Genetic Association Studies for Complex Traits: Relevance for the Sports Medicine Practitioner

William T. Gibson, MD, PhD, FRCPC, FCCMG

Assistant Professor, Department of Medical Genetics, University of British Columbia
Child and Family Research Institute
950 West 28th Avenue, Vancouver, BC, Canada V5Z 4H4
Telephone: 001-604-875-2000 ext. 5523
Fax: 001-604-875-2373
E-mail: wtgibson@cmmt.ubc.ca

Key words: Tendinopathy, Association Study, Single-Nucleotide Polymorphisms, Population Stratification, Hardy-Weinburg Equilibrium

In this issue of the Journal, September *et al.* [1] have studied DNA variants within the *COL5A1* gene among patients affected by Achilles tendinopathy (the cases) and among age- and country-of-origin-matched controls who have no tendinopathy (controls). Their findings suggest that common variation at the DNA level in the *COL5A1* gene it may be a risk factor for Achilles tendinopathy. Replication of their results among larger cohorts will be necessary to validate this finding.

It can be a challenge for the busy sports medicine practitioner to distill the clinical relevance of association studies such as this one. Though genetic testing for Mendelian (single-gene) disorders is widely available in many countries, genetic testing for single nucleotide polymorphisms (SNPs) is generally unavailable outside the research context. With certain notable exceptions (e.g. the link between apoE variants, cardiovascular and neurological disease),[2 3 4] statistical associations between common SNPs and complex diseases have not been borne out by further study. Numerous pitfalls occur in both the design and interpretation of these studies, which has resulted in a poor track record with respect to independent replication. Thus, a brief summary of these pitfalls may be useful to the Journal's readership.

By way of background, there are just over 14 million SNPs known to exist in the human genome.[5] Most of these exist in two possible forms, reflecting variation in the sequence that arose as a new mutation many generations ago. At some loci any of three or even all four DNA bases (adenine, guanine, cytosine and thiamine or A,G, C and T) may occur at measurable frequency in a population. Once a variant's frequency reaches 1% of all alleles in the population, such a variant is no longer considered a "mutation," but rather a "polymorphism." Recent advances in the rapidity and cost effectiveness of DNA sequencing technologies have enabled the assessment of such variants as risk factors for a broad range of medical conditions. Association studies have been used for many years to search for genes that predispose to aetiologically complex diseases such as obesity, type 2 diabetes, heart disease and adult-onset dementia. Case-control studies that examine SNPs at "candidate" genes are often used, wherein a gene is selected as a candidate based on the plausibility of its involvement in a molecular pathway relevant to the disease under study.

One often-overlooked pitfall of such an approach is that of the *a priori* equivalence of the two SNP variants. For most SNPs, no obvious functional effect will be discernible from visual inspection of the DNA sequence. If the SNP occurs in a protein-coding sequence and terminates the protein prematurely, one can have reasonable confidence that it has a true functional effect. Unfortunately, our ability to use theoretical algorithms to predict a SNP's biological effect is relatively poor, apart from this rare situation. Most SNPs occur outside of coding areas, and the majority of SNPs (even coding SNPs) will not change the amount or activity of any gene product. From a prior hypothesis standpoint, then, either SNP allele is equally likely to confer disease risk. But on a biological basis only *one* variant allele can confer increased risk, and if the risk factor were any other type of risk factor (smoking, viral exposure, etc.), this variant would have to be defined as "exposed" before the study were done. Consider the example of a C or T polymorphism at a specific locus. A human subject can then have SNP genotypes CC, CT or TT. With no obvious reason to prefer the C or T allele as the "at-risk" allele, a statistical association with the C

or with the T allele remains equally plausible. This is somewhat akin to a situation in which "exposure to virus" were equally believable as a risk factor for heart disease as "lack of exposure to virus." Epidemiological studies concluding that "lack of exposure to virus" conferred risk for heart disease (equivalent to saying that exposure to virus *protected* against heart disease) would be rejected as flawed, unless exemplary in methodology and highly statistically significant. Without a way to assess which SNP allele is likely to interfere with gene function until after the study is completed, roughly twice as many studies will achieve statistical significance (at whatever P-value is agreed upon) as "should" achieve it. This phenomenon of *a priori* equivalence acts to increase the false discovery rate if epidemiological data alone are considered. In this scenario, supportive experiments using *in vitro* systems such as transfected cells become very useful as an additional, independent line of evidence. In this article September *et al.* have identified two putative micro-RNA recognition sequences in the 3' untranslated region of the *COL5A1* gene. They suggest that suboptimal microRNA binding at these sites may promote increased translation of *COL5A1* mRNA into protein, thereby interfering with healing. Laboratory studies that examined the binding of microRNAs to transcripts bearing these recognition sequences have yet to be done, and would be useful to support the assertion that the identified SNPs truly affect *COL5A1* biology in a meaningful way (see figure 1).

Another pitfall is the fact that there are numerous potential sources of bias that may affect the SNP allele frequencies observed in the study population. Perhaps the most insidious of these is population substructure (also known as population stratification). Most human populations are genetically more heterogeneous than they appear at first glance. Thus, matching of cases and controls based on ethnic origin is a robust technique only in the context of populations that have minimal admixture (interbreeding) with other populations. Briefly, if two founding populations interbreed to create a third population, the genetic architecture of the new population will reflect the relative contributions of the different SNP allele frequencies that were present in each of the founding populations (see figure 2). If the two source populations differ in the prevalence of the medical condition under study (such as Achilles tendinopathy), a statistical association may arise between the condition and one of the SNP alleles. This association would arise from the allele frequencies in the source populations multiplied by the proportion of each that contributed to the admixed population. Given that the genome has many possible variants, the resulting statistical association is unlikely to reflect biological causality, even if the SNP-bearing candidate gene has a high biological plausibility. Among admixed populations such as South African Caucasians (known from historical records to include genetic contributions from Dutch Afrikaner, British and other European populations) and Australian Caucasians (known from historical records to include significant genetic contributions from the British Isles and Europe) spurious allelic associations can easily arise. This does erode the power to detect a significant effect of a SNP on a disease without bringing in additional evidence such as functional data. Debate exists regarding the actual magnitude of the bias introduced by population stratification, and whether this source of bias is significant when care is taken to match cases with controls from the same ancestral population (as has been done by September *et al.*). The topic of population stratification is usefully reviewed in an article by Cardon and Palmer,

which also discusses genetic methods of detecting and quantifying stratification.[6]

When considering the problem of population stratification, it is worthwhile noting the significance of Hardy-Weinberg equilibrium (HWE) and of departures from it. The presence of Hardy-Weinberg equilibrium is traditionally taken as evidence of random mating within a population over a period of time sufficient for the allele frequencies to equilibrate. If HWE is present, the proportion of homozygotes and heterozygotes for the alleles in question closely matches the proportion that would be predicted solely on the basis of the individual allele frequencies (i.e. there is neither a relative excess or deficiency of a particular genotype). If HWE is not present, there may be natural selection favoring a particular genotype (such as has been demonstrated for heterozygous carriers of sickle cell anaemia, thalassaemia and other haemoglobinopathies), but the most likely explanation for a lack of HWE is population substructure. Other possibilities include assortative mating (the tendency of individuals of like genotype to mate with each other) and errors in classification.[7] In case-control studies such as this one, artificial selection - the very process of ascertaining cases and controls from a sports medicine clinic - may result in some departure from Hardy-Weinberg equilibrium. The presence of HWE among the controls and its absence among the cases is reassuring, but is not in itself definitive proof that the statistical association of disease with SNP truly reflects a causative biological association. Such proof of causation must await replication studies among much larger populations. Consensus guidelines for such replication studies have been published.[8] Examples of high-quality studies include those recently published associating SNPs near the *FTO* and *MC4R* genes with obesity phenotypes.[9 10] Even then, the highest degree of confidence will be achieved only when long-term prospective follow-up studies are completed alongside other *in vitro* work that elucidates the biological mechanism whereby the SNP (or a closely linked DNA variant) influences cellular processes.

Because large studies require significant scientific and financial resources, it is worth considering whether other lines of genetic evidence might strengthen confidence in results obtained from case-control studies. Data from rare patients who manifest “extreme phenotypes” (including rare monogenic diseases) are useful in this regard.
Precisely what is considered to be an "extreme phenotype" will vary from one disorder to the next. Broadly speaking, features such as earlier onset and the presence of multiple affected family members often indicate a higher “genetic load.” In the case of overuse tendinopathy, it might be possible to determine the prevalence of an at-risk allele among patients with early-onset, bilateral, severe and/or recurrent tendinopathy, as well as among those with multiple other tendinous sites affected. If there exists a gender difference in the prevalence of the disease under study, one might also expect that the at-risk allele would be overrepresented among patients of the gender affected less frequently. Essentially, if the disease is found more rarely among women than among men, women who do manifest it are likely to harbour additional risk factors which offset the protection ordinarily offered by female gender. Cases with a family history of tendinopathy could be particularly valuable in this regard, because when additional family members (both affected and unaffected) can be collected, it is possible to apply tests such as the transmission disequilibrium test (TDT).[11] The TDT tests the hypothesis that the at-risk allele will be transmitted more often to offspring who manifest the disease

than to those do not.

Knowledge of rare Mendelian disorders can also be beneficial. If a gene's activity is critical for normal functioning, a complete loss-of-function mutation will almost always have a much more noticeable effect than a common variant such as a SNP. Were this not the case, the SNP itself would have been identified years ago and classified as a relatively common genetic disorder in its own right (as is arguably the case for ApoE). In the situation of *COL5A1*, loss-of function mutations cause Ehlers-Danlos syndrome, which includes recurrent tendinopathy as a feature. Because Ehlers-Danlos syndrome is rare, full loss-of function mutations in *COL5A1* are unlikely to account for a significant fraction of Achilles tendinopathy in the general population.[12] When reading SNP association studies, consideration should be given to whether mutations in the candidate gene cause any known Mendelian disorders, and whether the manifestations of these disorders include the common disease under study. If patients with a well-recognised Mendelian disorder do not show an increased prevalence of a common disease, it is unlikely that SNPs in or around that gene will confer a high population-attributable risk. Association studies that claim otherwise should be carefully scrutinised for potential sources of bias.

Using a similar rationale, data from transgenic mouse models can be incorporated into the selection of candidate genes for SNP association studies, and in the evaluation of their results. For example, mice deficient in the myostatin gene appear to have increased tendon fragility.[13] This suggests that the study of SNPs in or near the *MSTN* gene for association with Achilles tendinopathy may yield interesting results. The same is true for GDF-5 deficiency.[14 15] Conversely, if transgenic mice missing a candidate gene do not suffer from a particular disease, the likelihood that SNPs in or near that gene contribute to the disease in human populations is significantly reduced.

The hypothesis that genetic risk factors contribute to susceptibility to overuse injuries is highly plausible. Additional work needs to be done in order to identify specifically which genes and which SNPs (or other variants such as insertion/deletion polymorphisms) confer the greatest proportion of the genetic risk for these disorders. Ultimately, the clinical use of DNA-based testing to refine outcome predictions and/or modify rehabilitation regimens will have to wait until larger case-control studies and long-term follow-up studies have been done.

Competing Interests Statement:
The Corresponding Author declares no competing interests.

Figure Legends

Figure 1

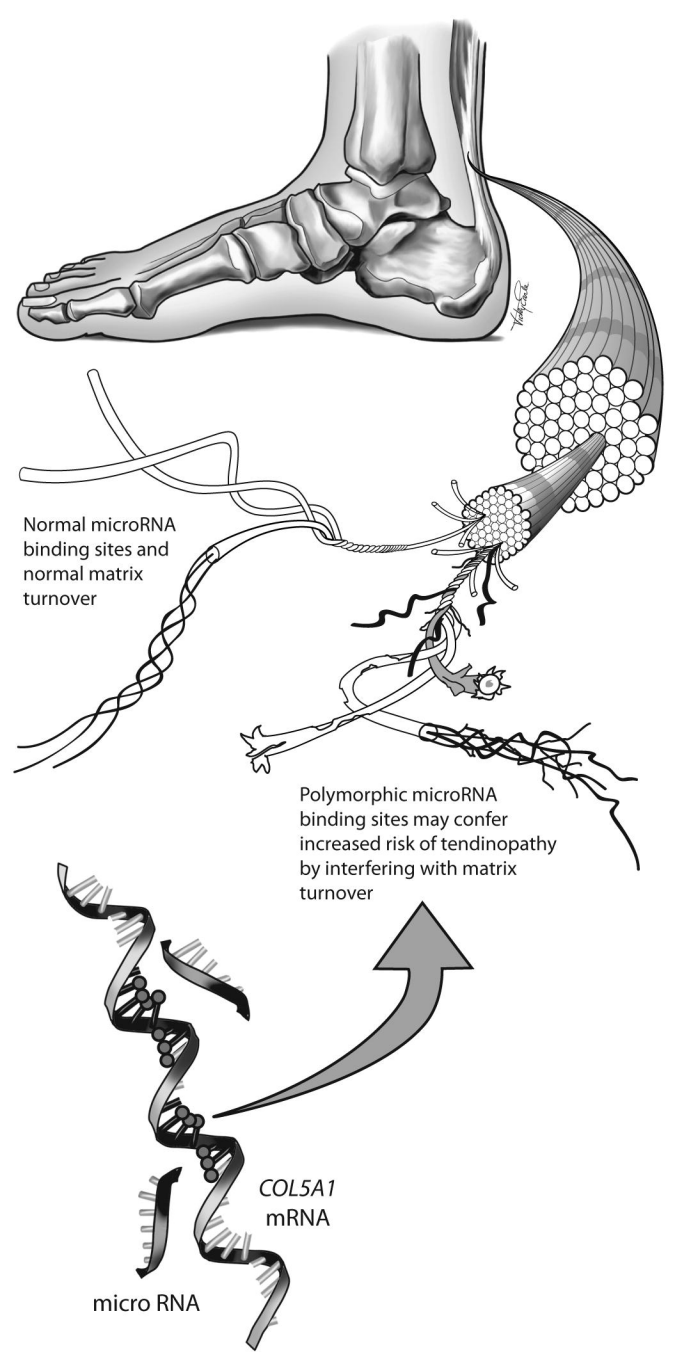
Postulated effect of *COL5A1* SNP on susceptibility to Achilles tendinopathy, according to September *et al.* (2008).

Figure 2

Diagram of the results of admixture between the two theoretical populations ("A" and "B") to create a third population "C." Population "A" has different allele frequencies than Population "B" at the locus of interest. Population "A" contributes 70% of the alleles to the new admixed population, and Population "B" contributes 30% of the alleles. The allele frequencies of both Population "A" and Population "B" are in Hardy-Weinberg equilibrium, but those of the new admixed population are not in HWE. In this situation, it is easy for a statistical association to arise between one of the alleles and a disease of interest, purely due to the relative allelic contributions from the founder populations to the population under study.

REFERENCES

1 **September M**, *et al.* Variants within the *COL5A1* gene are associated with Achilles tendinopathy in two populations. *Br J Sports Med* 2008; Apr 28.
2 **Saunders AM**, Strittmatter WJ, Schmechel D, *et al.* Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 1993;**43**:1467-72.
3 **Corder EH**, Saunders AM, Strittmatter WJ, *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 1993;**261**:921-23.
4 **de Knijff P,** van den Maagdenberg AM, Frants RR, *et al.* Genetic heterogeneity of apolipoprotein E and its influence on plasma lipid and lipoprotein levels. *Hum Mutat* 1994;**4**:178-94.
5 National Center for Biotechnology Information EntrezSNP database http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp&TabCmd=Limits Accessed July 19, 2008.
6 **Cardon LR**, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003;**361**:598-604.
7 **Cavalli-Sforza LL**, Bodmer WF. The Genetics of Human Populations. London, Constable and Company, 1999.
8 **Chanock SJ**, Manolio T, Boehnke M, *et al.* Replicating genotype-phenotype associations. *Nature* 2007;**447**:655-60.
9 **Frayling TM**, Timpson NJ, Weedon MN, *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;**316**:889-94.
10 **Loos RJ**, Lindgren CM, Li S, *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics* 2008;**14**:768-75.
11 **Ewens WJ**, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995;**57**:455-64.
12 **Järvinen M**, Józsa L, Kannus P, *et al.,* Histopathological findings in chronic tendon disorders. *Scand J Med Sci Sports* 1997;**7**:86-95.
13 **Mendias CL**, Bakhurin KI, Faulkner JA. Tendons of myostatin-deficient mice are small, brittle, and hypocellular. *Proc Natl Acad Sci USA* 2008;**105**:388-93.
14 **Mikic B**, Schalet BJ, Clark RT, *et al.* GDF-5 deficiency in mice alters the ultrastructure, mechanical properties and composition of the Achilles tendon. *J Orthop Res*. 2001;**19**:365-71.
15 **Warden SJ.** Animal models for the study of tendinopathy. *Br J Sports Med* 2007;**41**:232-40.


Normal microRNA binding sites and normal matrix turnover
Polymorphic microRNA binding sites may confer increased risk of tendinopathy by interfering with matrix turnover
COL5A1 mRNA
micro RNA

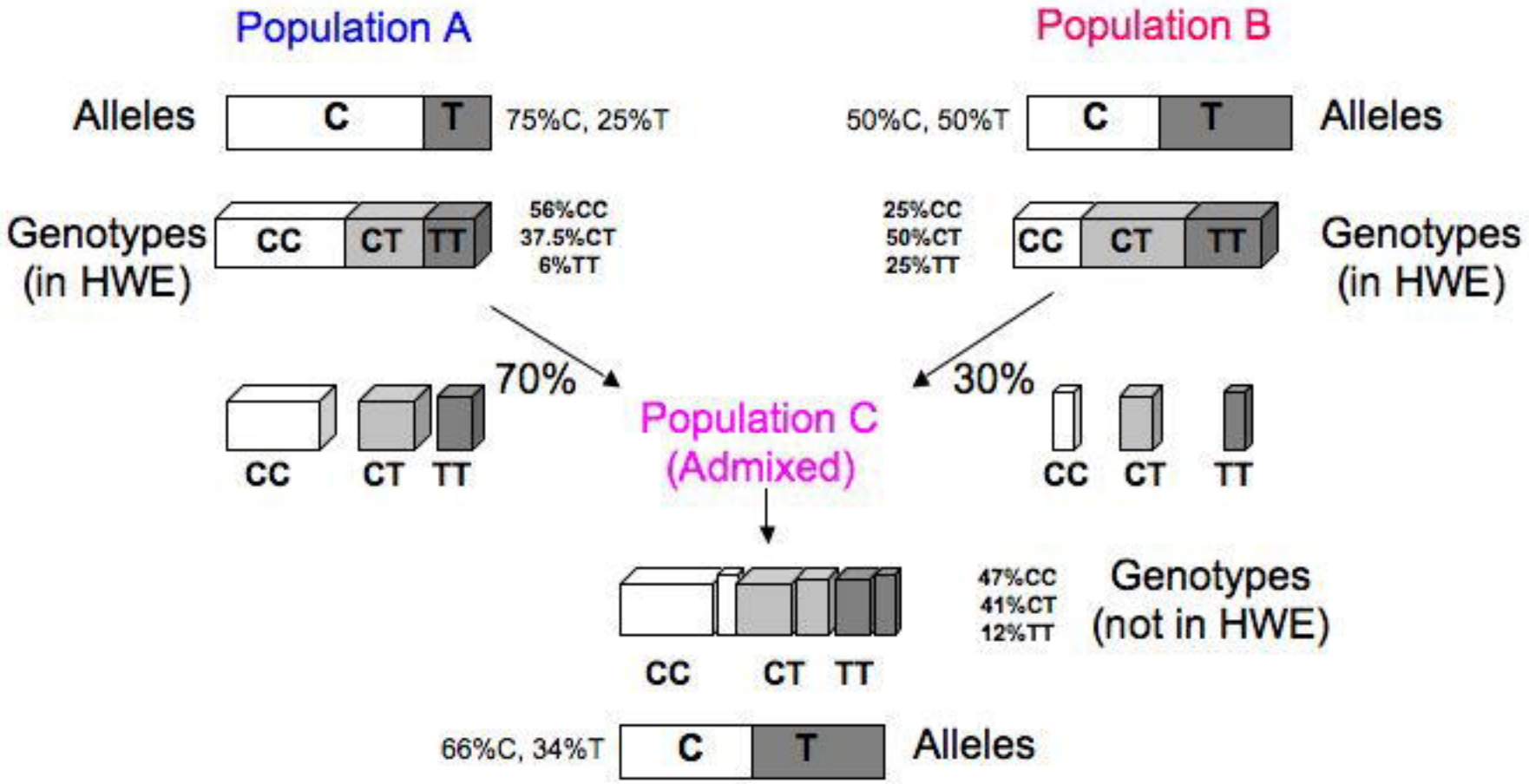

Population A
Population B
Alleles
C
T
75%C, 25%T
50%C, 50%T
Alleles
Genotypes (in HWE)
CC
CT
TT
56%CC
37.5%CT
6%TT
25%CC
50%CT
25%TT
Genotypes (in HWE)
70%
30%
Population C (Admixed)
47%CC
41%CT
12%TT
Genotypes (not in HWE)
66%C, 34%T
Alleles

Figure 2

Genetic Association Studies for Complex Traits: Relevance for the Sports Medicine Practitioner

William T Gibson

Br J Sports Med published online December 9, 2008