glycoprotein (9q33.1) and the α 1 chain of type XXVII collagen (*COL27A1*, 9q32), as well as matrix metalloproteinase 3 (*MMP3*, 11q22) and the α 1 chain of type I collagen (*COL1A1*, 17q21.33), have previously been associated with the risk of injury and were therefore prioritised for further interrogation. Previously associated variants within these regions and/or new candidate variants identified by whole exome sequencing (WES) and prioritised through the application of a customised, tiered filtering strategy, were genotyped in several previously recruited, self-identified White Achilles tendon injury and ACL rupture cohorts. The second aim of this study was to determine whether the observed risk-associated signatures in the self-identified White cohorts were similar to those underpinning injury in an ancestrally admixed sample, using ACL ruptures in a South African Coloured cohort as the phenotypic model.

The specific objectives of this thesis were:

- To select well-phenotyped participants (10 asymptomatic controls and 10 exemplar AT cases) for sequencing using an extended whole exome sequencing platform (Chapter 2).

- The development and application of a reusable bioinformatics analysis pipeline involving a tiered filtering strategy to select candidates for interrogation from the list of variants identified by WES (Chapter 2). The *TNC* (chr9q33.1), *COL27A1* (chr9q32), *MMP3* (chr11q22) and *COL1A1* (chr17q21.33) genes were prioritised for further interrogation using this approach.
- To test the association between the selected candidate variants and the risk of chronic AT and acute ACL ruptures using a case-control genetic association study design (Chapters 3 to 5). The candidates selected from the list of variants identified by WES included *TNC* rs1061494 (T>C), rs2104772 (T>A) and rs1061495 (T>C) and *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T). In addition, several variants previously associated with the risk of injury including *TNC* rs1138545 (C>T), *MMP3* rs3025058 (5A>6A), rs679620 (G>A), rs591058 (C>T) and rs650108 (G>A) and *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) were prioritised for interrogation.
- To functionally annotate the prioritised variants using a host of *in silico* bioinformatic analysis tools (Chapter 6).

RESEARCH PARTICIPANTS AND METHODS

Whole exome sequencing and data processing

Ten asymptomatic controls and ten chronic AT cases were selected for sequencing. Controls were all older than 47 years of age, were physically active and had not reported any previous tendon or ligament injuries. Cases were all younger than 35 years of age, had suffered bilateral AT and/or reported several Achilles tendon injuries. Paired-end WES was performed on the Illumina HiSeq 2000/2500 system at 30X coverage. A customised tiered filtering strategy was developed to screen for potentially functional candidates. All candidate variants were confirmed using Sanger Sequencing and genotyped, together with the other prioritised variants in the larger injury cohorts.

Case-control genetic association studies

Achilles tendon injury cohorts: Three independent cohorts were recruited from South Africa (SA), Australia (AUS) and the United Kingdom (UK). The South African White (SAW)-Achilles tendon injury cohort consisted of 165 controls (SAW-CON^{AT}), 123 cases with chronic AT (SAW-TEN) and 47 cases with acute Achilles tendon ruptures (SAW-RUP). The UK-Achilles tendon injury cohort consisted of 130 controls (UK-CON), 87 cases with chronic AT (UK-TEN) and 35 cases with acute Achilles tendon ruptures (UK-RUP). The AUS-Achilles tendon injury cohort included 210 controls (AUS-CON) and 85 cases with chronic AT (AUS-TEN).

Anterior cruciate ligament rupture cohorts: The first ACL rupture cohort consisted of 232 control participants (SAW-CON^{ACL}) and 234 cases with surgically diagnosed ACL ruptures (SAW-ACL), of which 135 were reportedly non-contact in mechanism (SAW-NON). All participants in this group were self-identified to be of SAW ancestry. The participants in the second ACL rupture cohort were of mixed ancestry and self-identified as being South African Coloured (SAC). This group consisted of 100 controls (SAC-CON) and 97 participants with surgically diagnosed ACL ruptures (SAC-ACL), of which 50 were reportedly non-contact in mechanism (SAC-NON).

The *TNC* and *COL27A1* genomic intervals were explored through the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants and the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the SAW- and UK-Achilles tendon injury cohorts (Section 3.2.2), in addition to the SAW- and SAC-ACL rupture cohorts (Section 3.2.3). The *MMP3* locus was explored using the previous risk-associated rs3025058 (5A>6A), rs679620 (G>A), rs591058 (C>T) and rs650108 (G>A) variants in the AUS-Achilles tendon injury (Section 4.2.1) and SAW-ACL rupture (Section 4.2.2) cohorts. Chapter 5 explored the *COL1A1* locus using the previous risk-associated *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the SAW- and UK-Achilles tendon injury cohorts (Section 5.2.1), in addition to the SAW-ACL rupture (Section 5.2.2) and SAC-ACL rupture (Section 5.2.3) cohorts. Statistical analyses were performed using the R programming environment, with statistical significance set at $P < 0.05$.

RESULTS

Chapter 3: A total of 204 807 variants were identified across the 20 sequenced exomes, of which 338 variants were called in the genes previously associated with the risk of AT. The main novel findings of this chapter were the independent and haplotype associations of variants within the *TNC* and *COL27A1* genes with the risk of AT and ACL ruptures. Specifically, the *TNC* rs1061494 CC genotype associated with increased risk for AT ($P=0.018$, OR: 2.5 95% CI: 1.2-5.1), with an underrepresentation in the SAW-CON^{AT} group compared to the SAW-TEN group (13% vs. 24% respectively). The *TNC* rs2104772 AA genotype was associated with ACL ruptures in the SAW-ACL rupture cohort, demonstrating an underrepresentation in the SAW-CON^{ACL} group compared to the SAW-ACL group when female participants were evaluated independently (18% vs. 33%, $P=0.035$, OR: 2.3, 95% CI: 1.1-5.5). Inferred haplotype analyses further implicated the *TNC* genomic region. Specifically, the T-T inferred haplotype (rs1061494-rs2104772) was associated with decreased risk for AT (SAW-CON^{AT} 30% vs. SAW-TEN 22%, $P=0.032$). The *COL27A1* rs2241671 GG genotype was associated with increased risk for AT ($P=0.037$, OR: 1.9, 95% CI: 1.0-3.6) in the UK-Achilles tendon injury cohort and acute Achilles tendon ruptures (ATR) in the combined-Achilles tendon injury cohort ($P=0.047$, OR: 2.4, 95% CI: 1.0-6.2). Furthermore, the rs2241671 GG genotype was associated with increased risk for ACL ruptures, demonstrating a significant underrepresentation in the SAC-CON group (39%) compared to the group of participants reporting a non-contact mechanism of ACL injury (SAC-NON: 58%, $P=0.021$, OR: 2.3, 95% CI: 1.01-5.1).

Chapter 4: Besides rs3025058, none of the genotyped *MMP3* variants were independently associated with the risk of AT or ACL ruptures in the investigated cohorts. The rs3025058 6A6A genotype was by inference associated with increased risk for AT in the previously genotyped SAW-Achilles tendon injury cohort (CON 19% vs. TEN 37%, $P=0.010$, OR: 2.5, 95% CI: 1.2-4.9). The 6A-G-C-G (rs3205058-rs697620-rs591058-rs650108) allele combination was significantly associated with decreased risk for AT in the AUS-Achilles tendon injury cohort (CON 29% vs. TEN 20%, $P=0.037$). However, this haplotype contrasts to that previously associated with decreased risk for injury (5A-A-T-G). Therefore, this genomic interval requires further interrogation in significantly larger cohorts.

Chapter 5: The previously reported association of the rare *COL1A1* rs1800012 TT genotype with reduced risk for ACL ruptures in cohorts of self-identified White ancestry was also observed in the meta-analysis presented in chapter 5, which involved combining the data from several previously published studies with that generated in the current study (CON: 3% vs. ACL: <1%, $P=0.014$). However, this variant was neither associated with chronic AT in the UK- and combined (UK and SAW)-Achilles tendon injury cohorts investigated in this study, or with the risk of ACL ruptures in the SAC-ACL rupture cohort. This may potentially reflect the small sample sizes of the individual injury cohorts. Similarly, the *COL1A1* functional promoter variant, rs1107946, was not associated with the risk of chronic AT in the SAW- and UK-Achilles tendon injury cohorts or with ACL ruptures in the SAW- and SAC-ACL rupture cohorts. An inferred haplotype did however implicate the 5' regulatory region of the gene encompassing the rs1107946 and rs1800012 variants. Specifically, the G-T inferred haplotype, previously associated with decreased risk of ACL ruptures in a Polish cohort, was also significantly associated with decreased risk for acute ATR in the combined (UK and SAW)-Achilles tendon injury cohort (CON 16% vs. ATR: 8%, $P=0.022$). Whereas, the alternate T-G inferred haplotype was significantly associated with increased risk for non-contact ACL ruptures in the SAC-ACL rupture cohort (CON 24% vs. NON 33%, $P=0.047$).

Chapter 6: The results of the bioinformatic *in silico* analyses suggest that, although not disease causing, several of the variants investigated may be impressionable at the functional level. The *TNC* rs1061494 (T>C) variant is predicted to overlap the sequence motifs of the muscle initiator nuclear protein, members of the myogenic family of transcription factors and RNA polymerase II subunit A. Furthermore, the rs1061494 variant demonstrated marked differences in its predicted pre-mRNA structures. The other *TNC* variant associated with injury risk, rs2104772 (T>A), was predicted to be deleterious by two independent annotation tools. The investigated *COL27A1* variants (rs2567706 A>G, rs2241671 G>A, rs2567705 A>T) were suggested to interact with several predicted enhancers of cellular function. However, this gene is still relatively uncharacterised in musculoskeletal soft tissue injuries, therefore, these variants will require further interrogation. The 8 kb *MMP3* genomic interval demonstrated high levels of linkage disequilibrium. Furthermore, *MMP3* was predicted to interact with the *MMP12* gene mediated by chromatin looping. The *COL1A1* rs1800012 (G>T) variant overlaps the recognition sequence of the specificity protein-1 (Sp1) transcription factor in intron 1. This

variant is also proposed to interact with the functional promoter variants, rs1107946 (G>T) and rs11327935 (indel/T). This interaction is suggested to be mediated by chromatin looping. Furthermore, the rare rs1800012 T allele is predicted to result in a looser conformation immediately surrounding the variant within the pre-mRNA sequence compared to the ancestral G allele.

CONCLUSION

These results provide proof of concept for the use of WES and a customised tiered filtering strategy to identify and prioritise potentially functional DNA sequence variants for further interrogation using traditional molecular techniques. This approach, utilising previous research to guide a targeted analysis of a large WES dataset has provided evidence for the potential contribution of several new variants in the *TNC* and *COL27A1* genes to the profiles predisposing for musculoskeletal injuries. Furthermore, haplotype analysis has implicated several signatures encompassing variants previously associated with risk of injury in the four investigated genes. Although no new candidate variants within the *MMP3* and *COL1A1* genes were independently associated with risk of injury, unique allele combinations were observed to co-segregate with an altered injury risk profile. Therefore, this study has genetically characterised several previously implicated loci in other independent injury cohorts and highlighted new sequence signatures, which may potentially contribute to the susceptibility of musculoskeletal soft tissue injuries. The next step would be to explore the functional significance of these sequence signatures *in vitro*; a process that would help further characterise the biological mechanisms underpinning the observed risk associations.

THESIS OUTLINE

Chapter 1 serves as an introductory chapter to this thesis; presenting the two clinical phenotypes under investigation, chronic AT and acute ACL ruptures. The functional anatomy of the affected systems has been illustrated, in addition to reviewing the current literature relating to the aetiology, prevalence and factors associated with risk of the aforementioned musculoskeletal phenotypes.

The central concept on which this work is based is the contribution of intrinsic risk factors, specifically genetic variants, to musculoskeletal soft tissue injury predisposition. Using previous research as a foundation, the primary aim of this thesis was to further define the genetic risk profiles predisposing for musculoskeletal soft tissue injuries mapping to specific genomic intervals. To address this aim, a key project objective involved the development and application of a customised WES analysis pipeline, which is described in chapter 2. Using this approach, four genes including *TNC*, *COL27A1*, *MMP3* and *COL1A1* were prioritised for further interrogation. Previously associated variants within these targeted genomic regions and/or new candidate variants identified by WES and prioritised through the application of a tiered filtering strategy were genotyped in several, previously recruited AT and ACL rupture cohorts. All participants within these cohorts were self-identified to be of White ancestry. The second aim of this thesis was to determine whether the molecular signatures predisposing for injury in the self-identified White cohorts were similar to those predisposing for injury in a self-identified Coloured cohort. All cohorts and the variants genotyped are described in detail in chapter 2. Chapters 3 to 5 of this thesis present the results of genetic association studies, which involved testing the association between the candidate variants and chronic AT and ACL ruptures. Lastly, using a host of bioinformatic tools, the genotyped variants were functionally assessed for their potential biological significance in chapter 6.

Chapter 7 summarises the main findings of this work and presents the limitations and strengths of the studies. High throughput sequencing technologies have rapidly become an important tool in the endeavour to understand the biological mechanisms underpinning disease susceptibility. Therefore, this work has important implications for future research within the field of sports genomics, as discussed in the concluding chapter.

***“Science knows no country, because knowledge belongs to humanity and is the torch which
illuminates the world”***

Louis Pasteur

CHAPTER 1: MUSCULOSKELETAL SOFT TISSUE INJURIES IN REVIEW

1.1. INTRODUCTION

Research suggests that regular participation in physical activity and exercise is associated with physical and mental well-being [1]. Specifically, regular exercise promotes longevity, retards disease onset and improves overall quality of life [1]. However, paralleled with the increase in physical activity, is the adverse increase in the incidence of associated injuries, such as those affecting musculoskeletal soft tissues [2].

'Musculoskeletal injury' is an encompassing term used to describe injuries affecting tendons, ligaments, muscles and skeletal structures. With a worldwide prevalence of over 100 million cases reported annually, these injuries present as a global burden [3–5]. Moreover, 30-50% of all sports-related musculoskeletal injuries are reported to affect tendons and ligaments [6]. These debilitating injuries lead to impaired sporting performance and may reduce functional capabilities in the workplace. Chronic Achilles tendinopathy (AT) and acute anterior cruciate ligament (ACL) ruptures are examples of musculoskeletal soft tissue injuries for which the prevalence appears to be increasing worldwide, despite increased injury awareness, improvements to training equipment and some advances in preventative strategies [7,8]. Furthermore, there is emerging evidence that suggests there is inter-individual variability in response to training, recovery and adaptation to subsequent loading, in addition to the presentation and resolution of symptoms [9–11]. For example, some athletes appear to be resistant to the development of tendinopathy despite excessive loading while others may sustain catastrophic ruptures prematurely, presumably resulting from the degeneration of the tendon. Consequently, the efficacy of treatment modalities to provide relief to patients has varied considerably, with several intervention and therapeutic options currently in practice [12]. Despite the lack of consensus concerning treatment standards, it is evident that these injuries require early intervention; whether of a surgical nature or that of conservative, rehabilitative therapies [11]. Failure in this regard may result in injuries that are unresponsive to treatment and recurrent in nature.

The paucity of knowledge regarding the underlying biological mechanisms resulting in musculoskeletal injuries can potentially explain the varying levels of treatment success. Furthermore, there are also inter-individual differences in the response of musculoskeletal systems to load. This response is partly determined by the physiological variation in the biomechanical and structural properties of these tissues [13]. It can be hypothesised that some individuals are, therefore, intrinsically predisposed to the development of tendon and ligament injuries. It is imperative that research is focused at understanding the structural and biomechanical properties of tendon and ligaments and how these properties provide for the functional roles of these structures. Moreover, when considering treatment strategies, it is important to understand that the clinical presentation of these injuries including alterations in structure, pain and dysfunction, may vary among individuals [10,11].

Although mechanical load is considered as an important contributor to pathology [14], it is agreed that these injuries have a multifactorial aetiology including the interplay of both extrinsic and intrinsic risk factors [15–17]. Evidence that DNA sequence variants contribute to the predisposition of musculoskeletal soft tissue injuries has predominantly been provided by case-control genetic association studies (reviewed in [18,19]), together with several familial-based studies [20,21]. These studies, which have predominantly focused on investigating the association of variants located within genes encoding structural and regulatory components of the extracellular matrix (ECM) with the risk of injury, provide only a limited level of evidence. Consequently, these results do not currently support the clinical implementation of genetic screening tests for musculoskeletal soft tissue injuries. However, they do provide support for prospective cohort studies in the future which may confirm the involvement of genetic risk factors in the development of musculoskeletal soft tissue injuries, in addition to determining the capability of genetic screening tests to effectively discriminate injury risk. Therefore, it is vitally important that we increase our understanding of the biological mechanisms underpinning injury development, subsequently providing the basis for future, high-evidence level studies which may investigate personalised injury risk profiling in the clinical setting.

1.2. TENDON AND LIGAMENT ANATOMY

Tendons and ligaments are both examples of dense regular connective tissues, with overlapping morphological and molecular features. However, differences in the macro- and microstructure of these fibrous tissues relate to their anatomical position and specialised mechanical functions [22].

1.2.1. Achilles tendon

1.2.1.1. Macro-structure and functional anatomy

The Achilles tendon, which is subjected daily to large mechanical stresses, is the largest and strongest tendon of the human body and is highly specialised to fulfill its anatomical function [16]. As the conjoined tendon of the soleus and gastrocnemius muscles of the lower leg (myotendinous junction) and inserting at the calcaneus bone of the ankle joint (osteotendinous junction or enthesis), this tendon promotes the efficient dissipation of force from its originating muscle system to its bony attachment (Figure 1.1) [23].

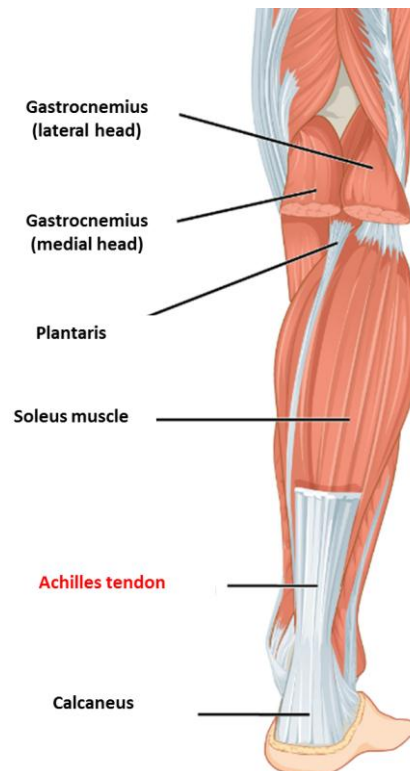


Figure 1.1: Basic anatomy of the lower limb

Schematic diagram showing the basic anatomy of the lower limb. The Achilles tendon is indicated in red. Source: OpenStax College via Wikimedia Commons

[\[https://commons.wikimedia.org/w/index.php?curid=64291156\]](https://commons.wikimedia.org/w/index.php?curid=64291156)

Tendons are highly tensile and resilient structures, capable of stretching to prevent damage during loading [24]. During physical activities such as running, hopping and jumping, the Achilles tendon has demonstrated *in vivo* loads up to 12 times an individual's total body weight [25,26]. Approximately 15 cm proximal to the calcaneal insertion, rotation of fibres up to 90° is observed, a phenomenon believed to enable greater elasticity and recoil of the structure [27].

In the absence of a true synovial sheath, the Achilles tendon is encapsulated in a thin layer of areolar tissue, the paratenon, which aids gliding of the structure during movement [28]. Furthermore, two bursae including the (i) retrocalcaneal bursa and (ii) subcutaneous bursa present around the calcaneal insertion. The retrocalcaneal bursa is a normal anatomical feature which is located between the Achilles tendon and the calcaneus bone to reduce

friction. Conversely, the subcutaneous bursa is an acquired feature located between the tendon and the superficial tissue in cases of inflammation and/or trauma (Figure 1.2) [29].

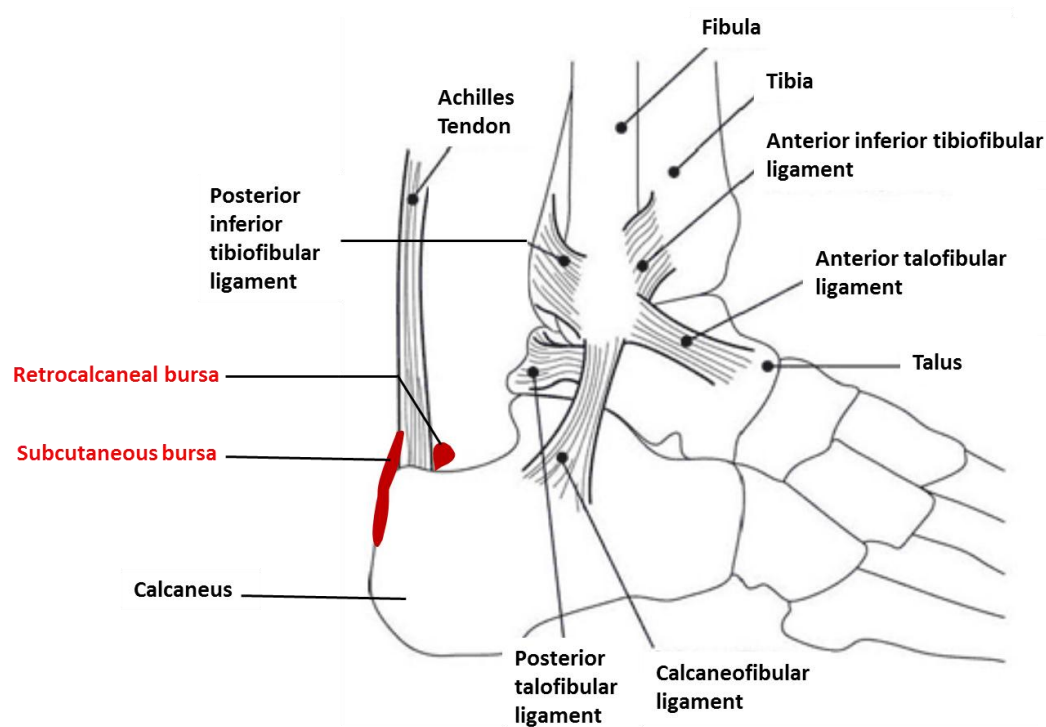


Figure 1.2: Basic anatomy of the ankle joint

Schematic diagram showing the basic anatomy of the ankle joint. The location of the retrocalcaneal and the subcutaneous bursae in relation to the Achilles tendon and its insertion on the calcaneus is demonstrated. The bursae are indicated in red. *Adapted from source: Wikimedia Commons*
[\[https://commons.wikimedia.org/w/index.php?curid=1176993\]](https://commons.wikimedia.org/w/index.php?curid=1176993)

1.2.1.2. Vascularisation

In comparison to other tissues, the Achilles tendon is relatively hypovascular, despite the presence of a vessel-rich paratenon [30]. However, angiographic methods have confirmed a blood supply to the structure that originates from three sources: the (i) myotendinous junction (MTJ), (ii) osteotendinous junction (OTJ) and (iii) vessels located within the surrounding connective tissue [31]. This vascular network predominantly originates from the posterior tibial artery with the greatest supply just proximal to the osseous insertion of the tendon [31]. In contrast, the mid substance is a region deficient in blood vessels and the site

with the greatest susceptibility to injury [30,32–34]. The Achilles tendon demonstrates a low metabolic rate presumably due to the relatively low number of residing tenocytes. The content of oxidative enzymes and mitochondria is, therefore, also relatively low [35]. The Achilles is one of two predominant energy-storing tendons in humans. This structure stretches and recoils with each stride to store and return energy, maximising energy efficiency. As a result, the Achilles tendon along with the energy-storing patellar tendon have specialised properties conferred by changes in matrix structure and composition including a relatively low vascular perfusion [36]. Even intense exercise is observed to result in only a subtle drop in oxygen content within the Achilles tendon and the effects of hypoxia are much less pronounced compared to areas of the tendon where blood flow is obstructed using a pneumatic cuff [37].

1.2.1.3. Innervation

Tendons are generally aneuronal which may explain, at least in part, their slow adaptation to repetitive mechanical loading, hindered healing properties and vulnerability to reoccurring injuries [38,39]. Innervation of the Achilles tendon is predominantly superficial and thought to involve nerves derived from the peritenon (paratenon and epitenon), cutaneous tissue and the associated muscle groups [40,41]. At the OTJ, all four neural receptor types are present: (i) Ruffini corpuscles (pressure and stretch), (ii) Vater-Pacini corpuscles (movement), (iii) Golgi tendon organs (tension) and (iv) free nerve endings (pain) [38]. Under normal homeostatic conditions the tendon proper is essentially devoid of nerve fibres. On the contrary, excessive neuronal ingress is observed after injury and chronic loading [39,42].

1.2.2. Anterior cruciate ligament

1.2.2.1. Macro-structure and functional anatomy

The ACL is an intra-articular ligament contributing to the stabilisation and movement of the knee joint together with the posterior cruciate ligament (PCL), the medial collateral ligament (MCL) and the lateral collateral ligament (LCL) (Figure 1.3). The PCL, MCL and LCL are not of

direct relevance to this thesis and are reviewed elsewhere [43,44]. At the macroscopic level, the ACL is a thick band of connective tissue encased in a membranous sheath, the synovium [45]. Following an oblique course, the ACL originates proximally at a fossa on the medial side of the lateral femoral condyle and terminates at the anterior intercondylar fossa on the tibial plateau [45,46]. This ligament is primarily tasked with stabilising against anterior translation of the tibia in relation to the femur while also controlling rotation during joint extension [46]. The anatomical complexity of this structure reflects the importance of its contribution to gross knee-joint movement.

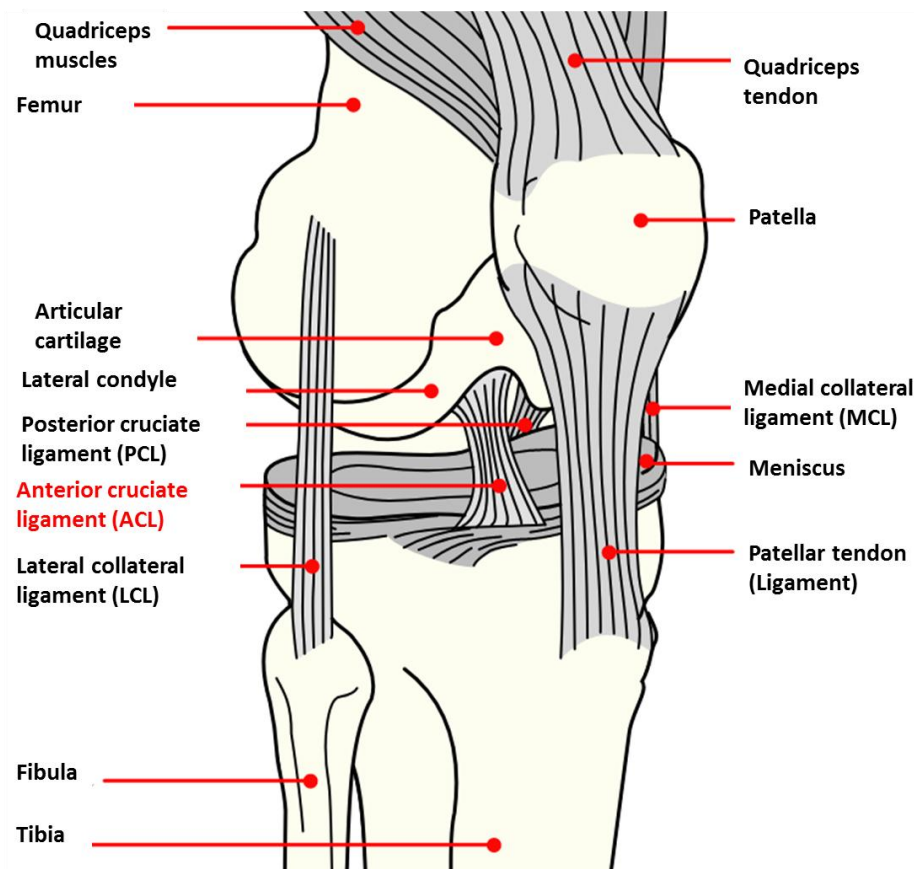


Figure 1.3: Basic anatomy of the knee

Schematic diagram showing the basic anatomy of the knee joint. The four major ligaments of the knee are demonstrated on the diagram including the (i) anterior cruciate ligament (ACL), (ii) posterior cruciate ligament (PCL), (iii) medial collateral ligament (MCL) and (iv) lateral collateral ligament (LCL). The ACL is indicated in red. Adapted from source: Mysid via Wikimedia Commons [https://commons.wikimedia.org/wiki/File%3AKnee_diagram.svg]

Arthroscopic investigations have demonstrated the ACL consisting of two functional bundles, the anteromedial (AM) and posterolateral (PL) bundles respectively [47,48]. Named as to indicate the orientation of their tibial insertions, these non-isometric bundles maintain a reciprocal relationship during knee movement. Specifically, at full extension, the PL bundle retains the greatest *in situ* force, however, upon increased flexion (up to 90°) this strain is reduced and the AM bundle becomes dominant in resisting anterior draw (Figure 1.4) [49].

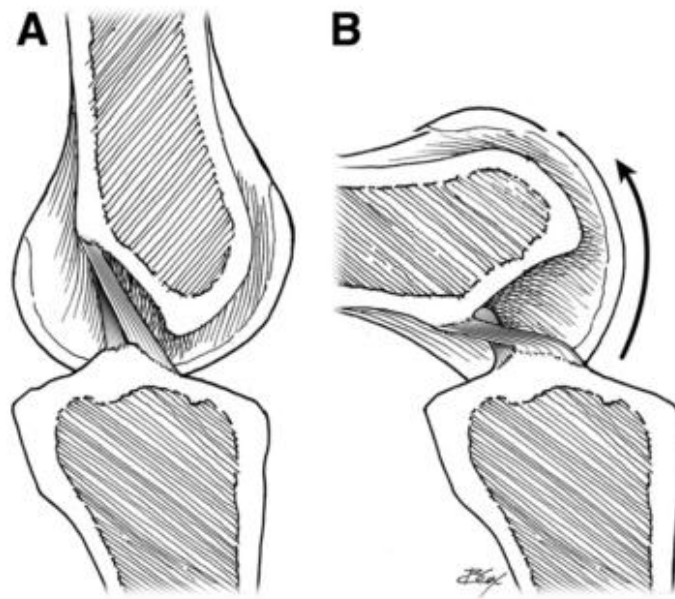


Figure 1.4: Orientation of the anterior cruciate ligament bundles

Schematic diagram demonstrating the crossing of the ACL bundles as the knee is flexed from extension. (A) The anteromedial and posterolateral bundles are parallel to one another during extension of the knee. (B) The anteromedial and posterolateral bundles are crossed during flexion of the knee. [Reprinted from Mall et al. (2013) [50] with permission from Elsevier]

1.2.2.2. Vascularisation

The ACL is known to be relatively hypovascular; an intrinsic factor believed to contribute to its poor healing potential [51,52]. Although the overlying synovial membrane contains blood vessels, very few penetrate the mid-substance [51,53]. Furthermore, the limited vascular architecture that is present in the ligament substance is not homogeneous. Specifically, the extremities are perfused to a greater degree compared to the central portion, while the proximal portion demonstrates a greater vascular density than the distal portion [53]. The blood

supply is provided by derivatives of the middle genicular artery (MGA) and the synovium. The proximal portion of the ACL, which receives its blood supply from the MGA, is perfused to a greater degree compared to the distal portion which is vascularised by infrapatellar branches of the inferior genicular arteries through the anterior fat pad [53]. These branches give rise to a web-like network of periligamentous vessels within the synovial membrane which ensheath the length of the ligament [48,53]. These transversely orientated vessels penetrate the ligament where they anastomose with the endoligamentous network of longitudinally orientated vessels coursing parallel to the collagen bundles (fascicles) in the connective tissue [48,53].

1.2.2.3. Innervation

The posterior articular branches of the tibial nerve serve as the primary innervation source of the ACL [54]. Penetrating the joint capsule posteriorly, these nerve fibres course the path of the synovial and periligamentous vessel network and proceed to reach the infrapatellar fat pad [54]. Most of these nerve fibres are vasomotor in function and are associated with the endoligamentous vasculature. However, studies have demonstrated the presence of other small isolated neural elements located among the collagen fascicles which course independently of the blood vessel [54,55].

Besides free nerve endings, the most prevalent mechanoreceptor type native to the ACL structure is the Ruffini corpuscle. These nerve fibres are sensitive to deformation, particularly stretching of the femoral portion of the ligament [56–58]. Other mechanoreceptors include the Vater-Pacini corpuscles and the Golgi-like endings [59]. Electron microscopy confirmed the presence of free-nerve endings in the vicinity of blood vessels and are suggested to serve as nociceptors in vasomotor control [56–58]. The proposed mechanism involves the local release of neuropeptides in response to chemical and/or mechanical deformation which consequently induce vasomotor responses [59]. Collectively, these receptors are involved in proprioception and initiate the afferent arc that signals for postural knee change [59,60] .

1.3. HIERARCHICAL ORGANISATION OF TENDONS AND LIGAMENTS

At the microscopic level, tendons and ligaments have a similar basic structure and constitution. These connective tissues consist of spindle-shaped cells (fibroblasts), interspersed within a network of collagen fibres [61]. Fibroblasts, also referred to as tenocytes (tendons) or ligamentocytes (ligaments), are responsible for the synthesis and maintenance of all ECM components, including collagen proteins (70-80% total dry weight), non-collagen proteins and the hydrated ground substance [28]. Type I collagen is predominant in both tendons and ligaments (~85%) and is responsible for the high tensile strength of these structures. This collagen is incorporated into the basic building block of tendons and ligaments, the collagen fibril, through a multi-stepped process, fibrillogenesis, orchestrated by the local fibroblasts [62–65]. Other less prevalent collagens including types III, V, VI, XI and XIV, in addition to other non-collagen proteins are associated with the fibrils. Collagen fibres and the fibroblasts are embedded in the ground substance which is a highly viscous interfibrillar saline gel consisting of water (60-80% wet weight), glycosaminoaglycans (GAG), structural proteoglycans and glycoproteins [66]. It functions to provide tendon and ligament fibres with viscoelastic properties; facilitating interfibrillar gliding and interactions between surrounding tissues. The continuous remodelling of the ground substance is also considered to be an important method for gas and nutrient diffusion [66].

These collagen-based tissues are hierarchically organised; the basic unit of which is a triple helical collagen molecule (Figure 1.5). The clustering of five collagen molecules gives rise to a single microfibril and bunches of several microfibrils form collagen fibrils. The aggregation of parallel microfibrils (imbedded in the ECM) to form collagen fibres is aided by the presence of cross-linking molecules, such as leucine rich proteoglycans [61,67]. Multiple fibres make up collagen fascicles that are individually encased by the endotenon/endoligament; a connective tissue sheath that accommodates vascular and neural elements. Finally, the epitenon/epiligament encompasses all fascicles to form a single gross structure. In the case of tendons, an additional superficial layer, the paratenon, is also present [28,66]. Both tendons and ligaments demonstrate a wave-like appearance (crimp) under conditions of little

or no tensile stress. This feature functions as a basic shock absorber along the length of the structure and allows for a certain amount of fibre elongation without the risk of damage [66].

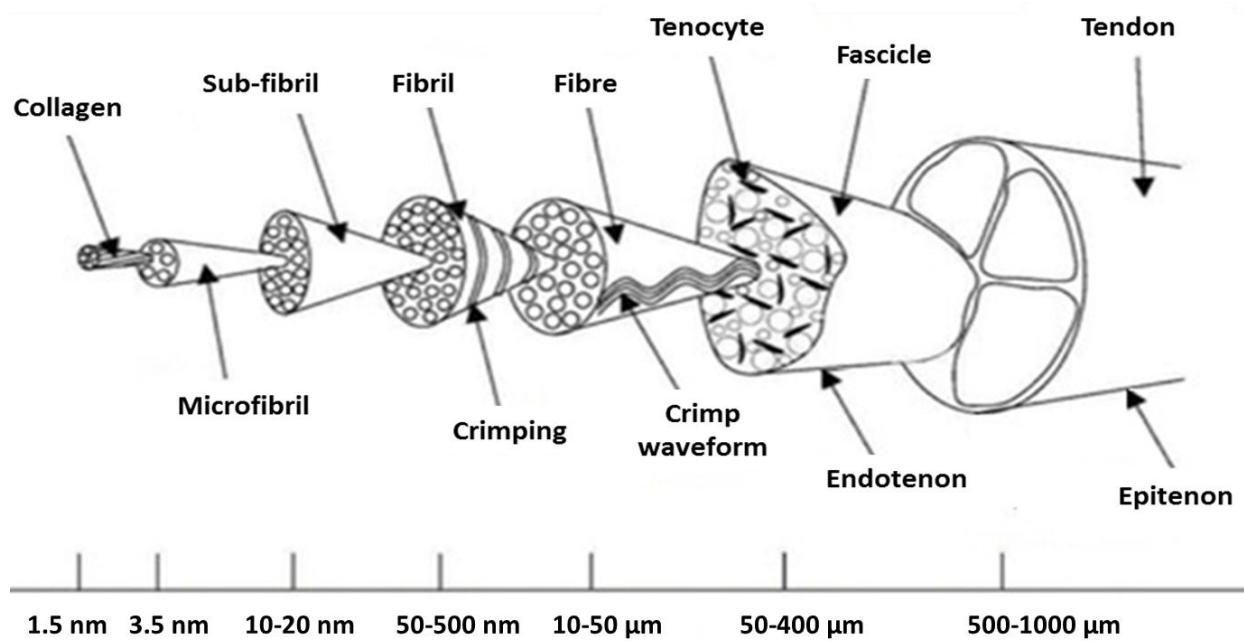


Figure 1.5: The hierarchical structure of a tendon

Collagen molecules aggregate to form fibrils, fibrils aggregate to form fibres and fibres group together to form fascicles. Fascicles bundle together to form tertiary fibre bundles; the primary unit of the tendon structure. The endotenon is a connective tissue that surrounds the bundled collagen fibrils, fibres and the fascicles. The epitenon is contentious with the endotenon and surrounds the entire tendon. [Reprinted from Screen et al. (2014) [68] with permission from Sage Publishing]

As previously described, the primary function of tendons is to transmit force generated through muscular contraction to bone, thus facilitating limb locomotion. This function requires tendons to be capable of withstanding high uniaxial tensile forces while resisting elongation. Conversely, ligaments primarily guide joint movement (bone on bone), promoting motion in some planes and preventing abnormal and/or excessive movement in others. This function requires ligaments to be able to resist high tensile forces that are multi-axial in origin. The differences in anatomical, functional and mechanical properties of these connective tissues are reflected in the precise composition and arrangement of their macromolecular components [22,69]. Tendon fibrils are tightly packed and align parallel to the long axis of the

structure. In comparison, collagen fibrils of ligament structures are less uniformly spaced and demonstrate an increased amount of surrounding ground substance. Furthermore, these connective tissues show marked differences in their overall collagen content, cell density and the concentration of non-collagenous molecules [22,69].

1.4. PATHOLOGIES OF THE ACHILLES TENDON AND ANTERIOR CRUCIATE LIGAMENT

A considerable proportion of morbidity associated with musculoskeletal injuries can be attributed to conditions affecting tendon and ligament structures [5]. These injuries are broadly classified as being either '*acute*' or '*chronic*' in nature. Acute injuries are spontaneous and result from a single macrotraumatic event as opposed to chronic injuries, which occur after considerable microscopic damage acquired through repetitive exposure over time [15]. Spontaneous ruptures of the Achilles tendon and ACL are examples of acute soft tissue injuries, whereas plantar fasciitis, tennis elbow and Achilles tendinosis are overuse chronic conditions. As described by van Mechelen et al. (1992), the initial step into injury prevention is to establish the magnitude of the problem, both in size and severity [70]. For this reason, a brief overview of the current knowledge regarding the epidemiology and aetiology of AT and ACL ruptures is provided below.

1.4.1. Achilles tendon injuries

1.4.1.1. Nomenclature

Several disorders are known to affect the Achilles tendon and its surrounds, fuelling debate as to the most appropriate terminology to describe these conditions. For the purposes of this thesis, the encompassing term '*Achilles tendon injuries*' will only refer to two common disorders affecting this structure including chronic AT and spontaneous Achilles tendon ruptures (ATR). Furthermore, the term '*tendinopathy*' will be used to define a syndrome in which pain and functional impairment are associated with swelling of the tendon area, in

addition to intratendinous alterations [71,72]. It is agreed that this term makes few assumptions as to the underlying pathology and entertains the idea that several pathological states may coexist within the diseased structure [10,11,72].

1.4.1.2. Prevalence

The Achilles tendon is particularly susceptible to injury with up to 18% of all diagnosed running-related tendon injuries localised to this structure [73]. Overuse tendinopathy is the most commonly diagnosed pathology of the Achilles tendon (55-65%) followed by insertional problems (20-25%). Other injuries include partial tears, undiagnosed ruptures and peritendinous perturbations [16]. The majority of Achilles tendon injuries can be attributed to a single overloading incident or degeneration over time. However, a small proportion of injuries present as comorbidities with systemic conditions, such as rheumatoid arthritis [74,75].

Overuse AT occurs most frequently amongst athletes participating in endurance sports, such as distance running, athletics and orienteering but may also be observed in sports involving repetitive jumping and sudden acceleration [76–81]. The lifetime cumulative incidence of AT amongst elite long-distance runners is reported at 52% [82], while annual incidence rates are reported between 7 and 9% [77,83]. In a sample of 57 725 Dutch participants, the incidence rate of midportion AT was observed to be 2.35 per 1000 adults [8]. Furthermore, Kujala et al. (2005) reported the lifetime incidence of AT in the general Finnish population to be approximately 6% [82]. Unfortunately, there is no data currently available describing the rates of Achilles tendon injuries in the South African populations.

The prevalence of AT increases with the advancement in age [84]. In a three-year follow-up study, Kannus et al. (1989) demonstrated the increased prevalence of degenerative, overuse injuries in veteran athletes (60-78 years) [84]. However, the majority of the elderly participants (84%, n=48) reported participation in endurance sports compared to only 25% (n=115) of the younger athletes analysed in the study. Moreover, most participants in the veteran and younger athlete groups were male (83% and 85% respectively). The gradual

deterioration of various tissues including musculoskeletal systems is observed with the advancement in age. Consequently, the type of injuries sustained by master athletes are more likely to have a degenerative basis [85].

The prevalence of AT is also reported to be higher in individuals with higher adiposity [86–92]. It is suggested that the mechanism linking adiposity to tendinopathy may be mechanical or systemic in nature. The mechanical theory proposes that an increase in body mass together with the misalignment of the foot and/or ankle may lead to increased forces and tensile strain experienced by the Achilles tendon [90]. Whereas, the systemic theory proposes that the bioactive peptides released by adipose tissue may drive a systemic state of chronic or low-grade inflammation; influencing tendon tissue remodelling. Alternatively, metabolic disorders such as diabetes, dislipidemia and hypercholesterolaemia, which have been associated with increased adiposity may indirectly affect the tendon structure leading to degeneration [90,93,94].

Achilles tendon ruptures are one of the most frequently presented tendon rupture phenotypes in the adult population [95], accounting for approximately 40% of all operated tendon ruptures in the study by Jozsa et al. (1989) [96]. Although the incidence of ATR varies according to geographical location, sports-related activities are identified as the predominant mechanism of injury in most countries (Table 1.1). Most sports-related ATR in central and northern Europe are reported to have occurred during badminton and football whereas basketball is most frequently associated with ATR in the United States of America (USA) [95,97–101]. It is therefore apparent that these injuries are often associated with strenuous activities involving abrupt repetitive jumping and sudden acceleration movements. Furthermore, some studies highlighted to seasonal variations in ATR rates [98,102,103].

The rate of ATR in males is reported to be three to six times greater than in females [95,98,100], yet females who have sustained Achilles tendon injuries are on average, two to six years older than male patients [100,101,104]. Furthermore, an apparent spike in injury rates between the ages of 30-49 years has been reported [95,98,101,104]. The results from a recent study that used a quasilinear viscoelastic model to test the association between Achilles tendon strain and age in rats, supported previous observations of an increased

incidence of ATR in middle-age individuals [105]. The stress at which middle-aged tendons failed was significantly lower (46%) than that of younger tendons. It was concluded that the reduced elasticity of older tendons may predispose older individuals to ATR at significantly lower levels of loading compared to younger athletes [105]. One may hypothesise that tendons of older individuals are at a higher risk of unresolved microtrauma with repetitive loading due to changes in tendon homeostasis. These changes may lead to incremental weakening of the structure which eventually fails at the macroscopic level [105].

Table 1.1: Country-specific incidence rates for acute Achilles tendon ruptures

Country	Previous incidence rate		Current incidence rate		Difference (%)	Sports-related (%)	Sporting code (%)	Reference
	Rate (10 ⁵ persons)	Year	Rate (10 ⁵ persons)	Year				
Scotland	4.7	1980	6.0	1995	28.0 %	-	Not provided	[106]
Finland	2.1	1989	21.5	2011	900.0 %	70.0 %	Badminton (30%), Volleyball (19%), Football (16%)	[101]
New Zealand	-	-	24.0	2011	-	79.0 %	Netball (31%), Squash (11%)	[107]
Sweden								
Male	47.0	2001	55.2	2012	17.0 %	-	Not provided	[108]
Female	12.9	2001	14.7	2012	22.0 %	-		
Denmark	27.0	1994	31.7	2013	16.0 %	73.0 %	Badminton (46%), Soccer (23%), Handball (16%)	[103]
Canada	18.0	2003	29.3	2013	63.0 %	-	Not provided	[109]

Incidence per 100 000 persons. Swedish study is presented as sex-specific rates of injury. Difference: Percentage change between the previous and current incidence rates. Sports-related: Percentage of injuries reported to be sports-related in mechanism. (-) No data provided for this parameter.

1.4.1.3. Histopathological characteristics

Histologically, healthy tendon is a brilliant white, highly organised tissue with a firm fibroelastic texture. The collagen bundles demonstrate uniform parallel packing with the individual fibres regularly sized and organised. Furthermore, tenocytes have spindle-shaped nuclei and are orientated with the collagen fibres. Conversely, tendinopathic tendon is dull in colour with increased ground substance [71,110,111]. Hypercellular areas are populated by metabolically active tenocytes with rounded nuclei while other sites can be completely devoid of any cellular elements [71,75,110,111]. Furthermore, the collagen fibres appear discontinuous and disorganised, and the characteristic crimp is abnormal and more pronounced due to increased loosening of the fibres. The accumulation of lipid cells can also be observed at sites of degeneration [71,75,112]. In contrast to the relatively avascular state of healthy tendons, degenerate tissues demonstrate increased vasculo-neural ingression [71,110,113,114]. Immune competent cells have been observed to accompany the infiltration of blood vessels, although the role of these elements in the disease course of tendinopathy remains unclear [115,116]. However, it has been hypothesised that the infiltration of immune cells may indicate an attempt at a healing response within the degenerate region of the tendon. However, if reparative attempts fail to result in adequate regeneration, the consequence is the production of a loose granulated tissue. This tissue is later converted to scar tissue, which cements the pathological changes [112].

1.4.1.4. Mechanism of injury

Currently, the underlying biological mechanisms resulting in injury are not completely understood. In saying this, several aetiological models have been proposed, all of which up until recently, predominantly focused on discrete events in the tendinopathic pathway. More recently, two integrative models have been presented which attempt to explain the clinical elements of tendinopathy in relation to the functional and structural properties of the tendon unit. Several of these models are briefly discussed below.

(A) Mechanical theory

The mechanical theory describes AT as the consequence of prolonged overloading and overuse of the musculoskeletal system; resulting in incremental fatigue, maladaptation and collagen fibre damage. This model, best described by the stress-strain relationship (Figure 1.6), proposes that repeated stressors, albeit within the structure's physiological range, may lead to the accumulation of microtrauma overtime and eventually present as painful symptoms. The incidence rate of AT is also said to increase with age; providing support for a more degenerative argument. Finally, stressors that surpass the structure's physiological threshold or that of its maximal capacity to withstand microtrauma, may potentially result in complete catastrophic failure to the system [117].

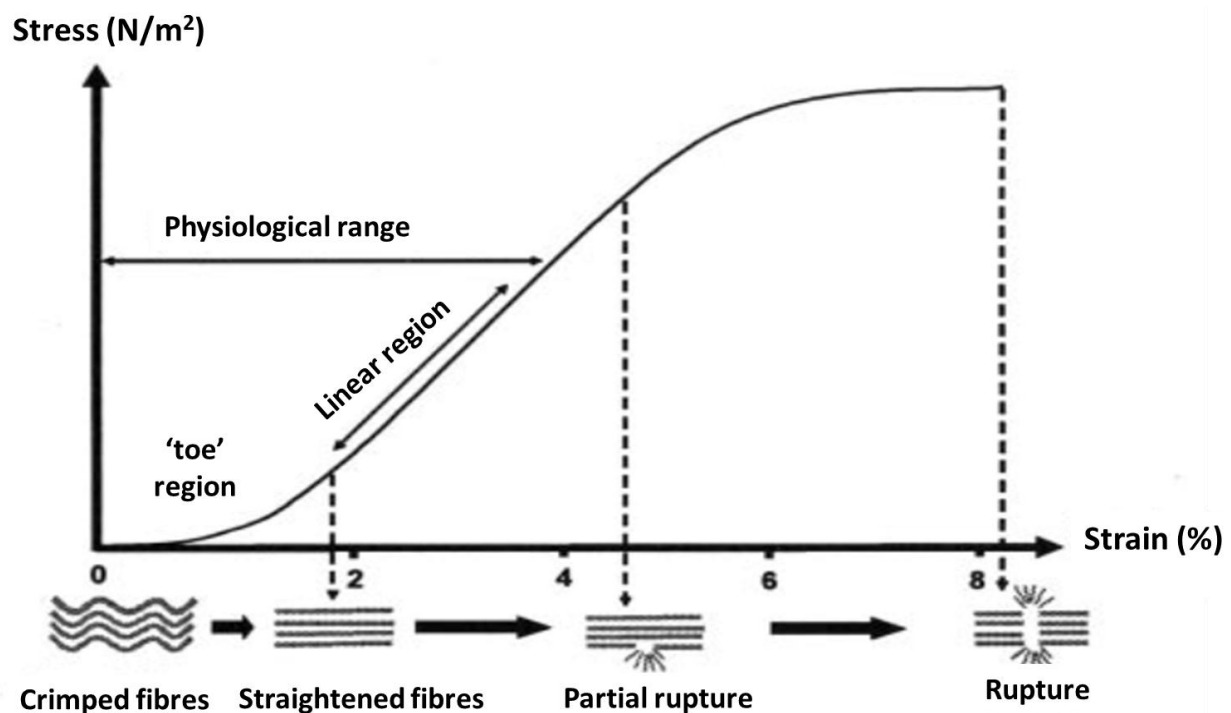


Figure 1.6: The stress-strain relationship of tendons

The application of strain to a tendon initially leads to the fibres losing their crimp; creating the 'toe region' of the curve. Further strain within the physiological range will induce elastic deformation. Any strain above the physiological range will result in partial or complete rupture of the tendon fibres. The precise levels of strain required to induce these changes along the curve may vary amongst tendons and is influenced by other biological factors. *[Reprinted from Riley et al. (2004) [118] with permission from Oxford University Press]*

More recently, Fu et al. (2010) proposed a ‘failed healing theory’ to describe tendinopathy [119]. To a certain extent, this theory acknowledges mechanical overload and overuse as pathological triggers but cautions against causative labelling. Instead this model focuses on failed healing; the second stage of the model which is elicited after the injury-inciting event and is characterised by an extended healing process due to the inability to fully resolve collagen damage. This prolonged healing cascade, together with other contributing factors including an inadequate healing environment, genetics, pharmaceutical influences and hormonal factors, may contribute to a tissue that is structurally disorganised and weaker, therefore, less able to respond effectively to further loading.

(B) Vascular theory

In contrast to load-induced mechanical failure, this model proposes alteration to the vascular supply as a driver of the pathological process. Several tendons are more susceptible to vascular deprivation including the Achilles tendon, particularly at the mid-portion [30–33,120]. It is proposed that insufficient vessel perfusion after an insulting event may hinder repair and lead to a weaker structure which is susceptible to degeneration and macroscopic rupture [31,121]. Furthermore, exercise-induced hyperthermia and hypoxia may further predispose avascular tissues to degeneration; a result of the tissues’ limited capability to regulate temperature and the effects thereof on matrix integrity [122]. Conversely, other investigations have demonstrated increased blood flow and neovascularisation in painful chronic tendon lesions [113,123,124]. This suggests that an unresolved hypervascular state may actually contribute to the clinical presentation of symptoms and chronicity of the condition [125].

(C) Neural theory

The studies by Alfredson and colleagues (1997 and 2003) noted nerve fascicles accompanying the blood vessels within symptomatic tendons, in addition to increased levels of glutamate signalling proteins, substance P and calcitonin gene-related peptide [113,126]. Similar features have also been observed by others [127,128]. These observations propose that

increased innervation in response to chronic loading may regulate a neurogenic inflammatory response, potentially contributing to matrix disorganisation and painful symptoms [129].

(D) Inflammation-associated, cell-signalling theory

The role of inflammation in the development of tendon injuries has long been a topic of considerable debate. Previously, '*tendinitis*' was used to describe chronically painful tendons [72]. This term implied injury with the accompaniment of inflammation. However, many opinions have since cautioned against the use of this terminology on the basis that several early studies that examined the late stages of the pathology had failed to identify chronic inflammation [71,111,130,131]. Features of a typical inflammatory cascade were however observed in response to acute tendon rupture including changes in localised protein expression, cell populations and tendon morphology [132,133].

Contrary to this traditional view of Achilles tendinopathy, we now understand that tenocytes are mechanoresponsive units, continuously responding to stimuli through biochemical signalling cascades. These cell-signalling pathways involve the local expression of certain pro-inflammatory cytokines to maintain the physiological properties of the tissue through the alteration of collagen synthesis and degradation. It is the loss in the homeostatic balance between synthesis and degradation that is likely to result in matrix disorganisation [134]. Furthermore, the results from a recent case-control genetic association study have implicated genes functioning within a signalling pathway involving inflammatory markers, with the risk of AT. Specifically, variants within several interleukin encoding genes including interleukin 1 beta (*IL1B*), interleukin 1 receptor antagonist (*IL1RN*) and interleukin 6 (*IL6*), in combination with a variant in the gene encoding the α_1 chain of type V collagen (*COL5A1*) were associated with the risk of AT [135]. This study supports the involvement of molecules within the inflammatory-associated, cell-signalling cascade in the development of tendinopathy.

In support of this genetic association study, several recent investigations have demonstrated the presence of inflammatory cells and key pro-inflammatory mediators in chronic tendinopathy [116,136–139]. Furthermore, the tenocyte hyperplasia and hypertrophy

observed under histological examination of pathological tendons may indirectly provide evidence for the upregulation of several cytokines and growth factors functioning within the inflammatory pathway, to which these cells are responsive [139]. This emerging research challenges the '*degenerative devoid of inflammation*' paradigm of tendon disease and instead proposes that treatment methods which target non-resolving inflammation may be therapeutically beneficial to patients seeking treatment for painful tendon injuries [116].

(E) Apoptosis theory

In response to mechanical overload, the normal healing cascade involves the removal of damaged ECM components and ageing tenocytes by cytokine-mediated apoptosis. However, repetitive overloading, or an acute bout of high tensile strain to which the tissue is not accustomed, may result in changes to the ECM and excessive apoptosis of tenocytes [140]. Consequently, the tissue may be compromised in its ability to regulate the repair process in response to further loading. Using DNA laddering techniques, Yuan and colleagues (2002) detected a significantly greater proportion of apoptotic cells in degenerative rotator cuff tissue (34%) compared to control tissue (13%) [141]. Furthermore, several genes encoding cytokines and proteins central to the caspase and nitric oxide pathways have been observed to be upregulated in tendinopathy [142–144]. Tumor necrosis factor (TNF) alpha, which induces apoptotic effects via its receptor, TNF receptor superfamily member 1A (TNFRSF1A), has been detected in tenocytes isolated from supraspinatus and Achilles tendinopathic tissue [144,145]. Furthermore, recent genetic association studies have also implicated several variants within the genes encoding caspase 8 (*CASP8*) [146] and caspase 9 (*CASP9*) [147], in addition to the Fas ligand (*FASLG*) and its receptor, the Fas cell surface death receptor (*FAS*) [148], with the risk of musculoskeletal soft tissue injuries. These studies further support the involvement of the apoptosis signalling pathway in the development of AT.

(F) Integrative models

Many aetiological models have attempted to describe the underlying biological mechanisms that precipitate Achilles tendon injuries. Although distinct in concept, these models may not

necessarily be mutually exclusive [11]. In fact, several mechanisms may contribute synergistically to the development of musculoskeletal soft tissue injuries. The consequence of mechanical loading is the mechanotransduction of signals by the residing tenocytes to regulate the synthesis of matrix proteins to maintain tissue homeostasis. The magnitude, frequency and duration of strain has been proposed to influence these homeostatic responses [149,150]. For example, at low levels of strain (4%) the intercellular communication between tenocytes is increased which promotes anabolic responses. Conversely, at higher strain levels (8%) the communication via gap junctions is decreased and inflammation and catabolic responses are favoured. Thus, changes in early cell-signalling could significantly alter downstream molecular cascades and potentially alter the physiological response of tenocytes to load [151]. Furthermore, factors including genetics, biomechanics, sex, age and systemic disease may also influence the threshold at which mechanical perturbations elicit cellular responses.

Two integrative models have been presented in the literature, the first by Fredberg et al. (2008) [152] and the second by Cook et al. (2009) [10], both of which present the development of AT as a pathological continuum but with some key differences. The Fredberg model, also referred to as the 'Iceberg theory', is presented in Figure 1.7 and has been further described by Abate and colleagues (2009) [153]. This theory proposes mechanical overuse as the instigator of injury. Physiological loading of tendons is beneficial resulting in increased tendon strength; however, once the individual set threshold is surpassed, any further loading may induce micro-disruption of collagen fibres. This initial micro-damage forms the base of the iceberg and represents the change that occurs within the tissue under physiological conditions. In response to micro-trauma, factors that aim to illicit an active healing response are produced, simultaneously to others which favour the disease cascade. With an adequate environment including sufficient recovery time devoid of any further loading, the healing mechanisms may prevail and result in complete repair. However, failure to resolve micro-damage leads to its accumulation (Phase 1) and the consequential weakening of the tendon structure due to degeneration. The increase in pro-inflammatory cytokines, vascular growth factors and free radicals within the tendon drives the pathological cascade into phase 2 of tendinopathy. This stage is still asymptomatic in patients until the amount of neovascularisation, neural ingression, and the consequence thereof including increased

glutamate levels, surpasses the tolerated threshold; evoking the perception of pain. This symptomatic stage represents the pinnacle of the iceberg. This theory also suggests a mechanism of relapse whereby brief periods of rest allow symptoms to recede below the detectable level; however, upon loading the threshold is automatically surpassed and painful symptoms are revived. This relapse phenomenon infers that the majority of intratendinous abnormalities are not resolved and that a relatively thin line exists between the asymptomatic and symptomatic stages of tendinopathy [152,153]. It must be highlighted that this theory confines pain and the presentation of other clinical symptoms to the end of the pathological spectrum.

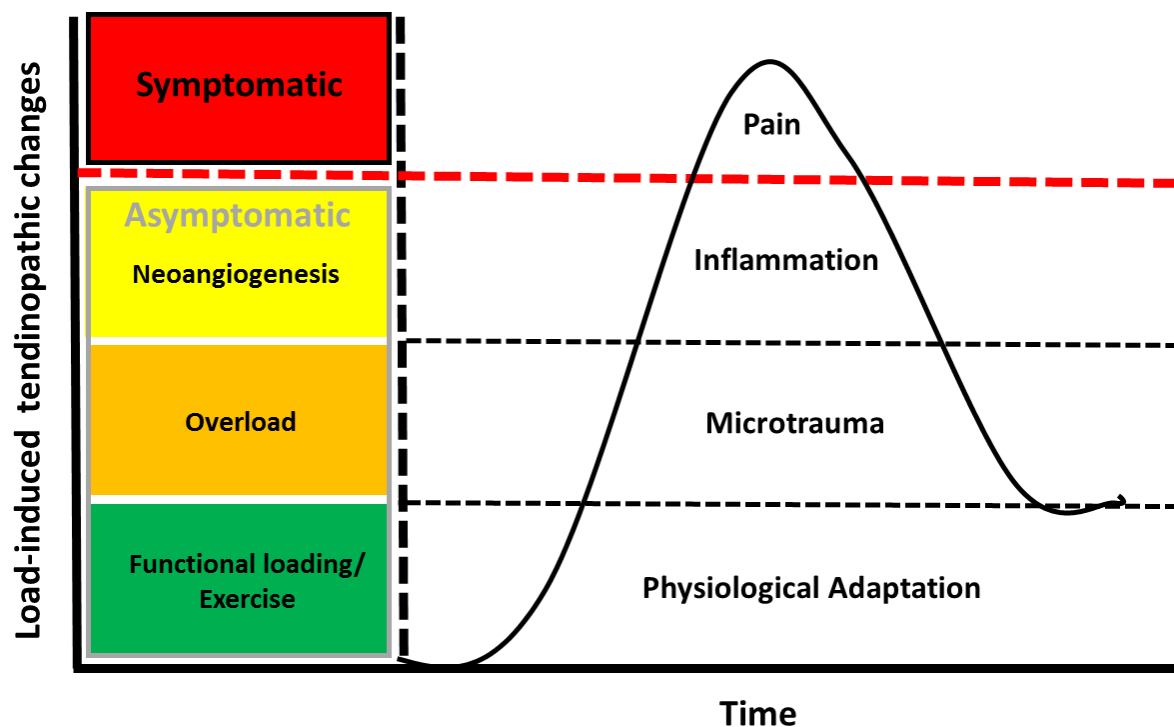


Figure 1.7: The 'Iceberg theory' of the pathogenesis of tendinopathy

[Compiled with modifications from Fredberg et al. (2008) [152] and Abate et al. (2009) [153]]

The second integrative model presented by Cook and colleagues (2009, 2016) is demonstrated in Figure 1.8 [10,11]. This model proposes the development of tendinopathy as a continuum with three stages:

- (i) Reactive tendinopathy
- (ii) Tendon dysrepair
- (iii) Degenerative tendinopathy

The addition or removal of a loading stressor, together with stage-specific treatment and the interaction with other risk factors, can drive the tendon back and forth through these stages along the continuum. Although the three stages are characterised by distinct tendon abnormalities, there is continuity between them. This is supported by the observation that discrete areas within a single tendinopathic structure may be affected by different stages within the pathology continuum. In contrast to the iceberg theory, this theory allows for the presentation of clinical symptoms including pain at any point along the continuum, taking into consideration the inter-individual variability of tissues in response to load.

Briefly, when a tendon is subjected to an acute mechanical overload, such as a single bout of physical exercise to which the tissue is not accustomed, short-term adaptive responses are initiated which lead to the thickening and stiffening of the structure. Reactive tendinopathy is a non-inflammatory-proliferative stage characterised by the rapid influx of large, water-bound glycoproteins and proteoglycans [154]. Any changes in imaging are mainly derived from the increase in water content, while collagen fascicles remain relatively unaltered. Importantly, the tendon may revert to normal physiology if loading is sufficiently reduced and adequate recovery time is provided between successive loading sessions. In the absence of these recovery measures, further loading of the tendon leads to tendon disrepair; an attempted healing stage similar to reactive tendinopathy but with more matrix disruption. Collagen fibres become separated due to the increase in cellular components and ECM proteins, in addition to the increase in vascular and neuronal ingressions [155]. Compared to tendinopathy in the reactive stage, reversibility of the pathology after tendon disrepair is limited; however, some alleviation is possible with controlled loading and exercise. At the furthest end of the continuum is degenerative tendinopathy. A tendon at this stage appears considerably

heterogeneous with some areas devoid of cells and collagen, while other regions may appear reactive or even normal [11]. Regions devoid of cells are occupied by blood vessels which are surrounded by remnants of the collagen matrix [71,111]. At this stage, the tendon has little capacity to recover. This model comprehensively details the shift from physiological adaptation into tendon degeneration but unlike the iceberg theory also provides compelling evidence for stage-specific intervention and rehabilitation. Management of tendinopathy may be optimised by phenotyping and treating patients individually using tailored interventions which target specific stages of pathology based on structure morphology, pain and mechanical load capacity. In this way, appropriate treatments would be directed at specific clinical factors driving alterations in matrix remodelling [10,11].

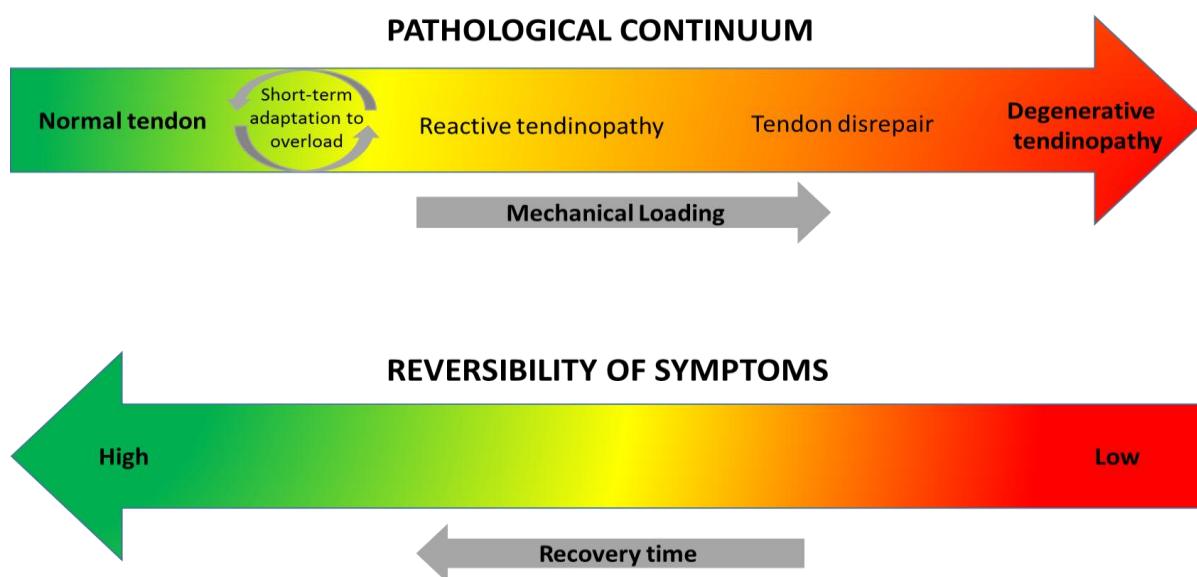


Figure 1.8: The pathological continuum model of tendinopathy as proposed by Cook and Purdam et al. (2009) [10]

1.4.2. Anterior cruciate ligament ruptures

1.4.2.1. Prevalence

Ruptures of the ACL are one of the most common injuries sustained during sporting activities. It is estimated that between 75 000 to 250 000 new ACL injuries occur annually in the USA alone, most of which are sustained by young athletes (15-25 years) participating in sports involving jumping and pivoting [156,157]. Interestingly, most of these injuries are reported to be non-contact in mechanism [158,159]. Furthermore, it is apparent that women are more predisposed to injury compared to males playing at similar sporting levels [160,161].

Moses et al. (2012) compared the annual incidence rates reported by 13 independent studies [162]. The national ACL incidence rates were observed to range between 0.01% and 0.05% annually, with the United Kingdom (UK) demonstrating the lowest incidence (0.01-0.02%) and Australia (AUS) reporting the highest (0.05%). Different levels of sports professionalism have also demonstrated varying incidence rates. Specifically, the annual incidence rates reported for amateur athlete groups were lower (0.03%-1.62%) compared to those reported for professional athletes (0.15%-3.67%); however, both were higher than that reported for the general population.

To date, three population-based registries have been established, all of which aim to gather information regarding ACL surgery and to monitor post-operative outcomes at a national level. In the Norwegian National Knee Ligament Registry, a total of 2 714 ACL reconstructive surgeries were registered between 2004 and 2006 [163]. Of these injuries, only one percent (n=28) required revision surgery. In comparison, a total of 16 351 reconstructive surgeries were recorded in the Swedish National ACL Registry between 2005 and 2010, of which four percent (n=654) required revision surgery [164]. Finally, 4 972 primary ACL reconstructions were recorded in the Danish ACL registry between 2005 and 2007, with three percent (n=149) of these primary injuries requiring surgical revisions [165]. However, in a recent study by Schilaty et al. (2017) the incidence of secondary rupture was considerably higher in the Olmsted county of Minnesota (USA) [166]. Specifically, 13.8% (n=126) of the 914 patients

diagnosed with primary ACL ruptures reported secondary tears. Approximately 50% were revisions of the primary surgery, while the other 50% were injuries sustained to the contralateral ligament. The outcome of revision surgery is considered significantly less effective than that of primary ligament reconstruction [167].

Evidence suggests an association between ACL ruptures and concomitant injuries (meniscal or other ligament damage) with the early onset of osteoarthritis within the knee joint [168–171]. However, the prevalence of osteoarthritis in ACL defective patients is reported with considerable variation throughout the literature [170,172,173]. Summarising the data from several follow up studies with documented osteoarthritis prevalence rates in ACL injury patients, Myklebust et al. (2005) inferred that half of all patients who have sustained ACL injuries would display radiographic signs of osteoarthritis after 10 years and almost all patients would suffer from the condition within 15-20 years [174].

1.4.2.2. Histological characteristics

In comparison to the MCL which is able to heal, the ACL retracts and does not demonstrate the capacity to form a bridging scar after rupture [175,176]. Although an active healing cascade is elicited in response to ACL trauma, it differs from that observed in other dense connective tissues. Specifically, there is an intermediate epiligamentous reparative stage between the inflammatory and proliferative stages which is characterised by the gradual growth of a thick synovial tissue over the ends of the ligament remnants; encapsulating the frayed tissue ends. This phase is not present in the healing of other tissues in which the inflammatory phase is immediately proceeded by fibroblast proliferation. Fibroblast proliferation is observed in the ruptured ACL; however, it occurs considerably later in the healing cascade [175]. The late stages of healing involve the retraction of the ACL remnants; however, no functional scar is observed connecting the two ends.

1.4.2.3. Mechanism of injury

It is evident from the literature that there has been a lack of consensus regarding the classification of ACL injuries by mechanism [177]. This problem has previously limited the ability of researchers to define ACL injuries using a standardised classification scheme. To address this problem, the American Orthopaedic Society of Sports Medicine, at the 2005 consensus conference on non-contact ACL ruptures (Atlanta, Georgia “Hunt Valley II meeting”), devised a scheme which has subsequently lead to the stratification of ACL injuries according to the type of contact experienced by the knee at the time of the inciting event (Figure 1.9) [177]. The three overarching terms used to define the mechanisms of ACL injury according to this system include (i) direct contact, (ii) indirect contact and (iii) non-contact.

Contact injuries result from the exposure of an athlete to an external force, such as a collision with an opponent or an object. These injuries are further stratified into those which are direct; force applied directly to the knee or indirect; force applied to the athlete but not directly to the knee. In contrast, a non-contact mechanism results from an athlete’s own movement and does not involve an external perturbation [177]. It can therefore be hypothesised that non-contact ACL injuries have a greater intrinsic contribution compared to injuries sustained through contact mechanisms. It is for this reason that it is important to analyse and report the frequency of non-contact ACL injuries independently to that for the other ACL injury mechanisms. This would increase the reliability of data and allow for meaningful comparisons between large datasets. Since the release of this classification scheme, modifications have been proposed [178]; however, for the purpose of the studies presented in this thesis, and in line with the work previously published by our research group, the scheme described above was used to classify ACL injuries. It must be noted that the mechanisms resulting in ACL injuries sustained during skiing are distinctly different to those that predominately result in injury during land-based sporting activities [179]. For this reason, ACL injuries sustained during skiing were not considered in this thesis.

Direct contact
Knee struck directly by an external force. Forceful contact the likely cause of injury. <i>“Was kicked by an opponent on the inside of the knee forcing it outwards. I felt and heard a popping noise.”</i>
Indirect contact
A physical perturbation to the athlete; however, the knee is not directly insulted. <i>“Was on the attack, planted my foot but was pushed by an opponent and my knee hyperextended. It was at this point I heard a pop.”</i>
Non-contact
A non-physical perturbation. Forces experienced by the knee at the time of injury are generated through an athlete’s own movement. No bodily contact with another player or object is experienced. <i>“Was running towards the line when an opponent stepped out in front of me. I side stepped and my knee buckled, causing me to collapse.”</i>

Figure 1.9: Proposed scheme for the classification of ACL injuries by type of contact, as devised by the American Orthopaedic Society for Sports Medicine. Figure modified from Marshall et al. (2007) [177]

A classic example of a patient’s narrative of the inciting event is provided.

1.5. RISK FACTORS

There is still much to be understood regarding the aetiology of musculoskeletal soft tissue injuries; however, it is generally accepted that they are complex multifactorial phenotypes influenced by a myriad of intrinsic and extrinsic risk factors (Figure 1.10). Intrinsic factors are those that influence an individual's risk from within, whereas extrinsic factors act on the body from without [15]. It is the net effect of the intrinsic and extrinsic risk factors which is suggested to determine injury susceptibility.

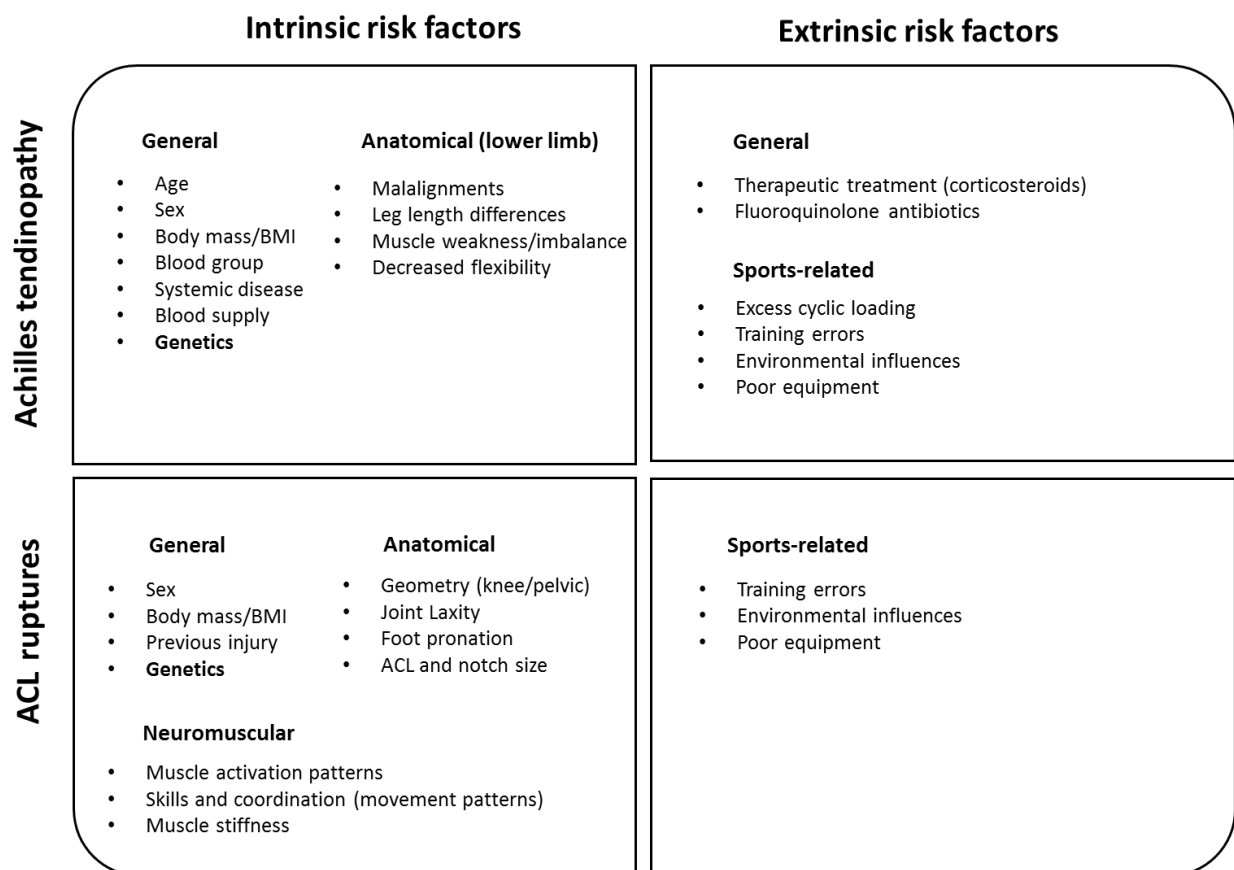


Figure 1.10: Intrinsic and extrinsic factors proposed to influence the risk of chronic Achilles tendinopathy and anterior cruciate ligament ruptures

Reviewed by Posthumus et al. (2011) [17], Uhorchak et al. (2003) [180] and Jarvinen et al. (2001) [16].

Most common sporting injuries follow the injury causation model of development, first presented by Meeuwisse in 1994 (Figure 1.11) [15]. Briefly, a combination of intrinsic risk factors may predispose an individual to injury. Although these factors are necessary, they rarely result in injury independently. Extrinsic factors are considered enablers of injury development and render an individual susceptible. However, similar to intrinsic factors, extrinsic factors are insufficient to bring about injury when acting in isolation. An inciting perturbation is the final link in the chain of causation and is associated with the definitive outcome of injury. These precipitating factors are regarded as necessary in the injury causation model. Although most injuries follow this model of development, the contribution of these factors to injury may vary among injury types. For example, it is hypothesised that extrinsic factors are the predominant instigators of acute traumatic injuries, whereas the development of chronic overuse injuries is more multifactorial in mechanism [15].

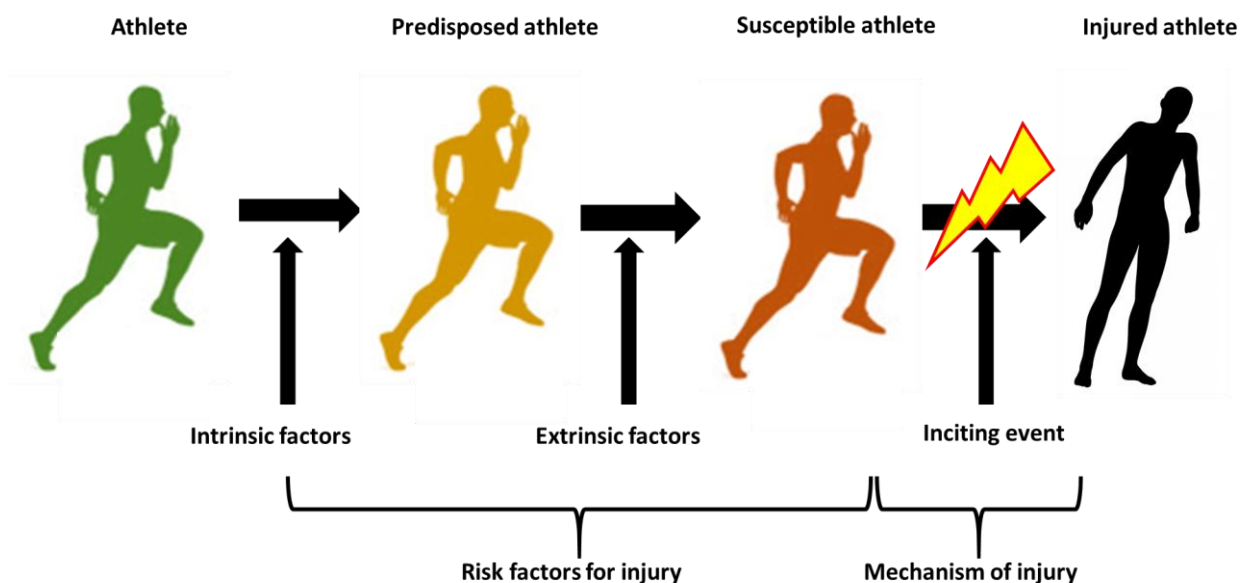


Figure 1.11: Multifactorial model of the aetiology of sports-related injuries as proposed by Meeuwisse (1994) [15]

1.5.1. Genetic predisposition

DNA sequence variants may contribute to the inter-individual differences in response to mechanical loading by influencing tissue integrity and consequently the function of tendon and ligament structures. By modulating the functional capacity of these tissues, the net effect of several risk factors may indirectly modulate injury susceptibility. Research of this nature has predominantly involved conducting familial or case-control genetic association studies. It was reported that the relative risk of sustaining full thickness rotator cuff tears (RCT) was significantly higher in siblings of participants with injuries compared to the spouses of the participants (OR: 2.4, 95% CI: 1.8-3.3) [181]. Furthermore, in a matched pair study, Kraemer et al. (2012) reported a positive family history of AT as a risk factor for AT (OR: 4.8, 95% CI: 1.1-21.4) but not for spontaneous ATR [21]. Similarly, Flynn et al. (2005) reported that individuals with ACL injuries were twice as likely to have had a family relative whom had suffered an acute ACL rupture compared to study participants without ACL injuries (OR: 2.0, 95% CI: 1.2-3.3) [20].

To date, numerous case-control genetic association studies have been conducted, several of which have found an association between the presence of variants within specific genes and injury susceptibility. Table 1.2 displays the loci previously implicated in the aetiology of chronic AT and acute ACL ruptures.

Table 1.2: Variants within genes encoding extracellular matrix proteins and cell-signalling factors previously associated with the risk of chronic Achilles tendinopathy and/or anterior cruciate ligament ruptures using case-control genetic association studies

Gene	Encoded protein	Chromosome position	Variants	Coding effect (location)	Injury	Association	Reference
Structural elements							
<i>COL1A1</i>	α_1 chain (I) collagen	17q21.33	rs1107946 (G/T)	Regulatory (Promoter)	ACL	G, H	[182,183]
			rs1800012 (G/T)	Synonymous (Intron 1)	ACL; acute STI	G, H; G	[183–187]
<i>COL3A1</i>	α_1 chain (III) collagen	2q31	rs1800255 (G/A)	Missense (Exon 3, A198T)	ACL	G	[188,189]
<i>COL5A1</i>	α_1 chain (V) collagen	9q34.3	rs12722 (T/C)	Regulatory (3' UTR)	AT; ACL	G, H; G, H	[188,190–194]
			rs71746744 (-AGGG)	Regulatory (3' UTR)	AT	G, H	[193,195]
			rs16399 (ATCT/-)	Regulatory (3' UTR)	AT	G	[195]
			rs1134170 (A/T)	Regulatory (3' UTR)	AT	G	[195]
			rs3196378 (C/A)	Regulatory (3' UTR)	AT	G, H	[191,193]
<i>COL11A1</i>	α_1 chain (XI) collagen	1p21.1	rs3753841 (T/C)	Missense (Exon 52, L1323P)	AT	H	[196]
			rs1676486 (C/T)	Missense (Exon 62, P1535S)		H	
<i>COL11A2</i>	α_2 chain (XI) collagen	6p21.3	rs1799907 (T/A)	Silent (IVS6-4)	AT	H	[196]
<i>COL12A1</i>	α_1 chain (XII) collagen	6q13-q14.1	rs970547 (A/G)	Missense (Exon 65, S3058G)	ACL	G, H	[188,197]
<i>COL27A1</i>	α_1 chain (XXVII) collagen	9q32	rs946053 (G/T)	Synonymous (Intron 41)	AT	H	[198]

Table. 1.2 continued.

Gene	Encoded protein	Chromosome position	Variants	Coding effect (location)	Injury	Association	Reference
Glycoproteins							
<i>TNC</i>	Tenascin-C	9q33.1	Z11654 (GT repeat)	Synonymous (Intron 17)	AT	G	[199]
			rs2104772 (T/A)	Missense (Exon 17)	AT	G, H	[198]
			rs1330363 (C/T)	Synonymous (Intron 15)	AT	G,	
			rs13321	Missense (Exon 24)	AT	H	
<i>FBN2</i>	Fibrillin 2	5q23.3	rs331079 (G/T)	Synonymous (Intron 7)	AT; ACL	G; G	[200]
Proteoglycans							
<i>ACAN</i>	Aggrecan	15q26.1	rs2351491 (C/T)	Silent (Exon 11)	ACL	H	[201]
			rs1042631 (C/T)	Silent (Exon 11)	ACL	H	
			rs1516797 (T/G)	Synonymous (Intron 12)	ACL	G, H	[201,202]
<i>BGN</i>	Biglycan	Xq28	rs1126499 (C/T)	Synonymous (Exon 4)	ACL	H	[194,201]
			rs1042103 (G/A)	Regulatory (3' UTR)	ACL	G, H	[194,201,202]
<i>DCN</i>	Decorin	12q21.33	rs13312816 (C/T)	Synonymous (Intron 1)	ACL	H	[201]
			rs516115 (T/G)	Synonymous (Intron 3)	ACL	G, H	[194,201]
<i>LUM</i>	Lumican	12q21.33	rs2268578 (T/C)	Synonymous (Intron 3)	ACL	H	[201]

Table 1.2 continued.

Gene	Encoded protein	Chromosome position	Variant	Coding effect (location)	Injury	Association	Reference
Matrix metalloproteases							
<i>MMP1</i>	Matrix metalloproteinase 1	11q22.3	rs1799750 (1G/2G)	Regulatory (Promoter)	ACL	H	[203]
<i>MMP3</i>	Matrix metalloproteinase 3	11q22.3	rs3025058 (5A/6A)	Regulatory (Promoter)	ACL; AT	G; G	[204,205]
			rs679620 (A/G)	Missense (Exon 2, E45K)	AT; ACL; ATR	G, H; H; H	[203,205–207]
			rs591058 (T/C)	Silent (Intron 4)	AT	G, H	[205,206]
			rs650108 (A/G)	Silent (Intron 8)	AT	G, H	[205,206]
<i>MMP10</i>	Matrix metalloproteinase 10	11q22.3	rs486055 (C/T)	Missense (Exon 1, R53K)	ACL	H	[203]
<i>MMP12</i>	Matrix metalloproteinase 12	11q22.3	rs2276109 (A/G)	Regulatory (Promoter)	ACL	G, H	[203]
Tissue inhibitors of metalloproteinases							
<i>TIMP2</i>	TIMP metalloproteinase inhibitor 2	17q25.3	rs4789932 (C/T)	Regulatory (Promoter)	ATP	G	[207,208]
Cytokines							
<i>IL1B</i>	Interleukin 1 beta	2q14.1	rs16944 (T/C)	Regulatory (Promoter)	AT	H	[135]
			rs1143627 (C/T)		AT	H	
<i>IL1RN</i>	Interleukin 1 receptor antagonist	2q14.1	rs2234663 (VNTR)	Synonymous (Intron 2)	AT	H	[135]
<i>IL6</i>	Interleukin 6	7p15.3	rs1800795 (G/C)	Regulatory (Promoter)	AT	H	[135]

Table 1.2. continued.

Gene	Encoded protein	Chromosome position	Variants	Coding effect (location)	Injury	Association	Reference
Other signalling factors							
<i>VEGFA</i>	Vascular endothelial growth factor A	6p21.1	rs699947 (C/A) rs1570360 (G/A) rs2010963 (G/C)	Regulatory (Promoter) Regulatory (Promoter) Regulatory (5' UTR)	ACL; AT ACL; AT ACL; AT	G, H; H G, H; H H; H	[209,210]
<i>KDR</i>	Kinase insert domain receptor	4q12	rs2071559 (G/A) rs1870377 (T/A)	Regulatory (Promoter) Missense (Exon 11)	ACL ACL	G, H H	[209,211]
<i>GDF5</i>	Growth differentiation factor 5	20q11.22	rs143383 (T/C)	Regulatory (5' UTR)	ATP	G	[212]
Apoptosis							
<i>CASP8</i>	Caspase 8	2q33.1	rs3834129 (indel)* rs1045485 (G/C)	Regulatory (Promoter) Missense (Exon 9, D302H)	AT AT	G, H G, H	[146]

STI: Musculoskeletal soft tissue injuries; ACL: Anterior cruciate ligament ruptures; AT: Chronic Achilles tendinopathy; ATR: Acute Achilles tendon rupture; ATP: Achilles tendon pathology; VNTR: variable number tandem repeat polymorphism; indel: insertion/deletion polymorphism. UTR: Untranslated region. Coding effect (location): Missense (amino acid change, exonic), synonymous (unchanged amino acid, intronic), silent (no amino acid change, exonic), regulatory (non-coding, promoter or untranslated region). Association: H: Variant associated with injury risk as part of an inferred haplotype; G: Variant genotype associated with injury risk. (*) rs3834129 is an insertion/deletion polymorphism (-/CTTACT) A: Alanine; D: Aspartic acid; E: Glutamic acid; G: Glycine; H: Histidine; K: Lysine; L: Leucine; P: Proline; R: Arginine; S: Serine; T: Threonine. [Adapted from Rahim et al. (2016) [18]]

Based on this previous research, four genes were prioritised for further interrogation in this thesis. The selected genes encode the tenascin-C glycoprotein (*TNC*; chr9q33.1), the α_1 chain of type (XXVII) collagen (*COL27A1*; chr9q32), matrix metalloproteinase 3 (*MMP3*; chr11q22) and the α_1 chain of type (I) collagen (*COL1A1*; chr17q21.33) respectively. These genes and their encoded protein products, which are proposed as important modulators of matrix remodelling or alternatively, are structural components of the ECM, will be reviewed briefly in the following sections. In addition, any previous literature implicating these genomic regions in the profiles predisposing for musculoskeletal soft tissue injuries will be discussed.

1.5.1.1. Tenascin-C

The human *TNC* gene, located on the long arm of chromosome 9 (chr9q33.1), is 97.6 kb and comprises of 30 exons (Figure 1.12). Mokone et al. (2005) was first to implicate a *TNC* variant with the risk of musculoskeletal soft tissue injuries [199]. Specifically, a guanine-thymine (GT) dinucleotide repeat within intron 17 of the gene was observed to associate with the risk of AT. Subsequently, Saunders et al. (2013) implicated two other variants, rs1330363 (A>G, intron 15) and rs2104772 (A>T, exon 17), both of which demonstrated independent allele associations with the risk of AT [198]. This study also inferred the association of a haplotype spanning the distal end of the *TNC* gene including the rs2104772 and rs13321 (G>C, exon 24) variants, together with the upstream *COL27A1* rs946053 (G>T) variant, with an altered AT risk profile. Interestingly, the rs2104772 variant has also been associated with asthma in an adult Japanese population [213]. Most recently, Kluger et al. (2017) identified several *TNC* variants associated with the risk of degenerative RCT [214]. This study highlighted the potential biological importance of one variant in particular, rs1138545 (C>T, exon 10), which alters the amino acid code (Arg1066His) close to the bridging sequence between two adjacent fibronectin type III repeat domains within the tenascin-C peptide. Following this initial study, the same research group identified a three variant *TNC* haplotype, C-A-G (rs1138545 C>T-rs2104772 A>T-s10759752 A>G) associated with large recurrent defects after RCT surgery [215].

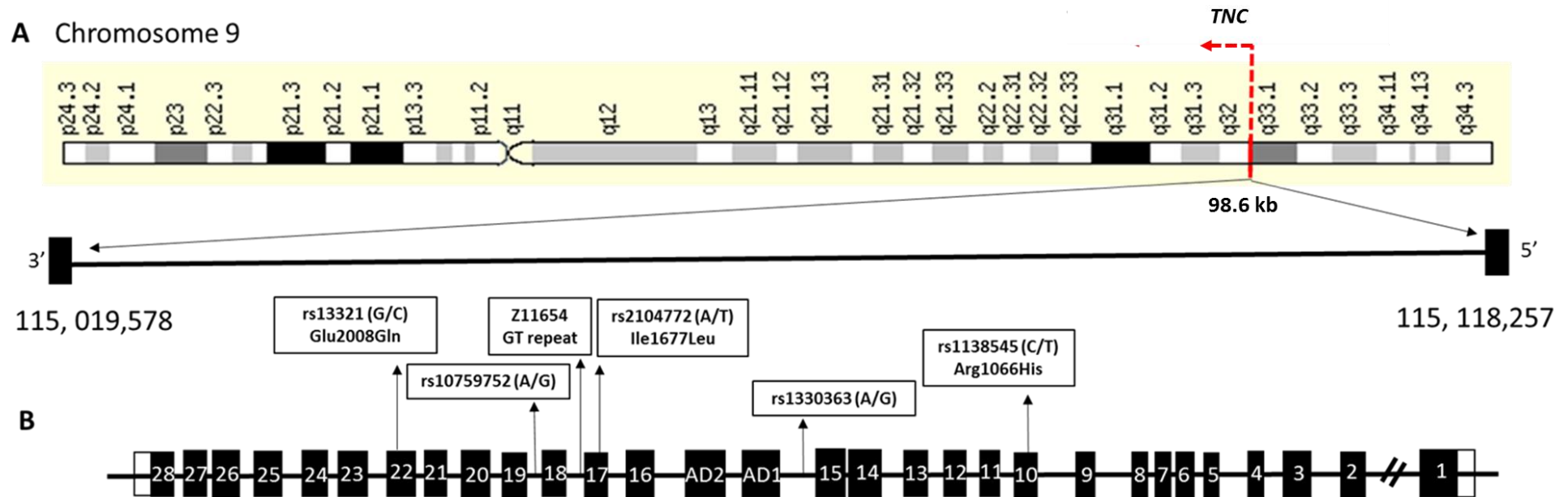


Figure 1.12: Gene encoding the tenascin-C glycoprotein

(A) Cytogenetic banding pattern of chromosome 9. The red arrow indicates the location of the *TNC* gene on chromosome 9 and direction of transcription (antisense strand). The genomic boundaries of *TNC* are provided according to the Human GRCh38.p10 genome release. (B) The numbered exons 1-28, AD1 and AD2 (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *TNC* are presented. Black arrows on gene demonstrated the location of variants previously associated with the risk of musculoskeletal soft tissue injuries, either independently or by inferred haplotypes analysis [198,199,214,215]. Exon 1 of *TNC* is not translated (double strikethrough). AD1 and AD2: The two exons later discovered to encode two additional alternatively spliced FNIII domains. Figure not drawn to scale. [Compiled from www.genecards.org and www.ensembl.org]

The tenascin-C protein contributes to the integrity of the matrix through its profound ability to influence cell behaviour and its interaction with a diverse range of ECM components [216,217]. Although tenascin-C demonstrates a restricted pattern of expression in most adult tissues, it is constitutively expressed in tendons, ligaments and at the MTJ [218]. For these reasons, *TNC* is a plausible candidate for further interrogation into the genetic basis of musculoskeletal soft tissue injuries. The tenascin-C hexamer (hexabrachion) has six highly symmetrical arms (polypeptide chains) that radiate from a central core (tenascin assembly domain). The proximal portions of the arms are rigid, thin and are composed of epidermal growth factor-like (EGFL) repeats. The distal ends are thick, flexible and are composed of fibronectin type III (FNIII) domains, in addition to a globular fibrinogen domain (Figure 1.13) [219,220].

Tenascin-C is an intriguing molecule in the sense that it can exist as many different naturally occurring isoforms of itself; resulting from modifications taking place between transcription and post-translational levels. Interestingly, more than half of the 17 FNIII repeats are subject to alternative splicing. This complex level of gene regulation profoundly influences the function of tenascin-C. Specifically, the FNIII repeats demonstrate specificity for different sites of proteolytic cleavage, post-translational modification and the protein partners to which they bind [221]. Thus far, over 25 different molecules have been identified to bind to the tenascin-C hexamer including a range of cell surface receptors, proteoglycans and other soluble factors [218].

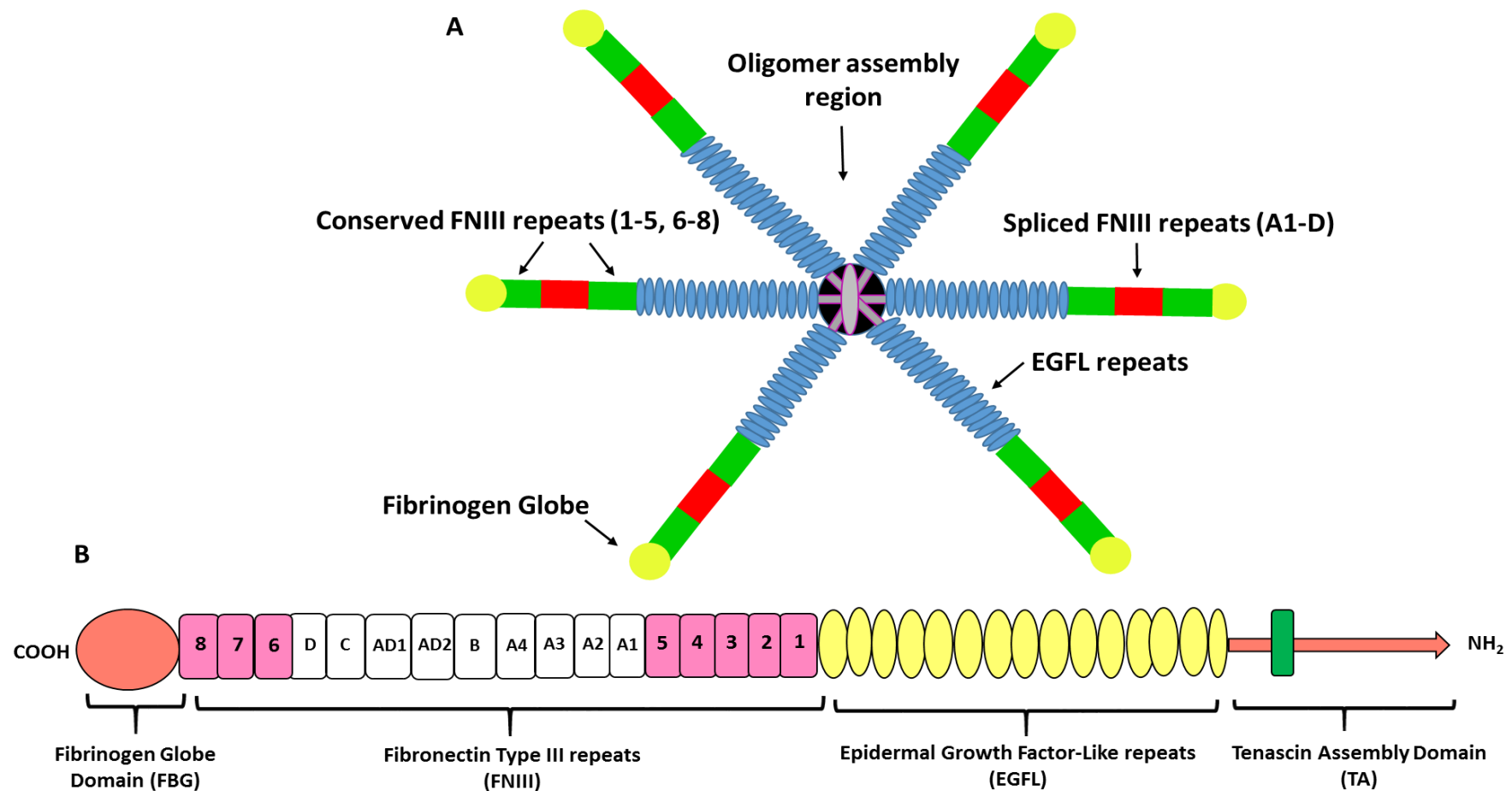


Figure 1.13: Structure of the tenascin-C hexamer

(A) Schematic diagram demonstrating the highly symmetrical structure of the tenascin- C hexabrachion. (B) Domain organisation of tenascin-C. The FNIII repeats (1-5 and 6-8) are constitutively expressed in all tenascin-C isoforms whereas the FNIII repeats, A1-D, are subject to alternative splicing. FNIII: Fibronectin type III repeats, EGFL: Epidermal growth factor-like repeats. COOH: Carboxyl terminal domain, NH₂: Amino terminal domain. [Complied from Jones & Jones (2000) [220] with permission from John Wiley and Sons]

The expression pattern of tenascin-C is under tight regulatory control with high levels detected during embryogenesis, transient expression during organogenesis and almost no expression in normal adult tissue. However, persistent tenascin-C expression is observed under pathological conditions including inflammation, tumorigenesis and infection [222,223]. Although predominantly absent from developed tissue, mechanical stress is a powerful regulator of tenascin-C [224,225]. This distinct mechanism of tenascin-C regulation supports its pattern of localised expression at sites subjected to mechanical loading including tendons and ligaments.

Tenascin-C has been described as an '*acute phase*' protein with its response to mechanical stress observed almost immediately [224]. Fluck et al. (2000) demonstrated the induction of tenascin-C expression within four hours of loading the chicken anterior latissimus dorsi muscle. After 36 hours, this expression was distributed along the surface of individual muscle fibres and remained elevated through seven days of loading but was significantly reduced on removal of the stimulus and disappeared completely within five days [226]. In agreement with these results, the re-exposure of unloaded rat soleus muscles to normal cage activity after a period of hindlimb suspension induced tenascin-C expression [227]. Similarly, Jarivinen et al. (1999) demonstrated that the removal of mechanical stress from the OTJ in rats; a site with high levels of tenascin-C expression under physiological loading, abolished expression almost completely [228]. However, upon re-exposure to mechanical loading, the immunoreactivity of tenascin-C was once again detected. This body of literature suggests that mechanical loading specifically and reversibly regulates the expression of tenascin-C *in vivo*.

Beside the regulation of tenascin-C expression, the action of this protein on a range of cells *in vitro* has been described as bewildering because its effects are often paradoxical in nature [223]. Although broadly classified as an adhesion-modulating protein, the function of tenascin-C extends beyond that of the classic anti-adhesive molecule. For example, tenascins may actively promote adhesion and spreading of one cell type and inhibit that of another [223]. It is recognised that tenascins interfere with integrin-dependent cell spreading. This process is however influenced by cell type and the cell-to-ECM contact [229]. The work by Chiquet-Ehrismann and co-workers has described a mechanism through which tenascin-C is proposed to regulate integrin signalling and, consequently, cell adhesion, cell spreading and

proliferation [223]. This mechanism involves the syndecan-4 heparan sulphate proteoglycan. For cells to spread completely on an adhesive molecule such as fibronectin, syndecan-4 is required in synergy with its coreceptor, integrin- $\alpha 5\beta 1$ [230]. According to this mechanism, tenascin-C prevents cells from interacting with fibronectin through syndecan-4 by actively binding both proteins [223,231]. Through this interference, tenascin-C may alter signalling pathways which are normally triggered upon the interaction of cells with adhesive ECM substrates [223]. Cells may still attach to fibronectin or other adhesive substrates via their integrin receptors but the formation of focal adhesions and synthesis of stress fibres is impaired [223,232,233].

The mechanical induction of tenascin-C does not require paracrine signalling factors such as transforming growth factor beta (TGFB) and mitogen-activated protein kinases (MAPKs). Instead, tenascin-C induction is dependent on actomyosin contractility, which is controlled by the RhoA/ROCK pathway [234]. Briefly, mechanical stimuli as sensed by cells, activates the RhoA/ROCK pathway through the integrin $\beta 1$ receptor and integrin-linked kinase (ILK) [235,236]. Depletion of the G-actin pool within the cytoplasm releases bound megakaryoblastic leukemia 1 (MKL1/MAL) which translocates into the nucleus and induces *TNC* transcription [237]. At this point, a negative feedback loop is established, whereby tenascin-C binds to both fibronectin and syndecan-4; inhibiting Rho-dependent signalling and consequently reducing focal adhesion and stress fibre formation, in addition to suppressing its own expression [237] (Figure 1.14).

With insight into the molecular mechanism through which this protein is modulated by mechanical stress, it is hypothesised that strained fibroblasts secrete tenascin-C to loosen their attachments to the ECM. In doing so, tenocytes release shear and tensile stress which may prevent overstretching [224]. Furthermore, tenascin-C is an elastic molecule that can be stretched several times its length *in vitro*; potentially contributing to the elasticity of the tissue in which it is expressed and preventing tissue damage through mechanical loading [238].

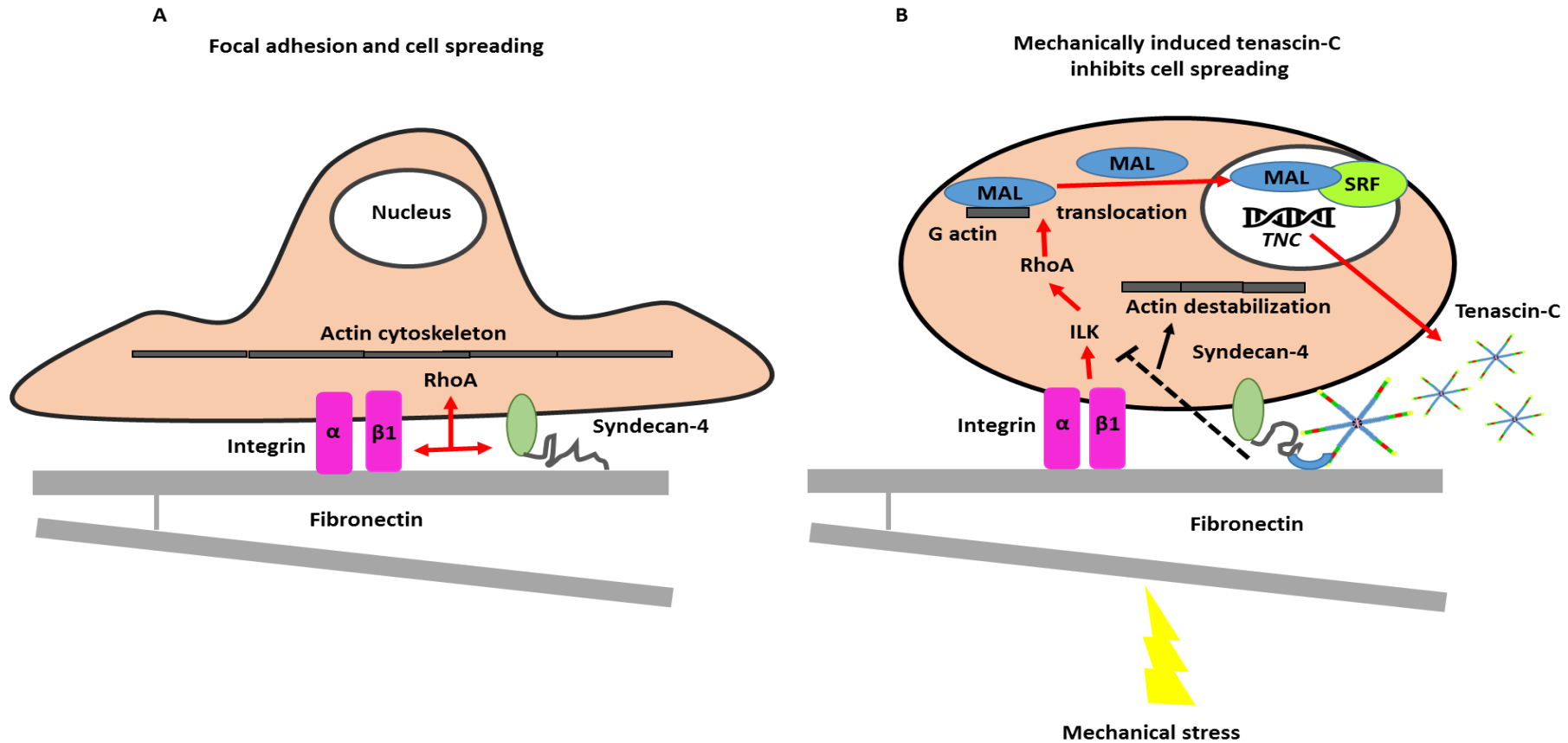


Figure 1.14: Proposed mechanism by which the mechano-induction of tenascin-C inhibits fibronectin-mediated cell spreading

(A) Fibroblasts adhere to fibronectin via integrin- $\alpha 5 \beta 1$ and syndecan-4. The synergistic binding of both these molecules leads to actin stress fibre assembly and cell spreading. (B) Mechanical stress activates the RhoA pathway. Depletion of the G-actin pool releases megakaryoblastic leukemia 1 (MKL1/MAL) which translocates into the nucleus and induces *TNC* expression. Tenascin-C binds syndecan-4 and fibronectin, inactivating RhoA signalling and consequently destabilises actin stress fibres and focal adhesions; reducing cell spreading. A negative feedback loop is established whereby tenascin-C represses its own expression through the inhibition of Rho-dependent signalling. *TNC*: gene encoding tenascin-C glycoprotein. ILK: Integrin-linked kinase. SRF: Serum response factor. [Compiled from Chiquet et al. (2003) [223] and Imanaka-Yoshida & Aoki (2014) [237]]

1.5.1.2. Type XXVII collagen

The *COL27A1* gene maps to the human chromosome 9 (chr9q32.33) and encodes the α_1 chain of type XXVII collagen (Figure 1.15). Expression of *COL27A1* is localised predominantly to cartilaginous structures [239]. This collagen, although classified as a fibrillar collagen, demonstrates features atypical of this group of proteins. Specifically, Pace et al. (2003) described three features specific to the pro- α_1 chain of (XXVII) collagen which included:

- (i) The presence of a single cysteine residue at position 633 (exon 3) and two tryptophan residues at positions 1392 (exon 45) and 1597 (exon 53)
- (ii) Two conserved imperfections of the triplet pattern at positions 679-687 (exon 6) and 1336-1339 (exon 42)
- (iii) Two splice variations within exon 16

With reference to gene organisation, *COL27A1* is similar to *COL24A1*, both of which are minor collagens. Although closely related to the other minor fibrillar collagens based on exon-intron organisation and C-propeptide sequence, these two collagens form a distinct phylogenetic clade [239].

Type XXVII collagen is one of the most recently identified members of the collagen family. However, based on its role in chondrogenesis and early skeletogenesis, in addition to its conservation in vertebrates, several groups have hypothesised that variants at this locus may undermine various chondrodysplasias [240,241]. In line with this, a founder mutation within *COL27A1* has recently been associated with Steel Syndrome, a Mendelian orthopaedic disorder characterised by skeletal abnormalities including bone dislocations, scoliosis and short stature [241]. However, no concomitant tendon and/or ligament injuries have been documented for this phenotype. In saying this, an inferred haplotype, G-C-A, spanning the *COL27A1* rs946053 (T>G) variant, together with the *TNC* 13321 (G>C) and rs2104772 (T>A) variants, was associated with increased risk for AT [198]. The *COL27A1* gene is still relatively uncharacterised in musculoskeletal soft tissue injuries. However, these studies, in addition to the proximity of *COL27A1* to *TNC* on chromosome 9, provided support for the prioritisation of this gene for further interrogation.

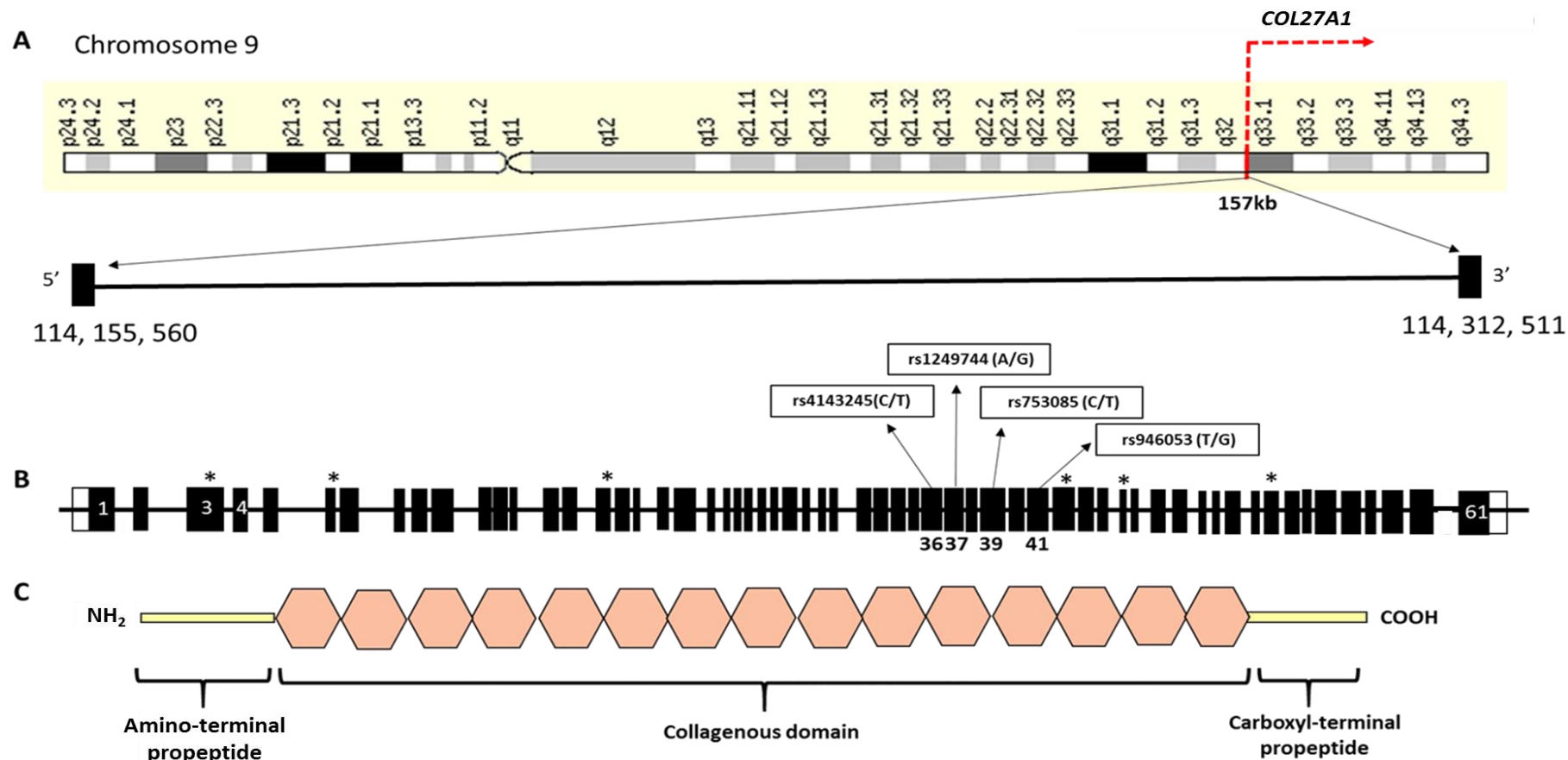


Figure 1.15: Gene encoding the α_1 chain of type XXVII collagen

(A) Cytogenetic banding pattern of chromosome 9. The red arrow indicates the location of the *COL27A1* gene on chromosome 9 and direction of transcription (sense strand). The genomic boundaries of *COL27A1* are provided according to the Human GRCh38.p10 genome release. (B) The exons 1-61 (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *COL27A1* are presented. Asterisk (*) indicate exons encoding known sequence variations specific to the *COL27A1* gene. Black arrows on the gene indicate variants previously tested for their potential association with the risk of chronic Achilles tendinopathy [198]. The rs946053 variant was associated with the risk of Achilles tendinopathy [198]. (C) Domain organisation of the α_1 chain of type XXVII collagen. COOH: Carboxyl terminal domain, NH₂: Amino terminal domain. Figure not drawn to scale.

[Compiled from www.genecards.org and Pace et al. (2003) [239]]

1.5.1.3. Matrix metalloproteinase 3

The *MMP3* gene is one of nine *MMP* genes clustered on the long arm of chromosome 11 (chr11q22) (Figure 1.16). This 8 kb gene encodes a zinc-dependent endopeptidase, also referred to as stromelysin 1 (MMP3) (Figure 1.17). When activated from its latent zymogen form, MMP3 has the ability to degrade several structural and modular proteins of the ECM [242], in addition to influencing the release of ECM-sequestered modulators of cell behaviour, such as growth factors, cytokines, chemokines and other bioactive agents [243]. Moreover, MMP3 may activate other MMP enzymes through cleavage of the propeptide unit [244]. For example, the activity of MMP3 on a partially processed proMMP1 molecule is critical for the generation of a mature MMP1 peptide, which subsequently digests the major fibrillar collagen types I-III and other matrix proteins [245]. Therefore, with its broad substrate specificity, MMP3 contributes to the control of cellular interactions in response to environmental stimuli; inducing conditions which promote connective tissue remodelling.

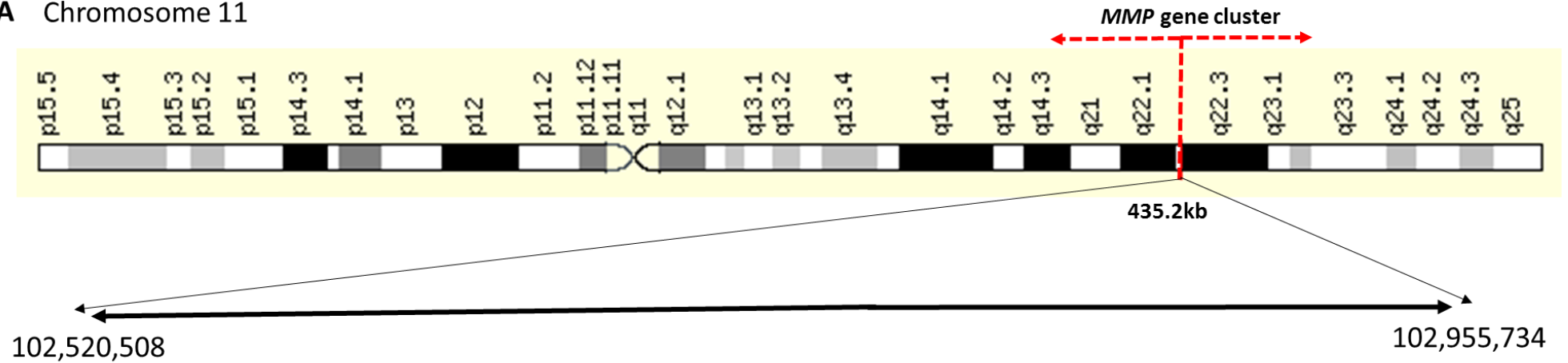
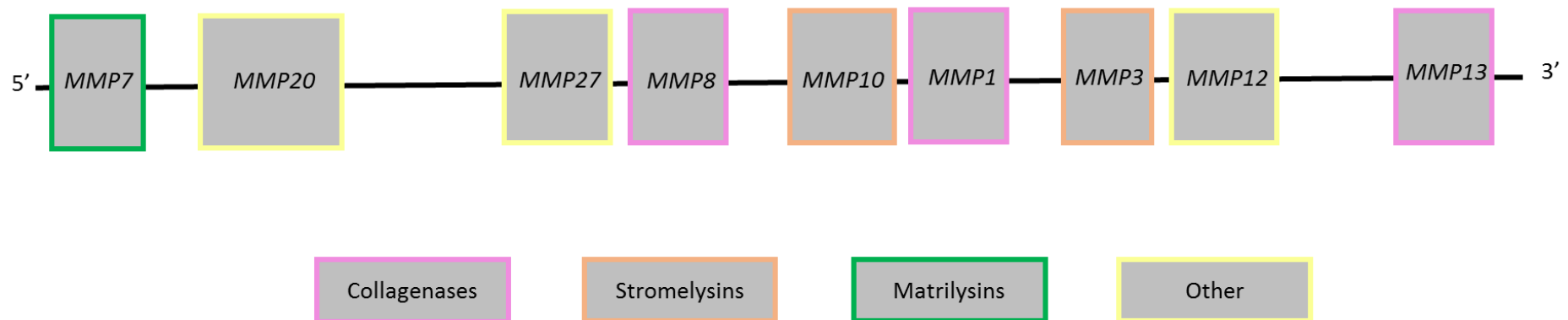
A Chromosome 11**B**

Figure 1.16: Matrix metallopeptidase gene cluster on chromosome 11q22

Matrix metallopeptidase gene cluster on the long arm of chromosome 11. Red arrows indicate the location of *MMP* genes on chromosome 11 and direction of transcription (sense or antisense strand). Only *MMP13* is encoded on the sense strand. The genomic interval encompassing the nine *MMP* genes is indicated. Matrix metallopeptidases are classified based on substrate preferences. Diagram not drawn to scale. [Compiled from www.genecard.org, www.ensembl.org and Somerville et al. (2003) [242]]

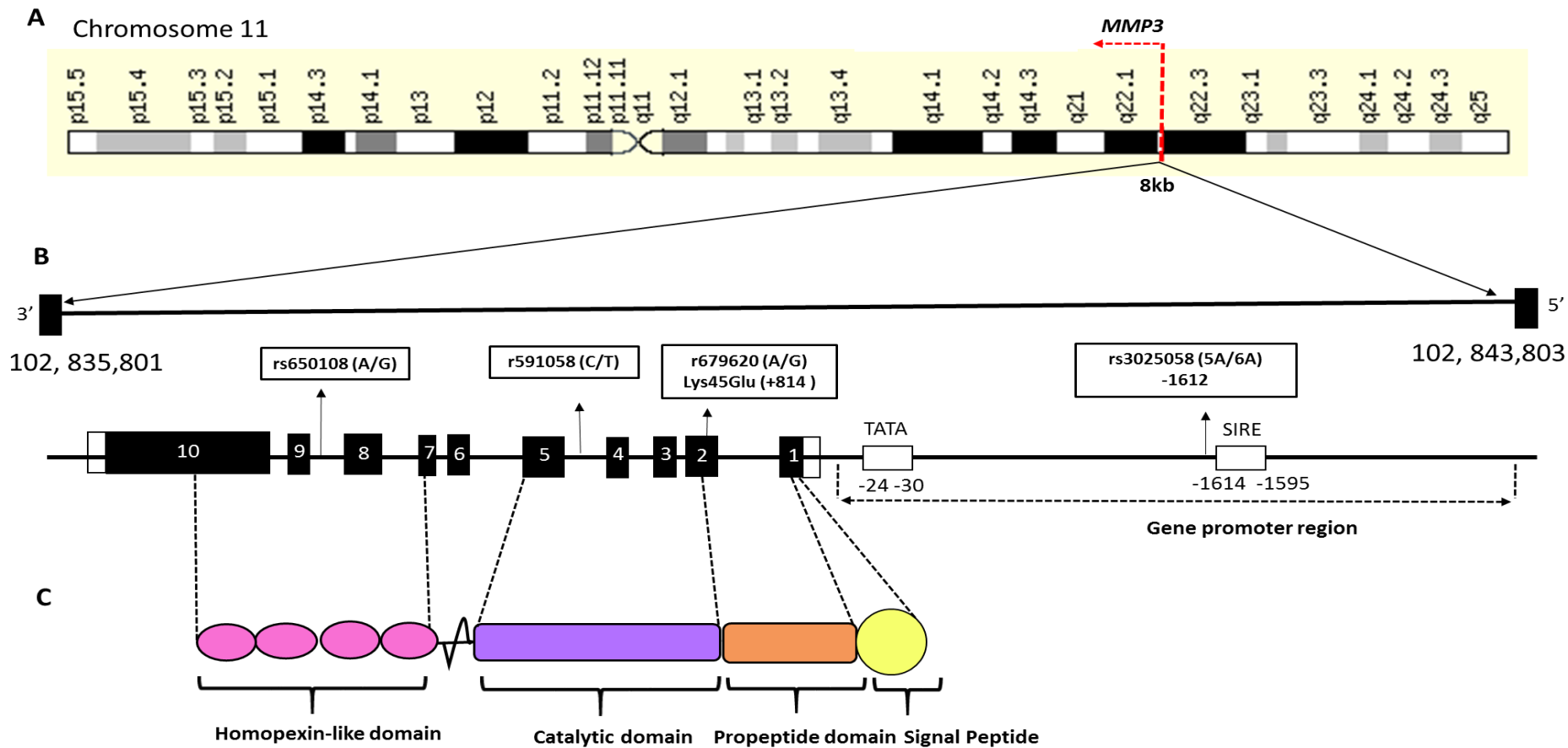


Figure 1.17: Gene encoding the MMP3 (Stromelysin 1) endopeptidase

(A) Cytogenetic banding pattern of chromosome 11. The red arrow indicates the location of the *MMP3* gene on chromosome 11 and direction of transcription (antisense strand). The genomic boundaries of *MMP3* are provided according to the Human GRCh38.p10 genome release. (B) The numbered exons 1-10 (black bars), introns (horizontal lines), untranslated regions (empty bars) and gene promoter (dashed line) of *MMP3* are presented. (-) Base pair number upstream from transcription start site. Black arrows on gene indicate the independent variants previously associated with the risk of chronic Achilles tendinopathy (rs679620, rs591058 and rs650108) [206], acute Achilles tendon ruptures (rs679620) [207] and ACL ruptures (rs3025058) [204]. TATA: TATA box, SIRE: Stromelysin interleukin-1 responsive element. (C) Domain organisation of the MMP3 peptide. Figure not drawn to scale. [Compiled from www.genegards.org and Ye et al. (2006) [246]]

Degenerative tendons demonstrate marked differences in their gene expression profile compared to healthy tendon specimens, with some studies reporting the greatest difference reflected in the level of *MMP3* expression [247,248]. Specifically, *MMP3* mRNA is observed to be downregulated in degenerative samples [247–249]. In contrast, significantly elevated and sustained *MMP3* mRNA levels were observed post ACL injury compared to ACL controls in the study by Beye et al. (2008) [250]. It is interesting to note that this study also reported marked differences in the levels of *MMP3* mRNA between the rabbit ACL and MCL post ligament transection; two ligaments known to have significantly different healing capacities within the knee joint. The basal expression of *MMP3* mRNA is much higher in the healthy ACL compared to the healthy MCL [251]. Therefore, it can be hypothesised that the differences observed between the ACL and MCL in the expression of key matrix remodelling factors such as *MMP3*, may predispose the ACL to injury and a poorer healing capacity post-trauma. Support for the theory that *MMP3* may contribute to ACL injury predisposition has been provided through several recent case-control genetic association studies.

The study by Raleigh et al. (2009) was the first to associate variants in the *MMP3* gene with the risk of musculoskeletal soft tissue injuries [206]. Specifically, significant differences in the frequency distribution of the rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants were noted between healthy control participants and cases with chronic AT in a South African White cohort. Furthermore, the three-variant haplotype, A-T-G (rs679620-rs591058-rs650108), was associated with decreased risk for AT. More recently, rs679620 was also associated with the risk of ATR in a self-identified British White cohort [207]. However, only 25 cases with acute injuries were analysed, therefore, these results would require validation in a larger cohort before any meaningful conclusions can be drawn. The interaction of the *MMP3* rs679620 variant with variants in other MMP genes on chromosome 11 has also been suggested to modify the risk of ACL ruptures [203]. Specifically, rs679620 in combination with rs485055 (*MMP10*; C>T), rs1799750 (*MMP1*; 1G>2G) and rs2276109 (*MMP12*; A>G) was suggested to modify the risk of ACL ruptures in a self-identified South African White cohort. Most recently, rs679620 demonstrated the strongest association with the risk of hamstring injuries in elite male soccer players of Spanish ancestry [252]. A total of 37 variants were investigated in this study; however, rs679620 was the only variant to associate with chronic, acute, severe and recurrent hamstring injuries. In line with this, the A allele of rs679620 was

also associated with the risk of post-operative stiffness, consequently decreasing range of motion after RCT in a Chinese Han population [253]. In addition, the functional *MMP3* promoter variant, rs3025058 (5A/6A), has also previously been associated with the risk of acute ACL ruptures. Specifically, the rs3025058 5A allele (5A5A+5A6A) was associated with increased risk of contact ACL ruptures in a cohort of Thai ethnicity [204].

Conversely to these findings, other studies have failed to identify these variants as potential modifiers of injury risk or associate these variants with connective tissue properties. Foster and colleagues (2014) concluded that the dimensional and mechanical properties of the patellar tendon *in vivo* are not influenced by the three variants investigated initially by Raleigh et al. (2009) [206,254]. In addition, the four variants contributing to the inferred haplotypes previously associated with ACL injury risk: rs485055 (*MMP10*; T>C), rs1799750 (*MMP1*; 1G/2G), rs679620 (*MMP3*; G>A) and rs2276109 (*MMP12*; T>C) were not associated with the risk of carpal tunnel syndrome [255]. Therefore, it is apparent that although some risk factors may be common, differences in the genetic profiles predisposing for clinically distinct musculoskeletal soft tissue injuries are also likely. Although differences in susceptibility have been noted, there remains a body of research supporting the potential involvement of the *MMP3* gene and its encoded protein in the development of musculoskeletal injuries. It is for these reasons, that the *MMP3* locus was prioritised for further interrogation in this thesis using an integrated approach.

1.5.1.4. Type I Collagen

Type I collagen is the predominant fibrillar collagen in tendons and ligaments and a heterotrimer consisting of two α_1 -(I) and one α_2 -(I) polypeptide chains. These α -chains are encoded by two different genes. Specifically, the 17.5 kb *COL1A1* gene (chr17q21.33) encodes the α_1 chains (Figure 1.18), whereas the 36.6 kb *COL1A2* gene (chr7q21.3) encodes the single α_2 chain [256,257]. Type I collagen is responsible for the hierarchical structure of non-cartilaginous tissues; providing tensile strength [28]. Rare mutations within the *COL1A1* gene have been associated with several severe musculoskeletal disorders including osteogenesis imperfecta [258,259], Ehlers Danlos Syndrome [260–262] and infantile cortical hyperostosis

[263]. In addition to these rare disorders, several complex multifactorial phenotypes have been associated with variants mapping to *COL1A1* and include among others osteoporotic fractures [264], osteoarthritis [265], urinary stress incontinence [266] and lumbar disc disease [267].

The most extensively investigated *COL1A1* variant, rs1800012 (G>T), maps within the first intron of the gene. This variant is known to locate in the binding site of the specificity protein-1 (Sp1) transcription factor. The first study to suggest this variant as a susceptibility marker for musculoskeletal soft tissue injuries reported the significant association of the rare TT genotype with the risk of shoulder dislocations and cruciate ligaments ruptures [184]. Based on these results, this variant was further investigated as a potential risk modifier for pathologies affecting the humerus (lateral epicondylitis or tennis elbow), the Achilles tendon (chronic AT and acute ATR) and the anterior cruciate ligament (ACL ruptures), in other independent cohorts [182,183,185–187,268–271]. The rs1800012 TT genotype was similarly found to associate with decreased risk of ACL ruptures [185], in addition to acute soft tissue injuries when cruciate ligament ruptures, shoulder dislocations and Achilles tendon ruptures were analysed collectively [187]. In accordance with these independent studies, a recent meta-analysis suggested that the rare rs1800012 TT was a protective factor for sports-related tendon and ligament injuries [272]. Conversely, this genotype was not observed to modulate the risk of chronic phenotypes including AT or tennis elbow [268,270] but was associated with increased risk for lumbar disc disease in two independent studies [267,273]. Interestingly, the study by Stępień-Słodkowska et al. (2013) reported the significant association of the alternate rs1800012 GG genotype with decreased risk of ACL ruptures sustained while skiing [186]. However, the mechanisms resulting in these ACL injuries are believed to be unique to the sport and different to the traditional non-contact mechanism, predominantly proposed to underpin sports-related ACL injuries [178]. Therefore, the level to which genetic factors influence these specific acute soft tissue injuries still requires further interrogation.

The other previously investigated *COL1A1* variant, rs1107946 (G>T), located within the distal promoter of the *COL1A1* gene is in high linkage disequilibrium with rs1800012. The rs1107946 variant has been associated with bone mineral density (BMD) and in studies of osteoporosis; an interaction with the Sp1 binding site variant has been reported [274]. In a recent study by

Stępień-Słodkowska et al. (2016), the GG genotype and G allele of this variant were associated with decreased risk for ACL ruptures sustained during skiing activities [182]. Furthermore, the G-T inferred haplotype (rs1107946 G>T- rs1800012 G>T) was associated with reduced risk for ACL ruptures in a cohort of professional football players [183]. These variants have also been explored functionally, with their effects proposed to influence DNA binding, gene transcription and protein production [275–277]. These studies will be discussed in greater detail in chapter 5.

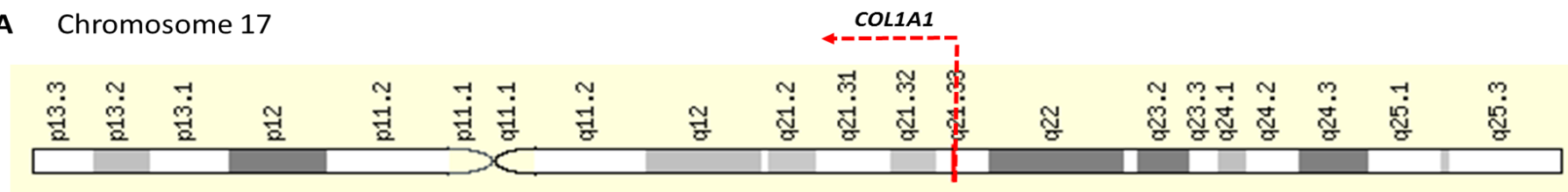
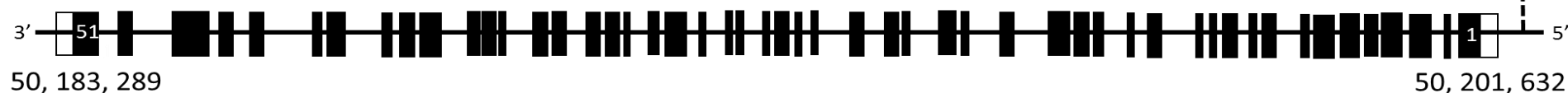
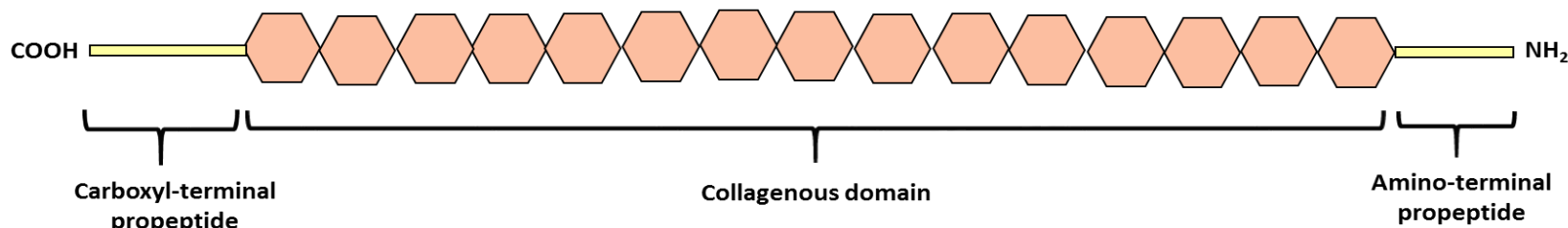
A Chromosome 17**B****C**

Figure 1.18: Gene encoding the α_1 chain of type I collagen

(A) Cytogenetic banding pattern of chromosome 17. The red arrow indicates the location of the *COL1A1* gene on chromosome 17 and direction of transcription (antisense strand). The genomic boundaries of *COL1A1* are provided according to the Human GRCh38.p10 genome release. (B) The exons 1-51 (black bars), introns (horizontal lines), untranslated regions (empty bars) and gene promoter (dashed line) of *COL1A1* are presented. Black arrows on gene demonstrate variants previously associated with the risk of musculoskeletal soft tissue injuries [182–187]. (-) Base pair number upstream from transcription start site; (+) base pair number downstream from transcription start site. (C) Domain organisation of the α_1 chain of type I collagen. COOH: Carboxyl terminal domain, NH₂: Amino terminal domain. Figure not drawn to scale. [Compiled from www.genecards.org]

Since the proposal of genetic factors as risk modifiers for musculoskeletal soft tissue injuries, the candidate gene, case-control genetic association study has been the predominant approach used to test the association of variants with the risk of injury susceptibility. This approach is hypothesis driven; investigating variants that are presented in the literature as being associated with biological processes known to be involved in the development of musculoskeletal soft tissue injuries. However, this approach has an important limitation. Our knowledge regarding the aetiology of musculoskeletal soft tissue injuries is limited. As a result, only a small number of genes within the genome have been prioritised for investigation and only a few variants mapping to these loci have been profiled in musculoskeletal soft tissue injuries. Furthermore, these association studies have demonstrated varying levels of replication success. Therefore, it is unlikely that all risk conferring loci predisposing for complex, multifactorial conditions such as AT and ACL ruptures will be identified using this method alone.

However, we have now entered the era of high throughput sequencing technologies, powerful tools able to address some of the limitations associated with the previous candidate variant approach. To date, only three studies have utilised these technologies to investigate the genetic signatures of common sporting injuries in humans. The details of these studies, the possible reasons for the slow uptake of sequencing in the field of sports genomics and a brief overview of several sequencing pipelines is provided below.

1.6. SEQUENCING TECHNOLOGIES

1.6.1. First generation sequencing

Sixteen years have elapsed since the completion of the first human whole genome sequence in 2001 [278,279], a project that cost approximately 70 million dollars. This project relied on the automation of Sanger sequencing, now considered a ‘first generation’ technology. In brief, linear nucleotide sequences are identified by electrophoretic separation of randomly terminated extension products [280]. To this day, Sanger sequencing remains the gold standard due to its accuracy. Despite this, Sanger sequencing is restricted in its application because of limitations in its technical workflow. Specifically, only a limited number of samples can be sequenced in parallel, reducing this pipeline’s functional throughput. Due to its dependency on electrophoretic separation of fragments, this is a lengthy process if the whole genome is to be sequenced [281]. Furthermore, the procedure is expensive, with an estimated cost of over 1 million dollars per genome. The time and expense associated with this method of sequencing has limited its accessibility and application. However, over the last ten to fifteen years, we have witnessed the evolution of automated sequencing; allowing for the in-depth exploration of the genome. These technologies, including genome-wide association studies (GWAS) and next generation sequencing (NGS) technologies have facilitated the exploration of genetic variation underpinning disease susceptibility but at a fraction of the cost and time of first generation sequencing. These tools will be discussed in further detail in the proceeding paragraphs.

1.6.2. Genome-wide association studies

Genome-wide association studies compare the difference in allele frequencies of variants across the genome between large groups of cases and controls. Using this method, hundreds of thousands of common variants are screened using microarray technologies [282]. For pathologies where the aetiology is well characterised, GWAS are sufficient to help further define the genetic basis of disease based on common variation. However, in disorders where

the aetiology is predominantly unknown, whole genome or exome sequencing is advantageous as they may elucidate signatures specific to the disease under investigation as opposed to a standard panel of common DNA sequence variants. Moreover, sequencing platforms are more equipped to identify rare coding variants and structural abnormalities [283].

Only recently have a small number of studies adopted the use of GWAS to further define the genetic risk profile of musculoskeletal injuries. The first GWAS study conducted by Baird et al. (2014) identified regions on the canine chromosomes 1, 3 and 33 associated with the risk of cranial cruciate ligament ruptures [284]. Two genes highlighted as interesting candidates included those encoding the sortilin related VPS10 domain containing receptor 2 (*SORCS2*) and semaphorin 5B (*SEMA5B*), both of which function within neurological pathways. These two genes are known to have human orthologs, suggesting the potential involvement of neurological signalling pathways in modulating the risk of ACL ruptures. The 99 loci associated with the risk of ACL ruptures in the most recent GWAS study by Baker et al. (2017), demonstrated little overlap with those highlighted in the previous study by Baird et al. (2014), albeit that both studies adopted domestic dog breeds to model cruciate ligament injuries [284,285]. It was proposed that these complex phenotypes are likely to be undermined by many genetic variants that may be concentrated differently in different dog breeds, similar to that observed when comparing susceptibility trait loci in ancestry independent human populations [285]. A similar study in humans identified two novel variants associated with the risk of degenerative RCT; rs820218 (G>A) in the gene encoding the SAP30-binding protein (*SAP30BP*) and rs12527089 (C>T) in the gene encoding the sterile motifs-and SH3 domain containing protein-1 (*SASH1*) respectively [286]. Both genes function within the cellular process of apoptosis, making them plausible candidates for further interrogation. Conversely, Kim and colleagues (2017) were unable to identify any risk conferring variants associated with either AT or ACL ruptures using a false discovery rate less than five percent [287]. Several limitations were however presented in this study. Firstly, all participants included in this study were mined from an electronic health record system. The International Classification of Disease (ICD) codes used to extract participants included amongst others, Achilles bursitis. Bursitis is the local inflammation of the bursa situated between the Achilles and calcaneal prominence, however it does not necessarily infer pathology of the Achilles tendon [288].

Similarly, ACL ruptures were not stratified into those that were non-contact in mechanism, injuries which are hypothesised to have a greater intrinsic factor contribution. Moreover, the mean age of participants was 62 years, which presents as a potential confounder because the risk of Achilles tendon injuries is known to increase with the advancement in age [85]. Therefore, a combination of factors may have contributed to the study failing to identify potential modifiers of risk; albeit the fact that a large injury cohort was investigated.

Even though GWAS have successfully identified many variants associated with a variety of phenotypic traits and diseases, these microarray-based genotyping platforms are limited in their resolution; meaning that a large proportion of the genome is not genotyped directly [289]. However, the advancement of next generation sequencing has significantly lowered the cost of sequencing, thereby providing a molecular tool for the interrogation of the genome at a base-pair resolution.

1.6.3. Next generation sequencing

Next generation sequencing, high throughput, deep or massively parallel sequencing are all terms used to describe DNA sequencing technologies which can sequence many different DNA molecules in a single reaction. Several NGS platforms are now commercially available, each presenting with a different design focus [281]. These technologies control the sequential addition of nucleotide bases to immobilised and spatially orientated DNA templates [281]. Briefly, the starting material (genomic DNA, cDNA, reverse-transcribed RNA) is converted into a library of reaction templates; a process that requires fragmentation, size selection and the ligation of adapters. Fragmentation and size selection involves the breaking of DNA templates into fragments, the size of which is specific to the sequencing platform. During the process of adaptor ligation, synthetic DNA sequences are added to the end of the library fragments, serving as primers for downstream amplification and the sequencing reactions. At this point, the library is sequenced either directly (single-molecule templates) or alternatively, amplified and then sequenced (clonal amplification templates). The initial analysis post sequencing involves base calling which is conducted by platform-specific software. Base calling is subsequently followed by the alignment of the sequenced data to a reference genome. Using

this approach, the entire genome or specific target regions are sequenced multiple times at a base-pair resolution to provide in-depth coverage of genetic variation [281].

1.6.3.1. Whole genome sequencing

Whole genome sequencing (WGS) is the most comprehensive and straightforward sequencing strategy. This method has the capacity to identify common single nucleotide polymorphisms (SNPs) and structural and copy number variations (CNVs) contributing to disease. However, large-scale WGS studies are still expensive, especially with respect to complex multifactorial phenotypes, in which large sample sizes are required to achieve sufficient power to detect significant associations using traditional statistical methods. Besides the expense associated with sequencing a large sample of well phenotyped participants, processing the raw data is a complex task, often requiring the assistance of an experienced bioinformatician. Furthermore, much of the genome remains relatively uncharacterised with many regions still of unknown function [290]. Therefore, exhausting financial resources to obtain a massive amount of data not immediately transferable to the pathology under investigation, has warranted careful consideration by many.

A cost effective and popular alternative to whole genome sequencing has involved targeting subregions of the genome, a targeted sequencing approach. The size of the target region may vary from a specific subset of genes known or suggested to be involved in the aetiology of the disease to the protein-encoding region of the genome, the exome. Moreover, the analysis of genome-sequenced data has predominantly resulted in the meaningful interpretation of DNA variation located within exonic structures of the genome [291]. Therefore, since many disease-associated variants have been identified in protein encoding regions; sequencing only the exome provides a cost-effective alternative to WGS.

1.6.3.2. Whole exome sequencing

Many laboratories are currently adopting next generation sequencing technologies and applying it to the small proportion (~2%) of the human genome encoding proteins

(approximately 230 000 exons). This approach requires an additional enrichment step whereby the coding region is selected from the total genome DNA and hybridised either in solution or on a microarray [292]. Although the data generated from sequencing platforms enriched for exonic regions is more manageable, exome sequencing is not without limitations. Firstly, although the depth of coverage is generally higher for exome-sequenced samples, it tends to be less uniform across the targeted region [293]. This non-uniform sequencing pattern may result in loss of sequencing power, in addition to the possibility of true risk-conferring variants being missed. Furthermore, whole exome sequencing (WES) is less equipped to detect large scale DNA rearrangements.

To date, only one study has used WES to identify genes or genetic variants associated with the risk of musculoskeletal injuries, specifically ACL ruptures. This family-based study included a pair of fraternal twins, both of whom had sustained spontaneous non-contact ACL ruptures [294]. It is interesting to note that, although several of the genes highlighted in this recent study are plausible candidates based on their biological function, none of them have previously been prioritised for interrogation using standard candidate variant driven approaches.

In all, it is evident from the literature that integrated approaches are required to facilitate the interrogation of the genome and assist in refining the biological signatures underpinning musculoskeletal soft tissue injuries. As far as we are aware, no studies have been published describing the application of NGS technologies to explore genetic variation underpinning chronic AT. Moreover, GWAS studies are still in their infancy in the field of musculoskeletal soft tissue injury genomics. Therefore, it is imperative that large-scale consortiums are established, which contain data of well-phenotyped individuals. It is through collaboration and the sharing of resources including sequenced data that we will be able to increase our understanding of the genetic loci contributing to injury predisposition.

1.7. LITERATURE REVIEW SUMMARY AND CONCLUSIONS

It is evident from the literature that the prevalence of musculoskeletal soft tissue injuries is on the rise worldwide. A plausible explanation for the observed increase in injury rates is the increase in the number of individuals participating in sports and physical activity from adolescents to senior citizens. This is likely due to the global promotion of physical activity to aid healthy living and longevity.

Several aetiological models have attempted to describe the underlying biological mechanisms of these injuries, focusing on key aspects of tissue remodelling. However, the inter-individual differences in symptom presentation and disease progression, in addition to the observed heterogeneity in response to treatment, suggests that the causative mechanisms of injury are still not fully understood. Furthermore, it is more likely the integration of several of these theories that best describes the aetiology of these complex multifactorial phenotypes as opposed to a single mechanism. The intrinsic and extrinsic risk factors have been presented in this chapter, with focus on the genomic loci proposed to predispose for injury. To date, the results from numerous candidate gene, case-control genetic association studies have been published; detailing the association of several DNA sequence variants with the risk of injury. However, these studies are known for their low reproducibility rate, which raises the question of whether they have truly identified risk-conferring variants. These studies are hypothesis driven, where the variants prioritised for interrogation are proposed to affect biological processes known to be involved in the development of musculoskeletal injuries or similar phenotypes.

Although the hypothesis driven candidate gene/variant approach holds merit and a certain level of overlap is expected between conditions affecting different connective tissues, musculoskeletal soft tissue injuries are complex disorders, presumably underpinned by complex genetic signatures, to which some loci may be unique. Next generation sequencing technologies have the capability to provide insight into the genetic basis underpinning specific traits and/or clinical phenotypes at a base pair resolution.

Currently, our knowledge regarding the biological mechanisms underpinning musculoskeletal soft tissue injury is limited. However, by prioritising genes previously implicated in the aetiology of injury, this project aimed to contribute to the body of evidence linking genetics to injury risk, by identifying other potentially functional, risk-associated markers using an integrated omics approach.

1.8. THESIS AIMS AND OBJECTIVES

The primary aim of this study was to further define the molecular signatures of musculoskeletal soft injuries mapping to specific genomic intervals encoding for several structural and regulatory components of the extracellular matrix. The *TNC* (chr9q33.1), *COL27A1* (chr9q32), *MMP3* (chr11q22) and *COL1A1* (chr17q21.33) genes have previously been associated with the risk of injury and were therefore prioritised for further interrogation. Previously associated variants within these regions or the new candidate variants identified by WES and prioritised through the application of a customised, tiered filtering strategy were genotyped in several previously recruited, self-identified White AT and ACL rupture cohorts. The second aim of this study was to determine whether the observed risk-associated motifs in the self-identified White cohorts were similar to those underpinning injury in an ancestrally admixed sample, using ACL ruptures in a South African Coloured cohort as the phenotypic model.

The specific objectives of this thesis were:

- To select well-phenotyped participants (10 asymptomatic controls and 10 exemplar AT cases) for sequencing using an extended whole exome sequencing platform (Chapter 2).
- The development and application of a reusable bioinformatics analysis pipeline, involving a tiered filtering strategy to select candidates for interrogation from the list of WES identified variants (Chapter 2). The *TNC* (chr9q33.1), *COL27A1* (chr9q32), *MMP3* (chr11q22) and *COL1A1* (chr17q21.33) genes were prioritised for further interrogation using this approach.

- To test the association between the selected candidate variants and the risk of two musculoskeletal phenotypes including chronic AT and acute ACL ruptures, using a case-control genetic association study design (Chapters 3 to 5). Candidates included those selected from the list of variants identified by WES: *TNC* rs1061494 (T>C), rs2104772 (T>A) and rs1061495 (T>C) and *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T), in addition to several variants previously associated with the risk of injury: *TNC* rs1138545 (C>T), *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) and *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) respectively.
- To functionally annotate the prioritised variants using a host of *in silico* bioinformatic analyses tools (Chapter 6).

CHAPTER 2: RESEARCH PARTICIPANTS AND STUDY METHODS

2.1. RESEARCH PARTICIPANTS

In total, five cohorts, which included banked DNA samples and information of ancestry matched control and case participants previously recruited for ongoing research within our research group, were used for the studies presented in this thesis (Table 2.1). Electronic capturing of all participant information involved the allocation of a unique code to each participant to ensure personal confidentiality and anonymity of personal information and banked DNA samples. The period of recruitment varied according to the cohort, with samples drawn independently from South Africa (SA), Australia (AUS) and the United Kingdom (British, UK) using the recommendations outlined by the genetic association study STREGA initiative [295]. This study was approved by the Human Research Ethics Committee of the Faculty of Health Sciences within the University of Cape Town (HREC references: 224/2013, 176/2015, 289/2004, 086/2005, 370/2013 and 124/2006) and the institutional review boards of the other international collaborating institutions (Appendix IA). Prior to participation, participants were provided with information regarding the nature of the study (Appendix IB), after which written informed consent was obtained in accordance with the Declaration of Helsinki (Appendix IC). All participants completed a questionnaire detailing personal particulars and medical history as well as their current and previous sports participation (Appendix ID).

Table 2.1: Overview of the (i) sequencing sample set and (ii) five genotyping cohorts used to investigate the genetic basis of chronic Achilles tendinopathy and anterior cruciate ligament ruptures using a case-control genetic association study design

Injury cohorts	Ethnicity	Recruited participants	Thesis chapter
(I) Sequencing sample set (SEQ)	White	10 SAW-CON ^{SEQ} , 10 SAW-TEN ^{SEQ}	Chapter 2, Chapter 3
(II) Genotyping cohorts			
<u>Achilles tendon injuries</u>			
(A) South African White (SAW)	White	165 SAW-CON ^{AT} , 123 SAW-TEN, 47 SAW-RUP	Chapter 3 (Section 3.2.2), Chapter 5 (Section 5.2.1)
(B) British (UK)	White	130 UK-CON, 87 UK-TEN, 25 UK-RUP	Chapter 3 (Section 3.2.2), Chapter 5 (Section 5.2.1)
(C) Australian (AUS)	White	210 AUS-CON, 85 AUS-TEN	Chapter 4 (Section 4.2.1)
<u>ACL ruptures</u>			
(D) South African White (SAW)	White	232 SAW-CON ^{ACL} , 234 SAW-ACL (135 NON)	Chapter 3 (Section 3.2.3), Chapter 4 (Section 4.2.2) Chapter 5 (Section 5.2.2)
(E) South African Coloured (SAC)	Coloured [†]	100 SAC-CON, 97 SAC-ACL (50 NON)	Chapter 3 (Section 3.2.3), Chapter 5 (Section 5.2.3)

Sequencing sample set: Sample of ten controls (SAW-CON^{SEQ}) and ten cases (SAW-TEN^{SEQ}) selected from the SAW-CON^{AT} and SAW-TEN groups for whole exome sequencing. Genotyping cohorts: South African White Achilles tendon injury cohort; British Achilles tendon injury cohort and Australian Achilles tendon injury cohort; SAW-ACL: South African White anterior cruciate ligament rupture cohort; SAC-ACL: South African Coloured anterior cruciate ligament rupture cohort. TEN: Cases with chronic Achilles tendinopathy, RUP: Cases with acute Achilles tendon ruptures, ACL: Cases with anterior cruciate ligament ruptures, NON: Subgroup of ACL cases reporting a non-contact mechanism of injury. [†] An admixed South African population with a complex ancestral history (Section 2.1.1).

2.1.1. South African cohorts

Three independent South African cohorts were included for genetic analysis in this thesis. The first cohort was recruited between 2004 and 2012 including participants of self-identified South African White (SAW) ancestry. This cohort consisted of 165 apparently healthy, asymptomatic controls (SAW-CON^{AT}, 104 males), 123 participants with clinically diagnosed chronic Achilles tendinopathy (SAW-TEN, 86 males) and 47 participants with acute Achilles tendon ruptures (SAW-RUP, 36 males). The recruitment criteria of this cohort have previously been described [199].

The second South African cohort also including self-identified SAW participants and was recruited between 2006 and 2013. This cohort consisted of 232 asymptomatic control participants (SAW-CON^{ACL}, 130 male) and 234 individuals with surgically diagnosed ACL ruptures (SAW-ACL, 171 males). Approximately 60% (n=135) of the SAW-ACL participants reported a non-contact mechanism of injury and were therefore additionally analysed as a subgroup (SAW-NON, 102 males). The recruitment criteria of this cohort have previously been described [185,201]. A third South African cohort was recruited between 2012 and 2016 and included participants who self-identified as being of mixed ancestry, specifically, South African Coloured (SAC) [211]. This cohort is representative of a population indigenous to the Western Cape region of South Africa with a complex ancestral history. This population is derived from the admixture of one or more of the indigenous African populations (Khoen- and San-speaking or Bantu-speaking) with immigrants from Western Europe or slave labourers from West Africa, Indonesia, Madagascar, Java, India and Malaysia [296,297]. The term '*Coloured*' in South Africa is therefore a name that encompasses a wide range of people who are unique to the country. This cohort consisted of 100 appropriately matched, asymptomatic and apparently healthy control participants (SAC-CON, 81 males) and 97 individuals with surgically diagnosed ACL ruptures (SAC-ACL, 80 males). Approximately 50% (n=50) of the participants in the SAC-ACL rupture diagnostic group reported injuries non-contact in mechanism (SAC-NON, 44 males).

The control participants within the three South African cohorts were recruited from various sports clubs and gyms around the greater Cape Town area. Whereas, case participants in the

first two cohorts were recruited from the Sports Science Orthopaedic Clinic at the Sports Science Institute of South Africa and other clinical practices in Cape Town. The participants in the SAC-ACL rupture diagnostic group were recruited from Cape Town's Groote Schuur and Victoria Hospitals. An experienced sports physician confirmed the diagnoses of degenerative AT (Appendix IE) and acute ACL ruptures. Briefly, a positive diagnosis for AT involved the patient reporting gradual but progressive pain of the posterior lower leg over an extended period (95%, n=105), in addition to fulfilling one of several other diagnostic criteria including (i) the early morning onset of pain and discomfort (48%, n=59), (ii) stiffness (67%, n=83), (iii) prolonged swelling of the Achilles tendon structure and surrounds (51%, n=63), (iv) palpation induced tenderness, (85%, n=104) (v) palpable nodular thickening of affected tendon (40%, n=49) and (vi) a positive shift test (17%, n=23). To reduce the effect of confounding factors, participants who reported either treatment with fluoroquinolone antibiotics or the local administration of corticosteroids were excluded. In addition, participants diagnosed with any connective tissue or systemic diseases known to influence the incidence of AT and/or ATR were excluded [199]. The diagnosis of ACL ruptures was confirmed using magnetic resonance imaging (MRI), arthroscopy or during reconstructive surgery. All SAW-ACL and SAC-ACL participants were asked to provide a descriptive account of the inciting event which had resulted in injury. The mechanism of injury was later classified by the investigator according to the ACL classification scheme as devised by the American Orthopaedic Society for Sports Medicine [177]. In line with this system, the information regarding sports participation was used to stratify participants by mechanism of injury.

2.1.1.1 Descriptive, medical and injury characteristics

(A) South African Achilles tendon injury cohort

The descriptive characteristics of participants within the SAW-Achilles tendon injury cohort are presented in Table 2.2. The participants in the SAW^{AT}-CON group were significantly younger at the time of recruitment (36.5 ± 10.9 years) compared to the age of participants in the SAW-TEN (41.5 ± 14.0 years, $P=0.002$) and SAW-RUP (40.4 ± 11.9 years, $P=0.038$) groups at the time of injury. Consequently, age was adjusted for in all logistic regression models. The mean age difference between age at the time of recruitment and age at the time of injury for

the SAW-TEN and SAW-RUP groups was 6.9 ± 8.4 years and 6.8 ± 8.4 years respectively. There were no significant differences in the distribution of sex between the SAW-CON^{AT} and SAW-TEN groups ($P=0.222$) and similarly between the SAW-CON^{AT} and SAW-RUP groups ($P=0.082$). The average height of participants in the SAW-CON^{AT} group was not significantly different to that of cases in the SAW-TEN and SAW-RUP groups ($P=0.161$ and $P=0.472$ respectively). However, the SAW-CON^{AT} participants weighed significantly less at the time of recruitment (72.1 ± 11.9 kg) compared to the cases in the SAW-TEN (78.6 ± 17.5 kg, $P<0.001$) and SAW-RUP (86.0 ± 15.3 kg, $P<0.001$) groups and remained significantly lighter after adjusting for age and sex ($P<0.001$). Consequently, and after adjustment, the BMI of the SAW-CON^{AT} group (23.5 ± 2.8 kg/m²) was significantly lower than that of the SAW-TEN (24.8 ± 3.3 kg/m², $P=0.002$) and SAW-RUP (27.8 ± 4.0 kg/m², $P<0.001$) groups. The proportion of participants who reported South Africa as their country of birth was not significantly different between the SAW-CON^{AT} group and the SAW-TEN and SAW-RUP groups ($P=0.998$ and $P=0.960$ respectively).

The average number of years involved in high-impact sports was significantly different between the participants in the SAW^{AT}-CON (16.4 ± 16.1 years) and SAW-TEN groups (31.1 ± 29.5 years, $P<0.001$) and similarly between the SAW-CON^{AT} and SAW-RUP (28.3 ± 26.7 years, $P<0.001$) groups (Table 2.3). Specifically, the number of years of running experience prior to recruitment was significantly less in the SAW-CON^{AT} (10.6 ± 7.6 years) group compared to the SAW-TEN (14.8 ± 9.0 years, $P=0.002$) and SAW-RUP (15.5 ± 11.1 years, $P=0.019$) groups prior to injury. Similarly, the number of years spent playing racket sports including squash, tennis and badminton was significantly greater in the SAW-TEN (23.6 ± 18.5 years) and SAW-RUP (24.9 ± 17.2 years) groups compared to the SAW-CON^{AT} group (12.8 ± 10.3 years, $P<0.001$) (Table 2.3). However, the number of hours per week involved in high-impact sports was not significantly different between the SAW-CON^{AT} and SAW-TEN groups ($P=0.317$) including the number of hours spent running ($P=0.640$) and playing racket sports ($P=0.922$). Conversely, participants in the SAW-CON^{AT} group spent a significantly greater number of hours per week running (4.5 ± 2.5 hours/week) and playing racket sports per week (4.7 ± 2.6 hours/week) compared to cases in the SAW-RUP group (Running: 2.8 ± 1.8 hours/week, $P=0.013$; Racket sports: 3.3 ± 2.0 hours/week, $P=0.025$) (Table 2.3).

The majority of the SAW-TEN cases suffered bilateral injuries (38%, n=41) compared to unilateral injuries of the left (34%, n=37) or right leg (28%, n=31). Furthermore, the mid-portion of the tendon was the site most commonly associated with painful degenerative changes (70%, n=57) compared to the distal (27%, n=22) and proximal regions (3%, n=4). Only five cases (11%) reported the spontaneous rupture of both tendons, compared to 42 cases reporting unilateral injuries (Left: 42%, n=20; Right: 47%, n=22). A total of 46 SAW-TEN (42%) and 40 SAW-RUP (87%) cases reported grade IV injuries; injuries that required athletes to significantly alter their training patterns or stop sports participation all together.

Table 2.2: Descriptive characteristics of the South African White participants in the control (SAW-CON^{AT}), Achilles tendinopathy (SAW-TEN) and acute Achilles tendon rupture (SAW-RUP) groups

	SAW-CON ^{AT}	SAW-TEN	P-values	SAW-RUP	P-values
	<i>165</i>	<i>123</i>		<i>47</i>	
Age (years)	36.5 ± 10.9 (158)	41.5 ± 14.0 (104)	0.002	40.4 ± 11.9 (46)	0.038
Sex (% male)	63.4 (164)	70.5 (122)	0.222	77.0 (47)	0.082
Height (cm)	174.9 ± 9.1 (158)	176.5 ± 8.8 (107)	0.161	176.0 ± 8.2 (46)	0.472
Body mass (kg) ^a	72.1 ± 11.9 (163)	78.6 ± 17.5 (113)	<0.001 (<0.001)	86.0 ± 15.3 (47)	<0.001 (<0.001)
BMI (kg/m ²) ^a	23.5 ± 2.8 (154)	24.8 ± 3.3 (84)	0.002 (0.002)	27.8 ± 4.0 (44)	<0.001 (<0.001)
Born SA (%)	74.2 (163)	74.6 (122)	0.998	78.3 (46)	0.960

Values are expressed as mean ± standard deviation; sex and country of birth (Born SA) are represented as percentages. Participant count (n) for each variable presented in parentheses with total count in italics. Body mass, height and body mass index (BMI) are self-reported values at the time of recruitment for control and cases. Age is reported at the time of recruitment for controls and at the time of injury for cases. P-values (CON vs. TEN, CON vs. RUP) in bold typeset indicate significance (P<0.05). Covaried P-values in parentheses. ^a Body mass and BMI adjusted for age and sex.

Table 2.3: Self-reported sports participation by the South African White participants in the control (SAW-CON^{AT}), Achilles tendinopathy (SAW-TEN) and Achilles tendon rupture (SAW-RUP) groups

	SAW-CON ^{AT}	SAW-TEN	P-values	SAW-RUP	P-values
	<i>165</i>	<i>123</i>		<i>47</i>	
High impact sports (Years)	16.4 ± 16.1 (127)	31.1 ± 29.5 (75)	<0.001	28.3 ± 26.7 (40)	<0.001
High impact sports (Hours/week)	5.1 ± 4.8 (127)	4.4 ± 5.1 (75)	0.317	4.5 ± 4.6 (26)	0.551
Distance running (Years)	10.6 ± 7.6 (97)	14.8 ± 9.0 (56)	0.002	15.5 ± 11.1 (19)	0.019
Distance running (Hours/week)	4.5 ± 2.5 (93)	4.7 ± 2.2 (44)	0.640	2.8 ± 1.8 (15)	0.013
Racket sports (Years)	12.8 ± 10.3 (101)	23.6 ± 18.5 (69)	<0.001	24.9 ± 17.2 (36)	<0.001
Racket sports (Hours/week)	4.7 ± 2.6 (95)	4.6 ± 2.2 (48)	0.922	3.3 ± 2.0 (23)	0.025

Values are expressed as mean ± standard deviation for the number of years and the number of hours per week involved in sport. Participant count (n) for each variable presented in parentheses with total count in italics. The number of hours/week value reflects that two years prior to recruitment for controls and two years prior to injury for cases. P-values (CON vs. TEN and CON vs. RUP) in bold typeset indicate significance (P<0.05).

(B) SAW-ACL rupture cohort

The descriptive characteristics of participants within the SAW-ACL rupture cohort are provided in Table 2.4. The SAW-CON^{ACL} participants were significantly older (29.9 ± 11.8 years) at the time of recruitment compared to participants in the SAW-ACL group (26.0 ± 10.8 years, $P=0.002$) and SAW-NON subgroup (26.3 ± 10.3 years, $P=0.004$). Moreover, the participants in the SAW-ACL group were 4.6 ± 8.0 years older and 1.2 ± 4.0 kg heavier at the time of recruitment than at the time of the first ACL rupture. A significantly larger proportion of SAW-ACL cases (73%, $n=171$) including those reporting a non-contact mechanism of injury (75.6%, $n=102$) were males compared to the SAW-CON^{ACL} group (56%, $n=130$, $P<0.001$). Participants in the SAW-CON^{ACL} group were significantly shorter than participants in the SAW-ACL group ($P<0.001$) and SAW-NON subgroup ($P<0.001$). However, after adjusting for the larger proportion of male cases, this difference was no longer significant (SAW-ACL: $P=0.512$ and SAW-NON: $P=0.224$). The SAW-CON^{ACL} participants weighed significantly less at the time of recruitment (73.5 ± 14.5 kg) compared to case participants in the SAW-ACL group (80.5 ± 16.9 kg, $P<0.001$) and SAW-NON subgroup (80.4 ± 16.1 kg, $P<0.001$). This difference remained significant after adjusting for age and sex ($P=0.002$ and $P=0.008$ respectively). In line with this observation, BMI was also significantly different between the diagnostic groups ($P<0.001$) and remained significantly different after adjusting for age and sex (SAW-ACL: $P=0.001$, SAW-NON: $P=0.011$). The proportion of participants that reported South Africa as their place of birth was similar between the SAW-CON^{ACL} and SAW-ACL groups ($P=0.185$) and between the SAW-CON^{ACL} group and SAW-NON subgroup ($P=0.441$).

Analysis of the data showed that a significantly larger proportion of males played contact sports compared to female participants (70% vs. 12%, $P<0.001$) (Table 2.5). Contact sports included: rugby, touch rugby, soccer, Gaelic football, Australian football, boxing and Muay Thai. Furthermore, the number of years' experience in contact sports was also significantly higher in male participants ($P<0.001$). Conversely, the proportion of male participants involved in non-contact jumping sports such as netball, volleyball, basketball, handball, ballet and gymnastics was significantly lower than that of female participants (12% vs. 41%, $P<0.001$) but the number of years' experience was similar ($P=0.224$). A similar proportion of

male and female participants participated in non-contact non-jumping, such as road running, cycling, swimming and field hockey ($P=0.620$) and were matched for number of years' experience ($P=0.387$).

Sex specific analysis showed that a significantly greater proportion of male SAW-ACL cases participated in contact and non-contact jumping sports compared to male SAW-CON^{ACL} participants (Contact: 81% vs. 55%, $P<0.001$; Non-contact jumping: 16% vs. 7%, $P=0.019$). However, the number of years' experience in these disciplines was not significantly different between these groups (Contact: $P=0.870$; Non-contact jumping: $P=1.000$). The number of years' experience in non-contact non-jumping sports was significantly greater in the male SAW-CON^{ACL} group compared to that in the male SAW-ACL group ($P=0.037$), although the proportion of male cases and controls playing non-contact non-jumping sports was not significantly different ($P=0.245$). The female SAW-CON^{ACL} and SAW-ACL participants were matched for sports participation including the proportion and number of years' experience in contact ($P=0.411$, $P=0.165$), non-contact jumping ($P=0.111$, $P=0.720$) and non-contact non-jumping sports ($P=0.436$, $P=0.756$).

Table 2.4: Descriptive characteristics of the South African White participants in the control (SAW-CON^{ACL}) group, anterior cruciate ligament rupture (SAW-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAW-NON)

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
	<i>232</i>	<i>234</i>		<i>135</i>	
Age (years)	29.9 ± 11.8 (228)	26.0 ± 10.8 (206)	0.002	26.3 ± 10.3 (129)	0.004
Sex (% Males)	56.0 (232)	73.0 (234)	<0.001	75.6 (135)	<0.001
Height (cm) ^a	174.5 ± 9.5 (224)	177.4 ± 9.8 (213)	<0.001 (0.512)	178.2 ± 9.1 (131)	<0.001 (0.224)
Body mass (kg) ^b	73.5 ± 14.5 (227)	80.5 ± 16.9 (213)	<0.001 (0.002)	80.4 ± 16.1 (132)	<0.001 (0.008)
BMI (kg/m ²) ^b	23.8 ± 4.0 (222)	24.7 ± 5.9 (207)	<0.001 (0.001)	24.7 ± 5.4 (128)	<0.001 (0.011)
Born SA (%)	85.7 (223)	80.9 (215)	0.185	82.6 (132)	0.441

Values are expressed as mean ± standard deviation; sex and country of birth (Born SA) are represented as percentages. Participant count (n) for each variable presented in parentheses with total count in italics. Age, body mass and body mass index (BMI) are self-reported values at the time of recruitment for control participants and at the time of the initial injury for participants within the ACL group and NON-subgroup. P-values (CON vs. ACL and CON vs. NON) in bold typeset indicate significance (P<0.05). Covaried P-values in parentheses. ^a Height is adjusted for sex. ^b Body mass and BMI adjusted for age and sex.

Table 2.5: Self-reported sports participation according to type of activity for the South African White participants in the control (SAW-CON^{ACL}) and anterior cruciate ligament rupture (SAW-ACL) groups

	Male vs. Female	Male participants		P-values	Female participants		P-values
	P-values	SAW-CON ^{ACL}	SAW-ACL		SAW-CON ^{ACL}	SAW-ACL	
		<i>130</i>	<i>171</i>		<i>102</i>	<i>63</i>	
Contact sports							
Participation (%)	<0.001	55 (72)	81 (140)	<0.001	14 (14)	10 (6)	0.411
Participation (years)	<0.001	12.6 ± 7.8	12.8 ± 7.9	0.870	5.3 ± 3.2	8.5 ± 7.4	0.165
Non-contact jumping sports							
Participation (%)	<0.001	7 (9)	16 (27)	0.019	36 (37)	49 (31)	0.111
Participation (years)	0.224	9.7 ± 5.2	9.7 ± 9.4	1.000	11.7 ± 7.3	11.2 ± 5.4	0.720
Non-contact non-jumping sports							
Participation (%)	0.620	87 (113)	82 (141)	0.245	83 (85)	79 (50)	0.436
Participation (years)	0.387	31.0 ± 34.84	23.6 ± 20.6	0.037	27.6 ± 21.3	26.4 ± 21.1	0.756

Values are expressed as a percentage with the number of participants (n) in parentheses and the total count in italics. The number of years participation is presented as a mean ± standard deviation. P-values (Male vs. Female and CON vs. ACL) in bold typeset indicate significance (P<0.05). Contact sports: Rugby, touch rugby, soccer, Gaelic football, Australian football, boxing and Muay Thai. Non-contact jumping sports: Netball, volleyball, basketball, handball, ballet, gymnastics, motocross and skateboarding. Non-contact non-jumping sports: Running, cycling, squash, swimming, waterpolo, cricket, tennis, field hockey, horse-riding, aerobics, yachting, athletics (excluding long and triple jump), golf, canoeing, surfing, windsurfing, badminton, gym training, bowls, triathlon and softball.

A total of 22 cases (17 males) reported bilateral ACL ruptures. Cases also reported other injuries associated with the primary ACL trauma including meniscal damage (42%, n=89) bone bruising (36%, n=76) and other knee ligament injuries (13%, n=28). A significantly greater proportion of female and male SAW-ACL participants reported ligament injuries compared to male and female SAW-CON^{ACL} participants (F: 47% vs. 23%, $P=0.017$; M: 46% vs. 38%, $P=0.020$) (Table 2.6). Similarly, the number of knee ligament injuries (excluding ACL) reported by SAW-ACL participants (F: 12%, M: 13%) was significantly greater compared to participants in the SAW-CON^{ACL} group; all of whom reported no previous knee ligament injuries ($P<0.001$) (Table 2.6). Furthermore, the familial prevalence of ligament injuries reported by SAW-ACL cases at the time of recruitment was significantly higher compared to that reported by SAW-CON^{ACL} participants (F: 49% vs. 29%, $P=0.010$; M: 42% vs. 29%, $P=0.002$). Participants in the female SAW-NON subgroup (10%), male SAW-ACL group (9%) and male SAW-NON subgroup (11%) reported a significantly greater proportion of Achilles tendon injuries compared to their respective SAW-CON^{ACL} groups, for which no Achilles tendon injuries were reported ($P\leq 0.001$). However, the proportion of participants that reported a family history of Achilles tendon injuries was not significantly different between the male and female SAW-ACL and SAW-CON^{ACL} groups.

Table 2.6: Self-reported history of musculoskeletal soft tissue injuries for the South African White participants in the control (SAW-CON^{ACL}) group, anterior cruciate ligament rupture (SAW-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAW-NON)

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
Female participants	<i>102</i>	<i>63</i>		<i>33</i>	
Ligament injury [†]	23 (98)	47 (58)	0.017	50 (32)	0.004
Bilateral ACL injury	-	8 (59)	<0.001	9 (33)	0.010
Knee ligament injury [‡]	-	12 (58)	<0.001	19 (32)	<0.001
Achilles tendon injury	-	5 (57)	0.092	10 (31)	0.001
Family ligament injury [*]	29 (98)	49 (51)	0.010	56 (32)	0.004
Family tendon injury [*]	9 (98)	14 (56)	0.443	13 (31)	0.551
Male participants	<i>130</i>	<i>171</i>		<i>102</i>	
Ligament injury [†]	38 (120)	46 (159)	0.020	40 (102)	0.653
Bilateral ACL injury	-	11 (158)	<0.001	13 (101)	<0.001
Knee ligament injury [‡]	-	13 (159)	<0.001	9 (102)	<0.001
Achilles tendon injury	-	9 (160)	<0.001	11 (102)	<0.001
Family ligament injury [*]	29 (118)	42 (159)	0.002	36 (101)	0.300
Family tendon injury [*]	6 (118)	11 (159)	0.117	9 (101)	0.409

Values are expressed as a percentage with the number of participants (n) whom had provided the relevant information in parentheses. Total number of female and male participants in italics. P-values (CON vs. ACL and CON vs. NON) in bold typeset indicate significance ($P < 0.05$).

[†] Excludes ACL injuries

[‡] Includes injuries of the lateral collateral ligament, medial collateral ligament and posterior cruciate ligament

^{*} Family members who are first degree relatives

(C) SAC-ACL rupture cohort

The descriptive characteristics of participants within the SAC-ACL rupture cohort are presented in Table 2.7. Except for age, no significant differences were noted for any of descriptive characteristics between participants in the SAC-CON group, SAC-ACL group and SAC-NON subgroup. Participants in the SAC-CON group were significantly older (27.4 ± 6.9 years) at the time of recruitment compared to the age of cases in the SAC-ACL group (24.6 ± 7.0 years, $P=0.007$) and SAC-NON subgroup (24.5 ± 7.8 years, $P=0.022$) at the time of injury. The participants in the SAC-ACL group were 2.5 ± 4.4 years older and 0.9 ± 9.6 kg heavier at the time of recruitment than at the time of the first ACL rupture.

It must be noted that the female representation in this group was insufficient to allow for independent analysis. Therefore, the results presented for sports participation and injury characteristics only reflect those for all participants (male and female) and the male subgroup. Analysis of the sports participation data showed that the SAC-CON and SAC-ACL groups were similarly matched for participation in contact and non-contact jumping sports, as well as the number of years' experience in these activities (Table 2.8). However, a significantly larger proportion of SAC-CON participants (All: 76%, $n=76$; Male: 73%, $n=59$) were involved in non-contact non-jumping sports compared to cases in the SAC-ACL groups (All: 52%, $n=50$, $P<0.001$; Male: 54%, $n=43$, $P=0.011$).

The majority of ACL ruptures were sustained whilst participating in team sports, such as soccer (39%, $n=36$) and rugby (35%, $n=32$). Only six cases reported bilateral ACL ruptures. However, other ACL-associated injuries were reported including meniscal damage (41%, $n=36$), bone bruising (26%, $n=24$) and other knee ligament injuries (2%, $n=2$). The SAC-CON participants (29%, $n=29$), together with the male SAC-CON subgroup (31%, $n=25$) sustained a significantly larger proportion of ligament injuries (excluding ACL) compared to SAC-ACL participants (All: 13%, $n=12$, $P=0.009$; Males: 14%, $n=10$, $P=0.008$) (Table 2.9). However, the familial prevalence of ligament injuries was significantly higher in the SAC-ACL group (35%, $n=31$) compared to the SAC-CON group (22%, $n=22$, $P=0.049$).

Table 2.7: Descriptive characteristics of the South African Coloured participants in the control (SAC-CON) group, anterior cruciate ligament rupture (SAC-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAC-NON)

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
	<i>100</i>	<i>97</i>		<i>50</i>	
Age (years)	27.4 ± 6.9 (99)	24.6 ± 7.0 (90)	0.007	24.5 ± 7.8 (50)	0.022
Sex (%Males)	81 (100)	82 (97)	0.790	88 (44)	0.280
Height (cm)	172.2 ± 8.3 (95)	173.7 ± 8.6 (85)	0.226	174.3 ± 8.8 (47)	0.168
Body mass (kg)	77.2 ± 15.6 (94)	77.7 ± 15.8 (87)	0.846	74.0 ± 13.0 (49)	0.215
BMI (kg/m ²)	25.9 ± 4.3 (93)	25.7 ± 4.7 (81)	0.791	24.3 ± 3.0 (46)	0.189
Born SA (%)	99 (100)	99 (89)	0.934	98 (46)	0.584

Values are expressed as mean ± standard deviation; sex and country of birth (Born SA) are represented as percentages. Participant count (n) for each variable presented in parentheses with total count in italics. Age, body mass and body mass index (BMI) are self-reported values at the time of recruitment for control participants and at the time of initial injury for participants within the ACL group and NON subgroup. P-values (CON vs. ACL and CON vs. NON) in bold typeset indicate significance (P<0.05).

Table 2.8: Self-reported sports participation according to type of activity for the South African Coloured participants in the control (SAC-CON) and anterior cruciate ligament (SAC-ACL) rupture groups

	All participants			Male participants		
	SAC-CON	SAC-ACL	P-values	SAC-CON	SAC-ACL	P-values
	<i>100</i>	<i>97</i>		<i>81</i>	<i>80</i>	
Contact sports						
Participation (%)	78 (78)	79 (77)	0.814	90 (73)	88 (70)	0.599
Participation (years)	16.8 ± 20.6	15.2 ± 8.8	0.540	15.2 ± 9.7	15.8 ± 8.9	0.727
Non-contact-jumping sports						
Participation (%)	16 (16)	15 (15)	0.918	9 (7)	10 (8)	0.768
Participation (years)	12.3 ± 9.3	10.2 ± 8.4	0.509	6.3 ± 3.3	7.5 ± 5.7	0.619
Non-contact non-jumping sports						
Participation (%)	76 (76)	52 (50)	<0.001	73 (59)	54 (43)	0.011
Participation (years)	19.1 ± 25.7	14.8 ± 13.7	0.283	15.4 ± 11.2	15.2 ± 14.7	0.935

Values are expressed as a percentage with the number of participants (n) in parentheses and the total count in italics. The number of years participation is presented as a mean ± standard deviation. P-values (CON vs. ACL) in bold typeset indicate significance (P<0.05). Contact sports: Rugby, touch rugby, soccer, Gaelic football, Australian football, boxing and Muay Thai. Non-contact jumping sports: Netball, volleyball, basketball, handball, ballet, gymnastics, motocross and skateboarding. Non-contact-non-jumping sports: Running, cycling, squash, swimming, waterpolo, cricket, tennis, field hockey, horse-riding, aerobics, yachting, athletics (excluding long and triple jump), golf, canoeing, surfing, windsurfing, badminton, gym training, bowls, triathlon and softball.

Table 2.9: Self-reported history of musculoskeletal soft tissue injuries for the South African Coloured participants in the control (SAC-CON) group, anterior cruciate ligament rupture (SAC-ACL) group and subgroup of participants reporting a non-contact mechanism of injury (SAC-NON)

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
All participants	<i>100</i>	<i>97</i>		<i>50</i>	
Ligament injury [†]	29 (100)	13 (89)	0.009	17 (48)	0.104
Bilateral ACL injury	-	7 (91)	0.775	8 (50)	0.773
Knee ligament injury [‡]	-	2 (89)	0.132	4 (48)	0.138
Achilles tendon injury	-	3 (89)	0.063	4 (48)	0.138
Family ligament injury [*]	22 (100)	35 (89)	0.049	36 (47)	0.068
Family tendon injury [*]	8 (100)	6 (89)	0.359	6 (47)	0.730
Male participants	<i>81</i>	<i>80</i>		<i>44</i>	
Ligament injury [†]	31 (81)	14 (74)	0.008	14 (42)	0.043
Bilateral ACL injury	-	7 (76)	0.793	7 (44)	0.792
Knee ligament injury [‡]	-	1 (74)	0.295	2 (42)	0.163
Achilles tendon injury	-	4 (74)	0.066	5 (42)	0.120
Family ligament injury [*]	25 (81)	36 (73)	0.139	41 (41)	0.055
Family tendon injury [*]	9 (81)	7 (73)	0.680	7 (41)	0.803

Values are expressed as a percentage with the number of participants (n) whom had provided the relevant information in parentheses. Total number of participants in italics. P-values (CON vs. ACL and CON vs. NON) in bold typeset indicate significance ($P < 0.05$).

[†] Excludes ACL injuries

[‡] Includes injuries of the lateral collateral ligament, medial collateral ligament and posterior cruciate ligament

^{*} Family members who were first degree relatives

2.1.2. Australian cohort

The AUS-Achilles tendon injury cohort included in this thesis consisted of asymptomatic control participants and cases with chronic Achilles tendinopathy, all of whom self-identified as being of Australian White ancestry. This study consisted of 210 apparently healthy, control participants (AUS-CON, 84 males) as well as 85 participants with clinically diagnosed chronic AT (AUS-TEN, 62 males). Participants with Achilles tendinopathy were recruited by the Musculoskeletal Research Centre at La Trobe University in Melbourne, Australia. One sports clinician confirmed the diagnoses of all AT cases using soft tissue ultrasound examinations. Exclusion criteria were the same as those described for the SAW-Achilles tendon injury cohort. The participants in the AUS-Achilles tendon injury cohort was recruited several years prior to the initiation of the current research project. Consequently, the DNA quality and quantity was low preventing the genotyping of all prioritised variants.

2.1.2.1. Descriptive, medical and injury characteristics

The descriptive characteristics of participants within the Australian cohort are presented in Table 2.10. The age at recruitment of the AUS-CON group did not differ significantly from the age at injury of the AUS-TEN group ($P=0.267$). The participants in the AUS-TEN group were however 8.8 ± 9.9 years older at the time of recruitment than at the time of injury. The distribution of sex was significantly different, with a significantly greater proportion of male cases (73%, $n=62$) compared to the male controls (40%, $n=84$, $P<0.001$). Participants in the AUS-TEN groups were also significantly taller (174.0 ± 9.9 cm) than participants in the AUS-CON group (171.6 ± 9.2 cm, $P=0.045$). However, this difference was not significant after adjusting for sex ($P=0.057$). Similarly, the participants in the AUS-TEN group were significantly heavier (80.5 ± 14.0 kg) and demonstrated a greater BMI (26.5 ± 3.9 kg/m²) at the time of recruitment compared to the AUS-CON participants (Mass: 72.7 ± 14.2 kg, $P<0.001$; BMI: 24.6 ± 3.9 kg/m², $P<0.001$). The diagnostic groups were however matched for mass and BMI after covarying for sex and age ($P=0.240$, $P=0.236$ respectively). Furthermore, the proportion of

control and case participants who reported Australia as their country of birth was not significantly different ($P=0.373$).

The medical and sporting history of the recruited Australian participants was unfortunately not obtained at the time of recruitment. However, most of the cases suffered bilateral AT (55%, $n=44$) compared to unilateral injuries of the left (24%, $n=20$) or right leg (25%, $n=21$). Furthermore, the majority of the cases (92%) suffered from mid-portion AT, whereas only 6% suffered from AT in the distal portion of the tendon.

Table 2.10: Descriptive characteristics of the Australian White participants in the control (AUS-CON) and Achilles tendinopathy (AUS-TEN) groups

	AUS-CON	AUS-TEN	P-values
	<i>210</i>	<i>85</i>	
Age (years)	38.5 ± 12.4 (205)	40.4 ± 14.2 (78)	0.267
Sex (% male)	40 (209)	73 (85)	<0.001
Height (cm) ^a	171.6 ± 9.2 (207)	174.0 ± 9.9 (82)	0.045 (0.057)
Body Mass (kg) ^b	72.7 ± 14.2 (208)	80.5 ± 14.0 (85)	<0.001 (0.240)
BMI (kg/m ²) ^b	24.6 ± 3.9 (207)	26.5 ± 3.9 (82)	<0.001 (0.236)
Born AUS (%)	82 (205)	77 (84)	0.373

Values are expressed as mean ± standard deviation; sex and country of birth (Born AUS) are presented as percentages. Participant count (n) for each variable presented in parentheses with total count in italics. Body mass, height and body mass index (BMI) are self-reported values at the time of recruitment for controls and cases. Age is reported at the time of recruitment for controls and at the time of injury for cases. P-values (CON vs. TEN) in bold typeset indicate significance ($P<0.05$). Covaried P-values in parentheses. ^a Height is adjusted for sex. ^b Body mass and BMI are adjusted for age and sex.

2.1.3. British cohort

The UK-Achilles tendon injury cohort included in this thesis included 130 apparently healthy, appropriately matched controls (UK-CON, 82 males) and 112 participants with clinically diagnosed Achilles tendon injuries. A total of 87 participants reported chronic AT (UK-TEN, 51 males) and 25 participants reported acute ATR (UK-RUP, 17 males). Recruitment of participants was conducted between 2011 and 2014. Control participants were recruited from the East Midlands region of the UK, whereas affected individuals were recruited from The County Clinic in Northampton, UK [298]. All participants within this cohort self-identified to be of British White ancestry. From the original cohort ($n=267$), 25 participants were excluded based on the exclusion criteria described by Mokone et al. (2005) [189].

2.1.3.1. Descriptive, medical and injury characteristics

The descriptive characteristics of participants within the UK-Achilles tendon injury cohort are presented in Table 2.11. The age at recruitment of participants in the UK-CON group (41.6 ± 11.7 years) was significantly younger than that at the time of injury for the UK-TEN group (45.2 ± 14.4 years, $P=0.043$) but matched the age at injury for the UK-RUP group (39.2 ± 10.4 years, $P=0.344$). The participants in the UK-TEN and UK-RUP groups were 8.3 ± 8.6 years and 10.8 ± 11.3 years older at the time of recruitment than at the time of injury respectively. At time of recruitment, the UK-TEN and UK-RUP participants were also 1.6 ± 6.0 kg and 5.0 ± 7.3 kg heavier than at the time of injury. There were no significant differences in the distribution of sex between the UK-CON and UK-TEN groups and between the UK-CON and UK-RUP groups ($P=0.511$ and $P=0.641$). Participants in the UK-TEN group were significantly shorter (172.3 ± 8.5 cm) than participants in the UK-CON group (174.9 ± 10.4 cm) after adjusting for sex ($P=0.021$). However, the UK-CON and UK-RUP groups were similarly matched for height ($P=0.834$). No significant differences were noted for body mass or BMI between the UK-CON and affected groups (UK-TEN: $P=0.261$, $P=0.984$; UK-RUP: $P=0.913$, $P=0.683$). Similar proportions of SAW-CON, SAW-TEN and SAW-RUP participants reporting the UK as their place of birth. ($P=0.840$ and $P=0.106$).

Unfortunately, very few case participants provided a detailed description of their sporting history prior to injury. However, at the time of recruitment, the number of hours per week spent running was not significantly different between the UK-CON (5.2 ± 4.0 hours/week, $n=30$) and UK-TEN (5.15 ± 4.4 hours/week, $n=24$, $P=0.963$) groups and likewise, between the UK-CON and UK-RUP (4.5 ± 1.7 hours/week, $n=4$, $P=0.183$) groups. Similarly, the number of hours per week spent playing racket sports including squash, tennis and badminton was not significantly different between participants in the UK-CON group (4.2 ± 3.2 hours/week), UK-TEN group (3.5 ± 1.7 hours/week) and UK-RUP group (2.4 ± 2.2 hours/week) ($P=0.473$, $P=0.240$ respectively).

The medical history of participants in the UK-Achilles tendon injury cohort has previously been discussed [298]. Briefly, the majority of UK-TEN cases suffered from bilateral Achilles injuries (44%, $n=39$) compared to unilateral injuries of the left (22%, $n=19$) or right leg (32%, $n=28$). Insertional AT was diagnosed in 29% of the cases ($n=25$) compared to the 55% ($n=48$) of cases reporting non-insertional injuries. A total of 14 UK-TEN cases (16%) demonstrated tendinopathic changes at both insertional and non-insertional sites. Cases in the UK-RUP group were diagnosed with partial ($n=7$) or complete ruptures ($n=18$). Nine participants with surgically diagnosed ATR were also diagnosed with AT but remained in the UK-RUP group for the purpose of analyses.

Table 2.11: Descriptive characteristics of the British White participants in the control (UK-CON), Achilles tendinopathy (UK-TEN) and acute Achilles tendon rupture (UK-RUP) groups

	UK-CON	UK-TEN	P-values	UK-RUP	P-values
	<i>130</i>	<i>87</i>		<i>25</i>	
Age (years)	41.6 ± 11.7 (123)	45.2 ± 14.4 (87)	0.043	39.2 ± 10.4 (25)	0.344
Sex (% male)	63.1 (130)	59.0 (87)	0.511	68 (25)	0.641
Height (cm)	174.9 ± 10.4 (123)	172.3 ± 8.5 (84)	0.056	174.4 ± 11.4 (25)	0.834
Body mass (kg)	80.3 ± 19.5 (124)	77.1 ± 15.0 (62)	0.261	79.8 ± 15.4 (20)	0.913
BMI (kg/m ²)	25.9 ± 4.5 (123)	25.9 ± 4.1 (62)	0.984	26.4 ± 3.9 (20)	0.683
Born UK (%)	96.0 (124)	96.5 (86)	0.840	88.0 (25)	0.106

Values are expressed as mean ± standard deviation; sex and country of birth (Born UK) are presented as percentages. Participant count (n) for each variable presented in parentheses with total count in italics. Age, body mass, height and body mass index (BMI) are self-reported values at the time of recruitment for controls and at the time of injury for cases. P-values (CON vs. TEN, CON vs. RUP) in bold typeset indicate significance (P<0.05).

2.2. SAMPLE COLLECTION AND DNA EXTRACTION

Through venipuncture of the forearm vein, each participant in the South African and Australian injury cohorts provided approximately 5 millilitres of venous blood, which was collected in an ethylenediaminetetraacetic acid (EDTA) vacutainer tube. The blood samples from the South African cohorts were stored at -20°C until total DNA extraction was performed using the protocol described by Lahiri and Nurnberger (1991) and later modified by Mokone et al. (2006) (Appendix IIA) [190,299]. With respect to the Australian cohort, DNA was extracted from the participant's blood using a sequenced extraction technique (FlexiGene DNA Kit, Qiagen P/L, Valencia, California, USA) as per the manufacturer's recommendations.

Approximately 2 millilitres of saliva was collected from each participant in the British cohort using the Oragene™ OG-500 DNA collection tubes. Participant DNA was subsequently extracted using the prepIT DNA extraction kit (DNA Genotek Inc., Ontario, Canada) as previously described [298]. Samples yielded A_{260}/A_{280} ratios between 1.8 and 2.0. All DNA extractions were conducted at the University of Northampton (UK).

2.3. WHOLE EXOME SEQUENCING

2.3.1. Selection of exemplar cases and controls

The injuries and physical activity participation of participants within the previously recruited SAW-Achilles tendon injury cohort have been well characterised. For this reason, chronic AT was considered the most appropriate phenotype for the application of WES in the exploration of the genetic basis of musculoskeletal soft tissue injuries.

From the original SAW-CON^{AT} and SAW-TEN groups, a subset of controls (SAW-CON^{SEQ}, n=10, 6 males) and cases (SAW-TEN^{SEQ}, n=10, 6 males), representing divergent extremes of the phenotypic spectrum were selected for WES. The selected SAW-CON^{SEQ} participants were unrelated, physically active, older than 47 years of age at the time of recruitment and had not sustained any tendon or ligament injuries prior to recruitment. All SAW-TEN^{SEQ} participants described an early onset of initial symptoms (35 years of age or younger) and suffered chronic bilateral mid-portion AT, with some reporting recurrent Achilles tendon injuries. Genomic DNA extracted from the whole blood of these 20 exemplar participants was prepared in accordance with the requirements of the WES service provider (Otogenetics Corporation Inc., Atlanta, USA).

2.3.2. Sample preparation

Total DNA was extracted from the blood of the SAW-CON^{SEQ} and SAW-TEN^{SEQ} participants using the protocol as previously described [190]. The DNA samples were stored in Tris-EDTA buffer (pH 8.0). All samples demonstrated DNA concentrations > 50 ng/μl with A₂₆₀/A₂₈₀ ratios ≥ 1,8 and A₂₆₀/A₂₃₀ ratios ≥ 1,9 (Appendix IIB). Prior to processing, all samples were screened for degradation. A 7 μl reaction containing 5 μl of unamplified DNA and 2 μl SYBR® Green nucleic acid gel stain (Invitrogen Molecular Probes™, Oregon, USA) was prepared for each sample before electrophoresis on a 2% agarose gel at 120 V for 30 minutes. Products were visualised under UV light using an Uvitec photodocumentation system (Uvitec Limited,

Cambridge, UK) (Appendix IIB). A total of 5 µg of each sample was prepared for Otogenetics. All samples passed Otogenetics' initial quality screening.

2.3.3. Sequencing and post-sequencing data processing

Paired end sequencing of samples was performed on the Illumina HiSeq 2000/2500 system at 30X coverage using the Agilent V5+UTR (71 Mbp) capture kit (Agilent Technologies, Santa Clara, USA). This platform has an average read length of 106 bp. The Agilent V5+UTR capture kit extends to target all exonic regions, in addition to untranslated regions and miRBase miRNA genes. The quality of sequencing reads was determined using FastQC (Babraham Bioinformatics, UK). All sequenced samples passed pre- and post-alignment quality control with no trimming of reads required. However, some samples did not achieve the required sequencing depth and were re-sequenced on the same platform.

For each sample, forward and reverse sequencing reads were aligned to the hg19 (GRCh37) human genome build using Novoalign (Novocraft Technologies, Selangor, Malaysia). For samples that were re-sequenced due to insufficient coverage during the initial sequencing run, the sequence alignment map (SAM) files were merged using Samtools 'reheader' and Picard 'MergeSamFiles' [300]. For condensed storage and further analysis, samtools was used to convert SAM files to BAM files using Picard (<http://broadinstitute.github.io/picard/>). BAM files are the binary version of SAM files. Several post alignment processing steps were conducted. These included marking PCR duplicates using Picard 'MarkDuplicates', as well as the realignment of reads around known insertions and deletions using the Genome Analysis ToolKit (GATK) 'RealignerTargetCreator' and 'IndelRealigner' (<http://software.broadinstitute.org/gatk/>). Mate-pair information was verified using Picard 'FixMateInformation', and bases were subsequently recalibrated using GATK 'BaseRecalibrator'. Depth of coverage statistics were determined for each sample using the GATK 'DepthOfCoverage' tool on post-alignment BAM files. Variants were called in all samples simultaneously using GATK 'HaplotypeCaller'. Variant call scores were then recalibrated using GATK 'VariantRecalibrator' and 'ApplyRecalibration'. Quality evaluation was performed using the GATK hard-filtering guidelines to produce a set of high-likelihood variants for each

exome. The resulting variant call format (vcf) file was annotated using the web version of ANNOVAR (wANNOVAR)[301].

2.3.4. Variant filtering

Due to the small sample size, variants with missing genotype data in two or more samples were excluded (n=1 916). Furthermore, variants called with more than two alleles at any locus were extracted and stored for analysis at a later stage (n=3 435). Given the complex, multifactorial nature of AT, and the fact that it is a common condition which lacks a clear inheritance pattern, no inheritance-based filters were applied to variant calls. Allele frequencies in the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups were calculated for each variant. An allele frequency difference of $\geq 30\%$ between the sequencing groups was selected as a cut-off to identify variants of interest for further analysis. Hereafter, this 30% allele frequency difference between sequenced cases and controls will be referred to as the “*allele frequency difference threshold*”. This conservative threshold was selected based on previous observations that a genotype or allele frequency difference of 10-20% between case and control groups in a genetic association study typically indicated a significant association using traditional statistical methods. Furthermore, acknowledging that (i) the aetiology of AT is complex and multifactorial, (ii) a certain percentage of risk-associated variants are likely to be common to both case and control participants and (iii) only a small number of exemplar case and control samples were sequenced (n= 20), an allele frequency difference of $\geq 30\%$ was used to prioritised variants for further interrogation.

2.3.5. Targeted gene analysis and variant prioritisation

All variant calls localising to genes previously associated with the risk of musculoskeletal soft tissue injuries were extracted from the WES dataset and noted. However, only four genomic regions previously associated with the risk of injury formed the focus of this thesis. These regions including the *TNC* (chr9q33.1), *COL27A1* (chr9q32), *MMP3* (chr11q22.2) and *COL1A1* (chr17q21.33) genes, were fine mapped using a customised, tiered filtering strategy (Figure

2.1). Given the limitations of the previous candidate variant approach, whereby previous literature was predominantly used to determine variant prioritisation within a gene, this new approach instead focused on screening for variants demonstrating potential functional properties which also met the allele frequency difference threshold. The previous association with a clinical phenotype and/or biological process proposed to be involved in the aetiology of musculoskeletal injuries supported the interrogation of a candidate but was not the primary determinant. Similarly, variants that localised to any of the 35 genes proposed as strong candidates for AT in a recently published *in silico* study using the BioOntological Relationship Graph (BORG) semantic database were also extracted and stored for interrogation at a later stage [302]. Variants that met the allele frequency difference threshold within the target genes were annotated using a host of freely available online tools and resources including the Ensembl (<http://www.ensembl.org/index.html>) and National Centre for Biotechnological Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>) genome browsers, LDlink [303] and the Geneious v9.1.5 (Biomatters, Auckland, New Zealand) next generation sequencing analysis tool. The general annotation of these regions allowed for the prioritisation of a subset of gene variants (Table 2.12). These variants were subsequently confirmed using Sanger sequencing (Inqaba Biotechnical Industries, Pretoria, South Africa). The specific oligonucleotide primer pair sequences are provided (Appendix IIC).

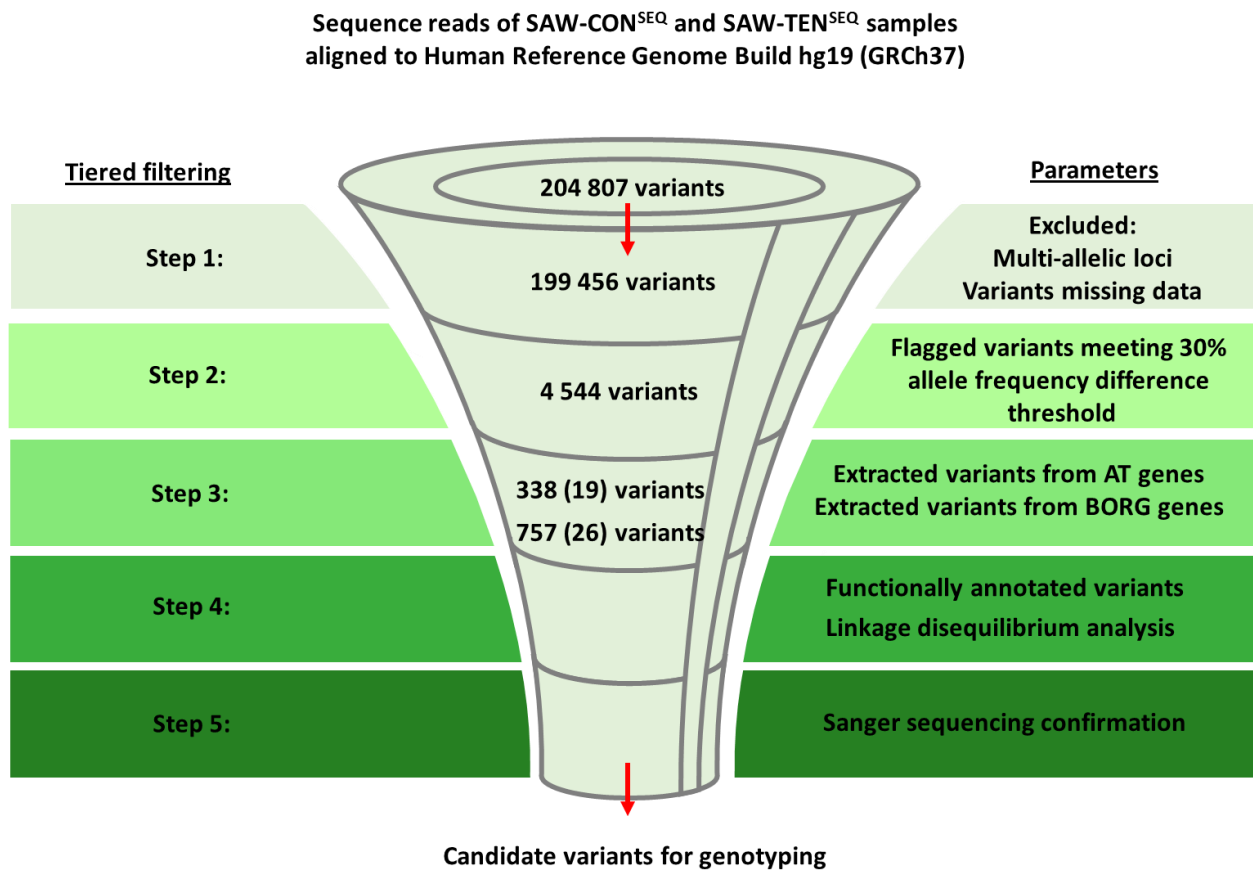


Figure 2.1: The customised, tiered filtering strategy used to prioritise loci for interrogation from the whole exome sequencing dataset

After alignment, a total of 204 807 variants were identified across the 20 sequenced exomes. Post-alignment processing resulted in the exclusion of all variants missing genotype data in two or more samples, in addition to the extraction of multi-allelic loci. The remaining 199 456 variants were mined for those variants which met the allele frequency difference threshold, specifically, variants mapping to the targeted genomic intervals including genes previously associated with the risk of Achilles tendinopathy and genes proposed as strong biological candidates for Achilles tendinopathy using the BORG semantic database. The number of variants meeting the allele frequency threshold in Achilles tendinopathy and BORG associated genes are indicated in parentheses. Candidate variants were functionally annotated, and their linkage patterns assessed. Tagged loci were identified. All variants prioritised for genotyping were confirmed using Sanger sequencing. Finally, participants in the larger injury cohorts were genotyped for the prioritised variants.

Table 2.12: Overview of variants prioritised for further interrogation in chapters 3 to 5 of this thesis. Candidates included new variants identified from the whole exome sequencing dataset using a tiered filtering strategy or were variants previously proposed as modifiers of risk

Gene	SNP ID	Alleles		MAF	Chromosomal Position	Variant location	Coding effect	AA substitution	Allele frequency difference (%)
		Ref	Alt						
<i>TNC</i>	rs944510*	C	T	T: 0.44	9: 115090743	Exon 2	Synonymous	-	0.35
	rs2992147*	C	T	C: 0.44	9: 115087035	Exon 3	Synonymous	-	0.35
	rs1061494	C	T	C: 0.44	9: 115074301	Exon 4	Non-synonymous	Q→R	0.35
	rs1138545	C	T	T: 0.18	9: 115073620	Exon 10	Non-synonymous	R→H	0.10
	rs2104772	T	A	A: 0.42	9: 115046506	Exon 17	Non-synonymous	I→L	0.45
	rs1061495	T	C	C: 0.26	9: 115042265	Exon 18	Synonymous	-	0.25
<i>COL27A1</i>	rs2567706	G	A	G: 0.27	9: 114167914	Exon 3	Non-synonymous	Q→R	0.30
	rs2241671	G	A	A: 0.39	9: 114168819		Non-synonymous	A→T	0.30
	rs2808770*	C	T	C: 0.27	9: 114169165		Non-synonymous	I→T	0.30
	rs2567705	T	A	T: 0.38	9: 114169386		Non-synonymous	I→F	0.30
	rs2808771*	T	C	T: 0.27	9: 114169457		Synonymous	-	0.30

Table 2.12 continued.

Gene	SNP ID	Alleles Ref Alt		MAF	Chromosomal Position	Variant location	Coding effect	AA substitution	Allele frequency difference (%)
MMP3	rs3025058*	6A	5A	-	11: 102845217-102845218	Promoter	Upstream variant	-	-
	rs679620	G	A	G: 0.41	11: 102842889	Exon 2	Non-synonymous	E→K	0.25
	rs602128*	G	A	G: 0.41	11: 102842734	Exon 2	Synonymous codon	D→D	0.30
	rs591058	T	C	C: 0.41	11: 102840607	Intron 4	Intronic	-	-
	rs520540*	G	A	G: 0.40	11: 102838694	Exon 8	Synonymous codon	A→A	0.30
	rs650108	G	A	A: 0.25	11: 102838056	Intron 8	Intronic	-	-
	rs639752*	A	C	A: 0.40	11: 102836608	Intron 9	Intronic	-	0.30
COL1A1	rs1107946	G	T	T: 0.16	17: 50203629	Promoter	Upstream variant	-	-
	rs1800012	G	T	T: 0.17	17: 50200388	Intron 1	Intronic	-	-

SNP ID: dbSNP accession number; Alleles: Ref (Reference) and Alt (Alternate); MAF: Minor allele frequency according to the 1000 Genomes Project data (phase-3) for the European Caucasian population; Chromosomal position: Chromosome number and location of variant according to the Human GRCh38.p10 genome release. AA substitution: Amino acid substitution at locus. Allele frequency difference: Allele frequency differences between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. (-) No information currently available for variant or no amino acid substitution. The *MMP3* rs3025058, rs591058 and rs650108 variants, in addition to the *COL1A1* rs1107946 and rs1800012 variants were not identified by WES due to their locations within intronic regions or the promoter of the genes. Consequently, no allele frequencies differences for these variants were determined between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. A: Alanine; D: Aspartic acid; E: Glutamic acid; F: Phenylalanine; H: Histidine; I: Isoleucine; K: Lysine; L: Leucine; Q: Glutamine; R: Arginine; T: Threonine; V: Valine. (*) Locus tagged by at least one other candidate variant prioritised for further interrogation.

2.3.5.1. TNC variants

A total of 25 variants mapping to the *TNC* gene were identified, five of which demonstrated allele frequency differences $\geq 30\%$. Of the five variants, three were observed to be in complete linkage disequilibrium (LD) with one another (Table 2.13). Specifically, the C allele of rs944510 (C>T) and the T allele of rs2992147 (T>C) corresponded to the T allele of rs1061494 (T>C). The measures of linkage disequilibrium (D') and the correlation coefficient (r^2) estimates were 1.000 for all pairwise comparisons ($P < 0.001$). From these three tagged loci, rs1061494 was selected as the proxy variant based on its potential functionality and was genotyped together with the other candidate variant identified by WES, rs2104772 (T>A), in the three South African cohorts and single British cohort. One other variant within the 3' UTR of the *TNC* gene was also identified by WES and met the allele frequency threshold (rs397761295, indel/T). However, this variant appears to be relatively uncharacterised, with no information regarding the allele frequencies available on either the NCBI or Ensembl resources. For this reason, this variant was not prioritised for interrogation in this current study.

Table 2.13: Identification of tagged loci from the list of *TNC* candidates generated by whole exome sequencing

Variants		Position 2			
		rs2992147	rs1061494	rs2104772	rs1061495
Position 1	rs944510	1.000	1.000	0.012	0.246
	rs2992147		1.000	0.012	0.246
	rs1061494			0.012	0.246
	rs2104772				0.017

Results based on the correlation coefficient estimates for the candidate loci in the European Caucasian population (<https://analysistools.nci.nih.gov/LDlink/>). Bold typeset indicates tagged variants.

Except for 64 samples, the *TNC* rs2104772 variant was previously genotyped in the SAW-CON^{AT} and SAW-TEN groups and the results described by Saunders et al. (2013) [198]. The other 64 samples were genotyped for this current study. Therefore, the previously published genotype data for the rs2104772 variant, together with newly genotyped samples were combined for comparative analyses (Chapter 3). The 47 SAW-RUP samples were not genotyped for the rs2104772 variant in the previous study and were therefore genotyped and the results described in this study. Although the rs1061495 variant (T>C) demonstrated only a 25% allele frequency difference between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups, this variant was also included for genotyping analysis due to its location in relation to variants previously associated with injury susceptibility and the newly identified WES candidates. Similarly, the rs1138545 (C>T) variant did not meet the allele frequency difference threshold; however, this variant was previously associated with the risk of RCT [214,215] and was therefore prioritised for interrogation together with the other *TNC* variants in chapter 3. Table 2.14 demonstrates the proportion of recruited participants successfully genotyped for the four prioritised *TNC* variants (rs1061494, rs1138545, rs2104772 and rs1061495) in the three South African cohorts and single British cohort.

Table 2.14: The proportion of recruited participants successfully genotyped for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts, in addition to the South African White (SAW)-and South African Coloured (SAC)-ACL rupture cohorts

Group	Participants (n)	TNC variants			
		rs1061494	rs1138545	rs2104772	rs1061495
SAW-Achilles tendon injury cohort [†]					
SAW-CON ^{AT}	165	87 (144)	70 (115)	96 (158)	89 (147)
SAW-TEN	123	80 (99)	65 (80)	92 (113)	83 (102)
SAW-RUP	47	87 (41)	68 (32)	72 (34)	89 (42)
Total	335	85 (284)	68 (227)	91 (305)	87 (291)
UK-Achilles tendon injury cohort					
UK-CON	130	63 (82)	93 (121)	78 (101)	82 (106)
UK-TEN	87	68 (59)	87 (76)	77 (67)	84 (73)
UK-RUP	25	76 (19)	96 (24)	80 (20)	96 (24)
Total	242	66 (160)	91 (221)	78 (188)	84 (203)
SAW-ACL rupture cohort					
SAW-CON ^{ACL}	232	96 (222)	88 (203)	98 (227)	98 (227)
SAW-ACL	234	99 (232)	90 (210)	99 (232)	99 (232)
Total	466	97 (454)	89 (413)	98 (459)	98 (459)
SAC-ACL rupture cohort					
SAC-CON	100	99 (99)	88 (88)	96 (96)	94 (94)
SAC-ACL	97	95 (92)	93 (90)	98 (95)	97 (94)
Total	197	97 (191)	90 (178)	97 (191)	94 (188)

Values are expressed as a percentage of the total number of participants recruited (n) with the number of participants successfully genotyped in parentheses. [†] A total of 124 SAW-CON^{AT} and 83 SAW-TEN samples were previously genotyped for the rs2104772 variant in the study by Saunders et al. (2013) [198]. SAW: South African White; UK: British; SAC: South African Coloured.

2.3.5.2. COL27A1 variants

A total of 66 variants were identified mapping to the *COL27A1* locus, five of which were observed to meet the allele frequency difference threshold. Of the five variants, the alleles of three were observed to be tightly correlated, allowing the inference of genotypes (Table 2.15). Specifically, the A allele of rs2567706 (A>G) correlated with the T allele of rs2808770 (T>C) and the C allele of rs2808771 (C>T). The D' measures were 1.000 and the r^2 estimates ranged between 0.980-1.000 for all pairwise comparisons ($P<0.001$). From the tagged loci, rs2567706 was selected as the proxy variant based on its potential functionality and was genotyped together with the other candidate variants identified by WES, rs2241671 (G>A) and rs2567705 (A>T), in the three South African cohorts and single British cohort.

Table 2.15: Identification of tagged loci from the list of *COL27A1* candidates generated by whole exome sequencing

Variants		Position 2			
		rs2241671	rs2808770	rs2567705	rs2808771
Position 1	rs2567706	0.539	0.980	0.599	0.980
	rs2241671		0.554	0.259	0.554
	rs2808770			0.615	1.000
	rs2567705				0.615
	rs2808771				

Results based on the correlation coefficient estimates for the candidate loci in the European Caucasian population (<https://analysistools.nci.nih.gov/LDlink/>). Bold typeset indicates tagged variants.

Table 2.16: The proportion of recruited participants successfully genotyped for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts, in addition to the South African (SAW)- and South African Coloured (SAC)-ACL rupture cohorts

Group	Participants (n)	COL27A1 variants		
		rs2567706	rs2241671	rs2567705
SAW-Achilles tendon injury cohort				
SAW-CON ^{AT}	165	84 (139)	78 (129)	73 (121)
SAW-TEN	123	76 (93)	59 (73)	65 (80)
SAW-RUP	47	85 (40)	68 (32)	40 (19)
Total	335	81 (272)	70 (234)	66 (220)
UK-Achilles tendon injury cohort				
UK-CON	130	98 (127)	98 (127)	93 (121)
UK-TEN	87	97 (84)	98 (85)	85 (74)
UK-RUP	25	96 (24)	100 (25)	100 (25)
Total	242	97 (235)	98 (237)	91 (220)
SAW-ACL rupture cohort				
SAW-CON ^{ACL}	232	97 (225)	94 (219)	94 (219)
SAW-ACL	234	99 (232)	96 (225)	95 (222)
Total	466	98 (457)	95 (444)	95 (441)
SAC-ACL rupture cohort				
SAC-CON	100	97 (97)	99 (99)	97 (97)
SAC-ACL	97	100 (97)	98 (95)	99 (96)
Total	197	98 (194)	98 (194)	97 (193)

Values are expressed as a percentage of the total number of participants recruited (n) with the number of participants successfully genotyped in parentheses. SAW: South African White; UK: British; SAC: South African Coloured.

2.3.5.3. MMP3 variants

Mining the WES data identified five variants mapping to the *MMP3* gene, three of which met the allele frequency difference threshold. The alleles at these loci were observed to be tightly correlated. Specifically, the A allele of rs602128 (A>G) correlated with the A allele of rs520540 (A>G) and the C allele of rs639752 (C>A). The D' measures were 1.000 and the r^2 estimates ranged between 0.960-1.000 for all pairwise comparisons ($P<0.001$) (Table 2.17). Although the rs602128 and rs520540 variants mapped within exonic regions, neither resulted in amino acid substitutions within the encoded protein sequence (silent variants). Further analysis revealed that these three variants were also in high LD with the *MMP3* rs679620 (A>G) variant, previously associated with the risk of AT [206]. The missense rs679620 variant, which results in a transversion substitution (Lys45Glu) within exon 2 of the gene was also identified by WES demonstrating a 25% allele frequency difference between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. The measure of LD (D') between rs679620 and the WES identified variants (rs602128, rs520540 and rs639752) was 1.000 and the correlation coefficient estimates ranged between 0.960-1.000 respectively (Table 2.17). Therefore, owing to the tight correlation of alleles, in addition to the previous association of rs679620 with the risk of chronic AT in the SAW-Achilles tendon injury cohort [206], the rs679620 variant was selected as the proxy variant and genotyped in the AUS-Achilles tendon injury and larger SAW-ACL rupture cohorts.

Table 2.17: Identification of tagged loci from the list of *MMP3* candidates generated by whole exome sequencing

Variants		Position 2		
		rs602128	rs520540	rs639752
Position 1	rs679620	1.000	0.960	0.960
	rs602128		0.960	0.960
	rs520540			1.000
	rs639752			

Results based on the correlation coefficient estimates for the candidate loci in the European Caucasian population (<https://analysistools.nci.nih.gov/LDlink/>). Bold typeset indicates tagged variants.

It must be noted that the SAW-ACL rupture cohort included in this thesis included 92 SAW-CON^{ACL} and 129 SAW-ACL rupture participants previously genotyped for the rs679620 (A>G) variant in the study by Posthumus et al. (2012) [203]. This previous study investigated a genomic interval spanning four *MMP* genes including *MMP3* at the 11q22 chromosomal locus. Since publication of this original research [203], an additional 140 SAW-CON^{ACL} and 105 SAW-ACL participants were recruited and genotyped for the current study. Consequently, the dataset presented in chapter 4 is larger and focuses specifically on the interrogation of the *MMP3* gene.

Two other *MMP3* variants, rs591058 (T>C) and rs650108 (G>A), although not identified by WES due to their intronic locations, were also included in this study for further interrogation based on their previous association with the risk of AT [206]. Furthermore, the functional *MMP3* rs3025058 (5A>6A) variant, which involves the insertion or deletion of an adenine (A) nucleotide 1 612 bp upstream from the transcription start site within the promoter of the gene, has been a locus of interest to many researchers investigating this gene and its association with disease predisposition. This variant was not identified by WES due to its location in the distal portion of the gene's promoter. However, because of its independent

association with other multifactorial diseases [204,246,304,305], in addition to its location within the regulatory region of the gene, this variant was included for analysis in the AUS-Achilles tendon injury and SAW-ACL rupture cohorts. Table 2.18 demonstrates the proportion of recruited participants successfully genotyped for the four *MMP3* variants (rs3025058, rs679620, rs591058 and rs650108) in the AUS-Achilles tendon injury and the SAW-ACL rupture cohorts.

Table 2.18: The proportion of recruited participants successfully genotyped for the *MMP3* rs3025058 (5A>6A) rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the Australian (AUS)- and South African White (SAW)-Achilles tendon injury cohorts, in addition to the South African White (SAW)-ACL rupture cohort

Group	Participants (n)	MMP3 variants			
		rs3025058	rs679620	rs5910581	rs650108
AUS-Achilles tendon injury cohort					
AUS-CON	210	91 (190)	91 (191)	91 (191)	76 (159)
AUS-TEN	85	91 (77)	91 (77)	91 (77)	79 (67)
Total	295	91 (267)	91 (268)	91 (268)	77 (226)
SAW-Achilles tendon injury cohort [†]					
SAW-CON ^{AT}	98		100 (98)	99 (97)	97 (95)
SAW-TEN	75	-	100 (75)	97 (73)	99 (74)
Total	173		100 (173)	98 (170)	98 (169)
SAW-ACL rupture cohort [‡]					
SAW-CON ^{ACL}	232	97 (226)	97 (225)	97 (225)	97 (224)
SAW-ACL	234	86 (200)	86 (200)	86 (200)	84 (197)
Total	466	91 (426)	91 (425)	91 (425)	90 (421)

Values are expressed as a percentage of the total number of participants recruited (n) with the number of participants successfully genotyped in parentheses. [†] All participants in the SAW-CON^{AT} and SAW-TEN groups were previously genotyped for the *MMP3* rs679620, rs591058 and rs650108 variants in the study by Raleigh et al. (2009) [206]. The rs3025058 variant was not genotyped in the SAW-Achilles tendon injury cohort in this previous study. [‡] Ninety-two SAW-CON^{ACL} and 129 SAW-ACL participants were previously genotyped for the rs679620 variant in the previous study by Posthumus et al. (2012) [203]. AUS: Australian; SAW: South African White.

2.3.5.4. COL1A1 variants

A total of 61 variants mapping to the *COL1A1* gene were identified by WES. However, none of these variants were observed to meet the allele frequency difference threshold. Therefore, variants implicated as risk modifiers of musculoskeletal injuries in previous studies were prioritised for further interrogation in the three South African cohorts and single British cohort in chapter 5. The two variants prioritised for investigation included the rs1107946 (G/T) proximal promoter variant and the rs1800012 (G/T) variant located within the first intron of the gene. The rs1800012 variant is located within the Sp1 transcription factor binding site and has previously been associated with the risk of acute musculoskeletal soft tissue injuries including ACL ruptures and shoulder dislocations [184–187]. Furthermore, an inferred haplotype constructed using genotype data from these two variants was proposed to modify the risk of ACL ruptures in a cohort of football players [183]. It must be noted that the rs1800012 variant was previously genotyped in SAW-CON^{AT}, SAW-TEN and SAW-RUP groups by Posthumus et al. (2009) [268]. Therefore, this data was included for combined and comparative analyses in chapter 5. In addition, the SAW-ACL rupture cohort analysed in chapter 5 included 130 SAW-CON^{ACL} and 117 SAW-ACL participants whom were previously genotyped and the results presented by Posthumus et al. (2009) [185]. Since publication of this manuscript, an additional 102 SAW-CON^{ACL} and 117 SAW-ACL participants were recruited and genotyped for the current study. Therefore, the results presented in chapter 5 pertain to a larger cohort, genotyped for rs1800012, in addition to the rs1107946 promoter variant. Table 2.19 demonstrates the proportion of recruited participants successfully genotyped for the *COL1A1* rs1107946 and rs1800012 variants in the three South African cohorts and single British cohort.

Table 2.19: The proportion of recruited participants successfully genotyped for the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts, in addition to the South African White (SAW)- and South African Coloured (SAC)-ACL rupture cohorts

Group	Participants (n)	COL1A1 variants	
		rs1107946	rs1800012
SAW-Achilles tendon injury cohort [†]			
SAW-CON ^{AT}	165	84 (138)	90 (150)
SAW-TEN	123	63 (85)	91 (112)
SAW-RUP	47	83 (39)	94 (44)
Total	335	78 (262)	91 (306)
UK-Achilles tendon injury cohort			
UK-CON	130	90 (117)	75 (98)
UK-TEN	87	97 (84)	84 (73)
UK-RUP	25	92 (23)	72 (18)
Total	242	93 (224)	78 (189)
SAW-ACL rupture cohort [‡]			
SAW-CON ^{ACL}	232	97 (225)	97 (225)
SAW-ACL	234	85 (199)	84 (197)
Total	446	95 (424)	95 (422)
SAC-ACL rupture cohort			
SAC-CON	100	91 (91)	88 (85)
SAC-ACL	97	88 (88)	86 (83)
Total	197	91 (179)	85 (168)

Values are expressed as a percentage of the total number of participants recruited (n) with the number of participants successfully genotyped in parentheses. [†] All SAW-CON^{AT}, SAW-TEN and SAW-RUP participants were previously genotyped for the rs1800012 variant in the study by Posthumus et al. (2009) [268]. [‡] The rs1800012 variant was also genotyped in 130 SAW-CON^{ACL} and 117 SAW-ACL participants in the study by Posthumus et al. (2009) [185]. SAW: South African White; UK: British; SAC: South African Coloured.

2.4. GENOTYPING AND ALLELIC DISCRIMINATION ASSAYS

2.4.1. Automated genotyping

Most variants investigated in this study were genotyped using inventoried Taqman® SNP Genotyping Assays (Applied Biosystems, Foster City, California, USA) including *TNC* rs1061494, rs1138545, rs2104772 and rs1061495; *COL27A1* rs2567706, rs2241671 and rs2567705; *MMP3* rs679620, rs591058 and rs650108 and *COL1A1* rs1107946 and rs1800012 respectively (Appendix IID). Automated amplification and calling of genotypes was conducted on the StepOnePlus™ or the QuantStudio™3 Real-Time PCR Systems (Applied Biosystems, Foster City, California, USA). In brief, 8 µl PCR reactions were formulated from 4 µl Taqman® Genotyping Master Mix (2X), 2.8 µl dH₂O, 1 µl of sampled DNA (approximately 100 ng/µl) and 0.2 µl Taqman® Assay Mix. The locus specific Taqman® Assay Mix consisted of a forward and reverse primer pair for the amplification of the polymorphic sequence, in addition to two labelled probes (VIC™ and FAM™ dyes) to detect and distinguish the reference allele (allele 1) from the alternate allele (allele 2). Real-time cycling conditions generally involved an initial hold step at 95°C for 10 minutes followed by 40 cycles of 92°C for 15 seconds and 60°C for 1 minute.

2.4.2. Restriction fragment length polymorphism analysis

The *MMP3* promoter variant, rs3025058 (5A>6A), was genotyped in all samples using standard polymerase chain reaction (PCR) and restriction fragment length polymorphism (RFLP) analysis (Appendix IID). The PCR reactions were performed in a final volume of 40 µl containing approximately 100 ng of DNA template, 1x reaction buffer, 2.5 mM of each dNTP (dATP, dCTP, dGTP, dTTP), 20 pmol of each primer (forward and reverse) and 0.5 units of Taq DNA polymerase (New England Biolabs Inc., Ipswich, Massachusetts, USA). Thermo-Pol® reaction buffer (New England Biolabs Inc., Ipswich, Massachusetts, USA) consisted of 10 mM KCl, 10 mM (NH₄)SO₄, 20 mM Tris-HCl, 2 mM MgSO₄ and 0.1% Triton X-100 at pH 8.8. Polymerase chain reaction amplification of fragments was conducted on a XP thermal block

cycler (Bioer Technology Co., Middlesex, UK). A 5-minute denaturation step at 94°C initiated the PCR amplification process, followed by 35 cycles progressing through 30 seconds at i) 94 °C, (ii) 65°C and (iii) 72 °C respectively. The final step involved a 5-minute hold at 72°C to ensure elongation of all PCR products. The specific oligonucleotide primer-pair sequences, reaction conditions and amplicon sizes for this PCR-RFLP are presented in Appendix IID viii. In a final reaction volume of 10 µl, 4 µl of PCR product was subjected to digestion using 5 units of the CutSmart® Tth111I restriction enzyme (New England Biolabs Inc., Ipswich, Massachusetts, USA), while incubated with 1X CutSmart® reaction buffer at 65°C for 60 minutes. The resultant products were treated with SYBR® Gold nucleic acid gel stain (Invitrogen Molecular Probes, Oregon, USA) and together with a 100 bp DNA size standard (Promega Corporation, Madison, Wisconsin, USA), separated on a 3% agarose gel using electrophoresis (120 V for 40 minutes). Gel images were captured under UV light using a photodocumentation system (Uvitec Limited, Cambridge, UK).

2.5. CASE-CONTROL GENETIC ASSOCIATION STUDIES

All case-control genetic association studies presented in this thesis were performed in accordance with the STREGA statement for conducting and reporting the results of genetic association studies [295]. For quality control purposes, negative (DNA-free) and positive controls (samples of known genotype) were included in each 96-well PCR plate during the genotyping process. Two investigators confirmed genotype calls with all laboratory work conducted at the Division of Exercise Science and Sports Medicine at the University of Cape Town.

2.6. STATISTICAL ANALYSES

The QUANTO v1.2.4 software (<http://biostats.usc.edu/software>) was used to determine the statistical power of all analyses based on the respective sample sizes. Assuming “risk” allele frequencies between 0.2 and 0.8, a sample size of 100 cases and 200 controls would be sufficient to detect an allelic odds ratio of ≥ 2.0 at a power of 80% and a significance level of 5%. All statistical analyses were conducted in the R programming environment using specific R packages [306]. Descriptive measures of quantitative variables were expressed as means and standard deviations, whereas categorical characteristics were presented as frequencies. One-way analysis of variance (ANOVA) was used to determine any characteristic differences amongst the case and control groups and the possibility of genotype associations with age, sex, body mass, height and BMI. The SNPAssoc [307] and Genetics [308] packages were used to determine genotype and allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities. Hardy-Weinberg equilibrium P-values are provided as exact P-values. Logistic regression models testing the association of gene variants with the risk of musculoskeletal soft tissue injuries were adjusted for potential non-genetic confounders. Confounding factors included participant characteristics which were observed to differ significantly between the diagnostic groups of the respective cohorts. These parameters were equally weighted in the logistic regression models and entered by means of a forward method. If any genotype categories demonstrated frequencies $< 10\%$, a Fisher’s exact test was conducted. The haplo.stats package [309] was used to test the association between inferred haplotypes and the risk of injury. The strength of the association between the inferred haplotypes and the risk of injury is reflected by the score statistic (Haplo.score). Negative values infer decreased risk, whereas positive scores infer an increased risk for injury. Statistical significance was accepted when $P < 0.05$. Adjustments for multiple testing were not performed on the genotype data as it has been suggested that certain statistical adjustments, such as Bonferroni corrections, may overcorrect for an inflated false positive rate and as a result, significantly reduce the power of genetic association studies, in which linkage disequilibrium exists between loci [310]. Furthermore, this study followed a hybrid approach in which several variants located within genes previously associated with injury risk were genotyped together with several new candidates identified by WES. Therefore, these factors

provide sufficient basis for an *a priori* hypothesis that these variants may be associated with musculoskeletal soft tissue injury risk in this study. However, it is acknowledged that a truly hypothesis-free approach to explore WES data in higher powered cohorts and using traditional measures of association will require correction for multiple testing.

2.7. FUNCTIONAL PREDICTIONS AND ANALYSIS

Several bioinformatic *in silico* tools were used to assess functional attributes of the variants prioritised for further interrogation from the WES dataset. Specifically, the 3DSNP database was used to annotate the regulatory function of the investigated variants by exploring their interactions with genes, other variants and regulatory elements in a three dimensional (3D) environment mediated by chromatin looping [311]. This resource quantitatively scored each variant based on functional categories including 3D interacting genes (chromatin looping), alterations to sequence motifs and transcription factor binding sites, in addition to the interaction with regulatory elements including gene promoters and enhancers in human cell lines. The greater the prediction score, the greater the annotation record (biological evidence) for a potential functional impact.

Secondly, the Sfold software (<http://sfold.wadsworth.org>) was used to predict changes in the pre-mRNA secondary structures induced by the presence of variants compared to that of the ancestral alleles. By algorithm, the Sfold program generates statistically representative samples from the ensemble of secondary structures in a population to their Boltzmann weights. The centroid structure, presented for each investigated variant, has the minimum base-pair distance to all structures within a given population and is believed to most accurately represent the folding tendency of the set, whether within a free energy cluster (cluster centroid) or alternatively the entire population of structures (ensemble centroid) [312]. These analyses examined apparent changes in the folding of the pre-mRNA structures including the formation or loss of features such as hairpin loops and/or single strandedness, all of which may potentially impact the accessibility of these regions to regulators of gene function including transcription factors and regulators of translation and/or post-translational

processes. Pre-mRNA in a state of single-strandedness or protruding as a hairpin loop, is presumably more accessible to gene regulators as opposed to double stranded pre-mRNA structures or structures subjected to complex folding.

Other bioinformatic tools including the (i) Sorting Intolerant from Tolerant (SIFT) database [313], (ii) Polymorphism phenotyping database (PolyPhen-2) [314], (iii) Combined Annotation Dependent Depletion (CADD) framework [315] and the Functional Analysis through Hidden Markov Models (Fathmm_MKL) resource [316] were used to predict the consequence (deleteriousness) of the single nucleotide variants on protein function. The Genotype-Tissue Expression (GTEx) consortium (version 7) was used to investigate the relationship between the genetic variants and gene expression in human tissues [317]. Furthermore, putative tagged loci as suggested by the allele frequencies generated by WES were confirmed using the 1000 Genomes Project data (phase-3) for the European Caucasian population and the LDlink online resource [303].

CHAPTER 3: PROFILING THE *TNC* AND *COL27A1* GENES IN MUSCULOSKELETAL SOFT TISSUE INJURIES

This chapter has been submitted for publication in a condensed format as:

Gibbon A, Saunders CJ, Collins M, Gamielien J, September AV. Defining the genetic risk profile of musculoskeletal soft tissue injuries: A whole exome sequencing approach (In press).

3.1. STUDY AIM AND OBJECTIVES

As previously discussed, several genomic loci have been suggested to modulate musculoskeletal soft tissue injury susceptibility [18,19]. Of these variants, several are known to map to the neighbouring *TNC* and *COL27A1* genes on the long arm of chromosome 9 [198,199,214,215]. Based on this previous research (Section 1.5.1), the primary aim of the current study was to further define the genetic risk profile of musculoskeletal soft tissue injuries mapping to the *TNC* (chr9q33.1) and *COL27A1* (chr9q32) genes using an integrated omics approach.

Objectives specific to this study were:

- (i) To prioritise *TNC* and *COL27A1* candidate variants for further interrogation from the WES dataset using a customised, tiered filtering strategy
- (ii) To test the association between the *TNC* candidate variants prioritised from the WES dataset: rs1061494 (T>C), rs2104772 (T>A) and rs1061495 (T>C) and the risk of Achilles tendon and ACL injuries in cohorts of self-identified White ancestry (SAW-, UK- and combined (SAW and UK)-Achilles tendon injury cohorts and SAW-ACL rupture cohort) and a single cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)
- (iii) To test the association between the *COL27A1* candidate variants prioritised from the WES dataset: rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) and the risk of

- Achilles tendon and ACL injuries in cohorts of self-identified White ancestry (SAW-, UK and combined (SAW and UK)-Achilles tendon injury cohorts and SAW-ACL rupture cohort) and a single cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)
- (iv) To test the association between the rs1138545 (C>T) variant, previously associated with the risk of RCT, and the risk of Achilles tendon and ACL injuries in cohorts of self-identified White ancestry (SAW-, UK and combined (SAW and UK)-Achilles tendon injury cohorts and SAW-ACL rupture cohort) and a single cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)
 - (v) To test the association between the inferred haplotypes constructed from the prioritised *TNC* candidate variants: rs1061494, rs1138545, rs2104772 and rs1061495 and the risk of Achilles tendon and ACL injuries in cohorts of self-identified White ancestry (SAW-, UK and combined (SAW and UK)-Achilles tendon injuries cohorts and SAW-ACL rupture cohort) and a single cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)
 - (vi) To test the association between the inferred haplotypes constructed from the prioritised *COL27A1* candidate variants: rs2567706, rs2241671 and rs2567705 and the risk of Achilles tendon and ACL injuries in cohorts of self-identified White ancestry (SAW-, UK and combined (SAW and UK)-Achilles tendon injury cohorts and SAW-ACL rupture cohort) and a single cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)

By addressing these objectives, this study aimed to refine the mapping of risk-associated variants for musculoskeletal soft tissue injuries in the *TNC* and *COL27A1* genes. Figure 3.1 and Figure 3.2. provide an overview of the *TNC* and *COL27A1* variants prioritised for further interrogation in this chapter.

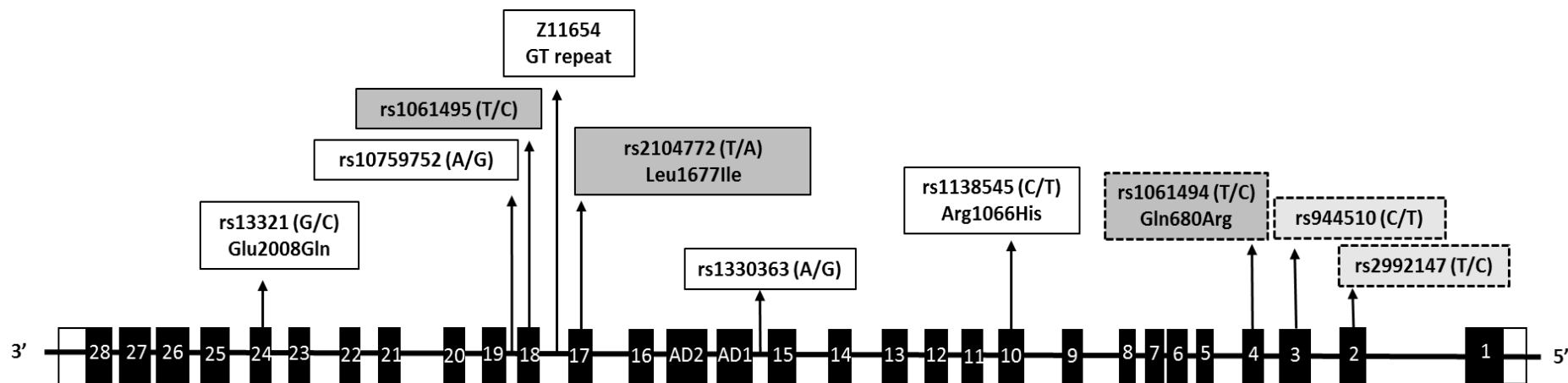


Figure 3.1: Map of *TNC* gene displaying candidate variants prioritised for interrogation from the whole exome sequencing dataset in relation to variants previously associated with the risk of musculoskeletal soft tissue injuries

Variants in white blocks in the *TNC* gene were previously associated with the risk of musculoskeletal soft tissue injuries but not identified by WES in the current study [198,199,214,215]. The exons 1-28, AD1 and AD2 (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *TNC* are presented. Variants in grey blocks (light and dark) were identified by WES and met the allele frequency difference threshold of 30%. Variants in dark grey blocks were prioritised for interrogation in the current study. Blocks with dashed outlines demonstrate tagged loci. Arg: Arginine; Gln: Glutamine; Glu: Glutamic acid; Hist: Histidine; Ile: Isoleucine; Leu: Leucine. Diagram not drawn to scale.

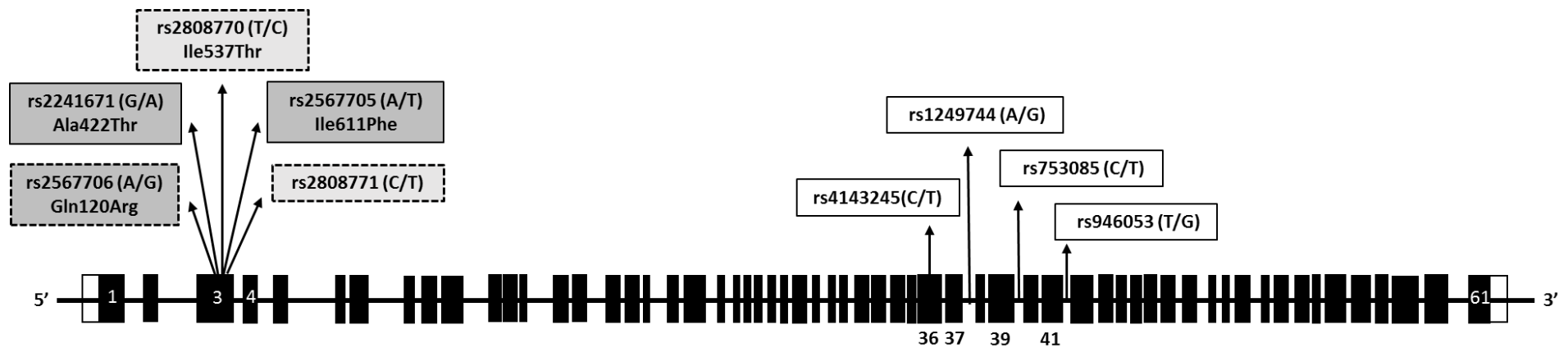


Figure 3.2: Map of *COL27A1* gene displaying candidate variants prioritised for interrogation from the whole exome sequencing dataset in relation to variants previously investigated for their potential association with the risk of musculoskeletal soft tissue injuries

Variants in white blocks within the *COL27A1* gene have previously been tested for their potential association with the risk of Achilles tendinopathy [198]. The exons 1-61 (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *COL27A1* are presented. Variants in grey blocks (light and dark) in the *COL27A1* gene were identified by WES and met the allele frequency difference threshold of 30%. Variants in dark grey blocks were prioritised for interrogation in the current study. Blocks with dashed outlines demonstrate tagged loci. Ala: Alanine; Arg: Arginine; Gln: Glutamine; Ile: Isoleucine; Phe: Phenylalanine; Thr: Threonine; Val: Valine. Diagram not drawn to scale.

3.2. RESULTS

3.2.1. General analysis of the whole exome sequencing dataset

The average depth of coverage across the 20 sequenced exomes was 49X, with 90% of target bases sequenced at a coverage greater than 15X. Table 3.1 categorises the variants captured using the Illumina HiSeq 2000/2500 sequencing system. A total of 204 807 variants meeting the quality score threshold were called across the 20 sequenced exomes. After the removal of multi-allelic variants (n=3 435) and those with missing genotype data (n=1 916), 199 456 variants were included in the analyses presented in this thesis.

The number of variants identified on each chromosome correlated with the size of the structure. Chromosome 1 demonstrated the highest variant density (19 921 variants), compared to the 2 874 variants mapping to chromosome 21; the smallest autosome (Figure 3.3). Of the 199 456 variants identified, 4 544 loci were observed to meet the allele frequency difference threshold of 30%, mapping to 2 127 genes. In total, 802 variants mapped to the genes previously implicated in the aetiology of AT and other musculoskeletal soft tissue injuries (Figure 3.3) [18,19]. Interestingly, the *TNC* rs2104772 variant demonstrated the highest allele frequency difference (45%) between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. In addition, 757 variants were highlighted to map to the 35 genes previously proposed as plausible candidates for AT risk using an *in silico* analysis of the BORG semantic database (Figure 3.4) [302]. Of these variants, 26 met the allele frequency difference threshold, with the majority mapping to the heparan sulfate proteoglycan 2 gene (*HSPG2*) on chromosome 1. This gene demonstrated the greatest variant density (n=74), with a total of 14 variants meeting the allele frequency difference threshold (Figure 3.4). Furthermore, a total of 36 variants meeting the allele frequency difference threshold were observed to map to 27 known miRBase miRNA genes.

Table 3.1: Overview of variants captured across the twenty sequenced exomes (SAW-CON^{SEQ} and SAW-TEN^{SEQ}) using the Illumina HiSeq 2000/2500 system

Type of Variant	Number of variants	
	All (n=204 807)	Allele frequency difference threshold (n=4 544)
Intronic (%)	28 (57 187)	29 (1 313)
Intergenic (%)	4 (8 203)	6 (208)
Exonic (%)	29 (60 282)	27 (1 223)
Untranslated regions (%)	30 (60 930)	31 (1 427)
Up/downstream (%)	3 (6 634)	4 (172)
Splicing (%)	1 (2 239)	<1 (3)
ncRNA (%)	5 (9 332)	4 (198)
Exonic Variants	(n=60 282)	(n=1 223)
Frameshift deletions (%)	<1 (354)	<1 (3)
Frameshift insertions (%)	<1 (214)	-
Non-frameshift deletions (%)	<1 (359)	<1 (5)
Non-frameshift insertions (%)	<1 (252)	<1 (3)
Non-frameshift substitutions (%)	<1 (3)	-
Non-synonymous SNV (%)	42 (25 033)	30 (366)
Synonymous SNV (%)	54 (32 671)	68 (828)
Stop-gain (%)	<1 (312)	<1 (2)
Stop-loss (%)	<1 (30)	-
Unknown (%)	2 (1 054)	1 (16)

For each variant category, a percentage of the total variant count (bold typeset) is provided. Count is in parentheses. Allele frequency difference threshold: Variants demonstrating an allele frequency differences $\geq 30\%$ between SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. ncRNA: non-coding RNA variant; SNV: Single nucleotide variant; Synonymous: Variant which does not result in an amino acid substitution (silent); Non-synonymous: Variant which results in an amino acid substitution within protein sequence; Frameshift: Addition or deletion of base pairs resulting in an alteration to the open reading frame of gene; Stop-gain: Variant results in the change from a codon to a stop-codon (terminates reading frame); Stop-loss: Variant which results in a stop-codon being substituted for a codon; (-) No variants in this category identified by whole exome sequencing.

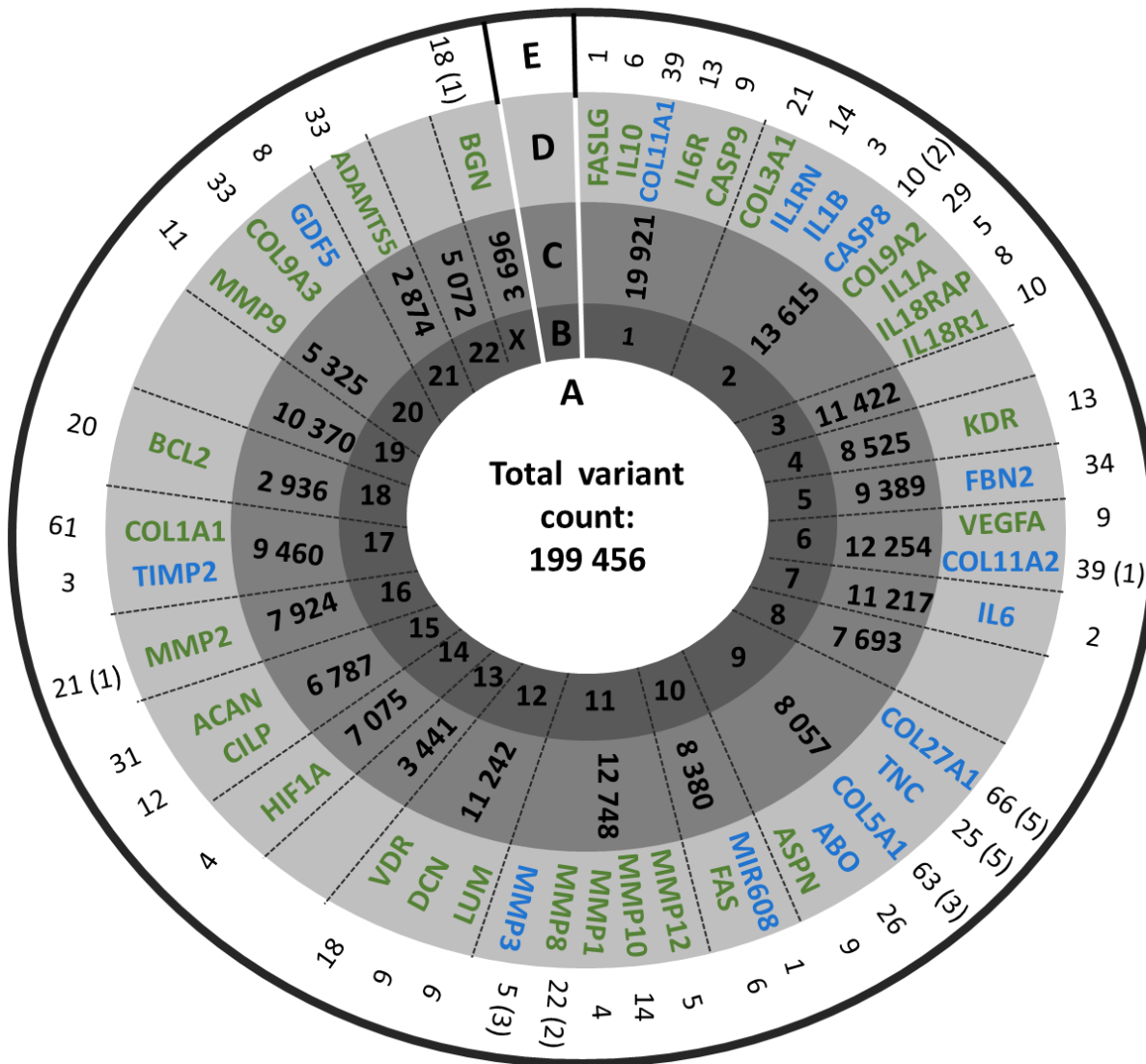


Figure 3.3: Distribution of variants identified by whole exome sequencing within genes previously associated with the risk of Achilles tendinopathy and other musculoskeletal soft tissue injuries

Wheel displays genes previously implicated in the aetiology of Achilles tendinopathy (blue) and other musculoskeletal soft tissue injuries (green). (A) Total number of variants identified across the 20 sequenced exomes. (B) The autosomes numbered 1-22 and the X gonosome. The Y gonosome is not presented due to its small size relative to the other chromosomes. Thirty-two variants were called on the Y chromosome. (C) Total number of variants identified mapping to each chromosome using WES. (D) Genes previously associated with the risk of injury mapping to the respective chromosomes. (E): Total number of variants identified mapping to these genes with number of variants meeting the allele frequency difference threshold in parentheses.

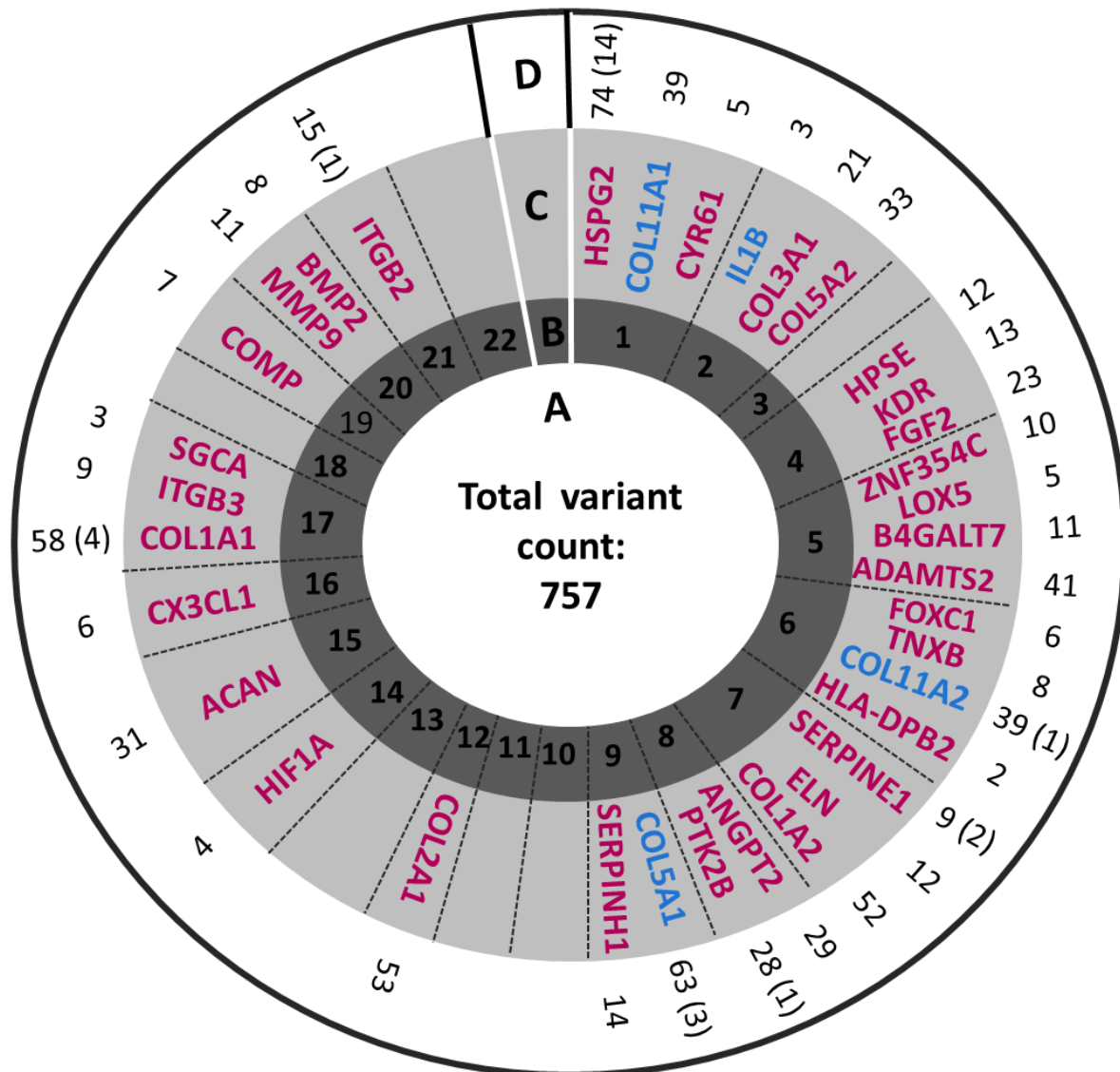


Figure 3.4: Distribution of variants identified by whole exome sequencing within genes proposed as plausible candidates for Achilles tendinopathy using the BioOntological Relationship Graph (BORG) semantic database

Wheel displays genes previously identified as plausible candidates for Achilles tendinopathy using the BORG semantic database (pink) [302]. Gene names in blue indicate those that have been also been associated with the risk of Achilles tendinopathy. (A) Total number of variants identified mapping to the 35 BORG gene candidates. (B) The autosomes numbered 1-22. (C) The BORG identified candidate genes. (D) Total number of variants identified mapping to BORG candidates with number of variants meeting the allele frequency difference threshold in parentheses.

Section 3.2.1 of this chapter presented the findings from the application of a customised analysis pipeline to the WES generated dataset, a process described in detail in chapter 2 (Section 2.3). The results presented above have characterised the type of variants identified by WES. In addition, these results also demonstrate the distribution of these variants across the (i) human chromosomes and (ii) genes previously implicated in the predisposition of AT, other musculoskeletal soft tissue injuries and those proposed as strong candidates for AT using the *in silico* BORG semantic database [302]. However, the interrogation of many of these gene variants extended beyond the scope of this thesis. Instead, and as previously discussed, this thesis focused specifically on refining the genetic risk signature of chronic AT and ACL ruptures mapping to four prioritised genomic intervals: *TNC* (chr9q33.1), *COL27A1* (chr9q32), *MMP3* (chr11q22) and *COL1A1* (chr17q21.33) respectively. Therefore, sections 3.2.2 and 3.2.3 of this current chapter, in addition to the subsequent chapters (Chapters 4 and 5) focused specifically on the results obtained from genotyping the variants prioritised for further interrogation within these targeted genomic regions, in several previously recruited AT and ACL rupture cohorts. These variants had either been identified as plausible candidates by WES and prioritised using a tiered filtering strategy or were previously proposed as modifiers of injury risk. A detailed description of the prioritised variants in each target region was provided in chapter 2 (Section 2.3.5).

3.2.2. Achilles tendon injuries

As outlined in chapter 2, the participants within the SAW-and UK-Achilles tendon injury cohorts were genotyped and the allele frequency distributions determined for the *TNC* variants prioritised from the WES dataset: rs1061494 (T>C), rs2104772 (T>A) and rs1061495 (T>C), in addition to the rs1138545 (C>T) variant, previously proposed to modulate the risk and chronicity of RCT [214,215]. Similarly, the results of genotyping the *COL27A1* variants prioritised from the WES dataset: rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) in the Achilles tendon injury cohorts are also presented in the current section.

3.2.2.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Section 2.1.1.1A and 2.1.3.1). The *TNC* rs1061494 (T>C) and *COL27A1* rs2241671 (G>A) variants were significantly associated with sex in the UK-Achilles tendon injury cohort ($P=0.006$ and $P=0.003$ respectively) (Table 3.2). The *TNC* rs1061494 CC genotype was significantly overrepresented in male participants (27%, $n=28$) compared to female participants (13%, $n=7$, $P=0.029$), whereas the CT genotype was significantly overrepresented in the female participants (61%, $n=34$) compared to male participants (36%, $n=37$, $P=0.002$). The rs2241671 GG genotype was significantly overrepresented in the female participants (43%, $n=38$) compared to male participants (26%, $n=39$, $P=0.007$), while a significantly greater proportion of male participants were heterozygotes at this locus (M: 56%, $n=83$, F: 34%, $n=30$, $P=0.001$). Additionally, rs1061494 was associated with body mass ($P=0.006$) and rs2241671 was associated with height ($P=0.001$) in the UK-Achilles tendon injury cohort. Specifically, participants with the *TNC* rs1061494 CC genotype were significantly heavier (87.3 ± 16.3 kg) than those with the CT + TT genotype (79.3 ± 18.3 kg, $P=0.025$) after adjusting for sex and age, while individuals with the rs2241671 GG genotype were significantly shorter (171.2 ± 8.9 cm) than participants with the GA + AA genotype (176.0 ± 9.9 cm, $P=0.002$) after adjusting for sex. No other genotype associations with participant characteristics were observed.

Table 3.2: Genotype associations with the *TNC* rs1061494 (T>C), rs11384545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in addition to the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants and the physiological characteristics of participants in the South African White (SAW)- and British (UK)-Achilles tendon injury cohorts

	<i>TNC</i> variants							
	rs1061494		rs11384545		rs2104772		rs1061495	
	SAW	UK	SAW	UK	SAW	UK	SAW	UK
Age	0.710	0.946	0.542	0.782	0.465	0.950	0.217	0.765
Sex	0.564	0.006	0.237	0.792	0.238	0.130	0.489	0.969
Body mass	0.429	0.006	0.255	0.803	0.581	0.366	0.314	0.353
Height	0.922	0.057	0.348	0.896	0.450	0.696	0.622	0.302
BMI	0.266	0.581	0.159	0.707	0.243	0.943	0.468	0.618

	<i>COL27A1</i> variants					
	rs2567706		rs2241671		rs2567705	
	SAW	UK	SAW	UK	SAW	UK
Age	0.818	0.554	0.863	0.807	0.440	0.685
Sex	0.917	0.129	0.814	0.003	0.853	0.177
Body mass	0.732	0.138	0.499	0.093	0.334	0.660
Height	0.542	0.793	0.337	0.001	0.076	0.102
BMI	0.562	0.736	0.243	0.758	0.457	0.087

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex. P-values in bold typeset indicate significance ($P < 0.05$).

3.2.2.2. Genotype and allele frequencies

Except for the *TNC* rs2104772 (T>A) variant, the other investigated *TNC* and *COL27A1* variants demonstrated similar genotype and allele frequency distributions between the SAW- and UK-Achilles tendon injury cohorts. Consequently, the genotype datasets from these two injury cohorts were combined to increase the statistical power of the study. The genotype and allele frequency distributions are therefore presented for the independent injury cohorts as well as the combined-Achilles tendon injury cohort except the data for the rs2104772 variant, which was only analysed in the independent cohorts. It must also be noted that for both cohorts, the RUP groups were too small for independent analyses. Therefore, testing the association of the *TNC* and *COL27A1* variants with the risk of ATR was only conducted in the combined-Achilles tendon injury cohort after adjusting for age, sex and country of birth.

No significant differences in the frequency distribution of the investigated *TNC* variants were noted between the CON and TEN groups in the UK- and combined-Achilles tendon injury cohorts (Table 3.3). This was also the case for the SAW-Achilles tendon injury cohort except for the rs1061494 variant, which was associated with AT risk ($P=0.049$) (Table 3.3). The CC genotype of rs1061494 was significantly underrepresented in the SAW-CON^{AT} group (13%, $n=18$) compared to the SAW-TEN group (24%, $n=20$) after covarying for sex and age ($P=0.018$, OR: 2.5, 95% CI: 1.2-5.1). Deviations from HWE were noted in the Tables 3.3. The CON and TEN groups demonstrated similar genotype and allele frequency distributions for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the SAW- and combined-Achilles tendon injury cohorts (Table 3.4). Similar results were observed for the UK-Achilles tendon injury cohort regarding the rs2567706 and rs2567705 variants; however, the GG genotype of rs2241671 (G>A) was significantly underrepresented in the UK-CON group (25%, $n=30$), compared to UK-TEN group (39%, $n=33$, $P=0.037$, OR: 1.9, 95% CI: 1.0-3.6) (Table 3.6).

Table 3.3: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the South African White (SAW)-, British (UK)- and combined-Achilles tendinopathy cohorts

	SAW		UK		P-values			
	CON ^{AT}	TEN	CON	TEN	Countries	Diagnostic Groups		
rs1061494	137	84	77	59		SAW	UK	SAW+UK
TT	29 (39)	27 (23)	33 (25)	37 (22)	0.143	0.049	0.745	0.345
TC	58 (80)	49 (41)	44 (34)	44 (26)				
CC	13 (18)	24 (20)	23 (18)	19 (11)				
C allele	42 (116)	48 (81)	45 (70)	41 (48)	0.814	0.213	0.484	0.697
HWE	0.025	0.546	0.371	0.589				
rs1138545	109	72	115	76				
CC	75 (82)	81 (58)	76 (87)	70 (53)	0.710	0.470	0.405	1.000
CT	22 (24)	18 (13)	23 (26)	28 (21)				
TT ^a	3 (3)	1 (1)	2 (2)	3 (2)				
T allele	14 (30)	10 (15)	13 (30)	16 (25)	0.460	0.432	0.376	1.000
HWE	0.402	1.000	1.000	1.000				
rs2104772[†]	152	94	94	67				
TT	25 (38)	18 (17)	44 (41)	37 (25)	<0.001	0.246	0.514	-
TA	43 (65)	42 (39)	39 (37)	49 (33)				
AA	32 (49)	40 (38)	17 (16)	13 (9)				
A	54 (163)	61 (115)	37 (69)	38 (51)	<0.001	0.137	0.823	-
HWE	0.149	0.550	0.138	0.800				
rs1061495	140	86	99	73				
TT	55 (77)	56 (48)	62 (61)	69 (50)	0.206	1.000	0.421	0.467
TC	37 (52)	35 (30)	29 (29)	25 (18)				
CC ^b	8 (11)	9 (8)	9 (9)	7 (5)				
C allele	26 (74)	27 (46)	24 (47)	19 (28)	0.158	1.000	0.400	0.579
HWE	0.673	0.202	0.065	0.119				

Genotypes are expressed as percentages with participant counts (n) in parentheses. P-values are for the difference between countries (SAW vs. UK) and diagnostic groups (CON vs. TEN) adjusted for each other, sex and age. P-values in bold typeset indicate significance (P<0.05). [†]A total of 124 SAW-CON^{AT} and 83 SAW-TEN samples were previously genotyped for the rs2104772 variant in the study by Saunders et al. (2013) [198]. (-): Unable to pool genotype data due to significant genotype and/or allele frequency differences between injury cohorts.

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

Table 3.4: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the South African White (SAW)-, British (UK)- and combined-Achilles tendinopathy cohorts

	SAW		UK		Countries	P-values		
	CON ^{AT}	TEN	CON	TEN		Diagnostic Groups		
rs2567706	133	76	120	84		SAW	UK	SAW+UK
AA	53 (71)	51 (40)	53 (63)	55 (46)	0.369	0.778	0.759	0.566
AG	38 (50)	41 (32)	34 (41)	35 (29)				
GG ^a	9 (12)	8 (6)	13 (16)	11 (9)				
G allele	28 (74)	29 (44)	30 (73)	28 (47)	0.663	1.000	0.685	0.823
HWE	0.532	0.625	0.057	0.184				
rs2241671	123	62	120	85				
GG	30 (37)	26 (16)	25 (30)	39 (33)	0.844	0.964	0.114	0.395
GA	47 (58)	48 (30)	52 (62)	41 (35)			(0.037)	
AA	23 (28)	26 (16)	23 (28)	20 (17)				
A allele	46 (114)	50 (62)	49 (118)	41 (69)	0.624	0.521	0.096	0.427
HWE	0.725	0.815	0.860	0.182				
rs2567705	116	71	114	74				
AA	42 (49)	34 (24)	35 (40)	37 (27)	0.632	0.626	0.778	0.716
AT	41 (48)	49 (35)	44 (50)	46 (34)				
TT	16 (19)	17 (12)	21 (24)	18 (13)				
T allele	37 (86)	42 (59)	43 (98)	41 (60)	0.391	0.460	0.683	0.825
HWE	0.445	0.654	0.579	0.809				

Genotypes are expressed as percentages with participant counts (n) in parentheses. P-values are for the difference between countries (SAW vs. UK) and between diagnostic groups (CON vs. TEN) adjusted for each other, sex and age. P-values in bold typeset indicate significance ($P < 0.05$). P-values in parentheses are generated from conducting post-hoc analyses based on an *a priori* hypothesis that the genotype is associated with the risk of injury.

^a SAW-CON^{AT} vs. SAW-TEN rs2567706 AA genotype vs. AG + GG genotype

No significant differences in the distribution of genotype or allele frequencies were noted between the combined-CON and combined-RUP groups for any of the investigated *TNC* variants (Table 3.5). Similarly, the genotype and allele frequencies for the rs2567706 and rs2567705 variants did not differ significantly between the combined-CON and combined-RUP groups (Table 3.6). However, following an *a priori* hypothesis and after adjusting for country of birth, sex and age, the rs2241671 GG genotype was significantly underrepresented in the combined-CON group (28%, n=67) compared to the combined-RUP group (42%, n=24, $P=0.047$, OR: 2.4, 95% CI: 1.0-6.2). The frequency distribution of the G allele was also significantly different between the diagnostic groups (combined-CON: 52%, n=122, combined-RUP: 64%, n=34, $P=0.030$).

Table 3.5: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) and Achilles tendon rupture (RUP) groups in the combined-Achilles tendon injury cohort

	SAW		UK		P-values	
	CON ^{AT}	RUP	CON	RUP	Countries	Diagnostic groups
rs1061494	137	41	77	19		SAW+UK
TT	29 (39)	27 (11)	33 (25)	26 (5)	0.050	0.840
TC	58 (80)	59 (24)	44 (34)	47 (9)		
CC	13 (18)	15 (6)	23 (18)	26 (5)		
C allele	42 (116)	44 (36)	45 (70)	50 (19)	0.399	0.664
HWE	0.025	0.344	0.371	1.000		
rs1138545	109	32	115	24		
CC	75 (82)	84 (27)	76 (87)	63 (15)	0.503	1.000
CT	22 (24)	16 (5)	23 (26)	38 (9)		
TT ^a	3 (3)	0 (0)	2 (2)	0 (0)		
T allele	14 (30)	8 (5)	13 (30)	19 (9)	0.535	0.878
HWE	0.402	1.000	1.000	0.551		
rs2104772	152	34	94	21		
TT	25 (38)	32 (11)	44 (41)	33 (7)	<0.001	-
TA	43 (65)	53 (18)	39 (37)	52 (11)		
AA	32 (49)	15 (5)	17 (16)	14 (3)		
A allele	54 (163)	41 (28)	37 (69)	40 (17)	0.002	-
HWE	0.149	0.731	0.138	1.000		
rs1061495	140	42	99	24		
TT	55 (77)	55 (23)	62 (61)	58 (14)	0.239	0.888
TC	37 (52)	38 (16)	29 (29)	25 (6)		
CC ^b	8 (11)	7 (3)	9 (9)	17 (4)		
C allele	26 (74)	26 (22)	24 (47)	29 (22)	0.594	0.672
HWE	0.673	1.000	0.066	0.060		

Genotypes are expressed as percentages with participant counts (n) in parentheses. P-values are for the difference between countries (SAW vs. UK) and diagnostic groups (CON vs. RUP) adjusted for each other, sex and age. P-values in bold typeset indicate significance ($P < 0.05$). (-): Unable to pool genotype data due to significant genotype and/or allele frequency differences between injury cohorts.

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

Table 3.6: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) and Achilles tendon rupture (RUP) groups in the combined-Achilles tendon injury cohort

	SAW		UK		P-values	
	CON ^{AT}	RUP	CON	RUP	Countries	Diagnostic Groups
rs2567706	133	40	120	24		SAW+UK
AA	53 (71)	50 (20)	53 (63)	71 (17)	0.306	0.575
AG	38 (50)	48 (19)	34 (41)	25 (6)		
GG ^a	9 (12)	3 (1)	13 (16)	4 (1)		
G allele	28 (74)	26 (21)	30 (73)	17 (8)	0.932	0.171
HWE	0.532	0.237	0.057	0.501		
rs2241671	123	32	120	25		
GG	30 (37)	41 (13)	25 (30)	44 (11)	0.650	0.074
GA	47 (58)	37 (12)	52 (62)	52 (13)		(0.047)
AA	23 (28)	22 (7)	23 (28)	4 (1)		
A allele	46 (114)	41 (26)	49 (118)	30 (15)	0.872	0.030
HWE	0.725	0.270	0.860	0.371		
rs2567705	116	31	114	25		
AA	42 (49)	42 (13)	35 (40)	44 (11)	0.376	0.257
AT	41 (48)	52 (16)	44 (50)	44 (11)		
TT	16 (19)	7 (2)	21 (24)	12 (3)		
T allele	37 (86)	32 (20)	43 (98)	34 (17)	0.186	0.212
HWE	0.444	0.436	0.579	1.000		

Genotypes are expressed as percentages with participant counts (n) in parentheses. P-values are for the difference between countries (SAW vs. UK) and between diagnostic groups (CON vs. RUP) adjusted for each other, sex and age. P-values in bold typeset indicate significance ($P < 0.05$). P-values in parentheses are generated from conducting post-hoc analyses based on an *a priori* hypothesis that the genotype is associated with the risk of injury.

^a rs2567706 AA genotype vs. AG + GG genotype

3.2.2.3. Haplotype analysis

Inferred haplotypes were constructed using the genotype data from the four *TNC* variants: rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C). Three *TNC* haplotypes (T-C-A-T, T-C-T-T and C-C-T-T) inferred with a frequency greater than 5%, were common to the SAW- and UK-Achilles tendon injury cohorts. It must be noted that the inferred haplotypes in the combined-Achilles tendon injury cohort were constructed from only three of the four genotyped *TNC* variants (rs1061494, rs1138545 and rs1061495) since rs2104772 demonstrated significant genotype and allele frequency differences between the injury cohorts, which prohibited the pooling of genotype data.

No significant differences in the distribution of inferred haplotypes constructed from all four variants were noted between the SAW-CON^{AT} and SAW-TEN groups ($P=0.642$) and similarly between the UK-CON and UK-TEN groups ($P=0.897$) (Figure 3.5A and Figure 3.5B). However, the frequency distribution of the T-T inferred haplotype constructed from the two variants independently associated with risk of AT (rs1061494 and rs2104772) was significantly different between the SAW-CON^{AT} and SAW-TEN groups. The T-T allele combination was significantly overrepresented in the SAW-CON^{AT} group (30%, $n=40$) compared to the SAW-TEN group (22%, $n=17$) ($P=0.032$, Haplo.score: -2.146) (Figure 3.5C). Conversely, the frequency distribution of this two-variant inferred haplotype was not significantly different between the diagnostic groups in the UK-Achilles tendon injury cohort (Figure 3.5D). None of the inferred haplotypes constructed from the three *TNC* variants in the combined-Achilles tendon injury cohort were observed to differ significantly between the CON and TEN groups ($P=0.344$) and similarly between the CON and RUP groups ($P=0.957$) (Figure 3.5E).

With respect to the *COL27A1* gene, inferred haplotypes were constructed using the genotype data from the three genotyped variants: rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T). Four *COL27A1* haplotypes (A-G-A, G-A-T, A-A-A and A-G-T), inferred with a frequency greater than 5%, were common to the SAW- and UK-Achilles tendon injury cohorts.

None of the four inferred *COL27A1* haplotypes were observed to differ significantly in distribution between the CON and TEN groups in the SAW- ($P=0.636$), UK- ($P=0.600$) and combined-Achilles tendon injury ($P=0.698$) cohorts (Figures 3.6A, 3.6B and 3.6C). Similarly, no significant difference in the frequency distribution of inferred haplotypes were noted between the combined-CON and combined-RUP groups ($P=0.405$) (Figure 3.6C).

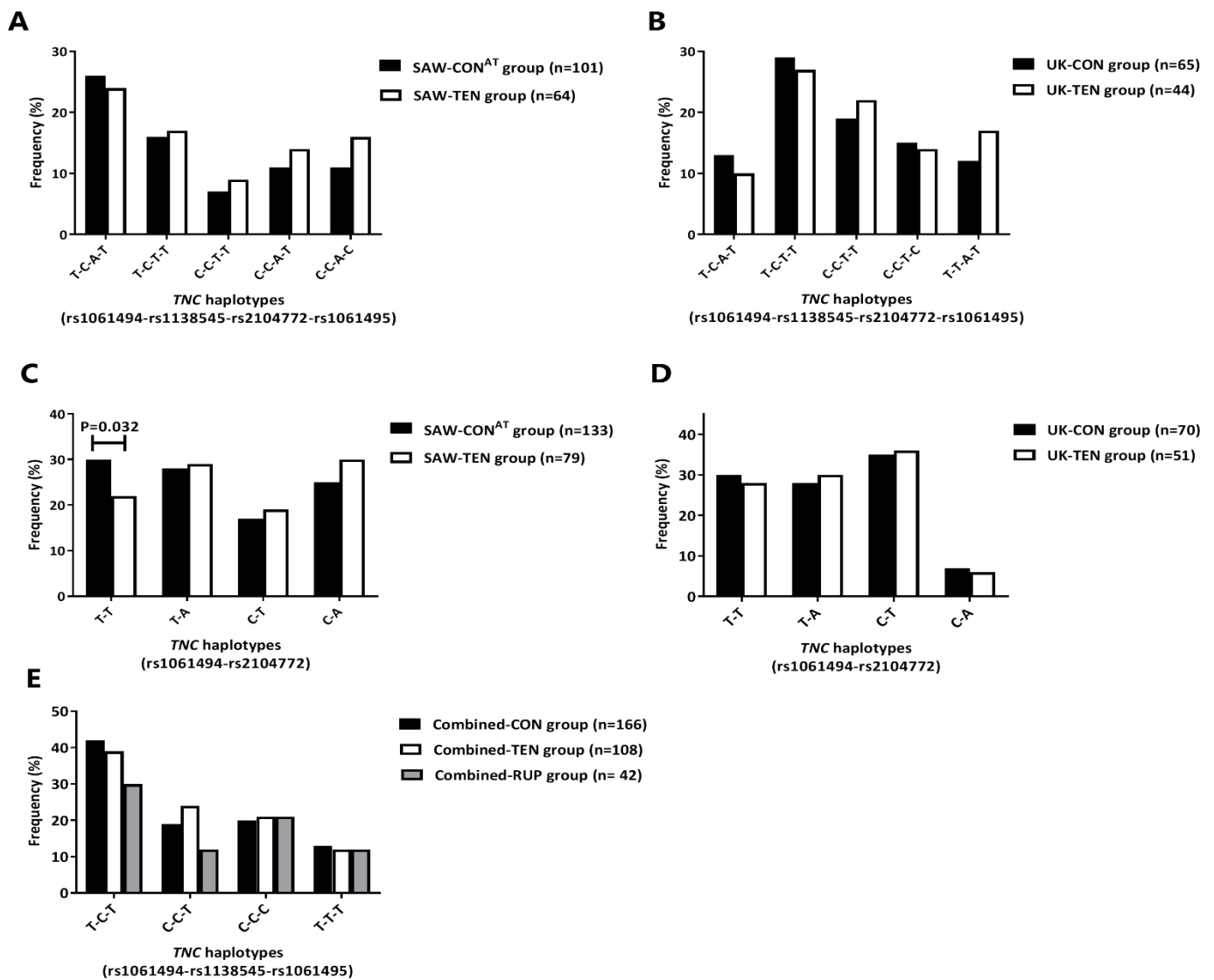


Figure 3.5: Inferred *TNC* haplotypes in the South African White (SAW)-, British (UK)- and Combined-Achilles tendon injury cohorts

Frequency distribution of the inferred *TNC* haplotypes constructed from the rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the (A) SAW- and (B) UK-Achilles tendon injury cohorts. (C) Frequency distribution of the inferred *TNC* haplotypes constructed from the rs1061494 (T>C) and rs2104772 (T>A) variants in the SAW-Achilles tendon injury cohort and (D) UK-Achilles tendon injury cohort. (E) Frequency distribution of the inferred *TNC* haplotypes constructed from the rs1061494, rs1138545 and rs1061495 variants in the combined-Achilles tendon injury cohort. CON: Group of control participants; TEN: Group of participants with chronic Achilles tendinopathy; RUP: Group of participants with acute Achilles tendon ruptures. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and sex, in addition to country of birth when the combined-Achilles tendon injury cohort was evaluated. The number of participants (n) in each group is in parentheses.

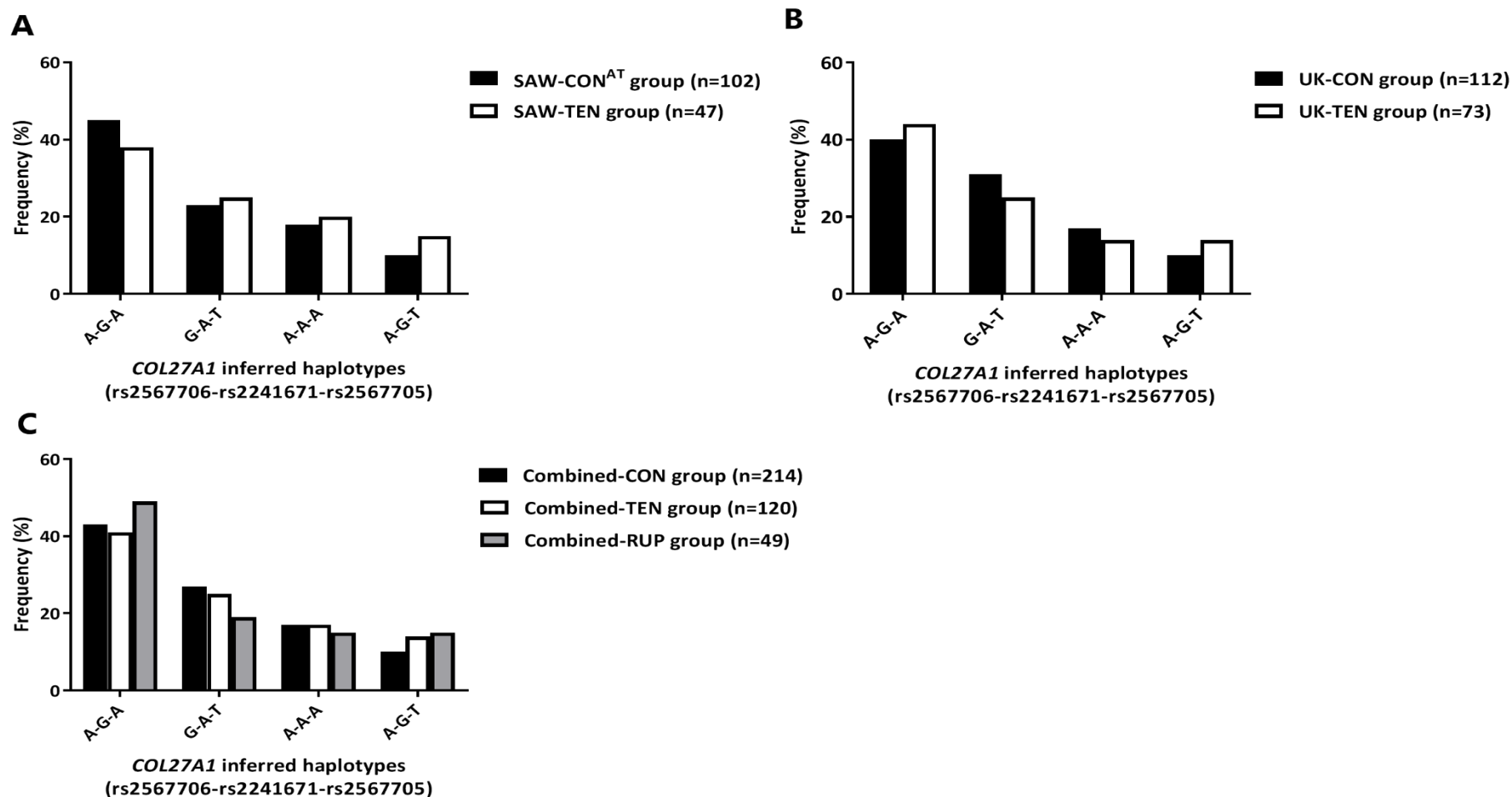


Figure 3.6: Inferred *COL27A1* haplotypes in the South African White (SAW)-, British (UK)- and combined-Achilles tendon injury cohorts

Frequency distribution of inferred *COL27A1* haplotypes constructed from the rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the (A) SAW- (B) UK- and (C) combined-Achilles tendon injury cohorts. CON: Group of control participants; TEN: Group of participants with chronic Achilles tendinopathy; RUP: Group of participants with acute Achilles tendon ruptures. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and sex, in addition to country of birth when the combined-Achilles tendon injury cohort was evaluated. The number of participants (n) in each group is in parentheses.

3.2.2.4. Linkage disequilibrium analysis

The rs1138545 variant demonstrated the strongest pairwise linkage with the other *TNC* variants: rs1061494, rs2104772 and rs1061495 in the SAW-Achilles tendon injury cohort ($D' \geq 0.890$). However, the alleles of the loci were weakly correlated ($r^2 \leq 0.218$) (Figure 3.7A). Similar results were noted for the UK-Achilles tendon injury cohort ($D' \geq 0.927$ and $r^2 \leq 0.521$) (Figure 3.8A). With respect to *COL27A1*, the rs2567706 variant was in high LD with rs2241671 and rs2567705 in both the SAW- and UK-Achilles tendon injury cohorts ($D' \geq 0.869$ and ≥ 0.920 respectively) and demonstrated alleles which were moderately associated ($r^2 \geq 0.658$ and ≥ 0.650 respectively). However, the *COL27A1* rs2241671 and rs2567705 variants were weakly linked in both Achilles tendon injury cohorts (SAW D' : 0.466 and r^2 : 0.398; UK D' : 0.490 and r^2 : 0.465) (Figures 3.7B and 3.8B).

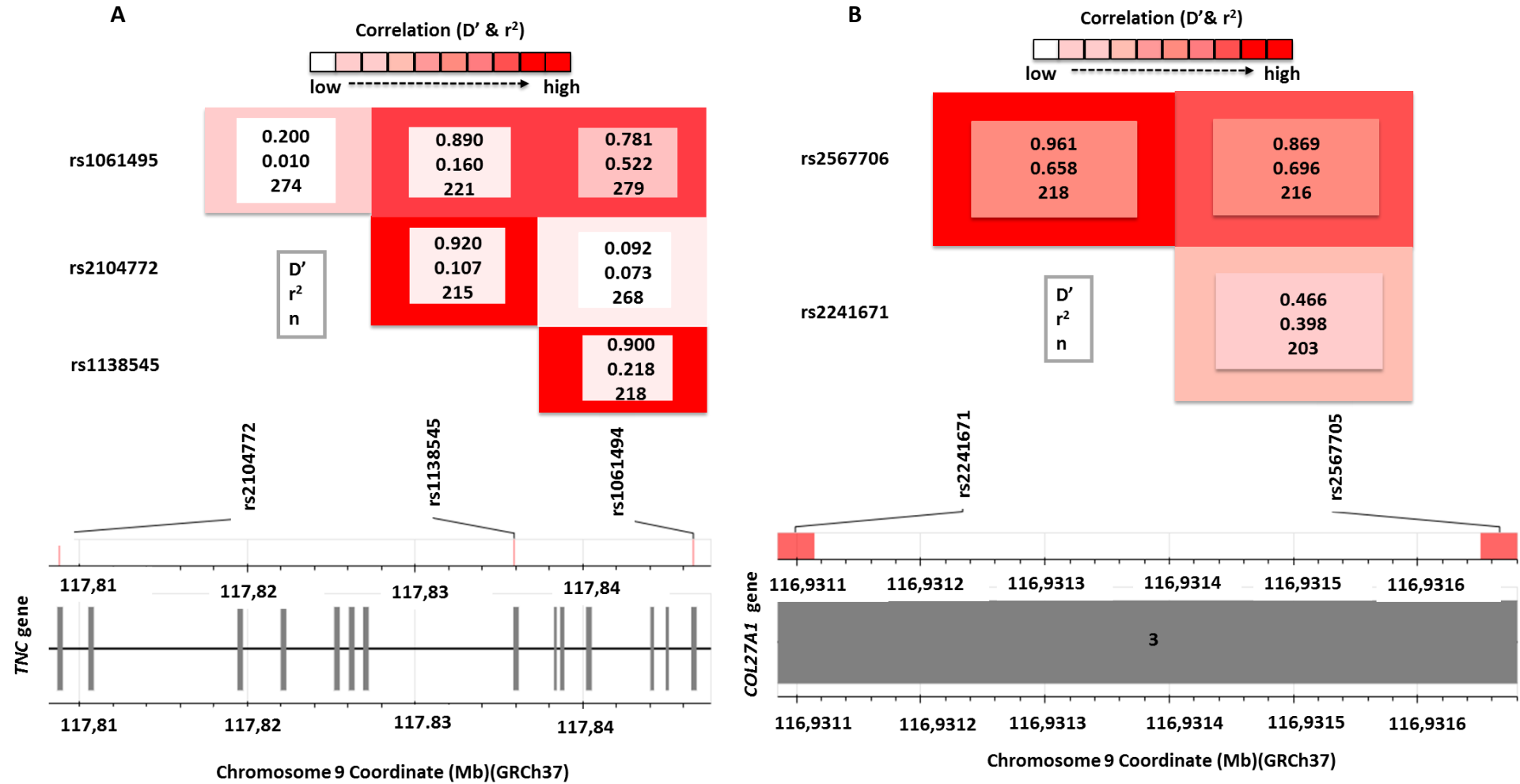


Figure 3.7: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *TNC* and *COL27A1* variants in the South African White (SAW)-Achilles tendon injury cohort

Results based on the (A) *TNC* and (B) *COL27A1* variants genotyped in the SAW-Achilles tendon injury cohort investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Total number of participants (n) for each pairwise analysis indicated. Grey blocks represent exons and horizontal lines represent introns.

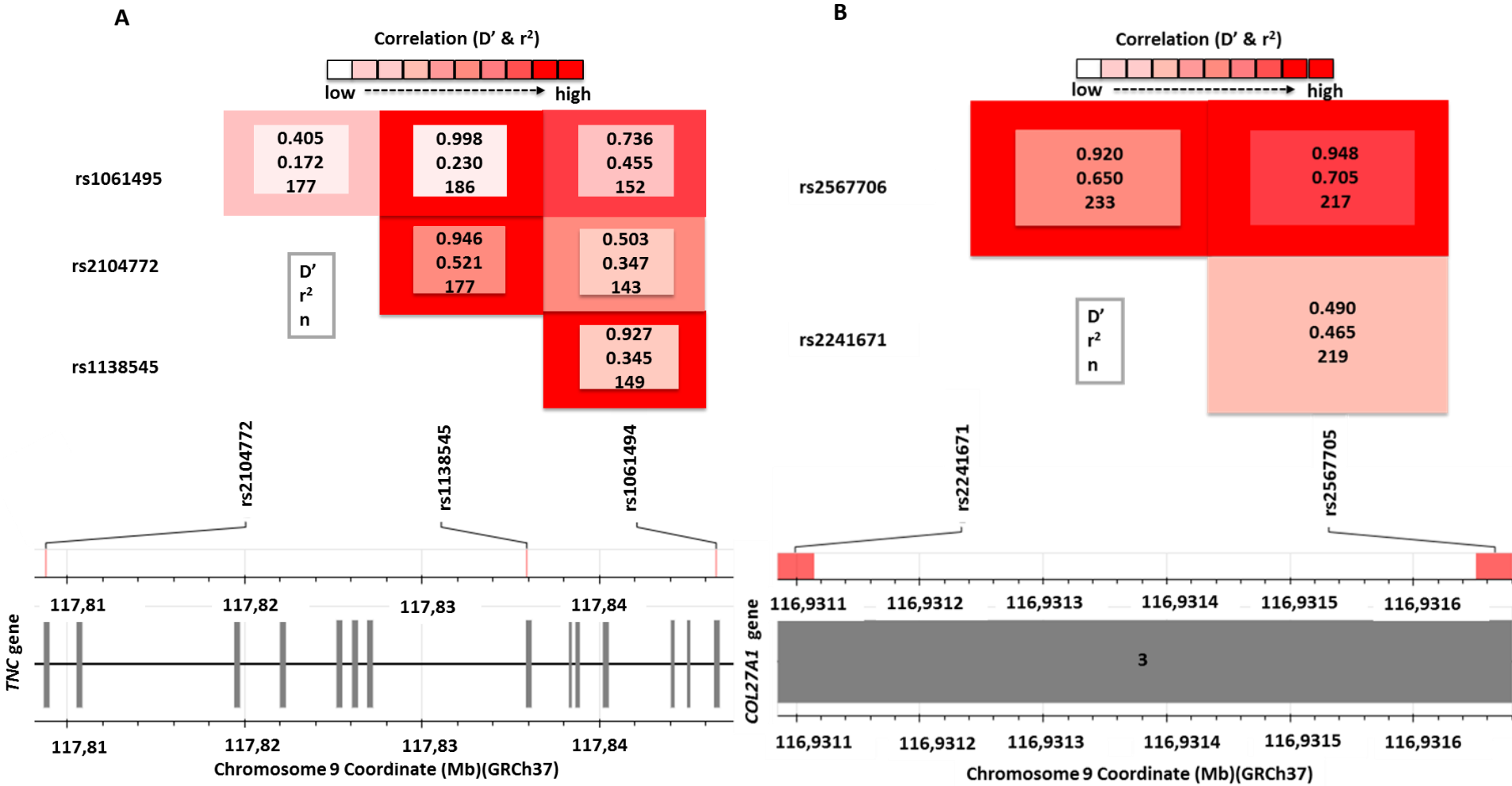


Figure 3.8: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *TNC* and *COL27A1* variants in the British (UK)-Achilles tendon injury cohort

Results based on the (A) *TNC* and (B) *COL27A1* variants genotyped in the UK-Achilles tendon injury cohort investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner block reflect the correlation (r^2) of alleles between loci. Total number of participants (n) for each pairwise analysis indicated. Grey blocks represent exons and horizontal lines represent introns.

3.2.3. Anterior cruciate ligament ruptures

Similar to the participants in the Achilles tendon injury cohorts, the participants within the SAW-and SAC-ACL rupture cohorts were genotyped for the four prioritised *TNC* variants: rs1061494 (T>C), rs2104772 (T>A), rs1138545 (C>T) and rs1061495 (T>C), in addition to the three prioritised *COL27A1* variants: rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T). The resultant genotype and allele frequency distributions are presented below.

3.2.3.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Sections 2.1.1.1B and 2.1.1.1C). None of the investigated variants demonstrated significant genotype associations with participant characteristics in either the SAW- or SAC-ACL rupture cohorts (Table 3.7).

Table 3.7: Genotype associations with the *TNC* rs1061494 (T>C), rs11384545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants, in addition to the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants and the physiological characteristics of recruited participants in the South African White (SAW)- and South African Coloured (SAC)-ACL rupture cohorts

	<i>TNC</i> variants							
	rs1061494		rs11384545		rs2104772		rs1061495	
	SAW	SAC	SAW	SAC	SAW	SAC	SAW	SAC
Age	0.767	0.270	0.241	0.239	0.403	0.471	0.231	0.824
Sex	0.153	0.420	0.757	0.112	0.462	0.236	0.884	0.349
Body mass	0.158	0.411	0.277	0.488	0.881	0.631	0.482	0.776
Height	0.227	0.151	0.961	0.597	0.993	0.998	0.855	0.101
BMI	0.551	0.082	0.197	0.913	0.098	0.415	0.592	0.494

	<i>COL27A1</i> variants					
	rs2567706		rs2241671		rs2567705	
	SAW	SAC	SAW	SAC	SAW	SAC
Age	0.655	0.161	0.271	0.191	0.662	0.624
Sex	0.166	0.162	0.773	0.127	0.508	0.465
Body mass	0.306	0.052	0.630	0.123	0.705	0.574
Height	0.975	0.230	0.227	0.732	0.624	0.941
BMI	0.682	0.468	0.477	0.086	0.266	0.826

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex.

3.2.3.2. Genotype and allele frequencies

The two South African ACL rupture cohorts were investigated independently from one another based on the knowledge that the SAC-ACL rupture cohort was sampled from a population with a unique genetic and social history [296,297].

After adjusting for age, sex and body mass, no significant differences in either genotype or allele frequencies were observed between the SAW-CON^{ACL} and SAW-ACL groups for the investigated *TNC* loci (Table 3.8). Similar results were observed when male participants were evaluated independently (Supplementary Table IIIA.1). However, when female participants were analysed separately (Supplementary Table IIIA.2), the rs2104772 variant was associated with ACL injury risk ($P=0.048$). Specifically, the AA genotype was significantly underrepresented in the SAW-CON^{ACL} group (18%, $n=18$) compared to the SAW-ACL group (33%, $n=18$, $P=0.035$, OR: 2.3, 95% CI: 1.1-5.5) (Figure 3.9). No significant differences in the distribution of genotype or allele frequencies were observed between the diagnostic groups in the SAW-ACL cohort for the three investigated *COL27A1* loci (Table 3.9). Similar results were observed when the male and female subgroups were evaluated independently (Supplementary Tables IIIA.3. and IIIA.4).

Table 3.8: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs1061494	216	198		125	
TT	34 (74)	31 (61)	0.917	35 (44)	0.598
TC	49 (106)	53 (105)		50 (62)	
CC	17 (36)	16 (32)		15 (19)	
C Allele	41 (178)	43 (169)	0.666	40 (100)	0.807
HWE	0.891	0.105		0.720	
rs1138545	197	180		112	
CC	74 (146)	71 (127)	0.489	69 (77)	0.356
CT	21 (41)	27 (48)		29 (32)	
TT ^a	5 (10)	3 (5)		3 (3)	
T Allele	11 (43)	16 (58)	0.850	17 (38)	0.668
HWE	0.014	0.804		1.000	
rs2104772	221	199		126	
TT	31 (69)	32 (64)	0.846	31 (39)	0.816
TA	44 (96)	43 (85)		44 (56)	
AA	25 (56)	25 (50)		25 (31)	
A allele	47 (208)	46 (185)	0.897	47 (118)	1.000
HWE	0.111	0.356		0.389	
rs1061495	221	198		126	
TT	53 (118)	60 (119)	0.169	65 (82)	0.129
TC	39 (85)	33 (66)		28 (35)	
CC ^b	8 (18)	7 (13)		7 (9)	
C Allele	27 (121)	23 (92)	0.188	21 (53)	0.080
HWE	0.618	0.460		0.124	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for sex, body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

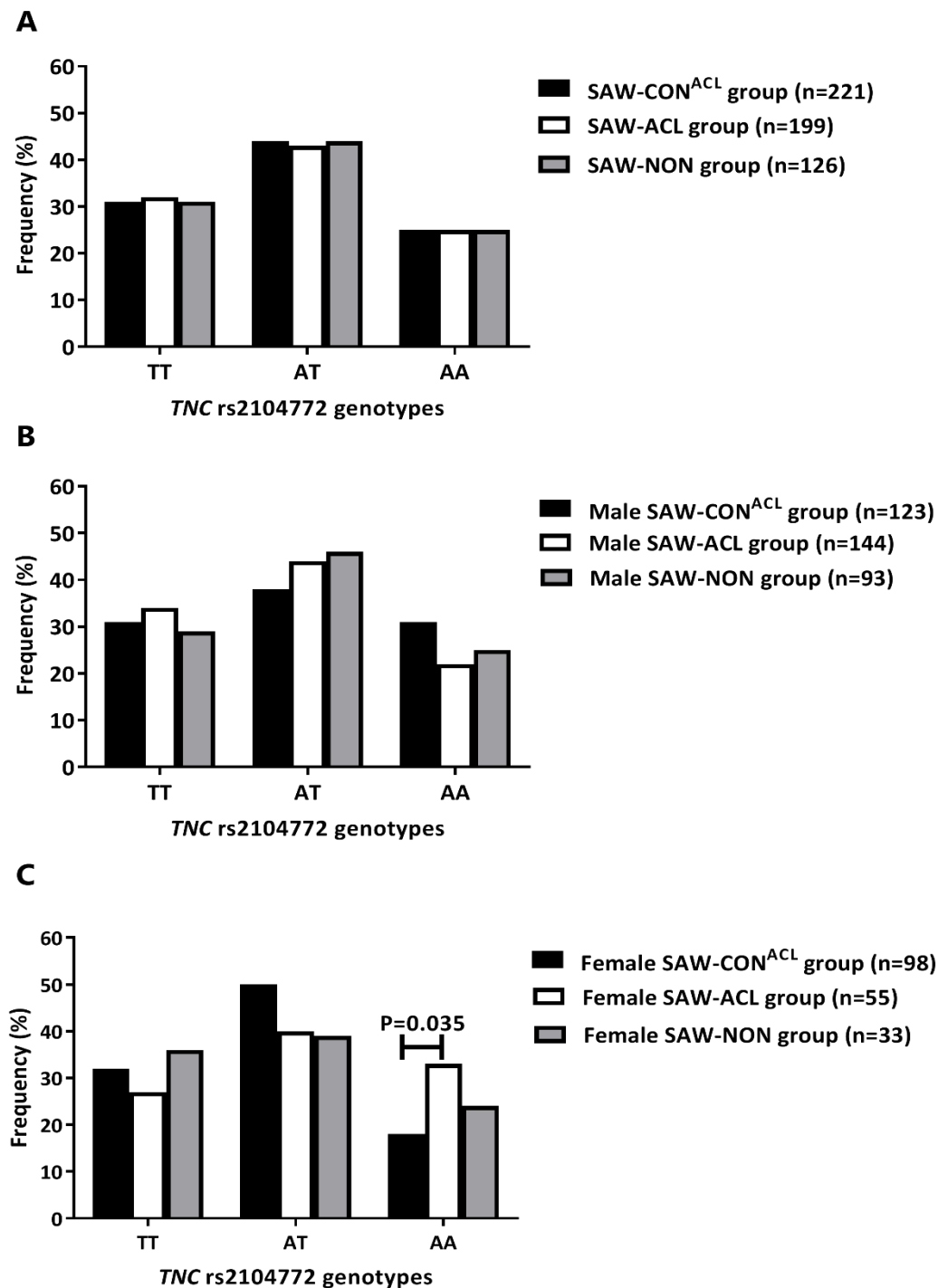


Figure 3.9: Genotype frequency distributions of the *TNC* rs2104772 (T>A) variant in the South African White (SAW)-ACL rupture cohort

Genotype frequency distributions of the *TNC* rs2104772 (T>A) variant in the SAW-ACL rupture cohort when (A) all participants, (B) the male subgroup and (C) the female subgroup was evaluated. CON: Group of control participants; ACL: Group of participants with ACL ruptures; NON: Subgroup of participants reporting a non-contact mechanism of ACL injury. Significant differences in genotype frequencies between groups are depicted on the graphs, with P-values adjusted for age and body mass, in addition to sex when all participants were evaluated. The number of participants in each group (n) is in parentheses.

Table 3.9: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs2567706	219	198		126	
AA	56 (122)	53 (105)	0.623	54 (68)	0.822
AG	37 (82)	37 (74)		35 (44)	
GG ^a	7 (15)	10 (19)		11 (14)	
G allele	26 (112)	28 (112)	0.402	29 (72)	0.437
HWE	0.860	0.202		0.098	
rs2241671	213	193		120	
GG	31 (66)	31 (59)	0.787	30 (36)	0.751
GA	51 (109)	52 (101)		50 (60)	
AA	18 (38)	17 (33)		20 (24)	
A allele	43 (185)	43 (167)	1.000	45 (108)	0.742
HWE	0.683	0.417		1.000	
rs2567705	213	189		120	
AA	42 (90)	34 (65)	0.762	41 (49)	0.504
AT	44 (93)	48 (90)		40 (48)	
TT	14 (30)	18 (34)		19 (23)	
T allele	36 (153)	42 (158)	0.102	40 (94)	0.473
HWE	0.555	0.783		0.102	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for sex, body mass and age at the time of recruitment or injury.

^a rs2567706 AA genotype vs. AG + GG genotype

With reference to the SAC-ACL rupture cohort, the female subgroup was too small for independent analysis. Therefore, only the male subgroup was investigated independently. After adjusting for age, no significant differences in the genotype or allele frequencies were observed between the diagnostic groups for the investigated *TNC* loci: rs1061494, rs1138545, rs2104772 and rs1061495 (Table 3.10). This observation was similar when male participants were evaluated independently (Supplementary Table IIIA.5). No significant differences in the distribution of genotype or allele frequencies were observed between the diagnostic groups in the SAC-ACL cohort and SAC-ACL male subgroup for two of the three investigated *COL27A1* loci: rs2567706 and rs2567705 (Table 3.11 and Supplementary Table IIIA.6). However, the genotype frequency distribution of the rs2241671 variant was observed to differ significantly. Following an *a priori* hypothesis, the rs2241671 GG genotype was found to be significantly less frequent in the SAC-CON group (39%, n=38) compared to the SAC-NON subgroup (58%, n=28, $P=0.021$, OR: 2.3, 95% CI: 1.1-4.6). In addition, the G allele was significantly underrepresented in the SAC-CON group (62%) compared to the SAC-NON subgroup (75%, $P=0.035$). Similarly, when male participants were evaluated independently (Supplementary Table IIIA.6), the GG genotype remained significantly underrepresented in the SAC-CON group (41%, n=32) compared to the SAC-NON subgroup (62%, n=26, $P=0.030$, OR: 2.4, 95% CI: 1.0-5.1) and the G allele was significantly less frequent in the SAC-CON group (63%) compared to the SAC-NON subgroup (77%, $P=0.027$) (Figure 3.10).

Table 3.10: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African Coloured (SAC)-ACL rupture cohort

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
rs1061494	98	84		46	
TT	37(36)	33 (28)	0.530	30 (14)	0.420
CT	46 (45)	43 (36)		46 (21)	
CC	17 (17)	24 (20)		24 (11)	
C Allele	40 (79)	40 (76)	0.417	47 (43)	0.325
HWE	0.675	0.207		0.564	
rs1138545	87	85		47	
CC	83 (72)	85 (72)	0.837	89 (42)	0.447
CT	17 (15)	14 (12)		11 (5)	
TT ^a	0 (0)	1 (1)		0 (0)	
T Allele	7 (15)	8 (14)	1.000	5 (5)	0.328
HWE	1.000	0.518		1.000	
rs2104772	95	88		49	
TT	35 (33)	30 (26)	0.245	29 (14)	0.775
AT	52 (49)	49 (43)		55 (27)	
AA	14 (13)	22 (19)		16 (8)	
A allele	39 (75)	46 (81)	0.239	44 (43)	0.509
HWE	0.671	1.000		0.563	
rs1061495	93	87		47	
TT	48 (45)	54 (47)	0.460	53 (25)	0.721
TC	45 (42)	39 (34)		40 (19)	
CC ^b	7 (6)	7 (6)		6 (3)	
C Allele	29 (54)	26 (46)	0.630	27 (25)	0.666
HWE	0.458	1.000		1.000	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for age at the time of recruitment or injury.

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

Table 3.11: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African Coloured (SAC)-ACL rupture cohort

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
rs2567706	96	90		50	
AA	58 (56)	64 (58)	0.452	68 (34)	0.290
AG	37 (35)	32 (29)		30 (15)	
GG ^a	5 (5)	3 (3)		2 (1)	
G allele	23 (45)	19 (35)	0.375	17 (17)	0.225
HWE	0.788	0.756		1.000	
rs2241671	98	88		48	
GG	39 (38)	48 (42)	0.442	58 (28)	0.066
GA	47 (46)	40 (35)		33 (16)	(0.021)
AA	14 (14)	13 (11)		8 (4)	
A allele	38 (74)	32 (57)	0.340	25 (24)	0.035
HWE	0.834	0.360		0.477	
rs2567705	96	89		50	
AA	33 (32)	39 (35)	0.752	46 (23)	0.336
AT	50 (48)	44 (39)		44 (22)	
TT	17 (16)	17 (15)		10 (5)	
T allele	42 (80)	39 (69)	0.602	32 (32)	0.126
HWE	1.000	0.666		1.000	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for age at the time of recruitment or injury. P-values in bold typeset indicate significance ($P < 0.05$). P-values in parentheses are generated from conducting post-hoc analyses based on an *a priori* hypothesis that the genotype is associated with the risk of injury.

^a rs2567706 AA genotype vs. AG + GG genotype

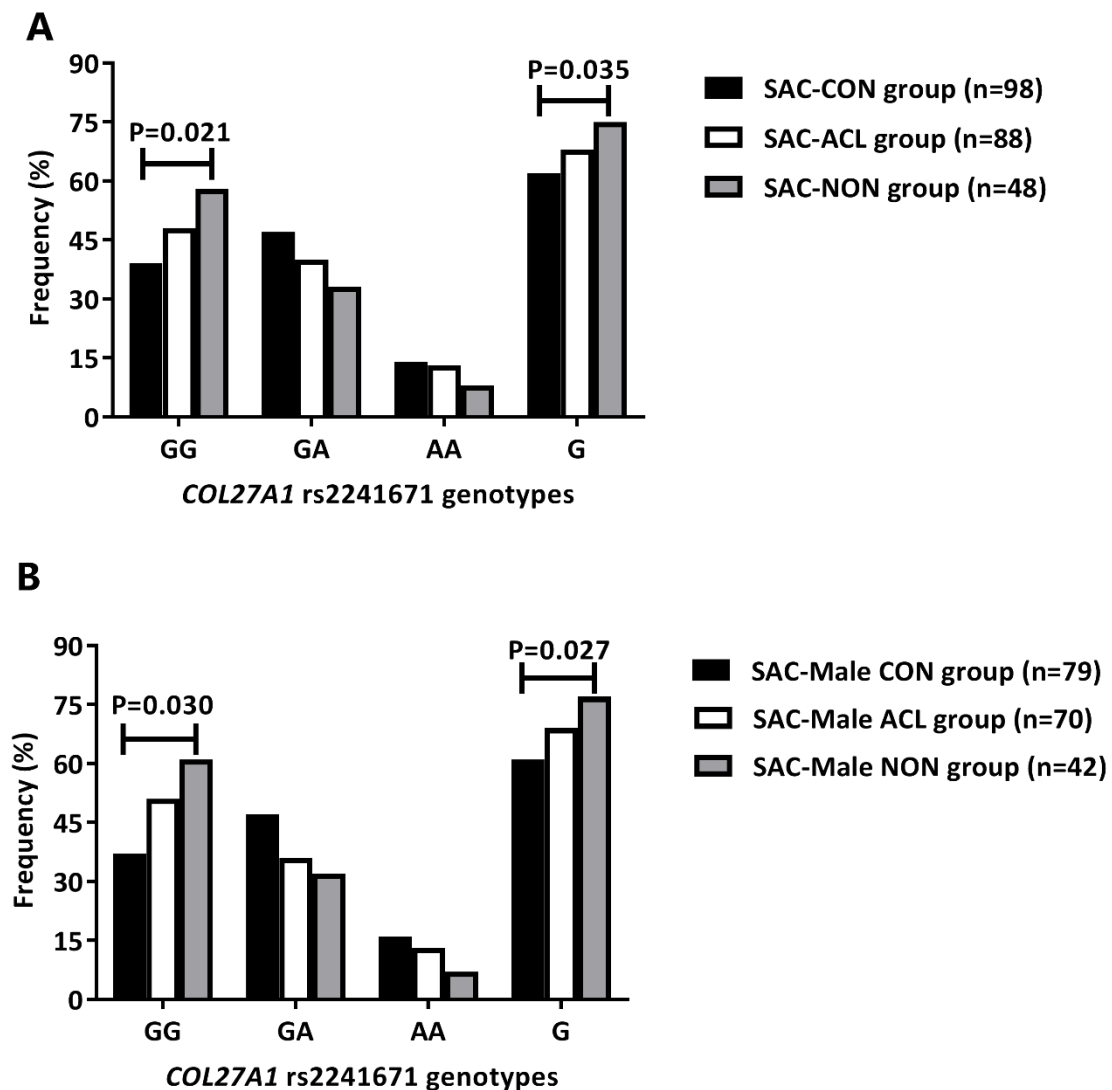


Figure 3.10: Genotype frequency distributions of the *COL27A1* rs2241671 (G>A) variant in the South African Coloured (SAC)-ACL rupture cohort

Genotype frequency distributions of the *COL27A1* rs2241671 (G>A) variant in the SAC-ACL cohort when (A) all participants and (B) the male subgroup was evaluated. CON: Group of control participants; ACL: Group of participants with ACL ruptures; NON: Subgroup of participants reporting a non-contact mechanism of ACL injury. Significant differences in genotype and/or allele frequencies between groups are depicted on the graphs, with P-values adjusted for age. The number of participants in each group (n) is in parentheses.

3.2.3.3. Haplotype analysis

Inferred haplotypes were constructed using the genotype data from the four *TNC* variants: rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) and the three *COL27A1* variants: rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T). Three *TNC* haplotypes (T-C-A-T, T-C-T-T and C-C-T-T) and four *COL27A1* haplotypes (A-G-A, G-A-T, A-A-A, A-G-T), inferred with a frequency greater than 10%, were common to the SAW- and SAC-ACL rupture cohorts.

There were no significant differences in the frequency distribution of the five inferred *TNC* haplotypes between the SAW-CON^{ACL} and SAW-ACL groups ($P=0.747$) and similarly between the SAW-CON^{ACL} group and SAW-NON subgroup ($P=0.600$), when all participants were analysed (Figure 3.11A). Similar results were observed when the male SAW-CON^{ACL} group was compared to the male SAW-ACL group ($P=0.735$) and male SAW-NON subgroup ($P=0.703$) (Figure 3.11B). Furthermore, the frequency distribution of the inferred haplotypes constructed from the two *TNC* variants independently associated with AT risk, rs1061494 and rs2104772, were not significantly different between the SAW-CON^{ACL} and SAW-ACL groups and similarly between the SAW-CON^{ACL} group and SAW-NON subgroup (Figure 3.11C and 3.11D).

No inferred haplotypes constructed from the three investigated *COL27A1* variants were significantly different between the SAW-CON^{ACL} and SAW-ACL groups ($P=0.234$). The frequencies of these haplotypes were also similarly distributed between the SAW-CON^{ACL} group and SAW-NON subgroup ($P=0.089$) (Figure 3.12A). Furthermore, no significant frequency differences were noted when male participants were evaluated independently (ACL: $P=0.211$, NON: $P=0.845$) (Figure 3.12B).

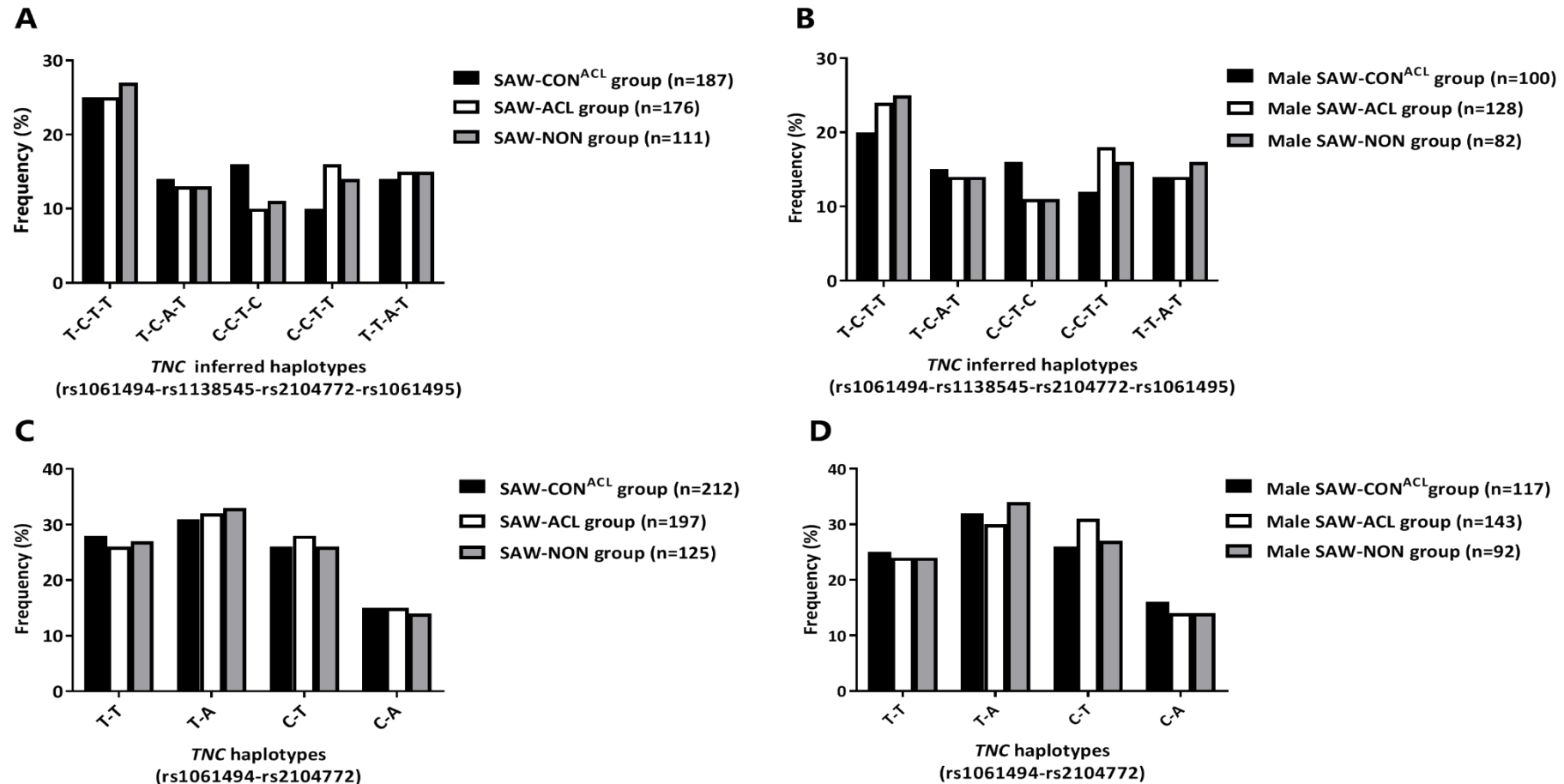


Figure 3.11: Inferred *TNC* haplotypes in the South African White (SAW)-ACL rupture cohort

Frequency distribution of inferred *TNC* haplotypes constructed from the rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the (A) SAW-ACL cohort and (B) male subgroup. Frequency distribution of inferred *TNC* haplotypes constructed from the rs1061494 (T>C) and rs2104772 (T>A) variants, in the (C) SAW-ACL cohort and (D) male subgroup. CON: Group of control participants; ACL: Group of participants with ACL ruptures; NON: Subgroup of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and body mass, in addition to sex when all participants were evaluated. The number of participants (n) in each group is in parentheses.

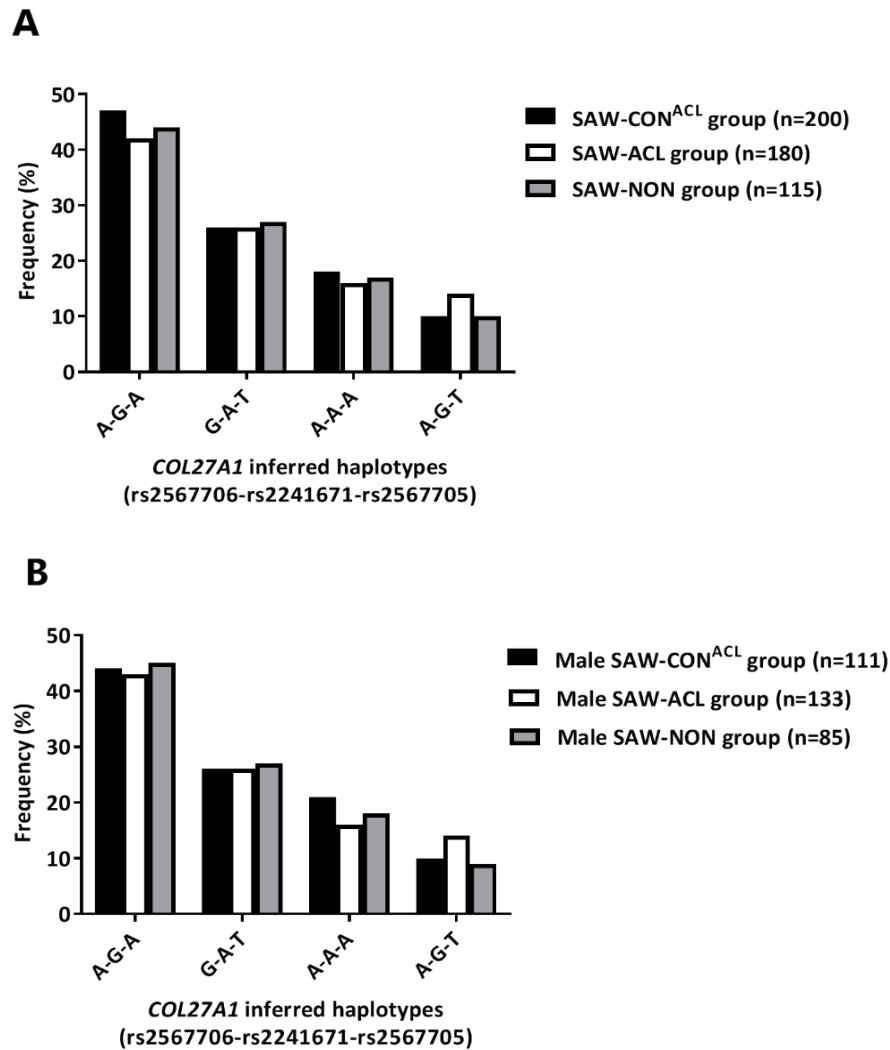


Figure 3.12: Inferred *COL27A1* haplotypes in the South African White (SAW)-ACL rupture cohort

Frequency distribution of inferred *COL27A1* haplotypes constructed from the rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the (A) SAW-ACL cohort and (B) male subgroup. CON: Group of control participants; ACL: Group of participants with ACL ruptures; NON: Subgroup of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and body mass, in addition to sex when all participants were evaluated. The number of participants (n) in each group is in parentheses.

No significant differences in the frequency distribution of inferred haplotypes constructed from the four genotype *TNC* variants were noted between the SAC-CON and SAC-ACL groups ($P=0.171$) and similarly between the SAC-CON group and SAC-NON subgroup ($P=0.544$) (Figure 3.13A). Similar results were observed when male participants were evaluated independently (ACL: $P=0.133$, NON: $P=0.453$) (Figure 3.13B). However, following an *a priori* hypothesis, the T-T (rs1061494-rs2104772) inferred haplotype, previously associated with reduced risk of AT (Section 3.2.2.3) was also associated with decreased risk for ACL ruptures in the SAC-ACL cohort. Specifically, the T-T allele combination was significantly overrepresented in the SAC-CON group (38%, $n=36$) compared to the SAC-ACL group (28%, $n=24$, $P=0.031$, Haplo.score: -2.157) and similarly when male participants were independently evaluated (CON: 40%, $n=30$, ACL: 27%, $n=19$, $P=0.022$, Haplo.score: -2.286) (Figure 3.13C and 3.13D). This haplotype also demonstrated a trend towards statistical significance when the SAC-CON group was compared to the SAC-NON subgroup ($P=0.052$) and likewise when male participants were analysed independently ($P=0.058$).

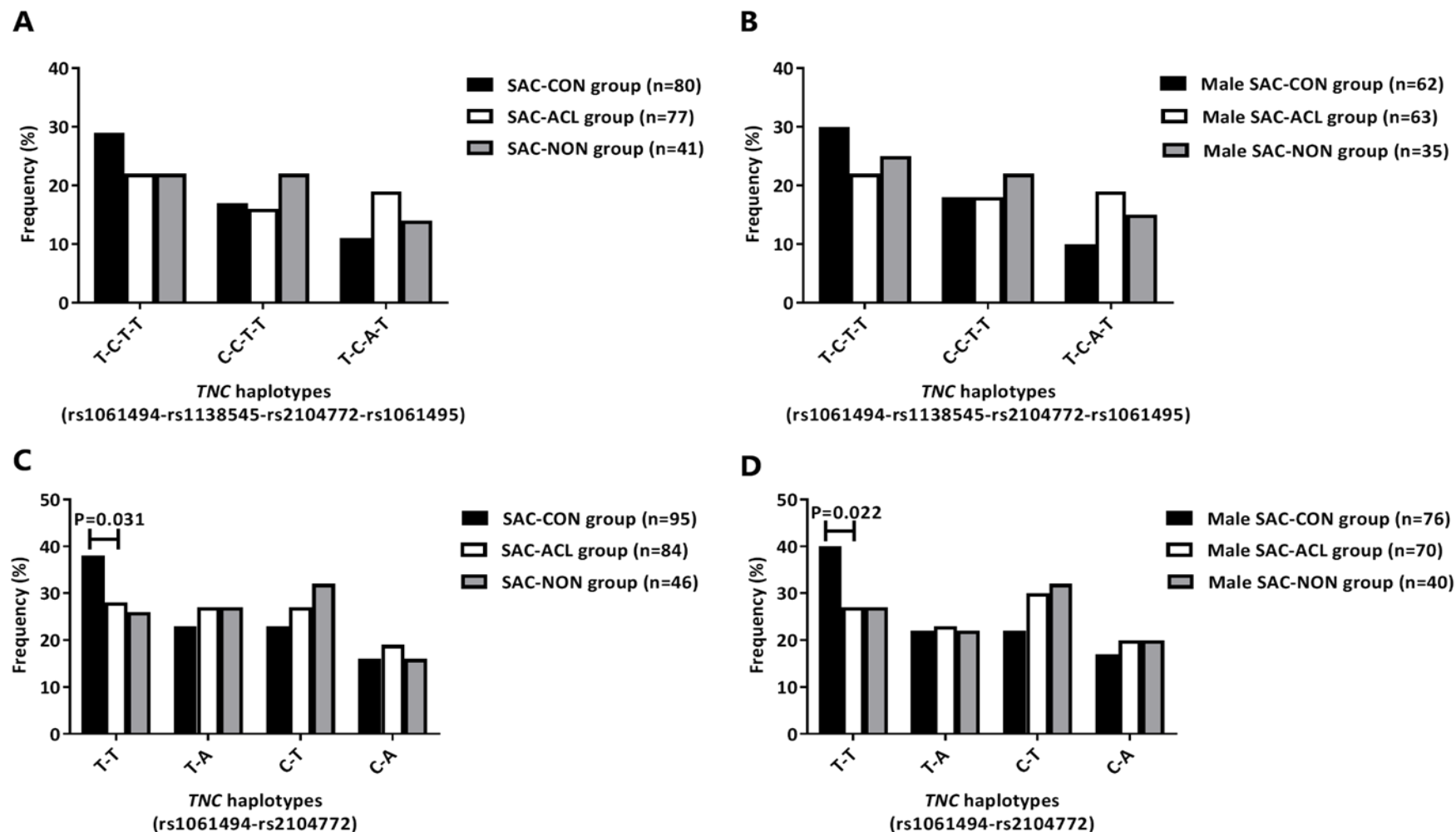


Figure 3.13: Inferred *TNC* haplotypes in the South African Coloured (SAC)-ACL rupture cohort

Frequency distribution of inferred *TNC* haplotypes constructed from the rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the (A) SAC-ACL rupture cohort and (B) male subgroup. Frequency distribution of inferred *TNC* haplotypes constructed from the rs1061494 (T>C) and rs2104772 (T>A) variants, in the (C) SAC-ACL cohort and (D) male subgroup. CON: Group of control participants; ACL: Group of participants with ACL ruptures; NON: Subgroup of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graphs, with P-values adjusted for age. The number of participants (n) in each group is in parentheses.

No significant differences were noted between the SAC-CON and SAC-ACL groups in the frequency of inferred haplotypes constructed from all three genotyped *COL27A1* variants when all participants were compared ($P=0.642$) as well as when the male subgroup was evaluated independently ($P=0.423$) (Figure 3.14A and 3.14B). However, the A-G-A inferred haplotype was significantly underrepresented in the SAC-CON group (41%, $n=40$) compared to the SAC-NON subgroup (53%, $n=25$, $P=0.049$, Haplo.score: 1.966). Moreover, the alternate G-A-T inferred haplotype was significantly overrepresented in the SAC-CON group (21%, $n=20$) compared to the SAC-NON group (11%, $n=5$, $P=0.045$, Haplo.score: -2.003) (Figure 3.14A). Similar results were also observed when the male SAC-CON group and male SAC-NON subgroup were compared (A-G-A CON: 42%, $n=32$, NON: 56%, $n=24$, $P=0.032$, Haplo.score: 2.149; G-A-T CON: 22%, $n=17$, NON: 10%, $n=4$, $P=0.039$, Haplo.score: -2.069) (Figure 3.14B).

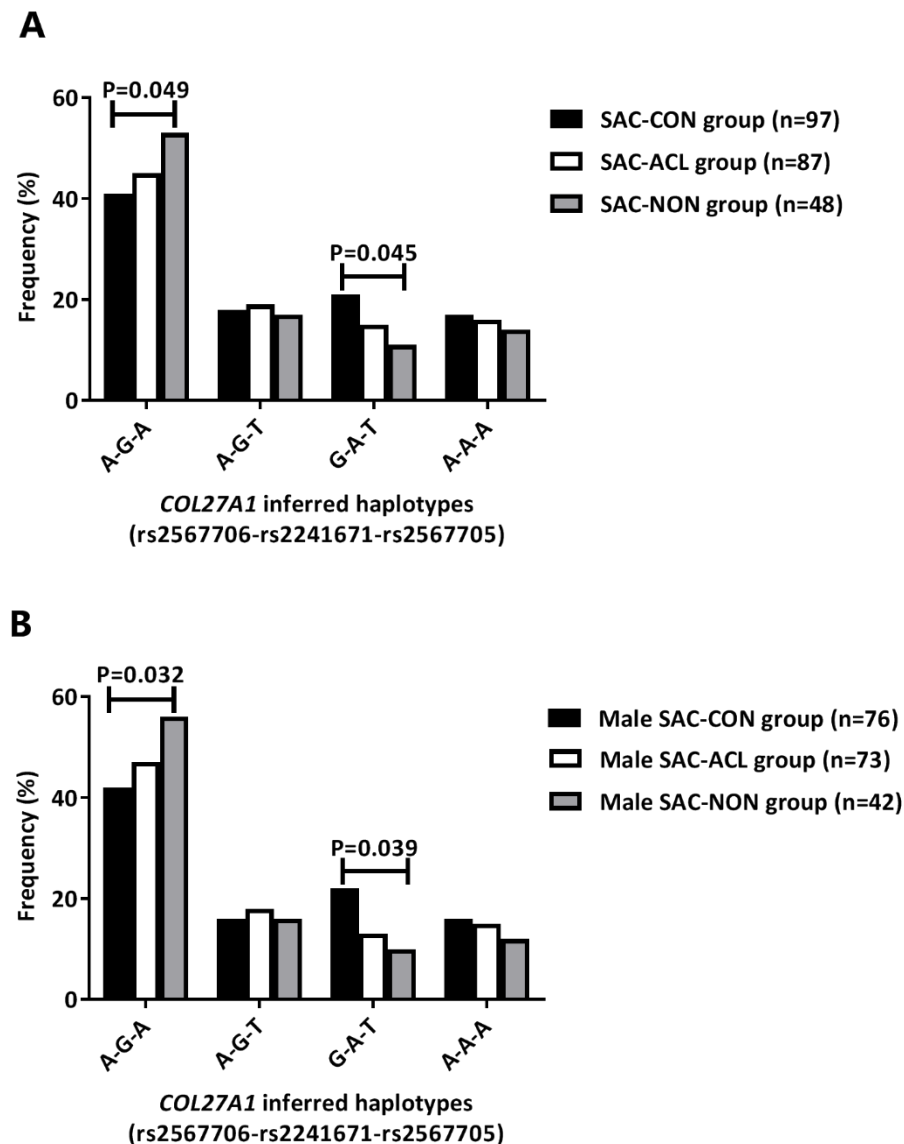


Figure 3.14: Inferred *COL27A1* haplotypes in the South African Coloured (SAC)-ACL cohort

Frequency distribution of inferred *COL27A1* haplotypes constructed from the rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the (A) SAC-ACL rupture cohort and (B) male subgroup. CON: Group of control participants; ACL: Group of participants with ACL ruptures; NON: Subgroup of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graphs, with P-values adjusted for age. The number of participants (n) in each group is in parentheses.

3.2.3.4. Linkage disequilibrium analysis

The *TNC* rs1138545 variant was in high LD with the rs2104772 and rs1061495 variants in both the SAW-and SAC-ACL rupture cohorts ($D' \geq 0.879$). However, the alleles of the loci were weakly correlated ($r^2 \leq 0.406$). The rs2104772 variant demonstrated the weakest pairwise linkage with the rs1061494 and rs1061495 variants in the SAW-ACL rupture cohort ($D' \leq 0.246$, $r^2 \leq 0.192$) (Figure 3.15A). These specific pairwise analyses were observed to be even weaker in the SAC-ACL rupture cohort ($D' \leq 0.069$, $r^2 \leq 0.049$) (Figure 3.16A). Based on the *COL27A1* genotype data for the SAW-ACL rupture cohort, the rs2567706 variant was in high LD with the rs2241671 and rs2567705 variants ($D' \geq 0.956$) and the alleles were moderately correlated ($r^2 \geq 0.662$) (Figure 3.15B). Similarly, the rs2567706 and rs2567705 variants were in high LD with moderate allele correlations ($D': 0.976$, $r^2: 0.636$) in the SAC-ACL rupture cohort but the rs2241671 variant was only moderately linked to rs2567706 ($D': 0.733$, $r^2: 0.527$) (Figure 3.16B).

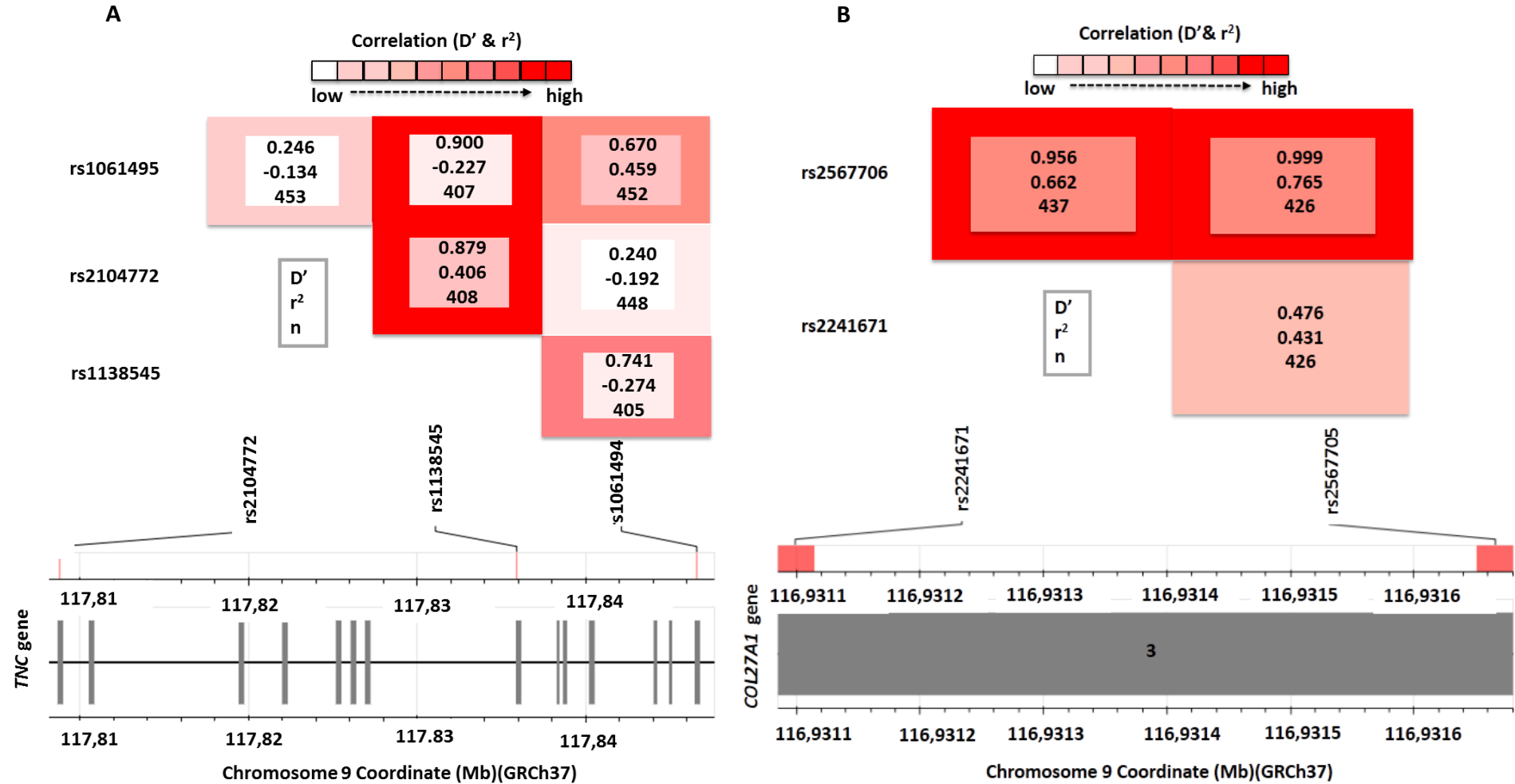


Figure 3.15: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *TNC* and *COL27A1* variants in the South African White (SAW)-ACL rupture cohort

Results based on the (A) *TNC* and (B) *COL27A1* variants genotyped in the SAW-ACL rupture cohort investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Total number of participants (n) for each pairwise analysis indicated. Grey blocks represent exons and horizontal lines represent introns.

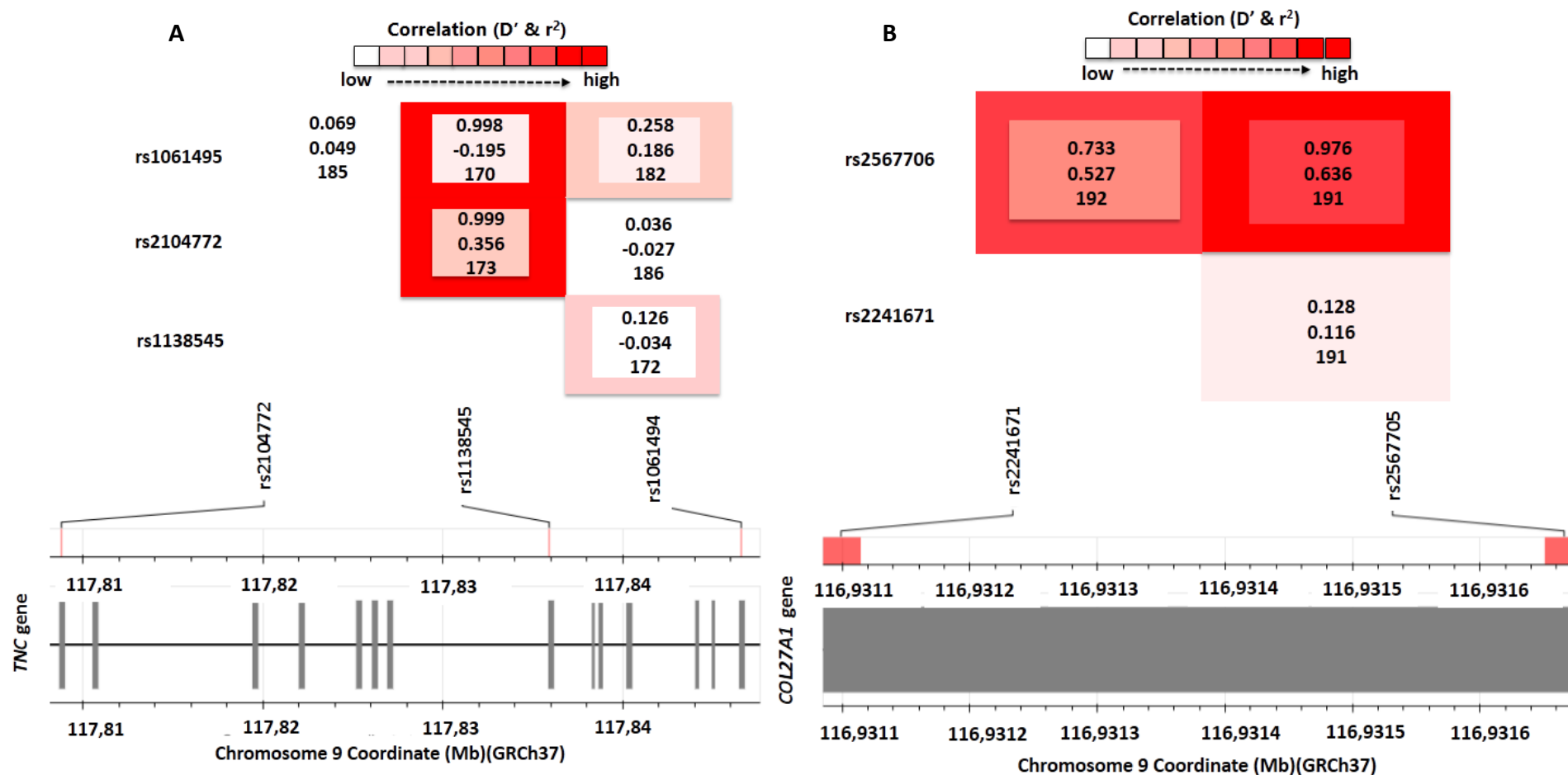


Figure 3.16: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the four genotyped *TNC* and three *COL27A1* variants in the South African Coloured (SAC)-ACL rupture cohort

Results based on the (A) *TNC* and (B) *COL27A1* variants genotyped in the SAC-ACL rupture cohort investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Total number of participants (n) for each pairwise analysis indicated. Grey blocks represent exons and horizontal lines represent introns.

3.3. DISCUSSION

Variants within the *TNC* and *COL27A1* genes have previously been associated with the risk of common sporting injuries including chronic AT and degenerative RCT [198,199,214,215]. Thus, this study aimed to further define the molecular signatures predisposing for musculoskeletal soft tissue injuries by revisiting these genomic intervals and genotyping several previously associated variants and/or candidates identified by WES and prioritised through the application of a customised tiered filtering strategy, in several independent AT and ACL rupture cohorts.

A total of 199 456 high quality variants were identified across the 20 sequenced exomes, 4 544 of which passed the initial filtering parameter, that of a 30% allele frequency difference between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. Only 22 variants (<1%) meeting the allele frequency threshold mapped to the genes previously implicated in the aetiology of chronic AT, 10 of which localised to the *TNC* and *COL27A1* genes. Interestingly, three of the six variants prioritised from the WES dataset; mapping to these two genes, were found to associate with the risk of either chronic AT, ACL ruptures or both.

The first novel finding was the association of the *TNC* rs1061494 (T>C) variant with injury risk. Specifically, the rs1061494 CC genotype inferred a 2.5-fold increased risk for chronic AT in the SAW-Achilles tendon injury cohort. This independent association was not replicated in the UK-Achilles tendon injury cohort or in either of the two South African ACL rupture cohorts. However, the *TNC* T-T haplotype, constructed from rs1061494 in combination with the downstream rs2104772 (T>A) variant, was associated with reduced risk for AT in the SAW-Achilles tendon injury cohort and ACL ruptures in the SAC-ACL rupture cohort (Figure 3.17). To our knowledge, this is the first study to present a *TNC*-specific haplotype, spanning approximately 38 kb of the gene associated with the risk of two clinically distinct musculoskeletal phenotypes. The association of this allele combination with injury risk provides support for the independent association of both variants with the risk of AT, rs1061494 in the current study and rs2104772 in a previous study by Saunders et al. (2013) [198]. Furthermore, the rs2104772 variant was observed to significantly associate with the

risk of ACL ruptures, the second novel finding of this study. Specifically, the AA genotype inferred a 2.3-fold increased risk for acute ACL ruptures in the female SAW-ACL rupture cohort. This variant is known to map within 2 kb of the GT dinucleotide repeat polymorphism, first proposed as a risk modifier of Achilles tendon injuries by Mokone et al. (2005) [199]. Moreover, in a recent study by Valdivieso et al. (2017), rs2104772 was associated with capillary remodelling following endurance exercise [318]. Specifically, carriers of the AA genotype demonstrated a significant increase in the density of capillaries within the skeletal muscle compared to carriers of the T allele (AT or TT, $P=0.004$). However, the sample size of this study was small ($n=61$), therefore, these results should be interpreted with caution. The novel findings of the current study support the growing body of evidence implicating the *TNC* gene and its encoded protein, a hexameric extracellular matrix glycoprotein, in pathologies characterised by dysregulated matrix remodelling. Although highly expressed during embryonic development and in tissues subjected to tensile loading including tendons and ligaments, *TNC* is under tight regulatory control, with negligible detection in most adult tissues under normal physiological conditions. However, transient expression is observed during wound healing, tissue remodelling and more persistently in pathological conditions [223].

In line with this, the association of the rs1061494 variant with injury risk, both independently and in combination with rs2104772 is an interesting observation. The rs1061494 variant is located within the first constitutively expressed FNIII repeat domain of the mature tenascin-C peptide and is a predicted expression quantitative trait locus (eQTL) in testis tissue according to the GTex consortium. By definition, an eQTL or a protein quantitative trait locus (pQTL) is a common DNA sequence variant with the ability to influence phenotypic traits including transcript levels (eQTL) or protein abundance (pQTL) [319,320]. Most of these variants are found close to or within the gene encoding the mRNA product; variants said to infer *cis*-effects, while others are off target and infer *trans*-effects.

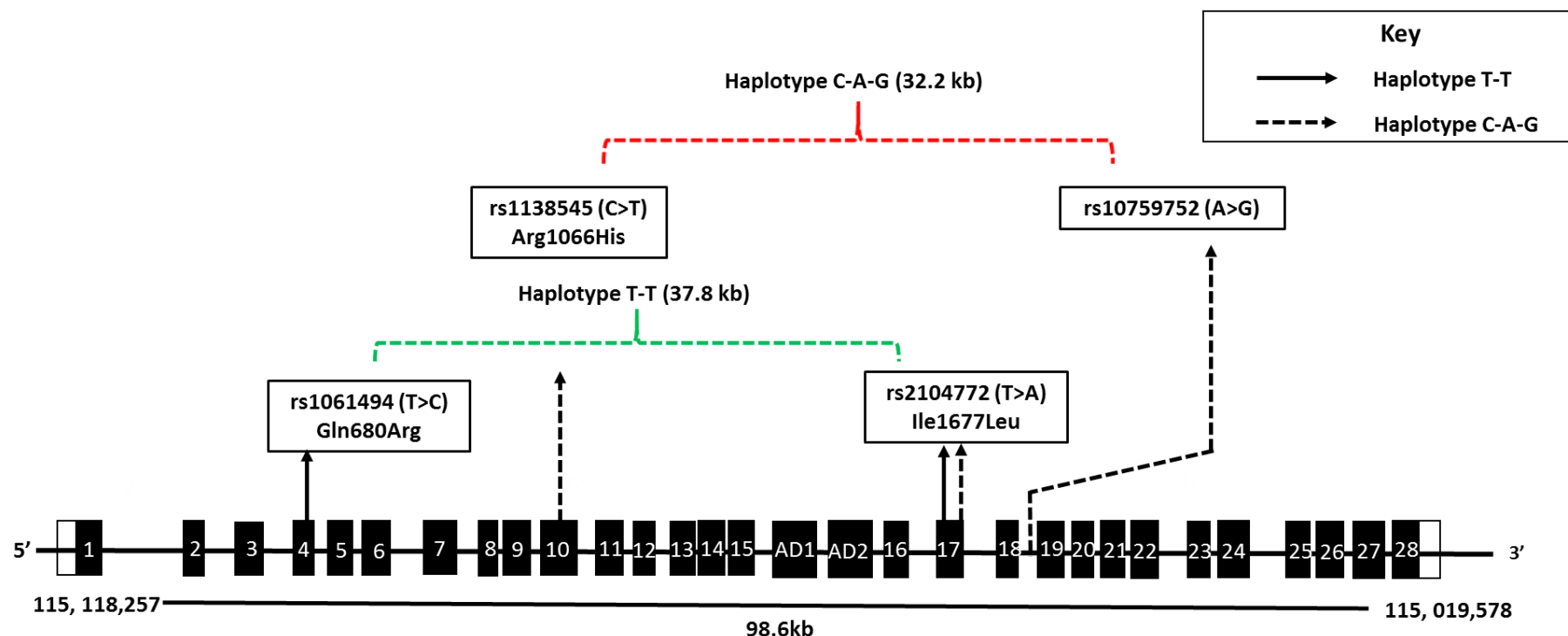


Figure 3.17: Inferred risk-associated haplotypes mapping to the *TNC* genomic interval

Diagram demonstrates two overlapping inferred haplotypes associated with the risk of musculoskeletal soft tissue injuries. Solid arrows indicate variant constituting the T-T haplotype (green); rs1061494 (T>C) and rs2104772 (T>A). This haplotype was associated with decreased risk for both chronic Achilles tendinopathy and acute ACL ruptures in the current study. Broken arrows indicate variants constituting the C-A-G haplotype (red); rs1138545 (C>T), rs2104772 (T>A) and rs10759752 (A>G). This haplotype was previously associated with large recurrent defects of the rotator cuff after corrective surgery [215]. The genomic boundaries of *TNC* are provided according to the Human GRCh38.p10 genome release. The exons 1-28, AD1 and AD2 (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *TNC* are represented. Although the genomic coordinates are provided in the 5' to 3' direction, the *TNC* gene is orientated on the antisense strand. Arg: Arginine Gln: Glutamine; Hist: Histidine; Ile: Isoleucine; Leu: Leucine. Diagram not drawn to scale.

With reference to the rs1061494 variant, the CC genotype was predicted to associate with increased *TNC* expression in testis tissue. In the current study, the CC genotype was significantly overrepresented in the group of cases with chronic AT; inferring increased *TNC* expression in the affected group. This is in line with the previous histological observation of increased tenascin-C in diseased tendons [321]. However, many cell and tissue types demonstrate distinct gene expression profiles, so it is uncertain whether the effects of rs1061494, as observed in testis tissue translate to changes in expression of *TNC* in dense connective tissue.

Deviations from HWE were noted for the *TNC* rs1061494 variant, in addition to rs1138545 in the SAW-CON^{AT} group. The inclusion criteria for control participants in the current study was highly selective, which potentially resulted in a sample that was not a true representation of the general population but instead represented a subset of the population towards the divergent end of the injury phenotype spectrum. The rs1138545 variant, although not independently associated with any of the investigated musculoskeletal soft tissue phenotypes, has previously been implicated in the aetiology of degenerative RCT [214]. This missense variant is located one amino acid away from the bridging sequence between the neighbouring FNIII5 and FNIII1A repeat domains of the mature tenascin-C peptide. Furthermore, rs1138545 is tagged by the *TNC* rs7021589 variant, a predicted *cis*-pQTL with significant regulatory effects on tenascin-C protein plasma levels [322]. More recently, a novel three-variant *TNC* haplotype inclusive of rs1138545, C-A-G (rs1138545 C>T-rs2104772 T>A-rs10759752 A>G), was associated with large recurrent defects after rotator cuff surgery (Figure 3.17) [215]. The *TNC* rs1138545 variant, described by the authors as the most “influential” of the *TNC* variants investigated, was neither independently or in combination with other *TNC* variants associated with the risk of AT or ACL ruptures. These results reiterate locus heterogeneity amongst clinically distinct musculoskeletal soft tissue injuries, whereby, some modifiers of risk are common to the predisposing genetic signatures, while other motifs are unique. In addition, the latter variant (rs10759752) of the haplotype inferred by Kluger et al. (2017) [215] was not identified by WES in the current study due to its intronic location. A limitation of our sequencing approach was therefore the inability to identify potential modifiers of risk located in regions other than that of the exons, captured microRNA genes and UTRs.

Another noteworthy finding pertained to the gene encoding perlecan (*HSPG2*), a heparan sulfate proteoglycan. This gene, a novel candidate in the investigation into the genetic basis of musculoskeletal soft tissue injuries, demonstrated the greatest burden of variants meeting the allele frequency difference threshold of 30% in the WES dataset (n=14). Perlecan was first proposed as a plausible candidate for interrogation into the genetic basis of Achilles tendinopathy by Saunders et al. (2016) using the *in silico* BORG semantic database [302]. However, to our knowledge, no variants in this gene have yet to be interrogated.

Interestingly, the function of tenascin-C is dependent on its incorporation into the ECM, which is suggested to be influenced by the presence of heparan sulfate proteoglycans including perlecan. A heparan knockout cell line demonstrated no tenascin-C incorporated into the ECM and the cells failed to assemble a fibronectin scaffold [323,324]. Furthermore, fibronectin-bound perlecan is known to co-localise and interact with tenascin-C, with the binding sites of perlecan situated within the constitutively expressed FNIII3-5 domains of tenascin-C [324]. Therefore, it may be hypothesised that variants located within the sequence motifs of these two proteins may influence their binding efficacy; altering the incorporation of the tenascin-C glycoprotein into the ECM and consequently effect matrix remodelling. This hypothesis does however warrant further investigation.

Changes in the pattern of tenascin-C expression have been documented in pathological tendons [321], a significant observation owing to tenascin-C strong ability to modulate the proliferation, migration and adhesion of cells and other extracellular matrix components [223]. Besides alterations in gene and protein expression by quantitative trait loci, missense variants within the sequence encoding the FNIII repeat domains of the mature tenascin-C peptide may modify the binding of the molecule to its preferential substrates including cell surface receptors, glycosaminoglycans and other ECM molecules. This may result in a shift in tissue modulation when *TNC* is transiently expressed during physiological processes involving healing, remodelling and matrix organisation, consequently, altering the tissue's capacity to recover in response to mechanical loading.

Mapping approximately 800 kb upstream from *TNC* is the *COL27A1* gene. Saunders et al. (2013) presented data implicating *COL27A1* in the risk signature predisposing for AT specifically the rs946053 (T>G) variant in combination with the two downstream *TNC* variants, rs13321 (G>C) and rs2104772 (T>A) [198]. Although atypical, the *COL27A1* gene encodes the α_1 chain of the newest member of the fibrillar collagen family, type XXVII collagen respectively. Characteristic of this family of proteins, the primary function of type XXVII collagen involves providing structural support and tensile strength to the ECM of cartilaginous structures [239].

Analysis of the WES data identified five non-synonymous variants spanning 1.6 kb within exon 3 of the *COL27A1* gene. Four of these variants met the allele frequency difference threshold and were observed to be in high LD. Tagged loci included rs2567707 (G>A), rs2567706 (A>G), rs2808770 (T>C) and rs2808771 (C>T). Demonstrating the strongest pairwise linkage, rs2567706 was selected as the proxy variant. However, analysis showed no association between the rs2567706 variant and the risk of either chronic AT or acute ACL ruptures in any of the investigated cohorts. A novel finding was however the significant association of the *COL27A1* rs2241671 variant with injury risk. Specifically, the GG genotype conferred a 1.9-fold increased risk of chronic AT in the UK-Achilles tendon injury cohort. Furthermore, following this *a priori* hypothesis, the GG genotype was also associated with a 2.4-fold increased risk of acute ATR when the genotype data from the SAW- and UK-Achilles tendon injury cohorts was pooled. However, due to the small number of participants in the combined-RUP group, this result should be interpreted with caution. Moreover, the GG genotype inferred a 2.3-fold increased risk for non-contact ACL ruptures in the SAC-ACL rupture cohort. Haplotype analysis further implicated the *COL27A1* interval with injury risk modulation through the association of the A-G-A (rs2567706-rs2241671-rs2567705) inferred haplotype with increased risk for non-contact ACL ruptures in the SAC-ACL rupture cohort while the alternate haplotype, G-A-T (rs2567706-rs2241671-rs2567705) was significantly associated with decreased risk.

These results support the idea that, although tendons and ligaments differ functionally, several genetic signatures may be common to the susceptibility profiles predisposing for

common musculoskeletal soft tissue injuries known to affect these structures. Several variants and allele combinations were associated with the risk of both chronic AT and ACL ruptures in this current study including the *COL27A1* rs2241671 variant and the *TNC* T-T (rs1061494-rs2104772) inferred haplotype. In saying this, it is also likely that certain genetic signatures underpinning risk are unique, such like that associated with AT and ACL ruptures in the current study compared to the three-variant *TNC* haplotype associated with large recurrent defects of the rotor cuff [215].

Lastly, several important considerations must be noted. Firstly, the representative admixed cohort consisting of self-identified South Africans Coloured individuals demonstrated similar risk-associated features to the self-identified White South African and British cohorts. However, no allele frequency data for this genetically diverse population currently exists; preventing the comparison between populations. However, LD analysis did not highlight any significant differences in the correlation of the alleles at the investigated *TNC* and *COL27A1* variants between the self-identified White and self-identified Coloured cohorts. Secondly, several inferred haplotypes constructed from genotype data in the SAW-and UK-Achilles tendon injury cohorts were unique to the respective population groups. The reason for this is not yet understood and the haplotypes that were found to be associated with the risk of injury require validation in larger injury cohorts as well as functional exploration *in vitro*. This study, together with some already published data provides evidence for the exploration of these variants *in vitro* with the aim of elucidating the biological mechanisms underpinning the observed *TNC* risk-associations for musculoskeletal soft tissue injuries. Although the results presented in this chapter provide support for further interrogation of these loci at the functional level, it is likely that important modifiers of risk within the gene or its surrounds still require elucidation.

3.4. CONCLUSION

Through the application of WES and a customised, tiered filtering strategy, several candidate variants within the targeted *TNC* and *COL27A1* loci were identified, prioritised and subsequently found to associate with the risk of chronic AT and/or acute ACL ruptures. These novel findings support the growing body of evidence supporting the involvement of the *TNC* gene and its encoded protein in the biological mechanisms underpinning injury development. In comparison to the *TNC* gene, the *COL27A1* gene is less well characterised in the profile of musculoskeletal soft tissue injuries. However, these results provide evidence for its potential contribution to an altered injury risk profile. This is the first study to have interrogated the contribution of genetic variants within the *TNC* and *COL27A1* genes to the genetic basis of ACL ruptures in both a self-identified White cohort and an ancestrally admixed Coloured cohort. Likewise, this was the first study to interrogate these specific loci in a self-identified white British cohort. The next step would be to explore the functional significance of these sequence signatures; a process that would hopefully help characterise the biological mechanisms underpinning the observed risk associations.

CHAPTER 4: INVESTIGATION OF VARIANTS WITHIN THE *MMP3* GENE AND MUSCULOSKELETAL SOFT TISSUE INJURIES

This chapter has been published in a condensed format as:

Gibbon A, Hobbs H, van der Merwe W, Raleigh SM, Cook J, Handley CJ, et al. The *MMP3* gene in musculoskeletal soft tissue injury risk profiling: a study in two independent sample groups. *J Sports Sci.* 2017;35(7):655-62.

4.1. STUDY AIM AND OBJECTIVES

DNA sequence variants within the *MMP3* gene have been suggested as markers of susceptibility for complex, multifactorial phenotypes including ischemic strokes [305], lumbar disc herniation [304] and coronary heart disease [246,325,326]. Furthermore, this gene has been implicated in the predisposition of chronic AT and acute Achilles tendon and ACL ruptures [203,204,206,207].

Based on this previous research implicating *MMP3* in the predisposition of musculoskeletal soft tissue injuries (Section 1.5.1), this gene was prioritised for further interrogation. Analysis of the WES dataset using the variant filtering strategy described in chapter 2 did not identify any new candidate variants for investigation. Instead, three *MMP3* variants: rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A), previously associated with chronic AT in a cohort of self-identified White ancestry [206], were prioritised for interrogation in two additional injury cohorts. In addition, the rs3025058 (5A>6A) promoter variant previously associated with ACL ruptures in a cohort sampled from a Thai population was also prioritised for further interrogation [204].

Therefore, the primary aim of this study was to further explore the association of the *MMP3* gene with the risk of musculoskeletal soft tissue injuries by genotyping four variants previously proposed as modifiers of risk: rs3025058 (5A/6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) in additional AT and ACL rupture cohorts.

Objectives specific to this study were:

- (i) To test the association between the *MMP3* candidate variants: rs3025058, rs679620, rs591058 and rs650108 and the risk of chronic Achilles tendinopathy and ACL ruptures in cohorts of self-identified White ancestry (AUS-and combined (SAW and AUS)-Achilles tendon injury cohorts and SAW-ACL rupture cohort)
- (ii) To test the association between the inferred haplotypes constructed from the prioritised *MMP3* candidate variants: rs3025058, rs679620, rs591058 and rs650108 and the risk of chronic Achilles tendinopathy and ACL ruptures in cohorts of self-identified White ancestry (AUS-and combined (SAW and AUS)-Achilles tendon injuries cohorts and SAW-ACL rupture cohort).

By addressing these objectives, this study aimed to refine the mapping of risk-associated variants for musculoskeletal soft tissue injuries in the *MMP3* gene. Figure 4.1 provides an overview of the loci prioritised for further interrogation in this chapter.

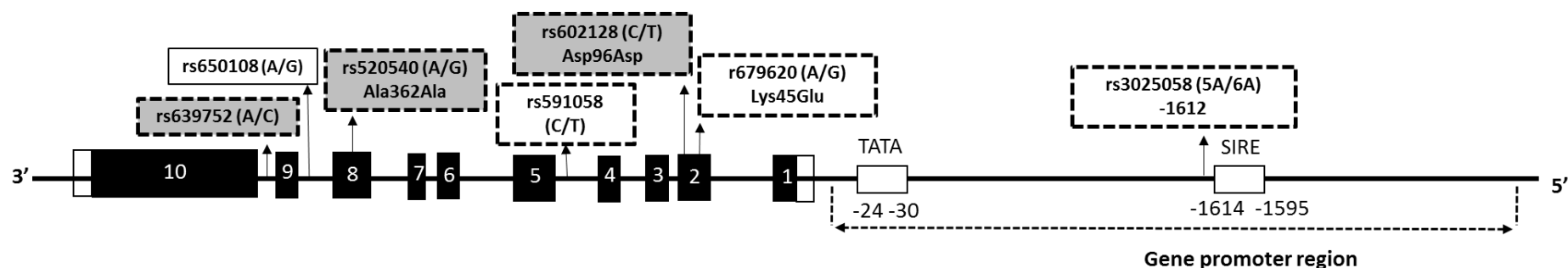


Figure 4.1: Map of *MMP3* gene displaying candidate variants identified by whole exome sequencing in relation to variants previously associated with the risk of musculoskeletal soft tissue injuries

Variants in white blocks have previously been associated with the risk of Achilles tendinopathy (rs679620, rs591058 and rs650108) [206], Achilles tendon ruptures (rs679620) [207] or ACL ruptures (rs3025058 and rs679620) [203,204] and were prioritised for interrogation in the current study. Variants in grey blocks were identified by whole exome sequencing and met the allele frequency difference threshold. However, due to the observed linkage across the *MMP3* genomic interval, none of the variants identified by whole exome sequencing were prioritised for investigation since they were tagged by variants previously associated with injury risk and prioritised for further interrogation in this study. The exons 1-10 (black bars), introns (horizontal lines), untranslated regions (empty bars) and the promoter region (dashed horizontal line) of *MMP3* are presented. Blocks with dashed outlines demonstrates tagged loci. The rs679620 variant was selected as the proxy variant. (-) Base pair number upstream of transcription start site. TATA: TATA box; SIRE: Stromelysin IL-1 responsive element; Lys: Lysine; Glu: Glutamic acid; Ala: Alanine; Asp: Aspartic acid. Diagram not drawn to scale.

4.2. RESULTS

4.2.1. Achilles tendinopathy

As outlined in the chapter 2, the participants within the AUS-Achilles tendon injury cohort were genotyped and the allele frequency distributions determined for the four prioritised *MMP3* variants: rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A). It must be noted that although the SAW-Achilles tendon injury cohort was included for comparative and combined analyses, the participants within this injury cohort were previously genotyped for the rs679620, rs591058 and rs650108 [206].

4.2.1.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Section 2.1.1.1A and Section 2.1.2.1). There were no significant genotype associations with sex, age, body mass, height and BMI in the AUS-Achilles tendon injury cohort. Similarly, no genotype associations with participant characteristics were noted for the previously genotyped SAW-Achilles tendon injury cohort (Table 4.1).

Table 4.1: Genotype associations with the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants and the physiological characteristics of participants in the Australian (AUS)- and South African White (SAW)-Achilles tendon injury cohorts

	<i>MMP3</i> variants						
	rs3025058	rs679620		rs591058		rs650108	
	AUS	AUS	SAW	AUS	SAW	AUS	SAW
Age	0.395	0.395	0.425	0.526	0.561	0.221	0.458
Sex	0.404	0.404	0.070	0.289	0.140	0.185	0.076
Mass	0.491	0.491	0.064	0.566	0.141	0.702	0.315
Height	0.728	0.728	0.120	0.768	0.324	0.266	0.142
BMI	0.806	0.806	0.449	0.862	0.496	0.645	0.793

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex. The rs3025058 variant was not genotyped in the SAW-Achilles tendon injury cohort in the previous study by Raleigh et al. (2009) [206].

4.2.1.2. Genotype and allele frequencies

It must be noted that the *MMP3* rs3025058 (5A>6A) promoter variant was not genotyped in the previous study by Raleigh et al. (2009) [206]; however, this variant is tagged by rs679620 (A>G) which allowed the inference of genotypes and the former locus in the SAW-Achilles tendon injury cohort. Furthermore, rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) demonstrated similar genotype and allele frequency distributions between the AUS- and SAW-Achilles tendon injury cohorts. Consequently, the genotype datasets from these cohorts were combined to increase the statistical power of the study. The genotype and allele frequency distributions are therefore presented for the independent injury cohorts as well as the combined-Achilles tendon injury cohort (Table 4.2).

After adjusting for age and sex, no significant differences in the genotype and allele frequencies were noted between the AUS-CON and AUS-TEN groups for all four investigated loci (Table 4.2). Similarly, the genotype and allele frequency distributions were not significantly different between the diagnostic groups in the combined-Achilles tendon injury cohort after adjusting for age, sex and country of birth. However, due to the association of the rs679620 variant with increased risk of AT in the SAW-Achilles tendon injury cohort [206], the rs3025058 variant was by inference also associated with injury risk ($P=0.034$). Specifically, the 6A6A genotype was significantly underrepresented in the SAW-CON^{AT} group (19%) compared to the SAW-TEN group (37%, $P=0.010$, OR: 2.5, 95% CI: 1.2-4.9).

4.2.1.3. Haplotype analysis

Haplotypes were inferred using the genotype data from the four investigated *MMP3* variants: rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A). Only three haplotypes (5A-A-T-G, 6A-G-C-A and 6A-G-C-G) were generated and inferred with a frequency greater than 10% in the AUS-, SAW- and combined-Achilles tendon injury cohorts.

The frequency distributions of two of the three inferred haplotypes, 5A-A-T-G and 6A-G-C-A, were not significantly different between AUS-CON and AUS-TEN groups (Figure 4.2). The distribution of the 6A-G-C-G haplotype was; however, significantly different between the diagnostic groups with an overrepresentation in the AUS-CON group (29%, $n=45$) compared to the AUS-TEN group (20%, $n=13$, $P=0.047$, Haplo.score: -1.987). No significant differences in the frequency distribution of the three inferred haplotypes were highlighted between the diagnostic groups when the combined-Achilles tendon injury cohort was analysed.

Table 4.2: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the Australian (AUS)-, South African White (SAW) - and combined-Achilles tendon injury cohorts

	AUS		SAW [†]		P-values			
	CON	TEN	CON ^{AT}	TEN	Countries	Diagnostic Groups		
rs3025058	191	77	96	74		AUS	SAW	AUS+SAW
6A6A	30 (58)	22 (17)	20 (19)	38 (28)	0.983	0.270	0.034	0.734
6A5A	49 (94)	52 (40)	52 (53)	43 (32)				
5A5A	20 (39)	26 (20)	25 (24)	19 (14)				
5A allele	45 (172)	52 (80)	52 (101)	41 (60)	1.000	0.182	0.029	0.614
HWE	0.776	1.000	0.412	0.468				
rs679620	191	77	96	74				
GG	30 (58)	22 (17)	20 (19)	38 (28)	0.983	0.270	0.034	0.734
GA	49 (94)	52 (40)	52 (53)	43 (32)				
AA	21 (39)	26 (20)	25 (24)	19 (14)				
A allele	45 (172)	52 (80)	52 (101)	41 (60)	1.000	0.182	0.029	0.614
HWE	0.776	1.000	0.412	0.468				
rs591058	191	77	93	73				
CC	31 (60)	23 (18)	20 (19)	36 (26)	0.804	0.246	0.056	0.816
CT	49 (93)	51 (39)	55 (53)	44 (32)			(0.017)	
TT	20 (38)	26 (20)	26 (25)	19 (14)				
T allele	44 (169)	51 (79)	55 (103)	41 (60)	0.477	0.154	0.047	0.773
HWE	0.773	1.000	0.417	0.473				
rs650108	159	67	94	73				
GG	57 (90)	51 (34)	54 (51)	52 (38)	0.478	0.465	0.876	0.460
GA	35 (56)	40 (27)	44 (41)	38 (28)				
AA ^a	8 (13)	9 (6)	2 (2)	10 (7)				
A allele	26 (82)	29 (39)	24 (45)	29 (42)	0.806	0.497	0.362	0.174
HWE	0.305	0.775	0.086	0.574				

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between countries (AUS vs. SAW) and diagnostic groups (CON vs. TEN) adjusted for each other, sex and age. P-values in bold typeset indicate significance ($P < 0.05$).[†]The rs679620, rs591058 and rs650108 variants were previously genotyped in the SAW-Achilles tendon injury cohort and the results presented by Raleigh et al. (2009) [206]. P-values in parentheses are generated from conducting post-hoc analyses based on an *a priori* hypothesis that the genotype is associated with the risk of injury. The rs3025058 variant was not genotyped in the SAW-Achilles tendon injury cohort but the genotypes were inferred from the rs679620 variant.

^a rs650108 GG genotype vs. GA + AA genotype

SAW-CON^{AT} vs. SAW-TEN rs650108 AA genotype vs. GA + GG genotype: $P = 0.043$

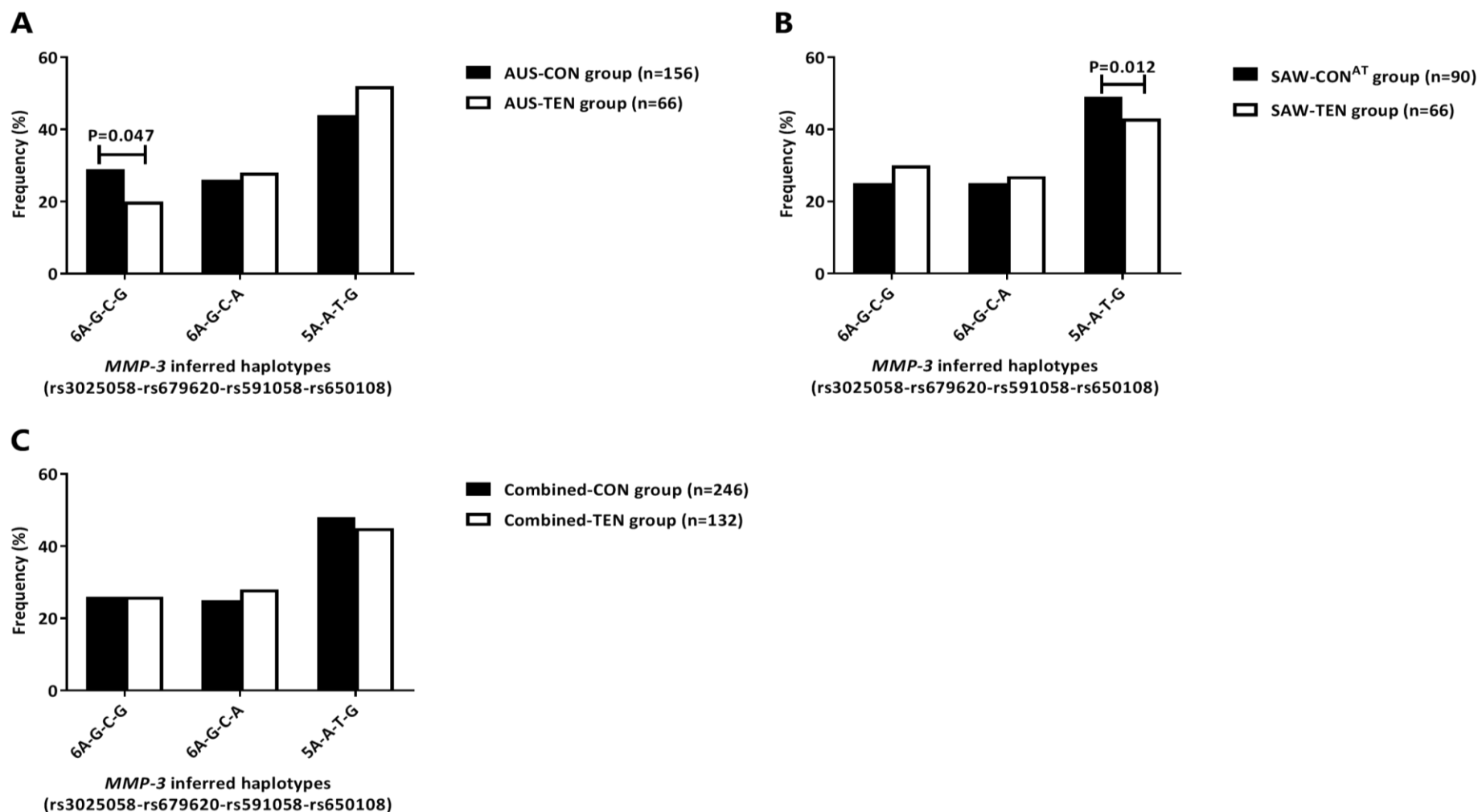


Figure 4.2: Inferred *MMP3* haplotypes in the Australian (AUS)-, South African White (SAW)- and combined-Achilles tendon injury cohorts

Frequency distribution of inferred *MMP3* haplotypes constructed from the rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the (A) AUS-, (B) SAW- and (C) combined-Achilles tendon injury cohorts. CON: Group of control participants; TEN: Group of participants with chronic Achilles tendinopathy. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and sex, in addition to country of birth when the combined-Achilles tendon injury cohort was analysed. The number of participants (n) in each group is in parentheses.

4.2.1.4. Linkage disequilibrium analysis

The LD between the genotyped variants in AUS- and SAW-Achilles tendon injury cohorts is presented in Figure 4.3. With respect to the AUS-Achilles tendon injury cohort, the genotyped rs3025058, rs679620, rs591058 and rs650108 variants were observed to be in high LD. The 5A allele of rs3025058 (5A>6A) and the T allele of rs591058 (T>C) strongly correlated with the A allele of rs679620 (A>G) (r^2 : 1.000 and 0.963 respectively). In addition, rs3025058, rs679620 and rs591058 were also observed to be in high LD with the most distal variant, rs650108, but the alleles were only moderately correlated ($r^2 < 0.600$) (Figure 4.3A). The three variants previously genotyped in the SAW-Achilles tendon injury cohort: rs679620, rs501058 and rs650108 were observed to be in high LD ($D' \geq 0.999$), with the alleles of the rs679620 (A>G) and rs591058 (T>C) loci strongly correlated (r^2 : 0.983). Conversely, rs650108 was observed to be only moderately correlated with the rs591058 and rs650108 variants ($r^2 \geq 0.563$) (Figure 4.3B).

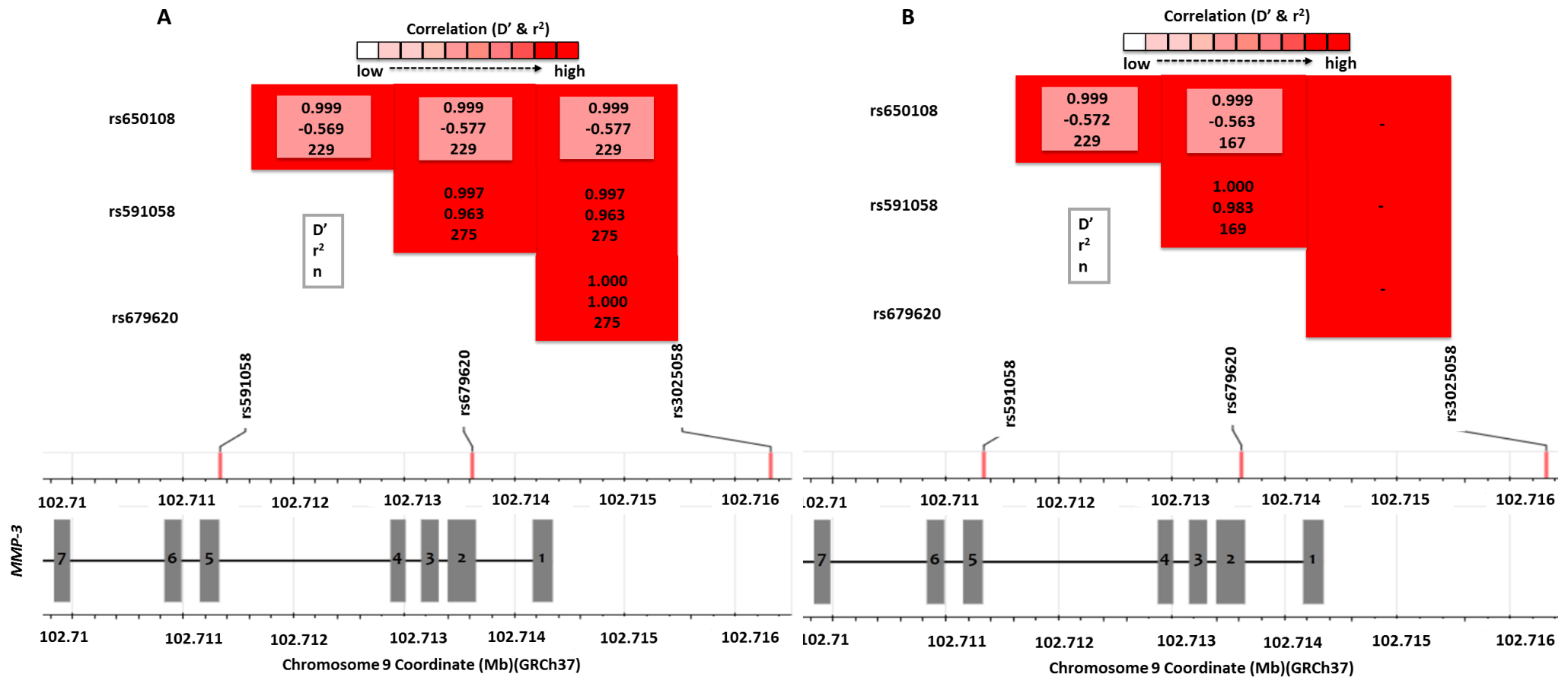


Figure 4.3: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *MMP3* variants in the Australian (AUS)- and South African White (SAW)-Achilles tendon injury cohorts

Results based on the (A) AUS-Achilles tendon injury cohort genotyped in the current study and (B) SAW-Achilles tendon injury cohort previously genotyped in the study by Raleigh et al. (2009) [205]. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Grey blocks represent exons and horizontal lines represent introns. (-) The rs3025058 variant was not genotyped in the SAW-Achilles tendon injury cohort in the previous study by Raleigh et al. (2009) [206].

4.2.2. Anterior cruciate ligament ruptures

Similar to the participants in the Achilles tendon injury cohorts, the participants within the SAW-ACL rupture cohort were genotyped and the allele frequency distributions determined for the four prioritised *MMP3* variants: rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A).

4.2.2.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Sections 2.1.1.1B). No genotype associations with age, sex, body mass, height and BMI were noted in the SAW-ACL rupture cohort (Table 4.3).

Table 4.3: Genotype associations with the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants and the physiological characteristics of participants in the South African White (SAW)-ACL rupture cohort

	<i>MMP3</i> variants			
	rs3025058	rs679620	rs591058	rs650108
Age	0.835	0.330	0.817	0.665
Sex	0.275	0.762	0.465	0.206
Mass	0.775	0.984	0.882	0.132
Height	0.934	0.988	0.978	0.314
BMI	0.315	0.953	0.770	0.185

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex.

4.2.2.2. Genotype and allele frequencies

After adjusting for age, sex and body mass, no significant differences in the distribution of genotype and allele frequencies were noted for any of the genotyped *MMP3* loci between the SAW-CON^{ACL} and SAW-ACL groups and similarly between the SAW-CON^{ACL} group and SAW-NON subgroup (Table 4.4). Similarly, no significant difference in genotype and allele frequencies were noted when male and female participants were analysed independently (Supplementary Tables IIIB.1 and IIIB.2).

4.2.2.3. Haplotype analysis

Haplotypes were inferred using the genotype data from the four investigated *MMP3* variants: rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A). Three haplotypes (5A-A-T-G, 6A-G-C-A and 6A-G-C-G) were generated and inferred with a frequency greater than 5% in the SAW-ACL rupture cohort.

No significant differences in the frequency distribution of the three inferred haplotypes were noted between the diagnostic groups of the SAW-ACL rupture cohort ($P=0.064$) (Figure 4.4A). Similarly, no significant differences were noted when male participants were evaluated separately ($P=0.395$) (Figure 4.4B).

Table 4.4: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of participants reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs3025058	226	200		126	
6A6A	27 (60)	20 (40)	0.177	19 (24)	0.178
6A5A	48 (108)	56 (113)		57 (72)	
5A5A	25 (57)	24 (47)		24 (30)	
6A allele	51 (228)	48 (193)	0.526	48 (120)	0.486
HWE	0.430	0.526		0.168	
rs679620[†]	225	200		126	
GG	27 (60)	20 (40)	0.256	19 (24)	0.217
GA	48 (107)	55 (110)		56 (71)	
AA	25 (58)	25 (50)		25 (31)	
G allele	50 (227)	48 (190)	0.431	47 (199)	0.459
HWE	0.357	0.600		0.227	
rs591058	225	200		126	
CC	27 (60)	20 (40)	0.211	19 (24)	0.199
CT	47 (106)	56 (111)		56 (71)	
TT	26 (59)	24 (49)		25 (31)	
C allele	50 (226)	48 (191)	0.515	47 (119)	0.494
HWE	0.292	0.513		0.227	
rs650108	224	197		125	
GG	54 (122)	63 (124)	0.092	65 (81)	0.070
GA	40 (90)	32 (63)		31 (39)	
AA ^a	5 (12)	5 (10)		4 (5)	
A allele	25 (114)	21 (83)	0.143	20 (49)	0.093
HWE	0.483	0.250		0.597	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for sex, body mass and age at the time of recruitment or injury. [†]Ninety-two SAW-CON^{ACL} and 129 SAW-ACL participants were previously genotyped for the rs679620 variant in the previous study by Posthumus et al. (2012) [203].

^a rs650108 GG genotype vs. GA + AA genotype

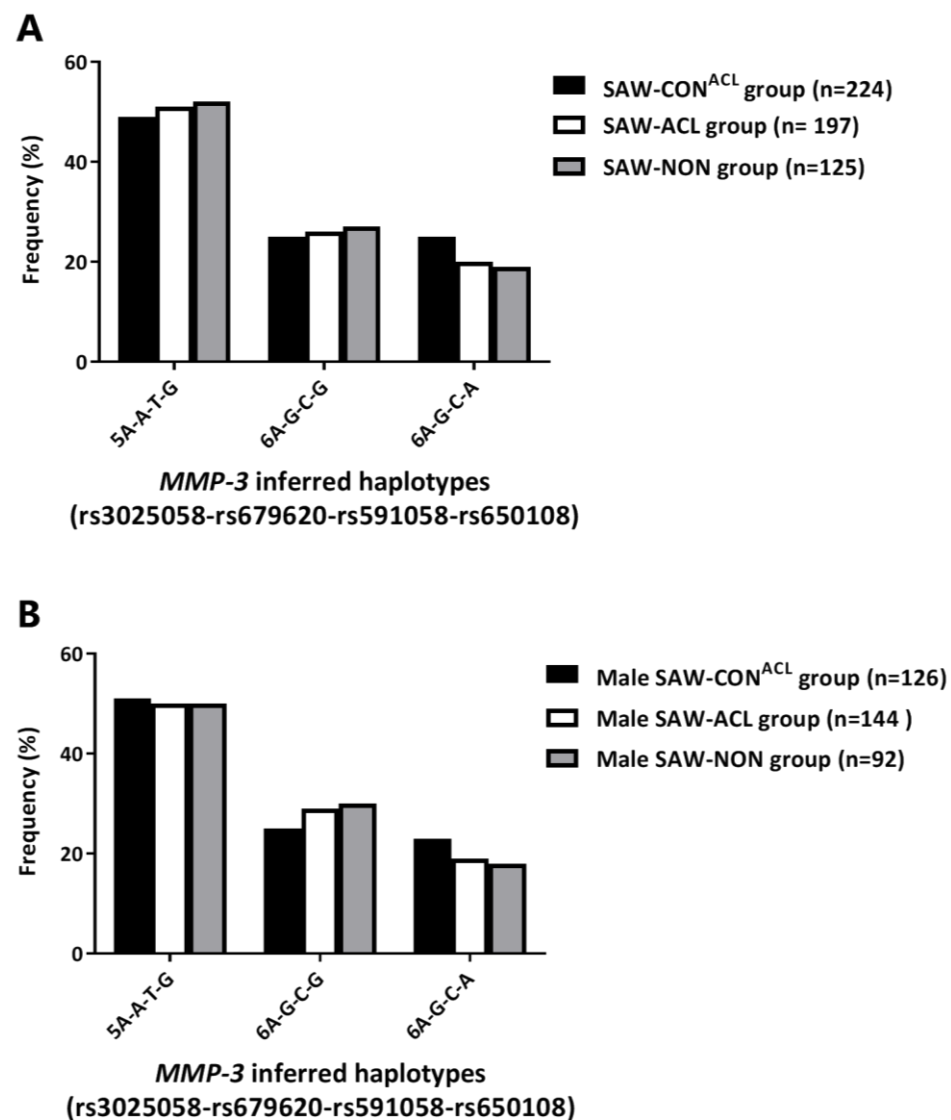


Figure 4.4: Inferred *MMP3* haplotypes in the South African White (SAW)-ACL rupture cohort

Frequency distribution of inferred *MMP3* haplotypes constructed from the rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the (A) SAW-ACL rupture cohort and (B) male subgroup. CON: Group of control participants; ACL: Group of participants with anterior cruciate ligament ruptures; NON: Group of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and body mass, in addition to sex when all participants were evaluated. The number of participants (n) in each group is in parentheses.

4.2.2.4. Linkage disequilibrium analysis

Linkage analysis conducted on the genotype data from the SAW-ACL rupture cohort demonstrated strong correlation between the 5A allele of rs3025058 (5A>6A) and T allele of rs591058 (T>C) with the A allele of rs679620 (A>G) (r^2 : 1.000 and r^2 : 0.963 respectively). Although these three variants were also in high LD with rs650108 ($D'=0.999$), the alleles were only modestly correlated ($r^2<0.600$) (Figure 4.5).

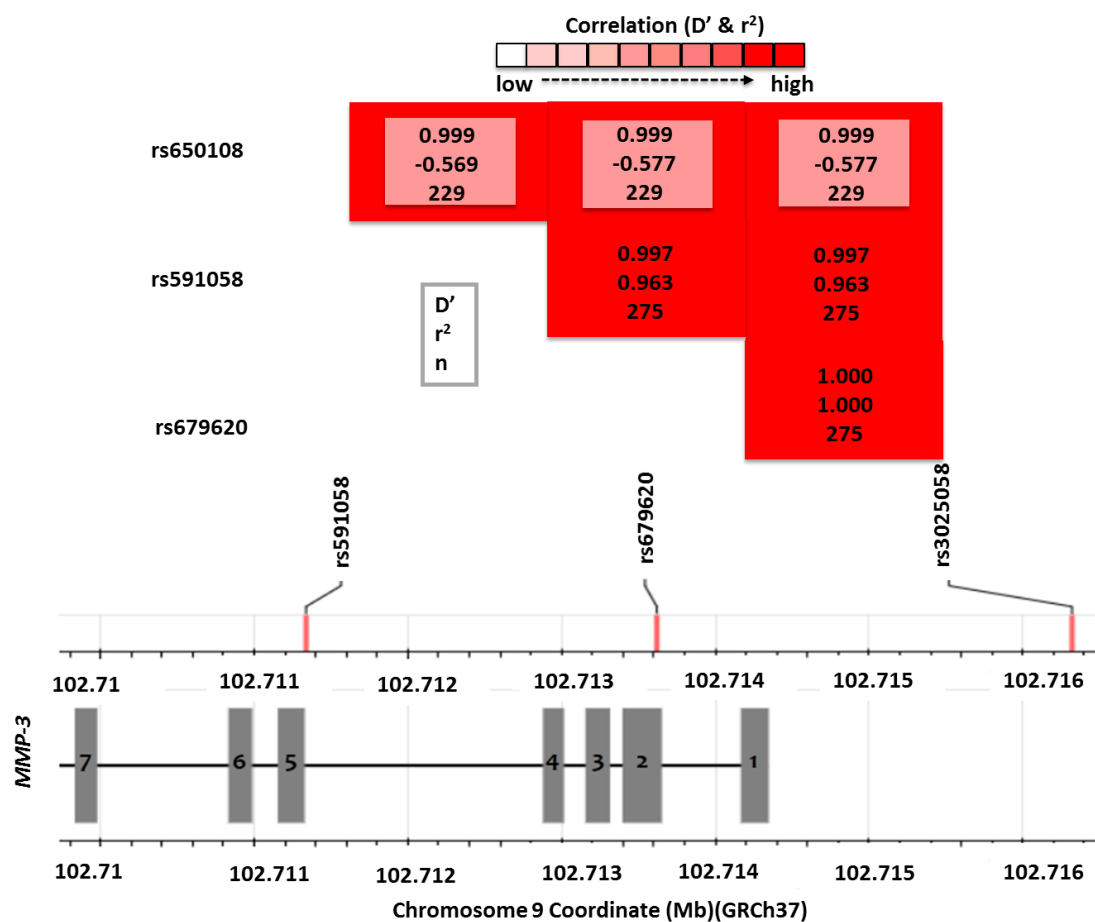


Figure 4.5: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *MMP3* variants in the South African White (SAW)-ACL rupture cohort

Results based on the SAW-ACL rupture cohort investigated in this study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Grey blocks represent exons and horizontal lines represent introns.

4.3. DISCUSSION

Raleigh and colleagues (2009) first proposed the *MMP3* gene as a risk locus for musculoskeletal soft tissue injuries by demonstrating the association of three single nucleotide variants, rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A), in addition to a three-variant haplotype, A-T-G (rs679620-rs591058-rs650108), with susceptibility to AT in a SAW cohort [206]. Furthermore, due to its linkage with the previously implicated rs679620 (A>G) variant, rs3025058 (5A>6A) was by inference associated with chronic AT in the SAW-Achilles tendon injury cohort. Specifically, the rs3025058 6A allele corresponded with the G allele at the rs679620 locus, the latter inferring a 2.5-fold increased risk for AT [206]. The rs3025058 variant has also previously been associated with the risk of ACL ruptures. However, it was the alternate 5A allele (5A5A+5A6A) that was associated with increased risk of contact ACL ruptures [204]. Contrary to these previous findings, none of these investigated *MMP3* variants demonstrated independent associations with the risk of AT in the AUS and combined cohorts or ACL ruptures in the SAW cohort.

These variants were postulated to be plausible candidates for interrogation into the genetic basis of musculoskeletal soft tissue injuries based on their proposed functionality, in addition to the fact that *MMP3* is known to contribute to the maintenance of tissue homeostasis. In particular, the distal promoter variant, rs3025058 (5A>6A), is known to significantly influence gene expression, with the 5A-allele promoter demonstrating a 2-fold increase in activity compared to the 6A-allele promoter [246]. In line with this, individuals homozygous for the 5A allele demonstrate the highest *MMP3* mRNA and protein levels compared to the intermediate levels detected in heterozygotes and the lowest levels in 6A homozygous individuals [327]. A possible mechanism by which this variant may influence gene expression is its possible interaction with the stromelysin interleukin-1 responsive element, which is known to overlap the run of five or six adenosines of rs3025058 [246]. This *cis* acting element has been characterised as a repressor of IL-1 induced stromelysin gene expression and that the binding of specific protein complexes to this sequence motif results in the repression of *MMP3* transcription. Specifically, nuclear factor kappa B (NFkB), a known repressor of gene transcription, demonstrates allele-specific preferences at the rs3025058 locus; binding to the

6A allele promoter more effectively [328]. Therefore, the influence of rs3025058 on DNA-protein interactions within the promoter of the gene may result in an altered *MMP3* expression profile.

As mentioned, this variant has previously been associated with the risk of ACL ruptures in a Thai population; however, we would caution against the over-interpretation of these results [204]. Firstly, this previous study reported the significant overrepresentation of the 5A allele in the group of participants with contact ACL ruptures compared to the group of participants with non-contact ACL ruptures. However, when compared to the healthy controls, the frequency distribution of the 5A allele in the ACL group was not significantly different. Conversely, the current study compared genotype and allele frequency differences between the healthy control group and group of participants with clinically diagnosed ACL ruptures. Moreover, the ACL group in the current study was only stratified for individuals reporting non-contact mechanisms of injury. Non-contact injuries are hypothesised to have the greater genetic contribution as they result from an individual's own movement, as opposed to contact injuries which results from an external force to the knee or body. It is plausible to assume that if a force experienced by the knee joint is large enough, regardless of an individual's genetic makeup, an injury may result; including ligament ruptures. Therefore, the failure of the current study to demonstrate a significant association with the risk of ACL injury, in addition to the only association being reported in the previous study being with that of contact ACL ruptures, one may propose that although this locus is functional, it may not be an informative marker for ACL injury risk.

With respect to chronic AT, the 6A6A genotype was, by inference, associated with increased risk in the previously genotyped SAW-Achilles tendon injury cohort [206]. With the knowledge that the 6A allele results in reduced gene promoter activity, the overrepresentation of the 6A6A genotype in the group of AT affected individuals compared to the healthy control group is in accordance with the histological observation of decreased *MMP3* mRNA expression in chronic, painful tendons [247,248]. However, this association was not observed in the AUS-Achilles tendon injury cohort. Therefore, this association warrants further investigation in additional cohorts.

Similarly, *MMP3* rs679620, rs591058 and rs650108 were previously associated with the risk of AT in the previously genotyped SAW-Achilles tendon injury cohort but not in the AUS-Achilles tendon injury cohort in the current study. However, the missense rs679620 variant (Lys45Glu) has been identified as a potential susceptibility locus for several other complex phenotypes including sporadic renal cell carcinoma [329], dental caries [330] and chronic periodontitis [331]. In addition, a study that aimed to functionally annotate all GWAS identified variants for Alzheimer's disease, reported rs679620 as one of two variants with the potential to influence *MMP3* protein phosphorylation [332]. Furthermore, rs679620 is known to tag rs591058, which explains the association of the latter non-coding variant within intron 4 with the risk of AT in the previous study by Raleigh et al. (2009) [206].

Besides the independent associations reported by Raleigh et al. (2009), the *MMP3* A-T-G (rs679620-rs591058-rs650108) inferred haplotype was also reported to associate with reduced risk for AT in the SAW-Achilles tendon injury cohort [206]. Conversely, the current study demonstrated the significant association of the alternate 6A-G-C-G inferred haplotype (rs3025058-rs679620-rs591058-rs650108) with reduced risk for AT in the AUS-Achilles tendon injury cohort. This phenomenon whereby alleles confer alternate effects to disease susceptibility in populations of independent ancestry has previously been documented but the exact biological reason for this still requires elucidation [333]. Owing to the non-deleterious nature of the investigated variants, it may be that these heritable differences have been maintained within the respective population structures. Alternatively, the inconsistency in risk-associated loci may suggest that the true risk conferring variants mapping to this gene or chromosomal region may still require identification and that these are in fact only statistical associations.

In saying this, the *MMP3* gene is still a plausible candidate for further interrogation due to the fact that the expression of this matrix metalloproteinase has been shown to be markedly different in pathological tissue specimens compared to that of healthy controls. Tissue from painful and ruptured Achilles tendons display reduced *MMP3* mRNA levels compared to healthy tendon samples [247–249]. Similarly, small tears of the rotator cuff have demonstrated reduced *MMP3* expression compared to controls. In contrast however, large or massive RCT have been associated with a significant increase in *MMP3* expression [334].

Moreover, Yoshihara et al. (2001) described the correlation between tear size and the level of MMP3 expression. Specifically, massive, full thickness RCT demonstrated significantly higher MMP3 levels in the synovial fluid sampled from the glenohumeral joints compared with that from isolated supraspinatus tears [335]. Furthermore, Leal et al. (2017) highlighted the relevance of epigenetic modification of *MMPs* and their inhibitors, tissue inhibitors of metalloproteinases (*TIMP*), in the development of acute tendon ruptures [336]. Therefore, these notable differences in the expression of *MMP3* and protein between healthy and diseased states provides support for the potential shift in control of *MMP3* gene regulation. In response to mechanical loading, the degradation of damaged collagen and worn out cells by proteolytic factors is essential in the maintenance of tissue homeostasis [134]. It may be hypothesised however, that if the balance between synthesis and degradation is distorted through alterations in gene activity and expression of key matrix regulators, the long-term outcome may be of disease.

Taken together, the inability of the current study to replicate the original findings; implicating *MMP3* in chronic AT and ACL injury, questions the biological significance of these investigated loci, particularly their contribution to musculoskeletal soft tissue injury predisposition. Furthermore, an important limitation was the inability of the sequencing platform to identify candidate variants for further interrogation within the introns and the promoter region of the gene. Therefore, other *MMP3* variants which are in fact ‘risk-conferring’ may have been overlooked.

4.4. CONCLUSION

To conclude, the customised WES analysis pipeline described in chapter 2 of this thesis was unable to detect any new exonic candidate variants within the *MMP3* gene, which were not tagged by the variants previously investigated by Raleigh et al. (2009) [206] and prioritised for further interrogation in this current study. Although the rs3025058 6A6A genotype was by inference associated with the risk of AT in the SAW-Achilles tendon injury cohort, in addition to the 6A-G-C-G (rs3025058-rs679620-rs591058-rs650108) inferred haplotype being associated with the risk of AT in the AUS-Achilles tendon injury cohort, we would caution

against the overinterpretation of these results. These results contrast with those reported by Raleigh et al. (2009) [206], which possibly suggests that other more informative markers of risk still require elucidation, and these results were simply statistical associations. However, several other *MMP* genes map to the 11q22 chromosomal region; supporting the hypothesis that other risk conferring loci within neighbouring genes may still require characterisation in the profile predisposing for musculoskeletal soft tissue injuries [203]. Therefore, a future study prospect would involve mining the WES dataset for new candidate variants located within the other *MMP* genes clustered on chromosome 11. These variants may be investigated for their potential contribution to musculoskeletal soft tissue injury predisposition, either independently or in combination with other proposed markers of risk.

CHAPTER 5: FUNCTIONAL *COL1A1* VARIANTS AS MODIFIERS OF RISK FOR ACUTE MUSCULOSKELETAL SOFT TISSUE INJURIES

5.1. STUDY AIM AND OBJECTIVES

Mutations within the *COL1A1* gene were previously shown to result in severe connective tissue disorders, such as Ehlers-Danlos Syndrome [261] and osteogenesis imperfecta [258]. In addition, variants within this gene have been associated with common multifactorial conditions including urinary stress incontinence [266], osteoarthritis [265,337] and osteoporotic fractures [264]. More recently, the two functional *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T), have been assessed for their potential contribution to the predisposition of common sporting injuries including shoulder dislocations and tennis elbow, in addition to Achilles tendon and ACL ruptures [182–187,268,271,272]. Interestingly, the evidence which has emerged from these studies is suggestive of mechanistic differences in the genetic profiles mapping to the *COL1A1* locus. Specifically, chronic injuries such as AT and tennis elbow demonstrate contrasting genetic patterns for the Sp1 transcription factor binding site variant (rs1800012 G>T, intron 1) compared to acute injuries such as ACL ruptures.

Therefore, based on this previous research (Section 1.5.1.), the *COL1A1* gene was prioritised for further interrogation in the current study with the aim of refining the genetic risk profile for chronic and acute musculoskeletal soft tissue injuries. Analysis of the WES dataset using the variant filtering strategy described in chapter 2 did not identify any new candidate variants for interrogation. Consequently, the *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T), were re-assessed for their potential contribution to the profiles predisposing for chronic musculoskeletal soft tissue injuries in two self-identified White cohorts for AT, in addition to acute musculoskeletal soft tissue injuries in self-identified White cohorts for ATR and ACL ruptures respectively. Furthermore, this study aimed to characterise these *COL1A1* variants in the profile predisposing for acute ACL ruptures in a cohort of self-identified Coloured ancestry. Based on the hypotheses that chronic and acute phenotypes

are undermined by different genetic profiles at this specific locus, the results are presented as discussed above, with each section reporting specific objectives.

Objectives specific to this study were:

Section 5.2.1: Chronic musculoskeletal soft tissue injuries-Achilles tendinopathy

- (i) To test the association between the *COL1A1* candidate variants, rs1107946 (G>T) and rs1800012 (G>T), and the risk of chronic Achilles tendinopathy in cohorts of self-identified White ancestry (UK-and combined (UK and SAW)-Achilles tendon injury cohorts)
- (ii) To test the association between the functional haplotypes constructed from the *COL1A1* candidate variants, rs1107946 (G>T) and rs1800012 (G>T), and the risk of chronic Achilles tendinopathy in cohorts of self-identified White ancestry (SAW-, UK- and combined (UK and SAW)-Achilles tendon injury cohorts)

Section 5.2.2: Acute musculoskeletal soft tissue injuries-Achilles tendon and ACL ruptures (Self-identified White cohorts)

- (i) To test the association between the *COL1A1* candidate variants, rs1107946 (G>T) and rs1800012 (G>T), and the risk of acute Achilles tendon and ACL ruptures in cohorts of self-identified White ancestry (combined (UK and SAW)-Achilles tendon injury cohort and SAW-ACL rupture cohort)
- (ii) To test the association between the functional haplotypes constructed from the *COL1A1* candidate variants, rs1107946 (G>T) and rs1800012 (G>T), and the risk of acute Achilles tendon and ACL ruptures in cohorts of self-identified White ancestry (combined (UK and SAW)-Achilles tendon injury cohort and SAW-ACL rupture cohort)

Section 5.2.3: Acute musculoskeletal soft tissue injuries in other populations- ACL ruptures (Self-identified Coloured cohort)

- (i) To test the association between the *COL1A1* candidate variants, rs1107946 (G>T) and rs1800012 (G>T), and the risk of ACL ruptures in a cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)

- (ii) To test the association between the functional haplotypes constructed from the *COL1A1* candidate variants, rs1107946 (G>T) and rs1800012 (G>T), and the risk of ACL ruptures in a cohort of self-identified Coloured ancestry (SAC-ACL rupture cohort)

By addressing these specific objectives, this study aimed to refine the mapping of risk-associated variants for chronic and acute musculoskeletal soft tissue injuries in the *COL1A1* gene in populations of White and Coloured ancestry. Figure 5.1 provides an overview of the loci prioritised for further interrogation in this chapter.

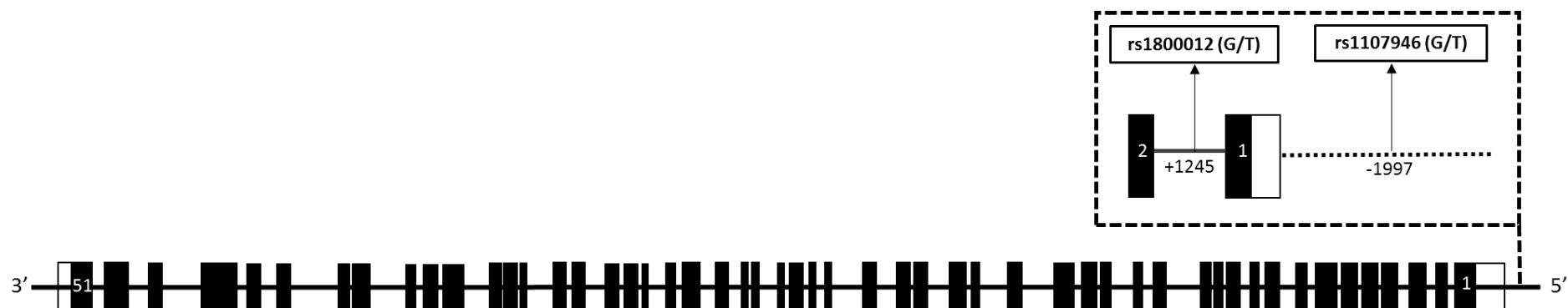


Figure 5.1: Map of *COL1A1* gene displaying variants previously associated with the risk of acute musculoskeletal soft tissue injuries and prioritised for further interrogation in the current study

The variants presented were prioritised for investigation in the current study based on their previous associations with the risk of acute musculoskeletal soft tissue injuries [182,184–187,272]. The exons 1-51 (black bars), introns (horizontal lines) untranslated regions (empty bars) and the promoter region (dashed horizontal line) of *COL1A1* are presented. The investigated variants were in high linkage disequilibrium. (-) Base pair number upstream of transcription start site. (+) Base pair number downstream of transcription start site. Diagram not drawn to scale.

5.2. RESULTS

5.2.1. Chronic musculoskeletal soft tissue injuries-Achilles tendinopathy

As outlined in the chapter 2, the participants with chronic AT within the SAW-and UK-Achilles tendon injury cohorts were genotyped, and the allele frequency distributions determined for the two prioritised *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T).

5.2.1.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Section 2.1.1.1A and 2.1.3.1). No genotype associations with sex, age, body mass, height and BMI were noted for the *COL1A1* variants in either the UK- or SAW-Achilles tendon injury cohorts (Table 5.1).

Table 5.1: Genotype associations with the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants and the physiological characteristics of participants in the British (UK)- and South African White (SAW)-Achilles tendon injury cohorts

	<i>COL1A1</i> variants			
	rs1107946		rs1800012	
	UK	SAW	UK	SAW
Age	0.320	0.859	0.563	0.574
Sex	0.877	0.796	0.131	0.547
Body mass	0.513	0.394	0.153	0.058
Height	0.789	0.082	0.057	0.161
BMI	0.345	0.054	0.080	0.294

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex.

5.2.1.2. Genotype and allele frequencies

The participants with clinically diagnosed chronic AT in the SAW- and UK-Achilles tendon injury cohorts demonstrated similar genotype and allele frequency distributions for the *COL1A1* variants investigated in the current study. Consequently, the genotype datasets from these groups were combined to increase the statistical power of the study. The genotype and allele frequency distributions are therefore presented for the independent injury cohorts as well as the combined-Achilles tendon injury cohort. The genotype and allele frequency distributions of the rs1107946 (G>T) and rs1800012 (G>T) variants did not differ significantly between the CON and TEN groups in the UK-, SAW- and combined-Achilles tendon injury cohorts (Tables 5.2).

Table 5.2: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) and Achilles tendinopathy (TEN) groups in the British (UK)-, South African White (SAW)- and combined-Achilles tendon injury cohorts

	UK		SAW		P-values			
	CON	TEN	CON ^{AT}	TEN	Countries	Diagnostic Groups		
rs1107946	117	84	131	71		UK	SAW	UK+SAW
GG	74 (86)	74 (62)	70 (91)	73 (52)	0.553	1.000	0.629	0.650
GT	25 (29)	24 (20)	24 (32)	27 (19)				
TT ^a	2 (2)	2 (2)	6 (8)	0 (0)				
T allele	14 (33)	14 (24)	18 (48)	13 (19)	0.392	1.000	0.232	0.380
HWE	1.000	0.668	0.078	0.344				
rs1800012[†]	98	73	143	94				
GG	75 (73)	73 (53)	70 (99)	70 (66)	0.532	0.861	0.886	1.000
GT	23 (23)	23 (17)	26 (37)	28 (26)				
TT ^b	2 (2)	4 (3)	5 (7)	2 (2)				
T allele	14 (27)	16 (23)	18 (51)	16 (30)	0.400	0.656	0.631	0.926
HWE	1.000	0.365	0.156	1.000				

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between countries (UK vs. SAW) and diagnostic groups (CON vs. TEN) adjusted for each other, sex and age. [†] The rs1800012 variant was previously genotyped in the SAW-CON^{AT} and SAW-TEN groups and the results presented in the study by Posthumus et al. (2009) [268].

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

UK-CON vs. UK-TEN rs1800012 TT genotype vs. GT + GG genotype: P=0.650

SAW-CON^{AT} vs. SAW-TEN rs1800012 TT genotype vs. GT + GG genotype: P=0.325

Combined-CON vs. Combined-TEN rs1800012 TT genotype vs. GT + GG genotype: P=0.787

5.2.1.3. Haplotype analysis

Inferred haplotypes were constructed using the genotype data from the two *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T). Three allele combinations (G-G, T-G and G-T) were common to the UK- and SAW-Achilles tendon injury cohorts and inferred with a frequency greater than 10%. None of the participants presented with the rare T-T inferred haplotype.

There were no significant differences in the distribution of the three inferred *COL1A1* haplotypes between the CON and TEN groups in the UK-($P=0.961$), SAW-($P=0.469$) and combined-Achilles tendon injury ($P=0.816$) cohorts (Figure 5.2A, 5.2B and 5.2C).

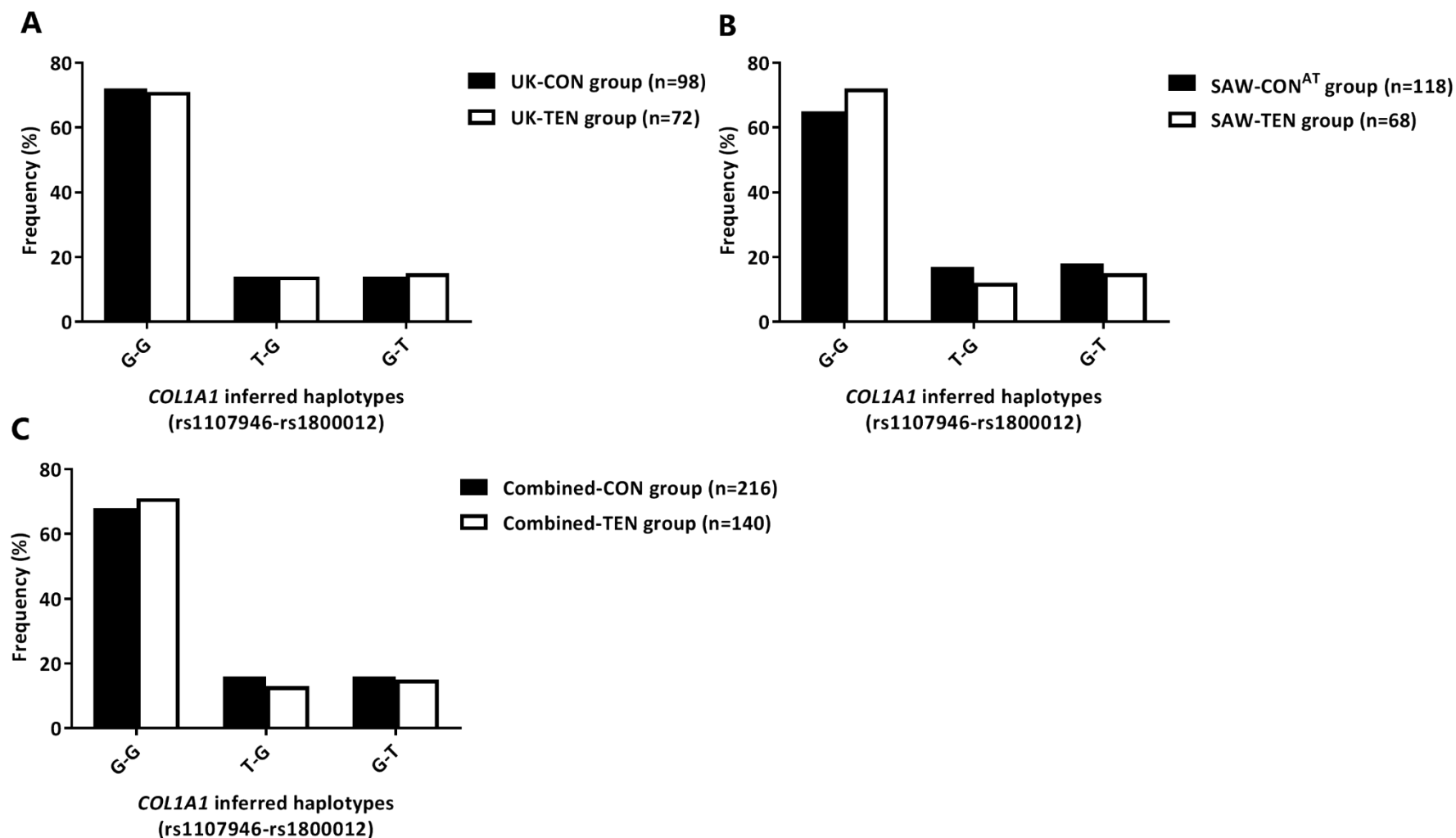


Figure 5.2: Inferred *COL1A1* haplotypes in the British (UK)-, South African White (SAW)- and combined-Achilles tendon injury cohorts

Frequency distribution of the inferred *COL1A1* haplotypes constructed from the rs1107946 (G>T) and rs1800012 (G>T) variants in the (A) UK-, (B) SAW and (C) combined-Achilles tendon injury cohorts. CON: Group of control participants; TEN: Group of participants with chronic Achilles tendinopathy. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and sex, in addition to country of birth when the combined cohort was evaluated. The number of participants (n) in each group is in parentheses.

5.2.1.4. Linkage disequilibrium analysis

The rs1107946 and rs1800012 variants were observed to be in high LD in both the UK- (D' : 0.997, $P=0.001$) and SAW-Achilles tendon injury (D' : 0.800, $P=0.008$) cohorts (Figure 5.3). Although these variants are located within the same haplotype block, their alleles are randomly associated ($r^2 < 0.200$).

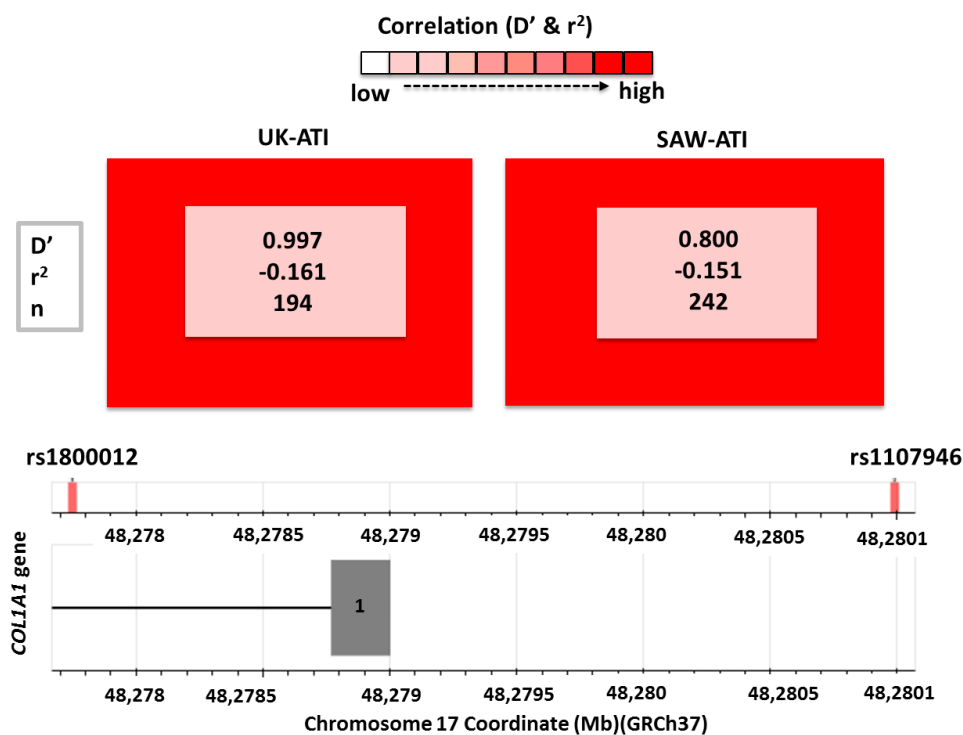


Figure 5.3: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *COL1A1* variants in the British (UK)- and South African White (SAW)-Achilles tendon injury cohorts

Results based on the UK- and SAW-Achilles tendon injury cohorts investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Grey blocks represent exons and horizontal lines represent introns.

5.2.2. Acute musculoskeletal soft tissue injuries-Achilles tendon and anterior cruciate ligament ruptures (Self-identified White cohorts)

As outlined in chapter 2, the participants with acute ATR in the UK- and SAW-Achilles tendon injury cohorts, in addition to the participants within the SAW-ACL rupture cohort were genotyped for the two *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T). The resultant allele and genotype frequencies are presented below.

5.2.2.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Sections 2.1.1.1A, 2.1.1.1B and 2.1.3.1). As previously presented (Table 5.1), no significant genotype associations were noted for participant characteristics in the UK- and SAW-Achilles tendon injury cohorts. Similarly, no significant genotype associations with participant characteristics were noted in the SAW-ACL rupture cohort (Table 5.3).

Table 5.3: Genotype associations with the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants and the physiological characteristics of participants in the South African White (SAW)-ACL rupture cohorts

	<i>COL1A1</i> variants	
	rs1107946	rs1800012
Age	0.402	0.473
Sex	0.512	0.553
Body mass	0.368	0.227
Height	0.277	0.712
BMI	0.138	0.326

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex.

5.2.2.2. Genotype and allele frequencies

The participants with acute ATR within the UK- and SAW-Achilles tendon injury cohorts demonstrated similar genotype and allele frequency distributions for the *COL1A1* variants investigated in this study. Consequently, the genotype datasets from these groups were combined to increase the statistical power of the study. It must be noted that the representative SAW- and UK- RUP groups were too small for independent analysis. Therefore, testing the association of the *COL1A1* variants with the risk of ATR was only conducted in the combined-Achilles tendon injury cohort after adjusting for age, sex and country of birth. The genotype and allele frequency distributions of the rs1107946 (G>T) and rs1800012 (G>T) variants did not differ significantly between the CON and RUP groups in the combined-Achilles tendon injury cohort (Tables 5.4).

With respect to ACL ruptures and after adjusting for age, sex and body mass, no significant differences in the distribution of genotype or alleles frequencies were noted between the SAW-CON^{ACL} and SAW-ACL groups (Table 5.5). In addition, the genotype and allele

frequencies were similarly distributed when male and female participants were analysed independently (Supplementary Tables IIIC.1 and IIIC.2).

Table 5.4: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) and Achilles tendon rupture (RUP) groups in the combined-Achilles tendon injury cohort

	UK		SAW		P-values	
	CON	RUP	CON ^{AT}	RUP	Countries	Diagnostic groups
rs1107946	117	23	131	38		UK+SAW
GG	74 (86)	74 (17)	70 (91)	71 (27)	0.255	1.000
GT	25 (29)	22 (5)	24 (32)	24 (9)		
TT ^a	2 (2)	4 (1)	6 (8)	5 (2)		
T allele	14 (33)	15 (7)	18 (48)	17 (13)	0.256	1.000
HWE	1.000	0.414	0.078	0.578		
rs1800012 [†]	98	18	143	43		
GG	75 (73)	94 (17)	70 (99)	77 (33)	0.212	0.106
GT	23 (23)	6 (1)	25 (37)	23 (10)		
TT ^b	2 (2)	0 (0)	5 (7)	0 (0)		
T allele	14 (27)	3 (1)	18 (51)	12 (10)	0.172	0.071
HWE	1.000	1.000	0.156	1.000		

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between countries (UK vs. SAW) and diagnostic groups (CON vs. RUP) adjusted for each other, sex and age. [†] The rs1800012 variant was previously genotyped in the SAW-CON^{AT} and SAW-RUP groups and the results presented in the study by Posthumus et al. (2009) [268].

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

Combined-CON vs. Combined-RUP rs1800012 TT genotype vs. GT + GG genotype: P=0.212

Table 5.5: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T), rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African White (SAW)-ACL cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs1107946	225	199		126	
GG	76 (172)	78 (155)	0.730	76 (96)	1.000
GT	23 (51)	22 (44)		24 (30)	
TT ^a	1 (2)	0 (0)		0 (0)	
T Allele	12 (55)	11 (44)	0.652	12 (30)	0.901
HWE	0.546	0.326		0.698	
rs1800012[†]	225	197		125	
GG	68 (154)	66 (130)	0.605	68 (85)	1.000
GT	29 (65)	33 (65)		30 (38)	
TT ^b	3 (6)	1 (2)		2 (2)	
T Allele	17 (77)	18 (69)	0.925	17 (42)	1.000
HWE	1.000	0.067		0.526	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for sex, body mass and age at the time of recruitment or injury. [†]A total of 130 SAW-CON^{ACL} and 117 SAW-ACL participants were previously genotyped for the rs1800012 variant and the results presented in the study by Posthumus et al. (2009) [185].

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

SAW-CON^{ACL} vs. SAW-ACL rs1800012 TT genotype vs. GT + GG genotype: P=0.293

SAW-CON^{ACL} vs. SAW-NON rs1800012 TT genotype vs. GT + GG genotype: P=0.717

5.2.2.3. Haplotype analysis

Inferred haplotypes were constructed using the genotype data from the two *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T). Three allele combinations (G-G, T-G and G-T) were inferred with a frequency greater than 10% in the combined-Achilles tendon injury and SAW-ACL rupture cohorts. None of the participants presented with the rare T-T inferred haplotype.

Two of the three *COL1A1* inferred haplotypes demonstrated similar frequency distributions between the combined-CON and combined-RUP groups. The G-T (rs1107946-rs1800012) allele combination was however significantly overrepresented in the combined-CON group (16%, n=35) compared to the combined-RUP group (8%, n=4, $P=0.022$, Haplo.score: -2.287) (Figure 5.4). Conversely, none of the *COL1A1* haplotypes demonstrated significant frequency differences between the SAW-CON^{ACL} and SAW-ACL groups ($P=0.749$) (Figure 5.5A) or between the SAW-CON^{ACL} group and SAW-NON subgroup ($P=0.929$). Similarly, no significant differences in the frequency distribution of the inferred haplotypes were noted when male participants in the SAW-ACL rupture cohort were analysed independently (ACL: $P=0.450$, NON: 0.423) (Figure 5.5B).

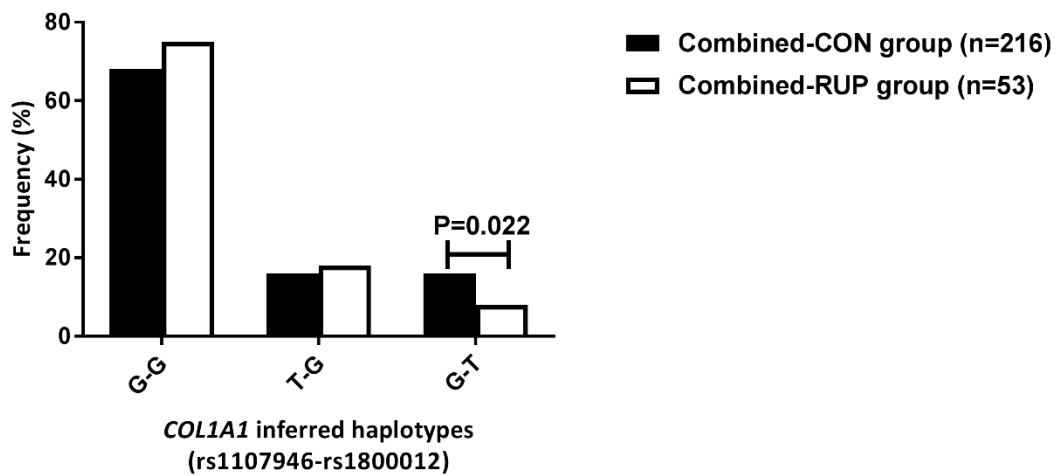


Figure 5.4: Inferred *COL1A1* haplotypes in the combined-Achilles tendon injury cohort

Frequency distribution of inferred *COL1A1* haplotypes constructed from the rs1107946 (G>T) and rs1800012 (G>T) variants in the combined-Achilles tendon injury cohort. CON: Group of control participants; RUP: Group of participants with acute Achilles tendon ruptures. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age, sex and country of birth. The number of participants (n) in each group is in parentheses.

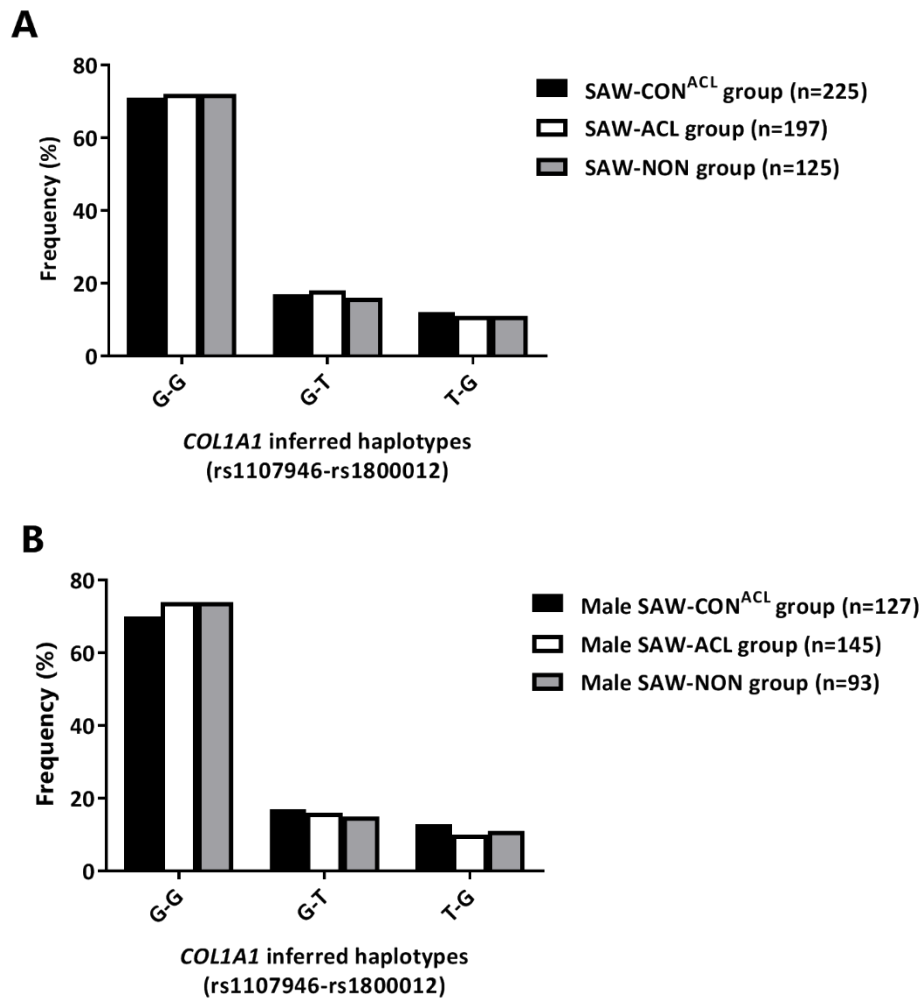


Figure 5.5: Inferred *COL1A1* haplotypes in the South African White (SAW)-ACL rupture cohort

Frequency distribution of inferred *COL1A1* haplotypes constructed from the rs1107946 (G>T) and rs1800012 (G>T) variants in the (A) SAW-ACL rupture cohort and (B) male subgroup. CON: Group of control participants; ACL: Group of participants with anterior cruciate ligament ruptures; NON: Group of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age and body mass, in addition to sex when the combined group was analysed. The number of participants (n) in each group is in parentheses.

5.2.2.4. Linkage disequilibrium analysis

The linkage relationship between the investigated *COL1A1* variants in the SAW-and UK-Achilles tendon injury cohorts was previously presented (Section 5.2.1.4). Therefore, this analysis informs only of the linkage between the rs1107946 (G>T) and rs1800012 (G>T) variants in the SAW-ACL rupture cohort. The rs1107946 and rs1800012 variants were observed to be in high LD ($D'=0.996$) but the alleles at this pair of loci were randomly associated ($r^2<0.200$) (Figure 5.6).

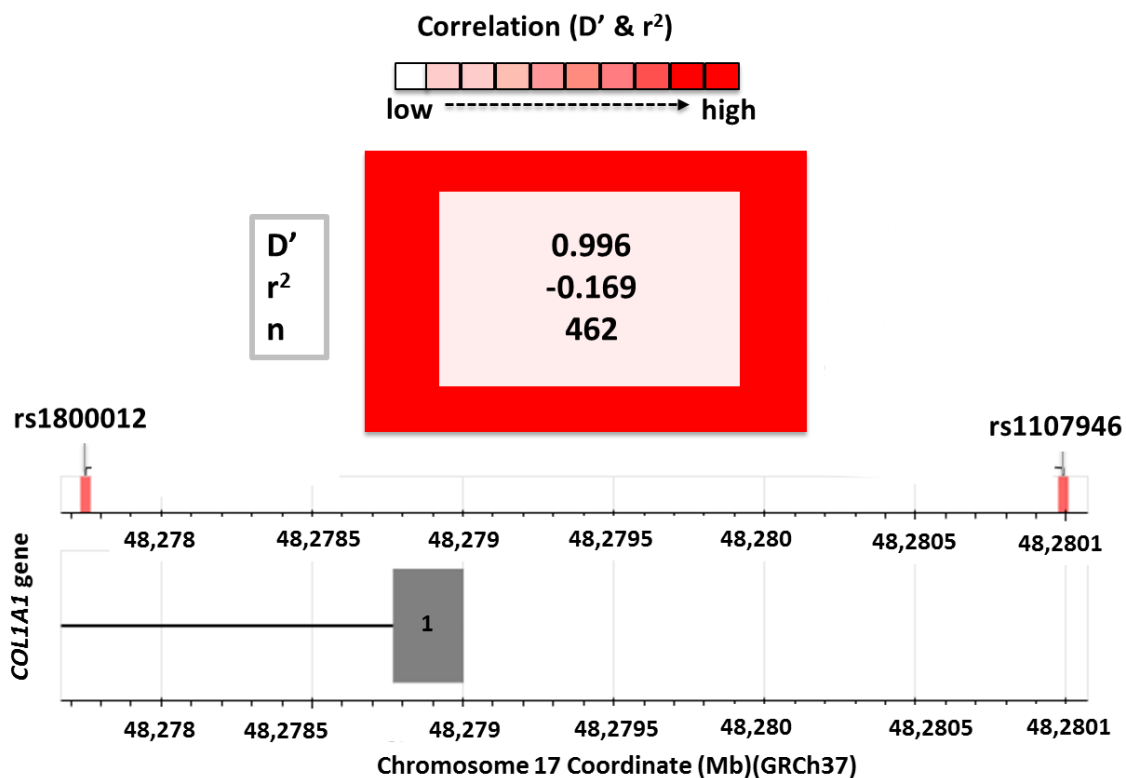


Figure 5.6: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *COL1A1* variants in the South African White (SAW)- ACL rupture cohort

Results based on the SAW-ACL rupture cohort investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Grey blocks represent exons and horizontal lines represent introns.

5.2.3. Acute musculoskeletal soft tissue injuries in other populations-Anterior cruciate ligament ruptures (Self-identified Coloured cohort)

As outlined in chapter 2, the participants within the SAC-ACL rupture cohort were genotyped for the two *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T). It must be noted that this cohort was sampled from a population with a unique genetic and social history and as a result, very few previously proposed markers of injury risk have been characterised in this population including the *COL1A1* variants [296,297]. Furthermore, the diverse genetic history of this population prevented the pooling of the genotyped data with the other previously recruited SAW-ACL rupture cohort. The resultant allele and genotype frequencies for the SAC-ACL rupture cohort are presented below.

5.2.3.1. Genotype associations with participant characteristics

General participant characteristics have been described in chapter 2 (Section 2.1.1.1C). Except for the effect of rs1800012 on BMI ($P < 0.001$), no other genotype associations with participant characteristics were observed in the SAC-ACL rupture cohort (Table 5.6). The BMI of participants with the rs1800012 GG genotype was significantly lower ($24.89 \pm 5.57 \text{ kg/m}^2$) compared to individuals with the GT genotype ($27.72 \pm 4.93 \text{ kg/m}^2$, $P < 0.001$) after adjusting for sex and age. Only one SAC-ACL participant demonstrated the TT genotype at this locus; however, this individual's body mass at the time of injury was not recorded.

Table 5.6: Genotype associations with the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants and the physiological characteristics of participants in the South African Coloured (SAC)-ACL rupture cohorts

	<i>COL1A1</i> variants	
	rs1107946	rs1800012
Age	0.963	0.551
Sex	0.063	0.652
Body mass	0.265	0.254
Height	0.331	0.563
BMI	0.652	<0.001

Values expressed as P-values. Associations with body mass and BMI are adjusted for age and sex. Associations with height are adjusted for sex. P-values in bold typeset indicate significance ($P < 0.05$).

5.2.3.2. Genotype and allele frequencies

After adjusting for age, no significant differences in the distribution of genotype or allele frequencies were noted between the SAC-CON and SAC-ACL groups and similarly between the SAC-CON group and SAC-NON subgroup for the rs1107946 and rs1800012 variants (Table 5.7). In addition, the genotype and allele frequencies were similarly distributed when male participants were analysed independently (Supplementary Table IIIC.3).

Table 5.7: Genotype and minor allele frequency distributions, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T), rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the South African Coloured (SAC)-ACL rupture cohort

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
rs1107946	97	89		49	
GG	55 (53)	57 (51)	0.768	47 (23)	0.388
GT	40 (39)	32 (28)		41 (20)	
TT ^a	5 (5)	11 (10)		12 (6)	
T Allele	25 (49)	27 (48)	0.733	33 (32)	0.213
HWE	1.000	0.040		0.744	
rs1800012	97	87		48	
GG	83 (80)	86 (75)	0.547	87 (42)	0.481
GT	18 (17)	13 (11)		13 (6)	
TT ^b	0 (0)	1 (1)		0 (0)	
T Allele	9 (17)	7 (13)	0.704	6 (6)	0.481
HWE	1.000	0.502		-	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for age at the time of recruitment or at the time of injury. P-values in bold typeset indicate significance (P<0.05).

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

SAC-CON^{ACL} vs. SAC-ACL rs1800012 TT genotype vs. GT + GG genotype: P=0.473

SAC-CON^{ACL} vs. SAC-NON rs1800012 TT genotype vs. GT + GG genotype: P=1.000

5.2.3.3. Haplotype analysis

Inferred haplotypes were constructed using the genotype data from the two *COL1A1* variants, rs1107946 (G>T) and rs1800012 (G>T). Three allele combinations (G-G, T-G and G-T) were inferred with a frequency greater than 10% in SAC-ACL rupture cohort. None of the participants presented with the rare T-T inferred haplotype.

Haplotype analysis conducted on the *COL1A1* genotype data in the SAC-ACL rupture cohort demonstrated no significant differences in the frequency distribution of two of the three inferred haplotypes, G-G and G-T (rs1107946-rs1800012), between the CON and ACL groups and similarly between the CON group and NON subgroup. However, the T-G haplotype was significantly underrepresented in the SAC-CON group (24%, n=23) compared to the SAC-NON subgroup (34%, n=16, $P=0.029$, Haplo.score: 2.182) (Figure 5.7A). No significant differences in the frequency distribution of all three allele combinations were noted when the male subgroup was analysed independently (ACL: $P=0.862$, NON: $P=0.350$) (Figure 5.7B).

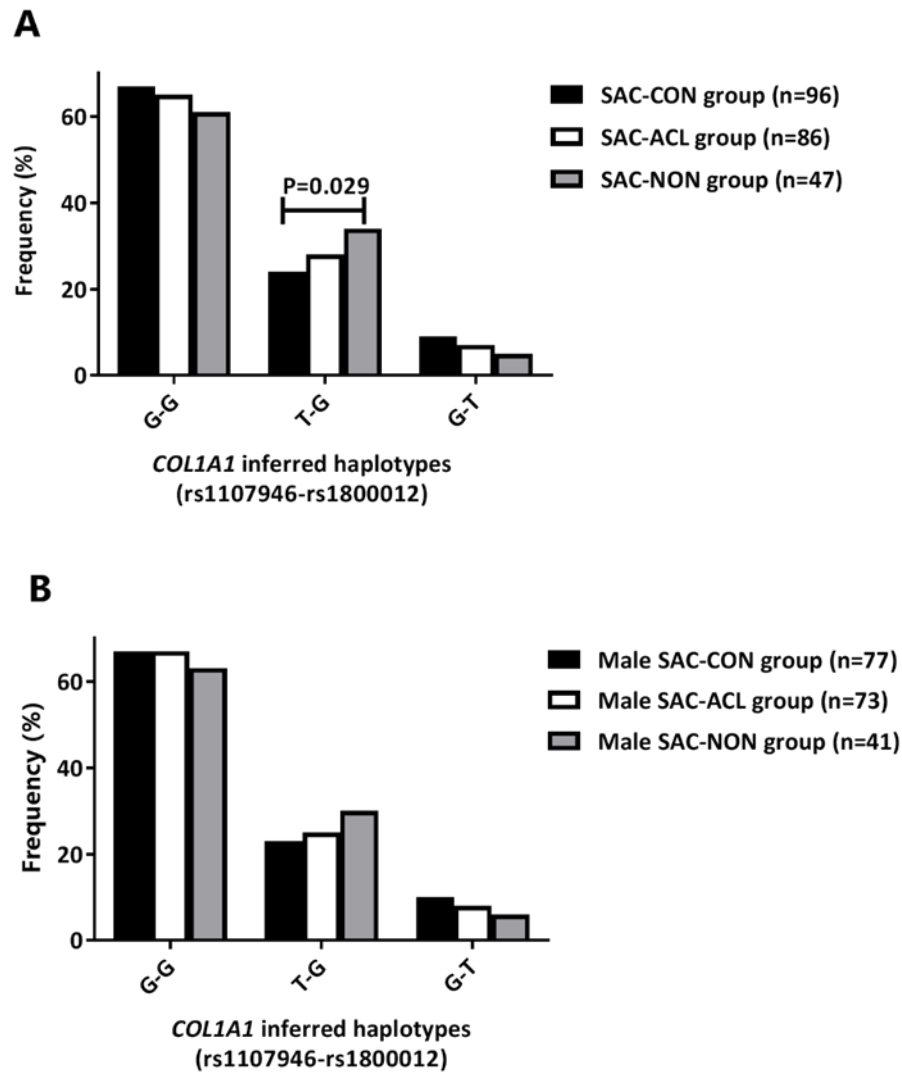


Figure 5.7: Inferred *COL1A1* haplotypes in the South African Coloured (SAC)-ACL rupture cohort

Frequency distribution of inferred *COL1A1* haplotypes constructed from the rs1107946 (G>T) and rs1800012 (G>T) variants in the (A) SAC-ACL rupture cohort and (B) male subgroup. CON: Group of control participants; ACL: Group of participants with anterior cruciate ligament ruptures; NON: Group of participants reporting a non-contact mechanism of injury. Significant differences in haplotype frequencies between groups are depicted on the graph, with P-values adjusted for age. The number of participants (n) in each group is in parentheses.

5.2.3.4. Linkage disequilibrium analysis

The *COL1A1* rs1107946 and rs1800012 variants demonstrated moderate linkage in the SAC-ACL rupture cohort (D' : 0.774) but the alleles showed little correlation (r^2 : 0.143).

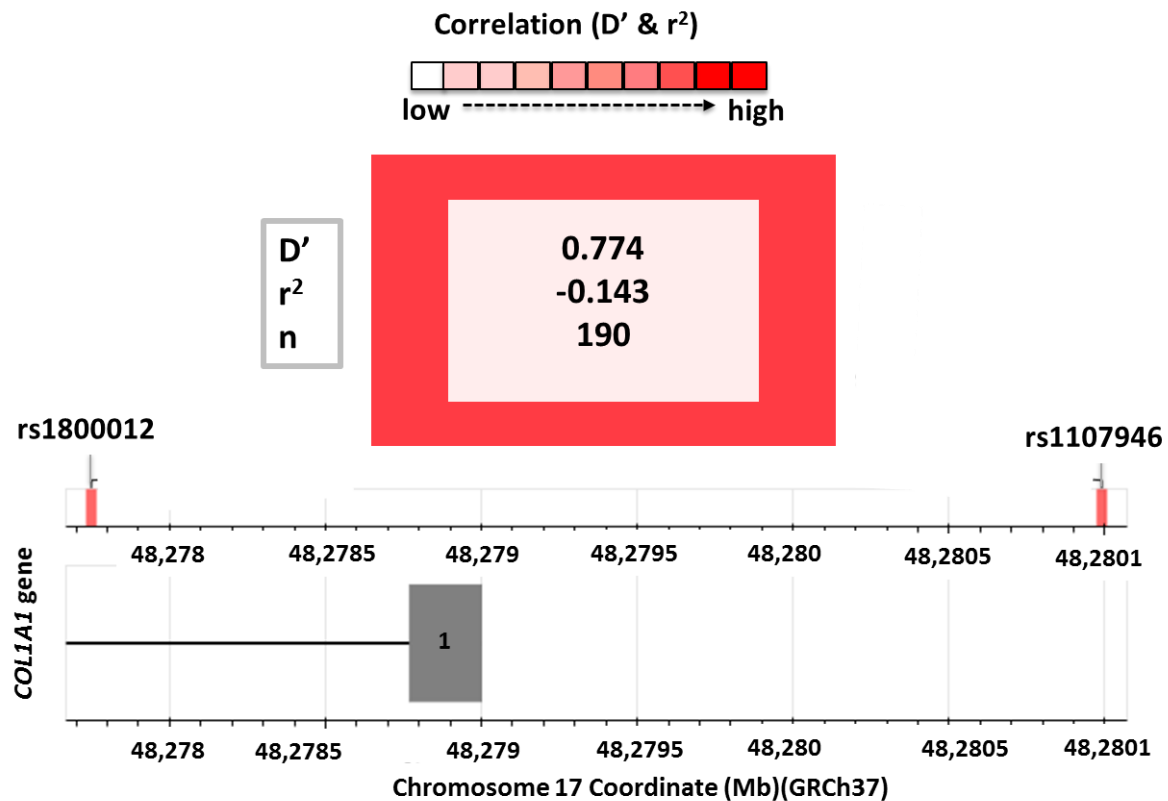


Figure 5.8: Heatmap showing linkage disequilibrium (LD) and correlation coefficient (r^2) estimates between the genotyped *COL1A1* variants in the South African Coloured (SAC)- ACL rupture cohort

Results based on the SAC-ACL rupture cohort investigated in the current study. Outer blocks reflect linkage disequilibrium (D') estimates and inner blocks reflect the correlation (r^2) of alleles between loci. Grey blocks represent exons and horizontal lines represent introns.

5.3. DISCUSSION

To date, several independent studies have published the results of genotyping either the *COL1A1* rs1107946 (G>T) promoter variant, the *COL1A1* rs1800012 (G>T) Sp1 transcription factor binding site variant or both in cohorts for the investigation of musculoskeletal soft tissue injuries [182–187,267,268,271–273,338]. These specific variants were prioritised as plausible candidates based on their proposed functionality and the fact that type I collagen is the predominant fibrillar collagen in non-cartilaginous connective tissues. The emerging evidence from these studies suggests that the genetic profile predisposing for chronic musculoskeletal soft tissue injuries mapping to the *COL1A1* locus may contrast with that underpinning acute musculoskeletal soft tissue injuries. Specifically, the rare rs1800012 TT genotype has repeatedly been associated with reduced risk of acute musculoskeletal soft tissue injuries, such as shoulder dislocations and ACL ruptures sustained during land-based sporting activities [184,185,187]. On the contrary, other studies have failed to demonstrate an association between this variant and the risk of chronic musculoskeletal soft tissue injuries [268,270]. Similarly, different mechanisms resulting in ACL injury have demonstrated contrasting associations involving the rs1107946 functional promoter variant. Specifically, this variant was not independently associated with the risk of ACL ruptures in a group of professional soccer players; all self-reported to be of Polish ancestry [183], yet the rs1107946 GG genotype was associated with increased risk for ACL ruptures sustained during skiing activities [182]. Furthermore, the rs1800012 GG genotype was associated with decreased risk of ACL ruptures in the same skiing cohort; in contrast to the proposed protective nature of the rs1800012 TT genotype in other studies [186]. This contrast in the genetic signature predisposing for ACL ruptures warranted further investigation, in addition to the fact that the rs1107946 variant had yet to be interrogated for its potential contribution to a chronic musculoskeletal soft tissue phenotype.

Therefore, based on this previous research, the current study aimed to further explore the association of the rs1107946 and rs1800012 variants in several injury cohorts representative of chronic (Achilles tendinopathy) and acute (ATR and ACL ruptures) musculoskeletal soft tissue injuries. Furthermore, this study aimed to determine whether the profile mapping to

the *COL1A1* locus, suggested to predispose for injury in cohorts of self-identified White ancestry was similar to that in a cohort sampled from a population of self-identified Coloured ancestry.

The rs1800012 variant is the most extensively investigated *COL1A1* locus due to its functional capacity to alter the recognition sequence of the Sp1 transcription factor within intron 1 of the gene and has been associated with several common, multifactorial phenotypes [264–267,273,337]. Conversely, this variant was not observed to associate with the risk of either chronic AT or tennis elbow in two independent studies [268,270]. In accordance with these previous observations, the current study failed to demonstrate the association of rs1800012 with the risk of chronic AT in the UK- and combined (SAW and UK)-Achilles tendon injury cohorts. The rs1107946 variant was also not associated with the risk of chronic AT in the three investigated cohorts (UK, SAW and combined). Furthermore, none of the inferred haplotypes constructed from these specific *COL1A1* variants were observed to associate with the risk of chronic AT.

With respect to the acute musculoskeletal soft tissue phenotypes investigated in the current study, neither of the investigated *COL1A1* variants were independently associated with the risk of ATR or ACL ruptures in the self-identified White cohorts. These results are similar to the results presented by Ficek et al. (2013) [183] but in contrast with those from previous studies, in which the rare rs1800012 TT genotype was independently associated with reduced risk for shoulder dislocations and ACL ruptures, in addition to ATR in a combined phenotype (shoulder dislocations, ACL ruptures and ATR) analysis [184,185,187]. However, an analysis which involved combining the data from the previously published Polish-ACL and Swedish-cruciate ligament (CL) rupture studies [183,184], together with the genotype data generated in the current study (SAW-ACL rupture cohort) demonstrated the significant association of the rare TT genotype (TT vs. GT + GG) with reduced risk for ACL ruptures ($P=0.014$, OR: 5.1, 95% CI: 1.3-45.0). Similar results were observed when all CL ruptures were analysed ($P<0.01$, OR: 6.2, 95% CI: 1.9-32.3) (Table 5.8). Conversely, the rs1107946 promoter variant was not associated with the risk of injury in the independent cohorts investigated in the current study

or in the combined cohort, which involved pooling the data form the current study and that from the previously published study by Ficek et al. (2013) [183] (Table 5.8).

Table 5.8: Comparative and combined analyses of the genotype and allele frequency distributions of the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants from the South African White-ACL rupture (SAW) cohort and the previously published Polish-ACL and Swedish-cruciate ligament (CL) rupture cohorts

	SAW cohort		Polish cohort		Swedish cohort		Combined cohort			P-values	
	CON	ACL	CON	ACL	CON	CL	CON	ACL	CL		
rs1107946	225	199	143	91			368	290		ACL	
GG	76 (172)	78 (155)	75 (107)	66 (60)			76 (279)	74 (215)		0.650	
GT	23 (51)	22 (44)	23 (33)	33 (30)		-	23 (84)	26 (74)	-		
TT ^a	1 (2)	0 (0)	2 (3)	1 (1)			1 (5)	<1 (1)			
T allele	12 (55)	11 (44)	14 (39)	18 (32)			13 (94)	13 (76)		0.864	
HWE	0.546	0.326	0.726	0.288							
rs1800012	225	197	143	91	325	233	693	288	521	ACL	CL
GG	68 (154)	66 (130)	67 (96)	71 (65)	71 (230)	70 (162)	69 (480)	68 (195)	69 (357)	0.650	0.802
GT	29 (65)	33 (65)	29 (41)	29 (26)	26 (83)	30 (70)	27 (189)	32 (91)	31 (161)		
TT ^{b, c}	3 (6)	1 (2)	4 (6)	0 (0)	4 (12)	<1 (1)	3 (24)	<1 (2)	<1 (3)		
T allele	17 (77)	18 (69)	19 (53)	14 (26)	16 (107)	15 (72)	17 (237)	16 (95)	16 (167)	0.790	0.501
HWE	1.000	0.067	0.578	0.201	-	-	-	-	-		

The previously published Swedish study by Khoschnau et al. (2008) [184] did not include genotyped data for the rs1107946 variant. Therefore, only the genotyped data from the SAW- (current) and Polish-ACL rupture (Ficek et al. 2013) [183] cohorts were combined and analysed for this variant. Furthermore, the original study by Khoschnau et al. (2008) [184] included participants with both anterior and posterior cruciate ligament ruptures. However, the number of ACL ruptures within the affected group was not indicated. Therefore, analysis of the rs1800012 variant in the combined cohort included control participants from all three studies (SAW, Polish and Swedish) but the ACL group only included genotype data from the SAW- and Polish-ACL rupture cohorts. The CL group; however, included data from the affected individuals from all three cohorts.

^ars1107946 CON vs. ACL: GG genotype vs. GT + TT genotype

^brs1800012 CON vs. ACL: GG genotype vs. GT + TT genotype

^crs1800012 CON vs. CL: GG genotype vs. GT + TT genotype

rs1800012 CON vs. ACL: TT genotype vs. GT + GG genotype (P=0.014, OR: 5.1, 95% CI: 1.3-45.0)

rs1800012 CON vs. CL: TT genotype vs. GT + GG genotype (P<0.01, OR: 6.2, 95% CI: 1.9-32.3)

These results show that the rare TT genotype of the Sp1 binding site variant is independently associated with the risk of acute soft tissue injuries such as CL ruptures, but more specifically ACL injuries. However, due to the low frequency of the rs1800012 TT genotype, these results should be interpreted with caution. Furthermore, an important limitation of the combined analysis was the inability to covary for country of birth, although the genotype and allele frequencies appeared to be similarly distributed across the representative cohorts.

Interestingly, the G-T (rs1107946-rs1800012) inferred haplotype was observed to associate with the risk of acute ATR in the current study. Specifically, the G-T allele combination was significantly associated with reduced risk for ATR in the combined (SAW+UK)-Achilles tendon injury cohort; implicating the same haplotype previously associated with reduced risk of ACL ruptures by Ficek et al. (2013) [183]. However, this haplotype was not associated with reduced risk for ACL ruptures in the self-identified SAW-ACL rupture cohort. Future studies should therefore reassess this specific allele combination and its possible influence on injury risk in larger self-identified White, acute injury cohorts. It must be noted that many of the independent studies analysing this variant have been insufficiently powered to detect the rare rs1800012 TT genotype. This may potentially explain the conflicting results when comparing the data from combined analyses as conducted in the current study and previously by Collins et al. (2010) [187] to that of smaller independent cohorts.

The third section of this chapter presented the results from genotyping these two *COL1A1* variants in an ACL-rupture cohort, the participants of which all reported to be of Coloured descent. None of the variants were observed to independently associate with the risk of ACL injury. However, the size of this cohort was a limiting factor. Since this cohort was sampled from a population with a diverse genetic history combining the genotyped data with that from the SAW-ACL rupture cohort was not encouraged. In support of this, the investigated variants demonstrated differences in their level of linkage when compared between the SAC and SAW cohorts. Specifically, the rs1800012 and rs1107946 variants were in high LD in the self-identified White cohorts investigated in this study (SAW-ACL, SAW- and UK-Achilles tendon injury cohorts) but to a lesser degree in the SAC-ACL rupture cohort. In a study by de Wit et al. (2010) [339], population structure analysis was used to characterise the genetic contributions to the SAC population. It was concluded that this population was highly

admixed, with no less than four ancestral contributors including the European, African, Asian and Khoisan populations. Understandably, variations in the genetic structure of these independent populations would potentially influence the complex genetic architecture of the SAC population including allele frequencies and the linkage relationships between markers.

The distribution of the genotype frequencies for the rs1107946 and rs1800012 loci are provided below in Figure 5.9. The allele frequencies for the main ancestral populations (Asian, African and European) are according to the 1000 Genomes Project data (phase 3). Conversely, there is no frequency data currently available for the SAC population. Therefore, the depicted frequencies are those obtained from genotyping the rs1107946 and rs1800012 variants in the SAC-ACL rupture cohort in the current study. The genotype and allele frequencies generated in this current study for the SAW- and UK-Achilles tendon injury and SAW-ACL rupture cohorts are in accordance with the frequencies provided for the general European population. Specifically, the rarity of the rs1800012 TT genotype was evident, with a frequency less than 5% observed in all three investigated self-identified White cohorts. Interestingly, the rs1800012 TT genotype was even more rare in the SAC-ACL rupture cohort (1%) and less heterozygosity was observed compared to the European population. With respect to the rs1107946 variant, the genotype frequencies obtained for the SAW- and UK-Achilles tendon injury and SAW-ACL rupture cohorts were similar to that provided for the European population (<5%). Conversely, the rs1107946 TT genotype frequency was higher in the SAC-ACL rupture cohort (8%) and a greater degree of heterozygosity was observed compared to the European population.

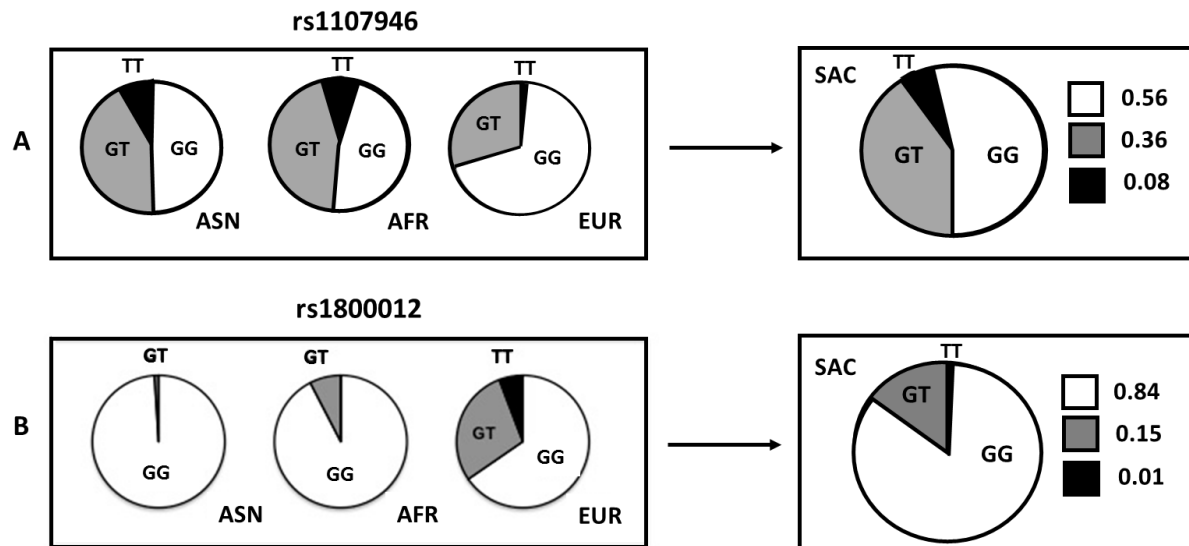


Figure 5.9: Genotype frequency distributions of the *COL1A1* (A) rs1107946 (G>T) and (B) rs1800012 (G>T) variants in the Asian, African, European and South African Coloured populations

Data for the main ancestral populations are according to that provided by the 1000 Genomes Project data (phase 3). Whereas, the frequencies for the SAC population are reflective of the genotype data generated in the current study. ASN: Asian population; AFR: African population; EUR: European population; SAC: South African Coloured cohort.

In summary, the SAC cohort demonstrated genotype frequencies for *COL1A1* rs1107946 and rs1800012 which were different to those of the self-identified White cohorts in the current study. These differences in allele frequencies may potentially explain the association of the T-G (rs1107946-rs1800012) inferred haplotype with the increased risk of ACL ruptures in the SAC cohort but not in any of the self-identified White cohorts. Therefore, this association warrants further interrogation in larger, ethnically diverse injury cohorts to determine whether it is indeed risk-conferring or simply a false positive association resulting from a small sample size.

Most recently, these two variants were also interrogated for their potential contribution to the genetic profile predisposing for ACL ruptures in a cohort of South Asian descent [271]. Like many of the small independent studies, no associations were reported for the *COL1A1* variants. Interestingly however, this study demonstrated the overrepresentation of the rs1800012 TT and rs1107946 TT genotypes in the investigated cohort compared to the GG

genotypes, which presented with very low frequencies (rs1800012: TT 85%, GG 1%; rs1107946: TT 47%, GG 15%). This data does conflict with the allele frequencies provided for the general South Asian population by the 1000 Genomes Project data (rs1800012: TT: 10%, GG: 90%; rs1107946: TT 24%, GG 76%). Furthermore, the sample size of this investigated cohort was an important limiting factor. Therefore, these results would require validation in a larger cohort before any conclusions can be drawn about the distribution of alleles in the South Asian population.

Another important acute phenotype to consider in this discussion are ACL ruptures sustained during skiing-related activities. Two recently published studies demonstrated the significant association of both investigated *COL1A1* variants with the risk of injury [182,186]. Specifically, the rs1107946 GG genotype (GG vs. GT + TT) and G allele were significantly underrepresented in the study's respective control group compared to a group of Polish skiers whom had sustained ACL ruptures. Hence, this variant was inferred to increase the risk of injury. On the other hand, the rs1800012 GG genotype (GG vs. GT + TT) was significantly overrepresented the control group compared to the affected group; inferring reduced risk for ACL injury. However, no inferred haplotypes constructed from both variants were observed to associate with the risk of injury [338]. It is interesting to note that these results contrast with those from previous studies, which described the association of the rs1800012 TT genotype with reduced risk for non-contact ACL ruptures sustained while participating in land-based sports.

Although ACL ruptures commonly present amongst skiers, the mechanisms resulting in these injuries are suggested to be unique to the sport. These injuries usually involve falling backwards after landing a jump, the hips lower than the knees while the feet remain engaged in the ski boots (phantom foot injury mechanism). This position is believed to result in excessive anterior tibial translation [340]. It may be proposed that external factors such as equipment and ground surface may influence the outcome of skiing-associated ACL injuries to a greater degree compared to non-contact ACL ruptures sustained on land, which are believed to occur because of an individual's own movement with no external perturbation. Therefore, the association of the rs1800012 GG genotype with decreased risk for ACL ruptures in skiers requires validation in other independent cohorts [186]. This association, to our

knowledge, has only been reported in the Polish population. Therefore, it would be of interest to investigate these variants in larger skiing-associated ACL injury cohorts, sampled from other populations.

Besides the many case-control genetic associations studies which have been used to explore these variants and their potential involvement in disease predisposition, they have also been explored functionally and have been proposed to influence DNA binding, gene transcription and protein production [275–277]. The rs1107946 variant demonstrates allele-dependent transcriptional differences. Specifically, the G allele is associated with higher transcriptional regulation compared to the T allele, both independently and in combination with a neighbouring promoter variant, rs11327935 (indel/T) [277]. Similarly, the rs1800012 T allele demonstrated increased binding affinity for the Sp1 transcription factor using gel shift assays, in addition to demonstrating a 3-fold increase in the level of primary *COL1A1* mRNA compared to the G allele [275,276]. The T allele was also associated with an increased α_1 (I) to α_2 (I) collagen protein ratio, reflective of the increased ratio of *COL1A1* mRNA relative to *COL1A2* (2.3:1), which is observed to result in the production of homotrimeric type I collagen molecules. In comparison, the production of collagen protein chains by osteoblasts derived from individuals with the rs1800012 GG genotype demonstrated ratios close to the expected (2:1 respectively) [275]. Furthermore, biomechanical properties of bone including strength and stiffness derived from heterozygote individuals were significantly lower compared with that of bone derived from individuals with the rs1800012 GG genotype [275]. Conversely, Jin et al. (2009) reported that the level of transcriptional activity of the *COL1A1* gene is influenced by the interaction of several variants located within the evolutionary conserved 5' regulatory region of the *COL1A1* gene, rather than that of the individual genotypes, for which there was no consistent effects on gene transcription observed [276]. This provides support for the observed association of the *COL1A1* inferred haplotypes with injury risk in the current study.

Jin and colleagues (2009) presented a three variant *COL1A1* haplotype, G-del-T (rs1107946-rs11327935-rs1800012), which drove a significantly higher level of gene transcription compared to the other investigated allele combinations [276]. It must however be noted that the rs1800012 and rs11327935 variants are tagged loci (D' : 1.000, r^2 : 1.000). Therefore,

genotyping only one of these loci as conducted in the current study was necessary. It was postulated that based on the effects on *COL1A1* transcription and collagen type I chain production, the G-del-T inferred haplotype reduced BMD and consequently predisposed for osteoarthritis. Interestingly, this haplotype appears to have the opposite effect within soft musculoskeletal tissues, as suggested by the results presented in the current study and that by Ficek et al. (2013) [183], in which the G-T inferred haplotype (rs1107946-rs1800012) was associated with decreased risk of Achilles tendon and ACL ruptures. It may be postulated that an increase in homotrimeric type I collagen molecules may increase the tensile strength of tendons and ligaments; thereby reducing the risk of acute soft tissue ruptures.

5.4. CONCLUSION

The original association between the *COL1A1* rs1800012 variant and the risk of acute soft tissue ruptures as reported by Khoschnau et al. (2008) [184], Posthumus et al. (2009) [185] and Collins et al. (2010) [187] was observed in the meta-analysis conducted in the current study. Specifically, the rs1800012 TT genotype was inferred to reduce the risk of ACL ruptures. Furthermore, in line with the functional consequence of the G-del-T (rs1107946-rs2412298-rs1800012) inferred haplotype *in vivo* and its previous association with reduced risk of ACL ruptures [183], this haplotype, spanning the 5' regulatory flank of the *COL1A1* gene was also associated with decreased risk for acute ATR. Although the alternate haplotype, T-G (rs1107946-rs1800012), was observed to associate with increased risk of ACL ruptures in the SAC cohort, this association warrants further interrogation based on the fact that it was not associated with injury risk in the larger self-identified White cohorts investigated in this study.

CHAPTER 6: FUNCTIONAL CHARACTERISATION OF THE GENETIC SIGNATURES UNDERPINNING INJURY SUSCEPTIBILITY

6.1. INTRODUCTION

Molecular genetics is a field of research that has rapidly evolved since the advent of next generation sequencing technologies. However, it is becoming increasingly apparent that high throughput sequencing technologies cannot be used in isolation and require an array of *in silico* tools and expertise to aid the meaningful interpretation of the generated data. For example, an important challenge facing researchers who have adopted this level of genomic interrogation to characterise the molecular mechanisms of multifactorial phenotypes, is the difficulty associated with prioritising variants for further investigation from the often-large list of plausible candidates. A cost effective and reasonable approach to filter and prioritise potential risk conferring variants involves the exploration of existing omics data using bioinformatic *in silico* tools [332,341]. This approach, exploited in the current study may provide compelling support for the interrogation of several of the identified sequence signatures *in vitro*.

The studies presented in this thesis used an integrated omics approach involving a tiered filtering strategy to identify plausible candidates from the list of 204 807 high quality variants identified by WES. A total of five new candidate variants mapping to the chr9q32-33.1 genomic interval encompassing the *TNC* (rs1061494, rs1061495) and *COL27A1* (rs2567706, rs2241671, rs2567705) genes were subsequently genotyped in several independent injury cohorts. In addition, eight variants, previously associated with the risk of musculoskeletal soft tissue injuries in the *TNC* (rs1138545, rs2104772), *MMP3* (rs3025058, rs679620, rs591058, rs650108) and *COL1A1* (rs1107946, rs1800012) genes, were prioritised for interrogation. The results of these genetic-association studies have provided support for further interrogation of several of these loci at the functional level. For this reason, all variants associated with the risk of injury in this thesis, either independently or in combination with the other candidates were analysed using an array of *in silico* tools.

6.2. STUDY AIM AND OBJECTIVES

This chapter aimed to characterise the genetic signatures implicated in injury predisposition in chapters 3 to 5 of this thesis at the functional level using several freely available *in silico* tools. A detailed description of the tools used in this study was provided in chapter 2 (Section 2.7); however, an overview is provided in Table 6.1 of this chapter.

In summary, the risk-associated variants identified in the *TNC*, *COL27A1*, *MMP3* and *COL1A1* genes were assessed for their potential impact on:

- Functional sequence motifs and transcription factor binding sites
- Interaction with other genes and variants through chromatin looping
- Pre-mRNA secondary structures and folding

Table 6.1: A brief overview of the *in silico* tools used to functionally characterise the genetic signatures associated with injury predisposition

Name	Description
3DSNP	A database for annotating the regulatory function of human variants by exploring their interactions with genes and other variants mediated by chromatin looping. The total score is based on the variants annotated record for the functional categories: 3D interacting genes (chromatin looping), alterations to sequence motifs and transcription factor binding sites and the interaction with regulatory elements including gene promoters and enhancers in human cell lines [311].
Sfold	A software used to predict changes in pre-mRNA secondary structures [312].
SIFT, PolyPhen-2, CADD and Fathmm_MKL	Tools used to predict the consequence (deleteriousness) of single nucleotide variants on protein function and stability [313–316].
GTex project	Database used to investigate the relationship between genetic variants and gene expression in human tissues [317].
LDlink	An online software used to interrogate linkage disequilibrium between variants within specific population groups [303].

The candidate variants prioritised for interrogation from the WES dataset were annotated using several annotation tools including SIFT: Sorting Intolerant from Tolerant (missense variant annotation); PolyPhen-2: Polymorphism Phenotyping (missense variant annotation); CADD: Combined Annotation Dependent Depletion framework (coding and non-coding variant annotation) and Fathmm_MKL: Functional Analysis through Hidden Markov Models (coding and non-coding variant annotation). The Genotype-Tissue Expression project (GTex consortium, version 7) was used to identify possible modulators of gene expression from the list of risk-associated variants.

6.3. RESULTS

6.3.1. *TNC* gene variants

Chapter 3 of this thesis presented the results from the interrogation of the *TNC* and *COL27A1* genes using a customised WES-analysis pipeline. In total, four *TNC* variants were prioritised for interrogation in two Achilles tendon injury cohorts and two ACL rupture cohorts (Figure 6.1). From these analyses, several findings were described including the association of the rs1061494 (T>C) and rs2104772 (T>A) variants with the risk of injury. Specifically, the rs1061494 CC genotype was associated with increased risk of AT in the SAW-Achilles tendon injury cohort (Section 3.2.2.2), while the rs2104772 AA genotype was associated with increased risk for ACL ruptures in females of the SAW-ACL rupture cohort (Section 3.2.3.2). Furthermore, the alternate rs2104772 T allele was previously reported to associate with reduced risk for AT in the SAW-Achilles tendon injury cohort in the study by Saunders et al. (2013) [198]. In addition, the T-T (rs1061494-rs2104772) inferred haplotype was associated with decreased risk for both chronic AT and acute ACL ruptures in the SAW-Achilles tendon injury and SAC-ACL rupture cohorts (Sections 3.2.2.3 and 3.2.3.3).

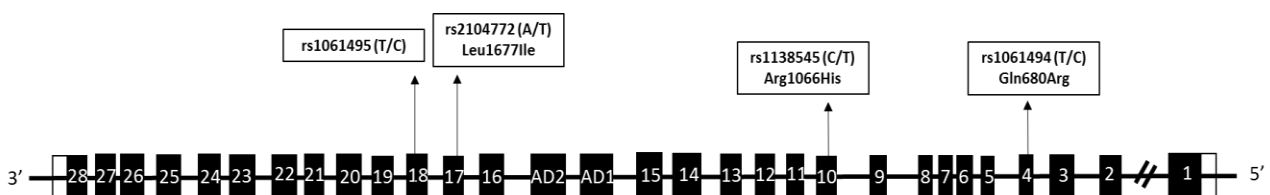


Figure 6.1: Overview of *TNC* variants genotyped in the study described in chapter 3

The *TNC* gene is encoded on the antisense strand. The exons 1-28, AD1 and AD2 (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *TNC* are presented. Arg: Arginine; Gln: Glutamine; His: Histidine; Ile: Isoleucine; Leu: Leucine. The first exon of *TNC* is not translated (double strikethrough). Diagram not drawn to scale.

The 3DSNP scores for the rs1061494 and rs2104772 variants are presented in Figure 6.2. The total score provided for the missense rs1061494 (T>C) variant, which encodes a glutamine to arginine substitution (Gln680Arg), was predominately attributed to its potential effects on DNA sequence motifs and transcription factor binding sites. Specifically, this variant was predicted to localise to the binding sites of several transcription factor types including the family of myogenic factors (Myf) and the muscle initiator (MUSCLE_INI_B) nuclear protein. In addition, this variant was also predicted to localise to the binding site of RNA polymerase II subunit A (POLR2A) in three tissue cell types. Based on intrachromosomal interactions, this variant was predicted to interact with twelve additional *TNC* variants within the European population. These twelve loci were confirmed to be tagged by rs1061494; however, the majority had not been identified by the current WES pipeline due to their location within intronic regions. The Sfold tool also predicted a looser, more single stranded pre-mRNA conformational structure in the presence of the ancestral C allele compared to the T allele (Figure 6.3). Furthermore, the GTex project data identified this variant as an eQTL in testis tissue. Specifically, the CC genotype is predicted to associate with increased levels of *TNC* expression.

The second variant of the risk-associated *TNC* T-T inferred haplotype, rs2104772 (T>A) (Figure 6.1), had a lower 3DSNP score which was largely attributed to its association with several predicted enhancers in nine different cell lines (Figure 6.2). It may be hypothesised that through chromatin looping this variant, located within exon 17 of the gene, may interact with *cis* acting enhancer-like elements or nuclear complexes within the regulatory regions of the gene. However, the variant itself was not identified to overlap any DNA sequence motifs or locate within the binding site of any transcription factors. Furthermore, analysis of the predicted pre-mRNA secondary structures demonstrated similar centroid conformations for the rs2104772 T and A alleles (Figure 6.3). However, rs2104772, together with the *TNC* rs1138545 variant which was also prioritised for interrogation in chapter 3 and previously associated with RCT [214,215], demonstrated CADD_phred scores >13 (Table 6.2). This score, which is used to estimate the pathogenicity of genetic variation, is equivalent to a 95% probability that the variants have a functional effect [315]. Moreover, the rs2104772 variant was predicted by the Fathmm-MKL database to be '*deleterious*' in nature.

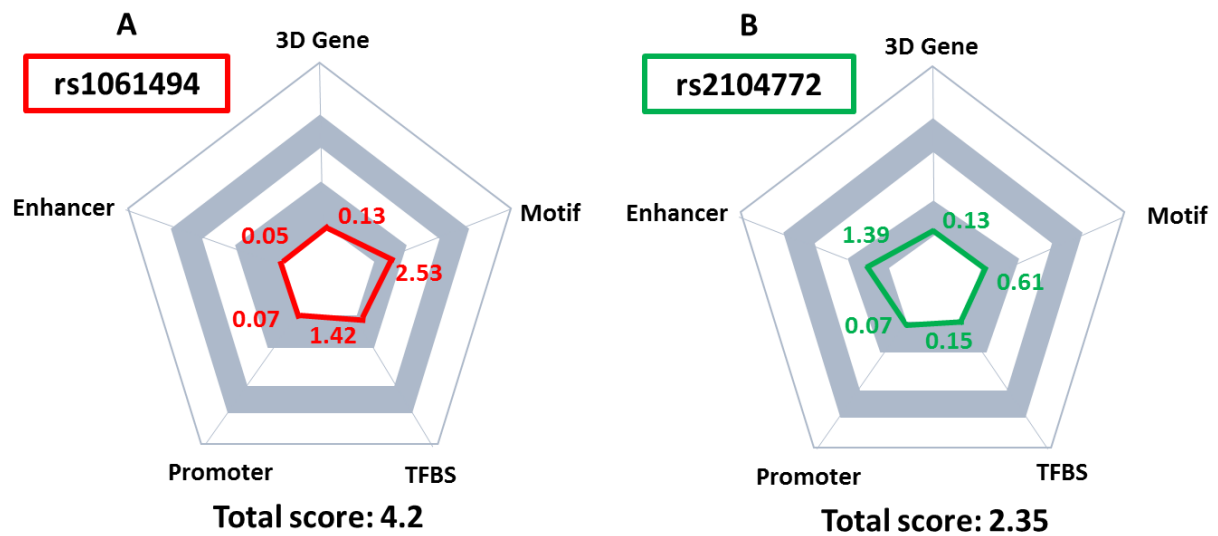


Figure 6.2: Functional annotation of the risk-associated *TNC* (A) rs1061494 (T>C) and (B) rs2104772 (T>A) variants using the 3DSNP annotation tool

The total score (bold typeset) for each variant is the sum of the individual quantitative scores for the five functional categories: 3D interacting genes (chromatin looping), alterations to sequence motifs and transcription factor binding sites and the interaction with regulatory elements including gene promoters and enhancers in human cell lines. The deformation of the centroid hexagon for a category, demonstrates the means through which the variant is predicted to influence function. The greater the individual score, the greater the annotation record (biological evidence) for the predicted effect [<http://cbportal.org/3dsnp/>].

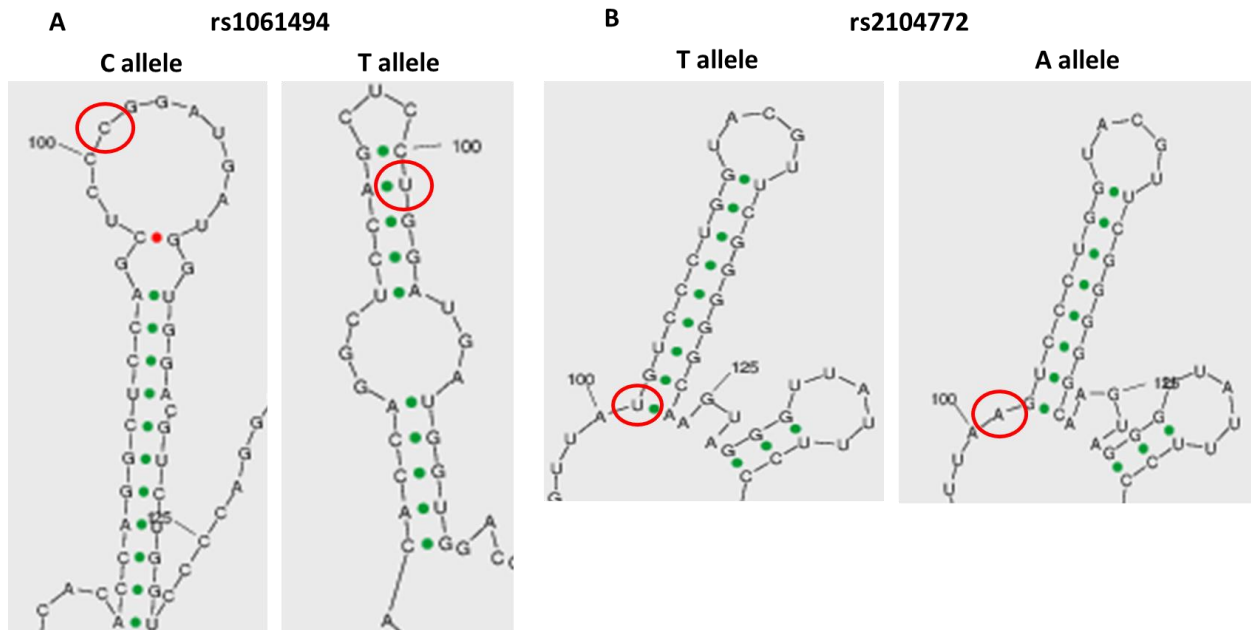


Figure 6.3: Sfold predicted pre-mRNA secondary structures immediately surrounding the *TNC* (A) rs1061494 (T>C) and (B) rs2104772 (T>A) variants

The diagrams illustrate the predicted three-dimensional architecture as dictated by the pre-mRNA sequence. Red circle indicates variant location. (A) Ancestral C allele of the rs1061494 variant results in a loose hairpin structure in the pre-mRNA structure compared to the T allele. (B) The rs2104772 T and A alleles demonstrated similar centroid conformations.

Table 6.2: Predicted functional effects of the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants

Gene	SNP ID	Coding effect	Functional Effect Predicted by:			
			SIFT	PolyPhen-2	CADD	Fathmm_MKL
<i>TNC</i>	rs1061494	NON-SYN	T	B	0.672	N
	rs1138545	NON-SYN	T	B	17.19	N
	rs2104772	NON-SYN	T	B	17.21	D
	rs1061495	SYN	-	-	-	N

SNP ID: dbSNP accession number. Coding effect: NON-SYN (non-synonymous); SYN (synonymous). SIFT: Sorting Intolerant from Tolerant (T: Tolerated; D: Deleterious); PolyPhen-2: Polymorphism Phenotyping (B: Benign; D: Deleterious); CADD: Combined Annotation Dependent Depletion; Fathmm_MKL: Functional Analysis through Hidden Markov Models (N: Neutral; D: Deleterious). Bold typeset indicates a score which predicts functional significance.

6.3.2. *COL27A1* gene variants

One of the candidates prioritised for further interrogation from the WES dataset in the study described in chapter 3, *COL27A1* rs2241671 (G>A), was implicated in the genetic profile predisposing for Achilles tendon injuries and ACL ruptures (Figure 6.4). Based on the results from the case-control genetic association studies presented in this thesis, the rs2241671 GG genotype was associated with increased risk for AT in the UK-Achilles tendon injury cohort (Section 3.2.2.2), spontaneous ATR in the combined-Achilles tendon injury cohort (SAW+UK) and ACL ruptures in the SAC-ACL rupture cohort (Sections 3.2.2.2 and 3.2.3.2 respectively). Furthermore, a three-variant *COL27A1* inferred haplotype, A-G-A (rs2567706-rs2241671-rs2567705) was also associated with increased risk for non-contact ACL ruptures in the SAC-ACL rupture cohort, whereas the alternate inferred haplotype, G-A-T (rs2567706-rs2241671-rs2567705) was associated with decreased risk (Section 3.2.3.3).

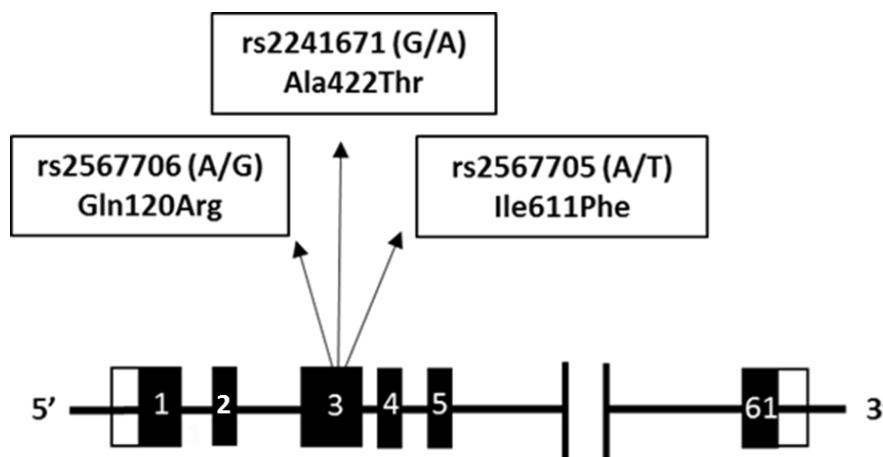


Figure 6.4: Overview of *COL27A1* variants genotyped in the study described in chapter 3

The *COL27A1* gene is encoded on the sense strand. The exons (black bars), introns (horizontal lines) and untranslated regions (empty bars) of *COL27A1* are presented. Exons 6-60 have not been included. Ala: Alanine; Arg: Arginine Gln: Glutamine; Ile: Isoleucine; Phe: Phenylalanine Thr: Threonine. Diagram not drawn to scale.

Figure 6.5 presents the 3DSNP scores for the three *COL27A1* variants prioritised from the WES dataset. These three missense variants: rs2567706 (Gln120Arg), rs2241671 (Ala422Thr) and rs2567705 (Ile611Phe) scored highest for their possible association with the enhancer state of cells, followed by their interaction with other neighbouring genes on chromosome 9 in a three-dimensional environment. Interestingly, all three variants are located within exon 3 of the *COL27A1* gene within a 1.4 kb stretch of each other and approximately 10 kb downstream from the transcription start site. Furthermore, these three variants are predicted to interact with other loci in at least six other genes located on chromosome 9 including the *MIR455* gene, which encodes a small non-coding RNA involved in post-transcriptional control. These interactions are proposed to be mediated by chromatin folding. The Sfold centroid structure for the rs2567706 G (G>A) and rs2241671 A (G>A) alleles are predicted to result in loosely assembled pre-mRNA secondary structures compared to that of the alternate alleles (Figure 6.6). The predicted centroid folding surrounding the rs2567705 variant appeared relatively unchanged in the presence of the T and A alleles. However, this latter variant did demonstrate a CADD phred scores > 13, equivalent to a 95% probability of having a functional effect [315] (Table 6.3).

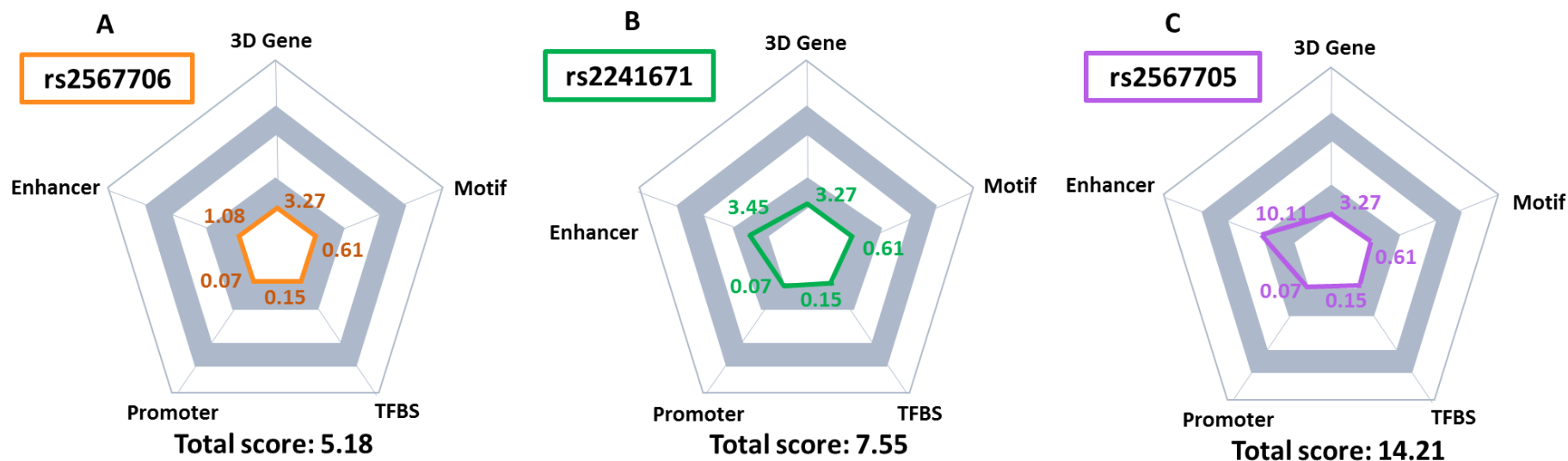


Figure 6.5: Functional annotation of the risk-associated *COL27A1* (A) rs2567705 (A>G), (B) rs2241671 (G>A) and (C) rs2567706 (A>T) variants using the 3DSNP annotation tool

The total score (bold typeset) for each variant is the sum of the individual quantitative scores for the five functional categories; 3D interacting genes (chromatin looping), alterations to sequence motifs and transcription factor binding sites and the interaction with regulatory elements including gene promoters and enhancers in human cell lines. The deformation of the centroid hexagon for a category, demonstrates the means through which the variant is predicted to influence function. The greater the individual score, the greater the annotation record (biological evidence) for the predicted effect [<http://cbportal.org/3dsnp/>].

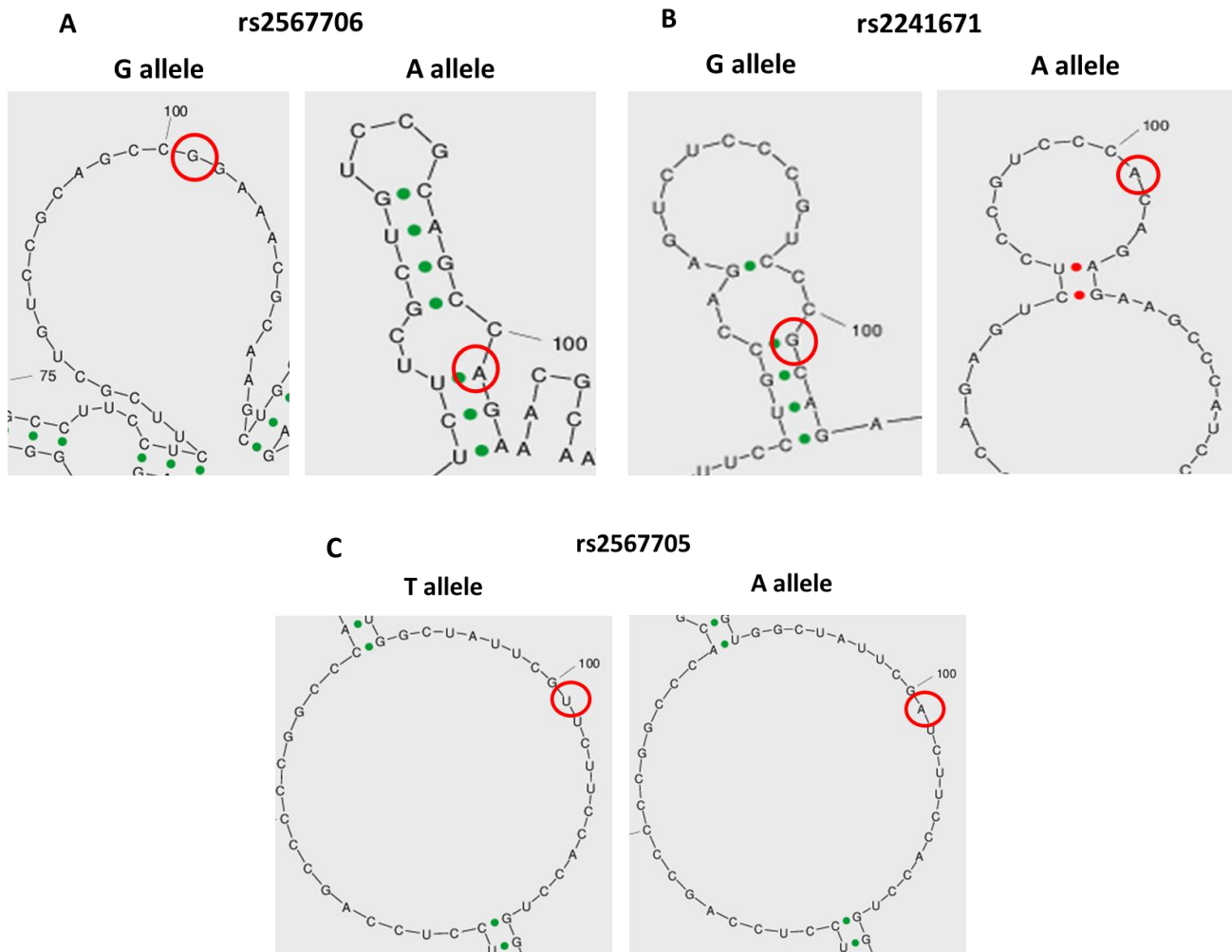


Figure 6.6: Sfold predicted pre-mRNA secondary structures immediately surrounding the *COL27A1* (A) rs2567706 (A>G), (B) rs2241671 (G>A) and (C) rs2567705 (A>T) variants

The diagrams illustrate the predicted three-dimensional architecture as dictated by the pre-RNA sequence. Red circle indicates variant location. (A) The centroid structure of the ancestral G allele of the rs2567706 variant is predicted to result in a single stranded hairpin assembly compared to the alternate A allele. Whereas, the alternate A allele of the (B) rs2241671 variant is predicted results in a looser, presumably more accessible structure compared to the ancestral G allele. (C) The predicted centroid folding surrounding the rs2567705 variant appeared unchanged in the presence of the T and A alleles.

Table 6.3: Predicted functional effects of the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants

Gene	SNP ID	Coding effect	Functional Effect Predicted by:			
			SIFT	PolyPhen-2	CADD	Fathmm_MKL
<i>COL27A1</i>	rs2567706	NON-SYN	T	B	2.00	N
	rs2241671	NON-SYN	T	B	0.001	N
	rs2567705	NON-SYN	T	B	16.51	N

SNP ID: dbSNP accession number. Coding effect: NON-SYN (non-synonymous); SYN (synonymous). SIFT: Sorting Intolerant from Tolerant (T: Tolerated; D: Deleterious); PolyPhen-2: Polymorphism Phenotyping (B: Benign; D: Deleterious); CADD: Combined Annotation Dependent Depletion framework; Fathmm_MKL: Functional Analysis through Hidden Markov Models (N: Neutral; D: Deleterious). Bold typeset indicates a score which predicts functional significance.

6.3.3. *MMP3* gene variants

Chapter 4 of this thesis presented the results from genotyping four *MMP3* variants: rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A), which had previously been associated with the risk of musculoskeletal soft tissue injuries, in two additional injury cohorts (Figure 6.7). Besides the rs3025058 variant, which was associated with chronic AT by inference in the previously genotyped SAW-Achilles tendon injury cohort [206], none of the *MMP3* variants were associated with the risk of chronic AT or acute ACL ruptures in the AUS-Achilles tendon injury and SAW-ACL rupture cohorts. Moreover, only the alternate inferred haplotype (6A-G-C-G) to that previously proposed to modulate risk (5A-A-T-G) [206] was associated with decreased risk for AT in the AUS-Achilles tendon injury cohort (Section 4.2.1.3). Despite the inability to replicate the original findings, the investigated *MMP3* variants, several of which have been suggested to influence functions including gene transcription, protein production and protein phosphorylation, were prioritised for further interrogation using *in silico* analyses.

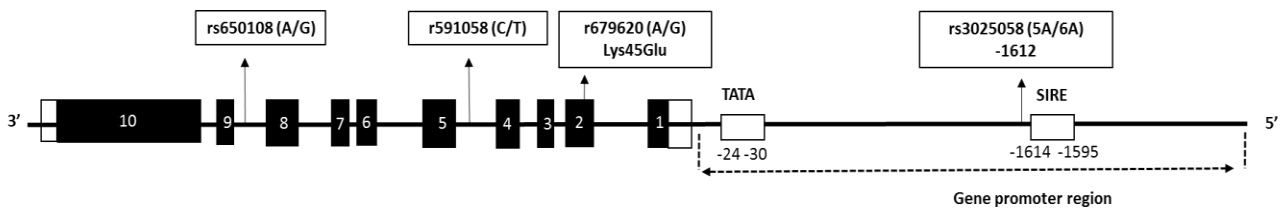


Figure 6.7: Overview of *MMP3* variants genotyped in the study described in chapter 4

The *MMP3* gene is encoded on the antisense strand. The exons 1-10 (black bars), introns (horizontal lines), untranslated regions (empty bars) and promoter region (dashed horizontal line) of *MMP3* are presented. (-) Base pair number upstream of transcription start site. (+) Base pair number downstream of transcription start site. TATA: TATA box; SIRE: Stromelysin IL-1 responsive element; Glu: Glutamic acid; Lys: Lysine. Diagram not drawn to scale.

The 3DSNP scores for the *MMP3* rs3025058, rs679620, rs591058 and rs650108 variants are presented in Figure 6.8. These variants scored highest for their possible association with the enhancer state of cells and other *MMP3* variants, in addition to their interaction with other genes on chromosome 11 in a 3D environment. Specifically, the missense rs679620 (Lys45Glu) and rs591058 (T>C) variants were observed to tag an additional 38 *MMP3* variants. Conversely, the most distally situated variant, rs650108 (G>A), tagged only seven loci. Furthermore, it was suggested that through chromatin looping these variants may potentially interact with other loci located within the neighbouring *MMP12* gene, situated 31 kb upstream. The rs3025058 (5A>6A) variant maps 1 612 bp upstream from the transcription start site within the distal promoter of the gene. This variant is known to overlap the sequence motif of the SIRE. The SIRE is an IL-1 inducible repressor of stromelysin gene expression, believed to aid in the fine-tuning of *MMP3* gene expression [328]. Despite its location within the SIRE sequence motif, analysis of the Sfold centroid structures for the rs3025058 variant suggests a relatively unchanged conformation in the presence of the 5A or 6A allele. The only variant which does demonstrate marked differences in the predicted centroid pre-mRNA secondary structure is the rs679620 variant, a tagged locus of the functional rs3025058 promoter variant (Figure 6.9). Specifically, the alternate rs679620 A allele is predicted to result in a large hairpin loop immediately surrounding the variant compared to that predicted for the ancestral G allele.

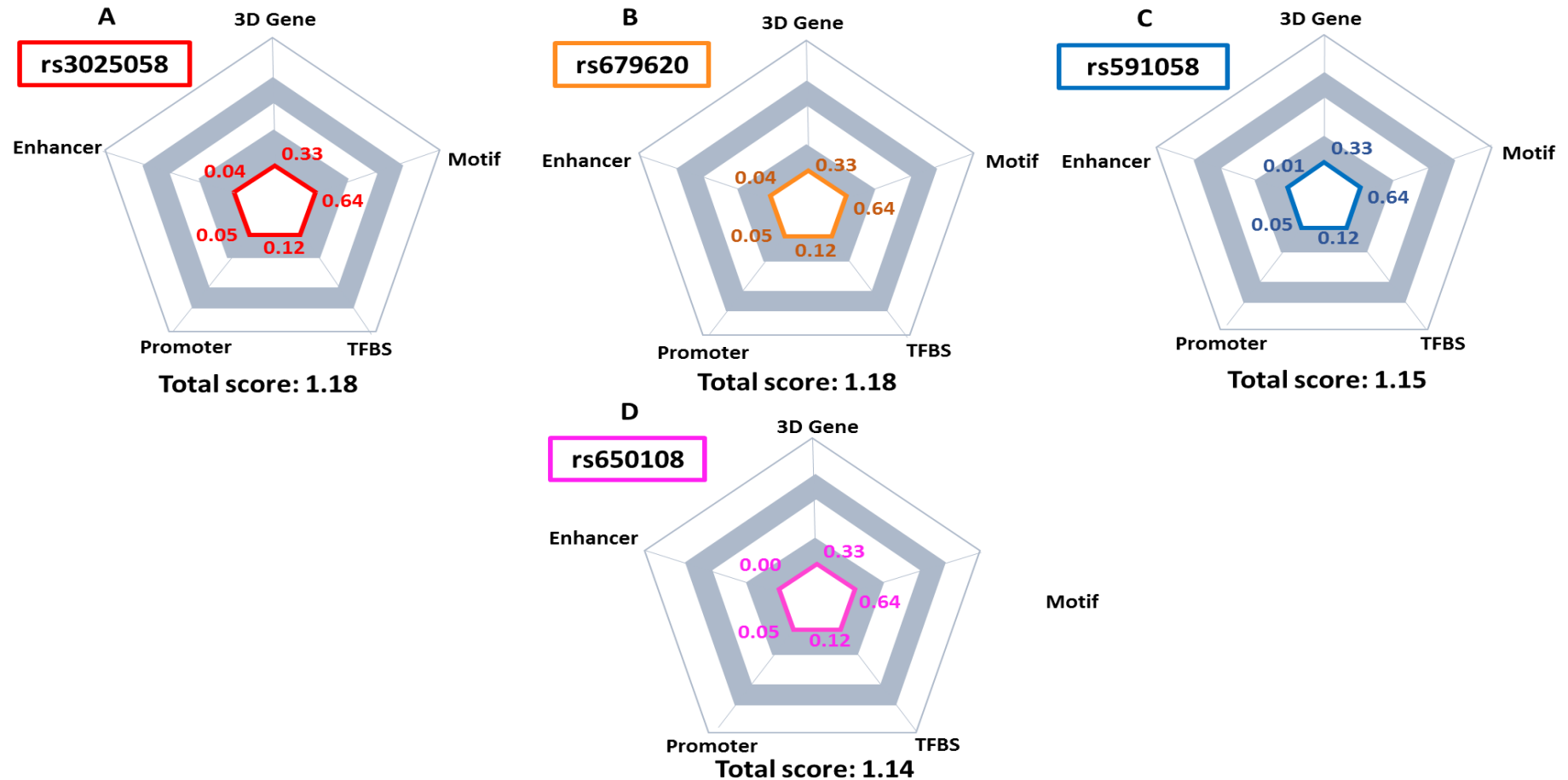


Figure 6.8: Functional annotation of the risk-associated *MMP3* (A) rs3025058 (5A>6A), (B) rs679620 (A>G), (C) rs591058 (T>C) and (D) rs650108 (G>A) variants using the 3DSNP annotation tool

The total score (bold typeset) for each variant is the sum of the individual quantitative scores for the five functional categories; 3D interacting genes (chromatin looping), alterations to sequence motifs and transcription factor binding sites and the interaction with regulatory elements including gene promoters and enhancers in human cell lines. The deformation of the centroid hexagon for a category, demonstrates the means through which the variant is predicted to influence function. The greater the individual score, the greater the annotation record (biological evidence) for the predicted effect [<http://cbportal.org/3dsnp/>].

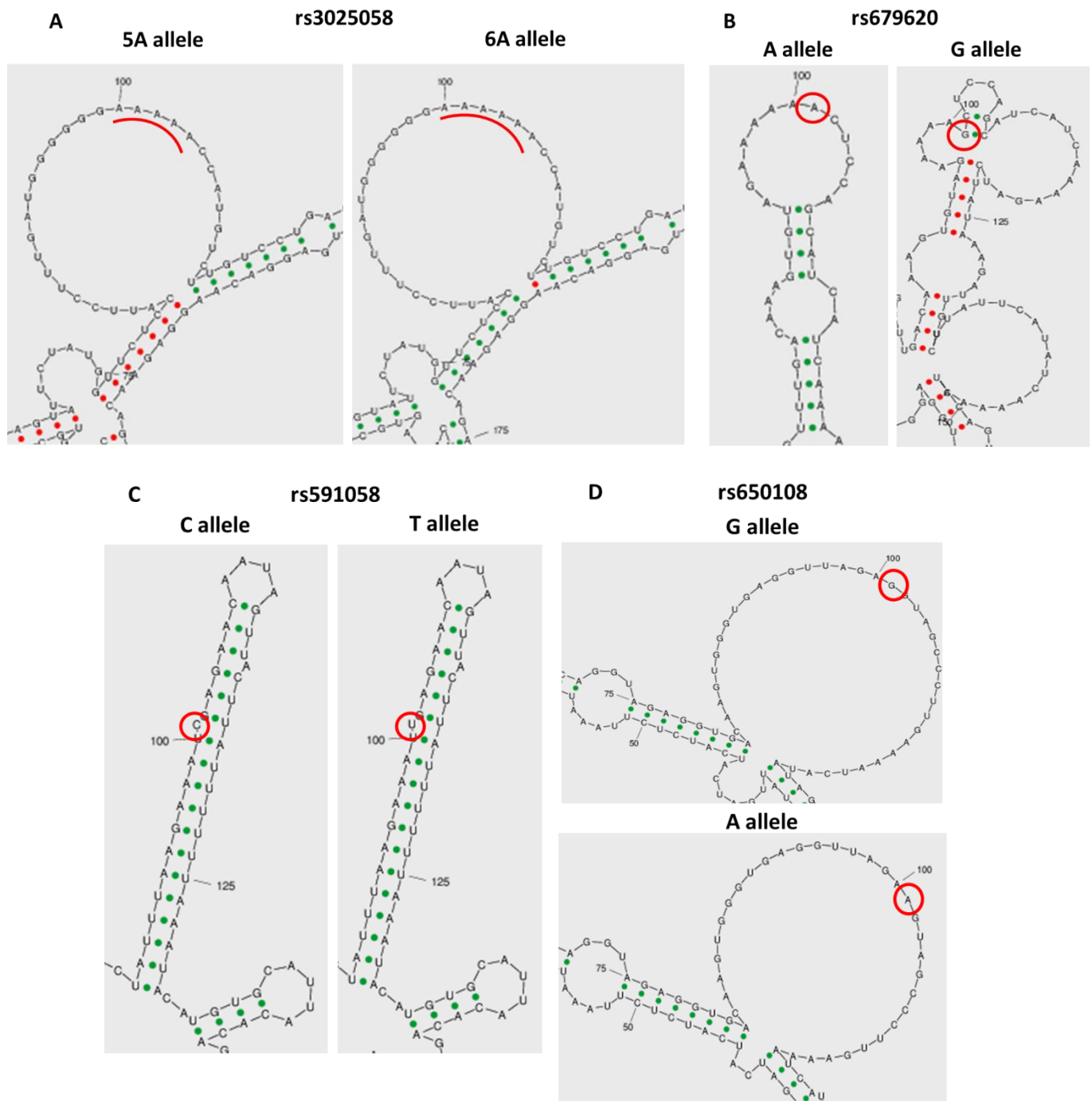


Figure 6.9: Sfold predicted pre-mRNA secondary structures immediately surrounding the *MMP3* (A) rs3025058 (5A>6A), (B) rs679620 (A>G), (C) rs591058 (T>C) and (D) rs650108 (G>A) variants

The diagrams illustrate the predicted three-dimensional architecture as dictated by the pre-mRNA sequence. Red circle indicates variant location. Only the rs679620 variant demonstrated differences in the centroid pre-mRNA secondary structures. The alternate rs679620 A allele is predicted to result in a large hairpin loop immediately surrounding the variant compared to that predicted for the ancestral G allele.

6.3.4. *COL1A1* gene variants

Chapter 5 of the thesis presented the results from genotyping two previous, risk-associated *COL1A1* variants (rs1107946 G>T and rs1800012 G>T) in four additional injury cohorts (Figure 6.10) [182–187]. Although not associated with injury in any of the prioritised cohorts, the rare rs1800012 TT genotype was associated with the risk of ACL ruptures in a meta-analysis conducted in chapter 5. Furthermore, the G-T (rs1800012-rs1107946) inferred haplotype, previously associated with reduced risk of ACL ruptures [183], was also associated with the decreased risk of ATR in the combined (SAW and UK)-Achilles tendon injury cohort. In accordance with this association, the alternate T-G (rs1107946-rs1800012) inferred haplotype was associated with increased risk of ACL rupture in the SAC-ACL rupture cohort (Section 5.2.2.3); however, this finding warrants further interrogation as it was not observed in any of the self-identified White injury cohorts.

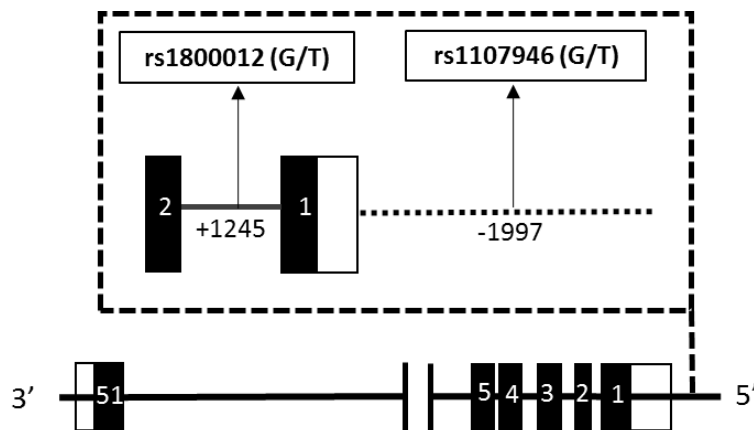


Figure 6.10: Overview of *COL1A1* variants genotyped in the study described in chapter 5

The *COL1A1* gene is encoded on the antisense strand. The exons (black bars), introns (horizontal lines), untranslated regions (empty bars) and promoter region (dashed horizontal line) of *COL1A1* are presented. Exons 6-50 have not been included. (-) Base pair number upstream of transcription start site. (+) Base pair number downstream of transcription start site. Diagram not drawn to scale.

The 3DSNP scores for the rs1107946 and rs1800012 variants are presented in Figure 6.11. The score provided for the rs1107946 (G>T) variant was attributed to its potential effect on the gene's promoter and enhancers of cell function, in addition to its interaction with other genes within a three-dimensional environment. Furthermore, through chromatin looping, rs1107946 was predicted to interact with at least 11 other genes on chromosome 17 and tagged three additional *COL1A1* variants, rs2269336 (G>C, promoter), rs2075554 (C>T, intron 11) and rs2075555 (G>T, intron 11) respectively. The Sfold centroid structures were similar when comparing the pre-mRNA secondary structures of the rs1107946 G and T alleles (Figure 6.12).

The *COL1A1* rs1800012 (G>T) variant demonstrated the highest predicted score of functionality of all those prioritised for interrogation in this thesis (Figure 6.11). This score was attributed to four main functional categories including promoter state, transcription factor binding sites, alterations to sequence motifs and gene interactions. Although this variant is located within the first intron of the *COL1A1* gene, it is predicted to interact with the genes promoter in several cell types. Furthermore, rs1800012 is predicted to localise to the recognition sequence of nine proposed transcription factors across a range of cell lines. In addition, rs1800012 was predicted to interact with 11 other genes on chromosome 17 modulated by chromatin looping but only demonstrated a tagged relationship with two other *COL1A1* variants in the European population, rs9898186 (C>T, intron 1) and rs11327935 (indel/T, promoter) respectively. The Sfold predicted pre-mRNA secondary structure in the presence of the rs1800012 T allele demonstrated a more relaxed, single stranded conformation immediately surrounding the variant compared to the double stranded conformation predicted for the G allele (Figure 6.12).

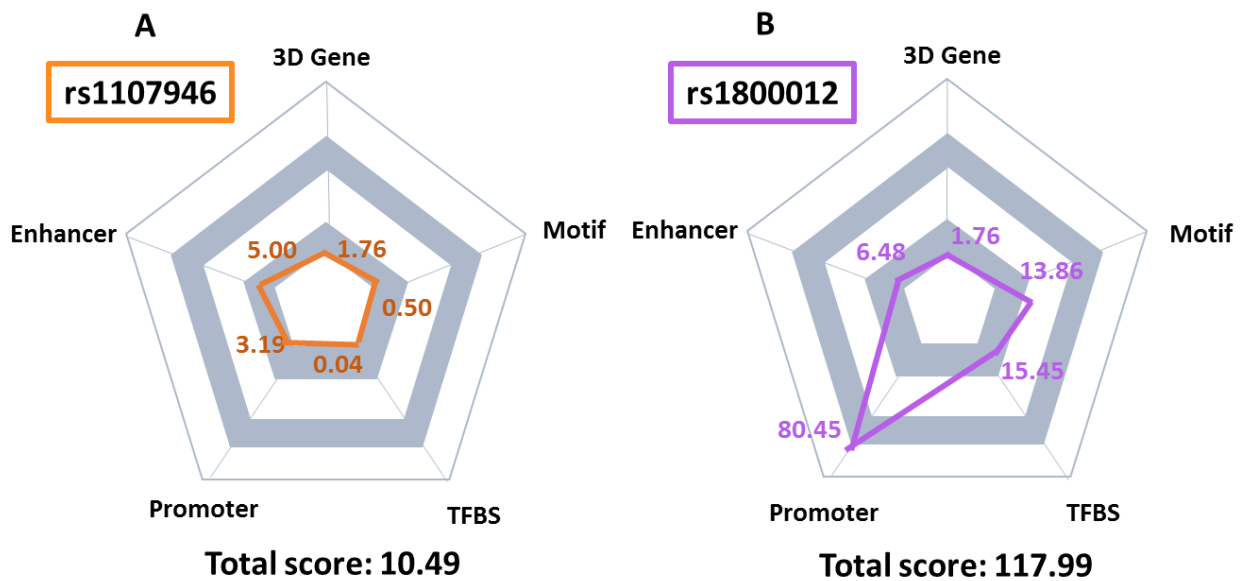


Figure 6.11: Functional annotation of the risk-associated *COL1A1* (A) rs1107946 (G>T) and (B) rs1800012 (G>T) variants using the 3DSNP annotation tool

The total score (bold typeset) for each variant is the sum of the individual quantitative scores for the five functional categories; 3D interacting genes (chromatin looping), alterations to sequence motifs and transcription factor binding sites and the interaction with regulatory elements including gene promoters and enhancers in human cell lines. The deformation of the centroid hexagon for a category, demonstrates the means through which the variant is predicted to influence function. The greater the individual score, the greater the annotation record (biological evidence) for the predicted effect [<http://cbportal.org/3dsnp/>].

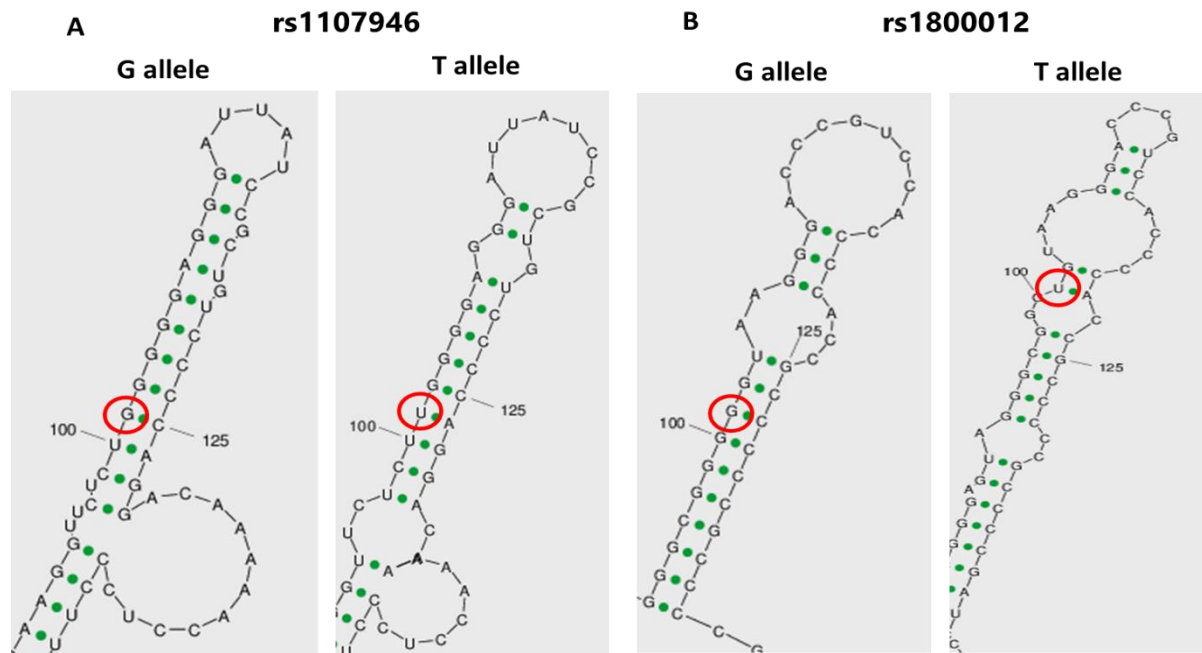


Figure 6.12: Sfold predicted pre-mRNA secondary structures immediately surrounding the *COL1A1* (A) rs1107946 (G>T) and (B) rs1800012 (G>T) variants

The diagrams illustrate the predicted three-dimensional architecture as dictated by the pre-mRNA sequence. Red circle indicates variant location. (A) The Sfold centroid structures were similar for the rs1107946 G and T alleles. (B) The Sfold predicted pre-mRNA secondary structure demonstrated a small 'pocket' in the presence of the rare rs1800012 T allele; potentially increasing accessibility to the sequence surrounding this locus, compared to the ancestral G allele, which demonstrated a double stranded confirmation at the locus.

6.4. DISCUSSION

The use of WES is not without limitation; however, it is a powerful molecular tool with the ability to identify functional variants that may contribute to the heritable predisposition to disease including musculoskeletal soft tissue injuries. Several novel risk associations have been presented in this thesis. Therefore, the aim of this study was to use an *in silico* approach to build on the statistical evidence provided by the case-control genetic association studies described in this thesis, implicating the prioritised loci in the profiles predisposing for injury. The results described in the current chapter suggest that, although not disease causing, several of the variants investigated in this thesis may have an effect at the functional level. It is vitally important that we understand the biological significance of genetic signatures in our efforts to develop multifactorial models that are capable of accurately assessing injury risk.

The *TNC* variants, which were moderately associated with the risk of AT (rs1061494 T>C, OR: 2.5, 95% CI: 1.2-5.1) and ACL ruptures (rs2104772 T>A, OR: 2.3, 95% CI: 1.1-5.5) in chapter 3 of this thesis, have demonstrated potential functional characteristics *in silico*. However, explaining these associations and the predicted functional effects of these variants within the context of tenascin-C biology is a complex task. It is known that the rs1061494 variant is present in all tenascin-C isoforms, whereas rs2104772 only presents in certain isoforms generated through the process of alternative splicing. Interestingly, the different tenascin-C isoforms exhibit contrasting functions in matrix remodelling [218]. The larger tenascin-C isoforms are expressed during active tissue remodelling. Within this environment they modulate cell proliferation and promote cell rounding and migration, while reducing focal adhesions (anti-adhesive). The smaller isoform on the other hand is expressed within more quiescent tissues, where it promotes cell attachment and the formation of focal adhesions, consequently increasing the stiffness and stability of the micro-environment (pro-adhesive) [222]. The rs1061494 variant maps to the fourth exon of the gene, encoding the first constitutively expressed FNIII repeat domain of the tenascin-C peptide. The rs2104772 (T>A) variant on the other hand is located within exon 17 of the gene, encoding the last variable FNIII repeat domain subject to alternative splicing (FNIIID). This specific FNIII domain is spliced

from the smaller tenascin-C isoform (~200kDa) but may be included in several of the larger tenascin-C isoforms (≥ 230 kDa). Several different sized tenascin-C isoforms may present in samples from degenerate and ruptured tendons, including protein fragments, believed to be the result of proteolytic degradation of the larger tenascin-C isoforms. Conversely, only the small, stable tenascin-C isoform, generated through the timely and regulated splicing of the variable FNIII repeat domains is found in normal tendons [321]. Therefore, the presence of several tenascin-C isoforms in pathological tendons may be hypothesised to result from the loss in effective regulation of gene expression and splicing, which may have a profound effect on protein production. This may consequently lead to chronic tissue remodelling which drives tissue degeneration. This hypothesis also supports the observed heterogeneity of the matrix of diseased tendons, in which portions of the tendon are degenerative while other portions appear normal [11].

At the level of transcription, the rs1061494 variant is a predicted eQTL in testis tissue, with the CC genotype associated with a significant increase in *TNC* expression. This may potentially involve the preferential binding of specific regulatory elements including the myogenic factor family of transcription factors and POLR2A to the sequence encompassing this variant. Furthermore, the proposed alterations in the pre-mRNA structures presented in this current chapter provide a hypothetical mechanism through which the investigated and constitutively expressed *TNC* rs1061494 variant may influence tenascin-C expression. Specifically, the looser, more single stranded pre-mRNA confirmation in the presence of the rs1061494 C allele suggests that the sequence surrounding the rs1061494 locus may be more accessible to potential regulators of gene function. Although it is yet to be determined whether the transcript-altering properties of rs1061494 in testis tissue translate to expression changes in dense connective tissue, the rs1061494 CC genotype was observed to be significantly overrepresented in participants diagnosed with chronic AT compared to asymptomatic control participants in chapter 3. This association, in addition to the predicted functionality of the rs1061494 variant, are in accordance with the histological observation of increased tenascin-C expression in diseased tendons [321].

With respect to rs2104772, although no transcription factor binding sites or sequence motifs were predicted to be influenced, this variant was predicted to be 'deleterious' by two independent annotation tools, CADD and Fathmm_MKL. Both tools predicted the change of the amino acid from a leucine residue to an isoleucine residue at position +1 677 downstream from the transcription start site to be functionally impressionable. In support of these predictions, Matsuda et al. (2005), using protein-structure modelling, predicted instability of the FNIIID domain structure in the presence of the isoleucine substitution (A allele) due to steric hindrance inside the fifth beta-sheet [213]. Conversely, the leucine residue (T allele) was predicted to form a stable hydrophobic interaction plane amongst the hydrophobic amino acids. It was hypothesised that this alteration may adversely affect the molecular elasticity of this FNIII domain, which may ultimately influence the ability of specific tenascin-C isoforms to modulate cell adhesion and spreading.

Considering this evidence, it would be beneficial to explore the net effect of these potentially functional variants within the *TNC* gene in mediating tenascin-C modulation of connective tissue matrix remodelling. Furthermore, it is interesting to note the potential biological effect of an 'altered' 3D conformational structure of tenascin-C, whereby concealed binding sites and/or specific sequence motifs may become exposed and thereby affect matrix component interactions and subsequent ECM function [218,323,342]. Therefore, future studies should aim to explore the combined effect of these variants *in vitro*, in addition to assessing their potential effect on the three-dimensional conformation of the protein.

The variants within the *COL27A1* gene, located upstream from *TNC*, also demonstrated several, potentially functional characteristics *in silico*. Compared to the other fibrillar collagens, the *COL27A1* gene has only recently been characterised [239]. The study described in chapter 3 is, to our knowledge, only the second to have prioritised the *COL27A1* gene for interrogation for its potential contribution to musculoskeletal soft tissue injuries [198], although this gene is known to be predominantly expressed in cartilage [239]. The *COL27A1* gene was predicted to interact with several other genes through chromatin looping; providing scope for gene-gene interaction studies in the future. Furthermore, all three WES prioritised missense mutations were scored highest for their potential interaction with enhancers of cell

function, although none were located within transcription factor binding sites or DNA sequence motifs. These functional predictions do however warrant further interrogation. This area of research has gained much attention within the last decade, with some proposing chromatin folding as a potential mechanism regulating the predisposition to multifactorial conditions [343,344]. It is for this reason that the phenomenon of chromatin looping remains a potential avenue of exploration for future studies, with the aim of characterising the interaction of functionally confirmed susceptibility loci for musculoskeletal soft tissue injuries.

The *in silico* analysis of the *MMP3* gene in the current chapter supported previous observations of high linkage across this genomic region [345]. Specifically, the rs679620 and rs591058 variants tagged one another and were predicted to tag an additional 38 loci spanning the 8 kb *MMP3* genomic interval. This level of linkage may be hypothesised to indicate genomic sequence conservation under evolutionary selection; however, other sequence features may contribute this complex relationship [346]. Stromelysin 1 is a matrix endopeptidase with broad substrate specificity. Therefore, changes in its pattern of expression brought about by the interaction of several functional, haplotype-tagging variants may influence matrix remodelling. However, the inability of the study described in chapter 4 to replicate the associations described by Raleigh et al. (2009) [206] and Malila et al. (2011) [204] suggests that other, potentially more informative markers of risk still require identification and that the associations may have been false positive discoveries. Significantly larger sample sizes are therefore required to further interrogate this chromosomal locus. However, it is interesting to note the predicted interaction between the *MMP3* and *MMP12* genes, the latter of which is located 31 kb upstream from the former. In support of this, Posthumus et al. (2012) previously demonstrated the association of an inferred haplotype spanning four *MMP* genes on chromosome 11 including *MMP10*, *MMP1*, *MMP3* and *MMP12* with the risk of ACL ruptures in the SAW-ACL rupture cohort [203]. This previous association, together with the proposed interactions of the *MMP3* and *MMP12* genes in a 3D environment provides support for the interrogation of the WES generated dataset for other potentially more informative markers of risk, mapping across the *MMP* gene cluster on chromosome 11q22.

The last two variants prioritised for interrogation using this *in silico* approach were the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) loci located on the long arm of chromosome 17. These variants have previously been associated with the risk of ACL ruptures [182–187]. Similarly, the rare rs1800012 TT genotype was moderately associated with reduced risk of ACL ruptures in chapter 5 of this thesis (OR: 5.1, 95% CI: 1.3–45.0) and an inferred haplotype spanning both variants was associated with acute ATR. Both *COL1A1* variants scored highly for their predicted interaction with the promoter state of cells using the 3DSNP annotation resource. This is not surprising since the rs1107946 variant locates within the distal promoter of the gene and the rs1800012 variant, located within intron 1 of the gene, is known to interact with elements within the 5' regulatory region.

In support of the functional predictions presented in this chapter, examination of the sequence surrounding the rs1107946 promoter variant (-1997 G>T) has previously revealed a consensus sequence for the Sp1 transcription factor [276]. Specifically, purified Sp1 was observed to bind to this region with the T allele demonstrating an increased affinity for the protein compared to the G allele. However, it was also suggested that although the Sp1 transcription factor has the ability to bind at this locus, it may not necessarily be responsible for DNA-protein binding in all cell types. On the other hand, the rs1800012 (+1245 G>T) variant is known to localise to the Sp1 transcription factor binding site in intron 1 of the gene. Besides the binding of the Sp1 transcription factor, the nuclear matrix protein-4 (Nmp4) is also observed to interact with the region overlapping the rs1800012 variant, although no Nmp4 consensus sequence is recognised [276]. Jin et al. (2009) presented this as evidence for a potential interaction between the Sp1 binding site variant in intron 1 and the *COL1A1* promoter variant, rs11327935 (-1663, indel/T), to which Nmp4 is known to bind and which is tagged by rs1800012 [276]. The 3DSNP resource provided support for the hypothesis by Jin et al. (2009), that this interaction may be regulated through chromatin looping (Figure 6.13).

Previous investigations have also demonstrated the preferential binding of the Sp1 transcription factor to the sequence encompassing the rare rs1800012 T allele compared to the ancestral G allele. Moreover, the T allele has been associated with a 3-fold increase in primary mRNA transcript compared to the G allele [275]. Interestingly, the Sfold predicted

pre-mRNA secondary structure for the rs1800012 variant suggested that the sequence surrounding this locus in the presence of the rare T allele may be more accessible to regulators of gene function compared to the sequence in the presence of the ancestral G allele (Figure 6.11). Although these predicted features require validation, they do support the observed increase in α_1 chains of type I collagen and the formation of homotrimeric type I collagen molecules; the product of the G-del-T (rs1107946- rs11327935-rs1800012) functional haplotype spanning the 5' regulatory flank of the *COL1A1* gene [275–277].

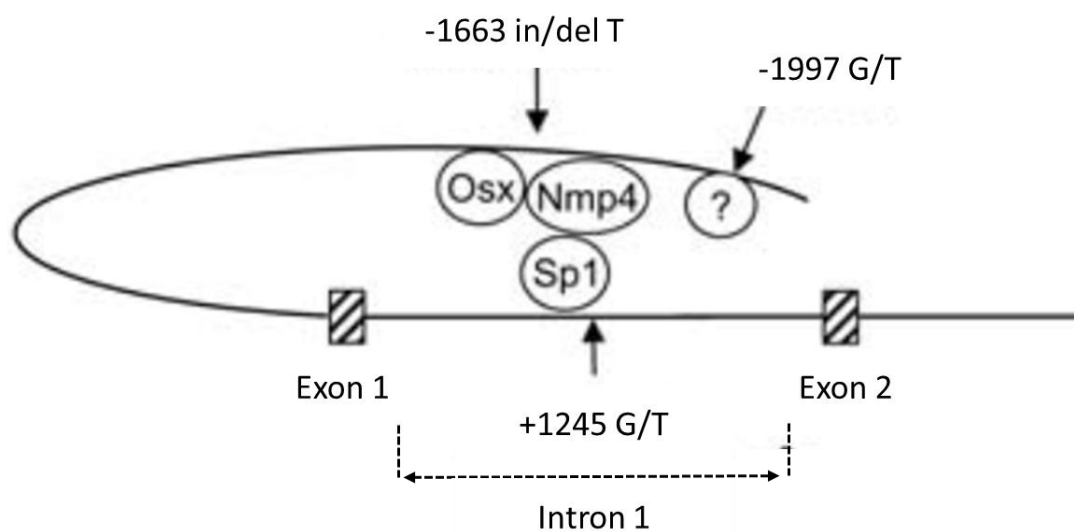


Figure 6.13: Proposed mechanism of DNA-protein binding and the interaction of variants in the 5' flank of the *COL1A1* gene regulated by chromatin looping [Compiled from Jin et al. (2009) [275] with permission from Oxford University Press]

The Sp1, Nmp4 and Osx proteins are transcription factors with binding sites located within the 5' flank of the *COL1A1* gene. It is hypothesised that the rs1800012 (+1245 G>T) variant interacts with a tagged locus, rs11327935 (-1663 indel/T), within the promoter of the gene. This interaction is mediated through chromatin looping and the presence of specific nuclear elements. (-) Base pairs upstream from the transcription start site. (+) Base pairs downstream from the transcription start site. Sp1: Specificity protein-1; Nmp4: Nuclear matrix protein-4; Osx: Osterix.

6.5. CONCLUSION

Based on the hypothesis that the accumulation of evidence at several levels of investigation may support the functional significance of genetic signatures in the predisposition to musculoskeletal soft tissue injuries, this study aimed to investigate the variants described in this thesis at the functional level using several freely available *in silico* tools. The results of these analyses suggest that although not disease causing, several investigated *TNC*, *COL27A1*, *MMP3* and *COL1A1* variants may have an effect at the functional level.

In summary:

- The *TNC* rs1061494 (T>C) variant maps to the sequence motifs identified as putative Myf and MUSCLE_INI_B transcription factor binding sites, in addition to the sequence recognised by the large subunit of RNA polymerase II. Furthermore, the rs1061494 CC genotype demonstrated changes in its predicted pre-mRNA structure, which suggest increased accessibility to the DNA surrounding this locus by regulators of gene function. Therefore, these predicted features provide support for the observed overrepresentation of the CC genotype in the group of cases with chronic AT (Chapter 3), which is in accordance with the increased tenascin-C expression in diseased tendons.
- The *TNC* rs2104772 (T>A) variant was predicted to be deleterious by two independent annotation tools. These predictions support the previous hypothesis of reduced instability of the FNIIID domain structure in the presence of the isoleucine substitution (A allele), which may affect the elastic property of tenascin-C. In accordance, the AA genotype was overrepresented in the group of females with acute ACL ruptures compared to female controls, inferring increased risk for injury (Chapter 3).
- The investigated *COL27A1* variants were predicted to interact with cellular enhancers *in silico*. Due to their proximity to the 5' regulatory region of the gene (<10 kb), they may interact with specific enhancers which are involved in regulating DNA-protein interactions within the chromosomal loop. However, the *COL27A1* gene is still relatively uncharacterised in musculoskeletal soft tissue injuries. These variants will

therefore require further interrogation in larger independent injury cohorts, in addition to being assessed *in vitro*.

- Except for the rs650108 variant, the alleles for the other investigated *MMP3* variants (rs3025058, rs679620 and rs591058) were highly correlated; proposing the linkage of these loci. This level of linkage was confirmed across the 8 kb *MMP3* genomic interval *in silico*. However, the fact that three of the four investigated loci were tagged, in addition to the fact that they were not independently associated with injury risk in two additional populations (Chapter 4), suggests that a more informative, risk-associated haplotype mapping to this gene may still require identification. Furthermore, the proposed interaction between loci in the *MMP3* and *MMP12* genes by chromatin looping, provides support for the investigation of gene-gene interactions in the future. Therefore, future studies should explore an extended interval encompassing all nine *MMP* genes mapping to chr11q22.
- The rs1800012 (G>T) variant localises to the recognition sequence of the Sp1 transcription factor in intron 1 of the *COL1A1* gene and is also proposed to interact with the functional promoter variants, rs1107946 (G>T) and rs11327935 (indel/T). This interaction may be regulated by chromatin looping. Furthermore, the presence of the rare rs1800012 T allele is predicted to result in single strandedness within the pre-mRNA conformation immediately surrounding the locus. This conformation may promote DNA-protein complexes and allele interactions; supporting the observed increase in *COL1A1* expression in the presence of the T- allele. However, this hypothesis would require validation. The rs1800012 TT genotype was overrepresented in the group of controls compared to ACL cases in the meta-analysis conducted in chapter 5. Furthermore, the G-T (rs1107946-rs1800012) inferred haplotype was also associated with reduced risk for acute ATR. These results suggest that an increase in the production of α_1 homotrimeric type I collagen molecules, may reduce the risk of acute musculoskeletal soft tissue injuries.

Considering the high prevalence of musculoskeletal soft tissue injuries, it is reasonable to hypothesise that these injuries are underpinned by a myriad of common, non-lethal variants, each of which contribute modestly to the overall predisposition to injury. These risk-associated signatures may include variants mapping to many genes within the genome,

several of which are still likely to require characterisation in the pathologies of musculoskeletal soft tissue injuries. While we are far from decoding these complex genetic signatures, understanding how chromatin folding occurs may provide insight into the intricate relationship that exists between chromatin structure, gene activity and cellular function. Furthermore, by harnessing the strength of next generation sequencing technologies together with the rapid evolution of *in silico* tools to assess the functional consequence of specific DNA sequence changes, we may further explore the biological mechanisms governing ECM regulation and improve our understanding of injury development.

CHAPTER 7: CONCLUDING REMARKS

7.1. THESIS SUMMARY

To date, the treatment of musculoskeletal soft tissue injuries remains challenging. These injuries are often non-responsive, recurrent and may predispose for other orthopaedic injuries and medical conditions [5]. This may be indicative of the paucity in knowledge regarding the biological mechanisms underpinning the development of these injuries. This problem is further compounded by the fact that there is inter-individual variation in response to training, presentation of symptoms and recovery [9,11]. It is however clear that these injuries are multifactorial in nature, influenced by the complex interaction of several environmental and intrinsic factors [15–17]. A notable intrinsic risk factor for chronic AT, RCT and ACL ruptures is a familial predisposition; supporting the contribution of genetic elements to the profile predisposing for musculoskeletal soft tissue injuries [20,21,181]. In line with this, several DNA sequence variants located within genes encoding structural components and regulators of the ECM have been associated with a range of musculoskeletal soft tissue injuries [18,19]. Taking into consideration the biological functions of the associated genes, it is evident that the underlying pathology involves a complex network of intracellular signalling cascades. Moreover, based on the polygenic nature of these conditions, it is more likely that a myriad of common DNA sequence variations contribute collectively to individual susceptibility to injury development.

To date, most studies investigating genetic risk factors for musculoskeletal soft tissue injuries have adopted the candidate gene approach using a case-control, genetic association design. These studies are hypothesis driven, where the variants prioritised for interrogation are proposed to affect biological processes known to be involved in the development of musculoskeletal injuries or similar phenotypes. In saying this, identifying candidate variants for interrogation within these regions has proved challenging. Furthermore, the ability to replicate the findings of the original studies has varied considerably, a possible indication that other, more informative markers of risk within the candidate genes still require identification.

In recent years, a small number of research groups have adopted GWAS and NGS technologies to identify genetic risk profiles predisposing for musculoskeletal soft tissue injuries [284–287]. Most of these studies identified several variants associated with injury predisposition, all mapping to genes that are plausible candidates based on their biological function but had yet to be prioritised for interrogation using the previous candidate variant approach. However, an important challenge with these strategies is the large sample numbers required to achieve sufficient power to detect significant associations using traditional statistical methods. Furthermore, the cohorts included in these studies need to be phenotypically well-defined. Unfortunately, many studies published to date have utilised public databases to establish their cohorts and consequently, may not represent homogenous groups of well-phenotyped individuals.

Although the cost of high throughput sequencing has steadily decreased over the last decade, these technologies are still prohibitively expensive when large sample sizes are required. This is particularly true for complex phenotypes in which it is likely that the effect size of individual associations will be small. It is for this reason that the current project adopted an integrated omics approach; harnessing the strength of NGS technologies together with the traditional case-control genetic association study to facilitate the interrogation of the genome in the most cost-effective manner.

In the current study, 20 well-phenotyped individuals were selected for whole exome sequencing. By traditional sequencing standards this sample set would be underpowered; however, the resulting data was not analysed using traditional statistical methods. Instead, the data was mined for plausible candidate variants that met a pre-determined allele frequency difference threshold. These variants were characterised using several bioinformatic tools and subsequently genotyped in several injury cohorts. It was for this reason that a crucial project objective involved the selection of individuals representing divergent extremes of the phenotype spectrum for sequencing. To provide proof of concept for this non-traditional method of exploring a sequencing dataset, it was decided that several

genomic regions already implicated in injury predisposition would be prioritised for further interrogation.

Therefore, the overarching aims of this thesis were:

- (i) To further define the molecular signatures of musculoskeletal soft injuries mapping to specific genomic intervals encoding for several structural and regulatory components of the extracellular matrix. The WES dataset was mined for new candidate variants mapping to the *TNC* (chr9q33.1), *COL27A1* (chr9q32-33), *MMP3* (chr11q22) and *COL1A1* (chr17q21.33) genes using a customised, tiered filtering strategy.
- (ii) To determine whether the genetic signatures associated with injury risk in individuals who self-identified to be of White ancestry were similar to those predisposing for injury in individuals who self-identified to be of Coloured ancestry.
- (iii) To functionally annotate the prioritised variants using a host of freely-available, online bioinformatic tools.

Tables 7.1 to 7.4 provide an overview of the novel genetic associations identified in the *TNC*, *COL27A1*, *MMP3* and *COL1A1* genes presented in this thesis. Any previous associations within these genomic regions that have been identified in the injury cohorts prioritised for interrogation in this thesis have also been noted.

Table 7.1: A summary of the associations between the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 3 are presented

Injury cohort	Association	Phenotype	TNC variants			
			rs1061494	rs1138545	rs2104772	rs1061495
Achilles tendon injuries						
SAW	Genotype	AT	CC			
	Haplotype	AT	T		T	
UK	-	-				
Combined (SAW+UK)	-	-				
Combined (SAW+AUS)	Genotype	AT	NI		T [†]	
ACL rupture						
SAW	Genotype	ACL			AA [‡]	
SAC	Haplotype	ACL	T		T	

Green shading indicates an association with reduced risk of injury. Red shading indicates an association with increased risk of injury. Grey shading indicates the lack of an association with the risk of injury. NI: Not investigated in this thesis; AT: Chronic Achilles tendinopathy; ATR: Acute Achilles tendon ruptures; ACL: Anterior cruciate ligament ruptures; NON: ACL injuries with a non-contact mechanism of injury.

[†]Previously reported association of the *TNC* rs2104772 T allele with decreased risk of Achilles tendinopathy in the Combined (SAW and AUS)-Achilles tendon injury cohort [198].

[‡]Independent association of the *TNC* rs2104772 AA genotype with increased risk of ACL ruptures in female participants within the SAW-ACL rupture cohort.

Table 7.2: A summary of the associations between the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 3 are presented

Injury cohort	Association	Phenotype	COL27A1 variants		
			rs2567706	rs2241671	rs2567705
Achilles tendon injuries					
SAW	-	-			
UK	Genotype	AT		GG	
Combined (SAW+UK)	Genotype and allele	ATR		GG and G	
ACL rupture					
SAW	-	-			
SAC	Genotype and allele	NON		GG and G	
	Haplotype	NON	A	G	A
			G	A	T

Green shading indicates an association with reduced risk of injury. Red shading indicates an association with increased risk of injury. Grey shading indicates the lack of an association with the risk of injury. AT: Chronic Achilles tendinopathy; ATR: Acute Achilles tendon ruptures; ACL: Anterior cruciate ligament ruptures; NON: ACL injuries with a non-contact mechanism of injury.

Table 7.3: A summary of the associations between the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 4 are presented

Injury cohorts	Association	Phenotype	MMP3 variants			
			rs3025058	rs679620	rs591058	rs650108
Achilles tendon injuries						
SAW	Genotype	AT	6A6A	GG [†]	CC [†]	AA [†]
	Haplotype	AT		‡ A	T	G
AUS	Haplotype	AT	6A	G	C	G
Combined (SAW+AUS)	-	-				
ACL rupture						
SAW	-	-				

Green shading indicates an association with reduced risk of injury. Red shading indicates an association with increased risk of injury. Grey shading indicates the lack of an association with the risk of injury. AT: Chronic Achilles tendinopathy; ATR: Acute Achilles tendon ruptures; ACL: Anterior cruciate ligament ruptures; NON: ACL injuries with a non-contact mechanism of injury.

[†]Previously reported association of the *MMP3* rs679620 GG genotype, rs591058 CC genotype and rs650108 AA genotype with increased risk for Achilles tendinopathy in the SAW-Achilles tendon injury cohort [206].

[‡] Previously reported *MMP3* haplotype association with decreased risk of Achilles tendinopathy in the SAW-Achilles tendon injury cohort [206].

Table 7.4: A summary of the associations between the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants and the risk of musculoskeletal soft tissue injuries. The independent variant genotype and inferred haplotype associations of chapter 5 are presented

Injury cohorts	Association	Phenotype	COL1A1 variants	
			rs1107946	rs1800012
Achilles tendon injuries				
SAW	-	-		
UK	-	-		
Combined (SAW+UK)	Haplotype	ATR	G	T
ACL rupture				
SAW	Genotype	ACL		TT [†]
Combined (Meta-analysis)	Genotype	ACL		TT
SAC	Haplotype	NON	T	G

Green shading indicates an association with reduced risk of injury. Red shading indicates an association with increased risk of injury. Grey shading indicates the lack of an association with the risk of injury. AT: Chronic Achilles tendinopathy; ATR: Acute Achilles tendon ruptures; ACL: Anterior cruciate ligament ruptures; NON: ACL injuries with a non-contact mechanism of injury.

[†]Previously reported association of the *COL1A1* rs1800012 TT genotype with decreased risk for ACL ruptures in the SAW-ACL rupture cohort [203].

In summary, three of the six candidate variants prioritised from the WES dataset using our customised, tiered filtering strategy were independently implicated in the predisposition of either Achilles tendon injuries, ACL ruptures or both (*TNC* rs1061494, rs2104772 and *COL27A1* rs2241671). Moreover, inferred haplotype analyses implicated larger genomic regions encompassing these variants in combination with the other prioritised loci from the WES dataset. No new candidate variants in the *MMP3* or *COL1A1* genes were prioritised for interrogation from the WES dataset. Several previously associated variants were; however, reassessed in additional injury cohorts. None of these previously implicated variants demonstrated independent associations in any of the investigated injury cohorts in this thesis, except for the inferred association of the *MMP3* rs3025058 variant with the risk of AT in the previously genotyped SAW-Achilles tendon injury cohort. Furthermore, a four-variant *MMP3* inferred haplotype was associated with the risk of AT in the AUS-Achilles tendon injury cohort; however, we must caution against the over-interpretation of these results. This risk-associated haplotype contrasts with that reported for AT by Raleigh et al. (2009) [206]. Therefore, it may be that more informative markers of risk, either located within the *MMP3* gene itself or in other *MMP* genes located within the cluster on chromosome 11q22, may still require identification. With respect to the *COL1A1* gene, the rare rs1800012 TT genotype was associated with reduced risk for ACL ruptures in the meta-analysis conducted in chapter 5. Besides this independent association, inferred haplotype analysis also implicated the 5' regulatory region of the *COL1A1* gene in the predisposition of both ATR and ACL ruptures which is in accordance with a previous observation [183].

Interestingly, several associations were common to the self-identified White and Coloured cohorts providing support for the functional interrogation of these variants as potential modifiers of risk for musculoskeletal soft tissue injuries. Specifically, the *TNC* T-T (rs1061494-rs2104772) inferred haplotype was associated with decreased injury risk in both the SAW-Achilles tendinopathy cohort and the SAC-ACL rupture cohort. Similarly, the rs2241671 GG genotype and G allele were significantly associated with increased injury risk in the self-identified White British and combined (SAW and UK)-Achilles tendon injury cohorts, in addition to the SAC-ACL rupture cohort. Although AT and ACL ruptures are clinically distinct phenotypes, the observed similarities in several injury-associated genetic signatures provides

support for the contribution of the investigated loci to the profiles predisposing for these musculoskeletal soft tissue phenotypes.

7.2. BIOLOGICAL SIGNIFICANCE OF FINDINGS

The use of WES is not without limitations; however, it is a powerful molecular tool with the ability to identify functional variants which may contribute to the heritable predisposition to disease including musculoskeletal soft tissue injuries. Several novel associations have been presented in this thesis. Referring to the Meeuwisse model of injury development (described in Section 1.5) [15], these individual variants are likely to contribute only modestly to an individual's overall predisposition. However, it may be hypothesised that the net effect of an array of markers across the genome, together with other intrinsic factors predispose for injury. If predisposed, extrinsic factors may act on these biological factors and potentially render an individual susceptible to an injury-inciting event [15]. Although the functions of the genes investigated in this thesis are distinct and cannot be confined to a single signalling cascade, there is evidence to suggest that they interact. Therefore, these factors may collectively influence the efficacy of matrix remodelling (Figure 7.1).

Connective tissues adapt dynamically to mechanical loading by either (i) triggering the release of autocrine- or paracrine cell-signalling factors, which consequently alter the transcription rate of specific mechanoresponsive genes or (ii) triggering intracellular signals through intergrin-ligand mechanotransduction [225]. Through these mechanisms, fibroblasts act to maintain tissue homeostasis by changing the content of fibrillar and network-forming collagens, together with other mechanoresponsive ECM proteins. The genes investigated in this thesis are known to encode several of these ECM proteins and although they have been investigated independently, there is evidence which suggests for their interaction.

As previously discussed, mechanical loading is a powerful and direct regulator of *TNC* expression through the activation of the RhoA/ROCK pathway [225,234]. This method of tenascin-C activation is rapid and occurs independently of any known autocrine or paracrine

signalling factors. However, some factors including platelet derived growth factor (PDGF) and TGF β are believed to increase the sensitivity of cells to mechanical stimulation, thereby synergistically enhancing the induction of tenascin-C [225]. Once expressed, the tenascin-C molecule is itself a powerful regulator of ECM modulation, with the ability to influence a wide range of ECM constituents and cells surface proteins [217,218]. These interactions are primarily determined through alternative splicing of the FNIII repeat domains, with the different tenascin-C isoforms showing preferences for specific binding partners, proteolytic cleavage sites and post-translational modifications [221].

Interestingly, MMP3 is known to directly interact with tenascin-C. This endopeptidase is capable of proteolytic degradation of tenascin-C *in vivo*, with cleavage sites located in the alternatively spliced FNIII Δ 3 domain [221,347]. Hence, the larger tenascin-C isoforms are more susceptible to proteolytic fragmentation than the smaller isoforms which lack this specific domain. It has previously been reported that MMP3 activity is important for normal tendon maintenance, especially in highly stressed structures such as the Achilles tendon and supraspinatus tendons [348], where it may contribute to the degradative regulation of constitutively expressed tenascin-C. It is therefore interesting to note that *MMP3* mRNA is significantly downregulated in degenerative tendon samples and the larger tenascin-C isoforms are more abundant [248,321]. Consequently, it may be hypothesised that a reduction in MMP3 may contribute to the altered tenascin-C profile observed in disease tissue [118].

Additionally, the *HSPG2* gene which encodes perlecan, a heparan sulfate proteoglycan (HSPG), was identified by WES in the current study and by Saunders et al. (2016) as a strong candidate for AT [302]. This modular proteoglycan is also known to interact with tenascin-C. The function of tenascin-C is dependent on its incorporation into the ECM; a process which is suggested to be influenced by the presence of HSPGs. A *HSPG* gene knockout cell line lacked the ability to incorporate tenascin-C into the ECM and failed to assemble a fibronectin scaffold [324]. Furthermore, tenascin-C knockout mice exhibit reduced fibronectin deposition. Fibronectin-bound perlecan is known to interact with tenascin-C. In support of this, tenascin-C, perlecan and fibronectin have been observed to co-localise [323,324]. The binding sites of

perlecan are situated within the constitutively expressed FNIII3-5 domains of tenascin-C. This protein interaction is suggested to be enhanced in the smaller tenascin-C isoform, whereas the presence of the alternatively spliced FNIII repeats attenuates perlecan to tenascin-C binding; possibly reducing adhesion in cells expressing the larger tenascin-C isoforms. Thus, it may be that HSPGs are required for the incorporation of the smaller tenascin-C isoform into the ECM which in turn stabilises fibronectin fibrils and the surrounding microenvironment [324]. Conversely, it may be hypothesised that variants located within the binding motifs of these proteins may influence their binding efficacy; altering the incorporation of the tenascin-C glycoprotein into the ECM and consequently effecting matrix remodelling. Based on this evidence, several variants within the *HSPG2* gene have been selected from the list of candidates identified by WES using the tiered filtering strategy and are currently being investigated.

The MMP3 endopeptidase is also known to influence type I collagen production, although these proteins do not directly interact with one another. Besides the degradative effects of MMP3 on ECM constituents, this endopeptidase can activate several other zymogens through cleavage of the propeptide unit [244]. Such is the case with MMP3 activity on a partially processed proMMP1 molecule; cleavage which is critical for the generation of a mature MMP1 peptide [245,349]. The MMP1 endopeptidase subsequently digests the major matrix collagen types I-III. Therefore, with its broad substrate specificity, MMP3 may contribute to the maintenance of tissue homeostasis by orchestrating the timely degradation of several ECM components including the major fibrillar collagens.

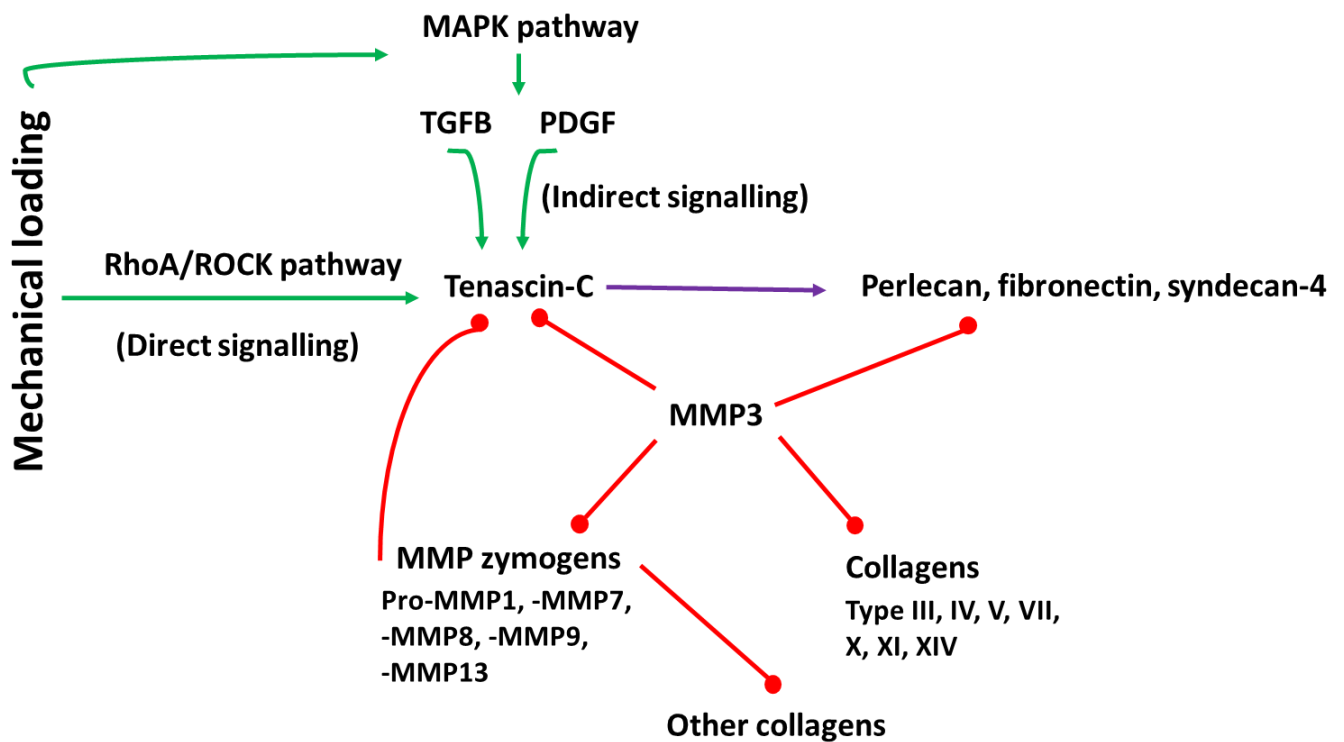


Figure 7.1: Proposed interactions between several proteins encoded by the genes investigated in the studies presented in this thesis

Green: Induces expression; Purple: Interaction of molecules; Red: Cleavage and/or degradation. Mechanical loading is known to directly induce tenascin-C expression via the RhoA/ROCK pathway. Although TGFβ and PDGF may synergistically enhance the induction of tenascin-C, they are not directly responsible for the effects of mechanical stress on this glycoprotein. Expressed tenascin-C isoforms are known to interact with a wide range of molecules. Several interactions have been highlighted in this thesis. Specifically, the small tenascin-C isoform interacts with perlecan within its constitutively expressed FNIII3-5 domains [324], leading to its incorporation into the matrix, whereas the larger tenascin-C isoforms interact with syndecan-4 within the FNIIIA2 domain, attenuating RhoA/ROCK signalling and decreasing focal adhesion formation [223]. Tenascin-C is cleaved by several matrix metalloproteinases including MMP1, MMP2, MMP3 and MMP7 which are believed to contribute to tenascin-C turnover in tissue. Although degradation is important for controlling levels of tenascin-C in tissue, MMP3 may pro-actively impact tenascin-C function by generating tenascin-C isoforms with unique biological functions [221]. The MMP3 cleavage site is located within the alternatively spliced FNIIIA3 domain. MMP3 is also known to activate other MMP from their latent zymogen forms including pro-MMP1 which consequently degrades the major fibrillar collagens [245,349]. RhoA/ROCK: Rho-GTPases and downstream Rho-dependent kinase signalling cascade. MAPK: Mitogen-activated protein kinase signalling cascade. TGFβ: transforming growth factor beta; PDGF: Platelet derived growth factor; MMP: Matrix metalloproteinases; MMP3: Matrix metalloproteinase 3 (Stromelysin 1).

Fibrocartilaginous metaplasia has been reported as a common feature of pathological tendons, although fibrocartilage may present as an adaptive feature in normal tendons in response to compressive forces [321,350]. Whether the development of fibrocartilage is pathological in nature or simply an adaptive response, the pattern of matrix synthesis and the remodelling of the existing matrix is altered [321]. Interestingly, tenascin-C has been implicated in the modulation of chondrocytes and fibroblasts at fibrocartilaginous regions of tendon, such as the OTJ [228]. Furthermore, the expression profile of these cells is typical of cartilaginous tissue with high levels of aggrecan and type II collagen detectable [13,14]. In line with this, collagen type XXVII is highly expressed during cartilage development, with continued expression into adult tissue [239]. Therefore, the *COL27A1* gene may be an interesting candidate for further explorations into the genetic basis of injuries, especially those that are characterised by fibrocartilaginous metaplasia.

Taken together, the complexity of the biological mechanisms underpinning the development of musculoskeletal soft tissue injuries is becoming increasingly apparent. Only four candidate genes were prioritised for interrogation in the current studies, yet all contained variants that were suggested to be implicated in the profile predisposing for injury. These specific genetic signatures will require functional characterisation before they may be considered as true risk conferring variants modulating injury predisposition. To date, several genes have been implicated in the predisposition to musculoskeletal soft tissue injuries [18,19]. Most of these regions have yet to be fine mapped to identify the most informative markers of injury. Furthermore, a significant proportion of the genome remains uncharacterised in the genetic profile predisposing for musculoskeletal soft tissue injuries. Therefore, it is evident that we are still far from fully understanding the genetic basis of musculoskeletal soft tissue injuries including signatures that are likely to span across multiple genes of diverse function, as illustrated from the results presented in this thesis. However, this work, together with other recently published studies using next generation sequencing technologies provides a foundation from where future studies may harness key strengths and build upon important limitations to establish pipelines which are effective in elucidating modulators of injury risk.

7.3. STUDY LIMITATIONS AND STRENGTHS

Like with any research, there are several important limitations but also some key strengths associated with this work. The number of individuals selected for WES is an important limitation (n=20). However, to maximise the amount of information obtained from these samples, the criteria used to select the exemplar cases and controls were highly selective. These individuals were specifically selected because they represented divergent extremes of the phenotype spectrum. The controls were physically active, mature athletes whom had not suffered from any ligament and/or tendon injuries prior to recruitment, whereas the cases were young individuals whom had suffered chronic bilateral and/or recurrent Achilles tendon injuries. It was hypothesised that sequencing only well-phenotyped individuals would increase the chances of identifying genetic signatures which potentially contribute to the genetic landscape of musculoskeletal soft tissue injuries.

A noteworthy limitation regarding the use of WES to explore the genetic basis of disease is its inability to identify variants residing in non-coding regions, in addition to its reduced ability to accurately detect structural and copy number variations. Consequently, all variants located within the intronic and promoter regions of the SAW-CON^{SEQ} and SAW-TEN^{SEQ} participants were not captured. It is therefore highly plausible that variants with important risk-modifying capabilities were not identified using this sequencing approach. However, an extended exome sequencing platform was used which included sequencing all known microRNA genes, UTRs and the exome. As there is still much to be understood regarding the biological mechanisms underpinning the development of musculoskeletal soft tissue injuries; mining only the protein encoding region of the genome for variants was considered an appropriate method for initial interrogation. Furthermore, the fact that these injuries are common phenotypes, it is unlikely that they are underpinned by large structural genetic defects. Although traditional statistical methods were not employed to identify variants from the WES dataset, a customised, tiered filtering strategy was devised based on results from previous genetic association studies and biological inferences. All prioritised candidate variants were subsequently interrogated in several independent injury cohorts with increased power to detect statistically significant effects.

Significant differences were also observed in the physiological and sporting characteristics between the respective diagnostic groups investigated in the studies described in this thesis. This data was however self-reported, and therefore, represents a limitation. The mean age of the cases in the SAW-and UK-Achilles tendon injury cohorts were significantly older at the time of injury compared to control participants at the time of recruitment. This is a confounding factor as the incidence of injuries to the Achilles tendon is known to increase with the advancement in age [84,85]. Tissues of the body including musculoskeletal soft tissues naturally undergo deterioration with age. This is presumably due to alterations in the repair and remodelling process which hinder the ability of tissues to modulate load [85,105]. Consequently, future studies would benefit from recruiting cohorts of individuals whom have sustained Achilles tendon injuries at a relatively young age compared to the age at recruitment of the apparently healthy, non-injured controls. This form of selective sampling was however practiced in the SAW-and SAC-ACL rupture cohorts. Specifically, the controls were significantly older at the time of recruitment compared to the age of the cases at the time of injury. This not only suggested that the cases were potentially biologically predisposed but also ensured that the controls had received adequate exposure to mechanical stressors prior to recruitment.

Similarly, sex has been proposed as an intrinsic risk for Achilles tendon pathology, with a predominance of males suffering chronic AT and sustaining acute ATR [84,98,103]. However, the relationship between sex and injuries of the Achilles tendon remains unclear. The higher number of male cases may not necessarily reflect greater susceptibility but may reflect a greater number of males participating in high-risk sports. With respect to our own Achilles tendon injury cohorts, a significantly greater proportion of AUS-TEN participants were male compared to AUS-CON participants. Unfortunately, no physical activity data was recorded for these individuals at the time of recruitment.

Conversely to the relationship between sex and Achilles tendon injuries, the effect of sex on knee ligament injuries is more evident, with females suggested to be at an increased risk for ACL ruptures [164,351,352]. Several biological and physiologic factors have been proposed to influence female susceptibility to ACL injury but many of these factors have yet to be

confirmed with high levels of certainty [17]. Genetics is one such factor; however, much work is still required in this field of research. Variants within the *COL5A1*, *COL12A1* and *VEGFA* genes have demonstrated associations with the risk of injury in female subgroups [188,192,197,209]. Similarly, the *TNC* rs2104772 variant in chapter 3 was associated with increased risk for ACL ruptures in the female subgroup of the SAW-ACL rupture cohort. However, due to the small female sample size in these sub-analyses, these associations require further validation. Therefore, future studies including case-control genetic association and large prospective-cohort studies should direct their recruitment efforts at obtaining a larger female representation.

Body mass and BMI are two related factors that have also been suggested as potential risk factors for musculoskeletal soft tissue injuries including tendon injuries and ACL ruptures [86–88,90,91]. Therefore, the significant differences noted for these factors between the respective diagnostic groups in the current study is a study limitation. Specifically, the SAW-TEN and SAW-RUP participants were significantly heavier at the time of recruitment compared to their SAW-CON^{AT} counterparts. However, the body mass recorded for cases was that at the time of recruitment and not at the time of initial injury, with many participants anecdotally reporting a body mass increase due to decreased physical activity after injury. Similarly, the SAW-ACL participants were significantly heavier than the SAW-CON^{ACL} participants. However, this may be reflective of the larger male representation in the SAW-ACL group compared to the SAW-CON^{ACL} group. The association between high adiposity and the development of Achilles tendinopathy has been discussed (Section 1.4.1.2). Briefly, obesity (high BMI) may increase the load to which the Achilles is subjected. Moreover, the release of bioactive peptides by adipose tissue may drive persistent, low-grade inflammation, which may ultimately influence musculoskeletal soft tissue remodelling overtime [90]. Similarly, studies have reported the relationship between high BMI and risk of ACL ruptures in female athletes. At an increased body mass, it is suggested that the knee is subjected to greater *in situ* loads during jump-landing manoeuvres, potentially predisposing for injury [353]. However, other studies have failed to replicate this finding [354,355]. Therefore, more research is required to validate this relationship.

Although efforts were taken to recruit participants matched for sports participation, several differences were observed between the diagnostic groups which is an important limitation. Specifically, the participants in the SAW-TEN and SAW-RUP groups reported a significantly greater number of years involved in high impact sports such as medium to long distance running and racket sports compared to the recruited SAW-CON^{AT} participants. Similarly, the percentage of male SAW-ACL participants participating in contact and non-contact jumping sports was significantly greater than the number of SAW-CON^{ACL} participants. Unfortunately, no data describing sporting and medical history was recorded for the participants in the AUS-Achilles tendon injury cohort at the time of recruitment.

It is evident that some of the cohorts investigated were small and were only sufficiently powered to detect large effect sizes. In these instances, cohorts of the same phenotype were combined after covarying for country of birth to increase the statistical power of the analyses. Furthermore, the earlier recruited injury cohorts demonstrated lower genotyping success rates compared to those most recently recruited. In particular, the SAW-and AUS-Achilles tendon injury cohorts were recruited several years prior to initiation of the current project, therefore, the lower genotyping success in these cohorts is likely indicative of the poorer DNA quality and quantity of the samples.

Besides these limitations, the studies described in this thesis also offer several key strengths which should be acknowledged. An integrated omics approach was exploited to further define the genetic risk profile of common musculoskeletal soft tissue injuries. Specifically, the selection of genes for interrogation was hypothesis driven based on biological information. However, WES provided the power to identify new candidate loci within these regions which were subsequently prioritised for further interrogation using a customised, tiered filtering strategy. This strategy highlighted potential candidate variants based on several functional predictions and differences in the distribution of allele frequencies between the SAW-CON^{SEQ} and SAW-TEN^{SEQ} groups. The interrogation of candidate variants using the traditional case-control genetic association study design was also strengthened by the fact that these studies were conducted in line with the recommendations outlined in the STREGA statement [295]. This checklist, adapted from the STROBE statement [356] on reporting observational studies,

provides guidelines for ethical approval, recruitment of participants and laboratory based work.

Due to the varying levels of success associated with repeating genetic association studies, hypothesis testing in multiple populations is important [357]. An association between a variant and injury risk in multiple populations provides compounding evidence for the potential biological relevance of the candidate. Every variant presented in this thesis was genotyped in at least two independent injury cohorts, including one for the investigation of chronic musculoskeletal soft tissue injuries (AT) and the other for that of an acute musculoskeletal soft tissue phenotype (ATR and/or ACL ruptures).

Another strength of this study was the inclusion of affected individuals with clinically, well-phenotyped injuries. All injury diagnoses in the South African cohorts were made by physicians and surgeons actively involved in the recruitment process for ongoing research within our research group. Furthermore, the study inclusion criteria were shared with our project collaborators in the UK and Australia, which ensured that the results were comparable across several independent cohorts. Consistency in the diagnosis of injuries allowed for the pooling of the data from multiple sources to increase the statistical power of the studies.

The unbiased nature of this integrated approach also presented as an important research strength. Specifically, the parameters used to select candidate variants were set *a priori*. In saying this, some parameters of the bioinformatic analysis pipeline may be adjusted to identify other, potentially important modifiers of risk. For example, the current studies presented in this thesis focused specifically on SNPs but future studies may consider mining the WES dataset for other variant types, such as insertion/deletion polymorphisms or copy number variants. In addition, large-scale LD analyses may be conducted across all identified loci, to identify haplotype tagging variants which may be genotyped in the injury cohorts.

Finally, the inclusion of a cohort sample from a population of mixed ancestry, the SAC population, is another novel component of this research. To advance our knowledge in the

molecular signatures predisposing for musculoskeletal soft tissue injury risk, it is imperative that we begin to explore these associations in under-studied populations. These admixed cohorts provide the means to identify robust markers of injury risk, which may ultimately contribute the development of injury prediction models.

Taken together, future studies should aim to recruit larger homogeneous cohorts matched for potential confounders including type of physical activity and level of participation. Emphasis should also be placed on recruiting affected individuals with clinically, well-phenotyped injuries; noting clinical data such as concomitant injuries. In addition, all the relevant information regarding participant characteristics at the time of initial injury should be recorded.

7.4. FUTURE DIRECTIONS

The use of NGS technologies to explore the genetic basis of injury is still in its infancy within the field of sports genomics. To our knowledge, this is the first study to have exploited WES to further define the genetic signatures contributing to the predisposition of chronic AT, along with one other research group which used WES to explore the genetic basis of ACL ruptures [294]. Although tendons and ligaments differ functionally, they are morphologically related [22]. It is therefore not surprising that there are similarities (shared) and differences (unique) in the loci implicated in the clinically distinct phenotypes known to affect these structures [18,19]. This observation has been presented in several of the chapters in this thesis and is likely to become more evident with the progression in genetic explorations. Therefore, the work presented in this thesis provides a foundation for the use of high throughput sequencing technologies to further define the genetic signatures underpinning musculoskeletal soft tissue injuries.

Through the process of mining the WES dataset, several key opportunities for future investigations were identified. The *HSPG2* gene on chromosome 1 was first discussed in

chapter 3. This gene was highlighted by the customised WES approach and was also previously identified as a strong candidate gene for AT using semantic interrogation of a biological database [302]. A total of 74 variants mapping to this gene were identified by WES, of which 14 met the allele frequency difference threshold. The encoded proteoglycan, perlecan, is also known to directly interact with the tenascin-C glycoprotein. Therefore, genotyping the *HSPG2* candidate variants identified by WES and analysing the data in combination with the *TNC* genotyped data may be informative. Besides the *HSPG2* gene, there are still many genomic regions to be explored using the customised, tiered filtering strategy. Approximately 4 500 variants were extracted from the WES dataset meeting the allele frequency difference threshold. Although some of these variants have previously been investigated using genetic association studies, most of these variants map to regions that remain uncharacterised in the aetiology of musculoskeletal soft tissue injuries.

Although some of the variants genotyped in the studies described in this thesis have previously been assessed for their independent effects on gene transcription, it would be interesting to assess the net biological consequence of these variants *in vitro*. As discussed in section 7.2 and presented in Figure 7.1, several of the proteins encoded by the prioritised genes are suggested to interact. With the advancement in molecular techniques, investigating these interactions, in addition to assessing their net biological effect is now possible. One such method involves the use of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9 genome editing technology [358]. This tool has provided a reliable way of making permanent and precise changes to the genome of living cells through RNA-guided, site-specific cleavage. Furthermore, multiple genes may be targeted using a multiplexed CRISPR array, allowing for the simultaneous editing of several sites within the mammalian genome [359]. Currently, such a project is underway in collaboration with researchers from the Wellcome Trust Centre for Cell Matrix Biology (Manchester, UK). Based on the new sequence signatures identified and the functional interrogation of these signatures *in silico* (Chapters 3 and 6), together with the work previously presented by Saunders et al. (2013) [198], this project has focused specifically on exploring the functional significance of specific *TNC* haplotypes *in vitro*. The results of this study, which will include assessing the impact of

these specific *TNC* signatures on gene transcription, cell migration and wound repair capabilities in a human fibroblast cell line, will be submitted for publication in the near future.

From a clinical perspective, the net evidence provided by these studies may potentially have broader implications which extend beyond that of sport and injury genomics. Tenascin-C is highly expressed in cancerous tissues and in tissues undergoing active wound repair, where it is hypothesised to control mechanisms leading to inflammation and in the case of cancer, oncogenic signalling [360]. Therefore, this study may contribute to a larger repository of research, aiming to identify the variants and/or functional haplotypes which regulate tenascin-C function. This knowledge regarding differential tenascin-C expression and function may also be useful in assessing the prognoses of patients after surgery, including the risk of sustaining recurrent musculoskeletal defects. As suggested by Chiquet-Ehrismann et al. (2003) [223], tenascin-C is a very useful marker indicative of early phase connective tissue remodelling. Although its precise function and mechanisms of regulation require further elucidation, it carries great diagnostic and therapeutic potential for many pathologies ranging from cancer to connective tissue disorders.

Another area which may be explored as part of the broader aim to elucidate the genetic signatures underpinning injury predisposition and which is gaining momentum in other fields such as cancer research [343,361], is 3D modelling of chromatin structure *in silico*. The analysis of chromatin structure using chromosome conformation capture (3C) techniques has provided novel insight into the formation of large chromatin loops and has hinted to a possible mechanism through which genes may be regulated. For example, long-range gene regulation may be mediated through the spatial folding of chromosomes to form chromatin loops which deliver regulatory elements to their target genes [362]. Furthermore, this also suggests a mechanism through which variants, apparently separated in the linear genome, may reside nearby and interact within a three-dimensional space. Chapter 6 presented the results from the *in silico* analysis of the variants genotyped in the independent studies presented in chapters 3 to 5 of this thesis. Although only a limited number of functional features were assessed, several intriguing findings emerged. These tools predicted the interaction of several of the prioritised genes and variants with other genomic loci *in cis*. Many

of the additional loci are currently uncharacterised in the profile predisposing for musculoskeletal soft tissue injuries; providing scope for future explorations. In addition, mediated by chromatin looping, several of the genotyped variants were predicted to interact with regulators of gene activity including enhancers and transcription factors. Therefore, this study provided proof of concept for the use of *in silico* tools to characterise new candidate variants identified by NGS or GWAS.

Lastly, investigations should also consider the use of transcriptomic techniques to explore phenotypic changes and inter-individual variability in response to loading. Specifically, by exploring total RNA, mRNA and non-coding RNA profiles specific to musculoskeletal tissues, we may begin to understand the role of different transcripts, different proteomic isoforms and small regulatory RNAs in the underlying mechanisms influencing injury predisposition [363]. Besides DNA sequence variants contributing to injury susceptibility, epigenetic modifications such as DNA methylation, modification of histones and microRNA silencing are likely to be important in modulating injury predisposition [363,364].

Taken together, more research is required to elucidate the true risk conferring variants contributing to the predisposition for musculoskeletal soft tissue injuries. Although many variants have now been associated with the risk of injury using case-control genetic association studies, it is imperative that a repository of evidence is established; demonstrating the potential significance of the candidates at several biological levels. Only once this data is available, can we begin to develop multifactorial models to understand the molecular mechanisms that cause musculoskeletal soft tissue injuries or other related pathologies.

7.5. CONCLUSION

Although the exact biological mechanisms underpinning musculoskeletal soft tissue injuries are not fully understood, the results presented in this thesis provide further support for the involvement of genetic factors in injury predisposition. Using an integrated omics approach, this thesis highlighted several novel genetic signatures associated with injury predisposition, in addition to providing empirical support for other variants previously proposed as risk markers. These findings, together with those from previous studies suggest that it is likely the net effect of many DNA sequence variants, located within a number of genes which collectively contribute to an individual's heritable predisposition to musculoskeletal soft tissue injuries.

It is evident from the number of variants identified by WES in relation to the number mapping to previously implicated genes that much of the genome remains uncharacterised in the pathology of common musculoskeletal soft tissue injuries. Therefore, some variants currently of unknown location and function may possess important risk modifying capabilities. Future studies are therefore required to interrogate the genome using integrated approaches to assist in the elucidation of these sequence signatures. Such investigations would also involve exploring the functional significance of these signatures, a process that would help further inform of the biological mechanisms underpinning the observed risk associations.

The establishment of multi-institutional consortia, such as the Genomics of Injury (GOINg) consortium [365] or the ATHLOME PROJECT [366], should be a priority within this field of research. It is imperative that we facilitate in-depth explorations of the genome, a process that is more likely to be accomplished through combining expertise and resources including the dissemination of data generated through GWAS and sequencing projects. It is only through collaborative efforts and the use of integrated research approaches that we may begin to establish multifactorial models to assess injury predisposition. Large prospective follow-up studies will also be required to evaluate the efficacy of genetic screening tests to accurately predict injury risk. However, it is plausible to assume that these models will be

used by clinicians in the future to develop personalised injury reduction programmes as well as personalised treatment and rehabilitation routines.

APPENDICES**(I) STUDY APPROVAL AND RECRUITMENT FORMS****(A) Original study approval letters from the Human Research Ethics Committee**

- (i) 224/2013 (PhD: 176/2015)
- (ii) 159/2011
- (iii) 086/2005
- (iv) 289/2004
- (v) 370/2013
- (vi) 164/2006

(B) Study information sheets

- (i) South African Achilles tendon injury study
- (ii) British Achilles tendon injury study
- (iii) Australian Achilles tendon injury study
- (iv) South African ACL rupture study

(C) Informed consent forms

- (i) South African Achilles tendon injury study
- (ii) British Achilles tendon injury study
- (iii) Australian Achilles tendon injury study
- (iv) South African ACL rupture study

(D) Medical history and sporting questionnaires

- (i) Achilles tendon injuries
- (ii) ACL ruptures

(E) Achilles tendinopathy clinical diagnostic criteria

(II) MATERIAL AND METHODS

(A) Protocol for DNA extraction from whole blood

(B) DNA quality assessment for whole exome sequencing

(C) Sanger sequencing oligonucleotide primer pair sequences

(D) Allelic discrimination assays

TNC: (i) rs1061494, (ii) rs1138545, (iii) rs2104772 and (iv) rs1061495

COL27A1: (v) rs2567706, (vi) rs2241671 and (vii) rs2567705

MMP3: (viii) rs3025058, (ix) rs679620, (x) rs591058 and (xi) rs650108

COL1A1: (xii) rs1107946 and (xiii) rs1800012

(III) SUPPLEMENTARY RESULTS

(A) Chapter 3: Profiling the *TNC* and *COL27A1* genes in musculoskeletal soft tissue injuries

(B) Chapter 4: The investigation of variants within the *MMP3* gene and musculoskeletal soft tissue injuries

(C) Chapter 5: Functional *COL1A1* variants as modifiers of risk for acute musculoskeletal soft tissue injuries

(I) STUDY APPROVAL AND RECRUITMENT FORMS

(A) Original study approval letters from the Human Research Ethics Committee

i.

UNIVERSITY OF CAPE TOWN



Faculty of Health Sciences
Human Research Ethics Committee
Room E52-24 Groote Schuur Hospital Old Main Building
Observatory 7925
Telephone [021] 406 6338 • Facsimile [021] 406 6411
e-mail: shurecta.thomas@uct.ac.za
Website: www.health.uct.ac.za/research/humanethics/forms

02 April 2013

HREC REF: 224/2013

Prof M Collins
Sport Science Institute
ESSM
Human Biology

Dear Prof Collins

PROJECT TITLE: INVESTIGATION OF THE MATRIX METALLOPROTEINASE GENE CLUSTER ON CHROMOSOME 11 IN CHRONIC ACHILLES TENDINOPATHY (Sub-study Linked to 159/2011)

Thank you for submitting your request to the Faculty of Health Sciences Human Research Ethics Committee for review.

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

Approval is granted for one year till the 15th April 2014

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.

(Forms can be found on our website: www.health.uct.ac.za/research/humanethics/forms)

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please quote the HREC. REF in all your correspondence.

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, FHS HUMAN ETHICS

Federal Wide Assurance Number: FWA00001637.

Institutional Review Board (IRB) number: IRB00001938

This serves to confirm that the University of Cape Town Human Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Convention on Harmonisation Good Clinical Practice (ICH GCP) and Declaration of Helsinki guidelines.

The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.

s.thomas



UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee



Room E52-24 Old Main Building
 Groote Schuur Hospital
 Observatory 792

Telephone [021] 6501236 • Facsimile [021] 406 641

Email: shahieda.amardien@uct.ac.za

Website: www.health.uct.ac.za/fhs/research/humanethics/form

26 March 2015

HREC/REF: 176/2015

A/Prof A September

ESSM, Human Biology
 SSISA Building
 1 Boundary Road
 Newlands
 7700

Dear A/Prof September

Project Title: THE APPLICATION OF NEXT GENERATION SEQUENCING TECHNOLOGIES AND BIOINFORMATIC TOOLS TO FURTHER CHARACTERISE THE GENETIC RISK PROFILE OF MUSCULOSKELETAL SOFT TISSUE INJURIES (PhD-CANDIDATE-A Gibbon) sub-study linked to 224/2013

Thank you for submitting your study to the Faculty of Health Sciences Human Research Ethics Committee (HREC) for review.

It is a pleasure to inform you that the HREC has **formally approved** the above mentioned study

Approval is granted for one year until the 30 March 2016.

Please submit a progress form, using the standardised Annual Report Form, if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.

We acknowledge that the following student:-Andrea Gibbon is also involved in this project.

Please note that the on-going ethical conduct of the study remains the responsibility of the principal investigator.

Please quote the HREC REF in all your correspondence.

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, HSF HUMAN ETHICS

HREC REF 176/2015

ii.



UNIVERSITY OF CAPE TOWN

Faculty of Health Sciences
Human Research Ethics Committee
Room E52-24 Groote Schuur Hospital Old Main Building
Observatory 7925
Telephone [021] 406 6626 • Facsimile [021] 406 6411
e-mail: shuretta.thomas@uct.ac.za

28 April 2011

Sent via internal mail

HREC REF: 159/2011

A/PROF M COLLINS,
HUMAN BIOLOGY
SPORT SCIENCE INSTITUTE
3RD FLOOR

Dear A/PROF COLLINS,

PROJECT TITLE: THE IDENTIFICATION OF GENETIC RISK FACTORS UNDERLYING ACHILLES TENDON INJURIES: A REPEATABILITY STUDY.

Thank you for submitting your new study to the Faculty of Health Sciences Human Research Ethics Committee

It is a pleasure to inform you that the Ethics Committee has formally approved the above-mentioned study.

Approval is granted until 30 April 2012

Please submit an annual progress report (FHS016) if the research continues beyond the expiry date. Please submit a brief summary of findings if you complete the study within the approval period so that we can close our file.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please quote the HREC REF in all your correspondence.

Yours sincerely

A/PROF MARC BLOCKMANCHAIRPERSON, FHS HUMAN ETHICS

Federal Wide Assurance Number: FWA00001637,
Institutional Review Board (IRB) number: IRB00001938

This serves to confirm that the University of Cape Town Human Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Convention on Harmonisation Good Clinical Practice (ICH GCP) and Declaration of Helsinki guidelines.

The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.

iii.

UNIVERSITY OF CAPE TOWN



Research Ethics Committee
E53 Room 44.1, Old Main Building
Groote Schuur Hospital, Observatory,
7925
Queries : Xolile Fula
Tel : (021) 406-6492 Fax: 406-6411
E-mail : Xfula@curie.uct.ac.za

21 February 2005

REC REF: 086/2005

Dr M Collins
Human Biology

Dear Dr Collins

THE GENETIC BASIS OF TENDINOPATHY

Thank you for submitting your study to the Research Ethics Committee for review.

It is a pleasure to inform you that the Ethics Committee has formally approved the above-mentioned study on the 10th February 2005

The contents are noted and added to our files.

Please quote the REC. REF in all your correspondence

Yours sincerely

PROF T. ZABOW
CHAIRPERSON

A handwritten signature in dark ink, appearing to be 'T. Zabow'.

iv.

UNIVERSITY OF CAPE TOWN



Research Ethics Committee
E53 Room 44.1, Old Main Building
Groote Schuur Hospital, Observatory,
7925
Queries : Xolile Fula
Tel : (021) 406-6492 Fax: 406-6411
E-mail : Xfula@curie.uct.ac.za

15 September 2004

REC REF: 289/2004

Dr M Collins
Human Biology
Exercise Science & Sports Medicine

Dear Dr Collins

**THE GENETIC BASIS OF ACHILLES TENDON PATHOLOGY IN AN AUSTRALIAN
POPULATION**

*Thank you for your letter to the Research Ethics Committee dated 13
September 2004.*

*It is a pleasure to inform you that the Research Ethics Committee has approved
the amended English Patient Information Leaflet and Informed Consent Sheet
dated 13 September 2004.*

Please quote the REC. REF in all your correspondence

Yours sincerely

PROF. T. ZABOW
CHAIRPERSON

A handwritten signature in blue ink, appearing to be 'T. Zabow'.

V.

UNIVERSITY OF CAPE TOWN



Faculty of Health Sciences
 Faculty of Health Sciences Human Research Ethics Committee
 Room E52-24 Groote Schuur Hospital Old Main Building
 Observatory 7925
 Telephone [021] 406 6338 • Facsimile [021] 406 6411
 e-mail: sumayah.ariiefdien@uct.ac.za
www.health.uct.ac.za/research/humanethics/forms

06 August 2013

HREC REF: 370/2013

Dr M Posthumus
 ESSM
 Sports Science Institute
 Newlands

Dear Dr Posthumus

PROJECT TITLE: THE IDENTIFICATION OF GENETIC SUSCEPTIBILITY LOCI UNDERLYING ACHILLES TENDON PATHOLOGY: A REPEAT STUDY IN A BRITISH COHORT

Thank you for your letter dated 29 July 2013, addressing the issues raised by the Human Research Ethics Committee.

It is a pleasure to inform you that the HREC has **formally approved** the above mentioned study.

Approval is granted for one year till the 28 August 2014.

Please submit a progress form, using the standardised Annual Report Form, if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.

However, in future please make the HREC aware of important issues such as the cohort having already been recruited. Such information is essential to facilitate the HREC review process.

Please note that the on-going ethical conduct of the study remains the responsibility of the principal investigator.

Please quote the REC. REF in all your correspondence.

Yours sincerely

PROFESSOR M/BLOCKMAN
CHAIRPERSON, HSF HUMAN ETHICS

Federal Wide Assurance Number: FWA00001637.
 Institutional Review Board (IRB) number: IRB00001938
 sAriefdien

vi.



UNIVERSITY OF CAPE TOWN

Health Sciences Faculty**Research Ethics Committee**

Room E52-24 Groote Schuur Hospital Old Main Building

Observatory 7925

Telephone [021] 406 6338 • Facsimile [021] 406 6411

e-mail: pr.watson@uct.ac.za

26 April 2006

REC REF: 164/2006

Dr M Collins
Human Biology

Dear Dr Collins

PROJECT TITLE: THE COL5A1 AND TNC GENES AND THEIR ASSOCIATION WITH ANTERIOR CRUCIATE LIGAMENT INJURIES

Thank you for submitting your study to the Research Ethics Committee for review.

It is a pleasure to inform you that the Ethics Committee has **formally approved** the above-mentioned study on the 21 April 2006.

This serves to confirm that the University of Cape Town Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Convention on Harmonisation Good Clinical Practice (ICH GCP) and Declaration of Helsinki guidelines.

The Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines *Re: Note for Guidance on Good Clinical Practice* (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.**Please quote the REC. REF in all your correspondence.**

Yours sincerely

DR. M. BLOCKMAN
CHAIRPERSON, HSE HUMAN ETHICS

lanjedi

(B) Study information sheets

i.

**Department of Human Biology**

UCT/MRC RESEARCH UNIT FOR EXERCISE SCIENCE & SPORTS MEDICINE
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 650 4561
 Fax: + 27 21 686 7530

THE GENETIC BASIS OF EXERCISE-INDUCED CHRONIC TENDON PATHOLOGY

Although there is a high incidence of tendon overuse injuries as a result of participation in exercise and sporting activities, the cause(s) of these injuries are poorly understood. Some researchers have suggested that there is a genetic component to exercise-induced tendon injuries. In an attempt to determine whether there is a genetic basis for tendon pathology, we are interested in studying whether certain genes are associated with chronic tendinopathies. This project is being done in collaboration with the UCT/MRC Research Unit for Exercise Science and Sports Medicine within the Department of Human Biology and the Division of Human Genetics within the Department of Clinical Laboratory Sciences at the University of Cape Town.

You will be required to visit the Sports Science Institute of South Africa (SSISA) in Boundary Road, Newlands. During the visit, which should take at least 30 minutes, you will be asked to donate 5 ml (1 teaspoon) of a blood sample for DNA analysis. You will also be required to complete personal particulars, sporting details, medical history and stretching and warm up questionnaires. At a later stage, some participants will be asked to visit a doctor (radiologist) for a tendon scan at no cost to themselves.

All the information retrieved from this study will be treated with the strictest confidentiality and will be used only for scientific research purposes. Your name and personal particulars will not be released under any circumstances and all data will be analysed anonymously. Your DNA sample will be destroyed on completion of the study on the genetic basis of tendon pathology. You are also free to request that your DNA sample be destroyed before the completion of the study.

If you are part of the tendon pathology group, we would appreciate it if you could help us by recruiting two other people of same (or similar) age whom you know and who has trained without suffering any tendon injuries for the control group.

We will keep you informed about the outcomes of this study and look forward to working together with you. If you have any questions about this study, please feel free to contact us at:-

Dr. Malcolm Collins, PhD
 (021) 650 4574
mcollins@sports.uct.ac.za

Prof. Martin Schwelinius, MBChB, MD
 (021) 650 4576
mschwell@sports.uct.ac.za

Colleen Saunders, MSc student
 (021) 650 4569
csanders@sports.uct.ac.za

ii.

PARTICIPANT INFORMATION

Background

Thank you for showing an interest in this new study that is going to be conducted at The University of Northampton. The study is being carried out by Dr Stuart Raleigh who is a Reader and a researcher in Molecular Biology, Professor William Ribbans who is a consultant orthopaedic surgeon and Mr Louis El Khoury who is a researcher and PhD student. All three are based at the University of Northampton's School of Health and Professor Ribbans is also based at the County Clinic, Northampton. The contact details of the team are found towards the end of this leaflet. Please feel free to contact them at any time if you require any additional information.

What is the study about?

We are trying to establish whether certain individuals carry particular versions of genes (known as alleles) that may increase their risk of Achilles tendon problems. We are particularly interested to learn more about the types of Achilles tendon problems called Achilles tendinopathy (swollen, painful tendon) and Achilles tendon rupture. The study is about Achilles tendon problems in humans – we never use animals – and we would greatly value you as a volunteer.

Why are we interested?

We are interested because many people with relatively active lifestyles as well as professional athletes experience Achilles tendon problems during or after sports activities. This can sometimes lead to long term disability. In some cases the injury needs surgery and some individuals are reluctant, or have difficulty returning to a sport, or physical activity, they once enjoyed. This research might eventually lead to a way of reducing the number of people who get tendon problems in the future or to better treatment for those already affected.

What's involved if I participate?

We would like individuals to donate a small sample of their spit (saliva) for DNA analysis. We only want about 2 ml of your saliva (the same as one teaspoonfull) which can be taken from simply spitting into a tube. The procedure would take about 3 minutes to perform. You don't even have to attend a clinic. We can send you the tube to spit into. We would also like you to answer some background questions and details about your health and any exercise that you do. If you would like to take part, please complete, sign, and return the consent form in the paid post envelope and we will send you a questionnaire and a saliva collection tube. The total time taken to collect the saliva sample and complete the questionnaire is about 10-15 minutes.

What will happen to my saliva sample?

Your saliva sample will ONLY be used for Achilles tendon studies. Once we receive your sample, it will be labelled with a unique number only and stored in a secure freezer at the University of Northampton's Park campus. We will isolate DNA from your saliva sample. We will then compare the genes between various groups of individuals to establish whether samples from

people who have had Achilles tendon problems contain different versions compared to people who don't have tendon problems. We would like to investigate many genes from your sample and in order for us to do this we would like to retain your sample indefinitely. There may be occasions when we might want to send a small amount of your sample to scientists that we do joint studies with. This would only be done if they could help us with our study. However, if this does happen the sample they receive will only have the number on it and it would not be possible for you to be identified.

Will my sample be used for studies investigating any other disease?

No. The genes that we are interested in analysing are only relevant to musculoskeletal soft tissue injuries (tendon injuries). It will not be possible for us to investigate other illnesses such as: cancer and cardiovascular diseases.

Will the information I provide be kept private?

Yes. We will not send any of your private contact details to any other party. Your personal contact details will only be seen by the three members of the study team. Indeed your personal contact details will be stored securely in a locked filing cabinet at the university and on a password protected computer. The data will not be on any networked computer drives. To make things even more secure your saliva sample and the DNA that we isolate from it will only be labelled with a number. It would not be possible for any person, other than the study team, to identify you from the number that your sample will be allocated.

How long will you be keeping my personal information?

We would like to retain your personal information for an indefinite amount of time. Nevertheless, if you want to end your participation in this project at any time we will immediately destroy your personal information as well as your DNA sample. However any data obtained from the analysis of your sample up until that point will remain part of the research and will be included in any published work.

Why do you need information about my previous medical history?

We need to ask these questions because certain ailments that you may have had in the past might be important to Achilles tendon studies. If we don't have the data we could make a false conclusion in our study.

Why do you need to know about my lifestyle?

This is because it may be lifestyle factors that are most important in causing people to have Achilles problems. Providing this information will help us to establish whether this is true or not.

Why do you need to know about my ethnicity?

This is because certain genetic factors might be different in people of different ethnicity. Having details about a person's ethnicity can help us build a more accurate assessment of the role of genes in causing any Achilles tendon problems. Without details on ethnicity the data we generate might be flawed.

Why do you need my name and contact details?

This is for two main reasons. Firstly we may want to invite you to be part of some future study and might send you a letter. Of course, it would be entirely up to you whether or not you replied to any future invites. You can always send us a request not contact you for any future studies. Secondly, very rarely, samples that are provided for a study can be accidentally damaged during the laboratory procedure. If this does happen we would like to think that we could contact you again for a replacement sample. If we were not able to contact you then the valuable sample that you kindly donated would have been lost from the investigation and this might lead to less accurate scientific results.

What happens if I want to be involved but later change my mind?

You simply contact us and let us know. We will then immediately destroy any sample and information that you have donated to us.

Will I be able to obtain information about my own DNA sample?

The study that we are conducting is aimed at understanding how genetic factors affect the Achilles tendon by studying a group of people. It will be difficult to predict any risk on an individual basis. This is why this study is not designed to test individual genetic status. For this reason we do not plan to give back individual results to participants. If you require addition information about this, please contact us on the numbers or e-mails provided on the last page.

How can I obtain information about the progress of the research?

After roughly 18 months, information about the preliminary results of the study will be posted on the University of Northampton School of Health website. For more information please visit:

(www.northampton.ac.uk/info/20038/school-of-health)

What will happen to the data obtained from this study?

We would like to publish this in scientific journals. The publications would never reveal any participant identity or contact details.

What benefit do I get from being involved in this study?

Many scientific studies rely on sample donations from the general population. By taking part in this study you will help to further our understanding of why some people, particularly active sportspeople, develop Achilles tendon problems. Eventually the findings may help people to be treated for the condition. Furthermore, it might be possible that in the future we could identify those most at risk of developing the problem. If this is the case we could educate people about the suitability of sports that they do and advise them on their risk. The finding may also be used to enhance our understanding of how human tendons work under stressed conditions.

We thank you for reading the above information. If you feel that you would like to be involved then please complete and return the consent form on the next page. Please remember that if you have any questions then please contact the study team who will be very pleased to talk you through any aspect of this study.

Contact Details

Louis El Khoury
 KN 101 Knowledge Exchange
 University of Northampton
 Boughton Green Road
 Northampton – UK
 NN2 7AL
 Tel: 01604 892512
 01604 892003
louis.elkhoury@northampton.ac.uk

Dr Stuart Raleigh
 School of Health
 Brampton Building
 Boughton Green Road
 Northampton – UK
 NN2 7AL
 Tel: 01640 892306
stuart.raleigh@northampton.ac.uk

Prof William Ribbans
 The County Clinic
 57 Billing Road
 Northampton – UK
 NN1 5DB
 Tel: 01604 795414
wjribbans@uk-consultants.co.uk

iii.



*Department of Human Biology
University of Cape Town
South Africa*

16 July 2004

STUDY ON THE GENETIC BASIS OF TENDON PATHOLOGY

The Musculoskeletal Research Centre at La Trobe University in Melbourne, Australia, in collaboration with the UCT/MRC Research Unit for Exercise Science and Sports Medicine at the University of Cape Town (UCT) in Cape Town, South Africa, are currently studying the genetic basis of chronic tendon pathology in the Australian population.

Studies have suggested that some individuals have a genetic predisposition to tendon injury and that genes (those traits which you inherit from your parents), such as Type V collagen (COL5A1) and Tenascin C (TNC), which encode for important components of tendons are associated with tendon pathology.

The aim of this study is to determine whether the TNC, the COL5A1 or other similar genes are associated with chronic tendon pathology.

In order to participate in this study, you will be required to donate five millilitres of venous blood after giving written consent. The blood sample will be used for the extraction and analysis of genetic material (DNA). The extracted DNA will be sent to UCT in South Africa for analysis. The samples will be shipped to and analysed by UCT anonymously. The DNA will be genotyped (analysed) for variations (polymorphisms) within the COL5A1, TNC and other candidate genes believed to be associated with tendon pathology. You will be required to complete a number of questionnaires regarding personal particulars, sporting participation and medical history, as well as a stretching and warm up questionnaire. In addition, you will be required to visit a doctor (radiologist) at a later stage for a tendon(s) scan free of charge. All the information collected during the study will be treated with the strictest confidentiality and will only be used for scientific research purposes. Your name and personal particulars will not be released under any circumstances and all the data obtained will be analysed anonymously.

If you would like to participate in the study and/or obtain any additional information, please contact Dr Jill Cook, on phone: 9479 5789, or e-mail: J.Cook@latrobe.edu.au.



Department of Human Biology
 UCT/MRC RESEARCH UNIT FOR EXERCISE SCIENCE & SPORTS MEDICINE
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 650 4561
 Fax: + 27 21 686 7530

THE GENETIC BASIS OF EXERCISE-INDUCED LIGAMENT INJURY

Although there is a high incidence of ligament injuries as a result of participation in exercise and sporting activities, the cause(s) of these injuries are poorly understood. Some researchers have suggested that there is a genetic component to exercise-induced ligament injuries. In an attempt to determine whether there is a genetic basis for ligament pathology, we are interested in studying whether certain genes are associated with ligament injuries. This project is being done in collaboration with the UCT/MRC Research Unit for Exercise Science and Sports Medicine within the Department of Human Biology of the University of Cape Town. You will be required to visit the Sports Science Institute of South Africa (SSISA) in Boundary Road, Newlands. During the visit, which should take 15 minutes, you will be asked to donate 5 ml (1 teaspoon) of a blood sample for DNA analysis. You will also be required to complete personal particulars, sporting details, medical history and stretching and warm up questionnaires.

All the information retrieved from this study will be treated with the strictest confidentiality and will be used only for scientific research purposes. Your name and personal particulars will not be released under any circumstances and all data will be analysed anonymously. Your DNA sample will be destroyed on completion of the study on the genetic basis of ligament injury. You are also free to request that your DNA sample be destroyed before the completion of the study.

If you are part of the ligament pathology group, we would appreciate it if you could help us by recruiting two other people of same (or similar) age whom you know and who has trained without suffering any ligament or tendon injuries for the control group. We will keep you informed about the outcomes of this study and look forward to working together with you. If you have any questions about this study, please feel free to contact us at:

Dr Alison September, PhD

(021) 650 4559

alison.september@uct.ac.za

Dr Mike Posthumus, PhD

(021) 650 4572

michael.posthumus@uct.ac.za

Prof Malcolm Collins, PhD

(021) 650 4574

malcolm.collins@uct.ac.za

(C) Informed consent forms

i.

**Department of Human Biology**

UCT/MRC RESEARCH UNIT FOR EXERCISE SCIENCE & SPORTS MEDICINE
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 650 4561
 Fax: + 27 21 686 7530

GENETIC BASIS OF EXERCISE-INDUCED CHRONIC TENDON PATHOLOGY**INFORMED CONSENT**

I, the undersigned, have been fully informed about the UCT/MRC Research Unit for Exercise Science and Sports Medicine within the Department of Human Biology and the Division of Human Genetics within the Department of Clinical Laboratory Sciences at the University of Cape Town's study on the genetic basis of exercise induced chronic tendon pathology. I have agreed to donate five millilitres of venous blood which will be used for the extraction and analysis of genetic material (DNA). I have also agreed to complete personal particulars, sporting participation, medical history, stretching and warm up questionnaires and understand that all the information that is collected during the study will be treated with the strictest confidentiality and will only be used for scientific research purposes. I also understand that my name and personal particulars will not be released under any circumstances and that all data will be analysed anonymously.

If requested, I am also prepared to visit a doctor (radiologist) at a later stage for a tendon scan at no cost to myself (please delete this sentence if not applicable). If requested, I am also prepared to visit the SSISA for measurements to determine musculo-tendinous stiffness.

I agree to participate in the study and I have been informed that I will be free to withdraw from the study at any time if I so wish. I understand that my DNA sample will be destroyed on completion of the study on the genetic basis of tendon pathology. I also understand that I will be free to request that my DNA sample be destroyed before the completion of the study.

FULL NAME OF SUBJECT: _____

SUBJECT'S SIGNATURE: _____

DATE: _____

INVESTIGATOR: _____

INVESTIGATOR'S SIGNATURE: _____

ii.



CONSENT FORM

If you would like to participate in this study then please complete the consent form below. If you have any further questions relating to the study, or if there is something that you wish to discuss with us then please contact us at the addresses provided above.

Please tick the box that corresponds to your choice.

I have read the study information sheet and understand what is involved.

Yes ☐ No ☐

I understand that my sample will be analysed anonymously and my contact details kept confidential. My contact details will only be used by the researcher if my sample is accidentally damaged. It would be up to me whether I provided another sample.

☐ ☐

I understand that I can withdraw my participation at any time and that my sample and information would be destroyed upon my request.

☐ ☐

I am willing to participate in this project.

☐ ☐

I would be interested in receiving information about future studies. My agreement to this would mean my contact details would be retained securely and for this purpose alone.

☐ ☐

Signed:

Date:

Name (Print)
Address

Phone Number

E-mail:

We appreciate your time and your willingness to participate in this study.

Participant number (to be entered by the research team)

iii.

13TH JULY 2004

GENETIC BASIS OF TENDON PATHOLOGY

INFORMED CONSENT

I, (the participant), have been fully informed about this study on the genetic basis of tendon pathology to be conducted by the Musculoskeletal Research Centre at La Trobe University in Melbourne, Australia, in conjunction with the UCT/MRC Research Unit for Exercise Science and Sports Medicine at the University of Cape Town in Cape Town, South Africa.

I have agreed to donate five millilitres of venous blood, which will be used for the extraction and analysis of genetic material (DNA), and will be taken by a registered medical physician or nurse. The DNA will only be used for scientific research purposes relating to the genetic basis of tendon pathology only. I have also agreed to complete personal particulars, sporting participation, medical history, stretching and warm up questionnaires and understand that all the information that is collected during the study will be treated with the strictest confidentiality and will only be used for scientific research purposes. I also understand that all data will be analysed anonymously and my DNA sample will be destroyed on completion of the study.

I understand that the DNA extracted from the donated blood sample will be sent to UCT in South Africa for analysis. I understand that the DNA samples will be shipped to and analysed by UCT anonymously. I understand that the DNA will be genotyped (analysed) for variations (polymorphisms) within the COL5A1 and TNC genes, as well as other similar genes relating to the genetic basis of tendon pathology only.

In addition, I am also prepared to visit a doctor (radiologist) at a later stage for a tendon(s) scan at no cost to myself.

I understand that whilst there is no direct benefit to myself, if a genetic predisposition for tendinopathy can be established, then future generations will be able to establish their risk for this condition. This may allow better prevention and treatment options in the future.

I have read (or, where appropriate, have had read to me) and understood the information about this study, and any questions I have asked have been answered to my satisfaction. I agree to participate in the study, realising that I have the right to request that my DNA sample be destroyed at anytime

and, further, to demand that data arising from my participation is not used in the research project provided that this right is exercised within four weeks of the completion of my participation in the project. I agree that research data provided by me or with my permission during the project may be included in a thesis, presented at conferences and published in journals on the condition that neither my name nor any other identifying information is used.

Any questions regarding this project may be directed to the Chief Investigator: **Dr Jill Cook**, of the **School of Physiotherapy** on telephone number **9479 5789**.

If you have any complaints or queries that the investigator has not been able to answer to your satisfaction, you may contact the Ethics Liaison Officer, Human Ethics Committee, La Trobe University, Victoria, 3086, (ph: 03 9479 1443, e-mail: humanethics@latrobe.edu.au).

Name of Participant: _____

Signature: _____

Date: _____

Name of Researcher: _____

Signature: _____

Date: _____

iv.



Department of Human Biology

UCT/MRC Research Unit for Exercise Science & Sports Medicine
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 850 4581
 Fax: + 27 21 686 7530

GENETIC BASIS OF EXERCISE-INDUCED LIGAMENT INJURY INFORMED CONSENT

I, the undersigned, have been fully informed about the UCT/MRC Research Unit for Exercise Science and Sports Medicine within the Department of Human Biology of the University of Cape Town's study on the genetic basis of exercise induced chronic ligament pathology. I have agreed to donate five millilitres of venous blood or a Buccal mouthwash sample, which will be used for the extraction and analysis of genetic material (DNA). I have also agreed to complete personal, sporting participation, medical history, stretching and warm up questionnaires and understand that all the information that is collected during the study will be treated with the strictest confidentiality and will only be used for scientific research purposes. I also understand that my name and personal particulars will be not released under any circumstances and that all data will be analysed anonymously.

I agree to participate in the study and I have been informed that I will be free to withdraw from the study at any time if I so wish. I understand that my DNA sample will be destroyed on completion of the study on the genetic basis of ligament pathology. I also understand that I will be free to request that my DNA sample be destroyed before the completion of the study.

I understand that the DNA will be genotyped (analysed) for variations (polymorphisms) within genes relating to the genetic basis of ligament injuries. I understand that whilst there is no direct benefit to myself, if a genetic predisposition for ligament injuries can be established, then future generations will be able to establish their risk for this condition. This may allow better prevention and treatment options in the future. I understand that I will receive the overall results of the study. I have read (or where appropriate, have had read to me) and understand the information about this study, and any questions I have asked have been

answered to my satisfaction. I agree to participate in the study, realising that I have the right to request that my DNA sample be destroyed at any time. I agree that research data provided by me or with my permission during the project may be included in a thesis, presented at conferences and published in journals on the condition that either my name not any other identifying information is used.

FULL NAME OF PARTICIPANT: _____

PARTICIPANT'S SIGNATURE: _____

DATE: _____

INVESTIGATOR: _____

INVESTIGATOR'S SIGNATURE: _____

(D) Medical history and sporting questionnaires

i.

**Department of Human Biology**

UCT/MRC RESEARCH UNIT FOR EXERCISE SCIENCE & SPORTS MEDICINE
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 650 4561
 Fax: + 27 21 686 7530

GENETIC BASIS OF TENDON INJURY QUESTIONNAIRES

A. PERSONAL PARTICULARS			
Surname			
First Name			
Postal Address			
	Code		
E-mail address	Phone (day time)		
Date of birth	Y Y Y Y / M M / D D	Cell	
Height (cm)	Gender	Male ☐	Female ☐
Weight (kg)			
Ethnic group (Only Required and Used for Research Purposes)	Black/African ☐	White ☐	Indian ☐
	Mixed Ancestry (Coloured) ☐	Asian ☐	Other ☐
Ancestry: Tribal or national background (eg Xhosa, Dutch, Zulu, German, Italian)	Father	Unknown ☐	
	Mother	Unknown ☐	
Country of Birth	Nationality		
Dominant Hand	Left ☐ Right ☐ Both ☐	Dominant Leg	Left ☐ Right ☐ Both ☐
Smoker	Yes (Current) ☐	Yes (Ex smoker) ☐	No, never ☐
	If yes, Number of years _____		If stopped, when _____
	If yes, number per day _____		
Do you know your blood group?	Yes ☐	A ☐ B ☐ AB ☐ O ☐	
	No ☐	Rh Pos ☐ Rh Neg ☐	

Genetic Basis of Tendon Injury Questionnaires

The University of Cape Town is committed to policies of equal opportunity and affirmative action
 which are essential to its mission of promoting critical inquiry and scholarship

(If you participate or have participated in more than 6 sports, please complete additional Sporting Details Questionnaires, Part B)

B. SPORTING DETAILS			
Please record your sporting activities in order of importance			
Type of sport(s) you have participated in	Sport 1	Sport 2	Sport 3
Current or past participation	Current ☐ Past ☐	Current ☐ Past ☐	Current ☐ Past ☐
Year started participation			
Years involved in the sport			
Current hours of training per week (1-12 months)			
Current hours of training per week (13-24 months)			
Hours of training per week prior to first injury (1-12 months)			
Hours of training per week prior to first injury (13-24 months)			

Type of sport(s) you have participated in	Sport 4	Sport 5	Sport 6
Current or past participation	Current ☐ Past ☐	Current ☐ Past ☐	Current ☐ Past ☐
Year started participation			
Years involved in the sport			
Current hours of training per week (1-12 months)			
Current hours of training per week (13-24 months)			
Hours of training per week prior to first injury (1-12 months)			
Hours of training per week prior to first injury (13-24 months)			

C. GENERAL MEDICAL DETAILS		If Yes, How long ago or how often	
Have you ever used oral corticosteroids (cortisone tablets)?	Yes ☐ No ☐	☐ 3 months ☐ 12 months	☐ 6 months ☐ 24 or more months
Have you ever been given an injection with corticosteroids?	Yes ☐ No ☐	☐ 3 months ☐ 12 months	☐ 6 months ☐ 24 or more months
Have you ever been given an injection of corticosteroids in or around a tendon?	Yes ☐ No ☐	☐ Once ☐ 3 times	☐ Twice ☐ >3 times
Have you ever used anabolic steroids?	Yes ☐ No ☐	☐ 3 months ☐ 12 months	☐ 6 months ☐ 24 or more months
Have you ever used fluoroquinolone antibiotics?	Yes ☐ No ☐	☐ 3 months ☐ 12 months	☐ 6 months ☐ 24 or more months
Do you suffer from any Connective Tissue and Rheumatological Diseases and Disorders?	Yes ☐ No ☐	If Yes, please select from the list below	
List of some Connective Tissue and/or Rheumatic Diseases and Disorders			
☐ Ankylosing Spondylitis ☐ Aspartylglycosaminuria (AGU) ☐ Behçet's Syndrome ☐ Crohn's Disease ☐ Discoid Lupus Erythematosus ☐ Ehlers Danlos syndrome (EDS) ☐ Eosinophilic Fascitis ☐ Giant Cell (Temporal) Arthritis ☐ Gout ☐ Hypersensitive vasculitis ☐ Lipid Storage Diseases ☐ Marfan Syndrome ☐ Menkes Kinky Hair Syndrome ☐ Mucopolysaccharidoses ☐ Myopathies and Dystrophies ☐ Ochronosis (Homocystinuria) ☐ Osteoarthritis ☐ Osteogenesis Imperfecta (OI) ☐ Polyarteritis Nodosa ☐ Polymyalgia Rheumatica ☐ Polymyositis & Dermatomyositis ☐ Pseudogout ☐ Reactive Arthritis ☐ Reiter's Syndrome ☐ Relapsing Polychondritis ☐ Rheumatic Arthritis ☐ Scleroderma ☐ Sjögren's Syndrome ☐ Systemic Lupus Erythematosus (SLE) ☐ Systemic Sclerosis ☐ Wegener's Granulomatosis			
Have any other members of your family suffered from any tendon pathology?	Yes ☐ No ☐	If Yes, please specify the family member (eg Mother, Son) and type of injury Acute Injury ☐ Chronic Pain and Swelling ☐ Other ☐	
Do you suffer from elevated blood cholesterol?	Yes ☐ No ☐	Do any other members of your family suffer from elevated blood cholesterol?	Yes ☐ No ☐
Have you been diagnosed with any of the following systemic diseases?	☐ No systemic disease ☐ Diabetes mellitus ☐ Adrenal disorders ☐ Thyroid disorders ☐ Amyloidosis ☐ Renal disease ☐ Other endocrine and metabolic disease (Specify)		

D. TENDON INJURY - MEDICAL DETAILS				
Symptoms				
How many times have you had tendon injuries?	Tendon injured	Date of Injury	Acute or Chronic Injury	Sudden ¹ or Gradual ² Onset
	1			
	2			
	3			
	4			
	5			
¹ Sudden onset is within a few seconds or minutes. ² Gradual onset is over days or weeks.				
Please complete a separate form, Part D only, for each Tendon Injury you have had				
Injury Number (1,2,3,4, or 5)	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ _____			
Which tendon did you injure?	☐ Rotator cuff tendon • Supraspinatus ☐ • Infraspinatus ☐ • teres minor ☐ ☐ Patellar tendon ☐ ☐ Wrist extensor tendons ☐ ☐ Achilles tendon ☐			
Which side was injured?	☐ Left ☐ Right ☐ Both			
Which region of your tendon was injured? Please indicate on a diagram. (Only if applicable)	☐ Upper 1/3 ☐ Middle 1/3 ☐ Lower 1/3			
To what extent was your Tendon ruptured?	☐ Complete ☐ Partial ☐ None			
How were you injured? (e.g., sport, walking)				
Grade of injury at the time of injury:	☐ pain only after exercise ☐ pain during exercise, but did not cause you to alter training ☐ pain during exercise, which causes you to alter training ☐ pain which causes you to stop training ☐ no pain ☐ not sure ☐ Other (Specify _____)			
Grade of injury currently	☐ pain only after exercise ☐ pain during exercise, but did not cause you to alter training ☐ pain during exercise, which causes you to alter training ☐ pain which causes you to stop training ☐ no pain ☐ not sure ☐ Other (Specify _____)			

Which of the following symptoms were present before the injury	☐ Pain (less than 1 week) ☐ Pain (1-4 weeks) ☐ Pain (> 4 weeks)	☐ Stiffness ☐ Swelling ☐ None
Which of the following symptoms were present after the injury	☐ Pain (less than 1 week) ☐ Pain (1-4 weeks) ☐ Pain (> 4 weeks)	☐ Stiffness ☐ Swelling ☐ None
If you have or had chronic tendon pain, what seems to alleviate the pain?		
Diagnosis		
Which type of Tendon Disease were you diagnosed with e.g. Rupture, Tendinitis, etc.		
Diagnosed by (Please indicate the name and contact number of the clinician who diagnosed you)	☐ Doctor ☐ Physiotherapist ☐ Biokineticist ☐ Podiatrist ☐ Other _____	
If you had a tendon rupture. How was it treated?	☐ Surgically ☐ Non-surgically	
If applicable, who was the surgeon?	Surgeon _____ Phone _____	
If applicable, what diagnostic imaging was performed?	☐ Ultrasound ☐ MRI ☐ CT Other _____	
If applicable, who did the imaging?	Clinician _____ Phone _____	
General Information		
	If Yes, How long ago or how often	
Have you ever used oral corticosteroids prior to your symptoms of tendon pathology (cortisone tablets)?	Yes ☐ No ☐	☐ 3 months ☐ 6 months ☐ 12 months ☐ 24 or more months
Have you ever been given an injection with corticosteroids prior to your symptoms of tendon pathology?	Yes ☐ No ☐	☐ 3 months ☐ 6 months ☐ 12 months ☐ 24 or more months
Have you ever been given an injection of corticosteroids in or around the injured tendon prior to your symptoms of tendon pathology?	Yes ☐ No ☐	☐ Once ☐ Twice ☐ 3 times ☐ >3 times
Have you ever used anabolic steroids prior to your symptoms of tendon pathology?	Yes ☐ No ☐	☐ 3 months ☐ 6 months ☐ 12 months ☐ 24 or more months
Have you ever used fluoroquinolone antibiotics prior to your symptoms of tendon pathology?	Yes ☐ No ☐	☐ 3 months ☐ 6 months ☐ 12 months ☐ 24 or more months

E. STRETCHING AND WARM UP (PRIOR TO YOUR INJURY)		
I usually stretch each week as follows: (Please tick ALL the appropriate boxes)		☐ Never ☐ Occasionally ☐ Before sport ☐ After sport ☐ Once daily ☐ Twice daily ☐ More than twice daily
Which of these muscle groups do you stretch?	Lower Back	☐ Always ☐ Occasionally ☐ Occasionally
	Buttock	☐ Always ☐ Occasionally ☐ Occasionally
	Hip Flexors	☐ Always ☐ Occasionally ☐ Occasionally
	Quads	☐ Always ☐ Occasionally ☐ Occasionally
	Hamstrings	☐ Always ☐ Occasionally ☐ Occasionally
	Calf Muscles	☐ Always ☐ Occasionally ☐ Occasionally
How many times do you stretch per week?		☐ Never ☐ < 5 min ☐ 5 min ☐ 10 min ☐ 15 min ☐ 20 min ☐ 25 min ☐ > 30 min
Do you warm up before exercise?		Yes ☐ No ☐
If yes, for how many minutes and how?		
Do you cool down after exercise?		Yes ☐ No ☐
If yes, for how many minutes and how?		

ii.



Department of Human Biology
 UCTMRC Research Unit for Exercise Science & Sports Medicine
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 650 4561
 Fax: + 27 21 686 7530

GENETIC BASIS OF LIGAMENT INJURY QUESTIONNAIRES

A. PERSONAL PARTICULARS			
First Name			
Surname			
Postal Address			
		Code	
E-mail address		Phone	
Date of birth		Age	
Height (cm)		Gender	☐ Male ☐ Female
Weight (kg)	Pre-Injury:	Current:	
Ethnic group (Only Required and Used for Research Purposes)	☐ Black/African ☐ White ☐ Indian ☐ Mixed Ancestry (Coloured) ☐ Asian ☐ Other		
Ancestry: Tribal or national background (eg Xhosa, Dutch, Zulu, German, Italian)	Father ☐ Unknown Mother ☐ Unknown		
Country of Birth			

B. SPORTING DETAILS			
Please record your sporting activities in order of importance			
Type of sport(s) you have participated in (please name)	Main sport 1	Other sport 2	Other sport 3
Number of years played prior to first ACL injury.			

Type of sport(s) you have participated in (please name)	Other sport 4	Other sport 5	Other sport 6
Number of years played prior to first ACL injury.			

C. ANTERIOR CRUCIATE LIGAMENT INJURY DETAILS	
Date of ACL injury?	
Which side was injured?	☐ Left ☐ Right ☐ Both
To what extent was your ligament ruptured?	☐ Complete ☐ Partial ☐ None ☐ Unknown
Investigation done to confirm the diagnosis	☐ MRI ☐ Surgery ☐ Scope
How was the ACL ruptured?	☐ Direct impact (directly to the injured knee) ☐ Twisting and bending with indirect contact (i.e. contact elsewhere on the body) ☐ Twisting and bending without contact (no external contact) ☐ Skiing ☐ Other (please specify).....
Please describe the exactly how the injury occurred (if uncertain please state that you do not know)	
Associated injuries?	☐ Meniscal tear ☐ MCL tear ☐ Other ligament tear ☐ Bone bruising ☐ Other.....

D. HISTORY OF OTHER LIGAMENT AND TENDON INJURIES IN THE PAST				
Have you ever injured a ligament in the past?	Yes ☐ No ☐			
If yes, please specify which ligaments? (You may tick more than one block, please select either L (left) or R (right))	L R		L R	
	Knee (ACL)	☐ ☐	Wrist ligaments ☐ ☐	
	Knee (MCL)	☐ ☐	Finger ligaments ☐ ☐	
	Ankle lateral ligaments ☐ ☐	Knee (PCL) ☐ ☐		
	Spinal ligaments ☐ ☐	Knee (LCL) ☐ ☐		
	Shoulder ligaments ☐ ☐	Ankle medial ligaments ☐ ☐		
	Elbow ligaments ☐ ☐	Other ligaments ☐ ☐		
To your knowledge, have any other members of your family suffered from any ligament injury?	Yes ☐ No ☐	If Yes, please specify the family member and condition: Please choose ligament injury from the list above ☐ Mother..... ☐ Father..... ☐ Sibling..... ☐ Son / daughter..... ☐ Other family member.....		
Have you ever injured a tendon in the past?	Yes ☐ No ☐			
If yes, please specify which tendon? (You may tick more than one block, please select either L (left) or R (right))	Foot and ankle:	L R		
		Achilles tendon ☐ ☐		
		Tibialis posterior ☐ ☐		
			Plantar fascia ☐ ☐	
	Knee:	Patellar tendon ☐ ☐		
	Elbow and wrist:	Wrist extensor tendons ☐ ☐		
	Shoulder:	Subscapularis ☐ ☐		
		Supraspinatus ☐ ☐		
		Infraspinatus ☐ ☐		
		Teres minor ☐ ☐		
Other.....				

To your knowledge, have any other members of your family suffered from any tendon pathology?	Yes ☐ No ☐	If Yes, please specify the family member and condition: Please choose tendon injury from the list above ☐ Mother..... ☐ Father..... ☐ Sibling..... ☐ Son / daughter..... ☐ Other family member.....
Have you ever suffered from any of the following joint capsule injuries?	☐ Acute shoulder dislocation ☐ Chronic shoulder instability ☐ Chronic ankle instability ☐ Other.....	

E. MEDICAL HISTORY		
Do you currently suffer from any of these medical conditions:		
☐ High Blood Pressure ☐ Emphysema ☐ Malignant disease (cancer) If Yes, what type? _____	☐ Angina/Heart Attack ☐ Rheumatoid arthritis ☐ Elevated Blood Cholesterol ☐ Diabetes mellitus ☐ Renal disease	☐ Asthma ☐ Osteoarthritis (wear & tear) ☐ Adrenal disorders ☐ Thyroid disorders ☐ Amyloidosis
Do you currently suffer from any other Connective Tissue & Rheumatological Diseases & Disorders?	Yes ☐ No ☐	If Yes, please select from the list below
List of some Connective Tissue and/or Rheumatic Diseases and Disorders		
☐ Ankylosing Spondylitis ☐ Aspartylglycosaminuria (AGU) ☐ Behcet's Syndrome ☐ Crohn's Disease ☐ Discoid Lupus Erythematosus ☐ Ehlers-Danlos syndrome (EDS) ☐ Eosinophilic Fascitis ☐ Giant Cell (Temporal) Arthritis ☐ Gout ☐ Hypersensitive Vasculitis	☐ Lipid Storage Diseases ☐ Marfan Syndrome ☐ Menkes Kinky Hair Syndrome ☐ Mucopolysaccharidoses ☐ Myopathies and Dystrophies ☐ Ochronosis (Homocystinuria) ☐ Osteogenesis imperfecta (OI) ☐ Polyarteritis Nodosa ☐ Polymyalgia Rheumatica ☐ Polymyositis & Dermatomyositis	☐ Pseudogout ☐ Reactive Arthritis ☐ Reiter's Syndrome ☐ Relapsing Polychondritis ☐ Scleroderma ☐ Sjogren's Syndrome ☐ Systemic Lupus Erythematosus (SLE) ☐ Systemic Sclerosis ☐ Wegener's Granulomatosis ☐ Other.....
What surgical operations have you had? (please list and give dates)	Operation	Date
If female:		

At what age did you start menstruating? (years)		
Are you currently using any type of contraception?	☐ Yes ☐ No	
If Yes, what type of contraception are you using?	☐ Pill ☐ Injection ☐ IUD	
Are you currently?	☐ Pre-menopausal (± 12 cycles per year at intervals of 23– 33 days & bleeding lasts 3-7 days) ☐ Menopausal (cycles are irregular and less frequent) ☐ Post-menopausal (no longer menstruating)	
Family History		
Do any other members of your family suffer from elevated blood cholesterol?	Yes ☐ No ☐	If Yes, which relative? ☐ Mother ☐ Father ☐ Sibling ☐ Son / daughter ☐ Other relative.....
Is there any history of arthritis in your family?	Yes ☐ No ☐	If Yes, which relative? ☐ Mother ☐ Father ☐ Sibling ☐ Son / daughter ☐ Other relative..... & What type of arthritis? Rheumatoid ☐ Osteoarthritis ☐ Other ☐

F. OCCUPATIONAL DETAILS	
What is your current occupation?	
What was your occupation prior to injuring your ligament?	

(E) Achilles tendinopathy clinical diagnostic criteria



Department of Human Biology
 UCT/MRC RESEARCH UNIT FOR EXERCISE SCIENCE & SPORTS MEDICINE
 Faculty of Health Sciences, University of Cape Town
 Private Bag, Rondebosch 7700, South Africa
 Tel: + 27 21 650 4561
 Fax: + 27 21 680 7530

CLINICAL DIAGNOSIS OF ACHILLES TENDINOPATHY

SUBJECT NUMBER/CODE: _____

Clinical criteria ^{1,2}	Present
Gradual progressive pain over the posterior lower leg - Achilles tendon area (> 6 weeks)	
Early morning pain	
Early morning stiffness	
History of swelling over the Achilles tendon area	
Tenderness to palpation over the Achilles tendon	
Palpable nodular thickening over the affected Achilles	
Positive "shift" test (movement of the nodular area with plantar- /dorsi-flexion)	
Other criteria	Present
Confirmation of the diagnosis by ultrasound *	
Confirmation of the diagnosis by MRI *	
Confirmation of the diagnosis by CT scan *	

*: One of these criteria must be present to confirm the diagnosis

Date: _____ / _____ / _____

Investigator: Prof M Schwellnus

Signature: _____

References:

1. Schepels AA, Jones H, Haas AL. Achilles tendon disorders in athletes. *Am J Sports Med* 2002; **30**: 287-305.
2. Kader D, Saxena A, Morin T, Mattioli N. Achilles tendinopathy: some aspects of basic science and clinical management. *Br J Sports Med* 2002; **36**: 239-49.

The University of Cape Town is committed to policies of equal opportunity and affirmative action which are essential to its mission of promoting critical inquiry and scholarship.

(II) MATERIAL AND METHODS

(A) Protocol for DNA extraction from whole bloodReagents**1. TKM1 Buffer (pH 7.6)**

	Final Conc.	MW	For 500 ml	For 1000 ml
Tris-HCl	10 mM	121.00	0.6056	1.2112
KCl	10 mM	74.56	0.3728	0.7456
MgCl ₂ ·6H ₂ O	10 mM	203.20	1.016	2.032
EDTA	2 mM	372.24	0.372	0.744
dH ₂ O			to 500 ml	to 1000 ml

2. TKM2 Buffer (pH 7.6)

	Final Conc.	MW	For 200 ml
Tris-HCl	10 mM	121.00	0.242
KCl	10 mM	74.56	0.149
MgCl ₂ ·6H ₂ O	10 mM	203.20	0.406
EDTA	2 mM	372.24	0.1488
NaCl		58.44	4.675
dH ₂ O			to 200 ml

3. 10% SDS

	Final Conc.	MW	For 200 ml
SDS	10%		20
dH ₂ O			to 200 ml

4. 1X TE buffer (pH 8.0)

	Final Conc.	MW	For 100 ml
Tris-HCl	10 mM	121.00	0.121
EDTA	1 mM	372.24	0.037
dH ₂ O			to 100 ml

5. 5M NaClO₄

	Final Conc.	MW	For 100 ml
NaClO ₄	5 M	122.4	61.2
dH ₂ O			to 100 ml

6. Other Chemicals and Reagents

- Chloroform (molecular grade)
- NP40
- Absolute ethanol

Protocol

1. Draw 5 ml of blood into an EDTA vacutainer tube
2. Transfer the blood to a sterile 15 ml polypropylene tube
3. Add 2 volumes (10 ml) of TKM1 buffer containing 2.5% Nonidet-P40
4. Mix by inverting several times and incubate at room temperature for 10 minutes to enhance the haemolysis of red blood cells
5. Centrifuge at 3000 rpm (1200Xg) at room temperature for 10 minutes
6. Decant off the supernatant leaving the white pellet at the bottom of the tube
7. Add 1 volume (5 ml) of TKM1 buffer
8. Invert and vortex the solution
9. Centrifuge at 3000 rpm (1200Xg) at room temperature for 10 minutes
10. Decant the supernatant leaving the white pellet in the bottom of the tube
11. Repeat steps 7-10 until the pellet in the bottom of the tube is clean and white
12. Add 800 µl of TKM2 buffer and 50 µl of the 10% SDS solution
13. Vortex and mix to assist in the lyses of the white blood cells
14. Incubate for 60 minutes at 55°C in a water bath
15. Add 150 µl of 5M NaClO₄
16. Add 500 µl of molecular biology grade chloroform
17. Vortex the solution
18. Transfer the solution to sterile 1.5 ml microfuge tubes
19. Centrifuge at 1300 rpm at room temperature for 5 minutes
20. Carefully transfer 500 µl of the top aqueous phase to a new sterile microfuge tube
21. Add 1 ml of absolute ethanol
22. Invert until DNA precipitates
23. Centrifuge at 1300 rpm at room temperature for 5-10 minutes
24. Carefully tip off supernatant leaving the pellet in the bottom of the tube
25. Allow pellet to air dry completely
26. Add 200 µl of 1XTE buffer
27. Incubate the tubes at 65°C for 15 minutes in a heating block
28. Store DNA at 4°C

Reference: Lahiri et al. (1991) [299]

(B) DNA quality assessment for whole exome sequencing

Table IIB. 1: Concentration and absorbance measurements made on a spectrophotometer used to assess the quantity and purity of DNA samples obtained from the 20 participants selected for whole exome sequencing on the Illumina HiSeq 2000/2500 system

Sample number	DNA concentration (ng/μl)	Nucleic Acid Ratios	
		A ₂₆₀ /A ₂₈₀	A ₂₆₀ /A ₂₃₀
SAW-TEN ^{SEQ}			
1	269.82	1.88	2.33
2	233.13	1.89	2.43
3	282.86	1.89	2.48
4	377.82	1.88	2.33
5	334.84	1.87	2.01
6	384.51	2.08	1.92
7	91.5	1.88	2.20
8	287.82	1.86	1.96
9	254.42	1.85	2.37
10	256.09	1.85	2.19
SAW-CON ^{SEQ}			
11	328.05	1.87	2.00
12	386.96	1.87	2.84
13	383.01	1.87	2.88
14	200.00	1.88	2.14
15	221.33	1.88	2.42
16	328.19	1.87	2.27
17	139.67	1.87	2.28
18	196.54	1.86	2.35
19	70.61	1.82	1.90
20	206.46	1.85	2.10

SAW-TEN^{SEQ} and SAW-CON^{SEQ}: The ten exemplar cases and ten exemplar controls selected from the SAW-Achilles tendon injury cohort for whole exome sequencing. Nucleic acid ratios were used to assess the purity of the DNA. Pure sample generally demonstrate A_{260}/A_{280} ratios of ~1.8 and A_{260}/A_{230} ratios between 2.0-2.2.

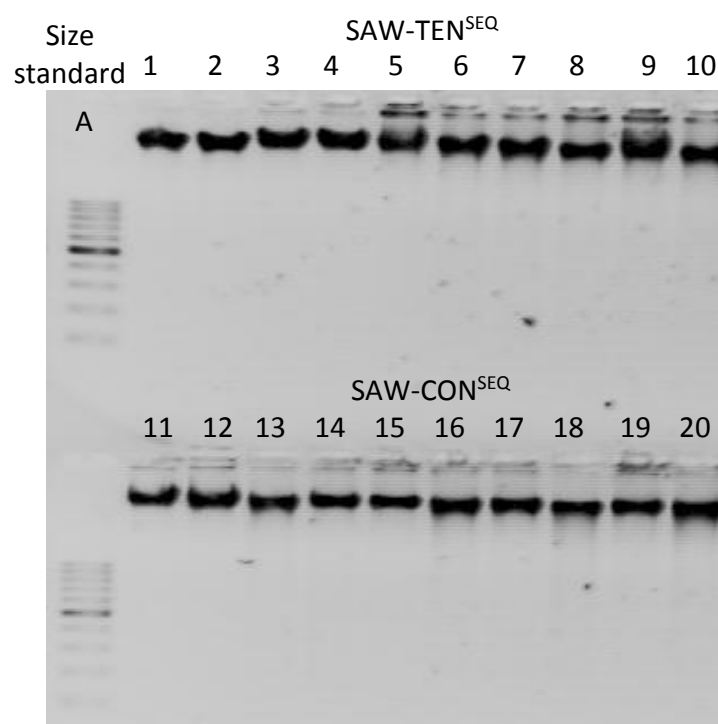


Figure IIB.1: Quality assessment of DNA samples obtained from the 20 participants selected for whole exome sequencing

A 5 µl DNA sample from each SAW-TEN^{SEQ} (Lanes 1-10) and SAW-CON^{SEQ} (Lanes 11-20) participant was assessed for degradation. Samples were visualised against a 100 bp DNA size standard (Lane A) on a photodocumentation system after electrophoresis at 120 V for 30 minutes on a 2% agarose gel.

(C) Sanger sequencing oligonucleotide primer pair sequences

Table IIC.1: Oligonucleotide primer pair sequences used to validate the new candidate variants identified by whole exome sequencing

SNP	Location	Primer pair	Fragment Size	Primers (5'- 3')	
				Forward	Reverse
<i>COL27A1</i>					
rs2567706	Exon 3	1	330	CTCAGATGTGGACATCCTCC	TCGTGCATGTCGAGGTCGAA
rs2241671		2	246	AAAACAGCTCCATCTTCTTTACAAAGTC	CGGCTTACTGGCATGGCT
rs2808770		3			
rs2567705		3	500	TGGTACCTCCAACCTTCGGGCA	CCCTGCTCTAAGCTGCTCCTTTTC
rs2808771		3			
<i>TNC</i>					
rs944510	Exon 2	1	216	TCAGCAGCTCCTTAACATCAGG	AGAACCAGCCAGTGGTGTTT
rs2992147	Exon 3	2	230	CTACACATGTGCCGTGCTC	CGAATGTCCAGGCAACTGTCA
rs1061494	Exon 4	3	559	GGAGAGCAAGGAGAAAGATTTGATGG	GACCTCTGTGGCTGACAGTGGAA
rs1061495	Exon 18	4	274	TTATCAGGGTCTTTTTTCATGAATCCATCC	AGGAATTCAGGAAGAACCATTGATGAA

COL27A1: Gene encoding the α_1 chain of type XXVII collagen; *TNC*: Gene encoding the tenascin-C glycoprotein. Primer pair: Primer pair sequences used to amplify the candidate variants. All primer sequences provided in the 5' to 3' direction.

(D) Allelic discrimination assays**(i) *TNC* rs1061494**

Chromosome position (GRCh38.10): chr9: 115084301

FASTA sequence

GTACTCCACACCAGGCTCCAGCTCC[T/C]GGATGATGGTGGACGTCTGGTCCCC

Location of variant indicated in parenthesis.

PCR conditions

Assay ID

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

C___3022586_1

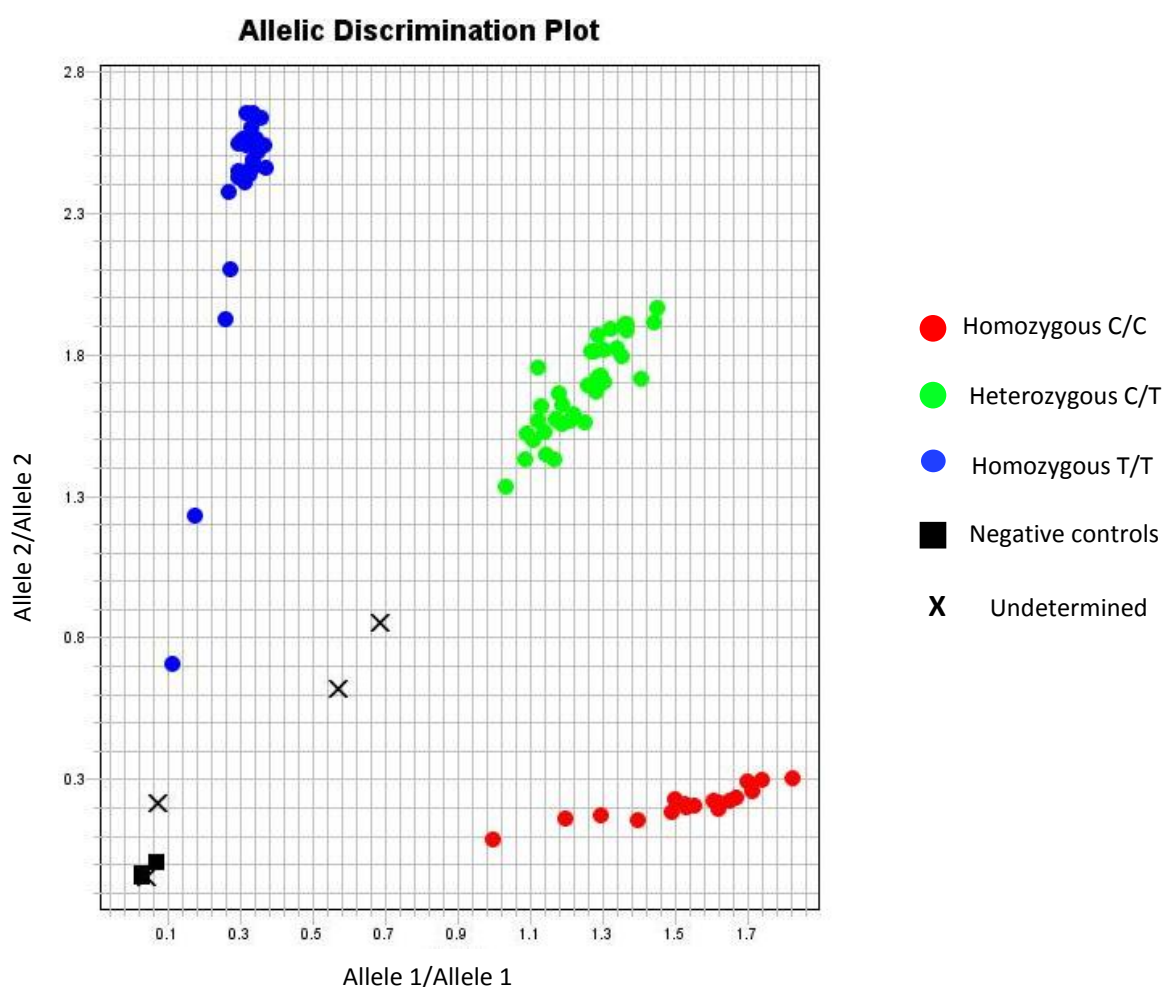


Figure IID.1: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs1061494) on the StepOnePlus™ Real-Time PCR System

(ii) TNC rs1138545

Chromosome position (GRCh38.10): chr9: 115073620

FASTA sequence

AGACTCACCAGTGGATGCCTTCACA[C/T]GTGCGGGCTTGCTCTTGTGTCTGCC

Location of variant indicated in parenthesis.

PCR conditions

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

Assay ID

C__25766586_20

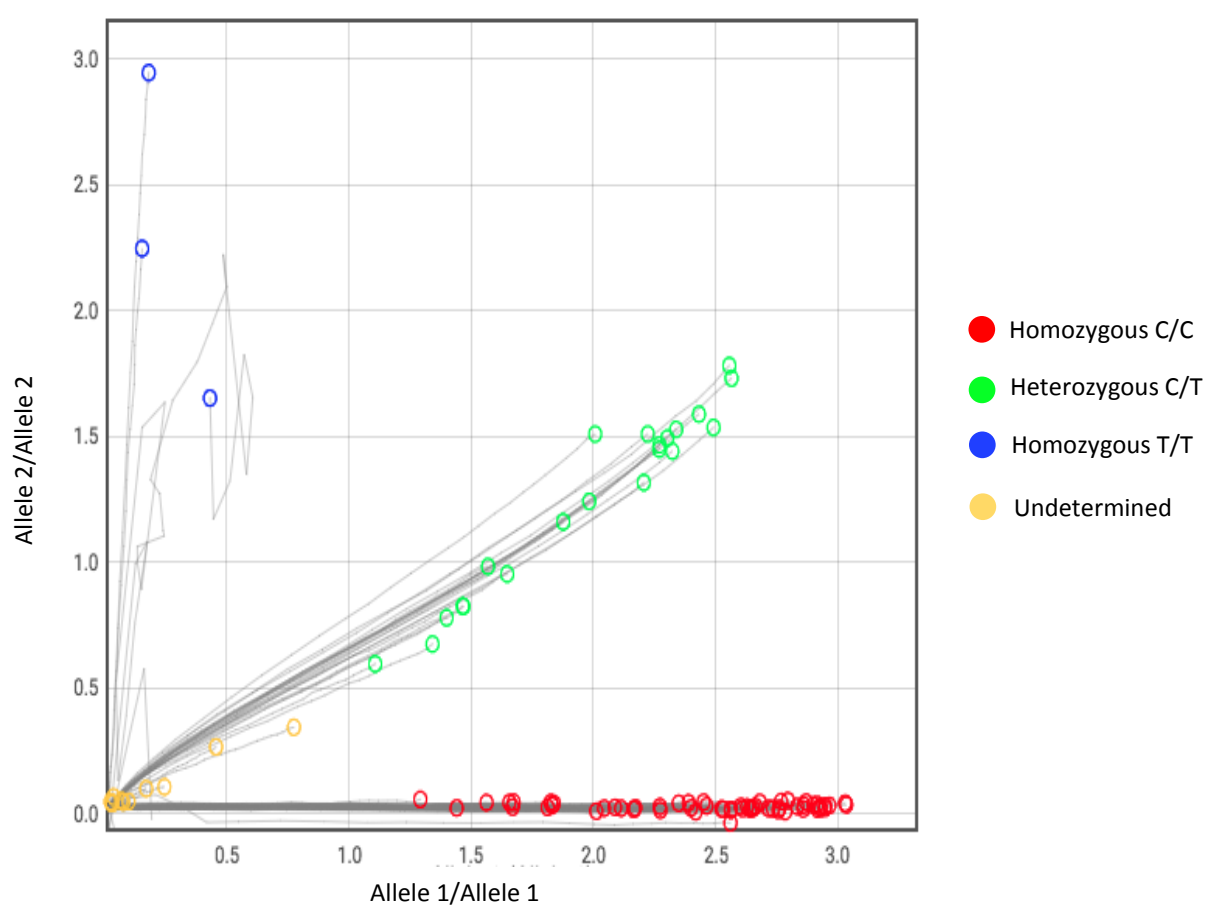
Allelic Discrimination plot

Figure IID.2: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs1138545) on the QuantStudio™3 Real-Time PCR system

(iii) **TNC rs2104772**

Chromosome position (GRCh38.10): chr9: 115046506

FASTA sequence

TCAGTAGCCTCTCTGAGACCTCTTA[T/A]GTCCCTGGTACGTTTCGGGGGCAAGT

*Location of variant indicated in parenthesis.***PCR conditions**

Assay ID

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

C___16182844_10

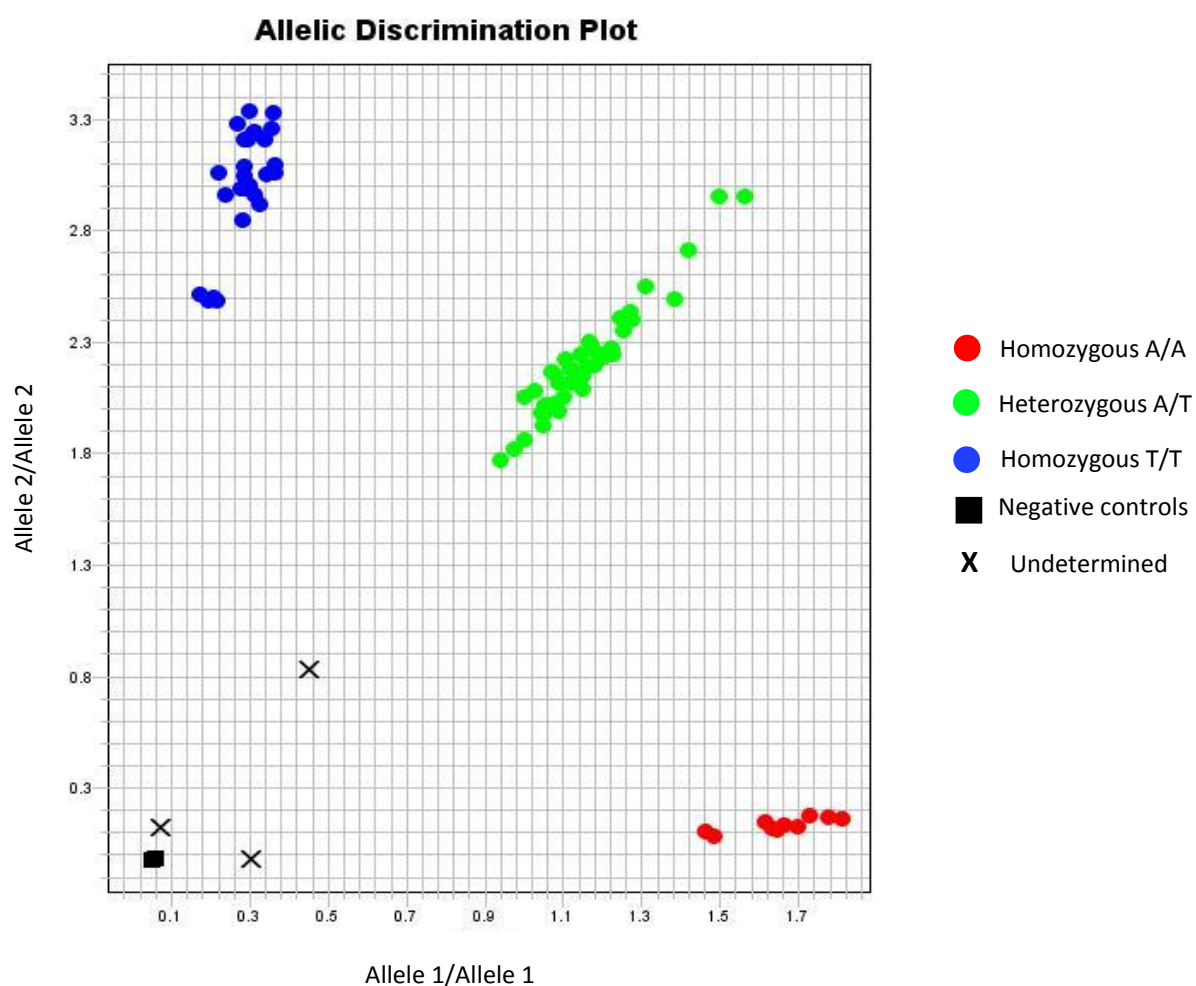


Figure IID.3: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs2104772) on the StepOnePlus™ Real-Time PCR System

(iv) *TNC* rs1061495

Chromosome position (GRCh38.10): chr9: 115042265

FASTA sequence

TAATCCGGAAGCTCTCCACTTGGGC[T/C]GTGGGTGCCCTCCAGCTGACAGTAG

Location of variant indicated in parenthesis.

PCR conditions	
TaqMan® SNP Genotyping Assays (Applied Biosystems™)	
Assay ID	
C___8787746_10	

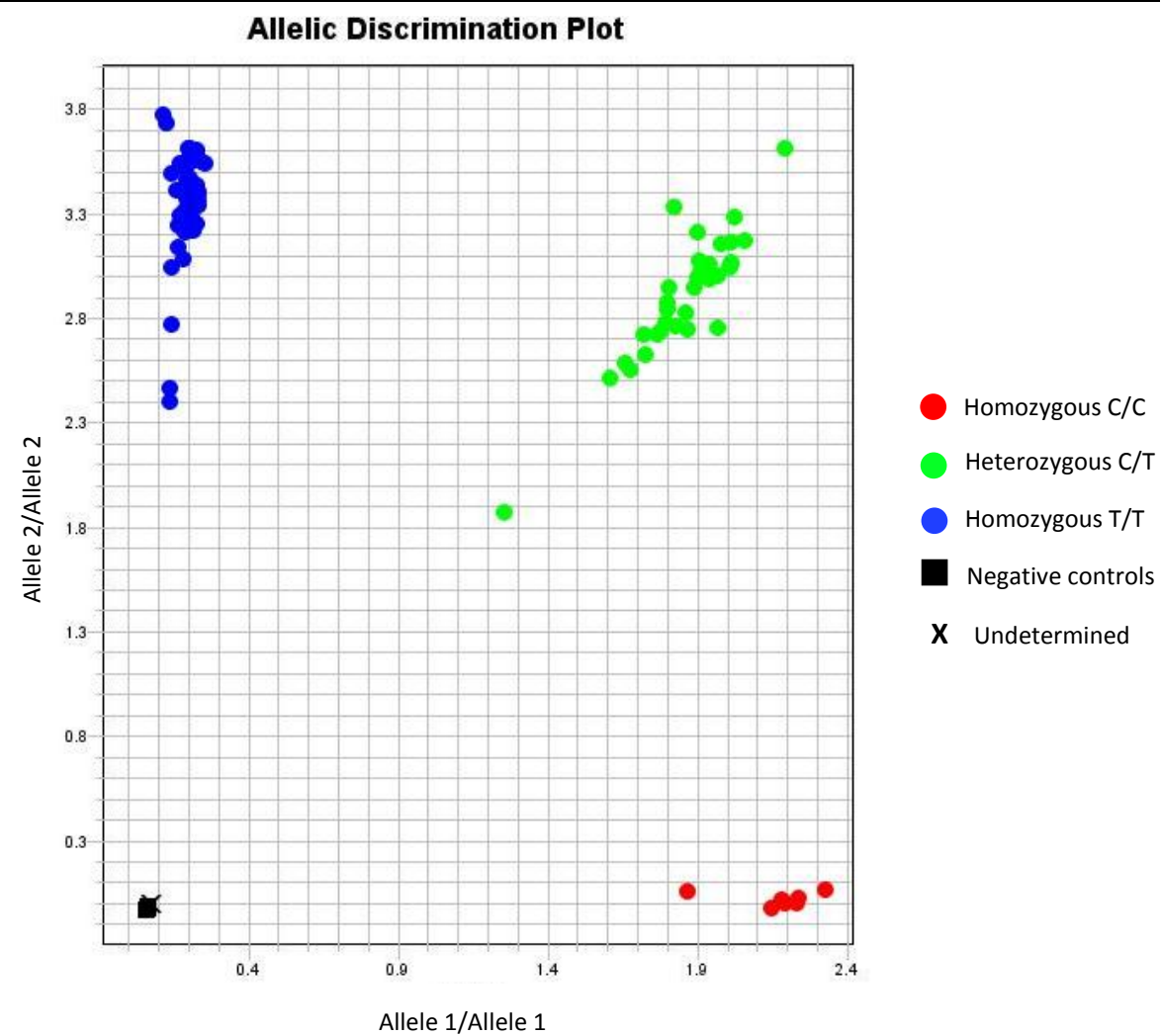


Figure IID.4: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs1061495) on the StepOnePlus™ Real-Time PCR System

(v) COL27A1 rs2567706

Chromosome position (GRCh38.10): chr9: 114167914

FASTA sequence

GCCTTCCTCTTCGCTGTCCGCAGCC[A/G]GAAACGCAAGCTGCAGCTGGGCCTG

*Location of variant indicated in parenthesis.***PCR conditions**

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

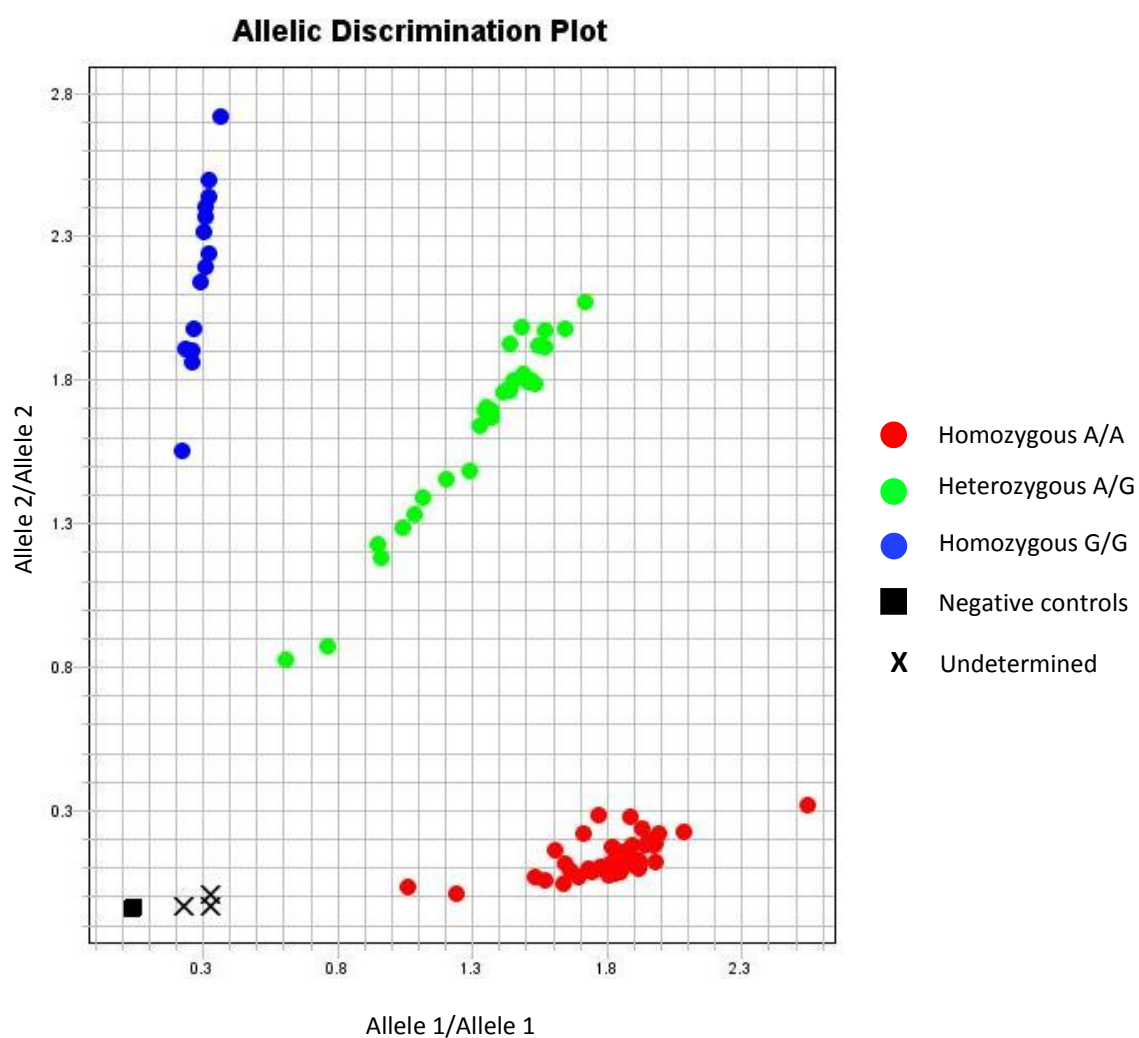
Assay ID
C__2060098_10

Figure IID.5: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs2567706) on the StepOnePlus™ Real-Time PCR System

(vi) COL27A1 rs2241671

Chromosome position (GRCh38.10): chr9: 114168819

FASTA sequence

AGTTCCTGCCAGAGTCTCCCGTCCC[G/A]CAGAGAAGCCCATCCAGAGGAACCC

Location of variant indicated in parenthesis.

PCR conditions

Assay ID

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

C___15906379_10

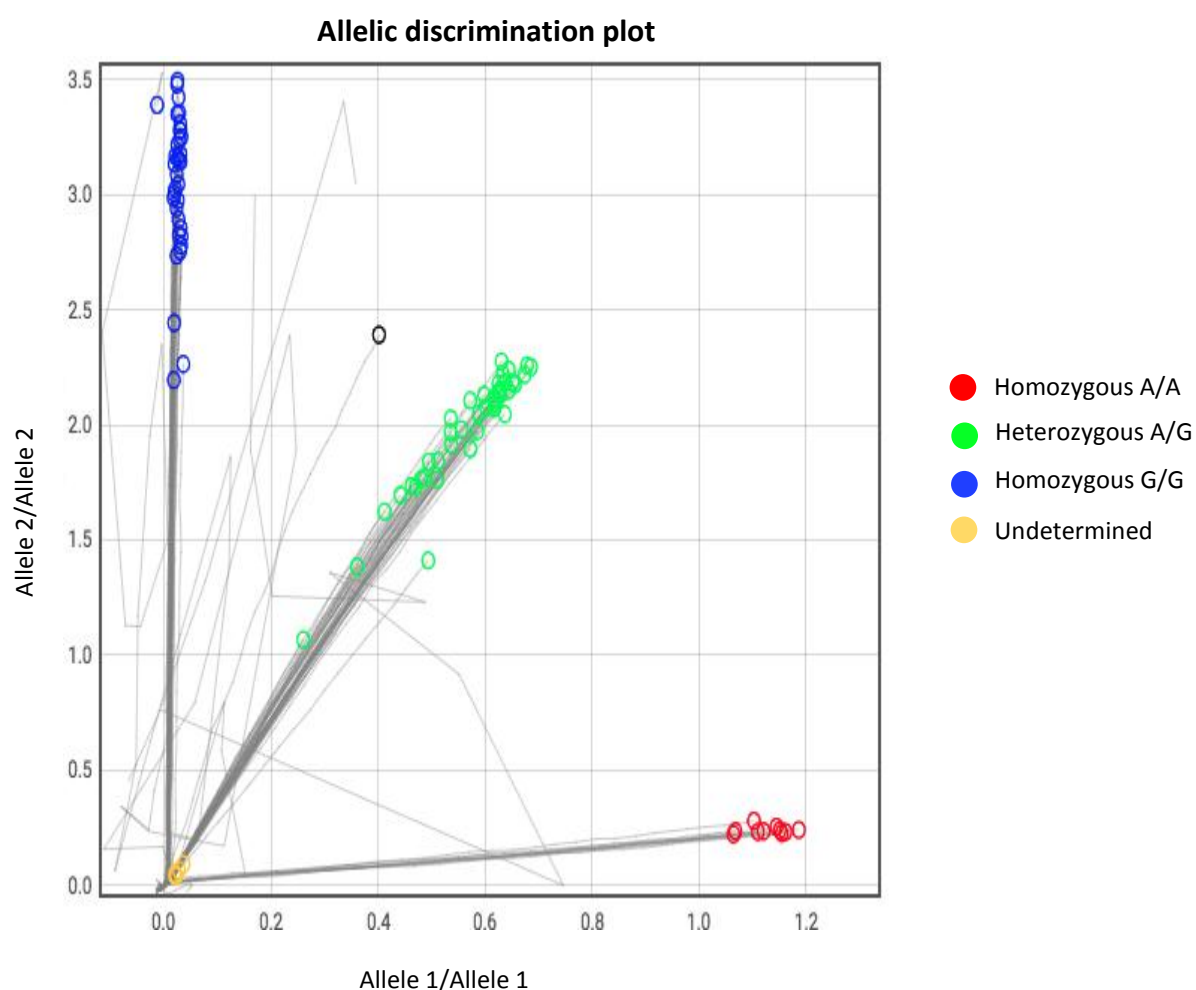


Figure IID.6: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs2241671) on the QuantStudio™3 Real-Time PCR system

(vii) COL27A1 rs2567705

Chromosome position (GRCh38.10): chr9: 114169386

FASTA sequence

CCGGCCCCACGAGCAGTGGCTATTCG[A/T]TCTTCCACCTGGCAGGATCTACGCC

Location of variant indicated in parenthesis.

PCR conditions

Assay ID

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

C__2060100_20

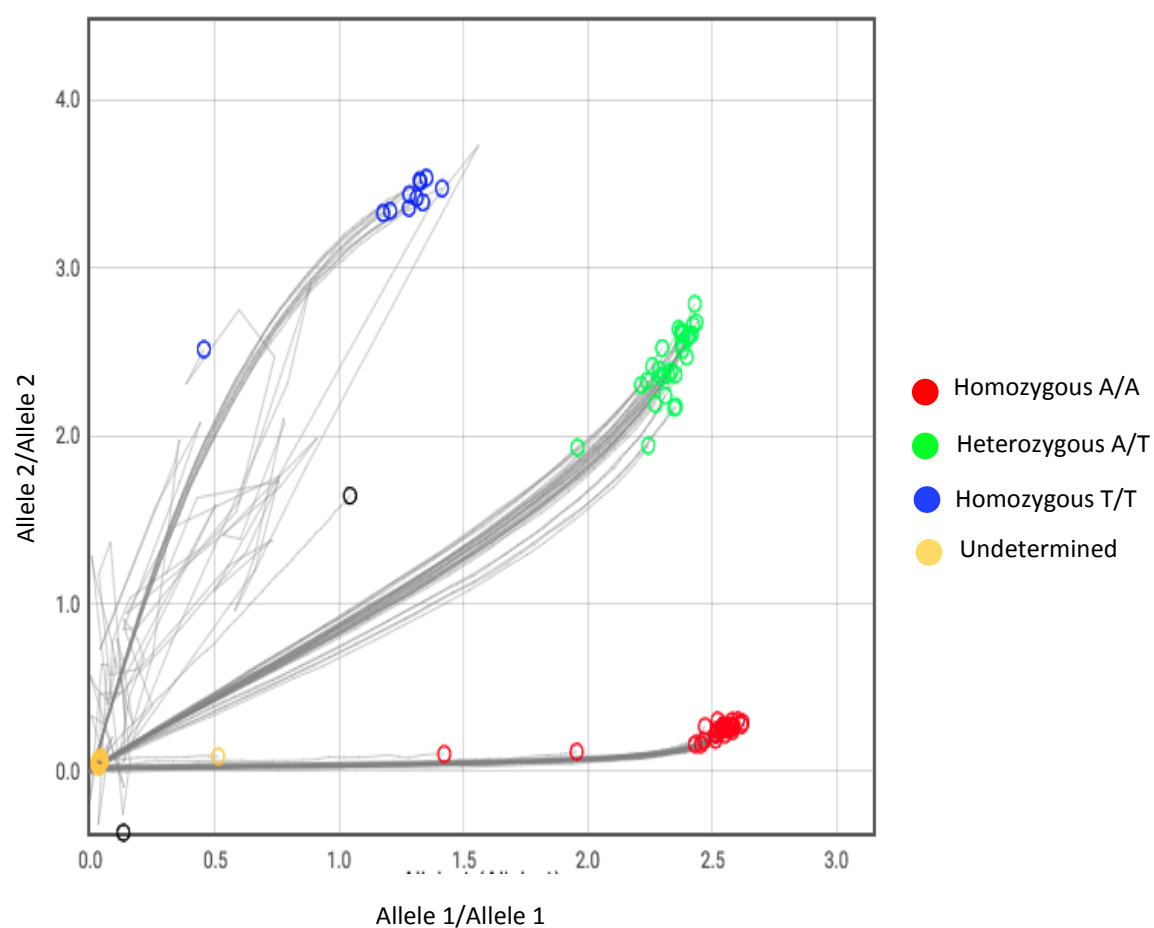
Allelic discrimination plot

Figure IID.7: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs2567705) on the QuantStudio™3 Real-Time PCR system

(viii) **MMP3 rs3025058**

Chromosome position (GRCh37.p13): chr11: 102715948-102715949

FASTA sequence

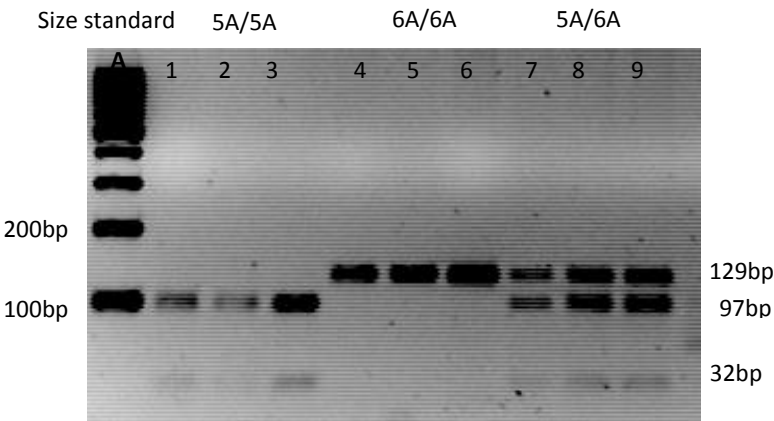
CCTCTCAAAGTGCTAGGATTACAGACATGGGTACGGCACCTGGCCTAAAGACATTTTAAACTAGT
ATTCTATGGTTCTCCATTCTTTGATGGGGGGA(-/A)AA[G]ACCATGTCTTGTCTGATTGAAATACA
GGGAAAATATTTGGCCACATTGATATGAGGACAAGGAGAACAGAGTGGTGGCAGTGATGTGAATT
CCAGGAAG

Underlined sequences indicate oligonucleotide primer sequences. Round parenthesis indicates position of indel (-/A). The highlighted sequence indicates the Tth111I restriction endonuclease recognition sequence. Squared parenthesis indicates position of induced mutation [A to G]. The oligonucleotides were created such that the A allele at the second base from the 3' end of the forward primer was substituted for a G allele, consequently creating a Tth111I recognition site in the case of the 5A allele. Therefore, the 6A allele would retain the uncut fragment size of 129bp, whereas the 5A allele would produce fragments 97bp and 32bp in size respectively.

PCR conditions	
Polymerase	Taq DNA polymerase (New England Biolabs Inc.)
Cycling conditions	Standard 35 cycles
Mg ²⁺ concentration	2.0 mM
Oligonucleotide sequences (5'-3')	<u>Forward</u> GGTTCTCCATTCTTTGATGGGGGAAA(G)A
Zhang et al. (2004) [367]	<u>Reverse</u> CTCCTGGAATTCACATCACTGCCACCACT
Amplicon size	129 bp
Restriction conditions	
Endonuclease	Tth111I (New England Biolabs Inc.)
Cutting site	5' GACN_NNGTC 3'
Reaction buffer	New England Biolabs CutSmart [®] Buffer
Incubation temperature	65°C
Incubation time	60 minutes
Resultant fragment sizes	6A allele: 129 bp 5A allele: 97 and 32 bp

Figure IID.8: Typical 3% agarose gel used to discriminate fragments generated after digestion with the Tth111I endonuclease and electrophoresed at 120 V for 40 minutes

Lane A is the 100 bp DNA size standard.
Lanes 1-3 demonstrate 5A/5A homozygous individuals.
Lane 4-6 demonstrate 6A/6A homozygous individuals.
Lane 7-9 demonstrate 5A/6A heterozygous individuals.
Size of fragments are indicated on the right side of gel.



(ix) **MMP3 rs679620**

Chromosome position (GRCh38.10): chr11: 102842889

FASTA sequence

GATTGTTTGACAAAGTGTAGAAAAA[G/A]CTCCAGCATCATCAAAAGATCTATA

Location of variant indicated in parenthesis.

PCR conditions	
	Assay ID
TaqMan® SNP Genotyping Assays (Applied Biosystems™)	C___3047717_1_

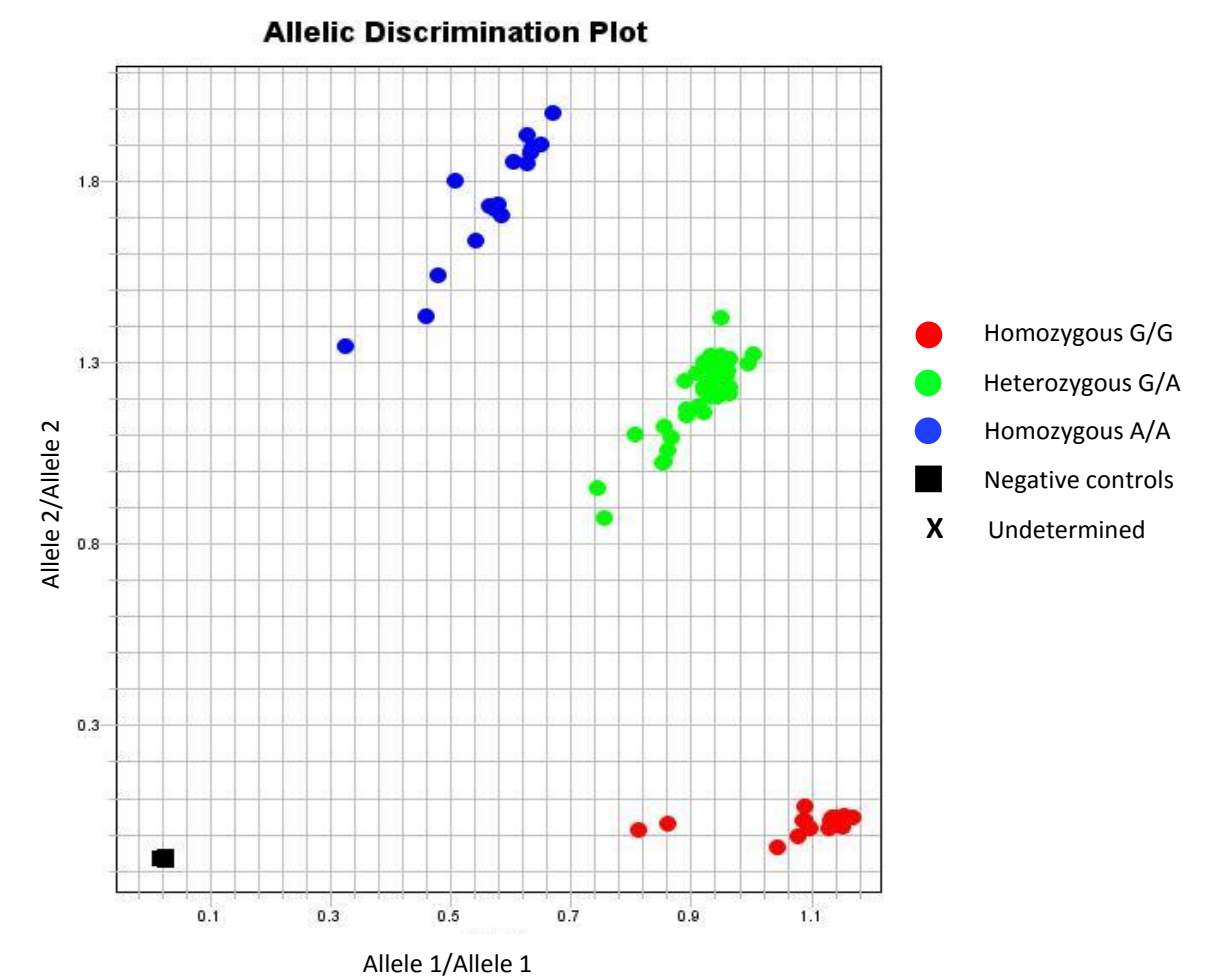


Figure IID.9: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs679620) on the StepOnePlus™ Real-Time PCR System

(x) MMP3 rs591058

Chromosome position (GRCh38.10): chr11: 102840607

FASTA sequence

ATAAATTGGTCCCTATTTAAGAAAT[C/T]GAGAACAATAGTTACTTATTTTTTA

*Location of variant indicated in parenthesis.***PCR conditions**

Assay ID

TaqMan® SNP Genotyping Assays (Applied Biosystems™)

C__785960_1_

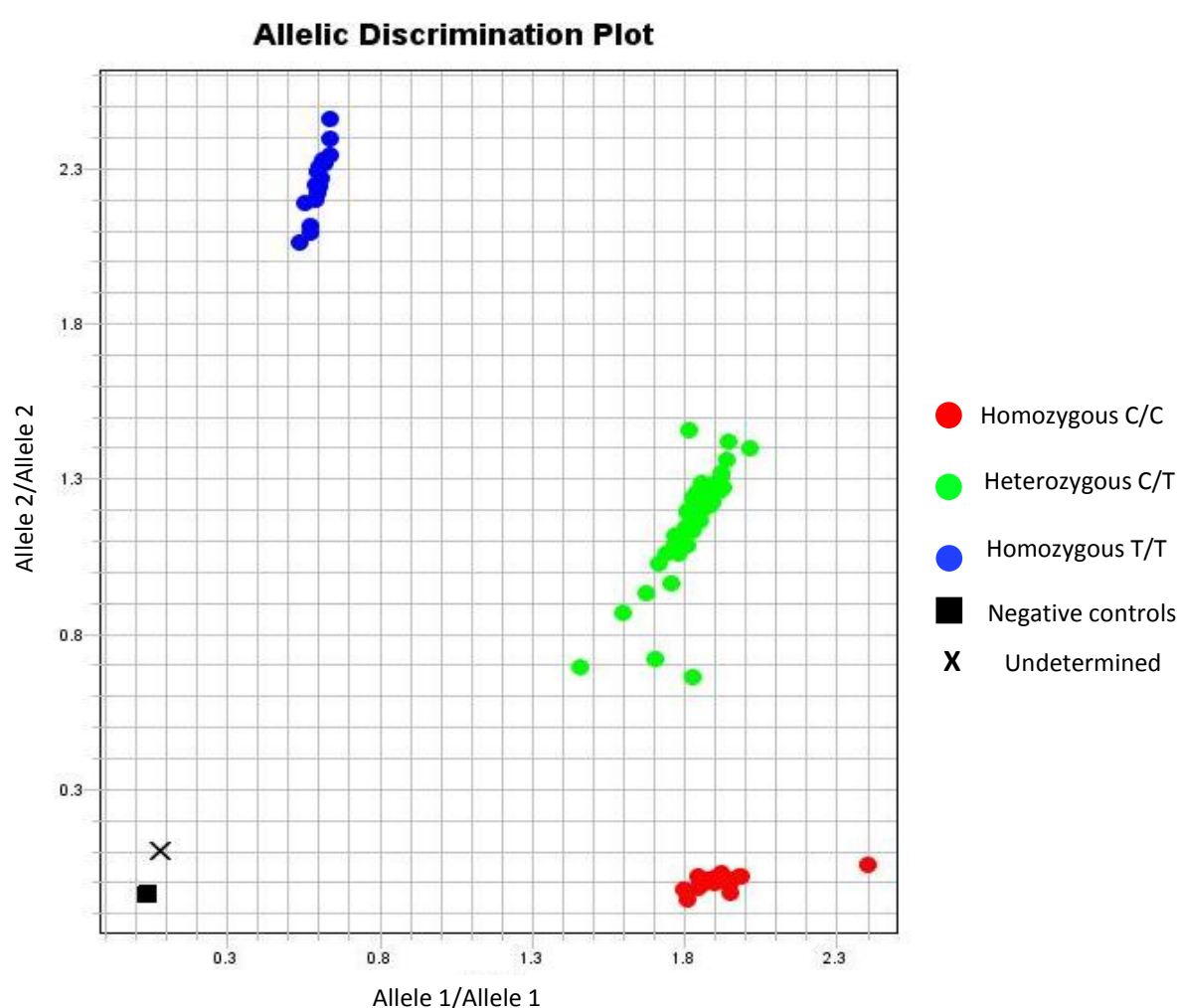


Figure IID.10: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs591058) on the StepOnePlus™ Real-Time PCR System

(xi) MMP3 rs650108

Chromosome position (GRCh38.10): chr11: 102838056

FASTA sequence

GAGGTGACAAGTGGGTGAGGTTAGA[G/A]GTAGCCCTTGAAAATCATATAGTAG

Location of variant indicated in parenthesis.

PCR conditions

	Assay ID
TaqMan® SNP Genotyping Assays (Applied Biosystems™)	C___785965_10

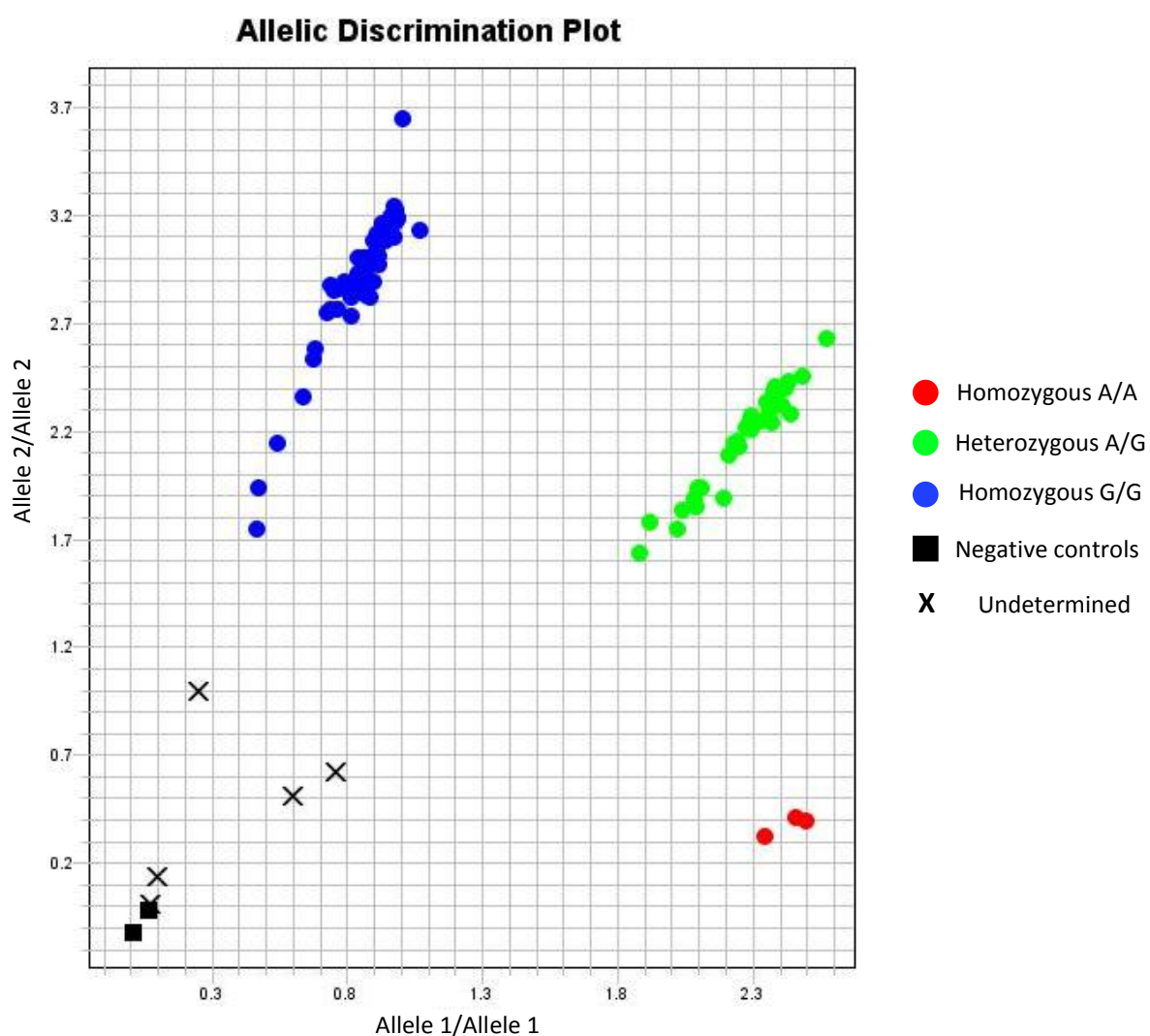


Figure IID.11: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs650108) on the StepOnePlus™ Real-Time PCR System

(xii) COL1A1 rs1107946

Chromosome position (GRCh38.10): chr17: 50203629

FASTA sequence

GGATGACACCCAGTCAAGGTTCTCT[G/T] GGGGAGGGATTATCCGCTGTCCCCA

Location of variant indicated in parenthesis.

PCR conditions	
TaqMan® SNP Genotyping Assays (Applied Biosystems™)	
Assay ID	
C___7477171_10	

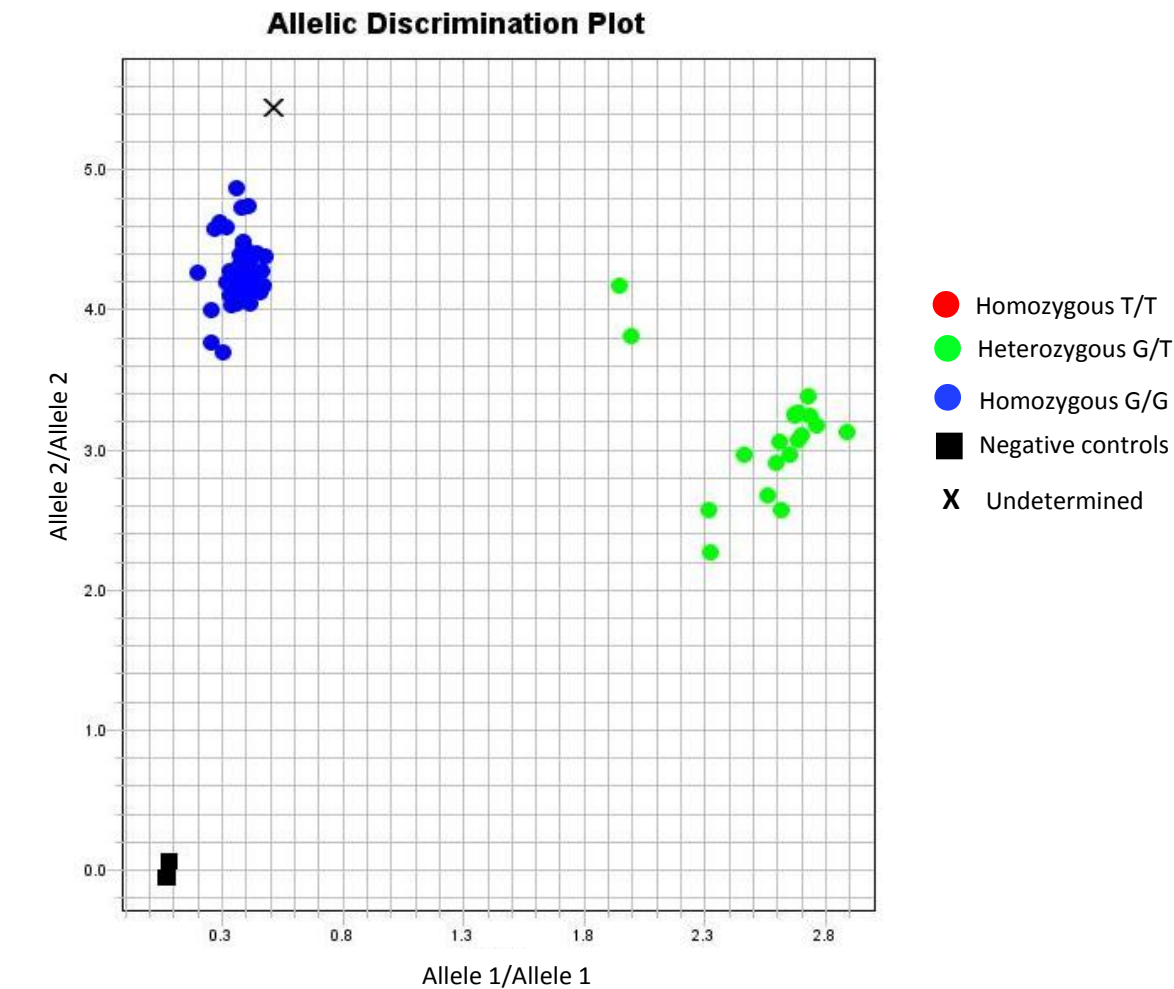


Figure IID.12: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs1107946) on the StepOnePlus™ Real-Time PCR System

(xiii) COL1A1 rs1800012

Chromosome position (GRCh38.10): chr17: 50200388

FASTA sequence

CCCTCCAGGTCGGGAGTAGGGCGGG [G/T] GTAAGGGACCCGTCCACCCCACCGC

Location of variant indicated in parenthesis.

PCR conditions	
TaqMan® SNP Genotyping Assays (Applied Biosystems™)	Assay ID C___7477170_30

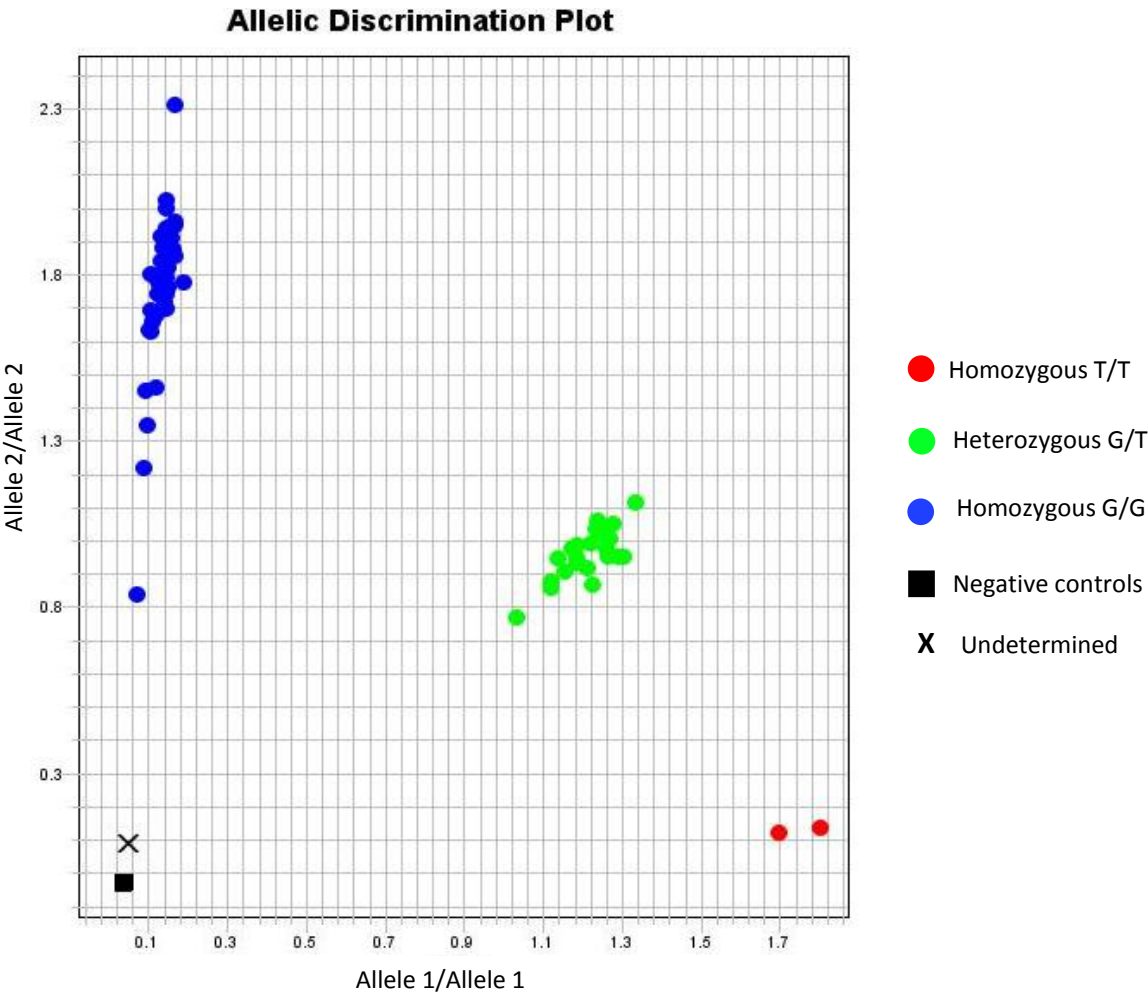


Figure IID.13: Allelic discrimination plot using the Taqman® SNP Genotyping Assay (rs1800012) on the StepOnePlus™ Real-Time PCR System

(III) SUPPLEMENTARY RESULTS

(A) Chapter 3: Profiling the *TNC* and *COL27A1* genes in musculoskeletal soft tissue injuries

Table IIIA.1: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs1061494	121	144		92	
TT	34 (41)	26 (38)	0.218	32 (29)	0.412
CT	46 (56)	57 (82)		54 (50)	
CC	20 (24)	17 (24)		14 (13)	
C Allele	43 (104)	45 (130)	0.661	41 (76)	0.767
HWE	0.718	0.043		0.303	
rs1138545	107	130		83	
CC	75 (80)	73 (95)	0.882	69 (57)	0.415
CT	20 (21)	25 (32)		29 (24)	
TT ^a	6 (6)	2 (3)		2 (2)	
T allele	15 (33)	15 (38)	0.903	17 (28)	0.791
HWE	0.030	0.751		1.000	
rs2104772	123	144		93	
TT	31 (38)	34 (49)	0.274	29 (27)	0.446
AT	38 (47)	44 (63)		46 (43)	
AA	31 (38)	22 (32)		25 (23)	
A allele	50 (123)	44 (127)	0.192	48 (89)	0.698
HWE	0.013	0.534		0.695	
rs1061495	125	145		93	
TT	54 (67)	61 (88)	0.268	67 (62)	0.070
TC	38 (47)	33 (48)		27 (25)	
CC ^b	8 (11)	6 (9)		6 (6)	
C Allele	28 (69)	23 (66)	0.197	20 (37)	0.086
HWE	0.509	0.825		0.228	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

Table IIIA.2: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs1061494	95	54		33	
CC	12 (12)	14 (8)	0.497	18 (6)	0.263
CT	53 (50)	43 (23)		36 (12)	
TT	35 (33)	43 (23)		46 (15)	
C Allele	39 (74)	36 (39)	0.710	36 (24)	0.770
HWE	0.529	1.000		0.261	
rs1138545	90	50		29	
CC	73 (66)	64 (32)	0.256	69 (20)	0.641
CT	22 (20)	32 (16)		28 (8)	
TT ^a	4 (4)	4 (2)		3 (1)	
T allele	16 (28)	20 (20)	0.432	17 (10)	0.847
HWE	0.223	1.000		1.000	
rs2104772	98	55		33	
AA	18 (18)	33 (18)	0.048	24 (8)	0.498
AT	50 (49)	40 (22)		39 (13)	
TT	32 (31)	27 (15)		36 (12)	
A allele	43 (85)	46 (185)	0.122	44 (29)	1.000
HWE	0.690	0.613		0.292	
rs1061495	96	53		33	
TT	53 (51)	59 (31)	0.607	61 (20)	0.544
TC	40 (38)	34 (18)		30 (10)	
CC ^b	7 (7)	7 (4)		9 (3)	
C Allele	27 (52)	25 (26)	0.681	24 (16)	0.747
HWE	1.000	0.486		0.345	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

Table IIIA.3: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs2567706	219	198		126	
AA	56 (122)	53 (105)	0.623	54 (68)	0.822
AG	37 (82)	37 (74)		35 (44)	
GG ^a	7 (15)	10 (19)		11 (14)	
G allele	26 (112)	28 (112)	0.402	29 (72)	0.437
HWE	0.860	0.202		0.098	
rs2241671	213	193		120	
GG	31 (66)	31 (59)	0.787	30 (36)	0.751
GA	51 (109)	52 (101)		50 (60)	
AA	18 (38)	17 (33)		20 (24)	
A allele	43 (185)	43 (167)	1.000	45 (108)	0.742
HWE	0.683	0.417		1.000	
rs2567705	213	189		120	
AA	42 (90)	34 (65)	0.762	41 (49)	0.504
AT	44 (93)	48 (90)		40 (48)	
TT	14 (30)	18 (34)		19 (23)	
T allele	36 (153)	42 (158)	0.102	40 (94)	0.473
HWE	0.555	0.783		0.102	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs2567706 AA genotype vs. AG + GG genotype

Table IIIA.4: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs2567706	96	54		33	
AA	55 (53)	50 (27)	0.610	52 (17)	0.840
AG	37 (35)	33 (18)		30 (10)	
GG	8 (8)	17 (9)		18 (6)	
G Allele	27 (51)	33 (36)	0.234	33 (22)	0.342
HWE	0.603	0.044		0.110	
rs2241671	94	52		31	
AA	15 (14)	21 (11)	0.719	23 (7)	0.615
AG	50 (47)	50 (26)		45 (14)	
GG	35 (33)	29 (15)		32 (10)	
A Allele	40 (75)	46 (48)	0.323	45 (28)	0.552
HWE	0.674	0.800		0.718	
rs2567705	94	50		31	
TT	15 (14)	22 (11)	0.825	26 (8)	0.708
AT	43 (40)	46 (23)		39 (12)	
AA	43 (40)	32 (16)		36 (11)	
T Allele	36 (68)	45 (45)	0.164	45 (28)	0.258
HWE	0.517	0.301		0.277	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance ($P < 0.05$).

Table IIIA.5: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *TNC* rs1061494 (T>C), rs1138545 (C>T), rs2104772 (T>A) and rs1061495 (T>C) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African Coloured (SAC)-ACL rupture cohort

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
rs1061494	79	70		40	
TT	38 (30)	27 (19)	0.205	25 (10)	0.219
CT	46 (36)	47 (33)		50 (20)	
CC	17 (13)	26 (18)		25 (10)	
C Allele	39 (62)	49 (69)	0.113	50 (40)	0.137
HWE	0.642	0.646		1.000	
rs1138545	69	71		41	
CC	83 (57)	84 (60)	0.742	90 (37)	0.302
CT	17 (12)	16 (11)		10 (4)	
TT ^a	0 (0)	0 (0)		0 (0)	0.403
T allele	9 (12)	8 (11)	0.822	5 (4)	
HWE	1.000	1.000		1.000	
rs2104772	76	74		43	
TT	36 (27)	28 (21)	0.490	28 (12)	0.635
AT	51 (39)	54 (40)		58 (25)	
AA	13 (10)	18 (13)		14 (6)	
A allele	39 (59)	45 (66)	0.330	43 (37)	0.561
HWE	0.814	0.371		0.351	
rs1061495	74	73		40	
TT	49 (36)	59 (43)	0.248	61 (25)	0.244
TC	45 (33)	34 (25)		32 (13)	
CC ^b	7 (5)	7 (5)		7 (3)	
C Allele	29 (43)	24 (35)	0.355	24 (19)	0.355
HWE	0.777	0.548		0.415	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for age at the time of recruitment or injury. P-values in bold typeset indicate significance ($P < 0.05$).

^a rs1138545 CC genotype vs. CT + TT genotype

^b rs1061495 TT genotype vs. TC + CC genotype

Table IIIA.6: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL27A1* rs2567706 (A>G), rs2241671 (G>A) and rs2567705 (A>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African Coloured (SAC)-ACL rupture cohort

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
rs2567706	78	72		44	
AA	58 (45)	68 (49)	0.237	68 (30)	0.333
AG	36 (28)	29 (21)		30 (13)	
GG ^a	6 (5)	3 (2)		2 (1)	
G allele	24 (38)	17 (25)	0.161	17 (15)	0.204
HWE	0.558	1.000		1.000	
rs2241671	79	70		42	
GG	41 (32)	51 (36)	0.217	62 (26)	0.083
GA	44 (35)	36 (25)		31 (13)	(0.030)
AA	15 (12)	13 (9)		7 (3)	
A allele	37 (59)	31 (43)	0.297	23 (19)	0.027
HWE	0.487	0.184		0.404	
rs2567705	77	71		44	
AA	34 (26)	45 (32)	0.339	48 (21)	0.311
AT	51 (39)	38 (27)		43 (19)	
TT	16 (12)	17 (12)		9 (4)	
T allele	41 (63)	36 (51)	0.419	31 (27)	0.125
HWE	1.000	0.331		1.000	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for age at the time of recruitment or injury. P-values in bold typeset indicate significance ($P < 0.05$).

^a rs2567706 AA genotype vs. AG + GG genotype

(B) Chapter 4: The investigation of variants within the *MMP3* gene and musculoskeletal soft tissue injuries

Table IIIB.1: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs3025058	126	128		79	
6A6A	25 (31)	20 (25)	0.278	18 (14)	0.281
5A6A	47 (59)	59 (75)		61 (48)	
5A5A	29 (36)	22 (28)		22 (17)	
6A allele	48 (121)	49 (125)	0.856	48 (76)	1.000
HWE	0.481	0.415		0.086	
rs679620	126	144		92	
GG	25 (32)	20 (29)	0.256	20 (18)	0.147
GA	46 (58)	58 (83)		61 (56)	
AA	29 (36)	22 (32)		20 (18)	
G allele	48 (122)	49 (141)	0.930	50 (92)	0.767
HWE	0.378	0.446		0.075	
rs591058	122	144		127	
CC	25 (32)	20 (29)	0.343	19 (18)	0.246
CT	47 (58)	57 (82)		59 (55)	
TT	29 (37)	23 (33)		22 (20)	
C allele	50 (122)	49 (140)	0.929	36 (91)	0.843
HWE	0.474	0.541		0.164	
rs650108	126	144		92	
GG	58 (73)	64 (92)	0.321	65 (60)	0.326
GA	37 (47)	33 (47)		33 (30)	
AA ^a	5 (6)	4 (5)		2 (2)	
A allele	23 (59)	20 (57)	0.338	18 (34)	0.226
HWE	0.806	0.358		1.000	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs650108 GG genotype vs. GA + AA genotype

Table IIIB.2: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *MMP3* rs3025058 (5A>6A), rs679620 (A>G), rs591058 (T>C) and rs650108 (G>A) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs3025058	96	49		28	
6A6A	29 (28)	22 (11)	0.658	21 (6)	0.709
5A6A	51 (49)	55 (27)		57 (16)	
5A5A	20 (19)	22 (11)		21 (6)	
6A allele	55 (105)	50 (49)	0.448	50 (28)	0.537
HWE	1.000	0.592		0.705	
rs679620	99	55		33	
GG	28 (28)	20 (11)	0.375	18 (6)	0.0256
GA	49 (48)	49 (27)		46 (15)	
AA	23 (23)	31 (17)		36 (12)	
G allele	53 (104)	45 (49)	0.199	41 (27)	0.128
HWE	0.556	1.000		0.727	
rs591058	89	53		32	
CC	30 (27)	19 (10)	0.187	19 (6)	0.407
CT	51 (45)	53 (28)		50 (16)	
TT	19 (17)	28 (15)		31 (10)	
C allele	56 (99)	45 (48)	0.106	44 (28)	0.109
HWE	1.000	1.000		1.000	
rs650108	98	53		33	
GG	50 (49)	60 (32)	0.237	64 (21)	0.227
GA	44 (43)	30 (16)		27 (9)	
AA ^a	6 (6)	9 (5)		9 (3)	
A allele	28 (55)	25 (26)	0.589	23 (15)	0.425
HWE	0.464	0.501		0.309	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs650108 GG genotype vs. GA + AA genotype

(C) Chapter 5: Functional *COL1A1* variants as modifiers of risk for acute musculoskeletal soft tissue injuries

Table IIIC.1: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs1107946	127	145		93	
GG	75 (95)	79 (115)	0.389	79 (73)	0.630
GT	25 (32)	21 (30)		22 (20)	
TT ^a	0 (0)	0 (0)		0 (0)	
T Allele	13 (32)	10 (30)	0.389	11 (20)	0.630
HWE	0.221	0.698		0.698	
rs1800012	127	145		93	
GG	69 (88)	70 (101)	1.000	72 (67)	0.765
GT	28 (35)	29 (42)		26 (24)	
TT ^b	3 (4)	1 (2)		2 (2)	
T Allele	17 (43)	16 (46)	0.814	15 (28)	0.606
HWE	0.791	0.258		0.526	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

SAW-CON^{ACL} vs. SAW-ACL rs1800012 TT genotype vs. GT + GG genotype: P=0.423

SAW-CON^{ACL} vs. SAW-NON rs1800012 TT genotype vs. GT + GG genotype: P=1.000

Table IIIC.2: Genotype and minor allele frequencies, in addition to Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the female South African White (SAW)-ACL rupture cohort

	SAW-CON ^{ACL}	SAW-ACL	P-values	SAW-NON	P-values
rs1107946	98	54		33	
GG	79 (77)	74 (40)	0.551	70 (23)	0.346
GT	19 (19)	26 (14)		30 (10)	
TT ^a	2 (2)	0 (0)		0 (0)	
T Allele	12 (23)	13 (14)	0.854	15 (10)	0.521
HWE	0.643	0.582		1.000	
rs1800012	98	52		32	
GG	67 (66)	56 (29)	0.213	56 (18)	0.290
GT	31 (30)	44 (23)		44 (14)	
TT ^b	2 (2)	0 (0)		0 (0)	
T Allele	17 (34)	22 (23)	0.317	22 (14)	0.430
HWE	0.730	0.264		0.295	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for body mass and age at the time of recruitment or injury. P-values in bold typeset indicate significance (P<0.05).

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

SAW-CON^{ACL} vs. SAW-ACL rs1800012 TT genotype vs. GT + GG genotype: P=0.544

SAW-CON^{ACL} vs. SAW-NON rs1800012 TT genotype vs. GT + GG genotype: P=1.000

Table IIIC.3: Genotype, minor allele frequencies and Hardy-Weinberg equilibrium (HWE) probabilities for the *COL1A1* rs1107946 (G>T) and rs1800012 (G>T) variants in the control (CON) group, anterior cruciate ligament rupture (ACL) group and subgroup of individuals reporting a non-contact mechanism of injury (NON) in the male South African Coloured (SAC)-ACL rupture cohort

	SAC-CON	SAC-ACL	P-values	SAC-NON	P-values
rs1107946	77	73		41	
GG	56 (43)	59 (43)	0.743	46 (19)	0.341
GT	42 (32)	33 (24)		46 (19)	
TT ^a	3 (2)	8 (6)		7 (3)	
T Allele	23 (36)	25 (36)	0.891	30 (25)	0.630
HWE	0.766	0.242		1.000	
rs1800012	77	73		41	
GG	80 (62)	84 (62)	0.523	88 (36)	0.441
GT	20 (15)	16 (10)		12 (5)	
TT ^b	0 (0)	0 (1)		0 (0)	
T Allele	10 (15)	7 (10)	0.689	6 (5)	0.441
HWE	1.000	0.473		0.473	

Genotypes are expressed as percentages with participant count (n) in parentheses. P-values are for the difference between diagnostic groups (CON vs. ACL and CON vs. NON) adjusted for age at the time of recruitment or injury. P-values in bold typeset indicate significance ($P < 0.05$).

^a rs1107946 GG genotype vs. GT + TT genotype

^b rs1800012 GG genotype vs. GT + TT genotype

SAC-CON^{ACL} vs. SAC-ACL rs1800012 TT genotype vs. GT + GG genotype: $P = 0.487$

SAC-CON^{ACL} vs. SAC-NON rs1800012 TT genotype vs. GT + GG genotype: $P = 1.000$

REFERENCES

1. Penedo FJ, Dahn JR. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Curr Opin Psychiatry*. 2005;18(2):189–93.
2. Morrow JR Jr, Defina LF, Leonard D, Trudelle-Jackson E, Custodio MA. Meeting physical activity guidelines and musculoskeletal injury: the WIN study. *Med Sci Sports Exerc*. 2012;44(10):1986–92.
3. Badley EM, Tennant A. Impact of disablement due to rheumatic disorders in a British population: estimates of severity and prevalence from the Calderdale Rheumatic Disablement Survey. *Ann Rheum Dis*. 1993;52(1):6–13.
4. Clayton RA, Court-Brown CM. The epidemiology of musculoskeletal tendinous and ligamentous injuries. *Injury*. 2008;39(12):1338–44.
5. Ljungqvist A, Schwellnus MP, Bachl N, Collins M, Cook J, Khan KM, et al. International Olympic Committee consensus statement: molecular basis of connective tissue and muscle injuries in sport. *Clin Sports Med*. 2008;27(1):231–9.
6. Järvinen TA, Kannus P, Maffulli N, Khan KM. Achilles tendon disorders: etiology and epidemiology. *Foot Ankle Clin*. 2005;10(2):255–66.
7. Renstrom P, Ljungqvist A, Arendt E, Beynon B, Fukubayashi T, Garrett W, et al. Non-contact ACL injuries in female athletes: an International Olympic Committee current concepts statement. *Br J Sports Med*. 2008;42(6):394–412.
8. de Jonge S, van den Berg C, de Vos RJ, van der Heide HJ, Weir A, Verhaar JA, et al. Incidence of midportion Achilles tendinopathy in the general population. *Br J Sports Med*. 2011;45(13):1026–8.
9. Coombes BK, Bisset L, Vicenzino B. A new integrative model of lateral epicondylalgia. *Br J Sports Med*. 2009;43(4):252–8.
10. Cook JL, Purdam CR. Is tendon pathology a continuum? A pathology model to explain the clinical presentation of load-induced tendinopathy. *Br J Sports Med*. 2009;43(6):409–16.
11. Cook JL, Rio E, Purdam CR, Docking SI. Revisiting the continuum model of tendon pathology: what is its merit in clinical practice and research? *Br J Sports Med*. 2016;50(19):1187–91.
12. Egger AC, Berkowitz MJ. Achilles tendon injuries. *Curr Rev Musculoskelet Med*. 2017;10(1):72–80.

13. Benjamin M, Ralphs JR. Fibrocartilage in tendons and ligaments--an adaptation to compressive load. *J Anat.* 1998;193(4):481–94.
14. Cook JL, Purdam C. Is compressive load a factor in the development of tendinopathy? *Br J Sports Med.* 2012;46(3):163–8.
15. Meeuwisse WH. Assessing causation in sport injury: a multifactorial model. *Clin J Sport Med.* 1994;4:166–70.
16. Järvinen TA, Kannus P, Paavola M, Järvinen TL, Józsa L, Järvinen M. Achilles tendon injuries. *Curr Opin Rheumatol.* 2001;13(2):150–5.
17. Posthumus M, Collins M, September AV, Schwellnus MP. The intrinsic risk factors for ACL ruptures: an evidence-based review. *Phys Sportsmed.* 2011;39(1):62–73.
18. Rahim M, Collins M, September A. Genes and musculoskeletal soft-tissue injuries. In Posthumus M, Collins M, editors. *Genetics and Sports.* 2nd ed. Basel: Karger; 2016. p.68–91.
19. September A, Rahim M, Collins M. Towards an understanding of the genetics of tendinopathy. In Ackermann PW, Hart DA, editors. *Metabolic influences on risk for tendon disorders. Advances in experimental medicine and biology (Vol 920).* Cham: Springer; 2016. p.109–16.
20. Flynn RK, Pedersen CL, Birmingham TB, Kirkley A, Jackowski D, Fowler PJ. The familial predisposition toward tearing the anterior cruciate ligament. *Am J Sports Med.* 2005;33(1):23–8.
21. Kraemer R, Wuerfel W, Lorenzen J, Busche M, Vogt PM, Knobloch K. Analysis of hereditary and medical risk factors in Achilles tendinopathy and Achilles tendon ruptures: A matched pair analysis. *Arch Orthop Trauma Surg.* 2012;132(6):847–53.
22. Rumian AP, Wallace AL, Birch HL. Tendons and ligaments are anatomically distinct but overlap in molecular and morphological features-a comparative study in an ovine model. *J Orthop Res.* 2007;25(4):458–64.
23. Doral MN, Turhan E, Dönmez G, Üzümcügil A, Kaymaz B, Ayvaz M, et al. Endoscopy and percutaneous suturing in the Achilles tendon ruptures. In Doral MN, Karlsson, J, editors. *Sports injuries: Prevention, diagnosis, treatment and rehabilitation.* Berlin Heidelberg: Springer; 2012. p.917-23.
24. O'Brien M. The anatomy of the Achilles tendon. *Foot Ankle Clin.* 2005;10(2):225–38.
25. Komi PV. Relevance of in vivo force measurements to human biomechanics. *J Biomech.* 1990;23 Suppl 1: 23–34.

26. Fukashiro S, Komi PV., Järvinen M, Miyashita M. In vivo achilles tendon loading during jumping in humans. *Eur J Appl Physiol Occup Physiol.* 1995;71(5):453–8.
27. Schepsis AA, Jones H, Haas AL. Achilles tendon disorders in athletes. *Am J Sports.* 2002;30(2):287–305.
28. Hoffmann A, Gross G. Tendon and ligament engineering in the adult organism: mesenchymal stem cells and gene-therapeutic approaches. *Int Orthop.* 2007;31(6):791–7.
29. Bleakney RR, White LM. Imaging of the Achilles tendon. *Foot Ankle Clin.* 2002;10(2):239-54.
30. Carr AJ, Norris SH. The blood supply of the calcaneal tendon. *J Bone Joint Surg Br.* 1989;71(1):100-1.
31. Ahmed IM, Lagopoulos M, McConnell P, Soames RW, Sefton GK. Blood supply of the Achilles tendon. *J Orthop Res.* 1998;16(5):591-6.
32. Stein V, Laprell H, Tinnemeyer S, Petersen W. Quantitative assessment of intravascular volume of the human Achilles tendon. *Acta Orthop Scand.* 2000;71(1):60-3.
33. Zantop T, Tillmann B, Petersen W. Quantitative assessment of blood vessels of the human Achilles tendon: an immunohistochemical cadaver study. *Arch Orthop Trauma Surg.* 2003;123(9):501-4.
34. Chen TM, Rozen WM, Pan WR, Ashton MW, Richardson MD, Taylor GI. The arterial anatomy of the Achilles tendon: anatomical study and clinical implications. *Clin Anat.* 2009;22(3):377-85.
35. Magnusson SP, Heinemeier KM, Kjaer M. Collagen homeostasis and metabolism. In Ackermann PW, Hart DA, editors. *Metabolic influences on risk for tendon disorders. Advances in experimental medicine and biology (Vol 920).* Cham: Springer; 2016. p.11–25.
36. Thorpe CT, Screen HR. Tendon structure and composition. In: Ackermann PW, Hart DA, editors. *Metabolic influences on risk for tendon disorders. Advances in experimental medicine and biology (Vol 920).* Cham: Springer; 2016. p.3-10.
37. Boushel R, Langberg H, Olesen J, Nowak M, Simonsen L, Bülow J, et al. Regional blood flow during exercise in humans measured by near-infrared spectroscopy and indocyanine green. *J Appl Physiol.* 2000;89(5):1868-78.
38. Ackermann PW. Neuronal regulation of tendon homeostasis. *Int J Exp Pathol.* 2013;94(4):271–86.

39. Ackermann PW, Li J, Lundeborg T, Kreicbergs A. Neuronal plasticity in relation to nociception and healing of rat achilles tendon. *J Orthop Res.* 2003;21(3):432-41.
40. Stilwell DL Jr. The innervation of tendons and aponeuroses. *Am J Anat.* 1957;100(3):289-317.
41. Andres KH, von Düring M, Schmidt RF. Sensory innervation of the Achilles tendon by group III and IV afferent fibers. *Anat Embryol (Berl).* 1985;172(2):145-56.
42. Messner K, Wei Y, Andersson B, Gillquist J, Räsänen T. Rat model of Achilles tendon disorder. A pilot study. *Cells Tissues Organs.* 1999;165(1):30-9.
43. Goldblatt JP, Richmond JC. Anatomy and biomechanics of the knee. *Oper Tech Sports Med.* 2003;11(3):172-86.
44. Bronstein RD, Schaffer JC. Physical examination of knee ligament injuries. *J Am Acad Orthop Surg.* 2017;25(4):280-7.
45. Bicer EK, Lustig S, Servien E, Selmi TA, Neyret P. Current knowledge in the anatomy of the human anterior cruciate ligament. *Knee Surg Sports Traumatol Arthrosc.* 2010;18(8):1075-84.
46. Zantop T, Petersen W, Sekiya JK, Musahl V, Fu FH. Anterior cruciate ligament anatomy and function relating to anatomical reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2006;14(10):982-92.
47. Girgis FG, Marshall JL, Monajem A. The cruciate ligaments of the knee joint. Anatomical, functional and experimental analysis. *Clin Orthop Relat Res.* 1975;(106):216-31.
48. Arnoczky SP. Anatomy of the anterior cruciate ligament. *Clin Orthop Relat Res.* 1983;(172):19-25.
49. Gabriel MT, Wong EK, Woo SL, Yagi M, Debski RE. Distribution of in situ forces in the anterior cruciate ligament in response to rotatory loads. *J Orthop Res.* 2004;22(1):85-9.
50. Mall NA, Lee AS, Cole BJ, Verma NN. The Functional and surgical anatomy of the anterior cruciate ligament. *Oper Tech Sports Med.* 2013;21(1):2-9.
51. Dunlap J, McCarthy JA, Joyce ME, Ogata K, Shively RA. Quantification of the perfusion of the anterior cruciate ligament and the effects of stress and injury to supporting structures. *Am J Sports Med.* 1989;17(6):808-10.
52. Bray RC, Leonard CA, Salo PT. Vascular physiology and long-term healing of partial ligament tears. *J Orthop Res.* 2002;20(5):984-9.

53. Toy BJ, Yeasting RA, Morse DE, McCann P. Arterial supply to the human anterior cruciate ligament. *J Athl Train*. 1995;30(2):149–52.
54. Kennedy JC, Alexander IJ, Hayes KC. Nerve supply of the human knee and its functional importance. *Am J Sports Med*. 1982;10(6):329–35.
55. Schultz RA, Miller DC, Kerr CS, Micheli L. Mechanoreceptors in human cruciate ligaments. A histological study. *J Bone Joint Surg Am*. 1984;66(7):1072–6.
56. Zimny ML, Schutte M, Dabezies E. Mechanoreceptors in the human anterior cruciate ligament. *Anat Rec*. 1986;214(2):204–9.
57. Haus J, Halata Z. Innervation of the anterior cruciate ligament. *Int Orthop*. 1990;14(3):293–6.
58. Çabuk H, Kuşku Çabuk F. Mechanoreceptors of the ligaments and tendons around the knee. *Clin Anat*. 2016;29(6):789–95.
59. Duthon VB, Barea C, Abrassart S, Fasel JH, Fritschy D, Menetrey J. Anatomy of the anterior cruciate ligament. *Knee Surg Sports Traumatol Arthrosc*. 2006;14(3):204–13.
60. Solomonow M. Sensory-motor control of ligaments and associated neuromuscular disorders. *J Electromyogr Kinesiol*. 2006;16(6):549–67.
61. Birk DE, Zycband E. Assembly of the tendon extracellular matrix during development. *J Anat*. 1994;184 (Pt 3):457–63.
62. Birk DE, Trelstad RL. Extracellular compartments in matrix morphogenesis: collagen fibril, bundle, and lamellar formation by corneal fibroblasts. *J Cell Biol*. 1984;99(6):2024–33.
63. Birk DE, Trelstad RL. Extracellular compartments in tendon morphogenesis: Collagen fibril, bundle, and macroaggregate formation. *J Cell Biol*. 1986;103(1):231–40.
64. Birk DE, Zycband EI, Winkelmann DA, Trelstad RL. Collagen fibrillogenesis in situ: fibril segments are intermediates in matrix assembly. *Proc Natl Acad Sci U S A*. 1989;86(12):4549–53.
65. Provenzano PP, Vanderby R Jr. Collagen fibril morphology and organization: implications for force transmission in ligament and tendon. *Matrix Biol*. 2006;25(2):71–84.
66. O'Brien M. Structure and metabolism of tendons. *Scand J Med Sci Sports*. 1997;7(2):55–61.

67. Gautieri A, Vesentini S, Redaelli A, Buehler MJ. Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. *Nano Lett.* 2011;11(2):757–66.
68. Screen HR, Lee DA, Bader DL, Shelton JC. An investigation into the effects of the hierarchical structure of tendon fascicles on micromechanical properties. *Proc Inst Mech Eng H.* 2004;18(2):109–19.
69. Birch HL, Thorpe CT, Rumian AP. Specialisation of extracellular matrix for function in tendons and ligaments. *Muscles Ligaments Tendons J.* 2013;3(1):12–22.
70. van Mechelen W, Hlobil H, Kemper HC. Incidence, severity, aetiology and prevention of sports injuries: A review of concepts. *Sports Med.* 1992;14(2):82–99.
71. Aström M, Rausing A. Chronic Achilles Tendinopathy. A survey of surgical and histopathologic findings. *Clin Orthop Relat Res.* 1995;(316):151–64.
72. Maffulli N, Khan KM, Puddu G. Overuse tendon conditions: time to change a confusing terminology. *Arthroscopy.* 1998;14(8):840–3.
73. Kingma JJ, de Knikker R, Wittink HM, Takken T. Eccentric overload training in patients with chronic Achilles tendinopathy: a systematic review. *Br J Sports Med.* 2007;41(6):e3. doi: 10.1136/bjsm.2006.030916.
74. Józsa L, Réffy A, Kannus P, Demel S, Elek E. Pathological alterations in human tendons. *Arch Orthop Trauma Surg.* 1990;110(1):15–21.
75. Kannus P, Józsa L. Histopathological changes preceding spontaneous rupture of a tendon. A controlled study of 891 patients. *J Bone Joint Surg Am.* 1991;73(10):1507–25.
76. Johansson C. Injuries in elite orienteers. *Am J Sports Med.* 1986;14(5):410–5.
77. Lysholm J, Wiklander J. Injuries in runners. *Am J Sports Med.* 1987;15(2):168–71.
78. Leppilahti J, Orava S, Karpakka J, Takala T. Overuse injuries of the Achilles tendon. *Ann Chir Gynaecol.* 1991;80(2):202–7.
79. Järvinen M. Lower leg overuse injuries in athletes. *Knee Surg Sports Traumatol Arthrosc.* 1993;1(2):126–30.
80. Kvist M. Achilles tendon injuries in athletes. *Sports Med.* 1994;18(3):173–201.
81. Fahlström M, Lorentzon R, Alfredson H. Painful conditions in the Achilles tendon region in elite badminton players. *Am J Sports Med.* 2002;30(1):51–4.

82. Kujala UM, Sarna S, Kaprio J. Cumulative incidence of achilles tendon rupture and tendinopathy in male former elite athletes. *Clin J Sport Med*. 2005;15(3):133–5.
83. Johannsen FE, Gam AN. Achillodynia is not just a sports injury. *Ugeskr Laeger*. 2010;172(48):3325–9.
84. Kannus P, Niittymäki S, Järvinen M, Lehto M. Sports injuries in elderly athletes: a three-year prospective, controlled study. *Age Ageing*. 1989;18(4):263–70.
85. Kannus P. Etiology and pathophysiology of chronic tendon disorders in sports. *Scand J Med Sci Sport*. 1997;7(2):78–85.
86. Gaida JE, Cook JL, Bass SL, Austen S, Kiss ZS. Are unilateral and bilateral patellar tendinopathy distinguished by differences in anthropometry, body composition, or muscle strength in elite female basketball players? *Br J Sports Med*. 2004;38(5):581–5.
87. Holmes GB, Lin J. Etiologic factors associated with symptomatic achilles tendinopathy. *Foot Ankle Int*. 2006;27(11):952–9.
88. Frey C, Zamora J. The effects of obesity on orthopaedic foot and ankle pathology. *Foot Ankle Int*. 2007;28(9):996–9.
89. Malliaras P, Cook JL, Kent PM. Anthropometric risk factors for patellar tendon injury among volleyball players. *Br J Sports Med*. 2007;41(4):259–63.
90. Gaida JE, Ashe MC, Bass SL, Cook JL. Is adiposity an under-recognized risk factor for tendinopathy? A systematic review. *Arthritis Rheum*. 2009;61(6):840–9.
91. Klein EE, Weil L Jr, Weil LS Sr, Fleischer AE. Body Mass index and Achilles tendonitis: a 10-year retrospective analysis. *Foot Ankle Spec*. 2013;6(4):276–82.
92. Scott RT, Hyer CF, Granata A. The correlation of achilles tendinopathy and body mass index. *Foot Ankle Spec*. 2013;6(4):283–5.
93. Abate M, Schiavone C, Salini V, Andia I. Occurrence of tendon pathologies in metabolic disorders. *Rheumatology (Oxford)*. 2013;52(4):599–608.
94. Abate M, Salini V, Antinolfi P, Schiavone C. Ultrasound morphology of the Achilles in asymptomatic patients with and without diabetes. *Foot Ankle Int*. 2014;35(1):44–9.
95. Leppilahti J, Puranen J, Orava S. Incidence of Achilles tendon rupture. *Acta Orthop Scand*. 1996;67(3):277–9.
96. Jozsa L, Kvist M, Balint BJ, Reffy A, Järvinen M, Lehto M, et al. The role of recreational sport activity in Achilles tendon rupture. A clinical, pathoanatomical, and sociological study of 292 cases. *Am J Sports Med*. 1989;17(3):338–43.

97. Levi N. The incidence of Achilles tendon rupture in Copenhagen. *Injury*. 1997;28(4):311–3.
98. Houshian S, Tscherning T, Riegels-Nielsen P. The epidemiology of achilles tendon rupture in a Danish county. *Injury*. 1998;29(9):651–4.
99. Davis JJ, Mason KT, Clark DA. Achilles tendon ruptures stratified by age, race, and cause of injury among active duty US military members. *Mil Med*. 1999;164(12):872–3.
100. Raikin SM, Garras DN, Krapchev PV. Achilles tendon injuries in a United States population. *Foot Ankle Int*. 2013;34(4):475–80.
101. Lantto I, Heikkinen J, Flinkkilä T, Ohtonen P, Leppilahti J. Epidemiology of Achilles tendon ruptures: increasing incidence over a 33-year period. *Scand J Med Sci Sports*. 2015;25(1):e133–8. doi: 10.1111/sms.12253.
102. Scott A, Grewal N, Guy P. The seasonal variation of Achilles tendon ruptures in Vancouver, Canada: a retrospective study. *BMJ Open*. 2014;4(2): e004320. doi: 10.1136/bmjopen-2013-004320.
103. Ganestam A, Kallemose T, Troelsen A, Barfod KW. Increasing incidence of acute Achilles tendon rupture and a noticeable decline in surgical treatment from 1994 to 2013. A nationwide registry study of 33,160 patients. *Knee Surg Sports Traumatol Arthrosc*. 2016;24(12):3730–7.
104. Nyyssönen T, Lühje P, Kröger H. The increasing incidence and difference in sex distribution of Achilles tendon rupture in Finland in 1987-1999. *Scand J Surg*. 2008;97(3):272–5.
105. Vafek EC, Plate JF, Friedman E, Mannava S, Scott AT, Danelson KA. The effect of strain and age on the mechanical properties of rat Achilles tendons. *Muscles Ligaments Tendons J*. 2018;7(3):548–53.
106. Maffulli N, Waterston SW, Squair J, Reaper J, Douglas AS. Changing incidence of Achilles tendon rupture in Scotland. *Clin J Sport Med*. 1999;9(3):157–60.
107. Gwynne-Jones DP, Sims M, Handcock D. Epidemiology and outcomes of acute Achilles tendon rupture with operative or nonoperative treatment using an identical functional bracing protocol. *Foot Ankle Int*. 2011;32(4):337–43.
108. Huttunen TT, Kannus P, Rolf C, Felländer-Tsai L, Mattila VM. Acute Achilles tendon ruptures: incidence of injury and surgery in Sweden between 2001 and 2012. *Am J Sports Med*. 2014;42(10):2419–23.

109. Sheth U, Wasserstein D, Jenkinson R, Moineddin R, Kreder H, Jaglal SB. The epidemiology and trends in management of acute Achilles tendon ruptures in Ontario, Canada: a population-based study of 27 607 patients. *Bone Joint J.* 2017;99-B(1):78–86.
110. Movin T, Gad A, Reinholt FP, Rolf C. Tendon pathology in long-standing achillodynia. Biopsy findings in 40 patients. *Acta Orthop Scand.* 1997;68(2):170–5.
111. Khan KM, Cook JL, Bonar F, Harcourt P, Astrom M. Histopathology of common tendinopathies: update and implications for clinical management. *Sports Med.* 1999;27(6):393–408.
112. Järvinen M, Józsa L, Kannus P, Järvinen TL, Kvist M, Leadbetter W. Histopathological findings in chronic tendon disorders. *Scand J Med Sci Sports.* 1997;7(2):86–95.
113. Alfredson H, Ohberg L, Forsgren S. Is vasculo-neural ingrowth the cause of pain in chronic Achilles tendinosis? An investigation using ultrasonography and colour Doppler, immunohistochemistry, and diagnostic injections. *Knee Surg Sports Traumatol Arthrosc.* 2003;11(5):334–8.
114. Lian Ø, Dahl J, Ackermann PW, Frihagen F, Engebretsen L, Bahr R. Pronociceptive and antinociceptive neuromediators in patellar tendinopathy. *Am J Sports Med.* 2006;34(11):1801–8.
115. Kraggsnaes MS, Fredberg U, Stribolt K, Kjaer SG, Bendix K, Ellingsen T. Stereological quantification of immune-competent cells in baseline biopsy specimens from achilles tendons: results from patients with chronic tendinopathy followed for more than 4 years. *Am J Sports Med.* 2014;42(10):2435–45.
116. Dakin SG, Newton J, Martinez FO, Hedley R, Gwilym S, Jones N, et al. Chronic inflammation is a feature of Achilles tendinopathy and rupture. *Br J Sports Med.* 2018;52(6):359-367.
117. Rees JD, Wilson AM, Wolman RL. Current concepts in the management of tendon disorders. *Rheumatology (Oxford).* 2006;45(5):508–21.
118. Riley G. The pathogenesis of tendinopathy. A molecular perspective. *Rheumatology (Oxford).* 2004;43(2):131–42.
119. Fu SC, Rolf C, Cheuk YC, Lui PP, Chan KM. Deciphering the pathogenesis of tendinopathy: a three-stages process. *Sports Med Athrosc Rehabil Ther Technol.* 2010;2:30. doi: 10.1186/1758-2555-2-30.
120. Schmidt-Rohlfing B, Graf J, Schneider U, Niethard FU. The blood supply of the Achilles tendon. *Int Orthop.* 1992;16(1):29–31.

121. Kvist M, Józsa L, Järvinen M. Vascular changes in the ruptured achilles tendon and paratenon. *Int Orthop*. 1992;16(4):377–82.
122. Birch HL, Wilson AM, Goodship AE. The effect of exercise-induced localised hyperthermia on tendon cell survival. *J Exp Biol*. 1997;200(Pt 11):1703–8.
123. Yu JS, Popp JE, Kaeding CC, Lucas J. Correlation of MR imaging and pathologic findings in athletes undergoing surgery for chronic patellar tendinitis. *AJR Am J Roentgenol*. 1995;165(1):115–8.
124. Öhberg L, Lorentzon R, Alfredson H. Neovascularisation in Achilles tendons with painful tendinosis but not in normal tendons: an ultrasonographic investigation. *Knee Surg, Sports Traumatol Arthrosc*. 2001;9(4):233–8.
125. Fenwick SA, Hazleman BL, Riley GP. The vasculature and its role in the damaged and healing tendon. *Arthritis Res*. 2002;4(4):252–60.
126. Alfredson H, Thorsen K, Lorentzon R. In situ microdialysis in tendon tissue: high levels of glutamate, but not prostaglandin E2 in chronic Achilles tendon pain. *Knee Surg Sports Traumatol Arthrosc*. 1999;7(6):378–81.
127. Schubert TE, Weidler C, Lerch K, Hofstädter F, Straub RH. Achilles tendinosis is associated with sprouting of substance P positive nerve fibres. *Ann Rheum Dis*. 2005;64(7):1083–6.
128. Molloy TJ, Kemp MW, Wang Y, Murrell GA. Microarray analysis of the tendinopathic rat supraspinatus tendon: glutamate signaling and its potential role in tendon degeneration. *J Appl Physiol*. 2006;101(6):1702–9.
129. Nho SJ, Yadav H, Shindle MK, Macgillivray JD. Rotator cuff degeneration: etiology and pathogenesis. *Am J Sports Med*. 2008;36(5):987–93.
130. Burry HC. The pathology of the painful heel. *Br J Sports Med*. 1971;6(1):9–12.
131. Burry HC, Pool CJ. Central degeneration of the achilles tendon. *Rheumatology*. 1973;12(4):177–81.
132. Dakin SG, Werling D, Hibbert A, Abayasekara DR, Young NJ, Smith RK, et al. Macrophage sub-populations and the lipoxin A4 receptor implicate active inflammation during equine tendon repair. *PLoS One*. 2012;7(2):e32333. doi: 10.1371/journal.pone.0032333.
133. Manning CN, Havlioglu N, Knutsen E, Sakiyama-Elbert SE, Silva MJ, Thomopoulos S, et al. The early inflammatory response after flexor tendon healing: A gene expression and histological analysis. *J Orthop Res*. 2014;32(5):645–52.

134. Magnusson SP, Langberg H, Kjaer M. The pathogenesis of tendinopathy: balancing the response to loading. *Nat Rev Rheumatol*. 2010;6(5):262–8.
135. September AV, Nell EM, O’Connell K, Cook J, Handley CJ, van der Merwe L, et al. A pathway-based approach investigating the genes encoding interleukin-1 β , interleukin-6 and the interleukin-1 receptor antagonist provides new insight into the genetic susceptibility of Achilles tendinopathy. *Br J Sports Med*. 2011;45(13):1040–7.
136. Fenwick SA, Hazelman BL, Harrall RL, Hackney R, Riley GP. Transforming growth factor- β isoform expression in chronic Achilles tendinopathy and their effects on tendon cell populations. *Int J Exp Pathol*. 2000;81(1):A11–2.
137. Legerlotz K, Jones ER, Screen HR, Riley GP. Increased expression of IL-6 family members in tendon pathology. *Rheumatology (Oxford)*. 2012;51(7):1161–5.
138. Behzad H, Sharma A, Mousavizadeh R, Lu A, Scott A. Mast cells exert pro-inflammatory effects of relevance to the pathophysiology of tendinopathy. *Arthritis Res Ther*. 2013;15(6):R184. doi: 10.1186/ar4374.
139. Rees JD, Stride M, Scott A. Tendons-time to revisit inflammation. *Br J Sports Med*. 2014;48(21):1553–7.
140. Scott A, Khan KM, Heer J, Cook JL, Lian O, Duronio V. High strain mechanical loading rapidly induces tendon apoptosis: an ex vivo rat tibialis anterior model. *Br J Sports Med*. 2005;39(5):e25. doi: 10.1136/bjsm.2004.015164.
141. Yuan J, Murrell GA, Wei AQ, Wang MX. Apoptosis in rotator cuff tendinopathy. *J Orthop Res*. 2002;20(6):1372–9.
142. Szomor ZL, Appleyard RC, Murrell GA. Overexpression of nitric oxide synthases in tendon overuse. *J Orthop Res*. 2006;24(1):80–6.
143. Millar NL, Wei AQ, Molloy TJ, Bonar F, Murrell GA. Heat shock protein and apoptosis in supraspinatus tendinopathy. *Clin Orthop Relat Res*. 2008;466(7):1569–76.
144. Millar NL, Wei AQ, Molloy TJ, Bonar F, Murrell GA. Cytokines and apoptosis in supraspinatus tendinopathy. *J Bone Joint Surg Br*. 2009;91(3):417–24.
145. Gaida JE, Bagge J, Purdam C, Cook J, Alfredson H, Forsgren S. Evidence of the TNF- α system in the human achilles tendon: expression of TNF- α and TNF receptor at both protein and mRNA levels in the tenocytes. *Cells Tissues Organs*. 2012;196(4):339–52.
146. Nell EM, van der Merwe L, Cook J, Handley CJ, Collins M, September AV. The apoptosis pathway and the genetic predisposition to Achilles tendinopathy. *J Orthop Res*. 2012;30(11):1719–24.

147. Sun ZM, Ling M, Huo Y, Chang Y, Li Y, Qin H, et al. Caspase 9 gene polymorphism and susceptibility to lumbar disc disease in the Han population in northern China. *Connect Tissue Res.* 2011;52(3):198–202.
148. Zhu G, Jiang XR, Xia CL, Sun YJ, Zeng QS, Wu XM, et al. Association of FAS and FAS ligand polymorphisms with the susceptibility and severity of lumbar disc degeneration in Chinese Han population. *Biomarkers.* 2011;16(6):485–90.
149. Arampatzis A, Karamanidis K, Albracht K. Adaptational responses of the human Achilles tendon by modulation of the applied cyclic strain magnitude. *J Exp Biol.* 2007;210(Pt 15):2743–53.
150. Arampatzis A, Peper A, Bierbaum S, Albracht K. Plasticity of human Achilles tendon mechanical and morphological properties in response to cyclic strain. *J Biomech.* 2010;43(16):3073–9.
151. Maeda E, Ohashi T. Mechano-regulation of gap junction communications between tendon cells is dependent on the magnitude of tensile strain. *Biochem Biophys Res Commun.* 2015;465(2):281–6.
152. Fredberg U, Stengaard-Pedersen K. Chronic tendinopathy tissue pathology, pain mechanisms, and etiology with a special focus on inflammation. *Scand J Med Sci Sport.* 2008;18(1):3–15.
153. Abate M, Silbernagel KG, Siljeholm C, Di Iorio A, De Amicis D, Salini V, et al. Pathogenesis of tendinopathies: inflammation or degeneration? *Arthritis Res Ther.* 2009;11(3):235. doi: 10.1186/ar2723.
154. Scott A, Cook JL, Hart DA, Walker DC, Duronio V, Khan KM. Tenocyte responses to mechanical loading in vivo: a role for local insulin-like growth factor 1 signaling in early tendinosis in rats. *Arthritis Rheum.* 2007;56(3):882–91.
155. Danielson P, Alfredson H, Forsgren S. Distribution of general (PGP 9.5) and sensory (substance P/CGRP) innervations in the human patellar tendon. *Knee Surg Sports Traumatol Arthrosc.* 2006;14(2):125–32.
156. Griffin LY, Agel J, Albohm MJ, Arendt EA, Dick RW, Garrett WE, et al. Noncontact anterior cruciate ligament injuries: risk factors and prevention strategies. *J Am Acad Orthop Surg.* 2000;8(3):141–50.
157. Huston LJ, Greenfield ML, Wojtys EM. Anterior cruciate ligament injuries in the female athlete. Potential risk factors. *Clin Orthop Relat Res.* 2000;(372):50–63.
158. Noyes FR, Mooar PA, Matthews DS, Butler DL. The symptomatic anterior cruciate-deficient knee. Part I: the long-term functional disability in athletically active individuals. *J Bone Joint Surg Am.* 1983;65(2):154–62.

159. Boden BP, Dean GS, Feagin JA Jr, Garrett WE Jr. Mechanisms of anterior cruciate ligament injury. *Orthopedics*. 2000;23(6):573-8.
160. Arendt EA, Agel J, Dick R. Anterior cruciate ligament injury patterns among collegiate men and women. *J Athl Train*. 1999;34(2):86-92.
161. Agel J, Arendt EA, Bershadsky B. Anterior cruciate ligament injury in national collegiate athletic association basketball and soccer: a 13-year review. *Am J Sports Med*. 2005;33(4):524-30.
162. Moses B, Orchard J, Orchard J. Systematic review: Annual incidence of ACL injury and surgery in various populations. *Res Sports Med*. 2012;20(3-4):157-79.
163. Granan LP, Bahr R, Steindal K, Furnes O, Engebretsen L. Development of a national cruciate ligament surgery registry: the Norwegian national knee ligament registry. *Am J Sports Med*. 2008;36(2):308-15.
164. Ahldén M, Samuelsson K, Sernert N, Forssblad M, Karlsson J, Kartus J. The Swedish national anterior cruciate ligament register: a report on baseline variables and outcomes of surgery for almost 18,000 patients. *Am J Sports Med*. 2012;40(10):2230-5.
165. Lind M, Menhert F, Pedersen AB. The first results from the Danish ACL reconstruction registry : epidemiologic and 2 year follow-up results from 5,818 knee ligament reconstructions. *Knee Surg Sports Traumatol Arthrosc*. 2009;17(2):117-24.
166. Schilaty ND, Nagelli C, Bates NA, Sanders TL, Krych AJ, Stuart MJ, et al. Incidence of second anterior cruciate ligament tears and identification of associated risk factors from 2001 to 2010 using a geographic database. *Orthop J Sports Med*. 2017;5(8):1-8.
167. Hewett TE, Di Stasi SL, Myer GD. Current concepts for injury prevention in athletes after anterior cruciate ligament reconstruction. *Am J Sports Med*. 2013;41(1):216-24.
168. Roos H, Adalberth T, Dahlberg L, Lohmander LS. Osteoarthritis of the knee after Injury to the anterior cruciate ligament or meniscus: the influence of time and age. *Osteoarthritis Cartilage*. 1995;3(4):261-7.
169. Lohmander LS, Östenberg A, Englund M, Roos H. High prevalence of knee osteoarthritis, pain, and functional limitations in female soccer players twelve years after anterior cruciate ligament injury. *Arthritis Rheum*. 2004;50(10):3145-52.
170. Lohmander LS, Englund PM, Dahl LL, Roos EM. The long-term consequence of anterior cruciate ligament and meniscus injuries: osteoarthritis. *Am J Sports Med*. 2007;35(10):1756-69.

171. Øiestad BE, Holm I, Aune AK, Gunderson R, Myklebust G, Engebretsen L, et al. Knee function and prevalence of knee osteoarthritis after anterior cruciate ligament reconstruction: a prospective study with 10 to 15 years of follow-up. *Am J Sports Med.* 2010;38(11):2201–10.
172. Lohmander LS, Roos H. Knee ligament injury, surgery and osteoarthritis: truth or consequences? *Acta Orthop Scand.* 1994;65(6):605–9.
173. Øiestad BE, Engebretsen L, Storheim K, Risberg MA. Knee osteoarthritis after anterior cruciate ligament injury: a systematic review. *Am J Sports Med.* 2009;37(7):1434–43.
174. Myklebust G, Bahr R. Return to play guidelines after anterior cruciate ligament surgery. *Br J Sports Med.* 2005;39(3):127–31.
175. Murray MM, Martin SD, Martin TL, Spector M. Histological changes in the human anterior cruciate ligament after rupture. *J Bone Joint Surg Am.* 2000;82-A(10):1387–97.
176. Hayashi K, Frank JD, Dubinsky C, Zhengling H, Markel MD, Manley PA, et al. Histologic changes in ruptured canine cranial cruciate ligament. *Vet Surg.* 2003;32(3):269–77.
177. Marshall SW, Padua DA, McGrath M. Incidence of ACL injury. In Hewett TE, Schultz SJ, Griffin LY, editors. *Understanding and preventing noncontact ACL injuries.* Champaign (Illinois): Human Kinetics; 2007. p.5–30.
178. Marshall SW. Recommendations for defining and classifying anterior cruciate ligament injuries in epidemiologic studies. *J Athl Train.* 2010;45(5):516–8.
179. Hunter RE. Skiing Injuries. *Am J Sports Med.* 1999;27(3):381–9.
180. Uhorchak JM, Scoville CR, Williams GN, Arciero RA, St Pierre P, Taylor DC. Risk factors associated with noncontact injury of the anterior cruciate ligament: a prospective four-year evaluation of 859 West Point cadets. *Am J Sports Med.* 2003;31(6):831–42.
181. Harvie P, Ostlere SJ, Teh J, McNally EG, Clipsham K, Burston BJ, et al. Genetic influences in the aetiology of tears of the rotator cuff. Sibling risk of full-thickness tear. *J Bone Joint Surg Br.* 2004;86(5):696–700.
182. Stępień-Słodkowska M, Ficek K, Kaczmarczyk M, Maciejewska A, Sawczuk M, Eider J, et al. Influence of biological factors on injuries occurrence in the Polish population. *Ann Agric Environ Med.* 2016;23(2):315–8.
183. Ficek K, Cieszczyk P, Kaczmarczyk M, Maciejewska-Karłowska A, Sawczuk M, Cholewinski J, et al. Gene variants within the COL1A1 gene are associated with reduced anterior cruciate ligament injury in professional soccer players. *J Sci Med Sport.* 2013;16(5):396–400.

184. Khoschnau S, Melhus H, Jacobson A, Rahme H, Bengtsson H, Ribom E, et al. Type I collagen alpha1 Sp1 polymorphism and the risk of cruciate ligament ruptures or shoulder dislocations. *Am J Sports Med.* 2008;36(12):2432–6.
185. Posthumus M, September AV, Keegan M, O’Cuinneagain D, van der Merwe W, Schwellnus MP, et al. Genetic risk factors for anterior cruciate ligament ruptures: COL1A1 gene variant. *Br J Sports Med*; 2009;43(5):352–6.
186. Stępień-Słodkowska M, Ficek K, Eider J, Leońska-Duniec A, Maciejewska-Karłowska A, Sawczuk M, et al. The +1245G/T polymorphisms in the collagen type I alpha 1 (col1a1) gene in polish skiers with anterior cruciate ligament injury. *Biol Sport.* 2013;30(1):57–60.
187. Collins M, Posthumus M, Schwellnus MP. The COL1A1 gene and acute soft tissue ruptures. *Br J Sports Med.* 2010;44(14):1063–4.
188. O’Connell K, Knight H, Ficek K, Leonska-Duniec A, Maciejewska-Karłowska A, Sawczuk M, et al. Interactions between collagen gene variants and risk of anterior cruciate ligament rupture. *Eur J Sport Sci.* 2015;15(4):341-50.
189. Stępień-Słodkowska M, Ficek K, Maciejewska-Karłowska A, Sawczuk M, Ziętek P, Król P, et al. Overrepresentation of the COL3A1 AA genotype in Polish skiers with anterior cruciate ligament injury. *Biol Sport.* 2015;32(2):143–7.
190. Mokone GG, Schwellnus MP, Noakes TD, Collins M. The COL5A1 gene and Achilles tendon pathology. *Scand J Med Sci Sports.* 2006;16(1):19–26.
191. September AV, Cook J, Handley CJ, van der Merwe L, Schwellnus MP, Collins M. Variants within the COL5A1 gene are associated with Achilles tendinopathy in two populations. *Br J Sports Med.* 2009;43(5):357–65.
192. Posthumus M, September AV, O’Cuinneagain D, van der Merwe W, Schwellnus MP, Collins M. The COL5A1 gene is associated with increased risk of anterior cruciate ligament ruptures in female participants. *Am J Sports Med.* 2009;37(11):2234–40.
193. Brown KL, Seale KB, El Khoury LY, Posthumus M, Ribbans WJ, Raleigh SM, et al. Polymorphisms within the COL5A1 gene and regulators of the extracellular matrix modify the risk of Achilles tendon pathology in a British case-control study. *J Sports Sci.* 2017;35(15):1475–83.
194. Willard K, Mannion S, Saunders CJ, Collins M, September AV. The interaction of polymorphisms in extracellular matrix genes and underlying miRNA motifs that modulate susceptibility to anterior cruciate ligament rupture. *J Sci Med Sport.* 2018;21(1):22–8.

195. Abrahams Y, Laguette MJ, Prince S, Collins M. Polymorphisms within the COL5A1 3'-UTR that alters mRNA structure and the MIR608 gene are associated with Achilles tendinopathy. *Ann Hum Genet.* 2013;77(3):204–14.
196. Hay M, Patricios J, Collins R, Branfield A, Cook J, Handley CJ, et al. Association of type XI collagen genes with chronic Achilles tendinopathy in independent populations from South Africa and Australia. *Br J Sports Med.* 2013;47(9):569–74.
197. Posthumus M, September AV, O'Cuinneagain D, van der Merwe W, Schwellnus MP, Collins M. The association between the COL12A1 gene and anterior cruciate ligament ruptures. *Br J Sports Med.* 2010;44(16):1160–5.
198. Saunders CJ, van der Merwe L, Posthumus M, Cook J, Handley CJ, Collins M, et al. Investigation of variants within the COL27A1 and TNC genes and Achilles tendinopathy in two populations. *J Orthop Res.* 2013;31(4):632–7.
199. Mokone GG, Gajjar M, September AV, Schwellnus MP, Greenberg J, Noakes TD, et al. The guanine-thymine dinucleotide repeat polymorphism within the tenascin-C gene is associated with achilles tendon injuries. *Am J Sports Med.* 2005;33(7):1016–21.
200. El Khoury L, Posthumus M, Collins M, van der Merwe W, Handley CJ, Raleigh SM. ELN and FBN2 gene variants as risk factors for two sports-related musculoskeletal injuries. *Int J Sports Med.* 2015;36(4):333–7.
201. Mannion S, Mtintsilana A, Posthumus M, van der Merwe W, Hobbs H, Collins M, et al. Genes encoding proteoglycans are associated with the risk of anterior cruciate ligament ruptures. *Br J Sports Med.* 2014;48(22):1640–6.
202. Ciężczyk P, Willard K, Gronek P, Zmijewski P, Trybek G, Gronek J, et al. Are genes encoding proteoglycans really associated with the risk of anterior cruciate ligament rupture? *Biol Sport.* 2017;34(2):97–103.
203. Posthumus M, Collins M, van der Merwe L, O'Cuinneagain D, van der Merwe W, Ribbans WJ, et al. Matrix metalloproteinase genes on chromosome 11q22 and the risk of anterior cruciate ligament (ACL) rupture. *Scand J Med Sci Sports.* 2012;22(4):523–33.
204. Malila S, Yuktanandana P, Saowaprut S, Jiamjarasrangsi W, Honsawek S. Association between matrix metalloproteinase-3 polymorphism and anterior cruciate ligament ruptures. *Genet Mol Res.* 2011;10(4):4158–65.
205. Gibbon A, Hobbs H, van der Merwe W, Raleigh SM, Cook J, Handley CJ, et al. The MMP3 gene in musculoskeletal soft tissue injury risk profiling : a study in two independent sample groups. *J Sports Sci.* 2017;35(7):655–662.
206. 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):514–20.

207. El Khoury L, Ribbans WJ, Raleigh SM. MMP3 and TIMP2 gene variants as predisposing factors for Achilles tendon pathologies: attempted replication study in a British case-control cohort. *Meta Gene*. 2016;9:52–5.
208. El Khoury L, Posthumus M, Collins M, Handley CJ, Cook J, Raleigh SM. Polymorphic variation within the ADAMTS2, ADAMTS14, ADAMTS5, ADAM12 and TIMP2 genes and the risk of Achilles tendon pathology: a genetic association study. *J Sci Med Sport*. 2013;16(6):493–8.
209. Rahim M, Gibbon A, Hobbs H, van der Merwe W, Posthumus M, Collins M, et al. The association of genes involved in the angiogenesis-associated signaling pathway with risk of anterior cruciate ligament rupture. *J Orthop Res*. 2014;32(12):1612–8.
210. Rahim M, El Khoury LY, Raleigh SM, Ribbans WJ, Posthumus M, Collins M, et al. Human genetic variation, sport and exercise medicine, and Achilles tendinopathy: role for angiogenesis-associated genes. *OMICS*. 2016;20(9):520–7.
211. Rahim M, Hobbs H, van der Merwe W, Posthumus M, Collins M, September AV. Investigation of angiogenesis genes with anterior cruciate ligament rupture risk in a South African population. *J Sports Sci*. 2018;36(5):551–7.
212. Posthumus M, Collins M, Cook J, Handley CJ, Ribbans WJ, Smith RK, et al. Components of the transforming growth factor-beta family and the pathogenesis of human Achilles tendon pathology--a genetic association study. *Rheumatology (Oxford)*. 2010;49(11):2090–7.
213. Matsuda A, Hirota T, Akahoshi M, Shimizu M, Tamari M, Miyatake A, et al. Coding SNP in tenascin-C Fn-III-D domain associates with adult asthma. *Hum Mol Genet*. 2005;14(19):2779–86.
214. Kluger R, Burgstaller J, Vogl C, Brem G, Skultety M, Mueller S. Candidate gene approach identifies six SNPs in tenascin-C (TNC) associated with degenerative rotator cuff tears. *J Orthop Res*. 2017;35(4):894–901.
215. Kluger R, Huber KR, Seely PG, Berger CE, Frommlet F. Novel tenascin-C haplotype modifies the risk for a failure to heal after rotator cuff repair. *Am J Sports Med*. 2017;45(13):2955–64.
216. Chiquet-Ehrismann R, Tucker RP. Connective tissues: Signalling by tenascins. *Int J Biochem Cell Biol*. 2004;36(6):1085–9.
217. Orend G, Chiquet-Ehrismann R. Tenascin-C induced signaling in cancer. *Cancer Lett*. 2006;244(2):143–63.
218. Giblin SP, Midwood KS. Tenascin-C: form versus function. *Cell Adhes Migr*. 2015;9(1–2):48–82.

219. Erickson HP, Inglesias JL. A six-armed oligomer isolated from cell surface fibronectin preparations. *Nature*. 1984;311(5983):267-9.
220. Jones FS, Jones PL. The tenascin family of ECM glycoproteins: structure, function, and regulation during embryonic development and tissue remodeling. *Dev Dyn*. 2000;218(2):235–59.
221. Midwood KS, Chiquet M, Tucker RP, Orend G. Tenascin-C at a glance. *J Cell Sci*. 2016;129(23):4321–7.
222. Jones PL, Jones FS. Tenascin-C in development and disease: gene regulation and cell function. *Matrix Biol*. 2000;19(7):581–96.
223. Chiquet-Ehrismann R, Chiquet M. Tenascins: regulation and putative functions during pathological stress. *J Pathol*. 2003;200(4):488–99.
224. Chiquet M, Sarasa-Renedo A, Huber F, Flück M. How do fibroblasts translate mechanical signals into changes in extracellular matrix production? *Matrix Biol*. 2003;22(1):73–80.
225. Chiquet M, Sarasa-Renedo A, Tunç-Civelek V. Induction of tenascin-C by cyclic tensile strain versus growth factors: Distinct contributions by Rho/ROCK and MAPK signaling pathways. *Biochim Biophys Acta*. 2004;1693(3):193–204.
226. Flück M, Tunc-Civelek V, Chiquet M. Rapid and reciprocal regulation of tenascin-C and tenascin-Y expression by loading of skeletal muscle. *J Cell Sci*. 2000;113(Pt 20):3583–91.
227. Flück M, Chiquet M, Schmutz S, Mayet-Sornay MH, Desplanches D. Reloading of atrophied rat soleus muscle induces tenascin-C expression around damaged muscle fibers. *Am J Physiol Regul Integr Comp Physiol*. 2003;284(3):R792-801.
228. Järvinen T, Jozsa L, Kannus P, Järvinen TL, Kvist M, Hurme T, et al. Mechanical loading regulates tenascin-C expression in the osteotendinous junction. *J Cell Sci*. 1999;112 (Pt 18):3157-66.
229. Orend G, Chiquet-Ehrismann R. Adhesion modulation by antiadhesive molecules of the extracellular matrix. *Exp Cell Res*. 2000;261(1):104–10.
230. Saoncella S, Echtermeyer F, Denhez F, Nowlen JK, Mosher DF, Robinson SD, et al. Syndecan-4 signals cooperatively with integrins in a Rho-dependent manner in the assembly of focal adhesions and actin stress fibers. *Proc Natl Acad Sci U S A*. 1999;96(6):2805–10.

231. Salmivirta M, Elenius K, Vainio S, Hofer U, Chiquet-Ehrismann R, Thesleff I, et al. Syndecan from embryonic tooth mesenchyme binds tenascin. *J Biol Chem.* 1991;266(12):7733–9.
232. Murphy UJ, Lightner VA, Aukhil I, Yan YZ, Erickson HP, Hook M. Focal adhesion integrity is downregulated by the alternatively spliced domain of human tenascin. *J Cell Biol.* 1991;115(4):1127–36.
233. Fischer D, Brown-Ludi M, Schulthess T, Chiquet-Ehrismann R. Concerted action of tenascin-C domains in cell adhesion, anti-adhesion and promotion of neurite outgrowth. *J Cell Sci.* 1997;110(Pt 13):1513–22.
234. Sarasa-Renedo A, Tunç-Civelek V, Chiquet M. Role of RhoA/ROCK-dependent actin contractility in the induction of tenascin-C by cyclic tensile strain. 2006;312(8):1361-70.
235. Chiquet M, Tunç-Civelek V, Sarasa-Renedo A. Gene regulation by mechanotransduction in fibroblasts. *Appl Physiol Nutr Metab.* 2007;32(5):967–73.
236. Maier S, Lutz R, Gelman L, Sarasa-Renedo A, Schenk S, Grashoff C, et al. Tenascin-C induction by cyclic strain requires integrin-linked kinase. *Biochim Biophys Acta.* 2008;1783(6):1150–62.
237. Imanaka-Yoshida K, Aoki H. Tenascin-C and mechanotransduction in the development and diseases of cardiovascular system. *Front Physiol.* 2014;5:283.
238. Oberhauser AF, Marszalek PE, Erickson HP, Fernandez JM. The molecular elasticity of the extracellular matrix protein tenascin. *Nature.* 1998;393(6681):181–5.
239. Pace JM, Corrado M, Missero C, Byers PH. Identification, characterization and expression analysis of a new fibrillar collagen gene, COL27A1. *Matrix Biol.* 2003;22(1):3–14.
240. Hjorten R, Hansen U, Underwood RA, Telfer HE, Fernandes RJ, Krakow D, et al. Type XXVII collagen at the transition of cartilage to bone during skeletogenesis. *Bone.* 2007;41(4):535–42.
241. Gonzaga-Jauregui C, Gamble CN, Yuan B, Penney S, Jhangiani S, Muzny DM, et al. Mutations in COL27A1 cause Steel syndrome and suggest a founder mutation effect in the Puerto Rican population. *Eur J Hum Genet.* 2015;23(3):342–6.
242. Somerville RP, Oblander SA, Apte SS. Matrix metalloproteinases: old dogs with new tricks. *Genome Biol.* 2003;4(6):216. doi: 10.1186/gb-2003-4-6-216.
243. Stamenkovic I. Extracellular matrix remodelling: the role of matrix metalloproteinases. *J Pathol.* 2003;200(4):448–64.

244. Suzuki K, Enghild JJ, Morodomi T, Salvesen G, Nagase H. Mechanisms of activation of tissue procollagenase by matrix metalloproteinase 3 (stromelysin). *Biochemistry*. 1990;29(44):10261–70.
245. Visse R, Nagase H. Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. *Circ Res*. 2003;92(8):827–39.
246. Ye S, Eriksson P, Hamsten A, Kurkinen M, Humphries SE, Henney AM. Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression. *J Biol Chem*. 1996;271(22):13055–60.
247. Ireland D, Harrall R, Curry V, Holloway G, Hackney R, Hazleman B, et al. Multiple changes in gene expression in chronic human Achilles tendinopathy. *Matrix Biol*. 2001;20(3):159–69.
248. Jones GC, Corps AN, Pennington CJ, Clark IM, Edwards DR, Bradley MM, et al. Expression profiling of metalloproteinases and tissue inhibitors of metalloproteinases in normal and degenerate human Achilles tendon. *Arthritis Rheum*. 2006;54(3):832–42.
249. Alfredson H, Lorentzon M, Bäckman S, Bäckman A, Lerner UH. cDNA-arrays and real-time quantitative PCR techniques in the investigation of chronic Achilles tendinosis. *J Orthop Res*. 2003;21(6):970–5.
250. Beye JA, Hart DA, Bray RC, McDougall JJ, Salo PT. Injury-induced changes in mRNA levels differ widely between anterior cruciate ligament and medial collateral ligament. *Am J Sports Med*. 2008;36(7):1337–46.
251. Zhang J, Yang L, Tang Z, Xue R, Wang Y, Luo Z, et al. Expression of MMPs and TIMPs family in human ACL and MCL fibroblasts. *Connect Tissue Res*. 2009;50(1):7–13.
252. Larruskain J, Celorrio D, Barrio I, Odriozola A, Gil SM, Fernandez-Lopez JR, et al. Genetic variants and hamstring injury in soccer: an association and validation study. *Med Sci Sports Exerc*. 2018;50(2):361–368.
253. Ling Y, Peng C, Liu C, Zhang N, Yue S. Gene polymorphism of IL-6 and MMP3 decreases passive range of motion after rotator cuff repair. *Int J Clin Exp Pathol*. 2015;8(5):5709–14.
254. Foster BP, Morse CI, Onambele GL, Williams AG. Variants within the MMP3 gene and patellar tendon properties in vivo in an asymptomatic population. *Eur J Appl Physiol*. 2014;114(12):2625–34.
255. Burger MC, De Wet H, Collins M. Matrix metalloproteinase genes on chromosome 11q22 and risk of carpal tunnel syndrome. *Rheumatol Int*. 2016;36(3):413–9.

256. Huerre C, Junien C, Weil D, Chu ML, Morabito M, Van Cong N, et al. Human type I procollagen genes are located on different chromosomes. *Proc Natl Acad Sci U S A*. 1982;79(21):6627-30.
257. Junien C, Weil D, Myers JC, Van Cong N, Chu ML, Foubert C, et al. Assignment of the human pro alpha 2(I) collagen structural gene (COL1A2) to chromosome 7 by molecular hybridization. *Am J Hum Genet*. 1982;34(3):381-7.
258. Willing MC, Deschenes SP, Scott DA, Byers PH, Slayton RL, Pitts SH, et al. Osteogenesis imperfecta type I: molecular heterogeneity for COL1A1 null alleles of type I collagen. *Am J Hum Genet*. 1994;55(4):638-47.
259. Pollitt R, McMahon R, Nunn J, Bamford R, Afifi A, Bishop N, et al. Mutation analysis of *COL1A1* and *COL1A2* in patients diagnosed with osteogenesis imperfecta type I-IV. *Hum Mutat*. 2006;27(7):716. doi: 10.1002/humu.9430.
260. D'Alessio M, Ramirez F, Blumberg BD, Wirtz MK, Rao VH, Godfrey MD, et al. Characterization of a COL1A1 splicing defect in a case of Ehlers-Danlos syndrome type VII: further evidence of molecular homogeneity. *Am J Hum Genet*. 1991;49(2):400-6.
261. Byers PH, Duvic M, Atkinson M, Robinow M, Smith LT, Krane SM, et al. Ehlers-Danlos syndrome type VIIA and VIIB result from splice-junction mutations or genomic deletions that involve exon 6 in the COL1A1 and COL1A2 genes of type I collagen. *Am J Med Genet*. 1997;72(1):94-105.
262. Nuytinck L, Freund M, Lagae L, Pierard GE, Hermanns-Le T, De Paepe A. Classical Ehlers-Danlos syndrome caused by a mutation in type I collagen. *Am J Hum Genet*. 2000;66(4):1398-402.
263. Gensure RC, Mäkitie O, Barclay C, Chan C, DePalma SR, Bastepe M, et al. A novel COL1A1 mutation in infantile cortical hyperostosis (Caffey disease) expands the spectrum of collagen-related disorders. *J Clin Invest*. 2005;115(5):1250-7.
264. Ralston SH, Uitterlinden AG, Brandi ML, Balcells S, Langdahl BL, Lips P, et al. Large-scale evidence for the effect of the COL1A1 Sp1 polymorphism on osteoporosis outcomes: The GENOMOS study. *PLoS Med*. 2006;3(4):e90. doi: 10.1371/journal.pmed.0030090.
265. Singh M, Singh P, Singh S, Juneja PK, Kaur T. A haplotype derived from the common variants at the -1997G/T and Sp1 binding site of the COL1A1 gene influences risk of postmenopausal osteoporosis in India. *Rheumatol Int*. 2013;33(2):501-6.
266. Skorupski P, Król J, Staręga J, Adamiak A, Jankiewicz K, Rechberger T. An alpha-1 chain of type I collagen Sp1-binding site polymorphism in women suffering from stress urinary incontinence. *Am J Obstet Gynecol*. 2006;194(2):346-50.

267. Tilkeridis C, Bei T, Garantziotis S, Stratakis CA. Association of a COL1A1 polymorphism with lumbar disc disease in young military recruits. *J Med Genet.* 2005;42(7):e44. doi: 10.1136/jmg.2005.033225.
268. Posthumus M, September AV, Schwellnus MP, Collins M. Investigation of the Sp1-binding site polymorphism within the COL1A1 gene in participants with Achilles tendon injuries and controls. *J Sci Med Sport.* 2009;12(1):184–9.
269. Pruna R, Artells R, Ribas J, Montoro B, Cos F, Muñoz C, 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.
270. Erduran M, Altinisik J, Meric G, Ates O, Ulusal AE, Akseki D. Is Sp1 binding site polymorphism within COL1A1 gene associated with tennis elbow? *Gene.* 2014;537(2):308–11.
271. Prabhakar S, John R, Dhillon MS, Anand A, Sharma K, Bammidi S. Are COL1A1 gene polymorphisms associated with anterior cruciate ligament tear in the Indian population? Results of a preliminary case-control study. *Muscles Ligaments Tendons J.* 2018;8(1):15–22.
272. Wang C, Li H, Chen K, Wu B, Liu H. Association of polymorphisms rs1800012 in COL1A1 with sports-related tendon and ligament injuries: a meta-analysis. *Oncotarget.* 2017;8(16):27627–34.
273. Pluijm SM, van Essen HW, Bravenboer N, Uitterlinden AG, Smit JH, Pols HA. Collagen type I 1 Sp1 polymorphism, osteoporosis, and intervertebral disc degeneration in older men and women. *Ann Rheum Dis.* 2004;63(1):71–7.
274. Jin H, Evangelou E, Ioannidis JP, Ralston SH. Polymorphisms in the 5' flank of COL1A1 gene and osteoporosis: meta-analysis of published studies. *Osteoporos Int.* 2011;22(3):911–21.
275. Mann V, Hobson EE, Li B, Stewart TL, Grant SF, Robins SP, 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.
276. Jin H, van't Hof RJ, Albagha OM, Ralston SH. Promoter and intron 1 polymorphisms of COL1A1 interact to regulate transcription and susceptibility to osteoporosis. *Hum Mol Genet.* 2009;18(15):2729–38.
277. García-Giralt N, Enjuanes A, Bustamante M, Mellibovsky L, Nogués X, Carreras R, et al. In vitro functional assay of alleles and haplotypes of two COL1A1-promoter SNPs. *Bone.* 2005;36(5):902–8.

278. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
279. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304–51.
280. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463–7.
281. Rizzo JM, Buck MJ. Key principles and clinical applications of “next-generation” DNA sequencing. *Cancer Prev Res (Phila)*. 2012;5(7):887–900.
282. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013;155(1):27–38.
283. Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. *Nat Genet*. 2012;44(6):623–30.
284. Baird AE, Carter SD, Innes JF, Ollier W, Short A. Genome-wide association study identifies genomic regions of association for cruciate ligament rupture in Newfoundland dogs. *Anim Genet*. 2014;45(4):542–9.
285. Baker LA, Kirkpatrick B, Rosa GJ, Gianola D, Valente B, Sumner JP, et al. Genome-wide association analysis in dogs implicates 99 loci as risk variants for anterior cruciate ligament rupture. *PLoS One*. 2017;12(4):e0173810.
286. Tashjian RZ, Granger EK, Farnham JM, Cannon-Albright LA, Teerlink CC. Genome-wide association study for rotator cuff tears identifies two significant single-nucleotide polymorphisms. *J Shoulder Elb Surg*. 2016;25(2):174–9.
287. Kim SK, Roos TR, Roos AK, Kleimeyer JP, Ahmed MA, Goodlin GT, et al. Genome-wide association screens for Achilles tendon and ACL tears and tendinopathy. *PLoS One*. 2017;12(3):e0170422. doi: 10.1371/journal.pone.0173810.
288. van Dijk CN, van Sterkenburg MN, Wiegerinck JI, Karlsson J, Maffulli N. Terminology for Achilles tendon related disorders. *Knee Surg Sports Traumatol Arthrosc*. 2011;19(5):835–41.
289. Lacey S, Chung J, Lin H. A comparison of whole genome sequencing with exome sequencing for family-based association studies. *BMC Proc*. 2014;8(Suppl 1):S38. doi: 10.1186/1753-6561-8-S1-S38.
290. Galperin MY, Koonin EV. From complete genome sequence to “complete” understanding? *Trends Biotechnol*. 2010;28(8):398–406.
291. Rabbani B, Tekin M, Mahdih N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet*. 2014;59(1):5–15.

292. Grody WW, Thompson BH, Hudgins L. Whole-exome/genome sequencing and genomics. *Pediatrics*. 2013;132(Suppl 3):S211-5.
293. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121–32.
294. Caso E, Maestro A, Sabiers CC, Godino M, Caracuel Z, Pons J, et al. Whole-exome sequencing analysis in twin sibling males with an anterior cruciate ligament rupture. *Injury*. 2016;47 (Suppl 3):S41–50.
295. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al. STrengthening the REporting of Genetic Association studies (STREGA)--an extension of the STROBE statement. *Eur J Clin Invest*. 2009;39(4):247-66.
296. Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, et al. Strong maternal Khoisan contribution to the South African Coloured population: a case of gender-biased admixture. *Am J Hum Genet*. 2010;86(4):611–20.
297. Patterson N, Petersen DC, van der Ross RE, Sudoyo H, Glashoff RH, Marzuki S, et al. Genetic structure of a unique admixed population: implications for medical research. *Hum Mol Genet*. 2010;19(3):411-9.
298. Rickaby R, El Khoury L, Ribbans WJ, Raleigh SM. Variation within three apoptosis associated genes as potential risk factors for Achilles tendinopathy in a British based case-control cohort. *Gene*. 2015;571(2):167–71.
299. Lahiri DK, Nurnberger JI Jr. A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nucleic Acids Res*. 1991;19(19):5444.
300. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93.
301. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi: 10.1093/nar/gkq603.
302. Saunders CJ, Jalali Sefid Dashti M, Gamielien J. Semantic interrogation of a multi knowledge domain ontological model of tendinopathy identifies four strong candidate risk genes. *Sci Rep*. 2016;6:19820. doi: 10.1038/srep19820.
303. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31(21):3555–7.

304. Eser B, Eser O, Yuksel Y, Aksit H, Karavelioglu E, Tosun M, et al. Effects of MMP1 and MMP3 gene polymorphisms on gene expression and protein level in lumbar disc herniation. *Genet Mol Res*. 2016;15(3).doi: 10.4238/gmr.15038669.
305. Huang X, Ye Q, Zhang Z, Huang X, Zhu Z, Chen Y, et al. Association of matrix metalloproteinase-3 gene 5A/6A polymorphism with the recurrence of ischemic stroke : a prospective observational study. *Brain Res*. 2017;1674:55–61.
306. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.
307. Gonzalez R, Armengol L, Guino E, Sole X, Moreno V. SNP-based whole genome association studies. Version 1.9-2. 2014.
308. Warnes G, Gorjanc G, Leisch F, Man M. Population Genetics. Version 1.3.8.1. 2013.
309. Sinnwell J., Schaid D. Statistical analysis of haplotypes with traits and covariates when linkage is ambiguous. Version 1.7.1. 2015.
310. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316(7139):1236–8.
311. Lu Y, Quan C, Chen H, Bo X, Zhang C. 3DSNP: a database for linking human noncoding SNPs to their three-dimensional interacting genes. *Nucleic Acids Res*. 2017;45(D1):D643–9.
312. Ding Y, Chan CY, Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*. 2005;11(8):1157–66.
313. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc*. 2016;11(1):1–9.
314. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248–9.
315. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310–5.
316. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*. 2013;34(1):57–65.
317. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648–60.

318. Valdivieso P, Toigo M, Hoppeler H, Flück M. T/T homozygosity of the tenascin-C gene polymorphism rs2104772 negatively influences exercise-induced angiogenesis. *PLoS One*. 2017;12(4):e0174864. doi: 10.1371/journal.pone.0174864.
319. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1620):20120362. doi: 10.1098/rstb.2012.0362.
320. Melzer D, Perry JR, Hernandez D, Corsi AM, Stevens K, Rafferty I, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet*. 2008;4(5):e1000072. doi: 10.1371/journal.pgen.1000072.
321. Riley GP, Harrall RL, Cawston TE, Hazleman BL, Mackie EJ. Tenascin-C and human tendon degeneration. *Am J Pathol*. 1996;149(3):933–43.
322. Lourdasamy A, Newhouse S, Lunnon K, Proitsi P, Powell J, Hodges A, et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum Mol Genet*. 2012;21(16):3719–26.
323. Chung CY, Zardi L, Erickson HP. Binding of tenascin-C to soluble fibronectin and matrix fibrils. *J Biol Chem*. 1995;270(48):29012–7.
324. Chung CY, Erickson HP. Glycosaminoglycans modulate fibronectin matrix assembly and are essential for matrix incorporation of tenascin-C. *J Cell Sci*. 1997;110(Pt 12):1413–9.
325. Lamblin N, Bauters C, Hermant X, Lablanche JM, Helbecque N, Amouyel P. Polymorphisms in the promoter regions of MMP2, MMP3, MMP9 and MMP12 genes as determinants of aneurysmal coronary artery disease. *J Am Coll Cardiol*. 2002;40(1):43–8.
326. Beyzade S, Zhang S, Wong YK, Day IN. M, Eriksson P, Ye S. Influences of matrix metalloproteinase-3 gene variation on extent of coronary atherosclerosis and risk of myocardial infarction. *J Am Coll Cardiol*. 2003;41(12):2130–7.
327. 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.
328. Borghaei RC, Rawlings PL Jr, Javadi M, Woloshin J. NF-kappaB binds to a polymorphic repressor element in the MMP3 promoter. *Biochem Biophys Res Commun*. 2004;316(1):182–8.
329. Ricketts C, Zeegers MP, Lubinski J, Maher ER. Analysis of germline variants in CDH1, IGFBP3, MMP1, MMP3, STK15 and VEGF in familial and sporadic renal cell carcinoma. *PLoS One*. 2009;4(6):e6037. doi: 10.1371/journal.pone.0006037.

330. Karayasheva D, Glushkova M, Boteva E, Mitev V, Kadiyska T. Association study for the role of matrix metalloproteinases 2 and 3 gene polymorphisms in dental caries susceptibility. *Arch Oral Biol.* 2016;68:9-12.
331. Letra A, Silva RM, Rylands RJ, Silveira EM, de Souza AP, Wendell SK, et al. MMP3 and TIMP1 variants contribute to chronic periodontitis and may be implicated in disease progression. *J Clin Periodontol.* 2012;39(8):707–16.
332. Han Z, Huang H, Gao Y, Huang Q. Functional annotation of Alzheimer’s disease associated loci revealed by GWASs. *PLoS One.* 2017;12(6):e0179677. doi: 10.1371/journal.pone.0179677.
333. Galan JJ, Guarducci E, Nuti F, Gonzalez A, Ruiz M, Ruiz A, et al. Molecular analysis of estrogen receptor alpha gene AGATA haplotype and SNP12 in European populations: potential protective effect for cryptorchidism and lack of association with male infertility. *Hum Reprod.* 2007;22(2):444–9.
334. Chaudhury S, Xia Z, Thakkar D, Hakimi O, Carr AJ. Gene expression profiles of changes underlying different-sized human rotator cuff tendon tears. *J Shoulder Elbow Surg.* 2016;25(10):1561–70.
335. Yoshihara Y, Hamada K, Nakajima T, Fujikawa K, Fukuda H. Biochemical markers in the synovial fluid of glenohumeral joints from patients with rotator cuff tear. *J Orthop Res.* 2001;19(4):573–9.
336. Leal MF, Caires Dos Santos L, Martins de Oliveira A, Santoro Belangero P, Antô Figueiredo E, et al. Epigenetic regulation of metalloproteinases and their inhibitors in rotator cuff tears. *PLoS One.* 2017;12(9):e0184141. doi: 10.1371/journal.pone.0184141.
337. Lian K, Zmuda JM, Nevitt MC, Lui L, Hochberg MC, Greene D, et al. Type I collagen alpha1 Sp1 transcription factor binding site polymorphism is associated with reduced risk of hip osteoarthritis defined by severe joint space narrowing in elderly women. *Arthritis Rheum.* 2005;52(5):1431–6.
338. Stępień-Słodkowska M, Ficek K, Ziętek P, Kaczmarczyk M, Łubkowska W, Szark-Eckardt M, et al. Whether the combination of COL1A1 gene polymorphisms may be a marker of the risk of injury? *J Sport Rehabil.* 2017;26 (3):234–8.
339. de Wit E, Delpont W, Rugamika CE, Meintjes A, Möller M, Van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum Genet.* 2010;128(2):145–53.
340. Ireland ML. Sport-specific injury mechanisms associated with pivoting, cutting and landing. In Hewett TE, Shultz SJ, Griffin LY, editors. *Understanding and preventing noncontact ACL injuries.* Champaign (Illinois): Human Kinetics; 2007. p.163-171.

341. Patnala R, Clements J, Batra J. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genet.* 2013;14:39. doi: 10.1186/1471-2156-14-39.
342. Chiquet-Ehrismann R, Matsuoka Y, Hofer U, Spring J, Bernasconi C, Chiquet M. Tenascin variants: differential binding to fibronectin and distinct distribution in cell cultures and tissues. *Cell Regul.* 1991;2(11):927–38.
343. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol.* 2010;30(6):1411–20.
344. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun.* 2015;6:10069. doi: 10.1038/ncomms10069.
345. Jin L, Weisman M, Zhang G, Ward M, Luo J, Bruckel J, et al. Lack of association of matrix metalloproteinase 3 (MMP3) genotypes with ankylosing spondylitis susceptibility and severity. *Rheumatology (Oxford).* 2005;44(1):55-60.
346. Kato M, Sekine A, Ohnishi Y, Johnson TA, Tanaka T, Nakamura Y, et al. Linkage disequilibrium of evolutionarily conserved regions in the human genome. *BMC Genomics.* 2006;7:326. doi: 10.1186/1471-2164-7-326.
347. Siri A, Knäuper V, Veirana N, Caocci F, Murphy G, Zardi L. Different susceptibility of small and large human tenascin-C isoforms to degradation by matrix metalloproteinases. *J Biol Chem.* 1995;270(15):8650-4.
348. Riley GP, Curry V, DeGroot J, van El B, Verzijl N, Hazleman BL, et al. Matrix metalloproteinase activities and their relationship with collagen remodelling in tendon pathology. *Matrix Biol.* 2002;21(2):185–95.
349. Bini A, Itoh Y, Kudryk BJ, Nagase H. Degradation of cross-linked fibrin by matrix metalloproteinase 3 (stromelysin 1): Hydrolysis of the γ Gly 404-Ala 405 peptide bond. *Biochemistry.* 1996;35(40):13056–63.
350. Vogel KG, Koob TJ. Structural specialization in tendons under compression. In Bourne GH, Jeon KW, Friedlander M, editors. *International review of cytology* (Vol 115). Elsevier; 1989. p.267–93.
351. Myklebust G, Maehlum S, Holm I, Bahr R. A prospective cohort study of anterior cruciate ligament injuries in elite Norwegian team handball. *Scand J Med Sci Sports.* 1998;8(3):149–53.
352. Ireland ML. The female ACL: why is it more prone to injury? *Orthop Clin North Am.* 2002;33(4):637–51.

353. Brown CN, Yu B, Kirkendall DT, Garrett WE. Effects of increased body mass index on lower extremity motion patterns in a stop-jump task. National Athletic Trainers Association annual meeting. Indianapolis, IN. *J Athl Train.* 2005;40(Suppl 2):S32.
354. Knapik JJ, Sharp MA, Canham-Chervak M, Hauret K, Patton JF, Jones BH. Risk factors for training-related injuries among men and women in basic combat training. *Med Sci Sports Exerc.* 2001;33(6):946–54.
355. Parkkari J, Pasanen K, Manila VM, Kannus P, Rimpelä A. The risk for a cruciate ligament injury of the knee in adolescents and young adults: a population-based cohort study of 46 500 people with a 9 year follow-up. *Br J Sports Med.* 2008;42(6):422–6.
356. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ.* 2007;335(7624):806-8.
357. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med.* 2002;4(2):45–61.
358. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc.* 2013;8(11):2281–308.
359. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science.* 2013;339(6121):819–23.
360. Midwood KS, Orend G. The role of tenascin-C in tissue injury and tumorigenesis. *J Cell Commun Signal.* 2009;3(3–4):287–310.
361. Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations. *Genome Res.* 2016;26(6):719-31.
362. Miele A, Dekker J. Long-range chromosomal interactions and gene regulation. *Mol Biosyst.* 2008;4(11):1046-57.
363. Collins M, September AV, Posthumus M. Biological variation in musculoskeletal injuries : current knowledge, future research and practical implications. *Br J Sports Med.* 2015;49(23):1497-503.
364. Ribbans WJ, Collins M. Pathology of the tendo Achillis: do our genes contribute? *Bone Joint J.* 2013;95-B(3):305–13.
365. Wang G, Tanaka M, Eynon N, North KN, Williams AG, Collins M, et al. The future of genomic research in athletic performance and adaptation to training. In Posthumus M, Collins M, editors. *Genetics and Sports.* 2nd ed. Basel: Karger;2016. p.55–67.

366. Pitsiladis YP, Tanaka M, Eynon N, Bouchard C, North KN, Williams AG, et al. Athlome Project Consortium: a concerted effort to discover genomic and other “omic” markers of athletic performance. *Physiol Genomics*. 2016;48(3):183–90.
367. Zhang J, Jin X, Fang S, Li Y, Wang R, Guo W, et al. The functional SNP in the matrix metalloproteinase-3 promoter modifies susceptibility and lymphatic metastasis in esophageal squamous cell carcinoma but not in gastric cardiac adenocarcinoma. *Carcinogenesis*. 2004;25(12):2519–24.