The Genetic and Physiological Correlates of Human Performance

by

Jason Withford-Cave

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in the Central Clinical School (Medicine) at

The University of Sydney

December 2006

Statement

The work undertaken for this thesis was performed in the Dept of Molecular and Clinical Genetics at the Royal Prince Alfred Hospital, Sydney and the Central Clinical School (Medicine) at The University of Sydney, between June 2002 and June 2006.

The content of this thesis is my own original work unless stated otherwise. It has not previously been presented for the purpose of obtaining any degree.

Jason Withford-Cave

BSpSc(HonI)

Acknowledgements

- Professor Ronald J Trent (Supervisor), Dr Bing Yu (Associate Supervisor) of the Faculty of Medicine, University of Sydney and Dept of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Sydney.
- The Rebecca L. Cooper Medical Research Foundation for my three year PhD research scholarship.
- Stuart Cole (Masters student) and Jennifer Henderson (PhD student) of the "Elite Athlete Project" of the Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Sydney and Faculty of Medicine, University of Sydney.
- Dr Nicole Sawyer (Research and Development Director) and other staff of the Sydney University Prince Alfred Macromolecular Analysis Centre (SUPAMAC).
- Edna Soriano, other staff and patients of the Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Sydney.
- Kevin McGeechan (Associate Lecturer, Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, University of Sydney) advised on multiple regression analysis using the SPSS™ program.
- Professor Colin Sullivan (Postgraduate Coordinator) and Professor Brian Morris, Faculty of Medicine, University of Sydney
- Professor Allan Hahn (Head of Department of Physiology) and Dr Jason Gulbin (National Talent Search Director), Donna Martin, other staff and athletes of the Australian Institute of Sport (AIS), Canberra.
- The staff and donors of the Australian Red Cross Blood Bank.
- The players and officials of the NSW Rugby Union and ACT Rugby Union.
- My wife, Chantelle Withford-Cave, and my father-in-law, Paul Cave, for their advice and support.

Abstract

The aims of the present study were to identify genes related to human performance and to assess if these genes affect standard human athletic performance variables. These genes were identified using a combination of *in silico* search techniques and association studies. The selected genes were screened for known polymorphisms and for new polymorphisms using a combination of PCR, RFLP, SNP and DHPLC technology. The variants found in athletes were compared to control groups for differences and compared within athlete groups with multiple regression analysis for correlation with phenotypic data.

The *ACE I/D* polymorphism was associated with groups of athletes within a variety of sports, including the novel sport of rugby. Multiple regression analysis consistently showed weight, but not *ACE I/D*, to be a significant predictor of performance. The *MYBPC3* e6 S236G variant was, for the first time, tested within a variety of sports and shown to be associated with rugby players and sprint runners. Multiple regression analysis consistently showed weight and elite, but not *MYBPC3* e6 S236G, to be significant predictors of performance. The *EPAS1* gene was identified as being associated with elite endurance athletes. A novel *EPAS1* i8/e9 C_7/C_5GC polymorphism was identified but did not appear to be associated with elite endurance athletes. Multiple regression analysis consistently showed weight, but not *EPAS1* i8/e9 C_7/C_5GC, to be a significant predictor of performance. Caution should be exercised in the assumption of exclusion of association for the non-significant results in the present study as most of the statistical tests were significantly under-powered.

The established technologies of PCR, RFLP and gel electrophoresis testing were used to test, a relatively large number of DNA samples for known polymorphisms (*ACE I/D* and *MYBPC3* e6 S236G). A novel gene (*EPAS1*) related to endurance athletes, was identified using a

combination of *in silico* search techniques, high-throughput SNP screening and association studies. The comparatively new technology of high-throughput SNP detection was used to test a large number of DNA samples for a SNP association study. The new technology of DHPLC DNA variant detection was used for screening a relatively large number of DNA samples for significant novel DNA variants. The identification of performance genes and understanding the function of these genes, will lead to exciting advances in sports science. The present study covers a wide range of popular international sports which makes it valuable to the field of sports genetics. This is the first time the genetic contribution to performance in rugby players has been examined.

Contents

Publications Arising From This Work

Henderson J, **Withford-Cave JM**, Duffy DL, Cole SJ, Sawyer NA, Gulbin JP, Hahn A, Trent RJ, Yu B (2005) The EPAS1 gene influences the aerobic-anaerobic contribution in elite endurance athletes. *Hum Genet* 118: 416-23

Abbreviations

A	adenine (nucleotide) or alanine (amino acid)
ABO	blood group system
ACE	angiotensin I converting enzyme
ACT	Australian Capital Territory
ACTN3	α-actinin 3
AMPD1	adenosine monophosphate deaminase 1
AIS	Australian Institute of Sport
ARNT	aryl hydrocarbon receptor nuclear translocator
ATP	adenosine triphosphate
C	cytosine (nucleotide) or cystine (amino acid)
cDNA	complementary DNA
cm	centimetre(s)
cM	centiMorgan
D	aspartic acid (amino acid)
D	deletion
DHPLC	denaturing high performance liquid chromotography
dH_2O	deionised water
DNA	deoxyribonucleic acid
E	glutamic acid (amino acid)
EDTA	ethylenediaminetetraacetic acid
EPAS1	endothelial PAS domain protein 1
EPO	erythropoietin
e4	exon 4
e5	exon 5
e6	exon 6
F	phenylalanine (amino acid)
FHC	familial hypertrophic cardiomyopathy
FnIII	fibronectin-like type III
g	gram(s)
G	guanine (nucleotide) or glycine (amino acid)
H	histidine (amino acid)
HIF	hypoxia inducible factor
HLH	helix-loop-helix
HRE	hypoxia responsive element
i	intron
I	isoleucine (amino acid)
I	insertion
IgI	immunoglobulin-type
Ironman	Ironman triathlon
K	lysine (amino acid)
kDa	kiloDalton
kg	kilogram
L	litre or leucine (amino acid)
lod	logarithm of odds
LV	left ventricular
M	methionine (amino acid)
min	minute(s)

mRNA	messenger RNA
MYBPC	myosin binding protein C
MYBPC3	cardiac myosin binding protein C
N	any nucleotide or asparagine (amino acid)
NSW	New South Wales (Australian state)
nt	nucleotide
O_2	oxygen molecule
OD	optical density
ODD	oxygen-dependent degradation domain
P	proline (amino acid)
PAS	per-arnt-sim
Pu	one of the two purine nucleotides (A or G)
Py	one of the two pyrimidine nucleotides (T/U or C)
PCR	polymerase chain reaction
PT-MAX	power-time-maximum
PT-SS	power-time-steady-state
Q	glutamine (amino acid)
QTL	quantitative trait loci
R	arginine (amino acid)
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rugby	Rugby Union
s	second(s)
S	serine (amino acid)
SNP	single nucleotide polymorphism
SUPAMAC	Sydney University Prince Alfred Macromolecular Analysis Centre
T	thymine (nucleotide) or threonine (amino acid)
T_m	melting temperature
U	uracil (nucleotide)
V	valine (amino acid)
VEGF	vascular endothelial growth factor
$\dot{V}O_2$ max	maximum oxygen uptake or consumption
wk	week(s)
W	tryptophan (amino acid)
X	stop codon
Y	tyrosine (amino acid)
yr	year(s)
↓	exon-intron boundary
→	changes to
Δ	delete
′	prime

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Overview

Understanding human variation is one reason why so much effort has gone into research in the area of genetics and to completing the Human Genome Project. Understanding the reasons behind human variation will provide avenues for progressing knowledge for many areas of human endeavour in the 21st century. The Human Genome Project was the first "big science" project in biology. Human molecular genetics not only forms the cutting edge of biomedical research, but it has immediate application to the diagnosis of disease, and has great potential for treating disease (Strachan and Read, 1999).

Sports Science will be an area where advances can be expected with the benefit of genetic information. Sports Science has had difficulty doing hypothesis-driven training-theory research. Sports Genetics is a new area of science which combines Sports Science and Genetics. Using the elite athlete model to understand human genetics could also lead to breakthroughs in the areas of genetics and medicine.

1.2 Sports Science

At least as far back as the ancient Olympic Games over 2,500 years ago, there has been an interest in Sports Science. The training of the ancient Olympic athletes was documented by the Greek and Roman writers: distance and sprint training in runners; ball exercises; and cross-training by running, weight lifting, and wrestling with animals. The ancient Greeks used the "tetrad" (four-day training cycle) with a specific type of training for each day (Grivetti and Applegate 1997). In the last century, exercise physiology has developed from physiology and medicine to further the understanding of human health. In the last few decades, an increasing interest and knowledge of genetics has begun to impact on exercise physiology in practical ways, e.g. it has now become possible to identify genes that are associated with human physical performance and understand the function of the gene variants.

1.2.1 Factors of Sports Performance

The Ss of sports performance include (Smith 2003):

1. stamina (endurance); 2. speed; 3. strength; 4. skill; 5. suppleness (flexibility); 6. (p)sychology; 7. stature (height, weight, somatotype and body composition); 8. sustenance (nutrition); 9. surroundings (acclimatisation); 10. socioeconomics; and 11. sex (gender).

The factors that will be the focus of this thesis are endurance, speed, stature (weight and height) and gender, which are the most commonly and easily measured athletic performance variables.

1.2.2 Exercise Physiology

Exercise physiology probably started with work of Hill and Lupton in the 1920s (Bassett and Howley 1997). They discovered and developed the concept of maximum oxygen uptake or consumption ($\dot{V}O_2$ max). They defined it as the maximum rate at which the body can use oxygen (O_2) during exhaustive exercise. $\dot{V}O_2$ max is considered the "gold standard" measurement of the maximal rate of function of the cardiovascular system.

Swedish researcher Per Olaf Astrand greatly expanded the knowledge of exercise physiology in the 1950s and 1960s (Astrand and Saltin 1961). Great work was done in the area of exercise training theory, quantification and modelling of performance by Banister (Banister et al. 1999; Banister et al. 1992; Morton et al. 1990) from the 1970s. The results of this research have yet to be widely implemented in the area of elite athlete training. A major reason for this may be that there is a chasm between university-based exercise physiology research, which generally uses case-control studies with student subjects and sports academy-based exercise physiology research, which revolves around the four year Olympic cycle and uses elite athletes. Untrained and/or non-elite student subjects will generally respond to most of the training that they are given. Many of the studies regarding optimal training may, therefore, have questionable relevance to elite athletes (Hopkins et al. 1999). The difficulty with elite athlete research is that much of the research is non-interventional because of the time restraints imposed by coaches whose careers rise or fall depending on results obtained in world and Olympic championships. Research on training theory on elite athletes, consequently, becomes extremely difficult to perform.

These difficulties gave rise to the burgeoning interest in using genetics to understand exercise physiology. It shows potential to find shortcuts to understanding the most important aspects

of the field, and to lead to improvements in the understanding of exercise training theory and the practical aspects of exercise training.

1.2.3 Energy Systems

Achieving survival goals required an efficient and powerful energy system able to produce energy on the continuum from high intensity to prolonged physical activity. Energy for skeletal muscle contraction is supplied by anaerobic and aerobic metabolic pathways. There are three distinct yet closely integrated processes that operate together to satisfy the energy requirements of muscle: aerobic, anaerobic lactate and anaerobic phosphate.

The aerobic energy is the predominant system for endurance athletes. The aerobic system is the most efficient adenosine triphosphate source for skeletal muscle. The aerobic energy system refers to the combustion of carbohydrates and fats in the presence of oxygen (De Feo et al. 2003).

Anaerobic energy is the predominant system for power athletes and is divided into phosphate (or alactic) and lactic components. It can allow short bursts of intense physical activity (60–90 s) and utilises as a source of energy the phosphocreatine shuttle and anaerobic glycolysis. The anaerobic pathways can regenerate adenosine triphosphate quickly but only for a limited duration. In contrast, the aerobic system cannot supply energy at nearly as high a rate but has an almost unlimited capacity, depending on concurrent nutrition. Energy is supplied in significant amounts from each of the energy systems during most exercise activities. The duration of maximal exercise at which equal contributions are derived from the anaerobic and aerobic energy systems appears to occur between 1–2 min and most probably around 75 s, a time that is considerably earlier than has traditionally been suggested (Gastin 2001).

The phosphate energy system entails splitting the stored phosphagens: adenosine triphosphate and phosphocreatine. The lactate energy system uses the non-aerobic breakdown of carbohydrate to lactic acid through glycolysis.

Another approach, which has not yet gained acceptance, is using power scaling laws for energy systems. Using these, there is a breakpoint at approximately 1000 m in athletic running events, at around 150–170 s duration. This may represent a transition from using all three energy systems to largely using the aerobic energy system with relatively little contribution from the anaerobic energy system. This was independent of gender and sport (Carbone and Savaglio 2001).

1.2.4 Limiting Factors of Endurance Performance

There has been much debate over what are the limiting factors of endurance performance. It is a key issue in sports genetics because it provides a starting point for looking for important genes related to the extreme phenotype needed for a strong genetic effect. The limiting factors of endurance performance are (Bassett and Howley 2000):

1. $\dot{V}O_2$ max;
2. Anaerobic Threshold;
3. Efficiency.

$\dot{V}O_2$ max is limited in humans by O_2 delivery to the exercising muscles because: 1) $\dot{V}O_2$ max changes when O_2 delivery is altered (by blood doping or hypoxia); 2) the increase in $\dot{V}O_2$ max with training mainly comes from an increase in maximal cardiac output (not arterio-venous O_2 difference); and 3) when only a small muscle mass is overperfused, O_2 extraction is not limiting. O_2 delivery, therefore, not skeletal muscle O_2 extraction, is the limiting factor for

$\dot{V}O_2$ max. Improvements in performance in experienced athletes are, however, thought to be due to increases in lactate threshold $\dot{V}O_2$, not $\dot{V}O_2$ max. The best predictor of distance running performance is considered to be the speed at lactate threshold (Bassett and Howley 2000).

1.2.4.1 Maximal Oxygen Uptake ($\dot{V}O_2$ max)

The fact that the $\dot{V}O_2$ max curve to the asymptote occurs so abruptly ("flat-lines") tends to indicate that there is at least one major limiting factor to $\dot{V}O_2$ max. Otherwise two breakpoints would be expected.

1.2.4.2 Anaerobic Threshold

"Anaerobic threshold is defined as the highest sustained intensity of exercise for which measurement of oxygen uptake can account for the entire energy requirement" (Svedahl and MacIntosh 2003). The anaerobic threshold (and various other lactate or ventilatory thresholds) is a highly controversial subject and will largely be avoided in this thesis.

1.2.4.3 Efficiency or Energy Cost of Exercise

The efficiency or energy cost of exercise is an important variable for endurance performance. The ratio of $\dot{V}O_2$ max / efficiency is thought to be a good indicator of performance, with r^2 correlation = 0.87 for 10 km run time (Telford 2003).

Some researchers have reported that in runners, better efficiency was associated with lower $\dot{V}O_2$ max and greater bodyweight (Pate et al. 1992). Others suggested that a high efficiency seems to compensate for a relatively low $\dot{V}O_2$ max in professional cyclists (Lucia et al. 2002). A critique noted, though, that some researchers had expressed $\dot{V}O_2$ max and sub-maximal $\dot{V}O_2$ (a common measure of efficiency) relative to body mass. Body mass was a divisor for ratios for both variables in the correlation analysis. Positive correlations would have been expected even for randomly selected values of $\dot{V}O_2$ max and submaximal $\dot{V}O_2$ (Atkinson et al. 2003). Any correlations between performance variables and efficiency have, therefore, to be interpreted cautiously. Other researchers disagreed with this critique by citing their own work which also used $\dot{V}O_2$ max values relative to body mass (Noakes and Tucker 2004). They may have, as a result, also suffered from spurious correlations. Some researchers cited in the original critique defended their position of scaling their data for runners relative to body mass but were not supportive of doing the same for cycling and other non-weight-bearing sports (Morgan and Pate 2004). Other researchers agreed with scaling efficiency relative to a power of body mass for runners (Saunders et al. 2004).

1.2.5 Limiting Factors of Maximum Oxygen Uptake

The $\dot{V}O_2$ max baseline value varies two-fold (Bouchard et al. 1998). The $\dot{V}O_2$ max response to 20 weeks of standardised training also varies two-fold (Kohrt et al. 1991; Lortie et al. 1984; Skinner et al. 2000). The highly trained $\dot{V}O_2$ max value probably varies, consequently, three- to four-fold. Consider the range of reported typical elite athlete values: $\dot{V}O_2$ max (male): 63–95 $mL.kg^{-1}.min^{-1}$ (1.50-fold range); anaerobic threshold: 80–90% of $\dot{V}O_2$ max (1.12-fold range); and efficiency: 22–27% (1.23-fold range). $\dot{V}O_2$ max has the largest range of typical values so it is the logical phenotypic trait to investigate. $\dot{V}O_2$ max comprises (Bassett and

Howley 2000): (a) pulmonary diffusion capacity; (b) cardiac output; and (c) oxygen carrying capacity of blood.

1.2.5.1 Lung Diffusion Capacity

All thoroughbred racehorses (and some humans) bleed into their lungs during racing indicating lung insufficiency (West and Mathieu-Costello 1995). Thoroughbred racehorses have registered $\dot{V}O_2$ max scores of approximately 180 $mL.kg^{-1}.min^{-1}$. In highly aerobic species, and possibly elite human athletes, the O_2 transport system may reach its capacity due to the lung which is the least malleable part. $\dot{V}O_2$ max seems to be limited by all parts of the O_2 transport system simultaneously (Jones and Lindstedt 1993).

1.2.5.2 O_2 Carrying Capacity of Blood

The O_2 carrying capacity of blood is determined by blood volume and %haematocrit. A high $\dot{V}O_2$ max with no history of training is mainly due to high (and possibly more haemodynamically active) blood volume that brings about a high maximum stroke volume and maximal cardiac output (Martino et al. 2002). A 50% haematocrit level is used as the medical disqualification-point in cycling. A higher level than this puts the cyclist at risk of suffering cardiovascular complications, including sudden death, during sleep after the race due to high blood viscosity.

1.2.6 Limiting Physical Factors of Performance

Much work has been done in the field of anthropometry, which is the science of body measurement. It has been used extensively in ergonometrics research, the military and sports. Morphological optimisation has been used in elite sport and is considered a limiting factor in

many sports. There is open upper-end optimisation for height and mass in sports such as Rugby Union (rugby) (Olds 2001), absolute optimisation for height in marathon (Norton and Olds 2001) and open lower-end optimisation for height and mass in women's gymnastics (Claessens et al. 1991). There are other sports where limb lengths are important such as boxing and weightlifting. For many professional and Olympic sports, the effect of genes related to height and mass are, as a result, very important and may confuse the signals coming from genes that are related to physiological performance abilities such as cardiovascular fitness.

1.2.6.1 Scaling Mass

Scaling mass for physiological variables has become a popular topic lately in exercise physiology. There are many conflicting studies. Most literature on exercise physiology reports $\dot{V}O_2$ max values using the absolute ($L.min^{-1}$) and/or the simple ratio ($mL.kg^{-1}.min^{-1}$) method. Evidence is accumulating that scaling methods such as the power function ratio method are more valid in many circumstances. Kleiber's law is the observation that, for the vast majority of animals, metabolic rate scales to the ¾ power of mass (Batterham et al. 1997). A cat which weighs 100 times more than a mouse, consequently, will have a metabolic rate only 30 times greater (Figure 1-1). Kleiber's law is a result of the physics and geometry of animal circulatory systems. Allometric law (or power-law) is relationships between living organism's body parts or process, usually expressed in power-law form: $y \sim x^a$.

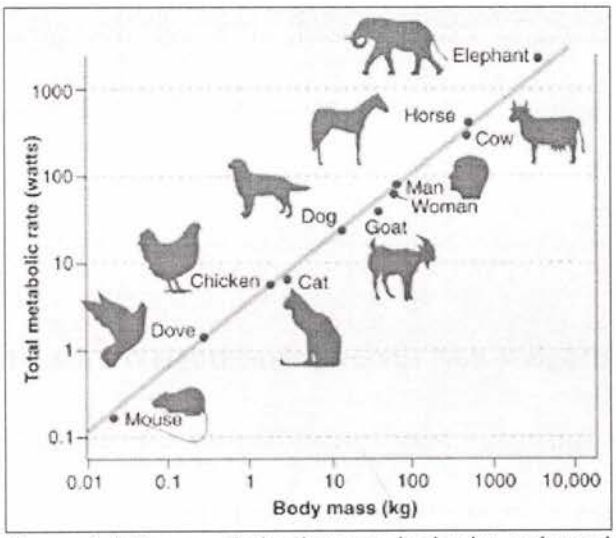


Figure 1-1 The correlation between body size and metabolic rate (http://biology.unm.edu/jhbrown/Research/Scaling/Scaling.htm).

Allometric scaling (geometric similarity) models overcome the heteroscedasticity and skewness observed with per ratio variables. If per ratio standards are to be incorporated in regression models to predict other dependent variables, the allometric or log-linear model form is often better than linear models (Nevill and Holder 1995). Allometric scaling theory shows that $\dot{V}O_2$ max should scale to a mass exponent of ⅔ instead of 1. In both athletic subjects and controls, body circumferences change more than by geometric similarity ($mass^{⅓}$) in fleshy sites and less in bony sites (Nevill et al. 2004b).

The majority of the research suggests that $\dot{V}O_2$ max increases in proportion to the $mass^{0.6–0.872}$ range, rather than in proportion to $mass^{1.0}$ (Batterham et al. 1997; Buresh and Berg 2002; Chamari et al. 2005; Eisenmann et al. 2001; Heil 1997; Jensen et al. 2001; Rogers et al. 1995; Weibel et al. 2004; Welsman et al. 1996). Mass exponents of 0.32 and 1 were, nonetheless, were found to be the best by some researchers for level and uphill cycling ability, respectively (Nevill et al. 2005; Padilla et al. 1999). Fitness tests that determine aerobic power in units relative to body mass, such as running, incur a bias against heavier subjects (Bilzon et al. 2001). Although the ratio standard $\dot{V}O_2$ max ($mL.kg^{-1}.min^{-1}$) is dependent on body mass,

some consider it the best predictor of running performance (Nevill et al. 2004a). Some studies indicate, however, that some scaling to a power less than 1 (0.71) is required for running (Bergh et al. 1991).

1.2.6.2 Weightbearing versus Non-weightbearing

Running is a weightbearing sport. Rowing, cycling and swimming are non-weightbearing sports. Rowing and cycling both involve external resistance. Ironman triathlon (Ironman) and rugby are sports that are both weight-bearing and involve external resistance. The external resistance can be air and water resistance (drag), mass of equipment and resistance to motion by opposition players. The complexity of resistive forces makes it difficult to categorise sports for whether body mass is a positive, negative or neutral factor for performance.

The question arises as to which group of athletes would most likely have a polymorphism of a gene associated with $\dot{V}O_2$ max. There are various arguments for each sport being more or less dependent on $\dot{V}O_2$ max: specialists versus generalists; weight-bearing versus non-weight-bearing, open versus weight-categories, etc. The genome-wide scan paper, which was the basis for the present study, used cycling for the training stimulus and testing $\dot{V}O_2$ max (Bouchard et al. 2000). Cycling is, accordingly, a sport that is appropriate for this investigation. Following on from this, sports such as rowing, Ironman and running are appropriate for the present study because they are whole-body endurance-oriented sports.

1.2.7 Anatomy and Physiology of Relevant Sports

1.2.7.1 Rowing

Typical $\dot{V}O_2$ max values for males are 6.1 $L.min^{-1}$ and females are 4.1 $L.min^{-1}$. The aerobic system provides 70–75% and the anaerobic system 25–30% of the energy for a 2000 m race. The muscle fibre distribution of male rowers is similar to distance runners while females tend to have slightly more fast-twitch fibres. The mechanical efficiency of rowers is approximately 20%. The pattern of race pacing for rowers begins with a vigorous anaerobic sprint start followed by an aerobic pace and another sprint at the finish (Hagerman 1984).

The Valsalva-like manoeuvre executed at the catch phase (oar enters water) of each stroke in rowing is the main cause of a transient increase in blood pressure. The associated blood pressure response could be the cause of cardiac hypertrophy in rowers (Clifford et al. 1994). A study of male club level rowers showed that $\dot{V}O_2$ max and lean body mass correlated best with the velocity for a 2000 m time-trial. Multiple regression analysis showed that $\dot{V}O_2$ max was the best predictor of the velocity for the 2000 m time-trial and accounted for 72% of the variability in performance (Cosgrove et al. 1999). A multiple regression study of competitive female rowers showed that peak power in a rowing Wingate test explained 76% of the variability in 2000 m indoor rowing performance time, while $\dot{V}O_2$ max explained 12% and fatigue explained 8% of the variability (Riechman et al. 2002). The heart rate response is attenuated in rowing compared to during running, even with a larger $\dot{V}O_2$ max in rowing. Central blood volume is lower in running than rowing due to posture (Secher 2003).

Ventilation probably does not limit $\dot{V}O_2$ max because there is greater ventilation during rowing than during running. The lungs may, however, fail to fully oxygenate arterial blood

(Yoshiga and Higuchi 2003). A reduction in the arterial O_2 tension is reported for some intensively trained runners and cyclists, but for all oarsmen during (ergometer) rowing (Hanel et al. 1994). During rowing, arterial O_2 saturation decreases to approximately 90%, similar to an altitude of over 3000 m.

1.2.7.2 Cycling

When elite-national rank cyclists were compared to state rank cyclists, $\dot{V}O_2$ max (~70 mL.kg^{-1}.min^{-1} and 5.01 L.min^{-1}) and lean body mass were not significantly different. The elite cyclists were, on the other hand, 10% faster, had a greater percentage of type I muscle fibres and a 23% greater muscle capillary density (Coyle et al. 1991). The energy cost for a cyclist is mainly related to two forces: air resistance (frontal drag) on flat terrain, and gravity (climbing). Other energy costs for a cyclist include rolling resistance and rotational kinetic energy. Some research found that the mass exponent for drag is ⅓, for $\dot{V}O_2$ max is ⅔ and for climbing is 0.79. Since the drag (⅓ exponent) varies little between small and large cyclists, whereas $\dot{V}O_2$ max (⅔ exponent) varies much, larger cyclists have an advantage in flat time trials. In climbing, the exponent of climbing is greater than that of $\dot{V}O_2$ max, therefore, the smaller cyclists have an advantage (Swain 1994). Computer modelling of actual and predicted road cycling times indicated that the main physiological factors contributing to road-cycling performance were $\dot{V}O_2$ max, fractional utilisation of $\dot{V}O_2$ max, mechanical efficiency, and projected frontal area (Olds et al. 1995). Some researchers concluded that there are no differences in efficiency between elite and recreational cyclists (Moseley et al. 2004). $\dot{V}O_2$ max is predictive of cycling performance when coupled with lactate, power, metabolic thresholds and efficiency measures (Faria et al. 2005).

1.2.7.3 Rugby

Rugby is a heavy body-contact (some say collision) amateur and professional ball sport, played with 15 players on each team who have well-defined positions and highly-specialised roles. The players have a wide variety of body types, heights and weights, compared to most sports, depending on the level or standard played and position played. The game is played for 80 min and is both highly aerobically and anaerobically demanding. The positions in rugby have traditionally been divided into two groups: forwards and backs. The forwards' main role is to win possession of the ball in scrums, lineouts, rucks and mauls. The backs' main role is to use the ball possession gained to score by means of fast running and passing movements forward and across the playing field. Rugby is predominantly played by males and all of the subjects in the present study were male.

A time-motion study of work-to-rest ratios in international rugby indicated that the game places greater demands on anaerobic glycolysis than previously reported (McLean 1992). Another match analysis has indicated that rugby is an intermittent sport requiring a large number of work periods of 5–15 s with recovery of less than 40 s. Rugby players have highly specialised anthropometric and physiological attributes for each position and playing standard (Nicholas 1997).

An analysis of elite Australian under-19 years rugby players showed that the outside backs covered a significantly greater total distance than forwards. Forwards expended more energy than backs, due to performing continuous static high-intensity activities (Deutsch et al. 1998). Rugby forwards are typically heavier, taller, fatter and have higher absolute aerobic and anaerobic power, and muscular strength. The intense efforts of rugby rely on anaerobic energy, while the aerobic system is used for recovery (Duthie et al. 2003). A study of elite international professional rugby players showed frequent short work periods (< 4 s) followed

by moderate rest periods (< 20 s) for forwards, and long rest periods (> 100 s) for backs (Duthie et al. 2005).

An amateur rugby club first grade team, immediately post-season had mean weight, 84 kg and mean $\dot{V}O_2$ max, 56.6 $mL.kg^{-1}.min^{-1}$ (Maud 1983). Twenty-nine South African club rugby players showed high absolute values for $\dot{V}O_2$ max in the forwards. Both backs and forwards had a higher than average percentage of fast-twitch muscle fibres (57% and 53% respectively compared to ~50% for normal population) (Jardine et al. 1988). New Zealand rugby players from school to senior level showed that forwards were taller and heavier allowing more momentum when sprinting, important in body contact; backs were aerobically fitter, faster, and more agile (Quarrie et al. 1995). Senior elite rugby players showed that mean $\dot{V}O_2$ max was higher in backs than in forwards (48.3 vs. 41.2 $mL.kg^{-1}.min^{-1}$) with no significant difference in mean exercise time to exhaustion (1306 vs. 1217 s) or mean times for 3 km run (667.5 vs. 699.0 s). The forwards were, however, taller and heavier (mean height 190.2 vs. 179.5 cm, mean body mass 104 vs. 86.3 kg) (Scott et al. 2003).

Fifty-four elite Rugby League (similar game to rugby) players showed that absolute strength varied by playing level. Research suggests that strength increases are in proportion to $mass^{0.6-0.7}$ rather than $mass^{1.0}$ (Atkins 2004). Even though $\dot{V}O_2$ max values appear relatively moderate for elite rugby players, if power-law scaling was used, elite rugby players would have $\dot{V}O_2$ max values comparable to elite endurance athletes such as rowers, cyclists and runners.

An historical study of data from high-standard rugby players from 1905 to 1999 showed that the rates of increase in body mass (2.6 $kg.decade^{-1}$) and body mass index (0.4 $kg.m^{-2}.decade^{-1}$)

were much greater than the rate of the general population of young males. The increases in body mass and body mass index since 1975 have been three to four times those between 1905 and 1975. Players have shown increased mesomorphy (+1.1 units.decade^{-1}) from 1975. Rugby World Cup 1999 ranking was significantly correlated with the average mass of the teams (Olds 2001).

1.2.7.4 Ironman

For events termed ultraendurance (i.e. >4 hr) such as Ironman, performance is difficult to predict. The Ironman (>8 hr) is a three-sport event consisting of a 3.8 km swim (>45 min) and a 180 km cycle (>5 hr), followed by a 42.2 km marathon run (>2.5 hr). The anaerobic threshold is too great an intensity to be maintained during an Ironman, and other factors such as fuel, fluid and electrolyte imbalances cause detriments in prolonged performance (Laursen and Rhodes 2001). A study of 14 Hawaiian Ironman triathletes showed the physique of triathletes to be most similar to that of cyclists. $\dot{V}O_2$ max was, for males and females, respectively: 68.8 mL.kg^{-1}.min^{-1}, 65.9 mL.kg^{-1}.min^{-1} on the treadmill; 66.7 mL.kg^{-1}.min^{-1}, 61.6 mL.kg^{-1}.min^{-1} on the cycle ergometer; and 49.1 mL.kg^{-1}.min^{-1}, 39.7 mL.kg^{-1}.min^{-1} on the arm ergometer. When comparing the highest $\dot{V}O_2$ max in any exercise modes, it was suggested that the male triathletes are comparable to swimmers, but have a lower $\dot{V}O_2$ max than cyclists or distance runners (O'Toole et al. 1987). These $\dot{V}O_2$ max scores are, however, comparable to cyclists and distance runners in other literature (Hue et al. 2000). In the sports of cycling, swimming and Ironman, drafting refers to closely following a competitor to reduce the drag forces from air or water resistance. $\dot{V}O_2$, heart rate and blood lactate were significantly lowered in a variety of drafting positions (Chatard and Wilson 2003).

Sub-maximal gas exchange values correlate with cycling performance for the half-Ironman but not the Ironman. The longer sections of the race have a greater impact upon the following sections in the full Ironman (Whyte et al. 2000). Twenty-nine male competitive triathletes showed no difference between mean cycling $\dot{V}O_2$ max and treadmill $\dot{V}O_2$ max in all triathletes (69.1 vs. 70.2 $mL.kg^{-1}min^{-1}$, respectively) and values of cycling $\dot{V}O_2$ max and treadmill $\dot{V}O_2$ max in elite triathletes (75.9 and 78.5 $mL.kg^{-1}min^{-1}$, respectively) that were comparable to those reported in elite single-sport athletes in these specialities (Hue et al. 2000).

Seventy-one elite and junior elite triathletes showed that race run time had the greatest variation. A regression equation including robustness, adiposity, body segmental lengths and skeletal mass, correlated significantly with total race time for all triathletes, accounting for 47% of the variance in total triathlon duration. Proportionally longer body segmental lengths were related to lower race swimming time (Landers et al. 2000).

1.2.7.5 Running

Running is not thought to be important in human evolution because humans are poor sprinters compared to most quadrupeds. Humans are, however, comparatively good at endurance running which may have influenced the evolution of the human body (Bramble and Lieberman 2004). Limitations in the running research include lack of: longitudinal studies, description of training, rest, nourishment and hydration, use of allometric scaling for $\dot{V}O_2$ max, anaerobic power and physical stature, central nervous system, field data, running economy, strength and training methods (Berg 2003). As stated previously, $\dot{V}O_2$ max (in $mL.kg^{-1}.min^{-1}$) is considered the best predictor of running performance (Nevill et al. 2004a).

1.2.7.6 Swimming

High cardiac output is probably not crucial for swimming since swimmers show higher values during running. Maximal heart rate is approximately 10 beats.min^{-1} lower during swimming than running. Most likely active muscle mass is smaller and rate of power production lesser in swimming, probably due to local muscle factors, requiring mostly swim-specific training (Holmer 1992). An analysis of the mechanics and energetics of swimming reveals that a 10% increase in propelling efficiency resulted an in improvement in performance greater than increasing the maximal aerobic or anaerobic power by 10% (Toussaint and Hollander 1994).

Immune system problems appear to be a feature of elite swimming training. The intensive training of elite swimmers over short and long timeframes suppresses systemic and mucosal immunity, and may lead to infection at competition time (Gleeson et al. 1995). Very high training volumes are a feature of elite swimming training compared to most sports, possibly due to the low impact of swimming, so perhaps genetic aspects of immunity and recovery are very important for elite swimmers.

Overall, the athletes selected for the present study come from sports where high physiological performance ability, and high $\dot{V}O_2$ max in particular, are clearly important requirements, rather than mainly high skill level. Caution has to be exercised in using blindly using these values without taking into account the often wide variation in body size in particular sports and events.

1.3 Basic Genetics

1.3.1 Cells and Chromosomes

Cells are the basic biological units of all organisms, with the exception of viruses, and under restricted conditions, cells are capable of existing independently. All cells are derived by cell division from other cells, going back approximately 3.5 billion years. Prokaryotes and eukaryotes are the two major classes. Prokaryotes have a simple internal structure, no defined nucleus and a single small circular chromosome of DNA (deoxyribonucleic acid). Eukaryotes have a complex intracellular structure with internal membranes, a membrane-bound nucleus, and an organised cytoskeleton. Eukaryotic cell nuclei have several linear chromosomes each containing a single extremely long DNA molecule, packaged by proteins. Eukaryotes probably first appeared about 1.5 billion years ago. Multicellular organisms begin as a single cell before multiplying through repeated cell division, cell differentiation and cell turnover. The metabolic activities of cells are controlled by diffusion and rates of diffusion fix some upper limits on cell size (Alberts et al. 2002; Strachan and Read 1999).

The nuclei of human cells hold over 99% of the cellular DNA. Most human cells are diploid and contain two copies of the human genome. Mature red blood cells are a rare exception of cell types without a nucleus and without, therefore, a copy of its genome. The ovary and testis have specialised diploid cells that divide by meiosis to produce haploid gametes (egg and sperm). Each gamete has 22 autosomes (nonsex chromosomes) as well as only one sex chromosome. Eggs always have an X, but sperm can have an X or a Y sex chromosome. After fertilisation the zygote is diploid 46,XX or 46,XY. The human genome contains approximately 3,200 Mb (million nucleotides) of DNA. Human chromosomes have between 50 Mb and 260 Mb. The nuclear base composition of the human genome is about 42%

guanine+cytosine (GC). There are approximately 25,000 genes in the human genome (Brown 2002; Strachan and Read 1999).

1.3.2 DNA, RNA and Proteins

Molecular genetics is the study of how DNA and RNA (ribonucleic acid) molecules synthesise the polypeptides and thence proteins. Certain sections of DNA molecules serve as templates for synthesising RNA molecules. Most RNA molecules are used to specify the synthesis of polypeptides, either directly or by altering different stages of gene expression. Proteins are the common end-points of the DNA template and constitute the majority of the dry weight of a cell. Proteins have vital roles in varied cellular functions including as enzymes, receptors, storage proteins, transport proteins, structural proteins, transcription factors, signalling molecules and hormones (Alberts et al. 2002; Brown 2002; Strachan and Read 1999).

Individual DNA molecules are present in the chromosomes of the nucleus and in mitochondria of eukaryotes. They consist of a backbone of alternating 5 carbon sugars (pentose) called deoxyribose, and phosphate residues (α, β and γ), linked by covalent phosphodiester bonds, forming large polymers. Carbon atom number 1′ (one prime) of each sugar residue is covalently bound to one of four types of nitrogenous base: adenine (A), cytosine (C), guanine (G) and thymine (T). The bases comprise heterocyclic rings of carbon and nitrogen atoms. They consist of two types: purines (A and G) have two rings; pyrimidines (C and T) have one ring. A nucleoside is a sugar with an attached base and with the addition of a phosphate group at carbon atom 5′ (five prime) or 3′ (three prime) becomes a nucleotide which is the basic repeating unit of DNA. RNA molecules differ from DNA molecules by having ribose sugar residues instead of deoxyribose and uracil (U) rather than thymine (Brown 2002; Strachan and Read 1999).

Polypeptide molecules contain polymers comprising a sequence of amino acids which consist of a positively charged amino group and a negatively charged carboxylic acid (carboxyl) group connected by a central carbon atom that has an identifying side chain. There are 20 different amino acids grouped into different classes according to their side chains: polar (acid or basic), nonpolar, hydrophobic or hydrophyllic. Proteins are made up of these polypeptide molecules which may be modified by the addition of various carbohydrate or other side chains (Strachan and Read 1999; Winter et al. 2002).

1.3.3 DNA structure and replication

The superstructure of single DNA and RNA molecules is an antiparallel (the polarity of one strand is oriented opposite to the other) double helix of alternating sugar residues and phosphate groups. It has a 2.37 nm helical diameter, 0.34 nm between base pairs lengthwise, and 3.4 nm between complete turns of the helix (ten base pairs). The sugar residues are linked together with 3′, 5′-phosphodiester bonds. A phosphate group links carbon atom 3′ of a sugar to carbon atom 5′ of the neighbouring sugar. During DNA synthesis (replication), the two DNA strands of each chromosome unwind to produce two identical daughter DNA duplexes. DNA polymerase enzyme catalyses the synthesis of new DNA strands with the four deoxynucleoside triphosphates (dATP, dCTP, dGTP, dTTP) as raw material (Alberts et al. 2002; Brown 2002; Strachan and Read 1999).

1.3.4 RNA transcription and gene expression

DNA codes for RNA and RNA codes for polypeptides (which constitute proteins), and is mostly a one-way information system. The DNA → RNA → polypeptide (protein) flow of genetic information is known as the central dogma of molecular biology (Figure 1-2).

Transcription is the first stage of gene expression and is the synthesis of RNA using a DNA-dependent RNA polymerase in eukaryotic cells' nuclei and in mitochondria. The transcriptome is the initial product of genome expression. Translation (polypeptide synthesis) is the second stage of gene expression and occurs in ribosomes in the cytoplasm and in also in mitochondria. Messenger RNA (mRNA) is RNA that specifies the polypeptide. RNA is decoded in groups of three nucleotides (codons) to provide a linear amino acid sequence for the polypeptide. The proteome is the final product of genome expression (Brown 2002; Strachan and Read 1999).

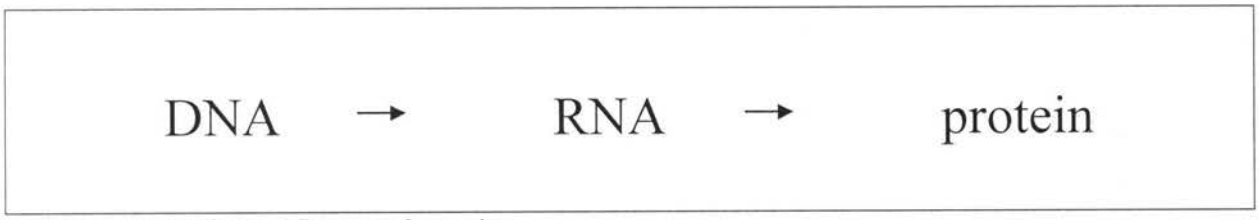


Figure 1-2 The Central Dogma of genetics.

Just a small part of the DNA in cells is transcribed and just a part of the RNA made by transcription is translated into polypeptide. The reasons for this are: (1) the expression of some transcription units give an RNA molecule other than mRNA such as ribosomal RNAs, transfer RNAs, small nuclear and cytoplasmic RNA molecules; (2) the primary transcript is put through RNA processing events; and (3) only the inner portion of the mRNA is translated; the ends remain untranslated. RNA constitutes 20–30 pg (1%) of cell mass in humans. The fraction of coding DNA in the genomes of complex eukaryotes is small due to the many noncoding and repeated sequences which are nonfunctional or not transcribed into RNA (Brown 2002; Strachan and Read 1999).

RNA polymerase enzyme performs RNA synthesis, with DNA as a template and using ribonucleoside triphosphates (rATP, rCTP, rGTP and rUTP) as precursors. The RNA is synthesized in the 5′ → 3′ transcription direction, as a single strand. The RNA chain is

elongated by adding the corresponding ribonucleoside monophosphate residues (AMP, CMP, GMP or UMP) to the free 3′ hydroxyl group at the 3′ end of the RNA chain. The extreme 5′ end nucleotide (the initiator nucleotide) has a 5′ triphosphate group. Double-stranded DNA is unwound during transcription and the DNA strand produces a temporary double-stranded RNA-DNA hybrid with the budding RNA chain. Upstream or downstream of a gene sequence refers to being towards the 5′ or 3′ end, respectively. The bulk of cellular genes, transcribed by RNA polymerase II, encode polypeptides. Eukaryotic RNA polymerases require special DNA sequences to initiate transcription. The promoter is a key group of short sequence elements often grouped upstream of the coding sequence of a gene. Various combinations of short sequence elements located near a gene provide recognition signals for transcription factors to bind to the DNA to activate the polymerase. An RNA polymerase binds to the transcription factor complex, after some general transcription factors bind to the promoter region, and the synthesis of RNA is initiated. The transcription factors are *trans*-acting, because they are synthesised by genes which are elsewhere. The promoter elements are *cis*-acting because they only function on their own DNA duplex (Brown 2002; Strachan and Read 1999).

The promoter always has a TATA box (frequently TATAAA) for genes which are actively transcribed by RNA polymerase II, about 25 bp upstream (−25) of the transcription start site. TATA DNA variants cause the startpoint of transcription to change position. The promoters of housekeeping genes and other genes commonly have a GC box instead of a TATA box, with variations of the consensus sequence (GGGCGG). The CAAT box (~ −80) is usually the strongest determinant of promoter efficiency. GC and CAAT boxes function in either orientation. Tissue-restricted transcription factors only recognise specific recognition elements (Strachan and Read 1999). Enhancers are groups of *cis*-acting sequences, positioned an unfixed distance from the transcriptional start site, which can enhance the transcription of

particular eukaryotic genes, independent of their orientation. They bind regulatory proteins causing the DNA between the promoter and enhancer to loop out. The enhancer-bound proteins can, as a result, interact with the promoter-bound transcription factors, or with the RNA polymerase. Silencers are similar regulatory elements but act to inhibit transcription. The DNA of specific eukaryotic cells from the same organism are virtually identical. The cell type variation is caused by the pattern of gene expression which defines the functions of the cell. Housekeeping gene functions are common and are vital for general cell functions, as opposed to the tissue-specific gene expression of other genes (Brown 2002; Strachan and Read 1999).

1.3.5 RNA processing

The RNA transcript of most eukaryotic genes is processed to remove unwanted internal segments (RNA splicing). The coding sequences contain segments (exons) which are separated by noncoding sequences (introns). RNA splicing of the RNA transcript causes the intronic RNA segments to be discarded and the exonic RNA segments to be spliced end-to-end to form a shorter RNA sequence. The nucleotide sequences of exon/intron boundaries (splice junctions) control RNA splicing. It is dependent on the GT-AG rule which is that introns almost always begin with GT (GU for RNA) and end with AG. Sequences adjacent to the GT and AG dinucleotides are important because they are highly conserved, as well as the branch site, which is usually within −40 of the terminal AG dinucleotide. Splicing involves: (1) cleavage at the 5′ splice junction; (2) nucleolytic attack by the terminal G nucleotide of the splice donor site at the invariant A of the branch site to form a lariat-shaped structure; and (3) cleavage at the 3′ splice junction, leading to release of the intronic RNA as a lariat, and splicing of the exonic RNA segments (Brown 2002; Strachan and Read 1999).

The splicing reactions are mediated by the spliceosome, an RNA-protein complex, which recognises a 5′ splice site and then scans the RNA sequence for the next 3′ splice site. The order of splicing is controlled by the shape of the RNA which probably affects 5′ splice site accessibility (Staley and Guthrie 1998). A particular nucleotide linkage is added, for RNA polymerase II transcripts, to the 5′ end of the primary transcript (capping), and adenylate residues are added to form a poly(A) tail to the 3′ end of mRNA (polyadenylation). A major element that signals 3′ cleavage for the majority of these transcripts is the AAUAAA sequence and cleavage occurs 15–30 bp downstream. Transcription often continues for thousands of nucleotides after this point until termination occurs at a later position where about 200 adenylate residues are added by the poly(A) polymerase to create a poly(A) tail (Brown 2002; Strachan and Read 1999).

A key means for control of gene expression is alternative splicing which gives a limited number of genes great proteomic complexity. The interaction of cis-acting sequences and trans-acting factors modulates the splicing of regulated exons (Caceres and Kornblihtt 2002; Stamm et al. 2005). Alternative splicing changes the sequence of transcripts and the structure of their proteins. More than 25% of all alternative exons, alongside nonsense-mediated decay, are predicted to regulate transcript abundance. Recent molecular analyses show that alternative splicing determines the binding properties, intracellular localisation, enzymatic activity, protein stability and posttranslational modifications of many proteins. The scale of the effects range from a loss or gain of function to slight modulations (commonly observed). Alternative splicing factors regulate multiple specific and congruent pre-mRNAs. Alternative splicing appears to control physiologically significant changes in protein expression and is a primary mechanism of complex organism variation (Stamm et al. 2005).

The important positions in GU-AG introns (which are all spliced similarly) are pointed to by conserved sequence motifs. The first two nucleotides of the intron sequence are 5′-GU-3′ and the last two are 5′-AG-3′, in most pre-mRNA introns. These conserved motifs were recognised soon after introns were discovered and it was immediately assumed that they must be important in the splicing process. As intron sequences started to accumulate in the databases it was realised that the GU-AG motifs are merely parts of longer consensus sequences that span the 5′ and 3′ splice sites. These consensus sequences in vertebrates are: 5′ splice site, 5′-AG↓GUAAGU-3′; and 3′ splice site, 5′-PyPyPyPyPyPyNCAG↓-3′ ('Py' is one of the two pyrimidine nucleotides (U or C), 'N' is any nucleotide, and the ↓ indicates the exon-intron boundary.) Other conserved sequences are present in some but not all eukaryotes.

In higher eukaryotes, introns usually have a polypyrimidine tract (pyrimidine-rich region) positioned slighly upstream of its 3′ end. In pre-mRNA of more than two introns, there is the chance of the incorrect splice sites being joined, since all splice sites are alike, resulting in exon skipping. Selection of a cryptic splice site, a location within an intron or exon that has sequence similarity with the consensus motifs of real splice sites, would also lead to incorrect splicing. There are cryptic sites in most pre-mRNAs and the splicing apparatus must disregard them. Over 35% of genes in the human genome use alternative splicing (Graveley 2001). It is not known how alternative splicing and the selection of its various pathways are regulated. Splicing factors called SR proteins (which are rich in serine, S and arginine, R) in combination with exonic splicing enhancers and exonic splicing silencers are thought to be involved, but how they direct splice site selection is not understood (Brown 2002). Pyrimidines, and the functional effects of the alternative splice sites and the polypyrimidine tract and branch site, are important for the DNA variants in the genes investigated in Chapters 4 and 5 (Wang and Marin 2005).

1.3.6 Translation and protein structure

The transcribed mRNA migrates to the cytoplasm where the ribosomes and other apparatuses control the synthesis of polypeptides. In a usual eukaryotic mRNA molecule, just the central section of it is translated. The flanking sequences, 5′ and 3′ untranslated regions (5′ UTR; 3′ UTR), are copied initially to aid in mRNA binding. Codons in the mRNA sequence are decoded to give individual amino acids. There are four possible bases at each of the three base positions in a codon and, therefore, 64 (= 4^3) possible codons, but only 20 different amino acids and just over 30 types of cytoplasmic transfer RNA with different anticodons. The normal A-U and G-C rules apply for the pairing of codon and anticodon for the first two base positions, but G-U base pairs are permitted at the third position. The decoding is mediated by transfer RNA molecules, which each have a particular amino acid covalently bound by a particular amino acyl transfer RNA synthetase. Each transfer RNA has a specific trinucleotide sequence (anticodon) at a specific spot and the relevant codon of the mRNA molecule must be recognised via base-pairing with a complementary anticodon of the correct transfer RNA molecule (Brown 2002; Strachan and Read 1999).

Ribosomes are big RNA-protein apparatus for polypeptide synthesis. One model of translation is that the ribosomal subunit initially recognises the 5′ cap through the proteins that bind to the cap and then scans along the mRNA until it finds the initiation codon, which is almost always AUG (methionine) and usually the first AUG. The AUG is recognised efficiently only when it is embedded in a suitable sequence, the optimal being the sequence: GCC**Pu**CC*AUG***G** with the purine (Pu; preferably A) preceding it by three nucleotides and the G following it (Kozak 1996). Amino acids are added to the polypeptide chain by the amino group reacting (condensation) with the carboxyl group of the previous amino acid in the chain, resulting in a peptide bonding (Brown 2002; Strachan and Read 1999).

Translation continues until a termination or stop codon is reached (UAA, UAG or UGA in nuclear-encoded mRNA). The backbone of the primary protein structure will, consequently, have a methionine with a free amino group (the N-terminal) end and an amino acid with a free carboxyl group (the C-terminal) end. Ribosome binding is the major stage directing translation. The 5′ UTR (usually <200bp) and 3′ UTR (usually >200bp), which may interact to enhance translation, both play critical roles in mRNA recruitment for translation, in addition to the 5′ cap. There are many *cis*-acting elements that are involved and some *trans*-acting factors which bind to these elements. The 3′ UTR has a primary role in translational regulation and signals for controlling translation, mRNA stability and localisation have all been found in this region (Strachan and Read 1999; Wickens et al. 1997).

Primary translation products are often modified covalently to the polypeptide chain during translation and post-translation such as by hydroxylation or phosphorylation of the side chains of single amino acids or the addition of carbohydrates or lipids. The proteins that are synthesised have different functions requiring them to be secreted from the cell or sent to specific intracellular locations, and a specific localisation signal (signal sequence) is embedded in the polypeptide so that it can be sent to the right place and is discarded by a signal peptidase after sorting. A human cell contains about 20,000 different proteins, accounting for approximately 0.5 ng (18–20%) of the cell weight. Proteins are often post-translationally modified. Specific cofactors (e.g. divalent cations or small molecules for functional enzyme activity) or ligands (protein binding molecule) can affect their conformation. There are at least four different levels of structural organisation for proteins: primary, secondary, tertiary and quaternary (Strachan and Read 1999; Winter et al. 2002).

There are many opportunities for hydrogen bonding between different residues within a single polypeptide chain. Irrespective of the side chains, the O_2 of a carbonyl group of the peptide

bond can hydrogen bond to the hydrogen of the nitrate group of another peptide bond. Fundamental structural units defined by hydrogen bonding between adjacent amino acid residues of a single polypeptide create the secondary structure:

- The α-helix is a stiff cylinder, common for the transcription factor DNA-binding domains. It is characterised by hydrogen bonding between the carbonyl oxygen of a peptide bond with the hydrogen atom of the amino nitrogen of a peptide bond four amino acids away.
- The β-pleated sheet contains hydrogen bond formation between opposed peptide bonds in parallel segments of the same polypeptide chain forming the core of most globular proteins.
- The β-turn contains hydrogen bonding between the peptide bond carboxyl group of amino acid residue n of a polypeptide with the peptide bond nitrate group of residue n+3 results in a hairpin turn allowing compact globular shapes, often in β-pleated sheets (Brown 2002; Strachan and Read 1999).

Complex tertiary structural motifs of two or more of the above structural modules form protein domains (compact regions of a protein of the primary structure forming adjacent elements of secondary structure). These domains often act as functional units involved in binding other molecules. Amongst the polypeptide chains are covalent disulfide bridges between the sulfhydryl groups of pairs of cysteine residues (Brown 2002; Strachan and Read 1999).

1.3.7 DNA Sequence Variants

1.3.7.1 Mutation and Polymorphism

DNA can undergo a variety of types of heritable change. Major chromosome defects involve loss or gain of chromosomes or rearrangement of chromatids. Minor DNA variants can be differently classed and can also be categorised on whether they involve a single DNA sequence (simple DNA variants) or they involve allelic or nonallelic sequence exchanges. Three classes of small-scale DNA variant can be distinguished:

- Base substitutions - involve replacement of usually a single base, which are the variants being investigated in Chapter 4 Cardiac Myosin Binding Protein C and Chapter 5 Endothelial PAS Domain Protein 1.
- Deletions - one or more nucleotides are eliminated from a sequence, which are the variants being investigated in Chapter 3 Angiotensin I Converting Enzyme I/D Polymorphism.
- Insertions - one or more nucleotides are inserted into a sequence which is the variant being studied in Chapter 3 Angiotensin I Converting Enzyme I/D Polymorphism (Brown 2002; Strachan and Read 1999).

New DNA variants arise in somatic cells or in the germline. A germline DNA variant can spread through a (sexual) population if it does not prevent the organism having offspring who can transmit the DNA variant.

A DNA polymorphism is when an allelic sequence variation occurs in a human population with a frequency greater than 1%. The mean heterozygosity for human genomic DNA is approximately 0.1–0.4% (Nickerson et al. 1998; Taillon-Miller et al. 1998). DNA variations drive evolution, but they can also be pathogenic by being the cause of a phenotypic

abnormality or increasing disease susceptibility. The low level of DNA variation permits a balance between occasional evolutionary changes at the cost of causing disease or death in a minority. Most DNA variants are copying errors from DNA replication because DNA polymerases are error-prone (Strachan and Read 1999).

1.3.7.2 Simple DNA variants

Under normal circumstances the greatest source of DNA variations is from endogenous mutation, mainly spontaneous errors in DNA replication and repair, but some are due to various mutagens in the environment. There are two classes of the frequently occurring base substitutions:

- Transitions are substitutions of a pyrimidine (C or T) by a pyrimidine, or of a purine (A or G) by a purine.
- Transversions are substitutions of a pyrimidine by a purine or of a purine by a pyrimidine (Strachan and Read 1999).

There are two options for transversion, but only one option for a transition, when a base is substituted. Transitions may be favoured over transversions in coding DNA because they usually result in a more conserved polypeptide sequence. The excess of transitions over transversions in coding and noncoding DNA is partially due to the high frequency of C→T transitions, resulting from instability of cytosine residues occurring in the CpG dinucleotide which is a hotspot for mutation (by an order of magnitude) in vertebrate genomes (Cooper and Youssoufian 1988).

Many DNA variants are produced randomly, so coding and noncoding DNA are about equally susceptible to mutation. Nevertheless, the chief consequences of mutation are within the

coding DNA (approximately 3% of the human genome). There are two types of DNA variants which occur there:

- Synonymous (silent) DNA variants that do not change the sequence of a gene product.
- Nonsynonymous DNA variants that cause an altered sequence in a polypeptide or functional RNA.

Synonymous DNA variants are thought to be effectively neutral, whereas nonsynonymous DNA variants can be grouped into three classes: harmful effect; no effect; and beneficial effect. Most new nonsynonymous DNA variants are likely to have a harmful effect on gene expression and so can result in disease or lethality but the frequency is reduced because of natural selection and hence is less common in coding DNA than in noncoding DNA. The coding DNA (and regulatory sequences) shows, therefore, high evolutionary conservation. Nucleotide substitutions in noncoding DNA usually have no net effect on gene expression, except when in regulatory elements such as promoter elements, splice sites (junctions) or branch sites. Substitutions occurring in coding DNA have a nonrandom pattern of substitutions so as to preserve polypeptide sequence and biological function (Brown 2002; Strachan and Read 1999).

Promoter DNA variants can lead to abolition or modulation of gene expression. Deletion, insertion or substitution of nucleotides within the promoter may alter expression. Complete deletion of the promoter abolishes function. Splice site DNA variants can also result in abolition or modulation of gene expression. Conserved GT and AG signals are critically important for normal gene expression. Splice site DNA variants may induce exon skipping or intron retention (Strachan and Read 1999).

The relative mutabilities of individual amino acids are affected by the design of the genetic code and functional similarity of amino acids. Some amino acids may have vital roles, such as disulfide bonding of cysteine for conformation of a polypeptide (see p185, C→G variant). Cysteine residues are strongly conserved and it is one of the least mutable amino acids, as it is the only amino acid that has a sulfhydryl group in its side chain (Collins and Jukes 1994). Serine and threonine (amongst others) have very similar side chains, on the other hand, and substitutions at both the first base (ACX→UCX; where X = any nucleotide) and second base positions (ACPy→AGPy; where Py = pyrimidine) can lead to serine→threonine substitutions. Understandably, serine and threonine are frequently mutated (Collins and Jukes 1994; Strachan and Read 1999).

Genes differ in the rate and type of substitution. Ubiquitin, some histones, calmodulin and ribosomal proteins sequences are highly conserved. In contrast, there are fibrinopeptides which are evolving quickly and without obvious selective constraint. They are discarded when the protein is activated. In the large majority of polypeptide-encoding genes, the rate of nonsynonymous substitution is moderate (Brown 2002; Strachan and Read 1999).

1.3.7.3 Loss of function versus gain of function DNA variants

There are two ways in which a phenotypic change may cause mutation of a gene: loss of function mutation (an amorph or hypomorph); or gain of function mutation (a hypermorph or neomorph). Loss of function mutations generally result in recessive phenotypes. The exact quantity is not essential for most gene products, and half the normal amount is sufficient. Half the normal level is not sufficient for normal function for some gene products and haploinsufficiency results in a dominantly-inherited abnormal phenotype. A nonfunctional mutant polypeptide can interfere with the function of the other normal allele, producing a

dominant negative effect (an antimorph). Dominant phenotypes are often produced by gain of function mutations, because the normal allele does not prevent the abnormal effect of the mutant allele, which can be a control system working incorrectly or a novel function of the gene product (Brown 2002; Strachan and Read 1999).

1.3.7.4 Pathogenic DNA Sequence Changes

Almost all sequence variants in an affected person are non-pathogenic. Screening a panel of 100 patients for mutations in a 3 kb coding sequence, for a genome-wide average heterozygosity of 0.32%, would give about 500 sequence changes. Despite high conservation of coding sequences, this screening would usually reveal rare nonpathogenic variants, along with pathogenic changes. The DNA variant is apt to be unambiguous if the pathogenic mechanism is gain of function. For any given disease, any sequence variant found is probably not pathogenic. Loss of function mutations are usually highly heterogeneous. Functional tests are the only way to see if a DNA variant affects the function, but the nature of the DNA variant can provide an indication.

Pathogenic changes:

- Nonsense DNA variants, deletions of the whole gene, and frameshifts are nearly sure to eliminate the gene function.
- DNA variants that change the conserved GT...AG nucleotides flanking most introns affect splicing, and usually abolish the gene function. The effects of other sequence changes on splicing are uncertain, and require an RT-PCR or an *in vitro* splicing assay.

- A missense DNA variant is more likely to be pathogenic if it affects a part of the protein known to be functionally important, such as in the key DNA-binding protein domains.
- Changing an amino acid often affects function if it is conserved across species (orthologs) or between members of a gene family (paralogs).
- Amino acid substitutions often affect function if they are nonconservative (replace a polar by a nonpolar amino acid, or an acidic by a basic one).
- A sequence change in a disease gene that is present in a *de novo* affected patient and not in the unaffected parents is often pathogenic (Strachan and Read 1999).

1.3.7.5 Single Base Substitution in Exonic DNA

On rare occasions, a single nucleotide substitution within exonic DNA causes defective gene expression by activating a cryptic splice site within an exon. Otherwise, single base substitutions can be classed as synonymous substitutions, missense DNA variants or nonsense DNA variants. Synonymous substitutions result in a new codon producing the same amino acid. Because they are usually neutral DNA variants and not under selection pressure, they are the most commonly observed in coding DNA. These substitutions are often found at the third base of a codon: the altered codon often codes for the same amino acid (third base wobble). Nonsense DNA variants are a type of nonsynonymous substitution where a stop codon replaces a codon specifying an amino acid. Because nonsense DNA variants usually result in a large reduction in gene function, selection pressure makes them rare. Missense DNA variants are nonsynonymous substitutions where the changed codon codes for a different amino acid and has two subgroups:

- A conservative substitution is substitution of an amino acid by another that is chemically alike and the effect on protein function is minimal if it has a similar side chain.
- A nonconservative substitution is replacement of one amino acid by another which has a dissimilar side chain such that there is a charge difference or the polarity of side chains is changed (Brown 2002; Strachan and Read 1999).

1.3.8 Gene Protein Structure to Function

Protein structure is best understood at the primary and secondary level. Amino acids commonly form α-helices or β-sheets, depending on the chemical properties of their side chains. A secondary structure forms around a group of amino acids that favour that particular secondary structure and initiate its formation. It then extends to include adjacent amino acids that either favour or allow the secondary structure, and finally ends when one or more blocking amino acids are reached. Biochemists have been able to infer rules for secondary protein folding, and can predict which secondary structures will be adopted by a polypeptide, by identifying which amino acids are most frequently located in which secondary structures, and by studying the structures taken up by small polypeptides of known sequence (Barton 1995; Rost 2001).

The tertiary and quaternary multi-subunit structures are more difficult to predict. The tertiary structures of most proteins are made up of two or more structural domains which are thought to fold independently. Not all proteins spontaneously refold and it is especially difficult for large proteins, probably because the protein can take alternative partially folded structures at various stages of the folding process, which may prevent the correctly folded tertiary configuration. The information encoded by the protein is determined by the spatial

arrangement of chemical groups on its surface and within its folded structure. They have a variety of functions: biochemical catalysis (enzymes which catalyse the central metabolic pathways, which provide the cell with energy); structure (cytoskeleton and extracellular proteins); movement (contractile proteins); transport (e.g. haemoglobin); regulation (e.g. signalling and activator proteins, hormones and cytokines); protection (antibodies and blood clotting proteins); and storage (e.g. ferritin, which stores iron) (Brown 2002; Strachan and Read 1999).

1.3.9 Gene Homology to Function

Elucidating the functions of unknown genes is done by computer analysis and experimental studies. One of the best tools available for computer analysis is homology searching, which locates genes by comparing a DNA sequence with all the other DNA sequences in the databases. The theory of homology searching is that related genes have similar sequences and so a new gene can be found through its similarity to an equivalent gene from a different organism. Homology analysis can be used to assign a function to a new gene. Homologous genes are ones that share a common evolutionary ancestor, revealed by sequence similarities between the genes. Homologous genes fall into two categories: orthologous genes are those homologs that are present in different organisms and whose common ancestor predates the split between the species; paralogous genes are present in the same organism and often members of a recognised multigene family. If the sequence of a newly discovered gene is similar to a known gene, then an evolutionary relationship can be inferred and its function is likely to be similar to that of the known gene. Usually the gene sequence is converted into the amino acid sequence because the amino acid differences will usually be greater for genes that are unrelated. The most popular software program for this is the online search tool BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990). A match to a known gene, or part of a gene, may indicate the function of the new gene or gene segment. The shared

sequence may encode domains within each protein that indicate the shared function (Brown 2002).

Experimental methods are needed to complement the results of homology studies. The strategies in use are not always able to determine the function/s of the numerous genes being discovered. In conventional genetic analysis, the genetics of a phenotype is usually studied by searching for mutant organisms with altered phenotypes. The mutants might be obtained experimentally with a mutagen, or found naturally. The gene/s that has been altered is then studied by genetic crosses, which can locate its position and determine if the gene is the same as another known gene. The gene can then be cloned and sequenced. The general principle is that the genes responsible for a phenotype can be identified by finding which genes are inactivated in a mutant phenotype. When the gene instead of the phenotype is the starting point for investigation, the strategy would be to mutate the gene and classify the phenotypic change (Brown 2002).

1.3.10 Non-Disease DNA Variants to Phenotype

Molecular variants of a gene may represent inherited predispositions to a phenotype. To comprehend the mechanism by which a genetic factor may predispose to a physiological phenotype is complex. Genetic variation vulnerability modulates response to environmental contact over time. The gene product may be very pleiotropic (influences multiple physiological processes in multiple tissues) (Lalouel 2001).

Many disease genes are not disease specific (common variants/multiple disease hypothesis). Common harmful alleles may be a factor in associated clinical phenotypes in diverse genetic backgrounds and under dissimilar environmental circumstances (Becker 2004).

1.3.11 Modifying genes

There is evidence for a genetic contribution to the pathophysiology of complex diseases. Variants of genes involved in a system are the logical candidate genes. A modifying role for many polymorphisms with weak or inconsistent association with disease seems more likely than as susceptibility genes. Gene-gene interactions and gene-environment interactions also need to be considered (Bleumink et al. 2004). The *ACE I/D* polymorphism may be a modifying variant for heart failure in hypertensive subjects (Schut et al. 2004). Several studies showed that renin-angiotensin system genotypes, including *ACE I/D*, may operate as modifying genes for hypertrophic cardiomyopathy in a disease gene-specific manner (Doolan et al. 2004; Lechin et al. 1995; Ortlepp et al. 2002; Perkins et al. 2005). The phenotype can vary greatly between different families with the same disease gene variants, suggesting that modifying genes and/or environment influence phenotype (Chung et al. 2003). The roles of potential modifying genes/gene variants in the cardiac system are investigated using the elite athlete model in Chapters 3, 4 and 5.

1.4 Gene Discovery

1.4.1 Genetic mapping of complex traits

Clarifying the genetic determinants of non-mendelian diseases is a major challenge in human genetics. The main genetic contribution to illness in the developed world is through common diseases. Identifying the genes involved may suggest new means of prevention or treatment. There are problems, however, in tackling complex diseases with the methods used for mapping mendelian traits (Mayeux 2005; Strachan and Read 1999).

1.4.2 Nonparametric linkage analysis

Model-free or nonparametric methods of linkage analysis ignore unaffected people, and look for alleles or chromosomal segments that are shared by affected individuals. Shared segment methods can be used within nuclear families, called sib (sibling) pair analysis. They can also be used within known extended families, or in whole populations. At the population level they constitute association studies.

1.4.2.1 Affected sib pairs allow model-free analysis

Picking a chromosomal segment at random, pairs of sibs are expected to share 0, 1 or 2 parental haplotypes with frequency ¼, ½ and ¼, respectively. If both sibs are affected by a genetic disease, though, then they are likely to share whichever segment of chromosome carries the disease locus. On the simplest assumption that everybody with the disease carries a mutant allele at this locus, then if the disease is dominant they will share at least one parental haplotype, and if the disease is recessive they will share both haplotypes. This allows a simple form of linkage analysis. Affected sib pairs are typed for markers, and chromosomal

regions sought where the sharing is above the random 1:2:1 ratios of sharing 2, 1 or 0 haplotypes identical by descent. If the sib pairs are tested only for identity by state, the expected sharing on the null hypothesis is a function of the gene frequencies. Multipoint analysis is preferable to single-point analysis because it more efficiently extracts the information about identical by descent sharing across the chromosomal region. The mapmaker/sibs program is widely used to analyse multipoint affected sib pairs data and produce nonparametric lod scores (Kruglyak and Lander 1995).

Because sib pair analysis is model-free, it can be performed without making any assumptions about the genetics of the disease. Thus it has been used as one of the main tools for seeking genes conferring susceptibility to common nonmendelian diseases like diabetes or schizophrenia. One drawback is that candidate regions defined by sib pair analysis are usually uncomfortably large for positional cloning. Sib pair analysis has no process analogous to the end-game of mendelian mapping, where closer and closer markers are tested until there are no more recombinants. It is not likely that a chromosomal segment can be defined that is shared by all affected sib pairs. If a susceptibility factor is neither necessary nor sufficient for disease, then not all affected sib pairs will share the chromosomal segment that contains the susceptibility locus. Moreover, sib pairs share many segments by chance, including segments that coincidentally lie close to a susceptibility locus (Mayeux 2005; Strachan and Read 1999).

1.4.3 Linkage versus association

In principle, linkage and association are totally different phenomena. Association is simply a statistical statement about the co-occurrence of alleles or phenotypes. Allele A is associated with disease D if people who have D also have A more (or maybe less) often than would be predicted from the individual frequencies of D and A in the population. An association can

have many possible causes, not all genetic. Linkage, on the other hand, is a specific genetic relationship between loci (not alleles or phenotypes). Linkage does not of itself produce any association in the general population. Linkage creates associations within families, but not among unrelated people. If two supposedly unrelated people with disease D have actually inherited it from a distant common ancestor, on the other hand, they may well also tend to share particular ancestral alleles at loci closely linked to D. Where the family and the population merge, linkage and association merge (Brown 2002; Strachan and Read 1999).

1.4.4 Linkage disequilibrium mapping

1.4.4.1 Linkage disequilibrium narrows candidate region

A population association may be used to narrow down a candidate region that was initially defined by standard parametric linkage analysis. A disease can be mapped to a broad locus. The initial markers may show no linkage disequilibrium, but new markers from the candidate region may show strong association between the haplotype and the disease. As more markers are isolated, the gradient of linkage disequilibrium indicates the location of the disease gene.

1.4.4.2 Linkage disequilibrium quantification

For positional cloning of a disease where a large number of patients are available, quantitative measures of linkage disequilibrium can be calculated for a series of markers across the target region. Hopefully the disease gene will be located at the peak of disequilibrium. The simplest measures of disequilibrium are affected by the gene frequencies. A better measure is the Yule coefficient (Krawczak and Schmidtke 1998). For two loci A and B with alleles A_1, A_2, B_1 and B_2, this is

$$(p_{1,1} - p_{1,2}) / (p_{1,1} + p_{1,2} - 2 p_{1,1}p_{1,2})$$

where $p_{1,1}$ and $p_{1,2}$ are the frequency of allele A_1 on chromosomes carrying alleles B_1 and B_2, respectively. A sophisticated method of analysis based on maximum likelihood estimation of multipoint data is one of several approaches that appear able to predict gene locations much better than the simple analysis (Strachan and Read 1999; Xiong and Guo 1997).

1.4.5 Significance thresholds in analysis of complex diseases

Whereas most mendelian loci localised by significant logarithm of odds (lod) scores have been successfully cloned, the history of complex disease analysis has been scored by a series of irreproducible results where the candidate regions in the different studies were different. A common problem is deciding when to call the results of a linkage or association study significant. A mendelian condition must map somewhere and so, in linkage analysis, no matter how many markers are used in finding the location, the risk of a false positive result remains manageably low. This is not the case for association studies. There may be no association to find, so each test carries an independent risk of a false positive and to correct for this a statistical adjustment such as Bonferroni correction is often used. The threshold of significance is set at $p = 0.05/n$, where n is the number of independent associations checked (Brown 2002; Strachan and Read 1999).

The difficulty of deciding appropriate thresholds of significance is partly technical and partly philosophical. The distinction between pointwise (or nominal) and genome-wide significance is important:

- The pointwise p value of a linkage statistic is the probability of exceeding the observed value at a specified position in the genome, assuming no linkage.
- The genome-wide p value is the probability that the observed value will be exceeded anywhere in the genome, assuming the null hypothesis of no linkage.

For a whole-genome study, the appropriate significance threshold is a value where the probability of finding a false positive anywhere in the genome is 0.05. Most complex disease studies avoid these theoretical approaches by basing the significance threshold on simulation. A whole-genome search is conducted in each simulated dataset and the maximum lod score noted. The genome-wide threshold of significance is taken as a score that is exceeded in less than 5% of replicates (Brown 2002; Strachan and Read 1999).

Proposed thresholds for linkage of disease susceptibility genes (Lander and Kruglyak 1995):

- Suggestive linkage is a lod score or p value that would be expected to occur once by chance in a whole genome scan.
- Significant linkage is a lod score or p value that would be expected to occur by chance 0.05 times in a whole genome scan (i.e. the conventional $p = 0.05$ threshold of significance)
- Highly suggestive linkage is a lod score or p value that would be expected to occur by chance 0.001 times in a whole genome scan.
- Confirmed linkage is when a significant linkage observed in one study is confirmed by finding a lod score or p value that would be expected to occur 0.01 times by chance in a specific search of the candidate region.

The pointwise p values for significant linkage work out at $1–5\times10^{-5}$ for different genome-wide study designs. Note that these values do not imply threshold lod scores of 4.3–5.0. A lod score of 5 means that the data are 10^5 times more likely on the given linkage hypothesis than on the null hypothesis; a p value of 10^{-5} means that the stated lod score will be exceeded only once in 10^5 times, given the null hypothesis. The two measures are

not the same. The lod scores for genome-wide significant linkage are in the range 3.3–4.0, again depending on the study design (Strachan and Read 1999).

Affected sib pair analysis would require unrealistically large samples to detect susceptibility loci for a complex disease conferring a relative risk of less than about 3, whereas transmission disequilibrium testing might detect loci giving a relative risk below 2 with manageable sample sizes. Susceptibility genes conferring a relative risk below 1.5 would be hard to find by either method (Brown 2002; Mayeux 2005; Risch and Merikangas 1996).

1.4.6 Strategies for complex disease mapping

Linkage and association can provide complementary data. Linkage operates over a long chromosomal range. Linkage analysis, whether parametric or nonparametric, can scan the entire genome in a few hundred tests. A typical study of 250 sib pairs with 300 markers would require $1.5–3\times10^5$ genotypes to be generated. Candidate regions defined by linkage are, however, usually too large for positional cloning. Association tests like the transmission disequilibrium test have the opposite characteristics. Linkage disequilibrium is seldom striking over more than a megabase, so a genome screen by transmission disequilibrium testing would involve huge numbers of tests; on the other hand, a positive result would localise the susceptibility factor rather accurately. A natural study design is, therefore, to start with a genome-wide screen by linkage, probably in affected sib pairs, and then, once an initial localisation has been achieved, to narrow the candidate region by linkage disequilibrium mapping (Strachan and Read 1999).

Linkage disequilibrium is not an inevitable result of tight linkage. Association due to disequilibrium will be seen only if a significant proportion of the disease chromosomes derive

from one not too distant common ancestor. Some serious dominant or X-linked mendelian diseases, however, show no linkage disequilibrium because natural selection ensures a rapid turnover of disease genes and most affected people are the result of independent mutations. For susceptibility factors in common disease, the problem is more likely to lie at the opposite end of the spectrum. Susceptibility factors may be common variants that have existed in the population at high frequency for a very long time, and that are non-pathogenic except under certain circumstances. A very old variant may have reached linkage equilibrium with adjacent markers. Equally, if many different changes to a given gene each act as a susceptibility factor, then there may be no linkage disequilibrium. Even if a susceptibility factor can, therefore, be localised by linkage, it does not necessarily follow that it can be fine-mapped by linkage disequilibrium or a method such as transmission disequilibrium testing that relies on it. Sib pair analysis will only detect rather strong susceptibility factors. Presently, transmission disequilibrium testing and other association testing is limited to testing candidate loci or regions (Brown 2002; Mayeux 2005; Strachan and Read 1999).

1.5 Sports Genetics

"I am convinced that anyone interested in winning Olympic gold medals must select his or her parents very carefully." (Per-Olaf Åstrand, sports physiologist, at a 1967 exercise symposium)

Early studies on the heritability of performance showed that physiological performance variables had high heritability (Klissouras 1971). Claude Bouchard has been the main driver of research into the genetics of performance with his original research on sets of twins (Bouchard and Malina 1983). Strong genetic contributions to exercise performance were found. Genetic tests may give priceless predictive information regarding physical or physiological characteristic that is difficult to measure or for the potential of children at unknown levels of maturity where other tests are only weakly predictive of adult performance (Macarthur and North 2005).

1.5.1 Phenotype to Genotype

Elite athletes have many phenotypic variables that could be related to their genotype: endurance/sprint athlete, sport, eliteness, $\dot{V}O_2$ max, speed, height, mass, professionalism, injury proneness and athleticism. Careful characterisation of the phenotype is important for analysing the effect of different genes on athletic performance.

1.5.2 Polygenic Theory of Quantitative Traits

In 1918, Fisher showed that the characters could be described in mendelian terms if they were polygenic (many gene loci). Any variable character that depends on a large number of small

independent causes will have a normal (Gaussian) population distribution. As more loci are included, the distribution begins to take the shape of a Gaussian curve, smoothed further due to variation caused by environment (Strachan and Read, 1999).

1.5.3 Spectrum of Diseases or Traits

There is a spectrum from pure mendelian to pure polygenic characters or traits (Figure 1-3). Oligogenic traits lie in the middle of the spectrum, controlled by a few major susceptibility loci, possibly with a polygenic background and varying environments. The main statistical tool for analysing the inheritance of nonmendelian characters is segregation analysis which can find major susceptibility loci and help indicate their properties. Whether it suggests that a complex disease is either oligogenic or polygenic, can show that linkage or association studies might be more useful in finding the underlying genes. Complex segregation analysis on a large collection of families to find a familial but non-mendelian trait is difficult because there could be a mix of genetic and environmental factors. The genetic factors could be polygenic, oligogenic or mendelian with any mode of inheritance, and the environmental factors could be both familial and non-familial. A range of genetic mechanisms, frequencies and penetrances are allowed. The maximum likelihood analysis finds which combination of parameter values gives the maximum likelihood for the data (Strachan and Read, 1999).

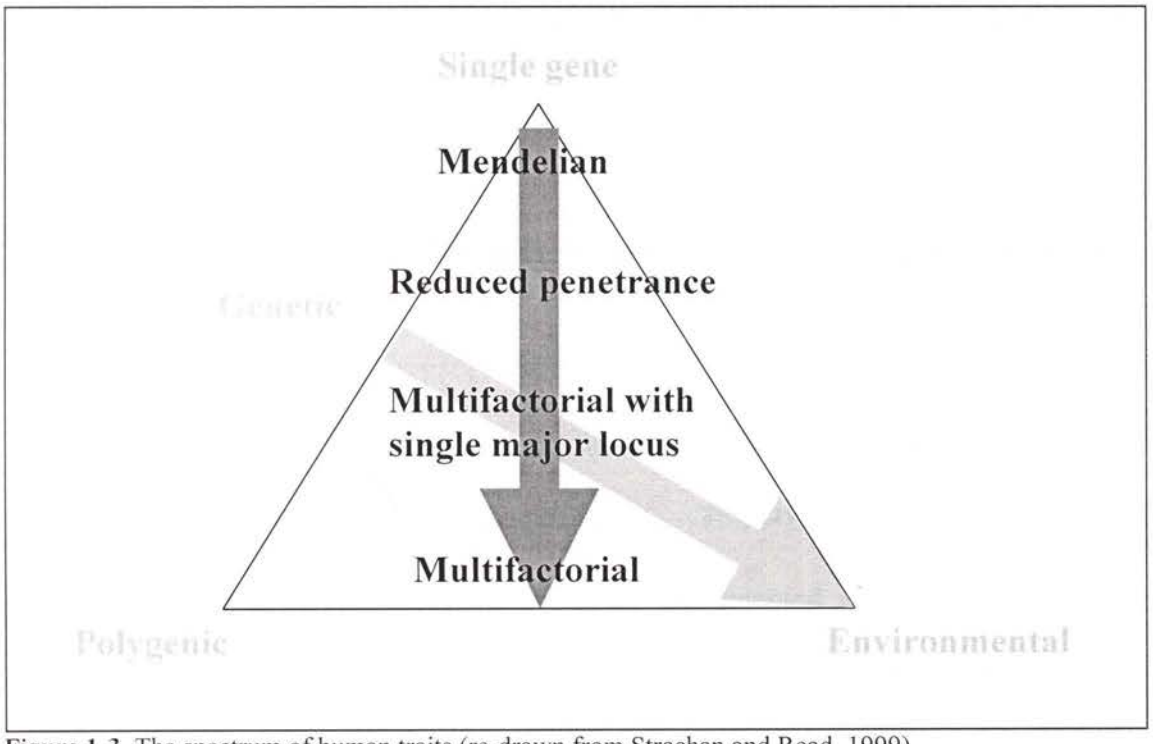


Figure 1-3 The spectrum of human traits (re-drawn from Strachan and Read, 1999).

1.5.4 Human Performance Gene Map

Review papers of the human physical performance gene map and health-related phenotypes have been published annually since 2001 (Perusse et al. 2003; Rankinen et al. 2001). They summarise peer-reviewed papers published by the end of the previous year and contain association studies with candidate genes, genome-wide scans with polymorphic markers, and single gene defects causing exercise intolerance. The latest update is based on peer-reviewed papers published by the end of 2004. The genes and markers with evidence of association or linkage with a performance or fitness phenotype in sedentary or active people, in adaptation to acute exercise, or for training-induced changes are positioned on the genetic map of all autosomes and the X chromosome. A new feature that has been incorporated is the genes whose sequence variants have been associated with either the level of physical activity or indicators of sedentarism. By the end of 2000, in the early version of the gene map, 29 loci were depicted. In contrast, the 2004 human gene map for physical performance and health-

related phenotypes includes 140 autosomal gene entries and quantitative trait loci, plus four on the X chromosome (Wolfarth et al. 2005).

1.5.5 The HERITAGE Family Study and Maximum Oxygen Uptake

There are large individual differences in the training response ranging from 5% to 88% increase in relative $\dot{V}O_2$ max (Lortie et al. 1984). The HERITAGE family study (HEalth, RIsk factors, exercise Training And GEnetics) documented the role of genotype in the cardiovascular, metabolic, and hormonal responses to aerobic exercise training. A total of 90 caucasian families and 40 African-American families with both parents and three or more biological adult offspring were recruited, tested, trained with the same program for 20 weeks, and re-tested. $\dot{V}O_2$ max and many other physiological variables were measured before and after training (Bouchard et al. 1995). Maximal familial heritability for $\dot{V}O_2$ max in the sedentary state was thought to be at least 50% but this estimate could have been overestimated due to nongenetic factors (Figure 1-4) (Bouchard et al. 1998). The maximal heritability estimate for $\dot{V}O_2$ max response is 47% (Figure 1-5) (Bouchard et al. 1999; Bouchard and Rankinen 2001). Perhaps surprisingly, one study showed that age, sex, race, and initial fitness level have little influence on $\dot{V}O_2$ max response (Skinner et al. 2001).

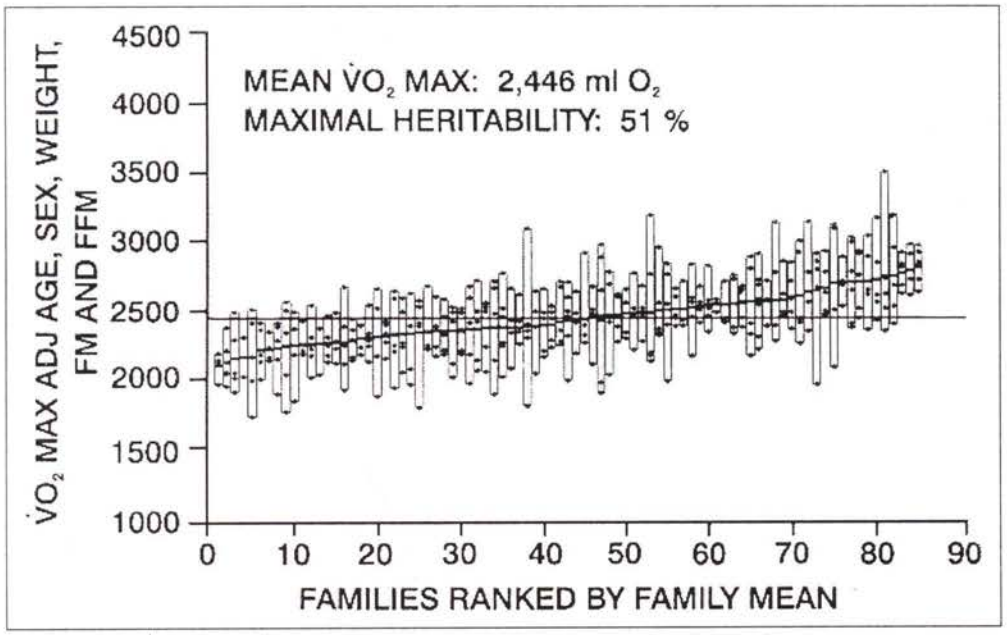


Figure 1-4 $\dot{V}O_2$ **max phenotype (y axis) plotted against family rank**
(families ranked by family mean within a column) (age, sex, weight, fat mass, and fat-free mass adjusted) (Bouchard et al. 1998).

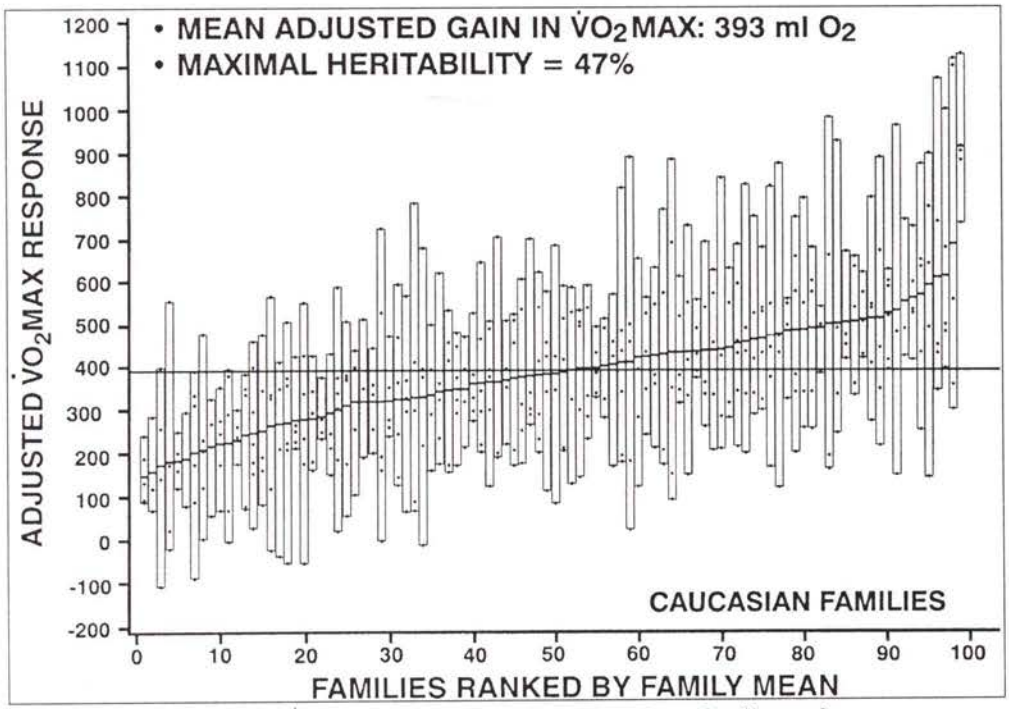


Figure 1-5 Response in $\dot{V}O_2$ **max phenotype plotted against family rank**
(families ranked by family mean within a column) (age- and sex-adjusted) (Bouchard et al. 1999).

1.5.6 Genome-wide Scans

Genome-wide scans for performance phenotypes have only recently been done. These show great promise for locating and identifying the genes with the greatest influence on performance.

1.5.6.1 Genomic Scan for Maximum Oxygen Uptake

Human genomic regions were found that are linked to the baseline $\dot{V}O_2$ max in sedentary individuals and to the responsiveness of $\dot{V}O_2$ max of these same individuals to an equivalent endurance training program. Significant linkages were found at markers on 4q, 8q, 11p, and 14q for $\dot{V}O_2$ max before training and at markers on 1p, 2p, 4q, 6p, and 11p for the change in $\dot{V}O_2$ max in response to a 20 week standardised endurance training program (Bouchard et al. 2000). There were 481 subjects from 99 two-generation caucasian families (236 men, 245 women) between the ages of 17 and 65 years. All subjects were sedentary at baseline with a body mass index $\leq$ 40 $kg.m^{-2}$ and blood pressure <160 mmHg (systolic) and <100 mmHg (diastolic). All subjects were given an equivalent aerobic exercise training program on a computer-controlled cycle ergometer. They trained for 30 min progressing up to 50 min per day three times per week for 20 weeks at a heart rate of 55% progressing up to 75% of $\dot{V}O_2$ max. The exercise intensity was monitored by portable heart rate monitors and was supervised at all times.

$\dot{V}O_2$ max tests were performed twice before and twice after the 20 week training period on bicycle ergometers using gas analysis. The criterion for $\dot{V}O_2$ max was that subjects had to reach one of: (1) respiratory exchange ratio > 1.1; (2) plateau in $\dot{V}O_2$; or (3) heart rate $\leq$ 10 $beats.min^{-1}$ of HRmax. This criterion is fairly loose and can result in subjects reaching

$\dot{V}O_2$ peak and not $\dot{V}O_2$ max. Using bicycle ergometers to test $\dot{V}O_2$ max usually results in scores 5–20% lower than the true value depending on training background specificity due to smaller muscle mass involved and local muscle fatigue in inexperienced and unconditioned subjects (Fernhall and Kohrt 1990). $\dot{V}O_2$ max values were adjusted for age, sex, body mass, fat mass, and fat-free mass using a stepwise multiple regression method. This has to be accounted for when analysing and comparing the genotype results with the $\dot{V}O_2$ max scores of subjects from the present study of elite athletes.

There were two types of statistical analysis used. Firstly, there was single-point linkage analysis using the sib pairs. Secondly, there was multipoint variance components linkage. The single-point linkage analysis used 415 pairs of siblings for baseline $\dot{V}O_2$ max and 327 pairs for response $\dot{V}O_2$ max. Multipoint linkage used all 99 families for linkage analysis. The multipoint linkage program performs path and segregation analysis jointly. It can provide increased power to detect linkages and reduce the likelihood of falsely detecting linkage (Province et al. 2003). Surprisingly, there was no concordance between the suggestive linkage results between single-point and multipoint linkage. The multipoint linkage results, however, could have been affected by the small sample size (99 families), the wide marker spacing and possible misspecification of marker location on the genetic map (Bouchard et al. 2000). The results of the genomic scan (Table 1-1) were used in Chapter 5 in the present study to pick a quantitative trait locus (QTL) (2p16.1) and thence a candidate gene for possible genetic association.

Table 1-1 Summary of the suggestive linkages ($p < 0.01$) with $\dot{V}O_2$ max.
Results for the sedentary state and response to training in the HERITAGE Family Study (Bouchard et al. 2000).

p Value	Chromosome	Marker	Map Position, cM
		Sedentary state	
SEGPATH			
0.0038	4q12	D4S3248	61.658
SIBPAL			
0.0054	8q24.12	D8S592	128.157
0.0084	11p15.1	*SUR*	21.174
0.0031	14q21.3	D14S587	49.255
		Training response	
SEGPATH			
0.0098	4q26	FABP2	127.793
0.0098	6p21.33	D6S2439	28.788
SIBPAL			
0.0090	1p11.2	D1S534	125.032
0.0095	2p16.1	D2S2739	62.602
0.0042	11p14.1	ATA34E08	31.272

$\dot{V}O_2$ max, maximal oxygen uptake; cM, centimorgans.

A similar study, by the same research group, for $\dot{V}O_2$ max and maximal power output found baseline $\dot{V}O_2$ max showed strong evidence of linkage was found on chromosomal regions 11p15 and 10q23 for $\dot{V}O_2$ max and maximal power output in the sedentary state and on chromosomes 1p31 and 5q23 for their responsiveness to training (Rico-Sanz et al. 2004).

1.5.6.2 Genomic Scan for Motor Coordination

Genome-wide linkage analysis of hand motor skill was undertaken in a group of 195 sibling pairs. Hand motor skill was significantly familial (maximum heritability = 41%) and the putative quantitative trait locus was near the chromosome 10p telomere (Francks et al. 2003).

The significance of finding motor coordination genes would be enormous as it would have an impact on practically every sport.

1.5.7 Candidate Performance Genes

When looking for candidate genes for complex traits with small effects, it is logical to look for genes that may affect the limiting factors of these traits.

1.5.7.1 ABO Blood Groups

The first hereditary factor to be examined in relation to physical performance was the ABO red blood cell group (Couture et al. 1986; deGaray et al. 1974). Tests for this at the 1968 Mexico Olympic Games showed no association with athletic performance. The histo-blood group ABO, the major human alloantigen system, involves three carbohydrate antigens (ABH). A, B and AB individuals express glycosyltransferase activities converting the H antigen into A or B antigens, whereas O(H) individuals lack such activity. The A and B genes differ in a few single-base substitutions, changing four amino-acid residues that may cause differences in A and B transferase specificity. A critical single-base deletion was found in the O gene, which results in an entirely different, inactive protein incapable of modifying the H antigen (Yamamoto et al. 1990).

1.5.7.2 *ACE I/D* Polymorphism

The angiotensin I converting enzyme (*ACE*) insertion/deletion (*I/D*) polymorphism has been of interest in sports physiology since 1997 and has been the most thoroughly studied gene in relation to performance. It has been studied in military recruits, mountaineers, Olympic rowers, swimmers, athletics, many other athletes, and sedentary subjects before and after

training (Gayagay et al. 1998; Myerson et al. 1999; Rankinen et al. 2000a). This gene has been controversially dubbed a sports gene but a definitive study remains to be performed. It will be investigated in Chapter 3 Angiotensin I Converting Enzyme (ACE) *I/D* Polymorphism.

1.5.7.3 α-Actinin-3 Gene (*ACTN3*) *R577X* Polymorphism

α-Actinin-3 (ACTN3) is a skeletal-muscle actin-binding protein related to dystrophin. It is completely deficient (577XX) in 18% of Caucasians, <1% of an African Bantu population and 25% of an Asian population. The deficiency is due to homozygosity for a common stop-codon polymorphism in the *ACTN3* gene (denoted R577X). *ACTN3* is specifically expressed in fast-twitch glycolytic myofibres that generate force at high speed. The frequencies of the 577R allele were significantly higher in elite sprint athletes and significantly lower in elite endurance athletes from the Australian Institute of Sport (AIS) compared to controls. The presence of ACTN3 probably increases skeletal muscle force at high speed and gives an evolutionary advantage through increased running speed (Yang et al. 2003). The α-actinin actin-binding proteins and ACTN3 are a highly conserved. They have structural and regulatory roles in the cytoskeleton and in muscle contraction. ACTN3 is the most specialised of the four α-actinins, being expressed mainly in the fast-twitch glycolytic myofibres. The functional/mechanistic significance of ACTN3 is thought to be: (1) to promote fast-twitch muscle fibre formation; (2) to alter glucose metabolism within fast-twitch fibres to affect glycolytic capacity; (3) to affect contractile force; (4) to reduce the muscle damage from high-impact, high-speed movement; or (5) to affect hypertrophy (MacArthur and North 2004). This result was recently replicated in a Finnish cohort (Niemi and Majamaa 2005).

Strenuous eccentric exercise causes muscle damage that shows increases in creatine kinase and myoglobin in the blood. Another recent study found no association of *ACTN3* R577X with the increase in creatine kinase and myoglobin due to eccentric exercise (Clarkson et al. 2005b). This contradicts the muscle damage hypothesis and makes the altered muscle fibre type, contraction force and hypertrophy hypotheses more likely. The same researchers found that about 2% of baseline and response-to-training strength were attributable to *ACTN3* genotype (Clarkson et al. 2005a). Athletes of West African ancestry are typically the fastest runners in the world. As of 2004, they held 494 out of the 500 fastest times in 100 m sprint running (Holden 2004). Muscle viscosity, elasticity and stiffness is greater in black than white athletes (Fukashiro et al. 2002). This supports the hypothesis that *ACTN3* R577R allele increases muscle contractile force.

There was evidence of familial aggregation for Type I fibre area in the sedentary state and enzyme activities of the main energy metabolism pathways of skeletal muscle in the sedentary state and in response to regular exercise (Rico-Sanz et al. 2003b).

1.5.7.4 Adenosine Monophosphate Deaminase 1 Gene (*AMPD1*) *C34T* Polymorphism

The associations of the C34T polymorphism of the adenosine monophosphate deaminase 1 (*AMPD1*) gene with cardiorespiratory phenotypes were tested during cycling exercise at absolute and relative power outputs progressing to exhaustion before and after endurance training for 20 weeks in the HERITAGE Family Study cohort. The TT genotype showed reduced exercise capacity and cardiorespiratory responses in the sedentary state and in the training response of ventilatory phenotypes during maximal exercise (Rico-Sanz et al. 2003a).

1.5.8 Genetic Association Studies

There are similarities between genetic association studies and classic epidemiological studies of environmental risk factors but there are also issues that are specific to studies of genetic risk factors such as the use of particular family-based designs, the need to account for different underlying genetic mechanisms, and the effect of population history (Cordell and Clayton 2005). Recent evidence suggests that common genetic variants will explain at least some of the inherited variation in susceptibility to common disease. Genetic association studies, in which the allele or genotype frequencies at markers are determined in affected individuals and compared with those of controls, may be an effective approach to detecting the effects of common variants with modest effects (Newton-Cheh and Hirschhorn 2005).

Replication of association study findings has a vital role in showing that associations that are identified reflect interesting biological processes rather than methodological quirks (Hattersley and McCarthy 2005). Study design and analysis should examine a range of settings of the important factors: the disease allele frequency or the difference of the disease allele frequency and the marker allele or haplotype frequency in linkage disequilibrium with the disease allele; genotype relative risk (effect size); and phenotype and genotype misclassification error rates. A simple procedure to help insure that genetic association studies will be more robust to error is specifying higher power values when computing sample size requirements (Gordon and Finch 2005). A study of the *ACE* gene showed that greater statistical power can be anticipated with association analysis versus linkage, when markers in strong linkage disequilibrium with a trait locus have been identified. Furthermore, allelic interaction may play an important role in the dissection of complex traits (Zhu et al. 2001).

Genetic association studies are important for disease gene candidacy but are variable in their design and use of statistics. Standardisation guidelines have been proposed. First, the

phenotype has to have already been shown to be heritable or have a reasonable assumption of genetic influence. Second, the candidate gene should be valid and supported by either positional and/or functional *a priori* arguments. Third, subjects should have been carefully matched confounding variables such as race in case-control studies. Fourth, the findings from an initial exploratory analysis should be replicated using an independent sample or through in vitro or in vivo functional studies, if the study was not verifying an established plausible hypothesis. Fifth, associations should be reported in the form of effect sizes and accuracy measures, such as odds ratios and their confidence limits, or if using *p* values, report the total number of associations investigated. Consideration should be made of correction for multiple testing. Sixth, priority for publication of negative results should be for results that favour rejection of a previously published claim of association. Power calculations should be given for negative data. Seventh, the publication of an association already published and independently confirmed is justified only for novel aspects (Cooper et al. 2002). Some would argue that some of these suggestions may introduce biases of their own or prevent stronger confirmation of controversial findings.

Results Chapters 3, 4 and 5 will be using genetic association studies to investigate candidate genes for their relationship to athletic performance. Significant associations will be further investigated by correlation of genetic and phenotype data to confirm the association results.

1.6 Aims of the Present Study

The aims of the present study were to identify genes related to human performance and to assess if these genes affect standard human athletic performance variables. These genes were investigated using a combination of *in silico* search techniques, association studies and multiple regression analysis. These genes were screened for known polymorphisms and for new polymorphisms using a combination of PCR, DHPLC and SNP technology. The variants found in athletes were compared to control groups for differences and compared within athlete groups for correlation with phenotypic data.

1.6.1 Hypotheses

The hypotheses for the present study are that there are genes responsible for human performance, and variations in these genes explain the differences in human performances. These variations should be in greater or lesser frequencies in elite athletes (cases) compared to controls and should correlate to phenotypic data within athlete groups.

1.6.2 Significance of the Present Study

The identification of performance genes and understanding the function of these genes should lead to great advances in Sports Science, especially in the area of talent identification and individualising training programs and event focus for athletes based on their genetic profile. The investigation of modifying genes of the cardiovascular system may lead to breakthroughs in understanding disease causation and also provide alternative targets for new therapies including gene therapy, drug treatments and embryo selection.

Chapter 2
General Materials and Methods

2.1 Subjects

2.1.1 Ethical Implications of the Project and Approval

The ethical implications of subject discomfort or danger were minimal. The blood samples of the elite athletes sourced from the AIS were sub-samples of blood taken for in-house routine testing. The rugby players only provided a buccal cell swab, with no reported discomfort. The Ironman athlete blood samples came from venesection, collected specifically for this project. The project was approved by the Sydney South West Area Health Service Ethics Review Committee, the Human Research Ethics Committee of the University of Sydney and the Human Research Ethics Committee of the AIS.

The ethics of testing genes in sport is controversial. The first question is one that is already applied to athlete tests is whether the testing is reliable and valid for the sport in question. The second question is whether or not genetic selection of athletes is ethical. This question also applies, however, to current physiological testing of athletes. The question of whether it is a form of discrimination to genetically select athletes, when other non-performance selection criteria are used, is valid. Conversely, it could be questioned whether it is reasonable to spend limited resources on athletes who are unlikely to be successful in a particular sport or event. In most situations, a coach or selection panel for a sport is unlikely to reject a potential elite athlete due to the results of testing alone, even if the testing is considered reliable and valid, e.g. measuring height for basketballers. Sports performance under competitive conditions is likely to be the main selection tool for elite sport.

Genetic testing can be justified, though, on the grounds of helping to improve athletic training and the understanding of human performance. This application of sports genetics would not

have too many detractors. It may be unethical to deny the knowledge and benefits that sports genetics could reveal about athletic traning and athletes in general, and apply that information to particular athletes and their training. Genetic testing of human performance is justified on medical grounds if it can help to identify potentially problematic conditions that may arise during certain physical activities or if it can help to understand and treat certain conditions. The Italian Olympic Committee encouranges electrocardiographic, and sometimes echocardiographic, testing for all their elite athletes, to help prevent the occurence of sudden death from undiagnosed cardiac disorders (Pelliccia and Maron 2001).

2.1.2 Elite Athletes

Subjects for the study were male and female athletes from various sports requiring a large proportion of cardiovascular fitness to attain success. The athletes were generally required to be competing at a very high level and in most sports only at the elite level. Elite level was defined as competing at the junior or senior national, international or professional level. These athlete DNA samples (from blood samples) and physiological/phenotype data were sourced from the AIS from 1996 to 2003. Olympians were defined as any athlete that has competed at the Olympic Games. The elite athletes' physical and physiological data were collected by the AIS according to the national sports laboratory testing guidelines. $\dot{V}O_2$ max was expressed as the simple ratio relative to body mass ($mL.kg^{-1}.min^{-1}$).

Ironman competitors' DNA samples were sourced from volunteers at the finish line of various races held in Australia. Only a small proportion of these competitors could be considered elite athletes. They had a very large variation in age, which was defined as age at last birthday, compared to all the other sports. Ironman race finish time was correlated to age, so linear regression was used in the SPSS Statistical Package to remove the confounder of age

and give a new finish time, which no longer correlated with age. Then the top 25% of the new finish time variable were the selected ironmen.

2.1.3 Rugby Subjects

Male rugby players (n = 301) (age ± SD = 23 ± 3 years) volunteered to participate in the study. They were from first division rugby clubs in Sydney and elite professional teams from NSW and the ACT. The players consisted of both amateur and professional players. The players were first divided into caucasian (n = 253) and non-caucasian (n = 48) groups. The caucasian players were then divided into elite (n = 112) and non-elite (n = 141) groups. Elite players were defined as having played at the junior or senior national, international or professional level. The rugby players' physical and physiological field-test data were collected by either the author or the clubs that they played for according to the relevant testing guidelines. Players were then divided into Front-five (set-piece specialists: positions #1–5) and Back-ten (runners: positions #6–15). Set-piece specialists are players who are selected mainly for their very large size and advantages at winning possession of the ball in set-piece (scrum and lineout) activities. At the senior international level, frontrow (#1–3) are very heavy (> 110 kg) and strong for scrummaging, and locks (#4–5) are very tall (> 2 m) and are good lineout jumpers. It was thought that this might initially be a more important factor for selection than their physiological attributes such as endurance or speed. Back-ten (runners: #6–15) are selected for their athletic ability (speed and endurance) and ball-handling skills. They can vary in size but are generally 1.75–1.95 m in height and 85–105 kg in weight. Back-ten players were thought to be more athletically talented since body size was not a defining characteristic of these positions.

2.1.4 Controls

Controls were obtained either from Australian Red Cross donors or from de-identified clinical diagnostic DNA samples that were designated as normal controls because they were partners of clinical patients who were being tested for various genetic diseases. All control subjects from the Red Cross had given written permission for their DNA samples to be used for research.

Controls were age-matched to cases where possible because the importance of age for SNP, and other, studies was uncertain. The aim was to have at least one age-matched control for every case. This was difficult because the initial group of controls came from blood donors and they tended to be older people (~ 40–60 years) whereas the majority of the elite athlete groups were aged 20–30 years. Younger controls were sought out and non-affected partners (< 40 years) of clinical patients were then selected. This enabled one-to-one age-matching for most of the athletes.

2.2 Materials

2.2.1 DNA

DNA came from two sources: white blood cells from peripheral blood or buccal (cheek) cells from a swab.

2.2.2 Chemicals

The chemicals used for this research study were analytical reagent grade.

2.2.3 Oligonucleotides

The oligonucleotides used in this project were designed by the author, unless otherwise stated.

2.3 Methods

2.3.1 Blood Sample Collection

Elite athlete blood samples were collected by the AIS during routine blood testing as well as for specific DNA testing. Ironman samples were collected at various Ironman events around the country by Dr Bing Yu and Dr Jason Gulbin. Buccal cell samples were collected from rugby players at various rugby clubs during training sessions.

2.3.2 Blood DNA Extraction

DNA was extracted from peripheral blood lymphocytes using the Qiagen™ Extraction system.

2.3.3 Buccal Cell Sample Collection

Buccal cell sample collection was performed using the Epicentre Sample Collection Kit™ swab.

2.3.4 Buccal Cell DNA Extraction

Buccal cell DNA was initially extracted using the Epicentre QuickExtract Kit™. There were contaminants, though, in the initial kit that interfered with the PCR (especially in SNP reactions which was probably due to the small reaction volumes) so the DNA was extracted using an in-house system. Samples which had contaminants were further ethanol precipitated to remove the contaminants. Some older DNA samples (from 1996 rowers) were found to need this clean-up procedure for them to work in SNP reactions.

Ethanol precipitation protocol (all steps at room temperature):

1. Add a volume of dH_2O, equal to initial volume, to each DNA sample tube.
2. Add 1/10 volume of above volume of sodium acetate/EDTA buffer (containing 1.5 M sodium acetate and 250 mM EDTA) to each tube and mix well.
3. Add a volume of 100% ethanol to each tube to make a final concentration of 70% ethanol and mix well.
4. Centrifuge the tubes at room temperature in a microcentrifuge for 15 min at ~ 12,000 rpm.
5. Remove supernatant by aspiration from each microcentrifuge tube.
6. Wash the DNA pellets with as large a volume as possible of 70% ethanol. Centrifuge briefly at room temperature for 1 min at ~ 12,000 rpm.
7. Remove the supernatants by aspiration. Air dry the DNA sample for 2–5 min. Do not overdry the pellets or they will become difficult to re-suspend.

The optical density (OD) or absorbance ratio of all DNA samples was checked at least once using a spectrophotometer before doing PCR. The Eppendorf Biophotometer™ was used. The $OD_{260/280}$ ratio > 1.6 was used to check the purity of the DNA. A ratio much lower than this indicated impurity in the sample that would be likely to cause problems with PCR amplification.

2.3.5 Polymerase Chain Reaction (PCR)

PCR has been used in all areas of biological research since 1989. PCR was performed with a variety of PCR machines, including Applied Biosystems GeneAmp PCR System™ 2400, 2700 and 9700. A fragment of DNA may be amplified over a million-fold by the polymerase

chain reaction (PCR) method, a procedure based on repeated cycles of denaturation, primer annealing, and extension by *Taq* DNA polymerase (an enzyme derived from the bacterium *Thermus aquaticus*) (Mullis et al. 1986; Mullis 1990). Target DNA is amplified by 20–40 cycles of DNA synthesis. Each cycle has three stages performed at different temperatures: (1) Denaturing at 94–96°C. (2) Annealing at 55–65°C. (3) Extension at 72°C (P=Polymerase) (Figure 2-1). Each molecule of target DNA acts as a template for the synthesis of new DNA in the next cycle.

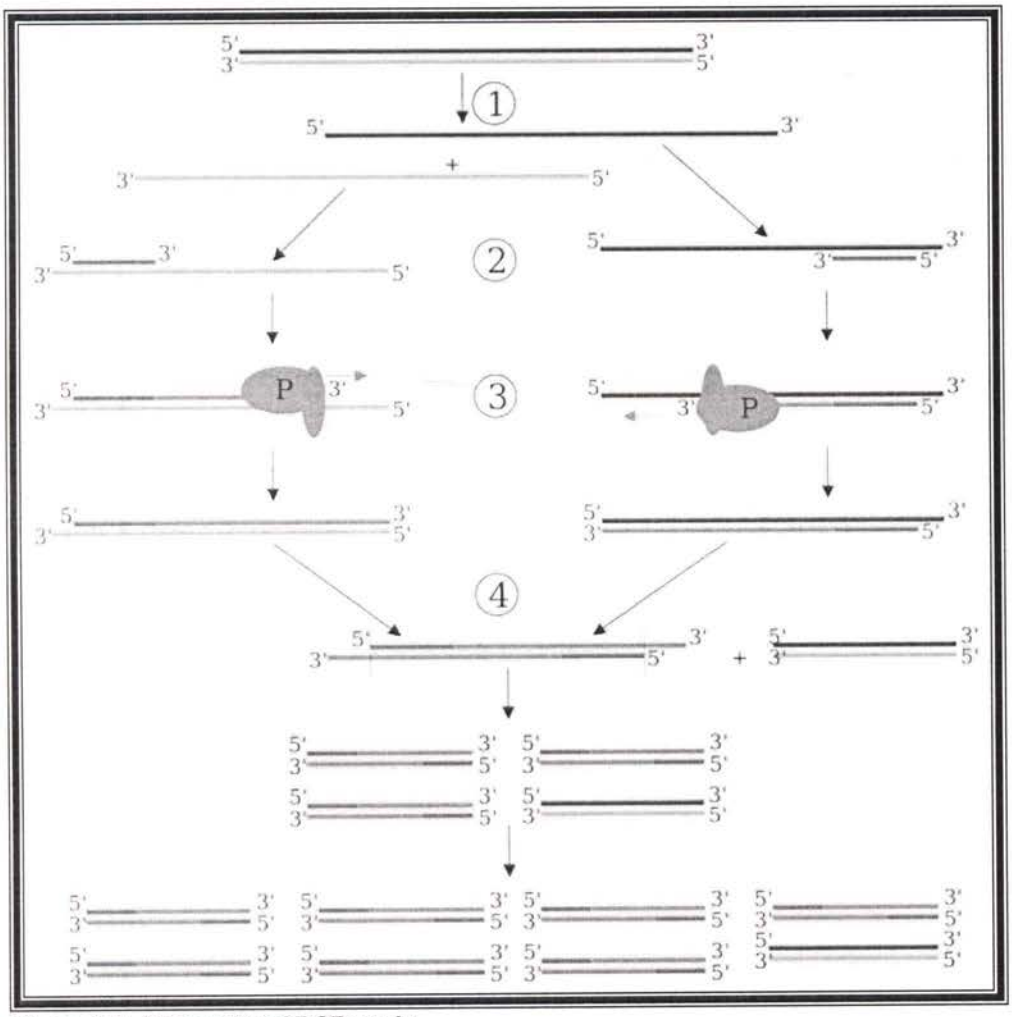


Figure 2-1 Schematic of PCR cycle.
(1) Denaturing at 94–96°C. (2) Annealing at 55–70°C. (3) Extension at 72°C (P = polymerase). (4) End of the first cycle. The two DNA strands produced make up the template DNA for the next cycle, doubling the amount of DNA duplicated in the following cycles (http://en.wikipedia.org/wiki/Image:Pcr.png).

2.3.6 PCR Genotyping

6% polyacrylamide gels were used for genotyping non-SNP assays. The PCR product was analysed using gel electrophoresis and ethidium bromide staining.

2.3.7 Sampling Methods

Athletes were divided into three groups for Chapter 5 SNPs using the power-time curve. If their sport or event duration was less than 50 s, they were considered to be sprint athletes. The other two groups were classified endurance athletes. If their event duration was greater than 50 s and less than 10 min, they were classified as power-time-maximum (PT-MAX). If their event duration was greater than 10 min, including events lasting several hours, it was defined as power-time-steady-state (PT-SS). These criteria were decided upon in conjunction with the AIS and are consistent with the findings of many researchers (Bassett and Howley 2000; De Feo et al. 2003; Jones and Carter 2000; Medbo and Tabata 1989). For the rugby players, it was more complicated. The short periods of activity (usually less than 20 s) are highly anaerobic and the recovery periods (usually less than 30 s) are aerobic. They would certainly be considered sprint/power athletes. The energetics of some positions also requires a high level of endurance.

2.4 Statistics

2.4.1 SigmaStat V1

The χ^2 test is a commonly used test for genetic data using the SigmaStat V1™ (Jandel Scientific Software, California) program.

2.4.2 CLUMP Program

The CLUMP program (Sham and Curtis 1995) (http://www.iop.bpmf.ac.uk) was used for χ^2 test for data where there were small allele or genotype frequencies. CLUMP was run through the DOS program and uses Monte Carlo simulation to perform the χ^2 test.

2.4.3 SPSS™ Regression Modelling

Kevin McGeechan (Associate Lecturer, Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, University of Sydney) advised on how to analyse the genetic results with the athlete physiological data using multiple regression analysis within the SPSS™ program. An SPSS™ file was created out of the various DNA and the de-identified athlete physiological files supplied by the AIS.

Multiple regression was used because there were several explanatory variables from the physiological data. The first step of regression analysis was to generate desriptive summary statistics of the variables. The second step was to identify how many subjects had all the variables. The third step was to generate scatterplots of pairs of continuous variables. The fourth step was to do boxplots of pairs of continuous variables. The fifth step was to turn the genotype categorical variable into continuous variables by creating dummy variables that

could be used as dependent variables in the multiple regression analysis. For example the *ACE I/D* polymorphism was recoded into two variables: ace1, 0/1; and ace2, 0/1. Dummy variables both have to be used at the same time in any regression. *ACE II* was recoded: ace1/ace2 = 0/1; *ACE ID* was recoded: ace1/ace2 = 0/0; and *ACE DD* was recoded: ace1/ace2 = 1/0.

The sixth step was to convert the fitness test and performance scores to z-scores and create an extra variable that combined all the z-scores so that all the sports could be compared using one performance variable. The seventh step was to perform the regression with a fitness test result as the outcome variable and all the other relevant variables as dependent variables. The eigth step was manual iterative backward removal of insignificant dependent variables until only the significant dependent variables remained, as well as the pair of dummy variables for the gene of interest. The ninth step was to perform the regression using syntax to do a subtest of whether the pair of gene dummy variables is significant when added to the model. If they were, then the gene was a significantly explanatory variable for the outcome performance variable.

The last step was to perform regression diagnostics to verify that the data met the assumptions of linear regression. These diagnostics included: was the r^2 value high enough to indicate a good model (closer to 1 than 0); were the residuals normally distributed; were there any plausible interactions; and was there any colinearity. Colinearity was checked by running the Colinearity Diagnostics option in SPSS™. If the variance inflation factor was greater than 10, then there was colinearity between variables and one of them had to be removed.

Chapter 3
Angiotensin I Converting Enzyme (*ACE*) *I/D* Polymorphism

3.1 Introduction

The angiotensin I converting enzyme (*ACE*) *I/D* polymorphism was tested in a large number of elite and non-elite athletes from a variety of sports and compared to a large number of controls for association. The genetic results showed small differences between the various male athlete groups and the age-matched controls. There were statistically significant results for subgroups of the male rowers, cyclists, runners and rugby players. The subgroup of Olympic rowers was highly significantly different (genotype $p = 0.004$; allele $p = 0.002$). The subgroup of track cyclists produced statistically significant results (allele $p = 0.033$). The subgroup of endurance runners produced statistically significant results (genotype $p = 0.046$; allele $p = 0.028$). The subgroup of elite rugby players (genotype $p = 0.020$) and Back-ten rugby players (genotype $p = 0.027$) produced statistically significant results. The comparison of subgroups of Front-five vs Back-ten rugby players was statistically significant (genotype $p = 0.044$).

Multiple regression analysis was performed to determine the significant predictors of performance for all the athletes combined because no specific order of importance was predicted by the literature on exercise physiology. Male and weight were the statistically significant predictors of $\dot{V}O_2$ max ($p < 0.001$). Weight and height were the statistically significant predictors of 2 km Row Time ($p < 0.001$). Male and age were the statistically significant predictors of Ironman Time ($p < 0.001$). The above three sports-specific measures of aerobic fitness were combined in a Fitness Z-score for analysis. Olympian, weight and male were the statistically significant predictors of Fitness Z-score ($p < 0.001$). Weight and age were the statistically significant predictors of 40 m Sprint Time in rugby players ($p <$

0.001). *ACE I/D* was not a significant predictor of any of the above performance measures when added to the models.

The genotyping described in this chapter was predominantly performed by the author. Some genotyping of Blood Bank controls was performed previously by laboratory staff. Some genotyping of newer controls was performed by Jennifer Henderson (PhD student).

3.1.1 Renin-Angiotensin System

The angiotensin I converting enzyme (ACE, or dipeptidyl carboxypeptidase 1, DCP1) is part of the renin-angiotensin system, which is a regulator of blood pressure, sodium and water homeostasis and tissue growth. Angiotensin II is produced by an enzyme cascade. Angiotensinogen is cut by renin to form the angiotensin I, which is then cut by ACE to produce angiotensin II, which is physiologically active (Lavoie and Sigmund 2003) (Figure 3-1).

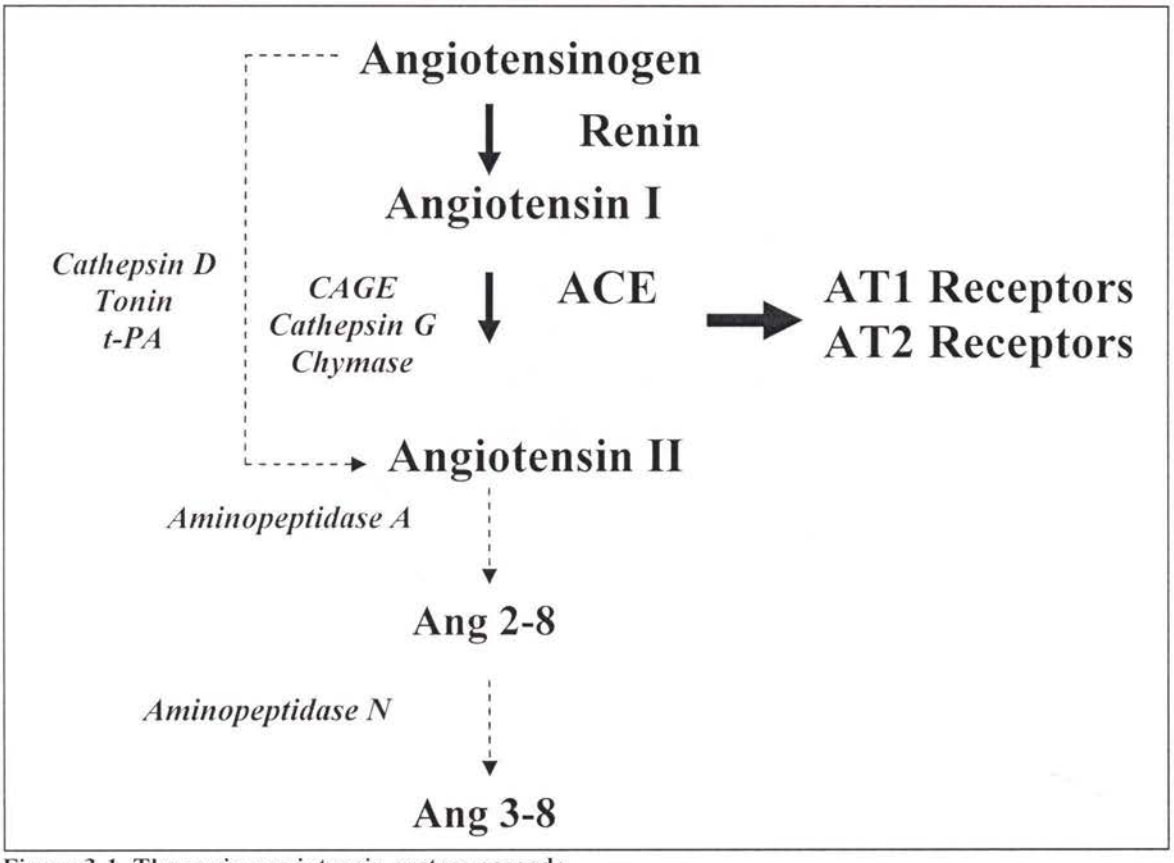


Figure 3-1 The renin-angiotensin system cascade.
The classic cascade of the renin-angiotensin system is shown in *block arrows*. Alternative means for the generation of Angiotensin II is indicated by *dotted arrows*. Ang 3–8, Angiotensin IV; Ang 2–8, angiotensin III; CAGE, chymostatin-sensitive Angiotensin II-generating enzyme; t-PA, tissue plasminogen activator (Lavoie and Sigmund 2003).

Renin was found to be a pressor substance in 1898. The renin-angiotensin system is of great interest as a controller of blood pressure. The functions of angiotensin II are vasoconstriction, renal sodium and water regulation, and thirst. Its regulation of blood pressure was thought to be through the endocrine pathway where blood-borne angiotensin would activate target tissues. A local autocrine or paracrine renin-angiotensin system may also act in many tissues, and help control blood pressure (Lavoie and Sigmund 2003) (Figure 3-2).

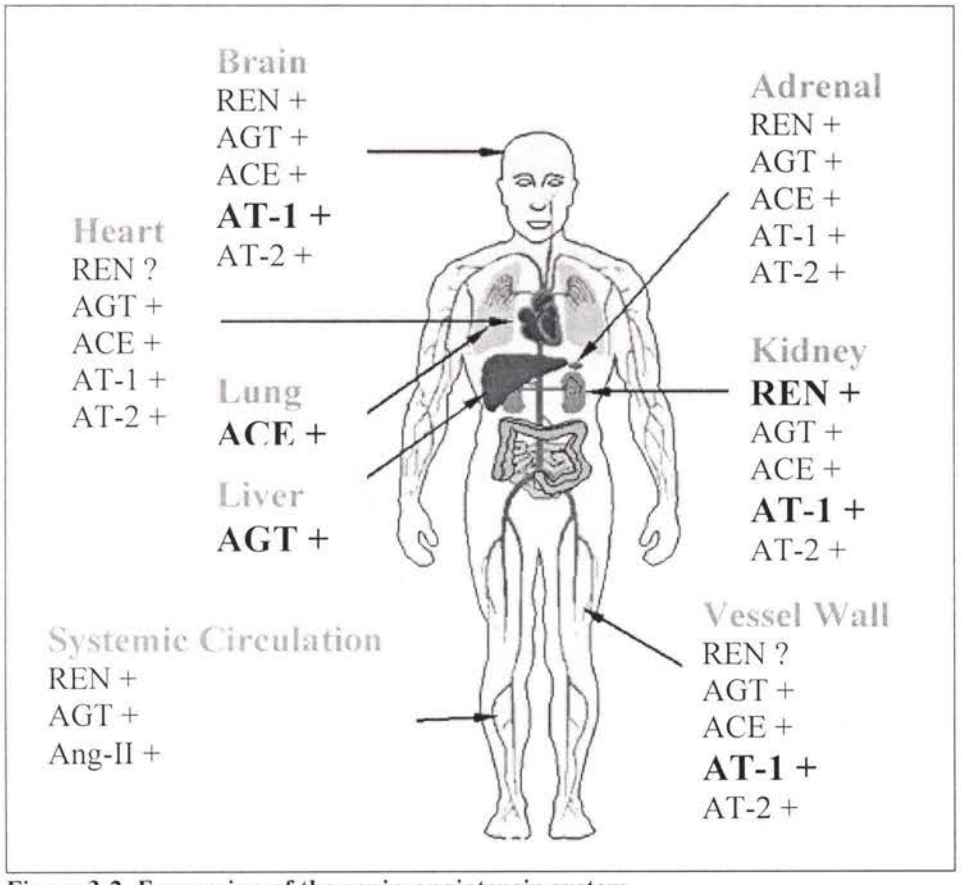


Figure 3-2 Expression of the renin-angiotensin system.
Sites of expression of the different components of the renin-angiotensin system are shown. Classical sites of synthesis for the endocrine renin-angiotensin system are in bold (Lavoie and Sigmund 2003).

3.1.2 *ACE* Gene

The *ACE* gene is found on chromosome 17q23 and contains 26 exons (Figure 3-3, Figure 3-4 and Figure 3-5). It is a zinc-dependent dipeptidyl carboxypeptidase with diverse physiological functions, including principally that of blood pressure regulation via angiotensin II production and bradykinin inactivation. The *ACE* gene encodes an enzyme involved in catalysing the conversion of angiotensin I into the physiologically active peptide angiotensin II. Angiotensin II is a strong vasopressor and aldosterone-stimulating peptide that directs fluid balance and blood pressure. *ACE* has a major role in the renin-angiotensin system. Many studies have correlated the insertion (*I*) or deletion (*D*) of a 287 bp Alu repeat element in intron 16 in *ACE* with the levels of circulating enzyme or cardiovascular events, both normal and pathologic (i.e. hypertension, myocardial infarction, left ventricular (LV) hypertrophy and hypertrophic cardiomyopathy) (Andrikopoulos et al. 2004; Danser et al.

1995; Doolan et al. 2004; Evans et al. 1994; Montgomery et al. 1997; Rigat et al. 1990; Rossi et al. 1999; Schunkert 1997; Schut et al. 2004; Tiret et al. 1992). Two of the most common alternatively spliced variants of this gene encode two isozymes - the somatic form (also known as isoform precursor 1) and the testicular form (also known as isoform precursor 2) (Figure 3-4). These are equally active. Many additional alternatively spliced variants have been identified but their full sequence and features are undetermined. *ACE* isoform precursors 1, 2 and 3 are shown (Figure 3-4).

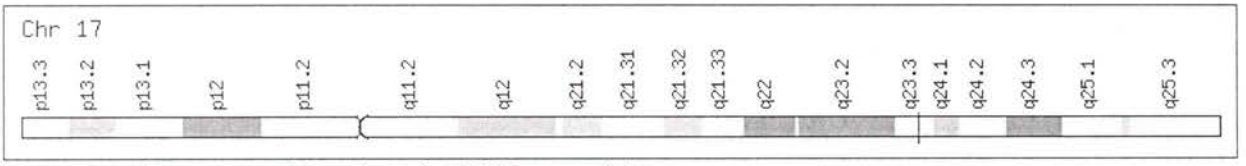


Figure 3-3 Chromosomal location of *ACE* (Genecards).
Start: 58,908,166 bp from pter. End: 58,952,935 bp from pter. Size: 44,769 bp.

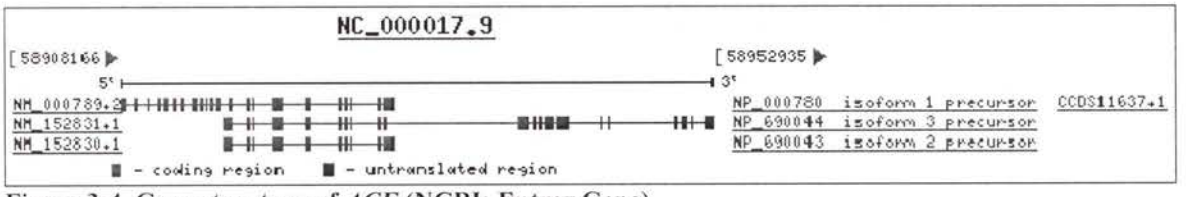


Figure 3-4 Gene structure of *ACE* (NCBI: Entrez Gene).

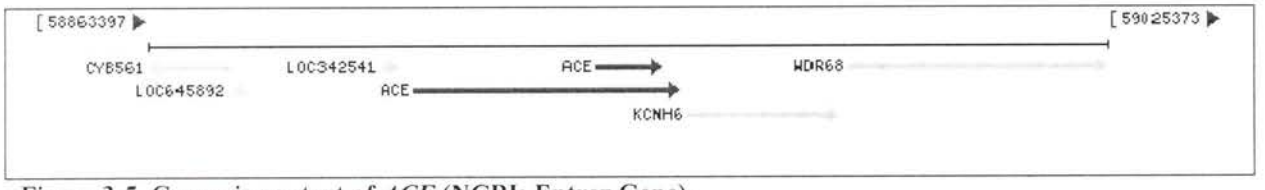


Figure 3-5 Genomic context of *ACE* (NCBI: Entrez Gene).

The protein sequence of ACE containing 1,306 amino acids is shown (Figure 3-6). The somatic form of ACE has two tandem active sites with separate catalytic properties. The function of testicular ACE is largely unknown and has only one active site. ACE also has a homolog, angiotensin I converting enzyme 2 (ACE2), in humans. ACE2 differs from ACE by preferentially removing carboxy-terminal hydrophobic or basic amino acids. ACE2 is considered to be important in cardiac function (Riordan 2003). Somatic ACE, a type I

transmembrane protein, is composed of two homologous catalytic domains (N and C domains), arising from a gene duplication event (Soubrier et al. 1988). The N and C domains have distinct physiological roles and possible negative cooperativity between them (Binevski et al. 2003).

```
        10          20          30          40          50          60
MGAASGRRGP GLLLPLPLLL LLPPQPALAL DPGLQPGNFS ADEAGAQLFA QSYNSSAEQV
        70          80          90         100         110         120
LFQSVAASWA HDTNITAENA RRQEEAALLS QEFAEAWGQK AKELYEPIWQ NFTDPQLRRI
       130         140         150         160         170         180
IGAVRTLGSA NLPLAKRQQY NALLSNMSRI YSTAKVCLPN KTATCWSLDP DLTNILASSR
       190         200         210         220         230         240
SYAMLLFAWE GWHNAAGIPL KPLYEDFTAL SNEAYKQDGF TDTGAYWRSW YNSPTFEDDL
       250         260         270         280         290         300
EHLYQQLEPL YLNLHAFVRR ALHRRYGDRY INLRGPIPAH LLGDMWAQSW ENIYDMVVPF
       310         320         330         340         350         360
PDKPNLDVTS TMLQQGWNAT HMFRVAEEFF TSLELSPMPP EFWEGSMLEK PADGREVVCH
       370         380         390         400         410         420
ASAWDFYNRK DFRIKQCTRV TMDQLSTVHH EMGHIQYYLQ YKDLPVSLRR GANPGFHEAI
       430         440         450         460         470         480
GDVLALSVST PEHLHKIGLL DRVTNDTESD INYLLKMALE KIAFLPFGYL VDQWRWGVFS
       490         500         510         520         530         540
GRTPPSRYNF DWWYLRTKYQ GICPPVTRNE THFDAGAKFH VPNVTPYIRY FVSFVLQFQF
       550         560         570         580         590         600
HEALCKEAGY EGPLHQCDIY RSTKAGAKLR KVLQAGSSRP WQEVLKDMVG LDALDAQPLL
       610         620         630         640         650         660
KYFQPVTQWL QEQNQQNGEV LGWPEYQWHP PLPDNYPEGI DLVTDEAEAS KFVEEYDRTS
       670         680         690         700         710         720
QVVWNEYAEA NWNYNTNITT ETSKILLQKN MQIANHTLKY GTQARKFDVN QLQNTTIKRI
       730         740         750         760         770         780
IKKVQDLERA ALPAQELEEY NKILLDMETT YSVATVCHPN GSCLQLEPDL TNVMATSRKY
       790         800         810         820         830         840
EDLLWAWEGW RDKAGRAILQ FYPKYVELIN QAARLNGYVD AGDSWRSMYE TPSLEQDLER
       850         860         870         880         890         900
LFQELQPLYL NLHAYVRRAL HRHYGAQHIN LEGPIPAHLL GNMWAQTWSN IYDLVVPFPS
       910         920         930         940         950         960
APSMDTTEAM LKQGWTPRRM FKEADDFFTS LGLLPVPPEF WNKSMLEKPT DGREVVCHAS
       970         980         990        1000        1010        1020
AWDFYNGKDF RIKQCTTVNL EDLVVAHHEM GHIQYFMQYK DLPVALREGA NPGFHEAIGD
      1030        1040        1050        1060        1070        1080
VLALSVSTPK HLHSLNLLSS EGGSDEHDIN FLMKMALDKI AFIPFSYLVD QWRWRVFDGS
      1090        1100        1110        1120        1130        1140
ITKENYNQEW WSLRLKYQGL CPPVPRTQGD FDPGAKFHIP SSVPYIRYFV SFIIQFQFHE
      1150        1160        1170        1180        1190        1200
ALCQAAGHTG PLHKCDIYQS KEAGQRLATA MKLGFSRPWP EAMQLITGQP NMSASAMLSY
      1210        1220        1230        1240        1250        1260
FKPLLDWLRT ENELHGEKLG WPQYNWTPNS ARSEGPLPDS GRVSFLGLDL DAQQARVGQW
      1270        1280        1290        1300
LLLFLGIALL VATLGLSQRL FSIRHRSLHR HSHGPQFGSE VELRHS
```

Figure 3-6 ACE protein sequence
(length: 1,306 amino acids, molecular weight: 149,715 Da)
(UniProtKB/Swiss-Prot: http://ca.expasy.org/uniprot/P12821).

The *ACE I/D* polymorphism was studied in elite athletes because it has been associated with LV hypertrophy (Danser et al. 1995; Diet et al. 2001; Doolan et al. 2004; Fatini et al. 2000; Hernandez et al. 2003; Montgomery et al. 1997; Nagashima et al. 2000; Rizzo et al. 2003) and elite male endurance athletes often show signs of athlete's heart, which is a physiological form of LV hypertrophy (described further on p100).

Elite athletes exhibit many electrocardiogram changes, including an increase of R or S wave voltage, either flat or inverted T waves, and deep Q waves, suggestive of structural cardiovascular disease, such as hypertrophic cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy, which are the most common causes of sudden death in young athletes (Pelliccia and Maron 2001).

Electrocardiograms of 1005 athletes from 38 different sports displayed abnormal electrocardiograms in 40%, but structural cardiac diseases in only 5%. In the absence of cardiac disease, factors responsible for abnormal electrocardiogram patterns were thought to include morphologic cardiac remodelling, participation in endurance sports, and male gender. Also, a small group of athletes showed electrocardiogram abnormalities that suggested cardiovascular disease in the absence of pathologic cardiac conditions or morphologic changes, which may have been due to the athletic training itself (Pelliccia and Maron 2001).

3.1.3 Physiology and Biochemistry

3.1.3.1 Plasma ACE Levels

The role of the intron 16 *I/D* polymorphism has been the focus of debate as a contributor to endurance performance in the last few years. The *ID* and *DD* genotypes correlate to an

almost one-and-a-half- and two-fold increase respectively in circulating angiotensin levels compared to *II* genotype (Rigat et al. 1990) (Table 3-1).

Table 3-1 *ACE* genotype and plasma ACE level (Rigat et al. 1990).

ACE Genotype	Serum immunoreactive ACE concentrations ($\mu g.L^{-1}$)
II	299 ± 49
ID	392 ± 67
DD	494 ± 88

The *ACE I/D* polymorphism explains 47% of the total phenotypic variance of serum ACE, showing that the *ACE* gene locus is the major locus that determines serum ACE concentration (Rigat et al. 1990; Rossi et al. 1999). Another study suggested that it was not directly responsible for the variation in serum ACE but was in linkage disequilibrium with a regulatory allele and accounted for 28% of the total variability in the ACE level (Tiret et al. 1992). There was a significant positive relationship between serum ACE activity and diastolic pressure. Serum ACE activity is related to the *ACE* gene *I/D* polymorphism in caucasian but not in black children and adolescents (Bloem et al. 1996). Other studies have consistently shown that there is lower plasma renin activity in Blacks than Caucasians (Kaplan et al. 1976; Price and Fisher 2003). Plasma renin activity, however, is not associated with *ACE I/D* in Caucasians (Rossi et al. 1999). Another study found that there was no significant association between *ACE* genotype and hypertension within Caucasians but that within Blacks there was an association between the frequency of the *D* allele and hypertension. There was no association between *ACE I/D* genotype and plasma renin activity in either group (Barley et al. 1996).

3.1.3.2 Bradykinin Metabolism

The *ACE D* allele is associated with improved degradation of bradykinin, a vasoprotective peptide. The *ACE* genotype determines bradykinin degradation and this could be another way

the *ACE D* allele negatively affects the cardiovascular system (Brown et al. 1998). A computer-based simulation study showed that a decrease in bradykinin plays an important role in the increased risk of diabetic nephropathy associated with genetically determined higher levels of ACE activity (Takahashi et al. 2003).

3.1.3.3 Disease States

Early studies of *ACE* and heart disease showed a significant relationship. A study of 213 autopsy cases demonstrated an increased frequency of the *ACE D* allele. The findings were consistent with the hypothesis that the *ACE I/D* polymorphism is a risk factor for fatal myocardial infarction and sudden cardiac death (Evans et al. 1994). Cardiac ACE activity was significantly higher in subjects with the *ACE DD* genotype compared with subjects with the *ID* and the *II* genotypes, of 71 subjects who died of noncardiac disorders. Elevated cardiac ACE activity in these subjects may result in increased cardiac angiotensin II levels, and this may be a mechanism underlying the reported association between the *ACE D* polymorphism and the increased risk for several cardiovascular disorders, such as hypertension, myocardial infarction and FHC (Danser et al. 1995). The early positive genetic association studies have not been replicated, often due to lack of statistical power (Sayed-Tabatabaei et al. 2006). It is also possible that the earlier studies were false positive associations.

The *ACE I/D* polymorphism is generally thought to play a modifying role as a susceptibility gene for various heart diseases (Bleumink et al. 2004). There was greater LV wall thickness in *MYBPC3*-hypertrophic cardiomyopathy patients with *DD* genotype compared with *ID* or *II* genotype, and only *ACE DD* had extreme hypertrophy. This is further evidence that ACE acts as a modifying gene for cardiomyopathy (Perkins et al. 2005).

3.1.3.4 Muscle Fibre Types

The proportion of slow-twitch, fatigue-resistant type I skeletal muscle fibres is often reduced in heart failure, while the proportion of fatigue-sensitive type II fibres increases. This maladaptation may be partially responsible for the exercise intolerance that characterises heart failure (Sabbah et al. 1996). The *ACE I* allele was associated with increased slow type I muscle fibres, which could explain the mechanism for the association between the *ACE* genotype and endurance athletes (Zhang et al. 2003).

3.1.4 Athlete Studies

3.1.4.1 Left Ventricular Hypertrophy

LV hypertrophy is common amongst male elite endurance athletes (Pelliccia and Maron 2001). Exercise-induced LV growth in young male caucasian military recruits appears to be strongly associated with the *ACE I/D* polymorphism (Montgomery et al. 1997). A study of 28 Italian elite male soccer players versus controls showed no difference for *ACE I/D* genotypes. Training-induced LV mass changes were, however, significantly associated with the *ACE D* allele (Fatini et al. 2000). LV end-diastolic diameter and LV mass were significantly greater in a group of 43 ultramarathon (i.e. >42 km) runners with *DD* and *ID* than in those with *II* genotype (Nagashima et al. 2000). An association of combined *ACE I/D* polymorphism genotypes, and angiotensinogen gene M235T polymorphism genotypes was found with LV hypertrophy following long-term athletic training in 83 male caucasian endurance athletes (Diet et al. 2001). The *ACE DD* genotype was associated with a higher LV mass index than the *ID* genotype performed in 61 male endurance athletes (age: 25–40 years), regardless of other confounder variables (Hernandez et al. 2003). LV hypertrophy was found in 17 out of

75 (23%) competitive adolescent soccer players. The *ACE I/D* polymorphism was associated with the level of cardiac hypertrophy but not its incidence (Rizzo et al. 2003).

A non-genetic study showed that 11 (2.5%) out of 442 (306 male, 136 female) elite British athletes from 13 sports including judo, skiing, cycling, triathlon, rugby and tennis, had a LV wall thickness >13 mm, i.e. equivalent to a diagnosis of hypertrophic cardiomyopathy. Systolic and diastolic function were normal for all athletes (Whyte et al. 2004). This was a much lower incidence of LV hypertrophy than the previous study. Perhaps the variety and types of sports involved were the reason. The majority of studies showed that the *ACE I/D* polymorphism was associated with LV hypertrophy and that significant, but reversible, LV hypertrophy is common in athletes.

3.1.4.2 Performance Level

There was a significantly increased proportion of the *ACE I* allele and the *ACE II* genotype in 64 Australian national-level rowers compared to normal controls. It was proposed that the mechanism explaining this association was that the *ACE I* allele and *II* genotype was related to a healthier cardiovascular system (Gayagay et al. 1998). In another study, 120 Australian national-level caucasian athletes in sports thought to require high aerobic fitness including team/ball sports, there was no difference in *ACE* genotype frequencies between athletes and controls (Taylor et al. 1999). The mixing of sports which were physiologically different (power and/or endurance) could have compromised the results.

The *ACE I* allele frequency was significantly higher in 60 professional athletes (25 cyclists, 20 distance runners, and 15 handballers) compared to 400 healthy controls. Plasma ACE levels demonstrated a high correlation with *I/D* genotype (Alvarez et al. 2000). In 447

caucasian male triathletes from the South African Ironman Triathlons and 199 caucasian male control subjects, the *I* allele of the *ACE* gene was associated with the endurance performance of the fastest 100 South African-born finishers in these triathlons (Collins et al. 2004). A study of 80 healthy Turkish athletes and 80 healthy sedentary controls who were genotyped for the *ACE I/D* polymorphism showed that there was a significant difference between athletes and controls (Turgut et al. 2004). Overall, the various research studies suggest that the *ACE I/D* polymorphism is inconsistently associated with athletic performance level.

3.1.4.3 Event Duration

A study of potential British Olympic Association athletes showed a significant linear trend for increased *I* allele frequency with distance run. This study, though, mixed males and females, and grouped runners in an unusual fashion – mixing 400 m (long sprint) with 3000 m (middle distance endurance) athletes (Myerson et al. 1999). Of the 103 (57 male, 46 female) caucasian swimmers from the European and Commonwealth championships and an American college team, there was a significant excess of the *ACE D* allele compared with the control group only in the elite swimmers of the European and Commonwealth championships. This association remained in those competing over shorter distances but not in the longer events. The researchers, therefore, suggested that a genetic association study of elite athletes requires a homogeneous cohort of subjects not only from the same sport, but also the same event type, i.e. sprint or endurance (Woods et al. 2001). Their negative results in some of their control groups and use of groups of varying mixed and non-mixed gender, on the other hand, make it difficult to be sure of the reliability of their result.

ACE I/D allele frequency was tested amongst 217 'outstanding' and 'average' Russian athletes (swimmers, skiers, triathletes and runners) with different event durations: short (<1

min), middle (1–20 min), and long (>20 min) distance athletes. There was no association for the *ACE* genotype for the whole cohort or for the group of 'outstanding' athletes. The *D* allele was significant for the 'outstanding' short distance athletes and the *I* allele was significant for the 'outstanding' middle distance athletes. No association was found for the 'outstanding' long distance athletes (Nazarov et al. 2001). Most research studies would suggest that the *ACE I/D* polymorphism is weakly associated with the duration of an athletic event.

One study of 63 male caucasian endurance athletes following natural exposure to moderate altitude (2,200 m) showed that the *ACE I/D* polymorphism did not influence the time course of the erythropoietic response (*DD*, 31 (49%); *ID*, 24 (38%); *II*, 8 (13%)) (Gonzalez et al. 2006). The distribution of genotypes is quite different from most studies of endurance athletes and from most matched control groups.

Baseline and change in $\dot{V}O_2$ max, body mass index, skinfold thickness, and serum lipids did not differ by *ACE* genotype for 110 subjects (14 *II*, 52 *ID*, and 44 *DD*), but adherence to exercise training was higher in *II* and *ID* than in *DD* subjects (Thompson et al. 2006).

3.1.4.4 Maximum Oxygen Uptake

The association between the *ACE I/D* polymorphism and fitness phenotypes was measured before and after 20 weeks of a standardised endurance training program in 476 sedentary caucasian and 248 black subjects. Only 11 out of 216 comparisons showed significant associations with the *ACE I/D* polymorphism. Unexpectedly, *DD* caucasian offspring showed a 14–38% greater training response for $\dot{V}O_2$ max, and various other fitness phenotypes, than did *II* (Rankinen et al. 2000a). The *ACE ID* polymorphism was again studied in 192 male

endurance athletes with $\dot{V}O_2$ max $\geq$ 75 ml.kg^{-1}.min^{-1} and 189 sedentary male controls from the GENATHLETE cohort. Both the genotype and allele frequencies were similar in the athletes and the controls (Rankinen et al. 2000b). One of the explanations for their differing result from (Gayagay et al. 1998) was that rowing mainly used the upper body muscles. This is, however, incorrect. Rowing is a whole body exercise, which places a heavy burden on the legs, particularly in the beginning of the catch phase of the stroke. A review paper concluded that an association seems likely and that it is probably due to a local muscle effect rather than a central cardiorespiratory mechanism (Woods et al. 2000). Research studies involving associations between the *ACE I/D* polymorphism and $\dot{V}O_2$ max remain contradictory.

3.2 Materials and Methods

3.2.1 Subjects

Elite male athletes included rowers (n = 108), cyclists (n = 57), swimmers (n = 29), endurance runners (n = 20) and sprint runners (n = 21) from the AIS. Elite (n = 107) and non-elite (n = 140) rugby players were from various amateur and professional rugby clubs in New South Wales and the Australian Capital Territory. Elite level was defined as competing at the junior or senior national, international or professional level (defined on p79). Rugby players were also subdivided into Front-five: #1–5 and Back-ten: #6–15 groups.

3.2.2 Materials

3.2.2.1 Oligonucleotides

The oligonucleotides used for this chapter were those designed and described elsewhere for *ACE I/D* (Evans et al. 1994). They were ordered from BioLabs™ (Table 3-2).

Table 3-2 *ACE I/D* Oligonucleotides.

Primer	Primer Sequence
ACE1	5′-CAT CCT TTC TCC CAT TTC TC-3′
ACE2	5′-TGG GAT TAC AGG CGT GAT ACA G-3′
ACE3	5′-ATT TCA GAG CTG GAA TAA AAT T-3′

3.2.3 Methods

3.2.3.1 PCR Amplification

The *ACE I/D* genotyping was performed using PCR. The 3-primer PCR was used to unambiguously distinguish between *II*, *ID* and *DD* genotypes (Evans et al. 1994) (Table 3-3).

Table 3-3 *ACE I/D* PCR conditions.

Primer1: ACE1 (20 pmol. μL^{-1})	0.5 µL
Primer2: ACE2 (20 pmol.μL^{-1})	0.25 µL
Primer3: ACE3 (20 pmol.μL^{-1})	0.5 µL
dNTP (2.5 mM)	1.0 µL
GeneAmpR 10× PCR (buffer II)	2.5 µL
GeneAmpR $MgCl_2$ (25 mM stock)	2.0 µL
dH_2O	16.15 µL
AmpliTaqR (5 U.μL^{-1})	0.1 µL
DNA (25 ng.μL^{-1})	2 µL
Total volume	25 µL
Thermal Cycling Conditions * product size: *I*: 65 bp; *D*: 84 bp	94 °C × 2 min × 1 (94 °C × 30 s; 55 °C × 30 s) × 30 72 °C × 7 min × 1

3.2.3.2 Statistical Analysis

3.2.3.2.1 SigmaStat V1

The χ^2 test is the most common test used for genetic data using SigmaStat V1™ (Jandel Scientific Software, California) program.

3.2.3.2.2 CLUMP Program

The CLUMP program was used for the χ^2 test for data where there were small allele or genotype frequencies. The CLUMP program was run through the DOS program and uses Monte Carlo simulation to perform the χ^2 test.

3.2.3.2.3 SPSS Regression Modelling

The SPSS™ program was used for the regression modelling analysis. Multiple regression analysis was used to compare genotype to phenotype data. Kevin McGeechan (Associate Lecturer, Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, University of Sydney) assisted with the preparation of the statistical results using the SPSS™ program.

3.3 Results

3.3.1 Genetic results

The genetic results showed small differences between the various male athlete groups and the age-matched controls (Table 3-4). There were statistically significant results for subgroups of the male rowers, cyclists, runners and rugby players. The full rowers group from 1996–2003 was not significantly different from controls. The subgroup of Olympic rowers was, on the other hand, highly significantly different (genotype $p = 0.004$; allele $p = 0.002$). The full cyclists group did not produce statistically significant results, although the subgroup of track cyclists did (allele $p = 0.033$). The full runners group did not produce statistically significant results, although the subgroup of endurance runners did (genotype $p = 0.046$; allele $p = 0.028$). The full group of rugby players did not produce statistically significant results, although the subgroup of elite rugby players (genotype $p = 0.020$) and Back-ten (genotype $p = 0.027$) did. The comparison of subgroups of Front-five vs Back-ten rugby players was statistically significant (genotype $p = 0.044$). The comparison of subgroups of elite vs non-elite rugby players was not statistically significant (genotype $p = 0.066$).

The male Ironman (range: genotype $p = 0.165$–0.831; allele $p = 0.327$–0.706) and male swimmers (range: genotype $p = 0.467$–0.634; allele $p = 0.476$–0.698) did not produce any statistically significant results.

Table 3-4 Male *ACE I/D* Results

(p values versus controls unless otherwise specified) (comparisons were age-matched).

Sport	Controls									
Sex	M		M							
Age	All		<40 yr							
Race	Caucasian		Caucasian							
II	60	0.21	48	0.21						
ID	143	0.50	110	0.49						
DD	81	0.29	68	0.30						
I	263	0.46	206	0.46						
D	305	0.54	246	0.54						
n	284		226							
Sport	Ironman		Ironman		Ironman		Ironman		Triathlon:AIS	
Sex	M		M		M		M		M	
Age	All		<40 yr		All		All		<40 yr	
Race	Caucasian		Caucasian		Caucasian		Caucasian		Caucasian	
Group	Ironman & Half-Ironman		Ironman & Half-Ironman		Top 25% Ironmen		Bottom 25% Ironman			
II	187	0.24	124	0.24	30	0.24	27	0.22	7	0.33
ID	380	0.48	257	0.50	61	0.48	51	0.41	9	0.43
DD	220	0.28	137	0.26	35	0.28	46	0.37	5	0.24
I	754	0.48	505	0.49	121	0.48	105	0.42	23	0.55
D	820	0.52	531	0.51	131	0.52	143	0.58	19	0.45
n	787		518		126		124		21	
genotype p	0.659		0.527		0.831		0.165		0.437	
allele p	0.545		0.285		0.706		0.332		0.327	
Sport	Rowers		Rowers							
Sex	M		M							
Age	<40 yr		<40 yr							
Race	Caucasian		Caucasian							
Group	All		All Olympians							
II	26	0.24	10	0.53						
ID	54	0.50	8	0.42						
DD	28	0.26	1	0.05						
I	106	0.49	28	0.74						
D	110	0.51	10	0.26						
n	108		19							
genotype p	0.694		0.004							
allele p	0.444		0.002							
Sport	Swimmers		Swimmers		Swimmers		Swimmers			
Sex	M		M		M		M			
Age	<40 yr		<40 yr		<40 yr		<40 yr			
Race	Caucasian		Caucasian		Caucasian		Caucasian			
Group	All distances		Sprint: 50/100 m		Mid-distance: 100/200 m		Long distance: 400+ m			
II	4	0.14	0	0.00	4	0.19	0	0.00		
ID	15	0.52	3	0.75	13	0.62	2	0.33		
DD	10	0.34	1	0.25	4	0.19	4	0.67		
I	23	0.40	3	0.38	21	0.50	2	0.17		
D	35	0.60	5	0.63	21	0.50	10	0.83		
n	29		4		21		6			
genotype p	0.634		n/a		0.467		n/a			
allele p	0.476		n/a		0.698		n/a			
Sport	Cyclists		Cyclists		Cyclists		Cyclists			
Sex	M		M		M		M			
Age	<40 yr		<40 yr		<40 yr		<40 yr			
Race	Caucasian		Caucasian		Caucasian		Caucasian			
Group	Track: Sprint/Endurance		Road: Endurance		All		All Endurance			
II	8	0.36	8	0.25	16	0.28	13	0.26		
ID	12	0.55	15	0.47	29	0.51	25	0.50		
DD	2	0.09	9	0.28	12	0.21	12	0.24		
I	28	0.64	31	0.48	61	0.54	51	0.51		
D	16	0.36	33	0.52	53	0.46	49	0.49		
n	22		32		57		50			
genotype p	0.071		0.888		0.317		0.619			
allele p	0.033		0.767		0.158		0.382			
Sport	Runners		Runners		Runners					
Sex	M		M		M					
Age	<40 yr		<40 yr		<40 yr					
Race	Caucasian		Caucasian		Caucasian					
Group	All		Sprint: 100–400 m		Endurance:800 m–42 km					
II	9	0.22	2	0.10	7	0.35				
ID	24	0.59	12	0.57	12	0.60				
DD	8	0.20	7	0.33	1	0.05				
I	42	0.51	16	0.38	26	0.65				
D	40	0.49	26	0.62	14	0.35				
n=	41		21		20					
genotype p	0.360		0.439		0.046					
allele p	0.411		0.441		0.028					
Sport	Rugby		Rugby		Rugby		Rugby		Rugby	
Sex	M		M		M		M		M	
Age	<40 yr		<40 yr		<40 yr		<40 yr		<40 yr	
Race	Caucasian		Caucasian		Caucasian		Caucasian		Caucasian	
Group	All		Elite		Non-Elite		Front-five		Back-ten	
II	44	0.18	13	0.12	31	0.22	21	0.24	23	0.15
ID	141	0.57	69	0.64	72	0.51	40	0.47	97	0.63
DD	62	0.25	25	0.23	37	0.26	25	0.29	35	0.23
I	229	0.46	95	0.44	134	0.48	82	0.48	143	0.46
D	265	0.54	119	0.56	146	0.52	90	0.52	167	0.54
n	247		107		140		86		155	
genotype p	0.187		0.020		0.753		0.832		0.027	
allele p	0.861		0.839		0.599		0.704		0.939	

* Rugby: Front-five versus Back-ten (genotype $p = 0.044$; allele $p = 0.818$); Elite versus Non-Elite (genotype $p = 0.066$; allele $p = 0.500$). Definitions on p105. Comparisons to all-age control group are shaded. Significant p values are shaded (p values are not Bonferroni-corrected).

3.3.2 Genetic and Physiological Results

Multiple regression analysis was performed to determine the significant predictors of performance for all the athletes combined because no specific order of importance was predicted by the literature on exercise physiology.

3.3.2.1 Maximal Oxygen Uptake

Weight, elite and sport were the statistically significant predictors of $\dot{V}O_2$ max ($p < 0.001$) and approximately 24% of the variance in $\dot{V}O_2$ max was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of $\dot{V}O_2$ max when added to the model ($p = 0.928$) (Table 3-5).

3.3.2.2 Two Kilometre Row Time

Age, weight and Olympian were the statistically significant predictors of 2 km Row Time ($p < 0.001$) and approximately 78% of the variance in 2 km Row Time was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of 2 km Row Time when added to the model ($p = 0.608$) (Table 3-5).

3.3.2.3 Ironman Time

Age was the only statistically significant predictor of Ironman Time ($p < 0.001$) and approximately 16% of the variance in Ironman Time was accounted for by this variable. *ACE I/D* was not a significant predictor of Ironman Time when added to the model ($p = 0.226$) (Table 3-5).

3.3.2.4 Fitness Z-score

The above three sports-specific measures of aerobic fitness were combined in a Fitness Z-score for analysis. Weight, Olympian, elite and sport were the statistically significant predictors of Fitness Z-score ($p < 0.001$) and approximately 23% of the variance in Fitness Z-score was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of Fitness Z-score when added to the model ($p = 0.440$) (Table 3-5).

3.3.2.5 40 m Sprint Time

Age, weight and elite were the statistically significant predictors of 40 m Sprint Time in rugby players ($p < 0.001$) and approximately 28% of the variance in 40 m Sprint Time was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of 40 m Sprint Time when added to the model ($p = 0.057$) (Table 3-5).

Table 3-5 Summary of multiple regression ANOVA related to performance (statistical results from SPSS™ program).

Outcome variable	Explanatory variables	Coefficients B	*p* value	r^2	genotype *p* value
$\dot{V}O_2$ max	weight	-0.026	< 0.001	24.5%	0.928
	elite	0.469			
	sport	n/a			
2 km Row Time	age	0.019	< 0.001	78.5%	0.608
	weight	0.037			
	Olympian	0.169			
Ironman Time	age	-0.046	< 0.001	15.5%	0.226
Fitness Z-score	weight	-0.019	< 0.001	22.6%	0.440
	Olympian	0.791			
	elite	0.445			
	sport	n/a			
40 m Sprint Time	age	-0.089	< 0.001	27.7%	0.057
	weight	-0.016			
	elite	0.732			

3.4 Discussion

3.4.1 Discussion of Genetic Results

The genetic results showed small differences between the various male athlete groups and the age-matched controls (Table 3-4). There were statistically significant results for subgroups of the male rowers, cyclists, runners and rugby players. The full rowers group from 1996–2003 was not significantly different from controls. The subgroup of Olympic rowers was highly significantly different (genotype $p = 0.004$; allele $p = 0.002$). This result confirms previous results (Gayagay et al. 1998) which found an association between the *ACE* gene *I/D* polymorphism and Olympic rowers because although the full rowers group results were not significant, many of these rowers were of a lesser standard than the Olympic rowers. This highlights the difficulty of selecting only elite athletes for study.

Another difficulty is selecting a large enough group of elite athletes to achieve statistically significant results. There are only a limited number of elite athletes in any country, appropriate age-matched control groups are hard to find and generalisability of results is not obvious (Sands et al. 2005). If the criteria for being elite are broadened, then there is a risk that the effect size being sought is too small to be found. If additional subjects are sought from overseas, that introduces the problem of finding matching controls from overseas as well.

Another explanation for the larger groups not being significant and the smaller groups being significant is that the smaller numbers may have produced spurious results.

The full cyclists group did not produce statistically significant results, but the subgroup of track cyclists did (allele $p = 0.033$) with an increased proportion of the *I* allele. The full group result contrasts with a previous study (Alvarez et al. 2000) which showed that the *ACE I* allele was associated with a group of elite athletes that predominantly included cyclists, but the track cyclists subgroup result confirms the previous study. Another study showed that the *D* allele and *DD* genotype were more common in cyclists, than other athletes or controls (Lucia et al. 2005).

The full runners group did not produce statistically significant results, although the subgroup of endurance runners did (genotype $p = 0.046$; allele $p = 0.028$). This supports a previous study (Nazarov et al. 2001) which showed an excess of the *D* allele in elite short distance athletes and an excess of the *I* allele in elite middle distance athletes.

The full group of rugby players did not produce statistically significant results, although the subgroup of elite rugby players (genotype $p = 0.020$) and Back-ten (genotype $p = 0.027$) did. The comparison of subgroups of Front-five vs Back-ten rugby players was statistically significant (genotype $p = 0.044$). The comparison of subgroups of elite vs non-elite rugby players was not statistically significant (genotype $p = 0.066$). This is the first time that this group of athletes has been investigated for the *ACE* gene. The consistent trend was for increased *ID* genotype at the expense of the other two genotypes (Elite: *II* = 12%, *ID* = 64%, *DD* = 23%; Back-ten: *II* = 15%, *ID* = 63%, *DD* = 23%). These results were unexpected. It was expected that the genotypes would be skewed towards more *DD* genotypes and *D* alleles because rugby players are considered to be power/sprint type athletes. These results might indicate that the mixed genotype was beneficial and that perhaps all-round sprint and endurance characteristics are required. Another study described how elite ice hockey players showed the increased ventricular size and wall thickness characteristic of combined

power/endurance sports (Bossone et al. 2004). Perhaps the greater level of cardiac hypertrophy associated with the *DD* genotype in many studies is a disadvantage in combined power/endurance sports such as rugby. As may have been the situation with the previous small group significant results, these results might also reflect an artefact.

The male Ironman (range: genotype p = 0.165–0.831; allele p = 0.327–0.706) and male swimmers (range: genotype p = 0.467–0.634; allele p = 0.476–0.698) did not produce any statistically significant results. The lack of an association between the *ACE* gene and Ironman triathletes contrasts with a previous study with triathletes (Collins et al. 2004; Nazarov et al. 2001) where there was an association between high performance level and *I* allele. It confirms, nevertheless, the results of another study (Nazarov et al. 2001) where *ACE* was not associated with the triathlon. The lack of an association between the *ACE* gene and swimmers contrasts with the results of previous studies (Nazarov et al. 2001; Tsianos et al. 2004; Woods et al. 2001) which showed an excess of the *D* allele in elite swimmers as well as in elite sprint swimmers. The lack of association for *ACE* and swimming, however, may be explained by the fact that $\dot{V}O_2$ max is not considered a limiting factor in swimming and that elite swimmers can reach higher levels in running tests than in swimming tests (Holmer 1992).

Overall, there was a pattern of full athlete groups results being not significant, with subgroups being significant. When Bonferroni correction is applied to the Table 3-4, the significant results become not significant. When many statistical comparisons are performed, it would result in some subgroups being significant. Perhaps there is merit, though, in making the phenotype narrower in an effort to find a larger effect size. Also, there are reasons to not do Bonferroni correction. Since the present study involves genetic testing of a new cohort of athletes, it could be considered hypothesis generating and Bonferroni correction would,

therefore, not be warranted. Further investigation of these results with larger groups of carefully selected elite athletes would be worthwhile.

3.4.2 Discussion of Genetic and Physiological Results

This section of the study explored the relationship between the *ACE I/D* genotype and various physical, physiological and performance variables. Multiple regression analysis yielded one-, three- and four-variable models, composed of age, weight, Olympian, elite and/or sport which accounted for 16–78% of the variance in performance variables (Table 3-5). Surprisingly, the addition of the *ACE I/D* genotype to the models added no significant prediction to the regression models. Perhaps the effect size is too small to be seen compared to the other predictive physical/phenotype variables.

Weight, elite and sport were the statistically significant predictors of $\dot{V}O_2$ max (z-score) ($p < 0.001$) and approximately 24% of the variance in $\dot{V}O_2$ max was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of $\dot{V}O_2$ max when added to the model ($p = 0.928$). It was surprising that the *ACE I/D* genotype was not a predictor of $\dot{V}O_2$ max. It was expected that weight (negative coefficient), elite (positive coefficient) and sport would be predictors of $\dot{V}O_2$ max. The sports which had $\dot{V}O_2$ max data were a mixture of weightbearing (running, rugby and shortcourse triathlon) and non-weightbearing (rowing and cycling) sports, and elite (rugby) and non-elite athletes (rugby), where these variables would be important.

Age, weight and Olympian were the statistically significant predictors of 2 km Row Time (z-score) ($p < 0.001$) and approximately 78% of the variance in 2 km Row Time was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of 2

km Row Time when added to the model ($p = 0.608$). Age (positive coefficient) would have been significantly related to 2 km Row Time because rowing had a large weight range. Weight (positive coefficient) was significant in the model because heavier rowers have an advantage in non-weightbearing sports. Rowing had large numbers of Olympic and non-Olympic athletes and this accounts for Olympian (positive coefficient) achieving significance in the model.

Age was the only statistically significant predictor of Ironman Time (z-score) ($p < 0.001$) and approximately 16% of the variance in Ironman Time was accounted for by this variable. *ACE I/D* was not a significant predictor of Ironman Time when added to the model ($p = 0.226$). The Ironman sample of athletes did not have much phenotype data, such as height, weight, Olympian or elite, and hence the model only explained 16% of the variance in Ironman Time. The athlete sample had a very large age range and that was why age (negative coefficient) was significantly associated with Ironman Time.

The above three sports-specific measures of aerobic fitness were combined in a Fitness Z-score (combined z-scores) for analysis. Weight, Olympian, elite and sport were the statistically significant predictors of Fitness Z-score ($p < 0.001$) and approximately 23% of the variance in Fitness Z-score was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of Fitness Z-score when added to the model ($p = 0.440$). The Fitness Z-score model encompassed most of the sports (rowing, cycling and rugby) and, therefore, weight (negative coefficient), Olympian (positive coefficient), elite (positive coefficient) and sport were all significant in the model.

Age, weight and elite were the statistically significant predictors of 40 m Sprint Time (z-score) in the regression model which only included rugby players ($p < 0.001$) and

approximately 28% of the variance in 40 m Sprint Time was accounted for by the linear combination of these variables. *ACE I/D* was not a significant predictor of 40 m Sprint Time when added to the model ($p = 0.057$). The effect of age (negative coefficient) may have reflected the high injury rate in rugby affecting older players more than younger players. This idea is supported by a study which found that hamstring injuries associated with sprinting in Australian Rules football players, were more common in players older than 23 years or with poor quadriceps flexibility (Hagel 2005). The effect of weight (negative coefficient) would be due to heavier players requiring more force to accelerate their heavier weight over the 40 m distance (from physics: force = mass × acceleration). The effect of elite was significant because there were large numbers of elite and non-elite athletes athletes in the rugby sample.

Age was a variable that appeared to be related to performance in some sports. One study showed that age was negatively correlated to *ACE* levels (Cambien 1988). This may have been a confounding factor in the present study especially with the older athletes (i.e. Ironman) because it may have reduced the effect of the genotype.

The lack of significance of the *ACE I/D* genotype to the model may have been due to group sizes being too small or due to the incomplete physiological data set reducing the statistical power. It also could have been because whatever might have been contributing to a difference in the genotype distributions between some of the athlete groups and controls was not measured by the standard physiological tests used in the various sports.

Some endurance sports require extreme levels of physical training which may have compromised the athletes' immune system (Gleeson et al. 1995; Gleeson et al. 2004) and this was not part of the physiological data set used for comparison used in this study.

Rugby, which is the most popular worldwide team contact sport involving collision, has one of the highest levels of injury of all team sports (Nicholl et al. 1995). One study showed that at the international level the injury rate was one injury per player every 3.44 matches, with an injury defined as requiring at least 24 h of restricted activity. The highest frequency of match injuries occurred in contact situations such as being tackled, rucks and mauls (Brooks et al. 2005). The sport has a high level of body contact which may require a high level of musculoskeletal robustness to minimise injury. This factor may not have been adequately measured in the physiological data set. Although height and weight were included in the data set, it may have been more enlightening to have taken further anthropometric measurements to produce somatotypes (i.e. endomorphy = fatness; mesomorphy = musculoskeletal robustness; and ectomorph = linearity) or simply skinfold %bodyfat tests for each player to be included in the statistical analysis. This may have been unrealistic due to the time constraints of obtaining samples in the amateur sport setting. Rugby players who are injury-prone are unlikely to continue playing at the senior level or to reach the elite level.

There are many other factors that influence elite athletic performance and their phenotypes were not measured here. Examples would include: big hands and feet for swimming (greater and more efficient propulsion); long arms for rowing (longer and more efficient stroke); big legs for cycling (greater power and cycling motion means that moment of inertia is irrelevant); thin legs for endurance running (decreased energy cost); and thin lower legs for sprint running (lower moment of inertia). Overall, this study of the *ACE* gene *I/D* polymorphism showed that it appeared to be associated with some sports but the underlying mechanism remains unclear.

This study adds to the rather large body of research into the association between the *ACE* gene *I/D* polymorphism and athletic performance. It does this because it is in the minority of

studies which separated the genders for analysis and had enough elite athletes to perform studies with many different types of athletes within a wide variety of sports, although not enough athletes to achieve highly statistically significant results. It showed that even small statistical significance can only be achieved when athlete groups are separated on the basis of duration and intensity of their event.

Chapter 4
Cardiac Myosin Binding Protein C (*MYBPC3*)

4.1 Introduction

The "athlete's heart" is a reversible cardiac hypertrophy which occurs in some athletes in response to prolonged endurance training. Familial hypertrophic cardiomyopathy (FHC) is an autosomal dominant disorder that is also characterised by LV hypertrophy but without haemodynamic overload. FHC is caused by mutations in several sarcomere genes including the cardiac myosin binding protein C (*MYBPC3*) gene on chromosome 11p11.2. FHC has similar characteristics (pronounced cardiac hypertrophy) to, as well as distinct differences (physiological versus pathological hypertrophy) from, "athlete's heart" cardiac hypertrophy. This similarity was one reason for investigating some nonsynonymous SNP variants in athletes and normal controls. Other reasons included the fact that cardiac myosin binding protein C is a major contributor to thick filament structure and regulation, and that *MYBPC3* gene knockout mice demonstrate that without MYBPC3 there is impaired contractile function and hypertrophy. These particular SNP variants were chosen because they are not associated with the aetiology of FHC. The present study investigated whether *MYBPC3* has a modifying effect on cardiac function by using an elite athlete model. For this work, the distributions of three nonsynonymous changes (V158M, V189I and S236G) in the *MYBPC3* gene were studied in elite athletes and normal controls.

In e6, the nonsynonymous SNP variant – 5190A→G; S236G was genotyped. The male sprinters and rugby players gave the most interesting results. The frequency of the *AG* genotype at 236 was significantly higher in male sprinters than 147 controls ($p = 0.040$), although the numbers were quite small. An increased frequency of the 236G allele and *AG* genotype was also found in Back-ten rugby players versus controls ($p = 0.044$, $p = 0.021$, respectively) and versus the Front-five rugby players ($p = 0.032$, $p = 0.049$, respectively).

Since rugby players gave the most interesting results with the largest numbers, fewer of the other athletes were genotyped for the e4 (V158M) variant because of a lack of promising results and concerns over a lack of statistical power. None of the groups gave statistically significant results. None of the other athletes were genotyped for the e5 (V189I) variant because even the large group of rugby players seemed to vary little from the controls. None of the groups gave statistically significant results.

Multiple regression analysis was performed to determine the significant predictors of performance for only the rugby players for e6 S236G variant because they gave the only statistically significant results and with enough subjects to achieve sufficient statistical power. Elite, height and weight were the statistically significant predictors of $\dot{V}O_2$ max ($p < 0.001$). Age, weight and elite were the statistically significant predictors of 40 m Sprint Time in rugby players ($p < 0.001$). *MYBPC3* e6 S236G variant was not a significant predictor of either performance variable when added to the models.

The genotyping described in this chapter was predominantly performed by the author. Some genotyping of rugby samples was performed by Edna Soriano (laboratory technician).

4.1.1 Anatomy and Structure

Myosin binding protein C (MYBPC) is a large sarcomeric protein found in striated muscle. It has multiple domains and is part of the intracellular immunoglobulin superfamily. Its role in the sarcomere remains to be fully explained. Its cardiac isoform accounts for 2% of the myofibrillar protein of the heart (Oakley et al. 2004) (Figure 4-1).

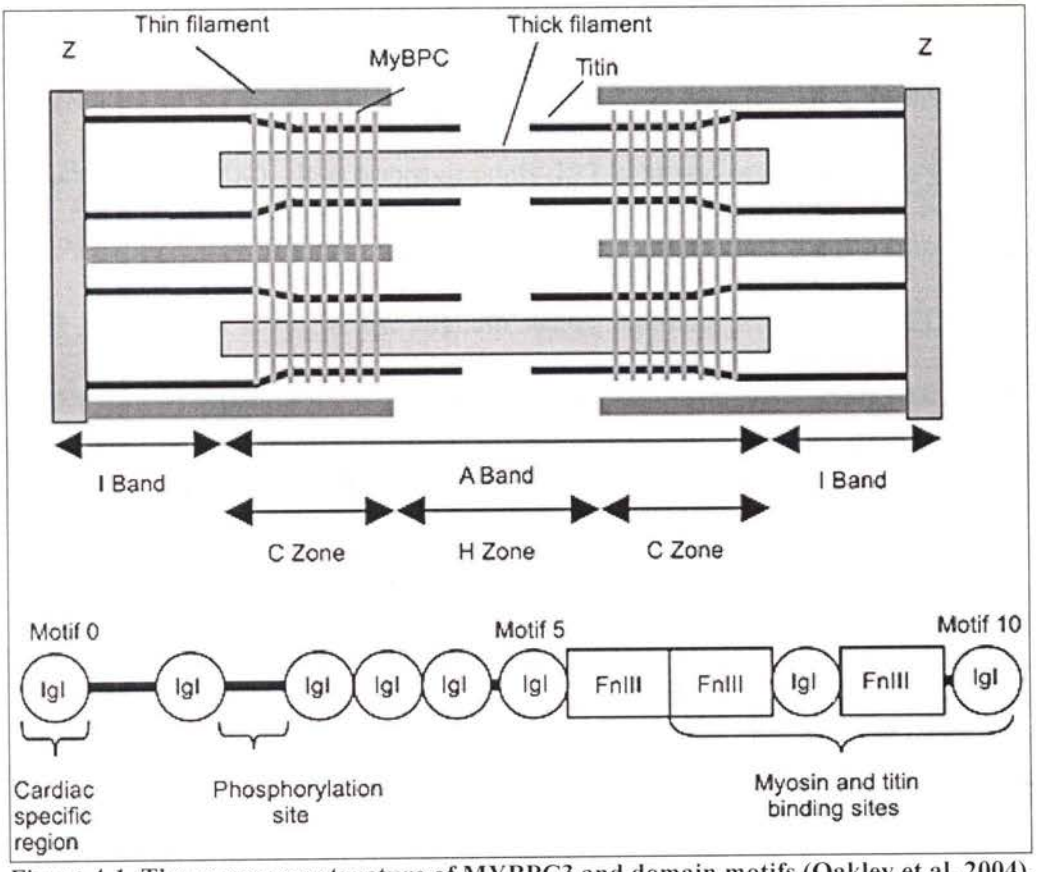


Figure 4-1 The sarcomere structure of MYBPC3 and domain motifs (Oakley et al. 2004).

Skeletal MYBPC has ten sub-domains (C1 to C10). There are seven are I-class immunoglobulin-type (IgI) domains and three (C6, C7 and C9) are fibronectin-like type III (FnIII) domains. In addition to these domains, cardiac MYBPC has an extra N-terminal domain (C0) and various insertions including phosphorylation sites. These are numbered Motifs 0 to 10 from the N-terminus to the C-terminus. It is thought that MYBPC3 has both structural and regulatory roles within the sarcomere. It is not known if the MYBPC3 actually binds to the myosin filament. MYBPC3 may increase the periodicity to about 435 Å, longer than the basic myosin filament repeat of 429 Å, as indicated by ultrastructural studies. The presence of this longer periodicity, as indicated by modelling, could be explained if the myosin-binding part of MYBPC3 binds to myosin with the expected 429 Å repeat. But this would only make sense if the N-terminal end of MYBPC3 interacts with the adjacent actin filaments in the hexagonal lattice of filaments in the A-band, in certain muscle states. Skeletal MYBPC contains a potential actin-binding domain in the Pro-Ala-rich sequence at

the N terminal region of skeletal MYBPC and in cardiac MYBPC between domains C0 and C1 (Squire et al. 2003). Interactions between adjacent MYBPC molecules have been noted (Oakley et al. 2004). It is approximately 137 kDa and is found in the C-zone sub-sections of the A-band. MYBPC is present in seven to nine out of the 11 transverse C-zone stripes. Only every third level of the myosin heads interacts with a MYBPC molecule due to the stripes 43 nm separation. Only the proportion of the myosin heads that are within the C-zone can interact directly with MYBPC.

4.1.1.1 Isoforms of Myosin Binding Protein

Three isoforms of MYBPC are known in adult muscle: fast skeletal, slow skeletal (originally MYBPX), and cardiac. Separate genes encode each isoform. The genes for the human fast (*MYBPC2*) and slow skeletal (*MYBPC1*) isoforms are on chromosomes 19q13.33 and 12q23.3, respectively, and the gene for human cardiac MYBPC (*MYBPC3*) is on chromosome 11p11.2 (Figure 4-2 and Figure 4-3). The mapping to different chromosomes indicates that the isoforms are not due to alternative splicing. The fast and slow skeletal isoforms can be found within the same sarcomere. The cardiac isoform was found during cardiac tissue phosphorylation studies (Flashman et al. 2004; Oakley et al. 2004). There is a smaller, related protein in skeletal muscle called myosin binding protein-H (MYBPH). It was discovered in the separation of myosin binding proteins. It is located in the third stripe in the C zone. It has high homology to four of the C-terminal domains of MYBPC, with 50% identity and 17% conserved amino acids (Flashman et al. 2004).

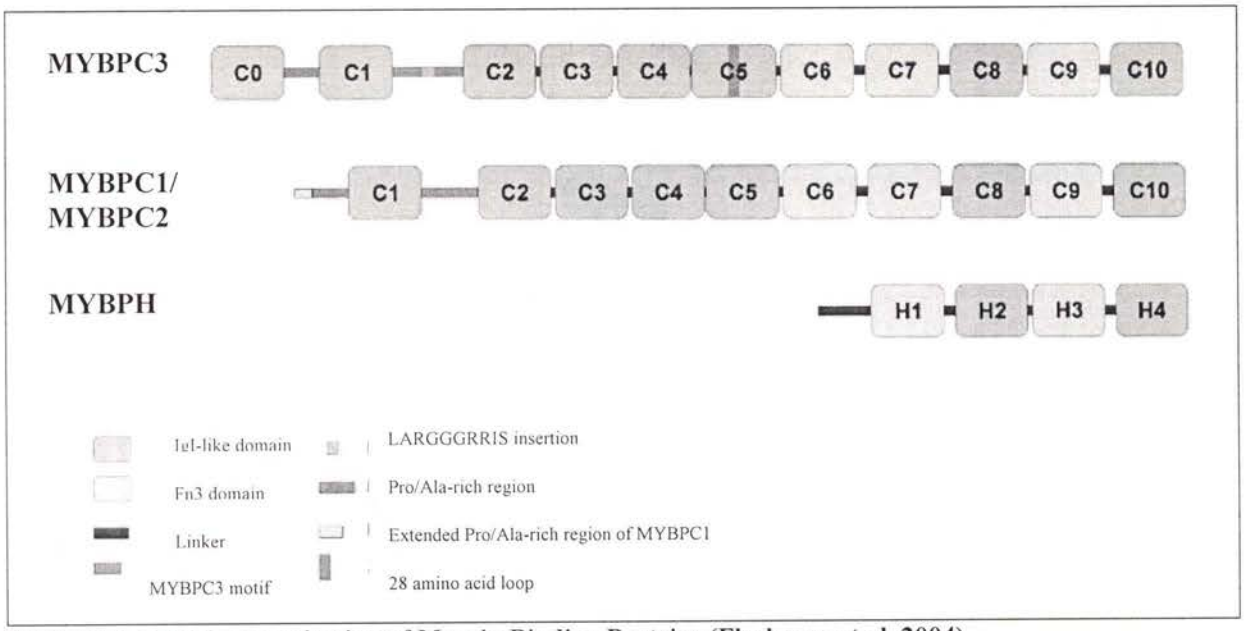


Figure 4-2 Domain organisation of Myosin Binding Proteins (Flashman et al. 2004).

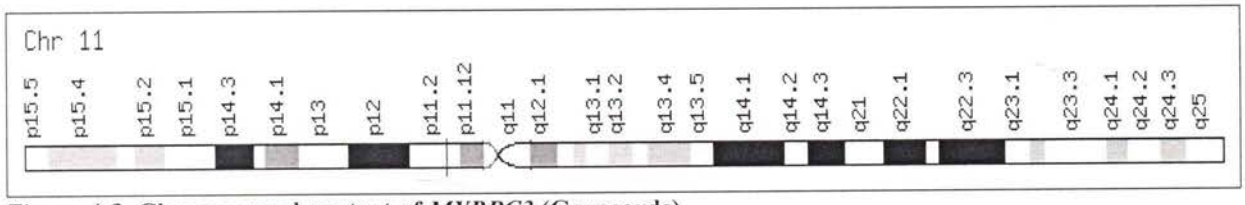


Figure 4-3 Chromosomal context of *MYBPC3* (Genecards).
Start: 47,309,527 bp from pter. End: 47,330,806 bp from pter. Size: 21,279 bp.

By promoting polymerisation of thick filaments, MYBPC is involved in sarcomere assembly, from the C-terminal domains binding to sites on titin and light meromyosin. It also binds to actin and myosin S2 subfragment. The regulatory role of MYBPC occurs primarily through the N-terminal domains, via phosphorylation-dependent interaction with the myosin crossbridges, which modulates muscle contraction. How the MYBPC attaches to the myosin thick filament is uncertain. It is thought that MYBPC3 proteins trimerise to form a collar around the thick filament with domains C5–C7 of one MYBPC3 overlying C8–C10 of the next (Figure 4-4). The cross-bridge formation may be increased or reduced, depending on whether the interaction is released or formed, respectively. The alterations of the MYBPC3 collar, caused by the FHC mutations, indicate its role in thick filament structure and regulation (Moolman-Smook et al. 2002). There is also a different model for MYBPC

amalgamation into the thick filament where there is no interaction between MYBPC molecules (Squire et al. 2003).

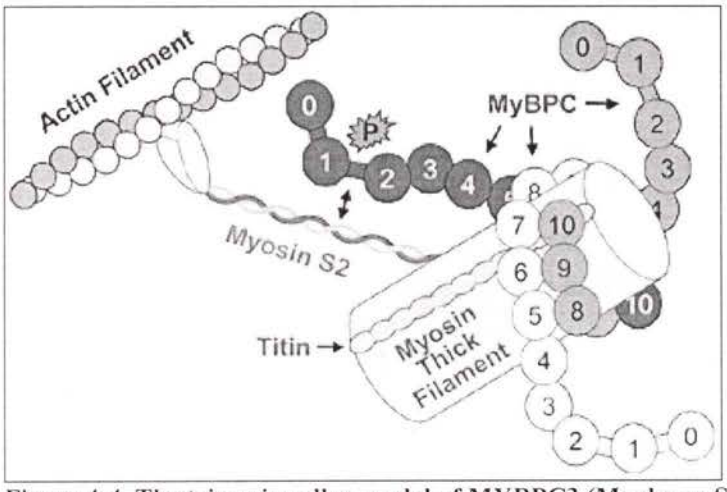


Figure 4-4 The trimeric collar model of MYBPC3 (Moolman-Smook et al. 2002).

4.1.2 MYBPC3

MYBPC3 is a sarcomeric protein and is part of the intracellular immunoglobulin superfamily. It has structural and regulatory roles, although its function is not fully understood. *MYBPC3* gene is made up of over 21,000 bp and contains 35 exons, with two exons only 3 bp each (Carrier et al. 1997) (Figure 4-5).

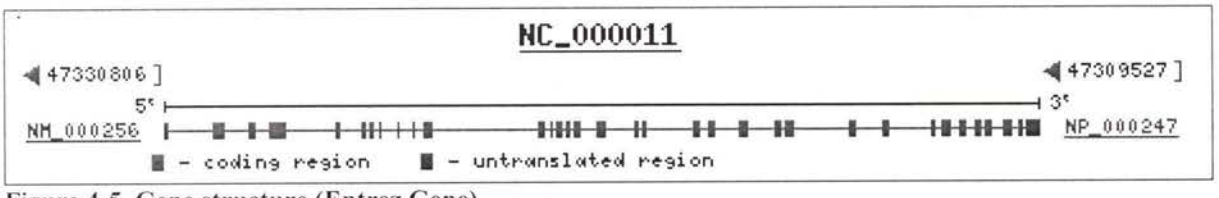


Figure 4-5 Gene structure (Entrez Gene).

MYBPC3 is important for myofibrillogenesis and regenerating muscle cells. *MYBPC3* gene knockout mice demonstrate that MYBPC3 is not necessary for the sarcomere to form, but that without it there is impaired contractile function and hypertrophy. There is a relationship between the extent of phosphorylation of MYBPC and increased systolic tension. This is caused by adrenergic stimulation. When MYBPC is extracted from fibres, calcium-ion sensitivity and shortening velocity is increased. The association of MYBPC3 with the heart

disease, FHC, has made it an object of intense interest. Mutations in ten sarcomeric proteins are known to cause FHC. How these mutations cause sarcomeric dysfunction is not clear and this has created much interest in understanding its underlying physiology (Oakley et al. 2004).

4.1.2.1 Domains and Motifs

Sequence alignments from various species and protein motifs have found important conserved regions for the classification of domains. The structure of these homologous domains, in conjunction with the location of FHC mutations, improves the understanding of how FHC point mutations can change the structure of its motifs. These subtle differences may have consequences for the stability, binding or function of the protein (Oakley et al. 2004). Exon 4 (e4) is located in the Motif C0–C1 Linker region, exon 5 (e5) in Motif C1 and exon 6 (e6) in Motif C1.

4.1.2.2 Phosphorylation Sites

There is evidence for as many as three binding sites within the region of Motifs 0–2. There is the well-known phosphorylation-dependant binding of the Motif 1–2 linker to S2 of myosin and recent evidence for Motif 0 and the Motif 0–1 linker binding to myosin or actin. Motif 0 has very few FHC-associated mutations. A mutant MYBPC knock-in mouse model indicated an interaction between Motif 0 and part of the myosin crossbridge. The MYBPC lacked both Motifs 0–1 linker and Motif 1; the Motif 1–2 linker was, however, still phosphorylatable. There are multiple FHC mutations in Motifs 1 and 2 and the Motif 1–2 linker. This linker has two additional phosphorylation sites (S284 and S304) in the cardiac MYPBC, including a cardiac-specific LAGGGRRIS sequence (S284) (Figure 4-6). Calcium/calmodulin-dependent kinases can phosphorylate these three phosphorylation sites (Oakley et al. 2004). Regulation of muscle contraction is thought to occur by the Motif 1–2 linker region by binding to the

myosin S2 region. The myosin S2 region can, by the phosphorylation of cardiac MYBPC, link more efficiently with actin, thereby increasing its force and systolic tension. This has been described as releasing of the braking effect on crossbridge cycling (Flashman et al. 2004). The flexibility of the myosin head allows more efficient force. This could also have an effect on diastolic filling of the heart (Kulikovskaya et al. 2003). Two of the mutations in the Motif 1–2 linker, G278E and G279A, are in the conserved, cardiac LAGGGRRIS sequence, near two of the phosphorylation sites (Figure 4-6). MYBPC can be phosphorylated by protein kinase C under non-physiological conditions. MYBPC1 and MYBPC2 are probably not phosphorylated by calcium/calmodulin-dependent kinases because they do not contain the LARGGGRRIS insertion. MYBPC3 regulation of cardiac muscle appears to entail hierarchical phosphorylation which does not apply to skeletal muscle (Flashman et al. 2004).

```
        10          20          30          40          50          60
MPEPGKKPVS AFSKKPRSVE VAAGSPAVFE AETERAGVKV RWQRGGSDIS ASNKYGLATE
        70          80          90         100         110         120
GTRHTLTVRE VGPADQGSYA VIAGSSKVKF DLKVIEAEKA EPMLAPAPAP AEATGAPGEA
                                    e4 [M]
       130         140         150         ↑          170         180
PAPAAELGES APSPKGSSSA ALNGPTPGAP DDPIGLF[V]MR PQDGEVTVGG SITFSARVAG
     e5 [I]                                                  e6 [G]
         ↑          200         210         220         230       ↑ 240
ASLLKPPV[V]K WFKGKWVDLS SKVGQHLQLH DSYDRASKVY LFELHITDAQ PAFTG[S]YRCE
       250         260         270         280         290         300
VSTKDKFDCS NFNLTVHEAM GTGDLDLLSA FRRTSLAGGG RRISDSHEDT GILDFSSLLK
       310         320         330         340         350         360
KRDSFRTPRD SKLEAPAEED VWEILRQAPP SEYERIAFQY GVTDLRGMLK RLKGMRRDEK
       370         380         390         400         410         420
KSTAFQKKLE PAYQVSKGHK IRLTVELADH DAEVKWLKNG QEIQMSGSKY IFESIGAKRT
       430         440         450         460         470         480
LTISQCSLAD DAAYQCVVGG EKCSTELFVK EPPVLITRPL EDQLVMVGQR VEFECEVSEE
       490         500         510         520         530         540
GAQVKWLKDG VELTREETFK YRFKKDGQRH HLIINEAMLE DAGHYALCTS GGQALAELIV
       550         560         570         580         590         600
QEKKLEVYQS IADLMVGAKD QAVFKCEVSD ENVRGVWLKN GKELVPDSRI KVSHIGRVHK
       610         620         630         640         650         660
LTIDDVTPAD EADYSFVPEG FACNLSAKLH FMEVKIDFVP RQEPPKIHLD CPGRIPDTIV
       670         680         690         700         710         720
VVAGNKLRLD VPISGDPAPT VIWQKAITQG NKAPARPAPD APEDTGDSDE WVFDKKLLCE
       730         740         750         760         770         780
TEGRVRVETT KDRSIFTVEG AEKEDEGVYT VTVKNPVGED QVNLTVKVID VPDAPAAPKI
       790         800         810         820         830         840
SNVGEDSCTV QWEPPAYDGG QPILGYILER KKKKSYRWMR LNFDLIQELS HEARRMIEGV
       850         860         870         880         890         900
VYEMRVYAVN AIGMSRPSPA SQPFMPIGPP SEPTHLAVED VSDTTVSLKW RPPERVGAGG
       910         920         930         940         950         960
LDGYSVEYCP EGCSEWVAAL QGLTEHTSIL VKDLPTGARL LFRVRAHNMA GPGAPVTTTE
       970         980         990        1000        1010        1020
PVTVQEILQR PRLQLPRHLR QTIQKKVGEP VNLLIPFQGK PRPQVTWTKE GQPLAGEEVS
      1030        1040        1050        1060        1070        1080
IRNSPTDTIL FIRAARRVHS GTYQVTVRIE NMEDKATLVL QVVDKPSPPQ DLRVTDAWGL
      1090        1100        1110        1120        1130        1140
NVALEWKPPQ DVGNTELWGY TVQKADKKTM EWFTVLEHYR RTHCVVPELI IGNGYYFRVF
      1150        1160        1170        1180        1190        1200
SQNMVGFSDR AATTKEPVFI PRPGITYEPP NYKALDFSEA PSFTQPLVNR SVIAGYTAML
      1210        1220        1230        1240        1250        1260
CCAVRGSPKP KISWFKNGLD LGEDARFRMF SKQGVLTLEI RKPCPFDGGI YVCRATNLQG
      1270
EARCECRLEV RVPQ
```

Figure 4-6 MYBPC3 protein sequence
e4–6 variants boxed, phosphorylation sites (S284 and S304) and mutations (G278E and G279A) in bold, conserved LAGGGRRIS sequence underlined (UniProtKB/Swiss-Prot: http://ca.expasy.org/uniprot/Q14896). Length: 1,274 AA, molecular weight: 140,762 Da.

4.1.2.3 Protein Folding and Structure

There is little information in the literature regarding how the changes in the amino acids in e4–6 may affect the folding of the protein and how this may affect the interactions of the

protein with myosin S2 to affect heart function. The e6 variant 5190A→G (S236G) is found in Motif 1, near the Motif 1-2 phosphorylatable linker. It is close to the phosphorylation sites and the LAGGGRRIS sequence (Figure 4-6). Glycine is the only residue that does not have a side chain. This means that it can have phi (φ) and psi (ψ) angles in the four quadrants of the Ramachandran plot and is the least restricted amino acid. If it is substituted, it may affect the three dimensional structure of the region or obstruct kinase binding. In the normal mouse heart, the 279 amino acid is an alanine (A) (Oakley et al. 2004). The e4 variant is 3634G→A (V158M). The e5 variant is 3817G→A (V189I).

4.1.3 Physiology and Biochemistry

4.1.3.1 Heart Function

MYBPC3 regulates myocardial work capacity by limiting power output (Korte et al. 2003). Contractile force is increased in conjunction with increasing heart rate in a normal heart. Phosphorylation of MYBPC3 changes the rates of force generation and troponin I changes the rates of relaxation, co-ordinately. The myofilament can be affected through a feedback mechanism in the crossbridges (Tong et al. 2004). MYBPC3 is important for the time course and extent of LV systolic stiffening. MYBPC3 deficient homozygous truncated MYBPC3 male mice had impaired sarcomere shortening velocity and reduced muscle stiffening (Palmer et al. 2004a). The same mice had, however, increased contractile efficiency, thought to result from an attenuated decrease of mechanical energy and an increase of phosphate-dependent oscillatory work (Palmer et al. 2004b).

The energy requirement of the physiologically hypertrophied, compared to the non-hypertrophied heart, in trained and untrained healthy individuals, is lower at rest and during exercise, than would be expected from differences in heart weight (Heiss et al. 1977).

$\dot{V}O_2$ max correlates with LV mass and inversely with LV wall stress. Myocardial glucose uptake relative to muscle mass is decreased due to decreased wall stress and energy requirements, and/or the use of additional fuel sources. Lactate becomes the prime energy source of the heart during blood lactate-producing exercise (Nuutila et al. 1994).

The MYBPC3 C5 motif may determine the folding of MYBPC3 and the manner in which it interacts with myosin. Its modification by mutation can affect contractility greatly. The effect of the motif is regulated by phosphorylation and appears to be important for the contractile control mechanism (Winegrad 2003, 2005). Data on the phosphorylation state in stressed and unstressed mouse hearts suggest that MYBPC3 phosphorylation is essential for normal cardiac function (Sadayappan et al. 2005). MYBPC3 is tris-phosphorylated at a conserved N-terminal domain (MYBPC3 motif) by cAMP-dependent protein kinase. The MYBPC3 motif is bound to a conserved section of sarcomeric myosin S2 through phosphorylation regulation (Kunst et al. 2000).

The phosphorylation in cardiac muscle probably has a specific regulatory function because its MYBPC has four phosphorylation sites compared to the one site of skeletal MYBPC. The effect of phosphorylation of MYBPC3 on cross bridges of isolated natural thick filaments from cardiac muscle was determined using electron microscopy and optical diffraction. The results indicate that phosphorylation of MYBPC3 extends the cross bridges from the backbone of the filament and affects their order and/or direction. This could affect the thin filament cycling rate and so modify cardiac force production (Weisberg and Winegrad 1996).

MYBPC3 has four sites that can be phosphorylated by a Ca^{2+}-calmodulin-controlled kinase, protein kinase A or protein kinase C. Electron microscopy and optical diffraction were used to study the structure of thick filaments isolated from rat ventricles with either the alpha or

beta isoform of myosin heavy chain and the effect of specific phosphorylation of MYBPC3 on the structure. Crossbridges were easily seen in thick filaments with alpha-myosin heavy chain. Phosphorylation of MYBPC3 by protein kinase A extended the crossbridges from the backbone of the filament, changed their orientation, increased the crossbridges order, and decreased the crossbridges flexibility. In filaments with beta-myosin heavy chain, the crossbridges were less ordered and less stiff. Phosphorylation of MYBPC3 in beta- myosin heavy chain-containing filaments did not extend the crossbridges and did not alter degree of order or flexibility. The crossbridge flexibility correlated with the rate of ATP hydrolysis which suggests that it is a key factor of crossbridge cycling rate, and MYBPC3-mediated control of the position and flexibility of crossbridges may regulate actomyosin ATPase activity by adjusting the kinetics of crossbridge cycling (Weisberg and Winegrad 1998).

4.1.3.2 Disease States

MYBPC3 mutations cause FHC (Figure 4-7). It is a heart disorder involving ventricular hypertrophy, usually asymmetric and commonly includes the interventricular septum. The disease is found in about 0.2% of the general population. The clinical manifestations are heterogeneous, ranging from benign to severe within and between families. There is a greatly increased risk of cardiac failure and sudden death. The *MYBPC3* gene mainly has truncated protein mutations, whereas the majority of mutations in FHC caused by mutations in other genes are from missense changes. Dominant negative and haploinsufficiency are two of the models that could explain the mechanism of FHC (Yu et al. 1998). Hypertrophic cardiomyopathy is an autosomal-dominant disorder. Many mutations across 11 genes are known: the beta-myosin heavy chain gene, alpha-myosin heavy chain, cardiac troponin T; cardiac troponin C, alpha-tropomyosin, *MYBPC3*, cardiac troponin, essential and regulatory light chain genes, cardiac alpha-actin gene and titin (Ramirez and Padron 2004). There

appears to be a gene dose effect in patients who have compound-heterozygous, double-heterozygous, or homozygous mutations. Screening for mutations should begin with the common mutations: *MYBPC3* and beta-myosin heavy chain gene and progress to cardiac troponin, cardiac troponin T, and cardiac myosin light chain 2. Multiple mutations should be screened in severe phenotypes (Richard et al. 2003). Hypertrophic cardiomyopathy is a cardiac disease which is multigenetic. Its pattern of inheritance is autosomal dominant and it has incomplete penetrance, except for mitochondrial genome mutations. Mutations have been found in four thick filament genes: beta myosin heavy chain, essential myosin light chains, regulatory myosin light chains, and *MYBPC3*; in five thin filament genes: cardiac actin, cardiac troponin T, cardiac troponin C, cardiac troponin I, and alpha-tropomyosin; and in the sarcomeric cytoskeletal protein titin. Mutations in other non-sarcomeric genes have been found, as well as sarcomeric mutations, in patients with mixed hypertrophic cardiomyopathy (Capek and Brdicka 2006).

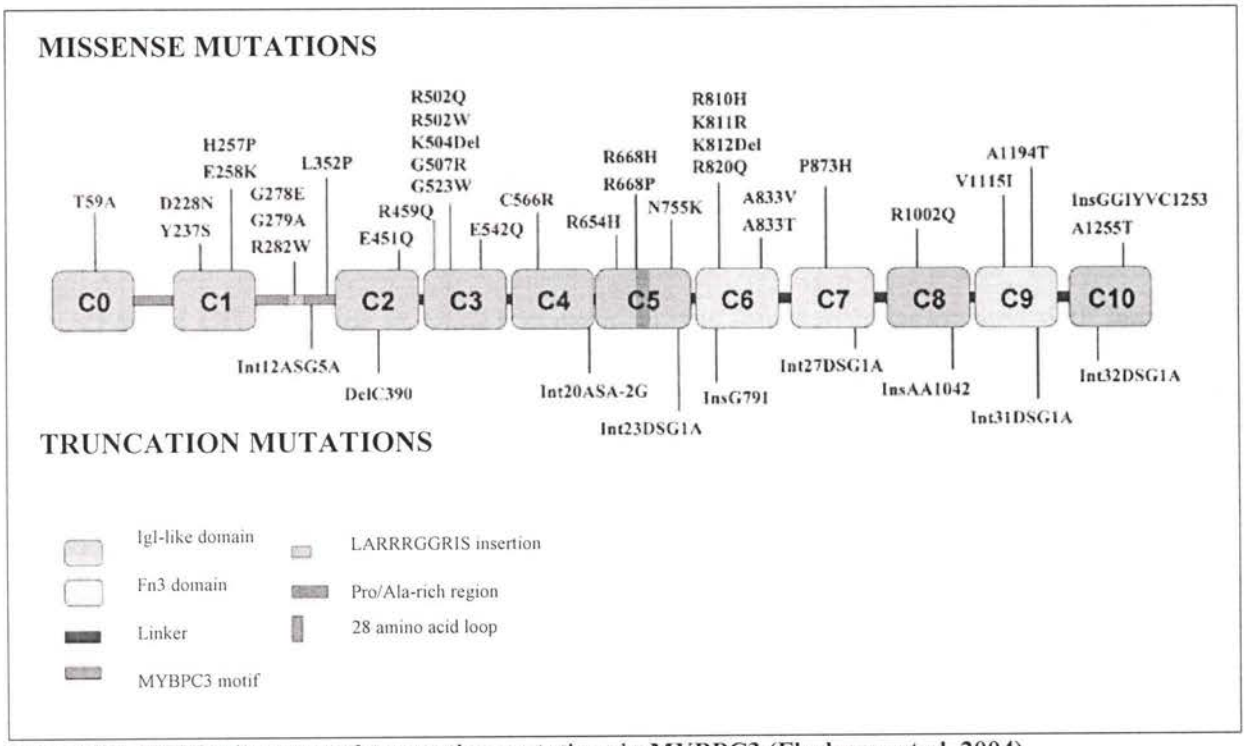


Figure 4-7 HCM missense and truncation mutations in MYBPC3 (Flashman et al. 2004).

Myocyte contractility disfunction is considered to be the mechanism for FHC mutations in *in vitro* studies. There is evidence for different mechanisms of disease for the different types of mutations. Another hypothesis is "energy compromise" from ATP inefficiency but there is little evidence in human FHC cases apart from a bioenergetic deficit (Crilley et al. 2003). Most FHC patients with an MYBPC3 mutation demonstrate low penetrance, late onset and a benign phenotype. If sudden death occurs, it is often as a result of stress and occurs later than 40 years of age. In a mouse model, exercise-related mortality (severe bradycardia) and sudden death were present (Yang et al. 2001). The clinical features of FHC due to heterozygous MYBPC3 mutations compared to myosin heavy chain gene mutations showed significantly better prognosis, older age of onset and lower penetrance before the age of 30, leading to a milder phenotype with less hypertrophy (Charron et al. 1998). Late-onset hypertrophic cardiomyopathy with no other family history has a late onset of symptom development and diagnosis at about 60 years of age. Mutations were found in the *MYBPC3*, troponin I, and alpha-cardiac myosin heavy chain genes and included missense mutations, truncating mutations, and splice mutations (Niimura et al. 2002). Knockout mice with deletion of *MYBPC3* exons 3–10 had cardiac development but showed obvious cardiac hypertrophy and impaired contractile function (Harris et al. 2002).

The *MYBPC3 GG* genotype at 18,443 in exon 30 was found to be associated with increased LV wall thickness in a patient group compared to the *AA* and *AG* genotypes. Patients and controls showed no difference in the genotype distribution. The *MYBPC3 GG* genotype may be associated with the severity of LV hypertrophy in patients with FHC, that is, a modifier gene for it (Wang et al. 2005).

A patient with a homozygous mutation in beta-myosin heavy chain gene (R403W) and a heterozygous variant in *MYBPC3* (V896M) showed a significant increase in sliding velocity

of actin filaments and an enhancement of mechanical and enzymatic properties of human mutant myosin. This suggested inefficient cardiac ATP utilisation and reduced mechanical efficiency (Keller et al. 2004).

DNA sequencing was performed in the Maine Coon cat *MYBPC3* gene and sequence alterations showed that they changed the amino acid produced, that the amino acid was conserved and that the protein structure was altered. There was a single base pair change (G→C) in the feline *MYBPC3* gene in affected cats that computationally alters the protein conformation of this gene and results in sarcomeric disorganisation. A causative, spontaneous mutation in the feline *MYBPC3* gene leads to the development of FHC (Meurs et al. 2005).

4.1.3.3 Athletes

The "athlete's heart" is a reversible cardiac hypertrophy which occurs in some athletes in response to prolonged endurance training. When the increased heart weight is normalised to bodyweight in some athletes, such as weightlifters, hypertrophy is often reduced but can present as concentric hypertrophy (increased wall thickness). With elite endurance athletes, there is usually significant eccentric hypertrophy (increased wall thickness and cavity size) (Czubryt and Olson 2004). Hypertrophy in athletes can show changes in electrocardiograms, although these do not usually correlate with disease (Oakley 2001). It seems to be beneficial in endurance athletes giving a balanced enlarged heart because stroke volume is increased with decreased resting heart rate, and with more-efficient pumping of blood (Scharhag et al. 2002).

Sudden cardiac death is rare among elite athletes and is less frequent than in the general population. The majority occur from undiagnosed genetic problems such as FHC (Futterman

and Myerburg 1998). In athletes under 35 years, the majority of sudden deaths are caused by congenital cardiac conditions. Hypertrophic cardiomyopathy is the most common accounting for 36% of sudden cardiac deaths in young competitive athletes, based on tracking of 158 athletes (Maron et al. 1996).

Distinguishing athlete's heart from FHC seems to only be a problem for male athletes (Pelliccia et al. 1996). Abnormal electrocardiograms have been detected in up to 40% of athletes, whereas cardiac disease has been found in less than 5% of athletes. A small proportion of athletes had electrocardiogram abnormalities that indicated cardiac disease, but they actually had no disease. These appear to have resulted purely from training (Pelliccia and Maron 2001). Significant LV wall thickening (>13 mm) in strength-trained athletes is indicative of pathologic hypertrophy, that is, FHC (Pelliccia et al. 1993). Elite ice hockey players showed the increased ventricular size and wall thickness characteristic of combined endurance and power sports (Bossone et al. 2004).

4.1.3.4 Exon 6 S236G Variant

Seven novel variants (Gln1061X, IVS5-2A→C, IVS14-13G→A, e25ΔLys, Pro147Leu, Ser236Gly, and Arg1138His) and two known variants (Arg326Gln, Val896Met) have been found in MYBPC3. They were all predicted to alter the structure of the protein. Some of the missense changes were, nevertheless, not considered to be disease-causing mutations. The Ser236Gly did not cosegregate with FHC-phenotype in family studies, was not within the highly conserved region of the MYBPC3 when compared with other species and had multiple distinct haplotypes in the FHC families and control subjects which indicated multiple ancestral origins (Jaaskelainen et al. 2002) (Figure 4-8).

Residue change from Ser (S) to Gly (G), S236G
Location on the sequence 216 ASKVYLFELHITDAQPAFTG S YRCEVSTKDKFDCSNFNLTV 256
↓
G

Figure 4-8 e6 S236G variant properties
(1,274 amino acids, physico-chemical property: changes from small size, polar, hydrophillic serine (S) to nonpolar, hydrophobic glycine (G))

4.1.3.5 Sequence Homology

The sequence homology of MYBPC3 is 46.8% in different species. It has three main differences from the skeletal isoforms: (1) an extra IgI motif at the N-terminus region (Motif 0); (2) two extra phosphorylation sites between Motifs 1 and 2; and (3) a proline/charge-rich insert in the central IgI domains (Motif 5). The cardiac isoform retained most of its defining features in the Japanese pufferfish, *Fugu rubripe*, except for one of the extra phosphorylation sites (Oakley et al. 2004). The sequence alignment of e6 S236G from BLAST search shows conservation between human and mouse DNA (Figure 4-9).

```
                                ASKVYLFELHITDAQPAFTGSYRCEVSTKDKFDCSNFNLTV
sp!Q14896!MYPC3_HUMAN 216 ASKVYLFELHITDAQPAFTGSYRCEVSTKDKFDCSNFNLTV 256
tr!Q9UM53_HUMAN       216 ASKVYLFELHITDAQPAFTGSYRCEVSTKDKFDCSNFNLTV 256
sp!O70468!MYPC3_MOUSE 214 ASKVYLFELHITDAQtTsAGGYRCEVSTKDKFDscNFNLTV 254
sp!Q90688!MYPC3_CHICK 206 -NKVYtFEMeIiEAnmTFAGGYRCEVSTKDKFDsSNFNLiV 245
tr!Q90X86_XENLA       222 -TKIYtFEIQIigAktTYAGGYRCEVSSKDKFDscNFNLAV 261
tr!Q6IP30_XENLA       222 -TKIYtFEIQIigAktTYAGGYRCEVSSKDKFDscNFNLAV 261
```

Figure 4-9 Alignment of e6 S236G from BLAST search.

4.2 Materials and Methods

4.2.1 Subjects

4.2.1.1 Elite athletes

Elite athletes included rowers, cyclists and sprint runners from the AIS. Elite and non-elite rugby players were from various amateur and professional rugby clubs in New South Wales and the Australian Capital Territory.

4.2.1.2 Controls

Controls were males and females selected from Blood Bank normal controls and de-identified genetic patients' (not athletes) non-affected partners. Comparisons between athletes and controls were gender-matched.

4.2.2 Materials

4.2.2.1 Oligonucleotides

The oligonucleotides used for this chapter were ordered from Invitrogen™ and BioLabs™.

4.2.3 Methods

4.2.3.1 PCR Amplification

Available PCR primers from our laboratory were used to amplify *MYBPC3* e6, e4 and e5, which contain the three nonsynonymous SNP variants investigated (Table 4-1 and Table 4-2).

Table 4-1 *MYBPC3* primer sequences.

Exon	Primer Sequences:	Amplicon Length
e6	6F: (*MYBPC3*) 5′-ATT ACA GGC CTG AGC CAC CG-3′ 6R: (*MYBPC3*) 5′-AGA CCA GGA CCC ATG GGG AG-3′	283 bp
e4	4F: (*MYBPC3*) 5′-TGG GAG GCG GAG CTT GCA GTG-3′ 4R: (*MYBPC3*) 5′-CCC CTT CCC ACC CCA ATG CTG-3′	246 bp
e5	5F: (*MYBPC3*) 5′-GCA GCA GGA CAC TCC CCA AG-3′ 5R: (*MYBPC3*) 5′-TGT CTC CAC GAC CCC GGT-3′	221 bp

Table 4-2 *MYBPC3* PCR conditions.

	e6	e4	e5
Primer1: F (20 pmol.μL^{-1})	1.0 μL	1.0 μL	1.0 μL
Primer2: R (20 pmol.μL^{-1})	1.0 μL	1.0 μL	1.0 μL
dNTP (2.5 mM)	2.0 μL	2.0 μL	2.0 μL
GeneAmpR 10× PCR (buffer II)	5.0 μL	5.0 μL	5.0 μL
GeneAmpR $MgCl_2$ (25 mM stock)	1.5 μL	3.0 μL	3.0 μL
dH_20	37.3 μL	35.8 μL	35.8 μL
AmpliTaq™ Gold (5 U.μL^{-1})	0.2 μL	0.2 μL	0.2 μL
DNA (25 ng.μL^{-1})	2.0 μL	2.0 μL	2.0 μL
Total volume	50.0 μL	50.0 μL	50.0 μL
Thermal Cycling Conditions	95°C × 12 min × 1 (95°C × 30 s; 60°C × 30 s; 72 °C × 10 s) × 37 72°C × 7 min × 1	95°C × 12 min × 1 (95°C × 30 s; 65°C × 30 s; 72 °C × 10 s) × 35 72°C × 7 min × 1	95°C × 12 min × 1 (95°C × 30 s; 60°C × 30 s; 72 °C × 10 s) × 35 72°C × 7 min × 1
*product size:	283 bp	246 bp	221 bp

4.2.3.2 Restriction Fragment Length Polymorphism Genotyping

Restriction enzymes were used for restriction fragment length polymorphism (RFLP) genotyping the athletes along with 147 age-matched caucasian male controls. Restriction enzymes were used to genotype *MYBPC3* for non-disease-causing nonsynonymous SNP variants present in e6, e4 and e5 (Table 4-3).

Table 4-3 RFLP conditions for genotyping.

Exon	Enzyme	Incubation conditions	Restriction fragments
e6 – 5190A→G; S236G	Enzyme *Alu*I – loss of restriction site.	Incubate at 37°C overnight.	Wild type = 283 bp Variant = 126 bp + 119 bp + 38 bp
e4 – 3634G→A; V158M	Enzyme *Nla*III – creates restriction site.	Incubate at 37°C for 3 h.	Wild type = 246 bp Variant = 246 bp + 173 bp + 73 bp
e5 – 3817G→A; V189I	Enzyme *Bcl*I – creates restriction site.	Incubate at 50°C overnight.	Wild type = 221 bp Variant = 221 bp + 117 bp + 104 bp

4.2.3.3 Statistical Analysis

The CLUMP program was used for the χ^2 test for data where there were small allele or genotype frequencies. The CLUMP program was run through the DOS program and uses Monte Carlo simulation to perform the χ^2 test.

The SPSS™ program was used for the regression modelling analysis. The linear regression was used to compare genotype to phenotype data. Kevin McGeechan (Epidemiology and Biostatistics, University of Sydney) assisted with the preparation of the statistical results using the SPSS™ program.

4.3 Results

4.3.1 Exon 6

In e6, the nonsynonymous SNP variant – 5190A→G; S236G was genotyped. The mixed male and female results were not statistically significant (Table 4-4). The male sprinters and rugby players gave the most interesting results. The frequency of the *AG* genotype at 236 was significantly higher in male sprinters than 147 controls (p = 0.040), although the numbers were quite small. An increased frequency of the 236G allele and *AG* genotype was also found in Back-ten rugby players versus controls (p = 0.044, p = 0.021, respectively) and versus the Front-five rugby players (p = 0.032, p = 0.049, respectively) (Table 4-5 and Figure 4-10).

Table 4-4 Males and Females: e6: 5190A→G; S236G
(p values vs. Controls) (p values from χ^2 distribution from CLUMP program).

Group	n	Genotypes			p value	Alleles		p value
		AA	GA	GG		A	G	
CONTROLS	212	167	43	2		377	47	
ENDURANCE: E.cyc., Rowers	237	175	60	2	0.448	410	64	0.272
SPRINT: Spr.cyc., Spr.run.	40	26	14	0	0.109	66	14	0.107

Table 4-5 Males only: e6: 5190A→G; S236G
(p values vs. Controls) (p values from χ^2 distribution from CLUMP program).

Group	n	Genotypes			p value	Alleles		p value
		AA	AG	GG		A	G	
CONTROLS	147	122	23	2		267	27	
ENDURANCE: E.cyc., Rowers	150	114	34	2	0.307	262	38	0.174
SPRINT: Spr.cyc., Spr.run.	28	18	10	0	0.040	46	10	0.053
Rugby: Front-five	74	63	9	2	0.628	135	13	0.890
Rugby: Back-ten	121	85	34	2	0.044	204	38	0.021
All RUGBY: Front-five, Back-ten	195	148	43	4	0.280	339	51	0.113
(Rugby: Front-five versus Back-ten)					(0.032)			(0.049)

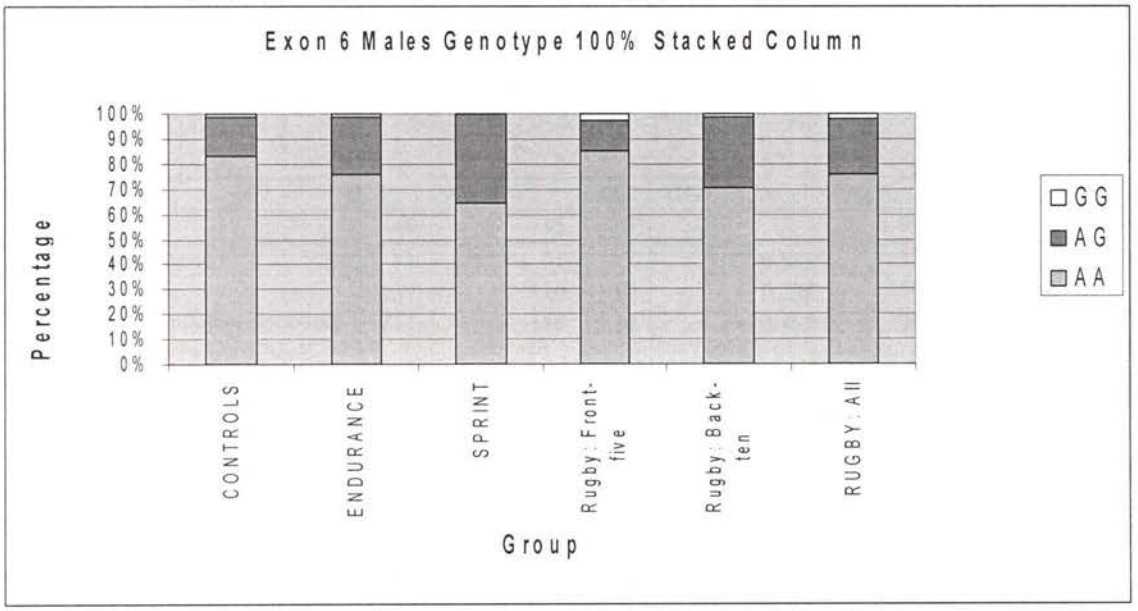


Figure 4-10 Bar graph of e6 Males only genotype results.

4.3.2 Exon 4

Since rugby players gave the most interesting results with the largest numbers, fewer of the other athletes were genotyped for the e4 (V158M) variant because of a lack of promising results and concerns over a lack of statistical power. The mixed male and female results were not statistically significant ($p > 0.050$) (Table 4-6). The rugby players versus the mixed male and female controls were not statistically significant ($p > 0.050$). The rugby players versus the male controls were not statistically significant ($p > 0.050$) (Table 4-7). The Back-ten versus the Front-five rugby players were not statistically significant ($p > 0.050$).

Table 4-6 Males and Females: e4: 3634G→A; V158M
(p values vs. Controls) (p values from χ^2 distribution from CLUMP program).

Group	N	Genotypes			p value	Alleles		p value
		AA	AG	GG		A	G	
CONTROLS	87	75	12	0		162	12	
ENDURANCE: E.cyc.	79	68	10	1	0.565	146	12	0.806
Rugby: Front-five	71	63	7	1	0.415	133	9	0.843
Rugby: Back-ten	137	116	18	3	0.380	250	24	0.480
All RUGBY: Front-five, Back-ten	214	185	20	4	0.257	390	28	0.930
(Rugby: Front-five versus Back-ten)					0.720			0.386

Table 4-7 Males only: e4: 3634G→A; V158M (*p* values vs. Controls)
(*p* values vs. Controls) (*p* values from χ^2 distribution from CLUMP program).

Group	N	Genotypes			*p* value	Alleles		*p* value
		AA	AG	GG		A	G	
CONTROLS	48	45	3	0		93	3	
Rugby: Front-five	71	63	7	1	0.549	133	9	0.266
Rugby: Back-ten	137	116	18	3	0.238	250	24	0.065
All RUGBY Front-five, Back-ten	214	185	20	4	0.468	390	28	0.185
(Rugby Front-five versus Back-ten)					0.720			0.386

4.3.3 Exon 5

None of the other athletes were genotyped for the e5 (V189I) variant because even the large group of rugby players seemed to vary little from the controls. The rugby players versus the mixed male and female controls were not statistically significant ($p > 0.050$) (Table 4-8). Tests of rugby players versus the male controls were not applicable (n/a) for e5 because there were no *AG* or *GG* genotypes in any groups (Table 4-9).

Table 4-8 Males and Females: e5: 3817G→A; V189I
(*p* values vs. Controls) (*p* values from χ^2 distribution from CLUMP program).

Group	N	Genotypes			*p* value	Alleles		*p* value
		AA	AG	GG		A	G	
CONTROLS	69	68	1	0		137	1	
Rugby: Front-five	72	72	0	0	0.591	144	0	0.306
Rugby: Back-ten	136	136	0	0	0.372	272	0	0.160
All RUGBY: Front-five, Back-ten	213	213	0	0	0.213	426	0	0.079
(Rugby Front-five versus Back-ten)					n/a			n/a

Table 4-9 Males only: e5: 3817G→A; V189I
(*p* values vs. Controls) (*p* values from χ^2 distribution from CLUMP program).

Group	N	Genotypes			*p* value	Alleles		*p* value
		AA	AG	GG		A	G	
CONTROLS	38	38	0	0		76	0	
Rugby: Front-five	72	72	0	0	n/a	144	0	n/a
Rugby: Back-ten	136	136	0	0	n/a	272	0	n/a
All RUGBY: Front-five, Back-ten	213	213	0	0	n/a	426	0	n/a
(Rugby Front-five versus Back-ten)					n/a			n/a

4.3.4 Genetic and Physiological Results

Multiple regression analysis was performed to determine the significant predictors of performance for only the rugby players for e6 S236G variant because they gave the only statistically significant results and with enough subjects to achieve sufficient statistical power.

4.3.4.1 Maximal Oxygen Uptake

Weight and elite were the statistically significant predictors of $\dot{V}O_2$ max ($p < 0.001$) and approximately 13% of the variance in $\dot{V}O_2$ max was accounted for by the linear combination of these variables. *MYBPC3* e6 S236G variant was not a significant predictor of $\dot{V}O_2$ max when added to the model ($p = 0.510$) (Table 4-10).

4.3.4.2 40 m Sprint Time

Age, weight and elite were the statistically significant predictors of 40 m Sprint Time in rugby players ($p < 0.001$) and approximately 28% of the variance in 40 m Sprint Time was accounted for by the linear combination of these variables. *MYBPC3* e6 S236G variant was not a significant predictor of 40 m Sprint Time when added to the model ($p = 0.901$) (Table 4-10).

Table 4-10 Summary of multiple regression ANOVA related to performance (results from SPSS™ program).

Outcome variable	Explanatory variables	Coefficients B	*p* value	r^2	genotype *p* value
$\dot{V}O_2$ max	weight elite	-0.032 0.626	< 0.001	13.0%	0.510
40 m Sprint Time	age weight elite	0.043 0.008 -0.355	< 0.001	27.7%	0.901

4.4 Discussion

4.4.1 Genetic results

This is the first time that athletes have been investigated for the *MYBPC3* gene. In e6, the nonsynonymous SNP variant – 5190A→G; S236G was genotyped. The mixed male and female results were not statistically significant (Table 4-4). The male sprinters and rugby players gave the most interesting results. The frequency of the *AG* genotype at 236 was significantly higher in male sprinters than 147 controls (p = 0.040), although the numbers were quite small and the result may simply reflect an artefact. An increased frequency of the 236G allele and *AG* genotype was also found in Back-ten rugby players versus controls (p = 0.044, p = 0.021, respectively) and versus the Front-five rugby players (p = 0.032, p = 0.049, respectively) (Table 4-5 and Figure 4-10).

The consistent trend for the male sprinters and Back-ten rugby players for e6 was for an increased *AG* genotype and *G* allele. This similarity in results would be expected because sprinters and Back-ten rugby players are considered to be power/sprint type athletes. As with the *ACE I/D* polymorphism, these results might indicate that for *MYBPC3* e6, the mixed genotype was beneficial. Perhaps there is a more moderate level of cardiac hypertrophy associated with the *AG* genotype and it is an advantage in combined power/endurance sports such as rugby. Both sports require phases of maximal distal skeletal muscle force/power combined with maximal isometric force/power of central respiratory and abdominal muscles during bouts of acceleration for sprinters and rugby players and during tackling/mauling for rugby players. These activities require the Valsalva maneouver to heighten central nervous system stimulation of muscle force/power (MacDougall et al. 1992). Optimisation of cardiac

muscle for these activities to ensure high cardiac output during phases of high blood pressure could be advantageous.

Since rugby players gave the most interesting results with the largest numbers, fewer of the other athletes were genotyped for the e4 (V158M) variant because of a lack of promising results and concerns over a lack of statistical power. The mixed male and female results were not statistically significant ($p > 0.050$) (Table 4-6). The rugby players versus the mixed male and female controls were not statistically significant ($p > 0.050$). The rugby players versus the male controls were not statistically significant ($p > 0.050$) (Table 4-7). The Back-ten versus the Front-five rugby players were not statistically significant ($p > 0.050$). The e4 variant does not appear to be important for athletic performance.

None of the other athletes were genotyped for the e5 (V189I) variant because even the large group of rugby players seemed to vary little from the controls. The rugby players versus the mixed male and female controls were not statistically significant ($p > 0.050$) (Table 4-8). Tests of rugby players versus the male controls were not applicable (n/a) ($p > 0.050$) (Table 4-9). The e5 variant does not appear to be important for athletic performance.

4.4.2 Genetic and Physiological Results

Multiple regression analysis was performed to determine the significant predictors of performance for only the rugby players for e6 because they gave the only statistically significant results and with enough subjects to achieve sufficient statistical power. Multiple regression analysis yielded two- and three-variable models, composed of age, weight and/or elite, which accounted for 13–28% of the variance in performance variables (Table 4-10).

Elite (positive coefficient) and weight (negative coefficient) were the statistically significant predictors of $\dot{V}O_2$ max (z-score) ($p = < 0.001$). *MYBPC3* e6 was not a significant predictor of $\dot{V}O_2$ max when added to the model ($p = 0.510$) (Table 4-10). Elite and weight would both be expected to be predictors of $\dot{V}O_2$ max. Unexpectedly, the e6 variant was not a predictor of $\dot{V}O_2$ max (z-score). This was the same situation, however, as the *ACE I/D* polymorphism. It is possible that the effect size is too small to be seen compared to the other predictive physical/phenotype variables.

Age (negative coefficient), weight (negative coefficient) and elite (positive coefficient) were the statistically significant predictors of 40 m Sprint Time (z-score) in rugby players ($p = < 0.001$). *MYBPC3* e6 was not a significant predictor of 40 m Sprint Time when added to the model ($p = 0.901$) (Table 4-10). Surprisingly, the addition of the *MYBPC3* e6 genotype to the either of the models added no significant prediction. Once again, this was the same lack of significance of the genotype to the physiological data as occurs with the *ACE I/D* polymorphism. Age, weight and elite would all be expected to be associated with 40 m Sprint Time. For age, younger players were faster and this can be explained by them generally being lighter and also carrying fewer chronic injuries, which can markedly reduce speed, than older players. Elite and lighter players would naturally be expected to be faster than other players.

The lack of significance of the *MYBPC3* e6 genotype to the model may have been due to group sizes being too small or due to the incomplete physiological data set reducing the statistical power. It also could have been because whatever might have been contributing to a difference in the genotype distributions between the athletes and controls was not measured by the standard physiological tests used.

MYBPC3 is considered to often act as a modifying gene (Wang et al. 2005) and so may be more likely to be a response gene rather than a baseline gene for performance. The response may, consequently, be more likely to be found in elite athletes than in lower level athletes and in older rather than younger players who have had more and longer training, allowing time for the differences to take effect. The fact that the heart muscle renews itself every few weeks (Sims et al. 1976; Zak 1977) may, on the other hand, contradict the idea that it takes a long time for the effect of any gene-environment interaction to affect the physiological impact of training on the heart.

The response due to being elite appears to have been reflected in the results of this study. The response due to age would, nevertheless, contradict this assertion of a modifying gene effect, since it was the younger players who were the fastest and not the oldest. The confounding factor of injuries was, perhaps, masking the effect.

This study adds to the body of research into the *MYBPC3* gene and cardiac hypertrophy. It showed that the *MYBPC3* e6, nonsynonymous SNP variant – 5190A→G; S236G, which was not associated with the aetiology of FHC, may have a modifying effect on cardiac function in the elite athlete model, although the evidence is weak. It also showed that the e4 and e5 variants probably did not have a modifying effect on cardiac function. It showed once again that even small statistical significance usually is only achieved when athlete groups are separated on the basis of duration and intensity of their event.

Chapter 5
Endothelial PAS Domain Protein-1 (*EPAS1*)

5.1 Introduction

The cardiovascular system plays a central role in many sports, especially endurance and $\dot{V}O_2$ max is the main measurement of cardiovascular function. The human map for genes related to athletic performance and health-related fitness traits mentioned in Chapter 1 shows hundreds of candidate genes related to human performance (Rankinen et al. 2001; Wolfarth et al. 2005). The genome-wide scan also mentioned in Chapter 1 (Bouchard et al. 2000) highlighted genomic regions with significant association with $\dot{V}O_2$ max, which is one of the most common and important measures of human performance. In the present study, the aim was to identify an underlying gene at the loci that had a suggestive linkage with $\dot{V}O_2$ max in this genome-wide scan. An *in silico* search was carried out by Dr Bing Yu to select plausible candidates.

SNPs

An association study was performed comparing elite endurance athletes classified into two groups: power–time-maximum (PT-MAX; n = 242, event duration 50 s to 10 min) and power–time–steady state (PT-SS; n = 151, event duration ~2–10 h), with controls (n = 444) using 12 SNPs across the *EPAS1* gene. Ordinal regression analysis of allele frequencies produced significant differences at SNPs 2 and 3 ($p = 0.01$). Haplotype analysis produced the haplotypes involving SNPs 2–5 that significantly differentiated ($p < 0.05$) the groups based on an ordinal ranking using the intensity classification. The same haplotypes differentiated the PT-MAX group in which a significant decrease in a haplotype (F: G-C-C-G; OR = 0.57, p = 0.02, 95% CI 0.36–0.92) and increase in a second haplotype (G: A-T-G-G; OR = 1.75, p = 0.03, 95% CI 1.05–2.91) was observed compared to controls. The PT-SS group was differentiated from the PT-MAX group by a third haplotype (H: A-T-G-A; OR = 0.46, p =

0.04, 95% CI 0.22–0.96). *EPAS1* functions as a complex sensor of O_2 availability. Since it differentiates between different groups of athletes, based on the intensity of their performance, it is proposed that DNA variants in *EPAS1* influence exercise metabolism and hence the maximum metabolic power.

DHPLC

The exons and exon-intron boundaries were analysed in eight amplicons (DNA amplified products) of *EPAS1* from representative DNA samples. After scanning through the amplicons using denaturing high performance liquid chromatography (DHPLC), six DNA variants were identified. The DNA samples from the 18 different SNP haplotypes produced three substitutions: a C→G transversion substitution at i8/e9, an A→C transversion substitution near i12/e13 and a G→A transition substitution is a synonymous codon in e15. Going from 18 SNP haplotypes to DHPLC-derived haplotypes produced three more substitutions: a T→C transition substitution at i8/e9, and a G→A transition missense substitution and a T→A transversion missense substitution resulting in two nonsynonymous codon variants in e9. A sample from each different waveform was sequenced and the variants were summarised. There were no statistically significant differences found.

Multiple regression analysis was performed to determine the significant predictors of performance. Only the *EPAS1* i8/e9 C_7/C_5GC polymorphism was used for this analysis because they were the only variants of sufficient quantity. Male, weight and Olympian were the statistically significant predictors of $\dot{V}O_2$ max ($p < 0.001$). Male, weight and height were the statistically significant predictors of 2 km Row Time ($p < 0.001$). Age was the only statistically significant predictor of Ironman Time ($p < 0.001$). The above three sports-specific measures of aerobic fitness were combined in a Fitness Z-score for analysis. Olympian and male were the statistically significant predictors of Fitness Z-score ($p < 0.001$).

EPAS1 i8/e9 C_7/C_5GC was not a significant predictor of any of the performance variables when added to the models.

The work described in this chapter involved state of the art genetic analysis technology. High throughput single nucleotide polymorphism (SNP) analysis, at the time that this work was being carried out, was only available at very few centres in Australia. This work required, therefore, input by a number of scientists and the relative contribution of each is summarised in Table 5-1.

Table 5-1. Summary of relative contribution to work in this chapter.

Activity	**Author's work**	**Others' work***
Obtain DNA samples	Rugby samples	RT, BY
Preparation of DNA samples	Ironman and rugby samples	BY, SC, ES, BG
Designing SNP probes	n/a	ABI
Preparation of SNP DNA plates	Most of this	Some contribution: JH
SNP genotyping	n/a	SUPAMAC
Troubleshooting failed SNP reactions	Much of this	BY, NS, SC
Organising repeat SNP reactions	Much of this	BY
Re-preparing SNP DNA samples	Much of this	SC, ES
DHPLC	8 amplicons	NS
Design probe amplicons	All of this	
Statistical analysis of SNP data	Nil	BY, JH, DD
Writing EPAS1 paper	Some contribution as second author	BY, JH, RT, SC
Statistical analysis of physiological data	Most of this	KM

* RT, Ron Trent; BY, Bing Yu, SC, Stuart Cole; JH, Jennifer Henderson; NS, Nicole Sawyer; DD, David Duffy; BG, Bamini Gopinath; ES, Edna Soriano; SUPAMAC, Sydney University Prince Alfred Macromolecular Analysis Centre; ABI, Applied Biosystems Incorporated; KM, Kevin McGeechan.

5.1.1 EPAS1

5.1.1.1 Anatomy and Structure

Endothelial PAS domain protein-1 (EPAS1) (also known as hypoxia inducible factor 2α; HIF-2α) is a PAS (per-arnt-sim) domain transcription factor. The PAS superfamily of basic helix-loop-helix (HLH) proteins is defined by the presence of two regions containing repeated sequences that share homology with the prototypical members from which the name of the family is derived. They are drosophila **per**iodic, the **a**ryl hydrocarbon receptor, the **a**ryl hydrocarbon **r**eceptor **n**uclear **t**ranslocator (ARNT, also known as hypoxia inducible factor 1β; HIF-1β) and drosophila **si**ngle **m**inded. This protein shares 48% sequence identity with hypoxia inducible factor 1α (HIF-1α) and lesser similarity with other members of the basic HLH/PAS domain family of transcription factors. The gene is located at 2p16.1 and consists of 16 exons and 89,273 bp. The protein has 870 amino acid residues (Tian et al. 1997) (Figure 5-1, Figure 5-2, Figure 5-3, Table 5-2 and Figure 5-4).

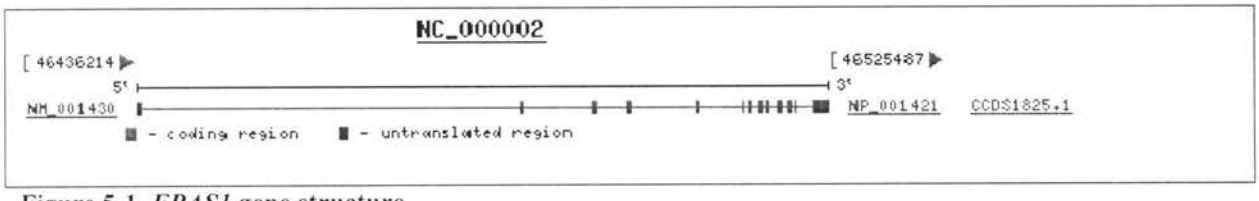


Figure 5-1 ***EPAS1*** **gene structure**
(Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez).

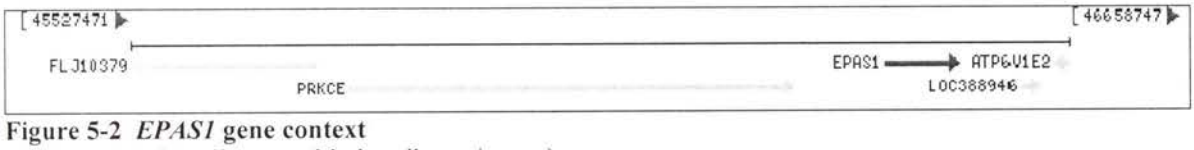


Figure 5-2 ***EPAS1*** **gene context**
(Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez).

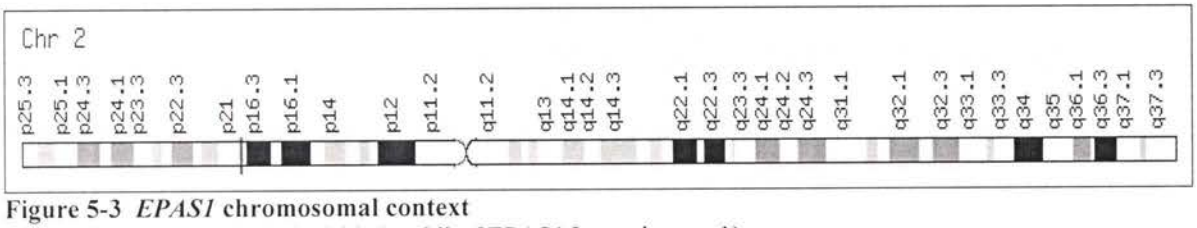


Figure 5-3 ***EPAS1*** **chromosomal context**
(http://www.genecards.org/cgi-bin/carddisp?EPAS1&search=epas1).

Table 5-2 *EPAS1* gene location and size
(http://www.genecards.org/cgi-bin/carddisp?EPAS1&search=epas1).

Start	46,436,214 bp from pter
End	46,525,487 bp from pter
Size	89,273 bp, 870 amino acids, 96,459 Da

```
        10         20         30         40         50         60
MTADKEKKRS SSERRKEKSR DAARCRRSKE TEVFYELAHE LPLPHSVSSH LDKASIMRLA
        70         80         90        100        110        120
ISFLRTHKLL SSVCSENESE AEADQQMDNL YLKALEGFIA VVTQDGDMIF LSENISKFMG
       130        140        150        160        170        180
LTQVELTGHS IFDFTHPCDH EEIRENLSLK NGSGFGKKSK DMSTERDFFM RMKCTVTNRG
       190        200        210        220        230        240
RTVNLKSATW KVLHCTGQVK VYNNCPPHNS LCGYKEPLLS CLIIMCEPIQ HPSHMDIPLD
       250        260        270        280        290        300
SKTFLSRHSM DMKFTYCDDR ITELIGYHPE ELLGRSAYEF YHALDSENMT KSHQNLCTKG
       310        320        330        340        350        360
QVVSGQYRML AKHGGYVWLE TQGTVIYNPR NLQPQCIMCV NYVLSEIEKN DVVFSMDQTE
       370        380        390        400        410        420
SLFKPHLMAM NSIFDSSGKG AVSEKSNFLF TKLKEEPEEL AQLAPTPGDA IISLDFGNQN
       430        440        450        460        470        480
FEESSAYGKA ILPPSQPWAT ELRSHSTQSE AGSLPAFTVP QAAAPGSTTP SATSSSSSCS
       490        500        510        520        530        540
TPNSPEDYYT SLDNDLKIEV IEKLFAMDTE AKDQCSTQTD FNELDLETLA PYIPMDGEDF
       550        560        570        580        590        600
QLSPICPEER LLAENPQSTP QHCFSAMTNI FQPLAPVAPH SPFLLDKFQQ QLESKKTEPE
       610        620        630        640        650        660
HRPMSSIFFD AGSKASLPPC CGQASTPLSS MGGRSNTQWP PDPPLHFGPT KWAVGDQRTE
       670        680        690        700        710        720
FLGAAPLGPP VSPPHVSTFK TRSAKGFGAR GPDVLSPAMV ALSNKLKLKR QLEYEEQAFQ
       730        740        750        760        770        780
DLSGGDPPGG STSHLMWKRM KNLRGGSCPL MPDKPLSANV PNDKFTQNPM RGLGHPLRHL
       790        800        810        820        830        840
PLPQPPSAIS PGENSKSRFP PQCYATQYQD YSLSSAHKVS GMASRLLGPS FESYLLPELT
       850        860        870
RYDCEVNVPV LGSSTLLQGG DLLRALDQAT
```

Figure 5-4 EPAS1 protein sequence
(UniProtKB/Swiss-Prot: http://ca.expasy.org/uniprot/Q99814). Length 870 AA, molecular weight: 96,459 Da.

EPAS1 is preferentially expressed in vascular endothelial cells. *EPAS1* shares high homology with HIF-1α and has also been shown to bind to the HIF-1-binding site and to activate its downstream genes such as vascular endothelial growth factor (*VEGF*) and erythropoietin (*EPO*). EPAS1 increases *VEGF* gene expression through the HIF-1-binding site (Maemura et al. 1999). Transactivation was improved by cotransfection of an ARNT expression plasmid. Deletion analysis of EPAS1 showed a strong activation domain (amino acids 486–639) vital

for transactivating the *VEGF* promoter. Its ability to activate transcription using a GAL4 fusion protein system was established. The truncated protein acts as a dominant-negative mutant because when missing the transactivation domain at amino acids 486–639, it abolished induction of the *VEGF* promoter by wild-type EPAS1. Infection of the cells with this mutant repressed the induction of *VEGF* mRNA under an environment that mimics hypoxia. This suggests that EPAS1 is an important regulator of *VEGF* gene expression (Maemura et al. 1999). Since VEGF plays a critical role in angiogenesis, the ability of dominant-negative EPAS1 to inhibit *VEGF* promoter activity suggests a new way to restrain pathological angiogenesis (Maemura et al. 1999).

5.1.1.2 EPAS1 Domains

There are three conserved domains in *EPAS1*: one HLH and two PAS domains (Figure 5-5).

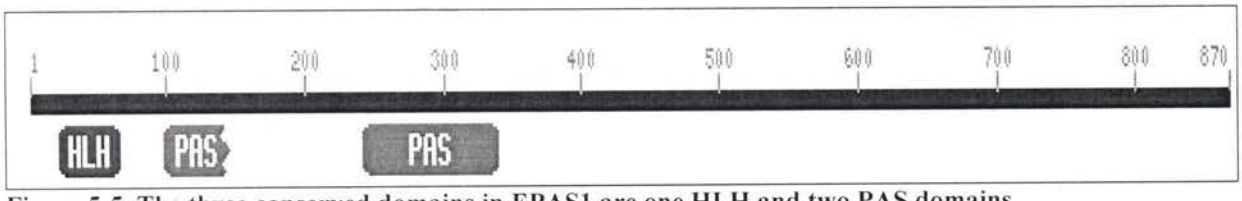


Figure 5-5 The three conserved domains in EPAS1 are one HLH and two PAS domains
(Entrez Gene:
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?INPUT_TYPE=precalc&SEQUENCE=40254439).

The HLH domain which is found in specific DNA-binding proteins that act as transcription factors is 60–100 amino acids long. A DNA-binding basic region is followed by two α-helices separated by a variable loop region; HLH forms homo- and heterodimers, dimerisation creates a parallel, left-handed, four helix bundle; the basic region N-terminal to the first amphipathic helix mediates high-affinity DNA-binding (from Entrez Gene) (Figure 5-6). The PAS domains have been found to bind ligands, and to act as sensors for light and oxygen in signal transduction (from Entrez Gene).

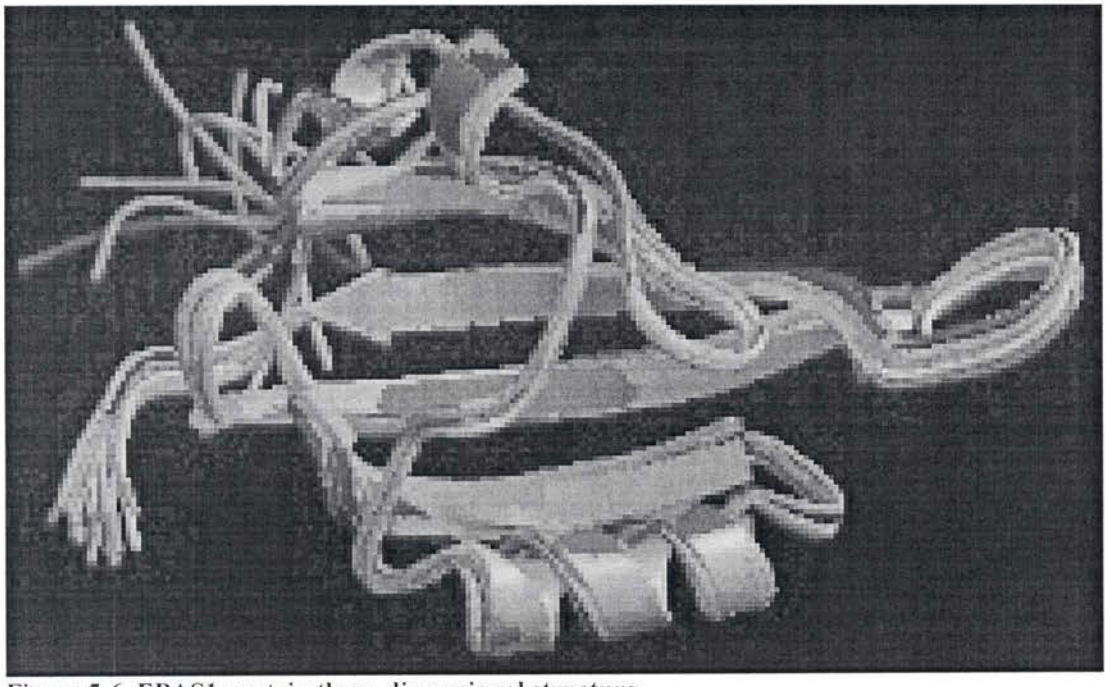


Figure 5-6 EPAS1 protein three-dimensional structure
(Protein Data Bank: http://www.pdb.org/pdb/explore.do?structureId=1P97).

Chimeric fusions of EPAS1 with a GAL4 DNA-binding domain, with or without the VP16 activation domain were used to analyse EPAS1 domains controlling transactivation and O_2-regulated function (O'Rourke et al. 1999). EPAS1 had two transactivation domains: a C-terminal domain (amino acids 828–870), and a large internal domain (amino acids 517–682). Interspersing these activation domains were functionally repressive sequences, some independently expressing O_2-regulated activity. N-terminal sequences overlapping the internal transactivation domain confer regulated repression on the VP16 transactivator. C-terminal sequences convey repression and O_2-regulated activity on the native EPAS1 C-terminal activation domain, but not the Gal/VP16 fusion. Fusions with internal but not C-terminal regulatory domains produced fusion protein level regulation. EPAS1 and HIF-1α proteins have a similar organisation with the C terminus containing a conserved RLL (arginine-leucine-leucine) motif required for inducibility. EPAS1 sequences were less inducible than those of HIF-1α and inducibility was greatly reduced as expression was increased. EPAS1 regulation was similar to HIF-1α, where distinct internal and C-terminal domains modulated protein level and activity (O'Rourke et al. 1999).

5.1.1.3 Physiology and Biochemistry

Like HIF-1α, EPAS1 binds to and activates transcription from a DNA element originally isolated from the *EPO* gene and containing the sequence 5′-GCCCTACGTGCTGTCTCA-3′. EPAS1 protein levels are low under normal conditions and increase in hypoxia, like HIF-1α levels. Transactivation is stimulation of transcription by a transcription factor which binds to DNA and extracts the activation of adjacent proteins. These factors then translocate to the nucleus and *trans*-activate target genes containing the sequence 5′-GCCCTACGTGCTGTCTCA-3′, the hypoxia response element (HRE) (Conrad, 1999). Activation by both HIF-1α and EPAS1 is stimulated by hypoxic conditions. EPAS1 forms a heterodimeric complex (two similar subunits or monomers linked together) with ARNT before transcriptional activation of target genes. EPAS1 is an important regulator of vascularisation, involving the regulation of endothelial cell gene expression in response to hypoxia (Tian et al. 1997).

The ability to sense and respond to varying oxygen levels is essential for survival. Oxygen-sensing systems maintain cell homeostasis and adapt to chronic hypoxia of diseases such as cancer. The major genes and pathways of oxygen sensing normoxia and hypoxia are not completely understood (Figure 5-7 and Figure 5-8) (Giaccia et al. 2004).

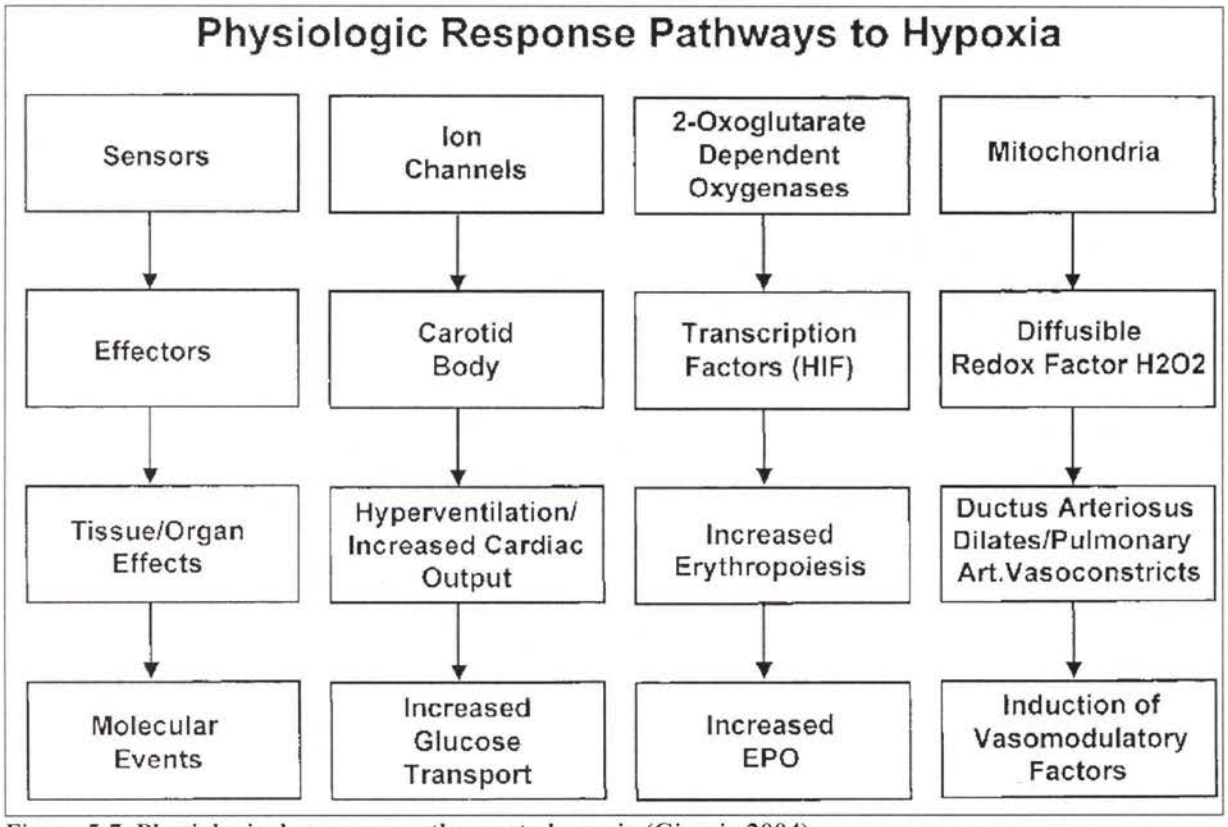


Figure 5-7 Physiological response pathways to hypoxia (Giaccia 2004).

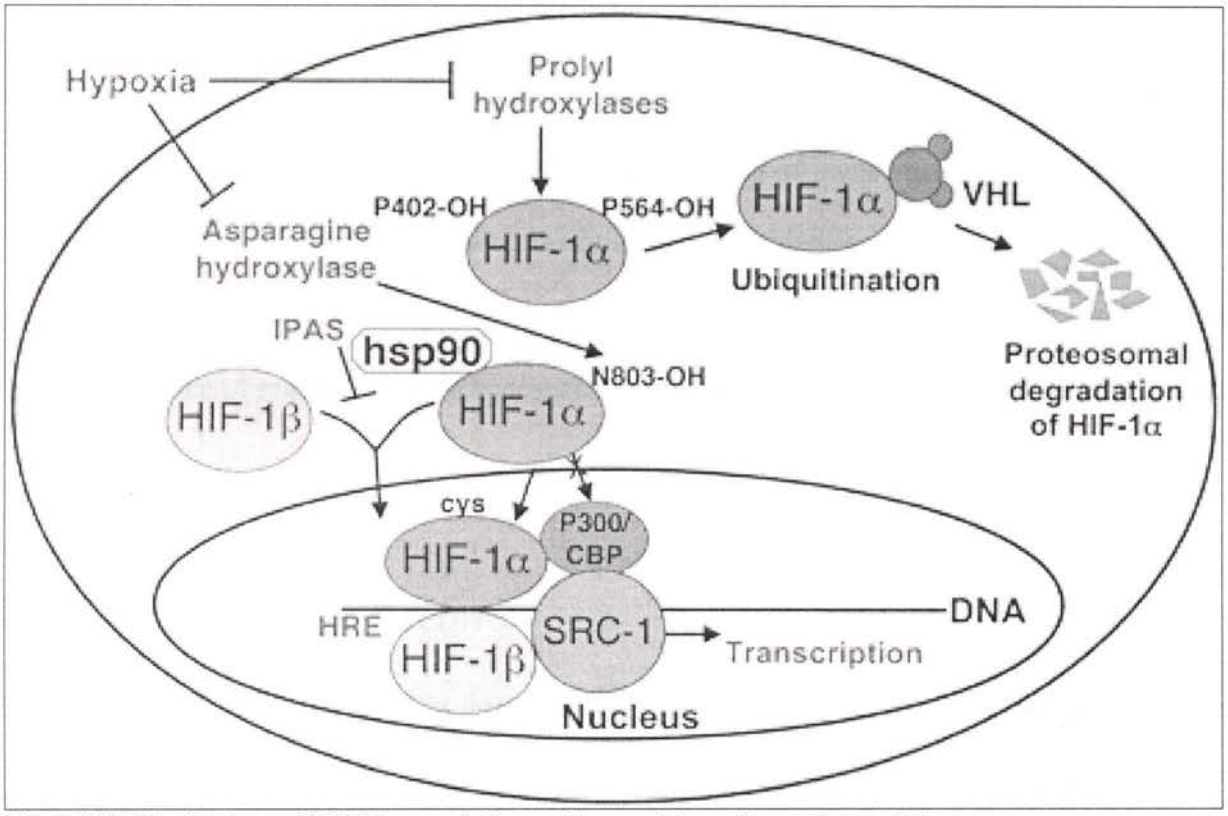


Figure 5-8 Mechanisms of HIF-1α regulation under aerobic and hypoxic conditions.
Under aerobic conditions, HIF-1α is hydroxylated on proline 402 and proline 564. The proline hydroxylations are necessary for binding to VHL and ubiquitin-mediated degradation by the proteasome. The asparagine hydroxylation prevents binding to p300/CBP. A splice derivative of HIF-3α called IPAS, as it only possesses the PAS domain, competes for HIF-1β binding. The TAD of HIF-1α binds p300/CBP and other coactivators such as SRC-1. HIF-1α and HIF-1β both translocate to the nucleus to transactivate genes such as VEGF that possess hypoxia responsive elements (HREs) (Giaccia 2004).

5.1.1.4 Disease States

Hypoxic stress is involved in pathophysiologic states such as myocardial infarction and stroke, as well as in normal development and haematopoiesis. Members of the HIF family act as transcriptional regulators of hypoxic response genes. After generating adult mice that globally lack EPAS1, the second member of the HIF family, characterisation of the haematopoietic cell population indicated that the loss of EPAS1 resulted in pancytopenia. Bone marrow reconstitution experiments of lethally irradiated hosts, detailed the haematopoietic injury in the EPAS1 null mice and suggested a vital role for EPAS1 in the bone marrow for useful haematopoiesis (Scortegagna et al. 2003b). The same research group showed that the mice have a syndrome of multiple-organ pathology, biochemical

abnormalities and altered gene expression. These changes include: retinopathy, hepatic steatosis, cardiac hypertrophy, skeletal myopathy, hypocellular bone marrow, azoospermia, mitochondrial abnormalities, hypoglycemia, lactic acidosis, altered Krebs cycle function, dysregulated fatty acid oxidation, improved production of reactive oxygen species, and decreased expression of genes encoding the primary antioxidant enzymes. EPAS1 was shown to transactivate the primary antioxidant enzymes' promoters and is thought to be involved in the maintenance of reactive oxygen species and mitochondrial homeostasis (Scortegagna et al. 2003a).

5.1.2 Transcription Factors

Transcription factors are the main intermediaries of the genetic programs that control human physiology. Mutations in genes that encode transcription factors or their targets may unfavourably affect gene expression and result in disease. Mutations in genes encoding transcription factors often have pleiotropic effects because some transcription factors are involved in the regulation of multiple genes (Semenza 1994b). Eukaryotic gene expression is controlled by a small number of transcription factors with a wide range of regulatory mechanisms. The molecular basis for the mechanism in which the transcription factor activates or represses transcription depends on the sequence of the DNA to which it binds. The mechanism by which the binding-site sequence regulates the activity of a gene is related to that of other transcription factors (Latchman 2001).

5.1.3 Erythropoietin (EPO)

EPO helps controls O_2 homeostasis by regulating blood O_2-carrying capacity. A hypoxia-inducible enhancer, identified in the *EPO* 3′-flanking sequence, contains binding sites for several transcription factors, including HIF-1. Binding of HIF-1 is required for *EPO*

transcriptional activation in response to hypoxia. There are HIF-1 binding sites in the *EPO* promoter, which could participate in hypoxia-inducible transcription. The hypoxia signal-transduction pathway leading to *EPO* transcriptional activation has not been established. Two hypothetical mechanisms of O_2 sensing are oxy-deoxy conformational changes of a haemoprotein and the production of reactive O_2 species from molecular O_2. The molecular mechanisms by which *EPO* transcription is regulated may also be utilised to control the expression of other genes responsible for cellular and systemic O_2 homeostasis (Semenza 1994a).

EPO production due to hypoxia is almost totally restricted to cells within the liver and kidney, yet the transcriptional enhancer lying 3′ to the *EPO* gene shows activity inducible into a variety of cultured cells. Many cells, which do not produce EPO contain, as a result, a very similar oxygen-regulated control system, suggesting that the same system regulates other genes (Firth et al. 1994). HIF-1 activates *EPO* gene transcription in cells subjected to hypoxia. HIF-1 activity is also induced by hypoxia in non-EPO-producing cells, suggesting a more general regulatory role. HIF-1 acts as a mediator of adaptive responses to hypoxia that underlie cellular and systemic oxygen homeostasis (Semenza et al. 1994).

5.1.4 Hypoxia Inducible Factor-1 (HIF-1)

HIF-1 is a basic HLH transcription factor which is expressed in mammalian cells in hypoxia and which activates transcription of genes encoding EPO, VEGF, and other oxygen homeostasis proteins. There is some evidence that HIF-1 also regulates transcription of glycolytic enzymes genes. HIF-1 activates transcription through these elements but a HIF-1 binding site alone is not enough to mediate transcriptional responses to hypoxia. Functional hypoxia-response elements consist of a pair of contiguous transcription factor binding sites at least one of which contains the core sequence 5′-RCGTG-3′ and are recognised by HIF-1.

The coordinate transcriptional activation of genes encoding glycolytic enzymes in hypoxia is, therefore, mediated by HIF-1 (Semenza et al. 1996).

HIF-1 activates transcription of hypoxia-inducible genes, including those encoding: EPO, VEGF, haeme oxygenase-1, inducible nitric oxide synthase, and the glycolytic enzymes aldolase A, enolase 1, lactate dehydrogenase A, phosphofructokinase I, and phosphoglycerate kinase 1. HIF-1α protein levels were induced *in vivo* when animals were subjected to anaemia or hypoxia. The *HIF-1α* gene was mapped to human chromosome 14q21–q24 and mouse chromosome 12 (Semenza et al. 1997).

HIF-1 DNA-binding activity, HIF-1α protein and HIF-1β protein each increased exponentially as cells were subjected to decreasing O_2 concentrations, with a half maximal response between 1.5 and 2% O_2 and a maximal response at 0.5% O_2. The biggest HIF-1 response was over O_2 concentrations associated with ischemic/hypoxic events *in vivo*. This indicates the involvement of HIF-1 in O_2 homeostasis and represents a functional characterisation of the putative O_2 sensor that initiates hypoxia signal transduction leading to HIF-1 expression (Jiang et al. 1996). A hypoxia-inducible enhancer spanning approximately 50 bp within the 3′-flanking region of the EPO gene is required for transcriptional activation in hypoxic cells. The binding of HIF1 is totally required for enhancer function. Factors binding to the enhancer may interact synergistically with factors binding to the *EPO* promoter to activate transcription in hypoxic cells. Indirect evidence suggests that O_2 tension may be sensed by a haemoprotein (Wang and Semenza 1996).

HIF-1α consists of 15 exons that are interrupted by introns at the same locations as in the mouse Hif1a gene, although sequences mediating alternative splicing and alternative translation initiation events in the mouse are not present in the human gene. Placement of

introns differs between *HIF-1α* and *EPAS1*, which encodes the human HIF-2α protein (Figure 5-9). Transcription of the *HIF-1α* gene was initiated over a 15 nt region downstream of two SP1 binding sites. In transient expression assays, a 0.7 kb region of 5′ flanking sequences operated as a powerful promoter. Comparison of 0.8 kb of 5′ flanking and 5′ untranslated sequences from the human and mouse *HIF-1α* genes revealed 70% identity. The proximal 300 bp of 5′ flanking sequences, including the SP1 binding sites and transcription initiation sites, had 83% identity. These results suggest evolutionary selection for maintenance of HIF-1α structure, function, and regulation (Iyer et al. 1998).

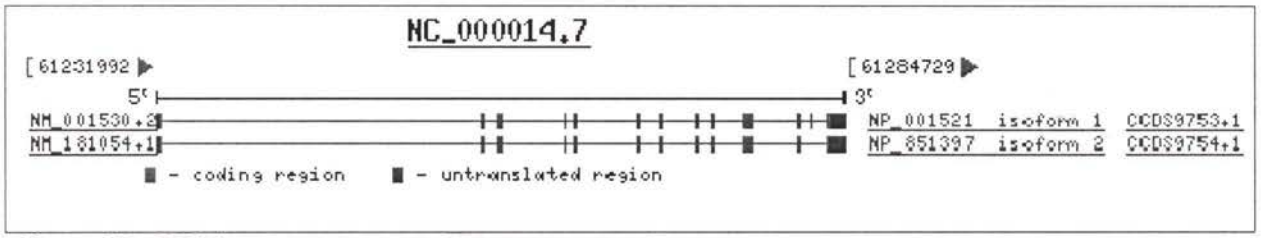


Figure 5-9 *HIF-1α* gene structure
(Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez).

HIF-1α sequence variation was associated with $\dot{V}O_2$ max before and after aerobic exercise training in older humans (Prior et al. 2003). A shorter HIF-1α isoform is three-fold less active than HIF-1α (full length), a result consistent with the lack of the C-terminal transactivation domain (Gothie et al. 2000). Two minimal domains within HIF-1α (amino acids 549–582 and amino acids 775–826) were defined by deletional analysis, each of which could independently convey inducible responses (Figure 5-10) (Pugh et al. 1997).

```
        10          20          30          40          50          60
MEGAGGANDK KKISSERRKE KSRDAARSRR SKESEVFYEL AHQLPLPHNV SSHLDKASVM
        70          80          90         100         110         120
RLTISYLRVR KLLDAGDLDI EDDMKAQMNC FYLKALDGFV MVLTDDGDMI YISDNVNKYM
       130         140         150         160         170         180
GLTQFELTGH SVFDFTHPCD HEEMREMLTH RNGLVKKGKE QNTQRSFFLR MKCTLTSRGR
       190         200         210         220         230         240
TMNIKSATWK VLHCTGHIHV YDTNSNQPQC GYKKPPMTCL VLICEPIPHP SNIEIPLDSK
       250         260         270         280         290         300
TFLSRHSLDM KFSYCDERIT ELMGYEPEEL LGRSIYEYYH ALDSDHLTKT HHDMFTKGQV
       310         320         330         340         350         360
TTGQYRMLAK RGGYVWVETQ ATVIYNTKNS QPQCIVCVNY VVSGIIQHDL IFSLQQTECV
       370         380         390         400         410         420
LKPVESSDMK MTQLFTKVES EDTSSLFDKL KKEPDALTLL APAAGDTIIS LDFGSNDTET
       430         440         450         460         470         480
DDQQLEEVPL YNDVMLPSPN EKLQNINLAM SPLPTAETPK PLRSSADPAL NQEVALKLEP
       490         500         510         520         530         540
NPESLELSFT MPQIQDQTPS PSDGSTRQSS PEPNSPSEYC FYVDSDMVNE FKLELVEKLF
       550         560         570         580         590         600
AEDTEAKNPF STQDTDLDLE MLAPYIPMDD DFQLRSFDQL SPLESSSASP ESASPQSTVT
       610         620         630         640         650         660
VFQQTQIQEP TANATTTTAT TDELKTVTKD RMEDIKILIA SPSPTHIHKE TTSATSSPYR
       670         680         690         700         710         720
DTQSRTASPN RAGKGVIEQT EKSHPRSPNV LSVALSQRTT VPEEELNPKI LALQNAQRKR
       730         740         750         760         770         780
KMEHDGSLFQ AVGIGTLLQQ PDDHAATTSL SWKRVKGCKS SEQNGMEQKT IILIPSDLAC
       790         800         810         820
RLLGQSMDES GLPQLTSYDC EVNAPIQGSR NLLQGEELLR ALDQVN
```

Figure 5-10 HIF-1α protein sequence
(UniProtKB/Swiss-Prot: http://ca.expasy.org/uniprot/Q16665). Length: 826 AA, molecular weight: 92,670 Da.

Sequence alignment of EPAS1 Exon 9 and HIF-1α Protein sequence reveals 45% identity (Figure 5-11).

```
EPAS1 [Homo sapiens] and HIF-1α [Homo sapiens]
Identities = 33/73 (45%), Positives = 44/73 (60%), Gaps = 7/73 (9%)

EPAS1   2     IEKNDVVFSMDQTESLFKP---HLMAMNSIFDSSGKGAVSEKSNFLFTKLKEEPEELAQL
58
              I ++D++FS+ QTE + KP     M M  +F        SE ++ LF KLK+EP+ L  L
HIF-1α  345   IIQHDLIFSLQQTECVLKPVESSDMKMTQLFTK----VESEDTSSLFDKLKKEPDALTLL
400

Query   59    APTPGDAIISLDF  71
              AP  GD IISLDF
Sbjct   401   APAAGDTIISLDF  413
```

Figure 5-11 EPAS1 Exon 9 and HIF-1α Protein sequence alignment
(NCBI BLAST: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi).

Interestingly, *HIF-1α* is located at 14q21–25 which happens to cover a QTL identified by the maximal oxygen uptake genome-wide scan (Bouchard et al. 2000) (Figure 5-12). There are three conserved domains in HIF-1α; one HLH and two PAS domains (Figure 5-13).

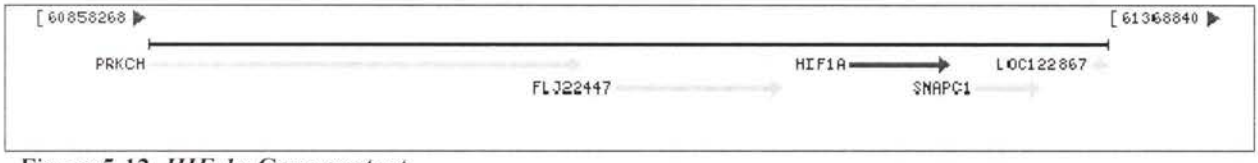


Figure 5-12 *HIF-1α* Gene context
(Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez).

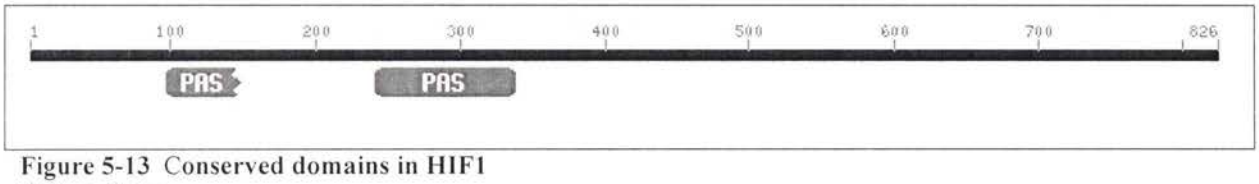


Figure 5-13 Conserved domains in HIF1
(Entrez Gene:
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?INPUT_TYPE=precalc&SEQUENCE=40254439).

5.1.5 Aryl Hydrocarbon Receptor Nuclear Translocator (*ARNT*)

Aryl hydrocarbon receptor nuclear translocator (*ARNT*) (also known as hypoxia inducible factor 1β; *HIF-1β*) gene contains 22 exons, varying in size from 25 to 214 bp, spans 65 kb and has 789 amino acids. The GT/AG consensus is followed by splice junctions except for intron 11 starting with GC at its 5′ end (OMIM).

5.1.6 Vascular Endothelial Growth Factor (*VEGF*)

High intensity training in hypoxia produces an increase of vascular endothelial growth factor (VEGF) mRNA, capillarity and myoglobin mRNA (Hoppeler 2001). VEGFs are a family of secreted polypeptides with a well conserved receptor-binding cystine-knot structure. The founding member of the family, VEGF-A, is highly conserved between animals. VEGFs operate through a family of cognate receptor kinases in endothelial cells to stimulate blood-vessel formation in vertebrates. VEGF-A has key functions in mammalian vascular

development and in abnormal blood vessel diseases; other VEGFs have roles in lymphatic vessels and disease-related angiogenesis. This family of growth factors appear very early during evolution because VEGF-like molecules and their receptors are in simple invertebrates without a vascular system (Holmes and Zachary 2005).

5.1.7 Aims

5.1.7.1 SNPs

The SNP data was analysed using Haplo.Stat software to produce haplotypes for the various athlete and control groups. These proportions of the different haplotypes were compared between the various athlete groups for significant differences.

5.1.7.2 DHPLC

DHPLC was used to screen DNA samples for DNA base variants. These changes, if present in heterozygous form, showed up as different DHPLC patterns. The different patterns were grouped and analysed for statistically different frequencies between cases and controls. Correlations between genotype and phenotype were investigated in the athlete groups.

5.2 Materials and Methods

5.2.1 SNPs

SNPs are single base pair changes, which are scattered regularly throughout the genome. SNPs are now thought to be crucial to further gene discovery and gene characterisation in genetics and genomics (Figure 5-14).

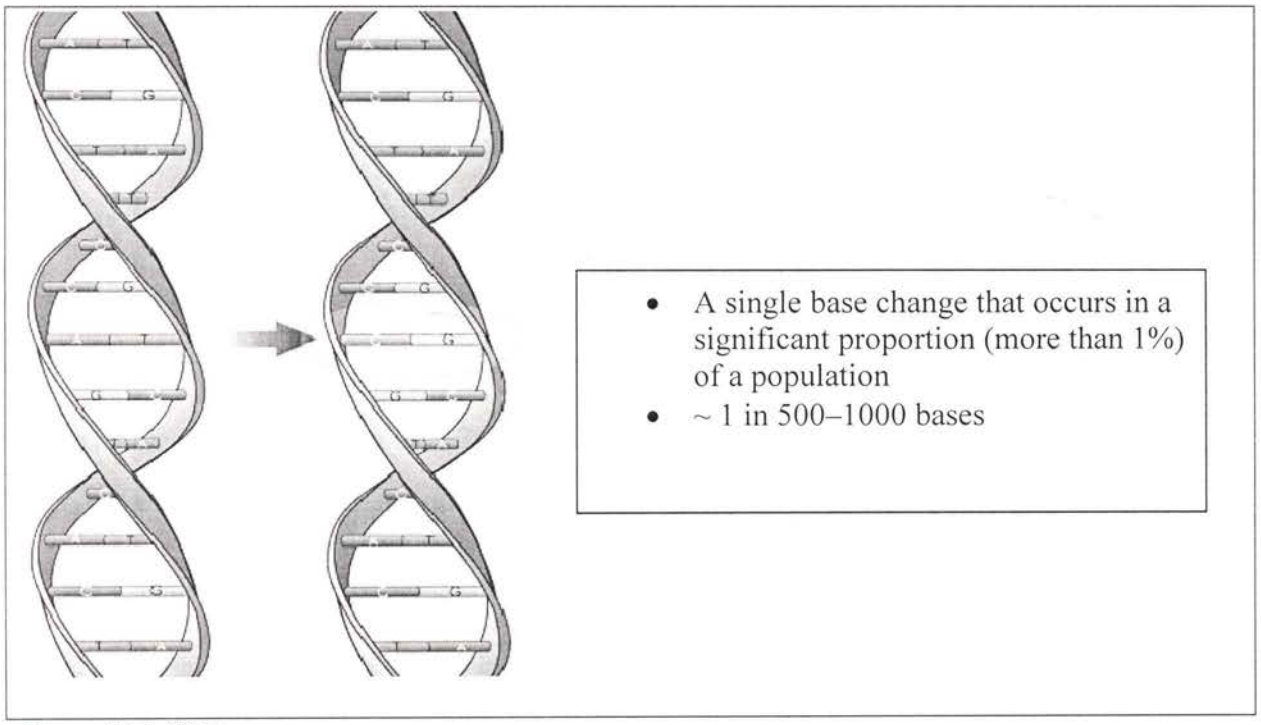


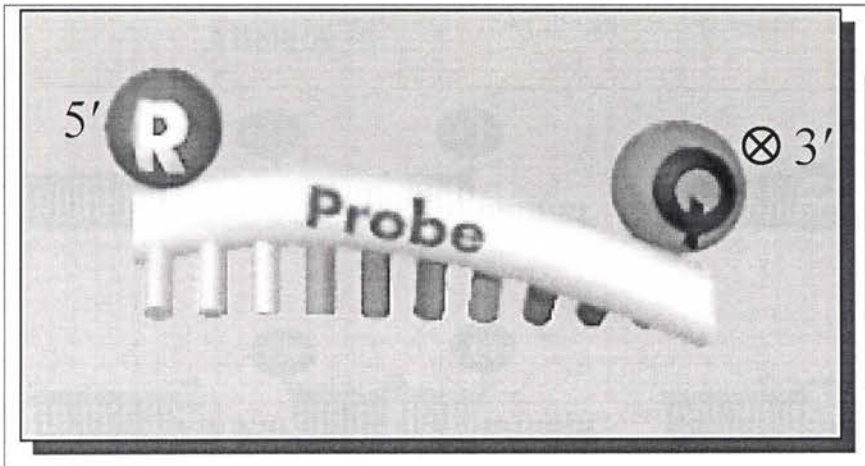
Figure 5-14 SNPs.

5.2.1.1 Beckman Coulter Biomek® FX robotic station

The Beckman Coulter Biomek® FX robotic station was used for setting up SNP reactions. This enabled small reaction volumes to be used, making it more economical to run and required less DNA from precious elite athlete samples. It also enabled high-throughput setups which are often needed for studies involving genes that may have a small but significant effect.

5.2.1.2 Applied Biosystems Prism® 7900HT Sequence Detection System

The Applied Biosystems Prism® 7900HT Sequence Detection System was used by Sydney University / Prince Alfred Macromolecular Analysis Centre (SUPAMAC) for the SNP assaying (Ranade et al. 2001). It is a high throughput, real time PCR instrument designed for automated, high-throughput detection of fluorescent PCR products. The fluorescent chemistry used for PCR product detection in the 7900HT was the sequence specific (Taqman®) probes. Taqman® probes are based on the 5′ nuclease allelic discrimination assay which uses the 5′-3′ exonuclease activity of *Taq* polymerase to cleave a sequence specific probe which has been designed to hybridise to the sequence that is flanked by the forward and reverse primers (Figure 5-15, Figure 5-16 and Figure 5-17). The Taqman® probe is a short oligodeoxynucleotide labelled with a reporter dye at the 5′ end and a quencher dye at the 3′ end of the probe. When the probe is intact, the reporter and quencher dyes are close causing suppression of the reporter dye fluorescence via energy transfer. If the target sequence is present during PCR, the probe anneals to its complement sequence. As the primers hybridise to their complement sequence, the *Taq* polymerase extends from the primers and 5′-3′ nucleolytic activity cleaves the probe, causing the reporter and quencher dyes to separate and increasing reporter dye emission intensity. The probe fragments are removed from the target sequence, and polymerisation of the strand continues. The 3′ end of the probe is phosphorylated to prevent it acting as primer and prevents *Taq* polymerase extension during PCR. The 7900HT uses an argon-ion laser for the excitation of the PCR chemistries and a CCD camera measures the fluorescence spectrum and intensity from each reaction. Minimum reaction volumes of 5μL can be used on the 7900HT (SUPAMAC 2005b).



- Hybridises with the target amplicon
- Has two fluorescent dyes attached:
  1. Reporter (**R**)
  2. Quencher (**Q**)

Figure 5-15 Taqman® probe (SUPAMAC).

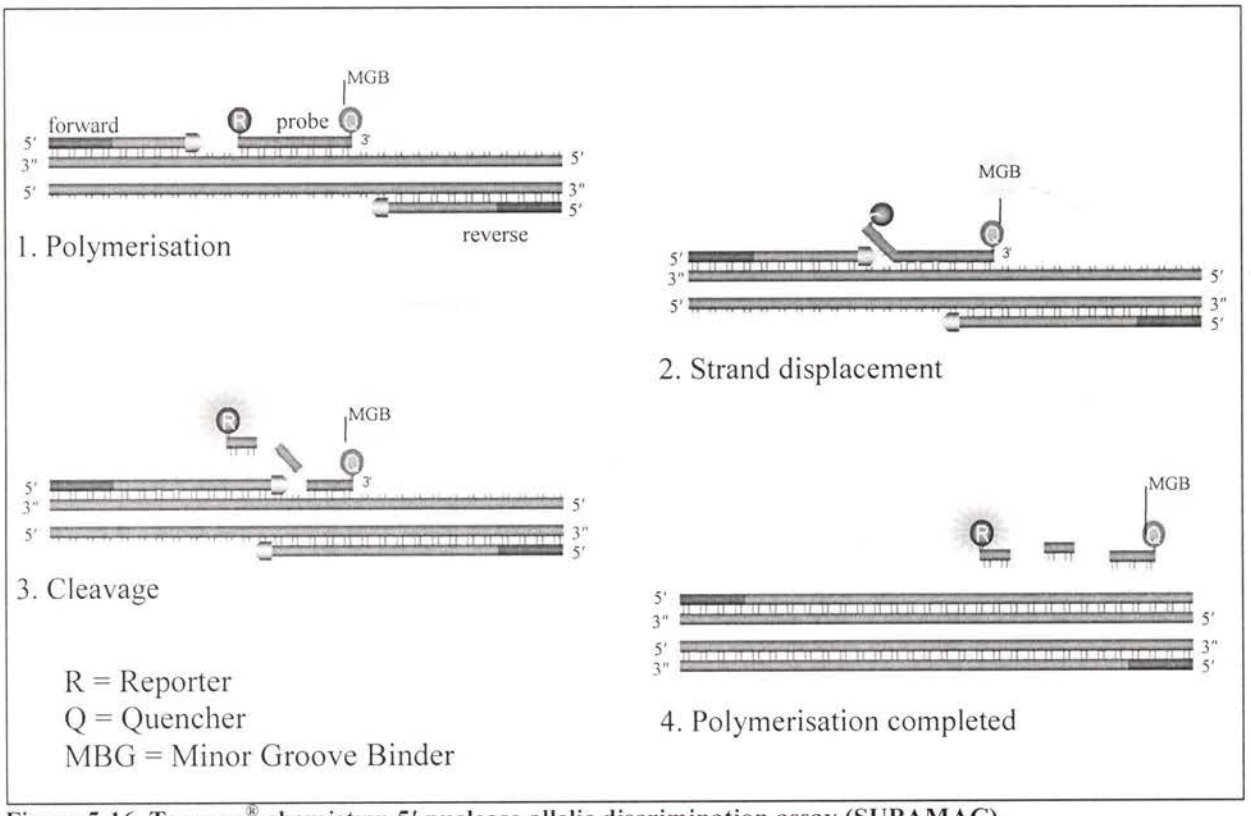


Figure 5-16 Taqman® chemistry: 5′ nuclease allelic discrimination assay (SUPAMAC).

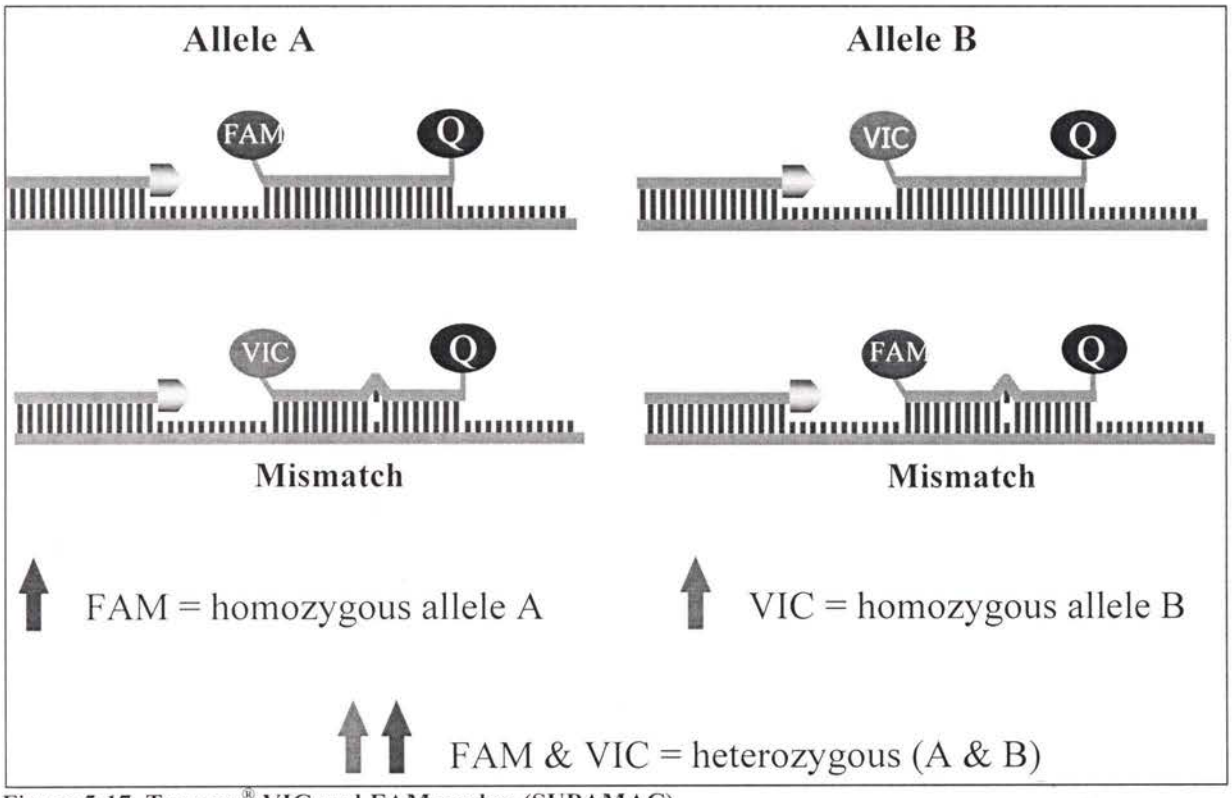


Figure 5-17 Taqman® VIC and FAM probes (SUPAMAC).

5.2.1.3 SNP Genotyping

The materials used for the SNP genotyping were sourced from Applied Biosystems TaqMan® SNP Genotyping Assay products. SUPAMAC performed the SNP genotyping using the TaqMan® SNP Genotyping Assay which consists of two primers for amplifying the target sequence and two TaqMan® Minor Groove Binder probes for detecting the target alleles (Table 5-3).

Each TaqMan® Minor Groove Binder probe contains:

- a reporter dye at the 5′ end of the probe
    - VIC dye for Allele 1 and FAM dye for Allele 2.
- A Minor Groove Binder to increase the melting temperature without increasing probe length, which increases accuracy.
- A non-fluorescent quencher which also increases accuracy.

Table 5-3. PCR conditions for TaqMan® SNP Genotyping Assay.

2× TaqMan® Universal Master Mix	2.50 µL
20× SNP Genotyping Assay Mix	0.25 µL
DNA (10 ng.µL^{-1})	2.25 µL
Total volume	**5.00 µL**
Thermal Cycling Conditions	95°C × 10 min × 1 (92°C × 15 s; 60°C × 60 s) × 40

5.2.1.4 SNP High-throughput Quality Control

Quality control is important in all genetic studies. Measures that are taken include: careful labelling of blood or buccal cell samples with unique identifiers, as soon as they arrive at the laboratory, if not before; having systems whereby when DNA samples are being extracted, or dilutions are being made, that samples are carefully lined up and checked-off so that samples are not mixed-up; and all PCR and restriction enzyme reactions include non-template and known controls for the known polymorphisms. An additional problem in the high-throughput genotyping is that the 96-well plates can be inadvertently placed backwards on the Beckman Coulter Biomek® FX robotic station platform, i.e. Sample 1 receives the result for Sample 96. To exclude, therefore, this source of error, non-template controls were used on each 396-well plate and a few samples were randomly repeated on each 96-well plate for every SNP.

5.2.2 DHPLC

DHPLC compares two or more DNA templates by using a mixture of denatured and reannealed PCR amplicons. This shows the presence of a change in DNA sequence by the differences between the homoduplex and heteroduplex DNA on reversed-phase chromatography. This is achieved using partial denaturation (Xiao and Oefner 2001). The optimum melting profile at a particular temperature (predicted by calculation) determines the

sensitivity. SNPs, deletions, and insertions can be detected by on-line UV or fluorescence monitoring within 2–3 min in unpurified amplicons 150–700 bp in size, although technically they can be as large as 1.5 kb. Sensitivity and specificity of DHPLC consistently exceed 96%. For short amplicons, under completely denaturing conditions, DHPLC can be used for the genotyping of known polymorphisms by utilising the ability of poly(styrene-divinylbenzene) to resolve single-stranded DNA molecules of identical size that differ in a single base, and it can be used to find all base substitutions except for C→G transversions.

5.2.2.1 WAVE® System DHPLC

The WAVE® Nucleic Acid Fragment Analysis System (Model 3500 HT; Transgenomic) using the Navigator® software (v1.5) is an automated system for nucleic acid fragment analysis (Figure 5-18). It was used in its DNA variant detection mode to screen *EPAS1* Exons 9–16 for functional variants relevant to performance.

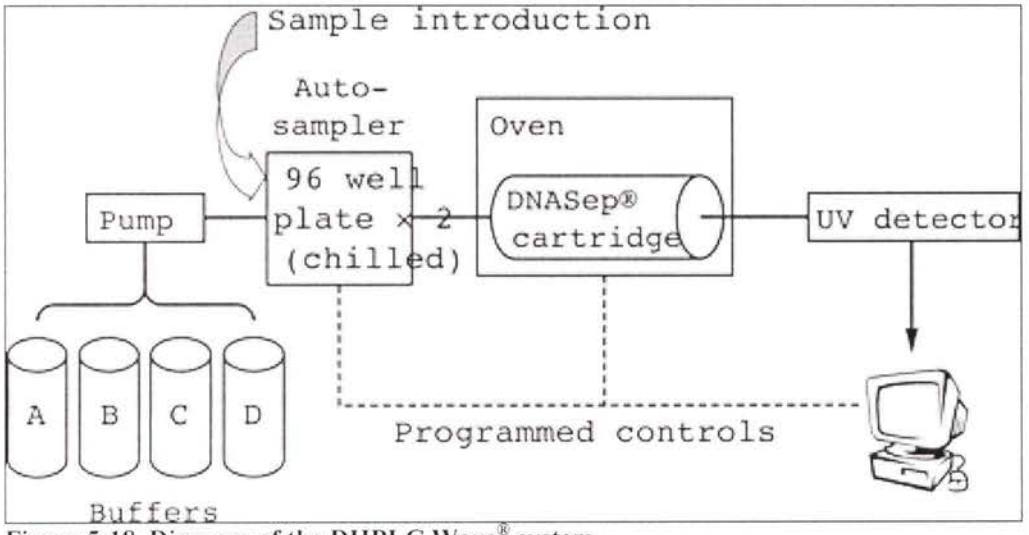


Figure 5-18 Diagram of the DHPLC Wave® system.
Buffers are pumped into the system individually or in a mixture through a multivalve mechanism. A sample is introduced via the autosampler, which holds 2 x 96 well chilled plates. The sample is eluted from the DNASep® Cartridge by linear gradient changes of buffers A and B (mobile phase) within an oven that has a precisely controlled temperature, and then travels through an ultraviolet detector. The amount of sample DNA and the elution time are automatically recorded.

WAVE® System uses a DNASep Prep HT column which is filled with a non-porous matrix consisting of polystyrene-divinylbenzene copolymer beads. The beads are alkylated with C-18 chains that form single C-C bonds and are electrostatically neutral and hydrophobic (do not readily react with DNA). Triethylammonium acetate buffer is an ion-pairing reagent which binds the DNA to the beads. The positively charged triethylammonium acetate part bonds to the negatively charged DNA backbone, while the hydrophobic triethylammonium acetate part interacts with the hydrophobic C-18 chains on the polystyrene-divinylbenzene beads forming a bridge between the DNA and the cartridge matrix. A 0.1 M solution of triethylammonium acetate mixed with 25% acetonitrile is used as the elution buffer. As acetonitrile flows across the cartridge matrix, the hydrophobic interaction between the column and the DNA/ triethylammonium acetate is broken and the DNA elutes (SUPAMAC 2005a).

5.2.2.2 Amplicon Design

The PCR primers were designed using OLIGO® 5.0 software (www.oligo.net). The target amplicon sequence, including 30–50 bp on either side, was copied and pasted into the software and its output suggested forward and reverse primers. This output included: primer length, position, GC content, amplicon length, melting and annealing temperature.

Amplicon design was optimised prior to PCR amplification of the exon. The most suitable primers were selected based on the following criteria:

- The amplicon must cover the intron/exon boundary and primers should be no closer than 30–50 bp to the end of the sequence.
- The amplicon must not include too much of the intron. This reduces the number of positive results caused by intron polymorphisms.
- Primers should be 18–25 bp long and should contain 45–60% GC.

- The melting temperature difference between primers should be less than 2°C.
- Amplicons should be from 150–700 bp long. Amplicons shorter than 150 bp tend to result in wide, flat peaks, and DNA variants in these fragments may remain undetected.

5.2.2.3 PCR Optimisation

The PCR was optimised so it was efficient (strong amplification) and highly specific (only one band is seen on an acrylamide gel). The annealing temperature, Mg^{2+} concentration, cycling conditions and template/primer concentrations were adjusted to optimise the PCR reaction. Although there is a protocol to follow for optimisation, optimising many reactions followed the 80/20 Rule: 80% of the PCRs were optimised in 20% of the time.

The initial conditions were tried for all eight exons in a 50 µL reaction (Table 5-4). The next step was to increase the annealing temperature for all the exons. This resulted in exons 9 and 15 being optimised. Then a $MgCl_2$ titration was performed with the remaining exons. This resulted in exon 13 being optimised. Subsequently, a DNA template titration was performed with negative results. After that the annealing temperature was increased again to 57°C, 58°C, 60°C and 62°C and deleted the extension cycle resulting in exon 14 being optimised at 62°C. Then the number of cycles was decreased to 28 and exon 10 was optimised. With the remaining three exons the annealing temperature, a $MgCl_2$ titration, primer titration, "touchdown" PCR program, AmpliTaq™ titration, annealing temperature, "touchdown" PCR program, primer titration and annealing temperature were sequentially varied. AmpliTaq™ was replaced with AmpliTaq™ Gold and more adjustments were made until exons 12, 11 and 16 were optimised.

Table 5-4 Initial PCR conditions for 8 amplicons.

	Initial	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	Exon 14	Exon 15	Exon 16
dH_2O	33.8 µL	33.8 µL	37.3 µL	31.8 µL	34.8 µL	35.3 µL	30.8 µL	33.8 µL	33.9 µL
10× PCR (buffer II)	5.0 µL	5.0 µL	5.0 µL	5.0 µL	5.0 µL	5.0 µL	5.0 µL	5.0 µL	5.0 µL
$MgCl_2$ (25 mM)	3.0 µL	3.0 µL	1.5 µL	6.0 µL	2.5 µL	1.5 µL	6.0 µL	3.0 µL	3.5 µL
dNTP (2.5 mM)	2.0 µL	2.0 µL	2.0 µL	2.0 µL	2.0 µL	2.0 µL	2.0 µL	2.0 µL	2.0 µL
P1 (20 $pmol.\mu L^{-1}$)	1.0 µL	1.0 µL	1.0 µL	0.5 µL	0.75 µL	1.0 µL	1.0 µL	1.0 µL	0.7 µL
P2 (20 $pmol.\mu L^{-1}$)	1.0 µL	1.0 µL	1.0 µL	0.5 µL	0.75 µL	1.0 µL	1.0 µL	1.0 µL	0.7 µL
AmpliTaq™ (5 $U.\mu L^{-1}$)	0.2 µL	0.2 µL	0.2 µL	**0.2 µL	**0.2 µL	0.2 µL	0.2 µL	0.2 µL	**0.2 µL
DNA (25 $ng.\mu L^{-1}$)	4.0 µL	4.0 µL	2.0 µL	4.0 µL	4.0 µL	4.0 µL	4.0 µL	4.0 µL	4.0 µL
Total volume	**50.0 µL**	**50.0 µL**	**50.0 µL**	**50.0 µL**	**50.0 µL**	**50.0 µL**	**50.0 µL**	**50.0 µL**	**50.0 µL**
Thermal Cycling Conditions	94°C × 2 min × 1 (94°C × 30 s, 55°C* × 30 s, 72°C × 10 s) × 30 72°C × 7 min × 1	94°C × 2 min × 1 (94°C × 30 s 56°C × 30 s 72°C × 10 s) × 30 72°C × 7 min × 1	94°C × 2 min × 1 (94°C × 30 s, 61°C × 30 s) × 28 72°C × 7 min × 1	95°C × 12 min × 1 (95°C × 30 s, 63°C × 30 s) × 28 72°C × 7 min × 1	95°C × 12 min × 1 (95°C × 30 s, 59°C × 30 s, 72°C × 10 s) × 32 72°C × 7 min × 1	94°C × 2 min × 1 (94°C × 30 s, 56°C × 30 s, 72°C × 10 s) × 30 72°C × 7 min × 1	94°C × 2 min × 1 (94°C × 30 s, 67°C × 30 s) × 30 72°C × 7 min × 1	94°C × 2 min × 1 (94°C × 30 s, 56°C × 30 s, 72°C × 10 s) × 30 72°C × 7 min × 1	95°C × 12 min × 1 (95°C × 30 s, 58°C × 30 s) × 30 72°C × 7min × 1

(*annealing temperature is 3–5°C lower than the average of forward and reverse primer melting temperatures; ** AmpliTaq™ Gold used)

5.2.2.4 DHPLC Application and Melting Profile

This system can operate in three modes (non-denaturing, partially denaturing and fully denaturing). Partially denaturing conditions (52–75°C; size dependant, sequence dependant separation) are used for:

- DNA variant detection (which is what was used in this chapter) (Table 5-5)
- SNP discovery

Table 5-5 Primer sequences and temperatures for 8 amplicons.

Exon	Primer sequences	Amplicon Length	Temperatures
9	Upper 5′-AAA TGT GGA AAG TCT GAA TGG-3′ Lower 5′-GGG AGG CCT GTT ATA GTA AG-3′	331 bp	60.2°C, 62.2°C
10	Upper 5′-CGA TGG TTG TGG GTG TTC AC-3′ Lower 5′-CCA TCT GGA AGC AGT CAT ACA-3′	301 bp	62.0°C, 65.5°C
11	Upper 5′-GGG AGC AGG CAC ACA CCC TAG-3′ Lower 5′-GAC AGC AGG CAC ACA CCC TAG-3′	206 bp	60.5°C
12	Upper 5′-GTT TGT GAG GTC GTA CCA A-3′ Lower 5′-CTG CTG GTA CAG CTG AGT ATC-3′	577 bp	61.2°C, 62.7°C
13	Upper 5′-GAC TGG AAG GGA CCC TAA GA-3′ Lower 5′-GCT ATG GAG AGG CCC CTG TG-3′	251 bp	63.3°C
14	Upper 5′-CCA TTT CCC CTT TCC ATC T-3′ Lower 5′-CTC AGA GAG GCA GGT ACG G-3′	262 bp	62.3°C
15	Upper 5′-CCT CAG GAA AAT GCT ACC GTC-3′ Lower 5′-CTT CCC CTG GCA TCG AAT CC-3′	272 bp	59.8°C, 63.8°C
16	Upper 5′-GAT TTA GGC CTT TAA GTT ATG-3′ Lower 5′-AGT GTG CTG GCG TTA GA-3′	293 bp	63.2°C, 64.8°C

The Navigator® software predicts the analysis conditions for DNA variant detection. Under partially denaturing conditions, fragments are separated based on sequence and analysis temperature. If the sequence of the gene or fragment to be analysed is known, sequences can be copied into the Navigator® software, and the melting profile of the fragment predicted. A well-designed amplicon should ideally have a single melt domain and the entire fragment should melt within a temperature interval of 5°C. This allows fragments to be analysed at a minimum number of temperatures. If a fragment had multiple domains, two temperatures were used to cover the whole fragment. If the segment of the amplicon that has a different melt domain was just part of the intron, only the temperature which covers the exon was used.

A DNA sample with a single DHPLC peak was used as a control to differentiate the homozygous DNA samples. The control was mixed with unknown homozygous samples to induce heteroduplex formation. Homozygous polymorphic variants gave heteroduplexes by mixing with the known fragment of the same amplified region. All samples were denatured and slowly reannealed (Table 5-6).

Table 5-6 Homoduplex- and Heteroduplex-formation Thermal Cycling Conditions:

95°C × 5 min	× 1 cycle
95–30°C × 45 s	× 1 cycle

5.2.2.5 Haplotypes

The Haplotyper program (Niu et al. 2002) produced 18 different haplotypes from the first 386-well plate of ironman, rower, cyclist and control DNA for the *EPAS1* SNPs. DNA samples representing these 18 different haplotypes (seven ironmen, five cyclists and six normal controls) were run through DHPLC to scan for DNA variants (Figure 5-19). Exons 9 to 16 were scanned with DHPLC for variants. The variants were identified by having different waveforms from the wild type sample for each exon. The samples were categorised into the different waveforms and were sequenced by SUPAMAC laboratory to find the variant/s causing the modified waveforms.

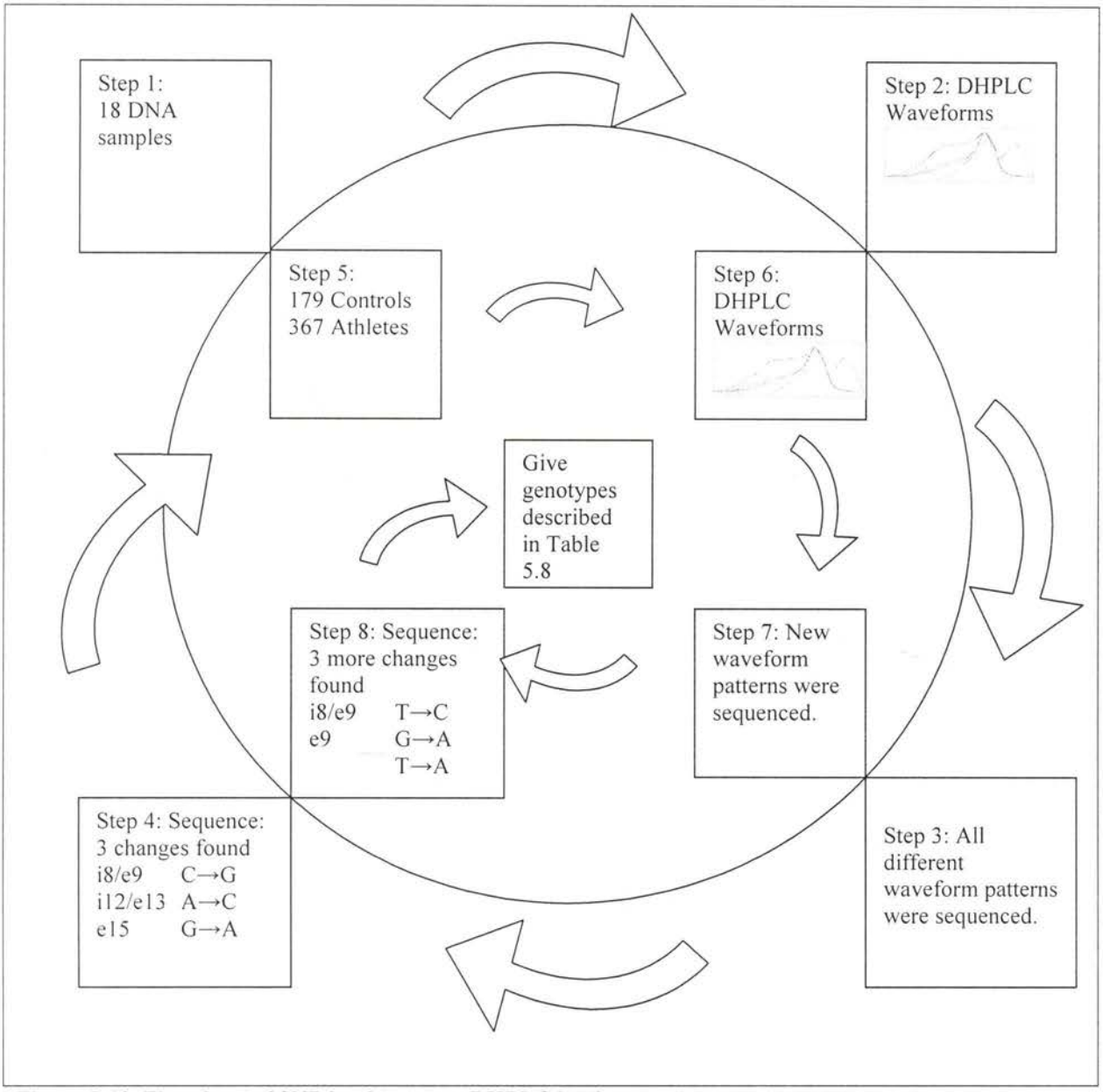


Figure 5-19 Flowchart of SNP haplotypes to DHPLC haplotypes.

5.2.2.6 DNA Sequencing

From each waveform pattern, an example was prepared for sequencing to cover all exons. DNA samples were purified to eliminate the nucleotides and the primers using Marligen Purification Kit. The optical density and concentration of purified DNA were measured. A sample of the purified DNA was checked with PCR and gel electrophoresis before sequencing. The Dye-Terminator chemistry of PCR sequencing uses labelled ddNTPs. DNA and primers were supplied to SUPAMAC pre-mixed in PCR tubes in strips of 8 (Interpath

Services) (Table 5-7). SUPAMAC used the Applied Biosystems Prism® 3700 Sequence Detection System.

Table 5-7 Dye-Terminator reaction mix.

Reagent	PCR products
Primer	5–10 pmol
DNA	50–100 ng per 300 bp
dH_2O	As required
Total volume	**16.0 µL**

The variants found in the sequence results were further investigated for their functional significance (Figure 5-20 and Figure 5-21). Larger groups of athletes and controls were scanned with DHPLC to look for statistically significant differences between them for the variant waveforms. Any new waveforms discovered with the larger sample sizes used, were sequenced and their functional significance was investigated.

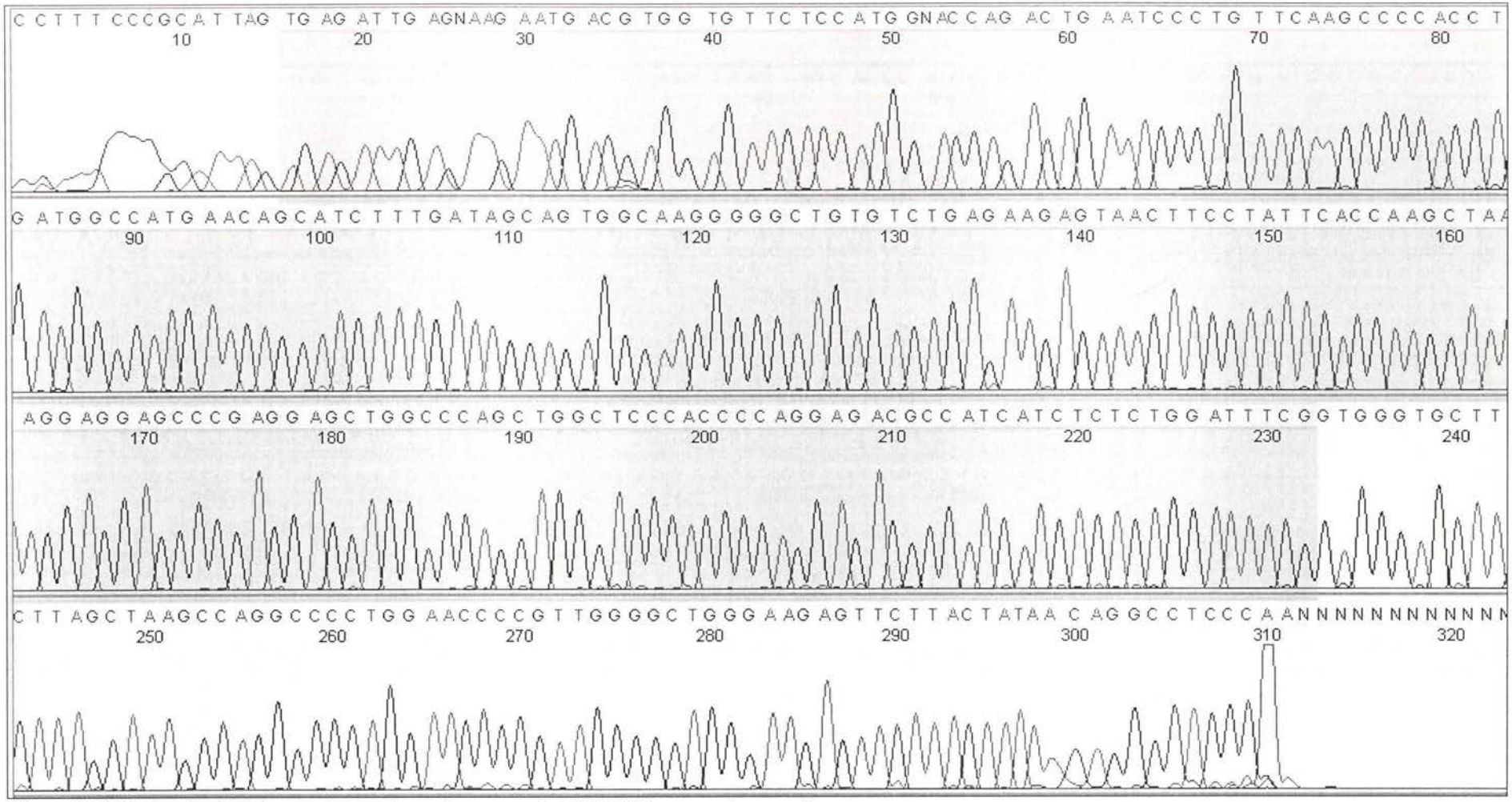


Figure 5-20 Example of e9 forward strand amplicon sample FHC1579 sequencing results.
(SUPAMAC) Exon highlighted; note the C→G change at position nt 9.

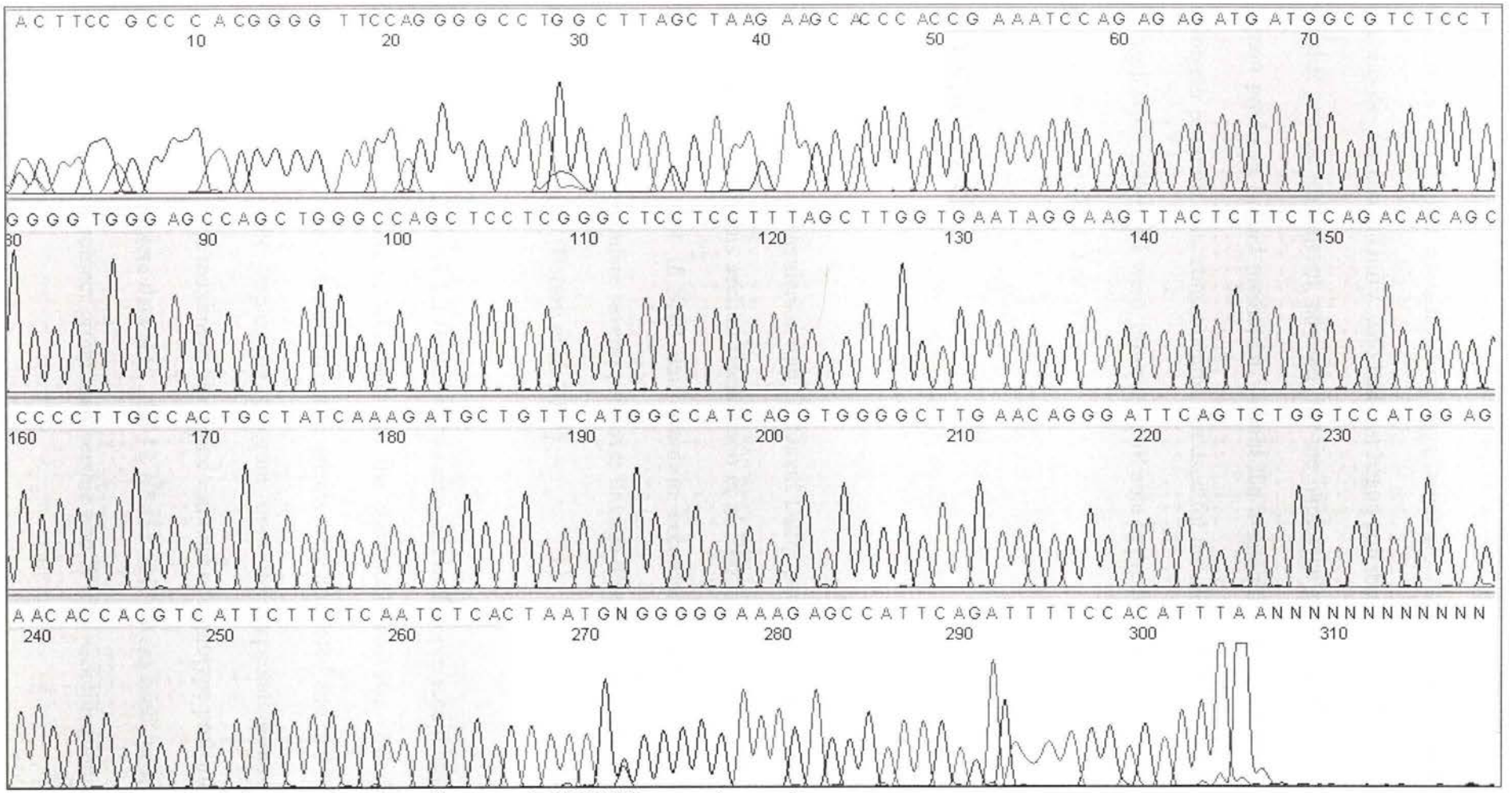


Figure 5-21 Example of e9 reverse strand amplicon sample FHC1579 sequencing results.
(SUPAMAC) Exon highlighted; note the G→C change at position nt 272.

The first 386-well plate contained four large target groups: ironmen, rowers, cyclists and controls. Dr David Duffy, who had just begun collaborating on this project as the specialist genetic statistician, advised that the haplotype prediction from Haplotyper program produces biased prediction without the availability of family studies. The Haplotyper program was, consequently, not used for the main SNP statistical analysis. The haplotypes it defined were, however, still able to produce a variety of valid DNA variants through DHPLC analysis.

5.2.3 Statistics

5.2.3.1 SNPs

Dr Bing Yu, Jennifer Henderson and Dr David Duffy performed the SNP statistical analysis described in this section (Henderson et al. 2005). The Haploview program v3.0 was used to test if SNP data were in accordance with Hardy-Weinberg Equilibrium and to calculate several pair-wise linkage disequilibrium measures in the *EPAS1* block structures (Barrett et al. 2005).

The haplo.stats program v1.1.1 (http://www.mayo.edu/hsr/people/schaid.html), which runs in the R-environment, was used for the statistical analysis. For haplotype association study, it was assumed that all subjects were unrelated and that haplotypes were ambiguous. The haplo.stats program uses the expectation-maximisation algorithm to compute the maximum likelihood estimate of haplotype probabilities that can be used for incomplete data sets. All 12 *EPAS1* SNPs were used for the ordinal regression analysis of subject group membership and SNPs. Combinations of SNPs

were analysed in the regression model for their contribution to significance. Between-groups binomial regression analysis showed which group contributed most to the statistical significance and identified the within-group haplotypes (Henderson et al. 2005).

5.2.3.2 DHPLC

The Haplotyper program (Niu et al. 2002) uses a Partition-Ligation algorithm to reconstruct individual haplotypes from population genotype data. It generated the haplotypes that were used in the initial stage to select representative DNA samples for the DHPLC analysis. Initially, the MENDEL program v4.1.1 (Goradia et al. 1992) was used for the case/control analysis of the haplotypes generated by Haplotyper program. It was discarded later, however, for the haplotype analysis since the input of the haplotypes from the Haplotyper program did not provide the posterior probability. As a result, the MENDEL program tends to exaggerate the differences between the case and control group because the program takes all the haplotypes as definite instead of estimated ones.

The DHPLC results from the Wave® system's Navigator® software (v1.5) were analysed using the CLUMP program. The CLUMP program was used for the χ^2 test for data where there were small allele or genotype frequencies. The CLUMP program was run through the DOS program and uses Monte Carlo simulation to perform the χ^2 test.

The SPSS™ program was used for the regression modelling analysis. The linear regression was used to compare genotype to phenotype data. Kevin McGeechan

(Epidemiology and Biostatistics, University of Sydney) assisted with the preparation of the statistical results using the SPSS™ program.

5.2.4 Candidate Gene *In Silico* Search

The *in silico* search was performed using the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) website. Using the genome-wide scan paper of Bouchard as a starting point, Dr Bing Yu selected the QTL on 2p. He searched the NCBI Human Genome Resources to examine all genes within 5 Mb upstream and downstream of the QTL marker D2S2739. A spreadsheet was made of all these genes and they were included as possible candidate genes if they were expressed in the heart, not housekeeping genes, and possibly involved in athletic performance.

Dr Bing Yu produced a shortlist of four candidate genes and these were examined in more depth for functional importance in the cardiovascular system of athletes (Table 5-8). A search of the literature showed that *EPAS1* gene may be involved in hypoxia response. The author confirmed to Dr Bing Yu the importance of hypoxia response and how achieving a temporary state of local muscle hypoxia was one of the main aims of cardiovascular exercise training. At the cellular level at $\dot{V}O_2$ max there is depletion of O_2 similar to death or complete anoxia (Duhaylongsod et al. 1993). Discussions with various other people led to Dr Bing Yu selecting the *EPAS1* gene to be investigated.

Table 5-8 Four candidate gene shortlist.

Gene name	Gene symbol	Alternative name	Alternative gene symbol	Location
Endothelial PAS domain protein 1	*EPAS1*	Hypoxia-inducible factor 2, alpha subunit	*HIF-2α*	2p21-p16
Protein phosphatase, magnesium-dependent, 1, beta isoform	*PPM1B*	Protein phosphatase 2C, beta isoform	*PP2CB*	2p21
Calmodulin 2	*CALM2*	n/a	*PHKD2*	2p21
Solute carrier family 8, member 1	*SLC8A1*	Sodium-calcium exchanger 1	*NCX1*	2p23-p22

5.3 Results

5.3.1 SNPs

See Section 5.5 *EPAS1* SNP Paper.

5.3.2 DHPLC

5.3.2.1 Initial Screening of 18 SNP Haplotypes with DHPLC

The DNA samples from the 18 different SNP haplotypes produced three substitutions: one at the intron 8/exon 9 junction (i8/e9), one near the intron 12/exon 13 junction (i12/e13) and one in exon 15 (e15) (Figure 5-22, Figure 5-23 and Figure 5-24).

i8/e9

The C→G transversion substitution is very close to the splicing site of e9, so it could be important (Figure 5-22).

i12/e13

This A→C transversion substitution is near the i12/e13 junction but is too far away from e13 to be part of the 3′ splice site or branchpoint sequence (Figure 5-23).

e15

This G→A transition substitution is a synonymous codon: GTG and GTA both code for amino acid valine (Figure 5-24). It will not alter, therefore, charge or conformation and is unlikely to be functionally significant (Chapter 1: Section 1.3.7.4). It is possible that it could act as a cryptic splice site. The sequence change

from AAGGTGTCAG→AAGGTATCAG is similar to the consensus sequence 3′ splice site 5′-PyPyPyPyPyPyNCAG↓-3′ (Py = pyrimidine; N = any nucleotide; nucleotides consistent with consensus shaded; changes underlined). This is unlikely, however, because there are only five out of ten nucleotides that fit the consensus sequence and the nucleotide change does not increase this proportion.

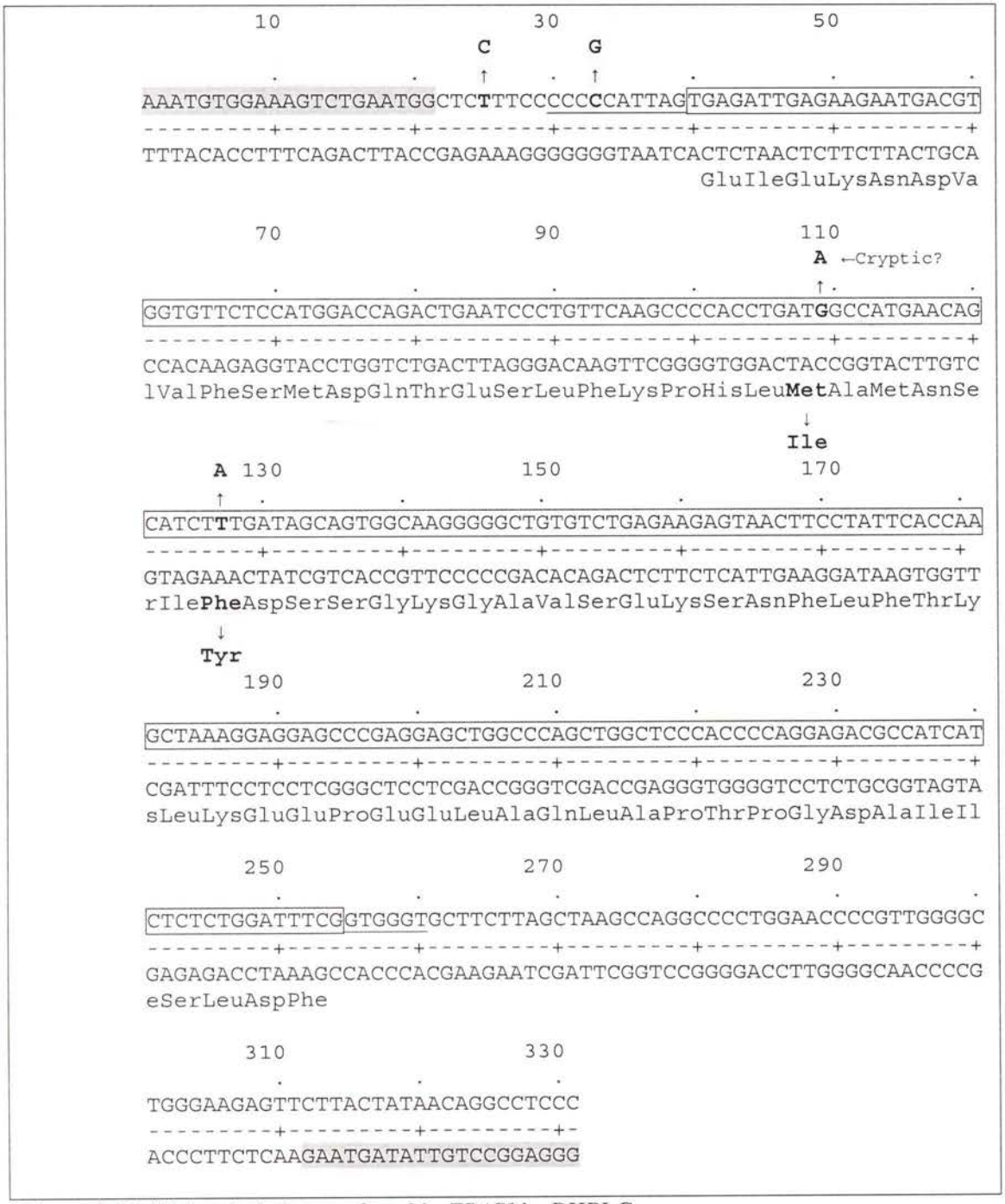


Figure 5-22 i8/e9 and e9 changes found in *EPAS1* by DHPLC.
(primer shaded, exon boxed, intron donor and acceptor sites underlined and variants marked with arrow). The G→A change could be a cryptic splice site.

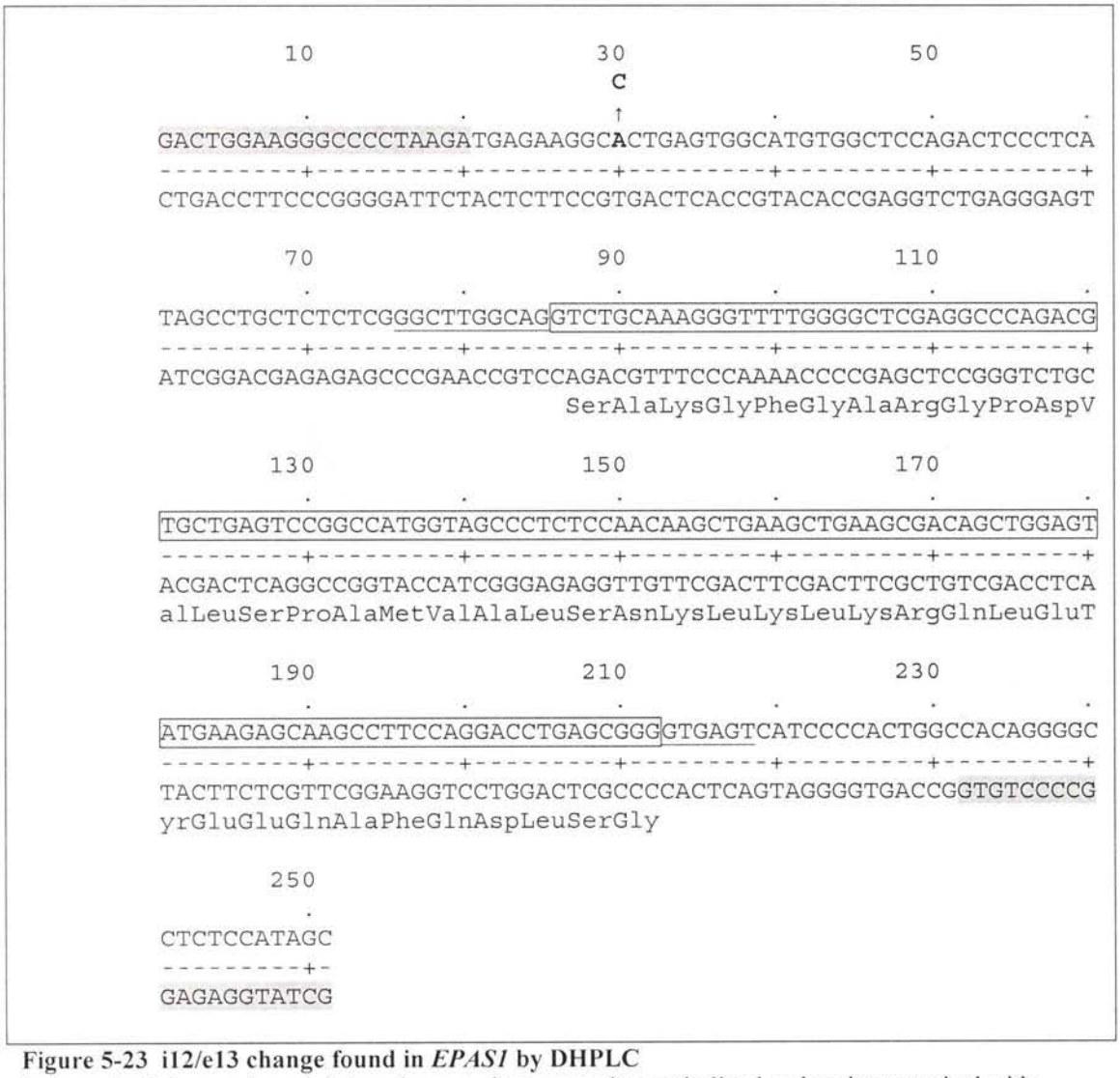


Figure 5-23 i12/e13 change found in *EPAS1* by DHPLC
(primer shaded, exon boxed, intron donor and acceptor sites underlined and variants marked with arrow).

Figure 5-24 e15 G→A change found in *EPAS1* by DHPLC
(primer shaded, exon boxed, intron donor and acceptor sites underlined and variants marked with arrow).

5.3.2.2 Going from 18 SNP Haplotypes to DHPLC-derived Haplotypes

Going from 18 SNP haplotypes to DHPLC-derived haplotypes produced three more substitutions: one at i8/e9 and two in e9 (Figure 5-22).

i8/e9

The T→C transition substitution is very close to the **splicing site** of e9, so it could be important.

e9

There was a G→A transition missense substitution and a T→A transversion missense substitution resulting in two nonsynonymous codon variants in e9. The G→A transversion missense substitution changes the codon from Met→Ile (M368I) which are both non-polar, hydrophobic amino acids. This is a conservative codon change and hence is less likely to affect function. This G→A change could also lead to the creation of a cryptic splice site, CACCTGATGG→CACCTGATAG since it does not differ from the consensus 3′ splice site (5′-PyPyPyPyPyPyNCAG↓-3′) by more than a few substitutions (Py = pyrimidine; N = any nucleotide; nucleotides consistent with consensus shaded; changes underlined). For a cryptic splice site, part of an exon might be lost from the mRNA or if the cryptic site lies within an intron then a segment of that intron will be retained in the mRNA.

The T→A transversion missense substitution is interesting because it changes the codon from Phe→Tyr (F374Y) which is from a non-polar, hydrophobic to a polar, hydrophilic amino acid. This is a nonconservative codon change and hence is more likely to affect function (Chapter 1: Section 1.3.7.4).

The i8/e9 and e9 waveforms clearly show the different DHPLC patterns (Figure 5-25, Figure 5-26, Figure 5-27, Figure 5-28 and Figure 5-29). A sample from each different waveform was sequenced and the variants were summarised (Table 5-9 and Table 5-10). There were no statistically significant differences found with males and females mixed or males only (Table 5-10). The differences between the cyclists and the controls for Waveform D is interesting but the sample size is too small (Table 5-10).

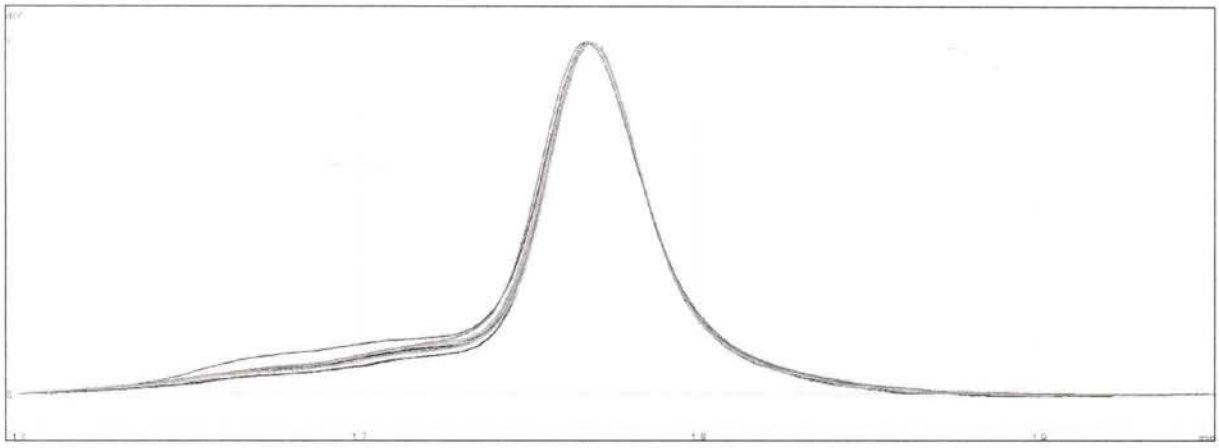


Figure 5-25 i8/e9 and e9 DHPLC Controls Single Peak
(Waveform A: C7/C7) at 60.2°C (Project: Locus B; Tray 021).

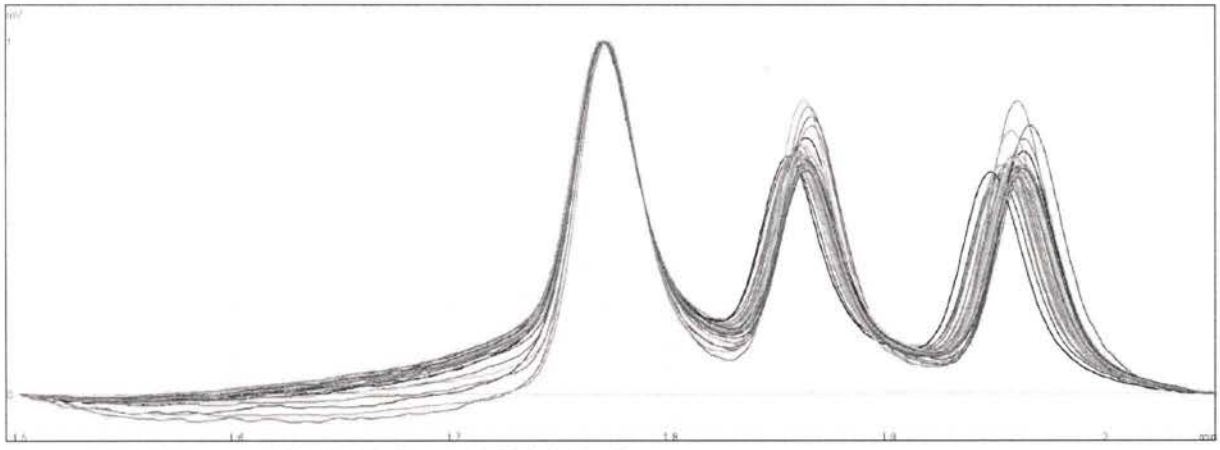


Figure 5-26 i8/e9 and e9 DHPLC Controls Triple Peak
(Waveform B: C_7/C_5GC) at 60.2°C (Project: Locus B; Tray 021).

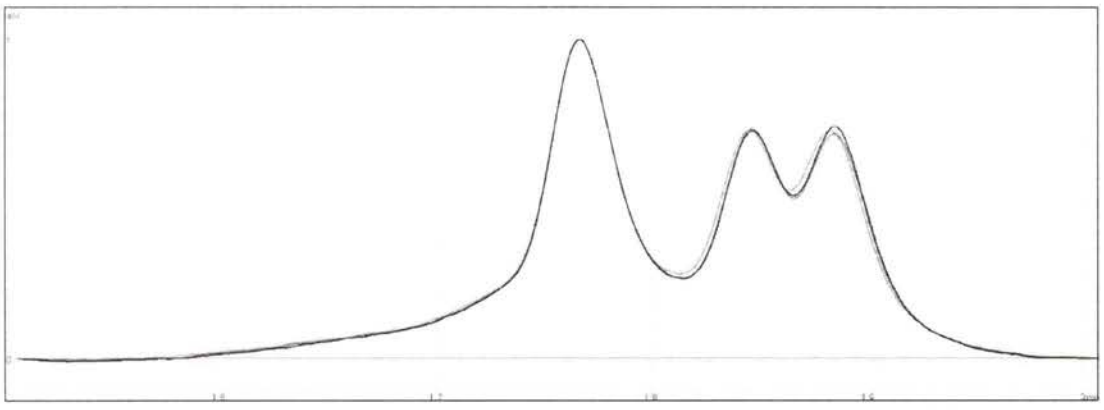


Figure 5-27 i8/e9 and e9 DHPLC Controls Triple Peak Narrow Gap
(Waveform C: C_7/$CTTC_7$) at 60.2°C (Project: Locus B; Tray 021).

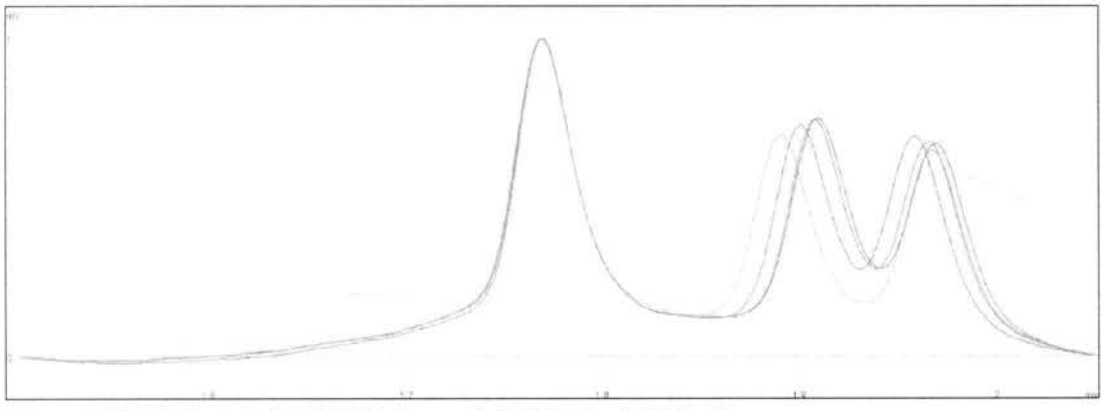


Figure 5-28 i8/e9 and e9 DHPLC Controls Triple Peak Wide Gap
(Waveform D: C_5GC/$CTTC_7$) at 60.2°C (Project: Locus B; Tray 021).

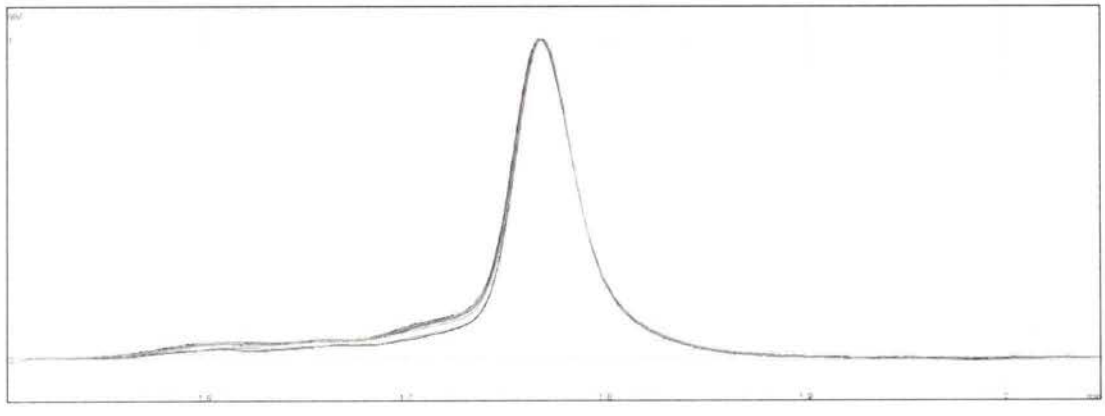


Figure 5-29 i8/e9 and e9 DHPLC Controls Single Late Peak
(Waveform E: C_5GC/CGC_5) at 60.2°C (Project: Locus B; Tray 021).

Table 5-9 i8/e9 and e9 complete waveform summary.

i8/e9 and e9 Waveforms	A	B	C	D	E	F	G	H	I	J	K	L	M
Controls (n=179)	46	76	5	10	35	1	0	0	1	0	0	1	4
Cyclists (n=73)	14	37	4	0	16	0	1	1	0	0	0	0	0
Ironman (n=125)	27	60	2	3	28	1	0	0	2	1	1	0	0
Rowers (n=169)	33	73	9	14	38	0	0	0	0	0	0	2	0
All Endurance athletes (n=367)	74	170	15	17	82	1	1	1	2	1	1	2	0

* A: C_7 / C_7 i.e. $TTTC_7ATTAG$
B: C_7 / C_5GC i.e. C_7/ $TTTC_5GCATTAG$
C: C_7 / $CTTC_7$ i.e. C_7/ $CTTC_7ATTAG$
D: C_5GC / $CTTC_7$ i.e. $TTTC_5GCATTAG/CTTC_7ATTAG$
E: C_5GC / C_5GC i.e. $TCTTTC_5GCATTAG$
F: $CTTC_7$ / $CTTC_7$ i.e. $CTTC_7ATTAG$

Table 5-10 Males and Females: i8/e9 simplified waveform summary
(*p* values from χ^2 distribution from CLUMP program)

i8/e9 Waveforms	A	B	C	D	E	F	*p* v ctrl
Controls (n=179)	**50**	**76**	**5**	**10**	**36**	**1**	
Males only (n=120)	**31**	**54**	**3**	**8**	**23**	**1**	
Cyclists (n=73)	16	37	4	0	16	0	0.214
Males only (n=46)	11	22	3	0	10	0	0.397
Ironman (n=125)							n/a
Males only (n=125)	30	60	2	3	29	1	0.639
Rowers (n=169)	33	73	9	14	40	0	0.275
Males only (n=100)	16	44	5	12	23	0	0.274
All Endurance athletes (n=367)	**79**	**170**	**15**	**17**	**85**	**1**	**0.544**
Males only (n=271)	**57**	**126**	**10**	**15**	**62**	**1**	**0.806**

* A: C_7 / C_7 i.e. $TTTC_7ATTAG$
B: C_7 / C_5GC i.e. C_7/ $TTTC_5GCATTAG$
C: C_7 / $CTTC_7$ i.e. C_7/ $CTTC_7ATTAG$
D: C_5GC / $CTTC_7$ i.e. $TTTC_5GCATTAG/CTTC_7ATTAG$
E: C_5GC / C_5GC i.e. $TCTTTC_5GCATTAG$
F: $CTTC_7$ / $CTTC_7$ i.e. $CTTC_7ATTAG$

For e15, the control DHPLC waveforms showed six samples that have waveforms that were different from the normal single peak whereas the cyclists did not show any (Figure 5-30, Figure 5-31 and Figure 5-32). The slight bulge in the majority of the

waveforms to the left of the peak is an artefact. Since there were so few variant waveforms, no further samples were run, since it would not achieve statistical significance.

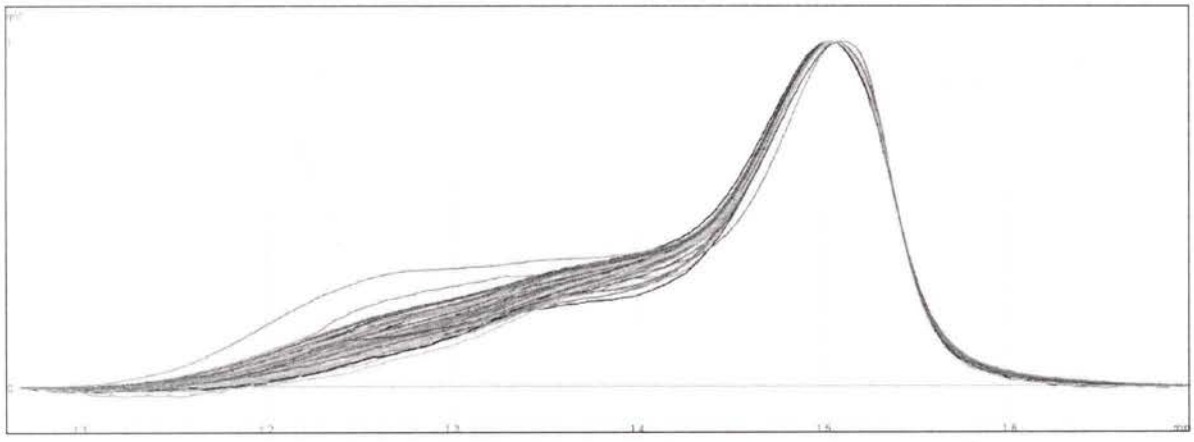


Figure 5-30 e15 DHPLC Controls Single Peak at 63.8°C
(Project: Locus B; Tray 033).

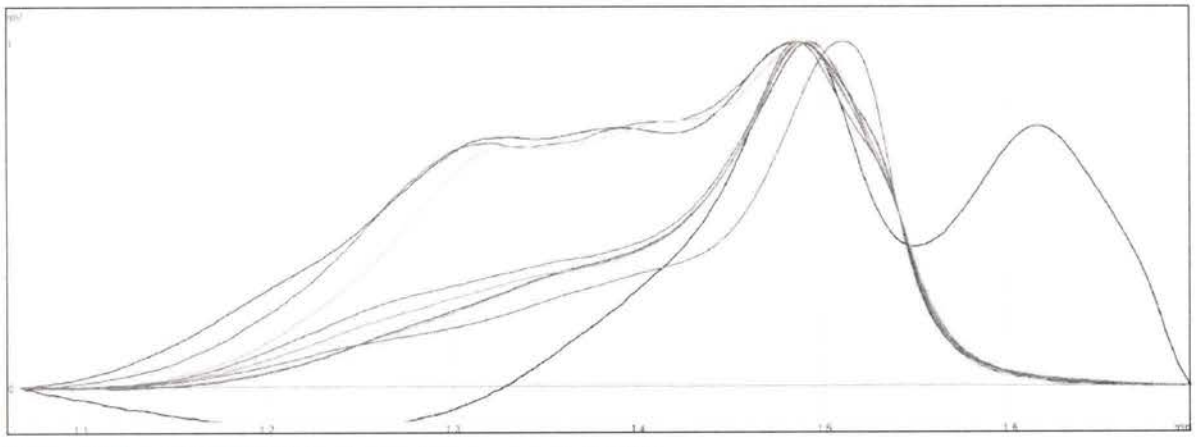


Figure 5-31 e15 DHPLC Controls other curves at 63.8°C
(Project: Locus B; Tray 033) (the double peak sample is the cyclist from the first 18 samples screened).

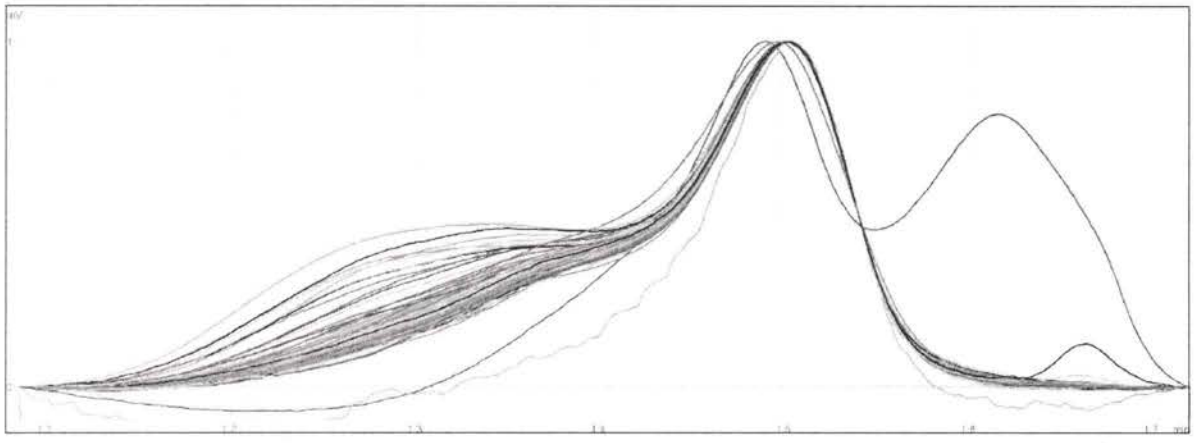


Figure 5-32 e15 DHPLC All Cyclists curves at 63.8°C
(Project: Locus B; Tray 032 and 033) (the double peak sample is the cyclist from the first 18 samples screened).

5.3.2.3 Genetic and Physiological Results

Multiple regression analysis was performed to determine the significant predictors of performance for all the male athletes combined because no specific order of importance was predicted by the literature on exercise physiology. Only the *EPAS1* i8/e9 C_7/C_5GC (i8/e9 Waveforms A, B and E from Table 5-10) polymorphism was used for this analysis because they were the only variants of sufficient quantity. Ignoring the i8/e9 T→C transition substitution variant: Waveforms A, C and F became C_7/C_7 genotype, B and D became C_7/C_5GC genotype and E was C_5GC/CC_5GC genotype.

5.3.2.3.1 Maximal Oxygen Uptake

Weight and cyclist were the statistically significant predictors of $\dot{V}O_2$ max ($p < 0.001$) and approximately 35% of the variance in $\dot{V}O_2$ max was accounted for by the linear combination of these variables. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of $\dot{V}O_2$ max when added to the model ($p = 0.964$) (Table 5-11).

5.3.2.3.2 Two Kilometre Row Time

Weight was the only statistically significant predictor of 2 km Row Time ($p < 0.001$) and approximately 70% of the variance in 2 km Row Time was accounted for by this variable. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of 2 km Row Time when added to the model ($p = 0.809$) (Table 5-11).

5.3.2.3.3 Ironman Time

Age was the only statistically significant predictor of Ironman Time ($p < 0.001$) and approximately 58% of the variance in Ironman Time was accounted for by this variable. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of Ironman Time when added to the model ($p = 0.402$) (Table 5-11).

5.3.2.3.4 Fitness Z-score

The above three sports-specific measures of aerobic fitness were combined in a Fitness Z-score for analysis. Age, Olympian, Ironman and cyclist were the statistically significant predictors of Fitness Z-score ($p < 0.001$) and approximately 44% of the variance in Fitness Z-score was accounted for by the linear combination of these variables. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of Fitness Z-score when added to the model ($p = 0.795$) (Table 5-11).

Table 5-11 Summary of *EPAS1* i8/e9 C_7/C_5GC multiple regression ANOVA related to performance
(statistical results from SPSS™ program).

Outcome variable	**Explanatory variables**	**Coefficients B**	***p* value**	**r^2**	**subtest genotype *p* value**
$\dot{V}O_2$ max	weight sport	-0.035 n/a	< 0.001	35.0%	0.964
2 km Row Time	weight	-0.041	< 0.001	69.8%	0.809
Ironman Time	age	-0.047	< 0.001	58.3%	0.402
Fitness Z-score	age Olympian sport	-0.042 0.781 n/a	< 0.001	44.1%	0.795

5.4 Discussion

5.4.1 SNPs

See Section 5.5 *EPAS1* SNP Paper.

5.4.2 DHPLC

5.4.2.1 DNA Variants

The DNA samples from the 18 different SNP haplotypes produced six substitutions: two at i8/e9, two in e9, one near i12/e13 and one in e15 (Figure 5-22, Figure 5-23 and Figure 5-24).

The i8/e9 C→G transversion substitution and T→C transition substitution are very close to the splicing site of e9, so they could be important. Individual genes can generate multiple protein isoforms through alternative splicing. Alternative splicing has a hidden role in gene control by nonsense-mediated decay of RNA (Matlin et al. 2005). The i8/e9 C→G transversion substitution could be very important because cysteine residues are strongly conserved and are considered to have a vital role because of the disulfide bonding of cysteine for conformation of the polypeptide. It is one of the least mutable amino acids because it is the only amino acid that has a sulfhydryl group in its side chain (Collins and Jukes 1994).

The e9 G→A transition missense substitution and a T→A transversion missense substitution resulted in two nonsynonymous codon variants in e9. The G→A transversion missense substitution changes the codon from Met→Ile (M368I) which

are both non-polar, hydrophobic amino acids. This is a conservative codon change and hence is less likely to affect function. This often has no significant effect on the biological activity of the protein because most proteins can tolerate at least a few amino acid changes without noticeable effect on their ability to function in the cell, but changes to some amino acids, such as those at the active site of an enzyme, have a greater impact (Brown 2002). This G→A change could also lead to the creation of a cryptic splice site, CACCTGATGG→CACCTGATAG since it does not differ from the consensus 3′ splice site (5′-PyPyPyPyPyPyNCAG↓-3′) by more than a few substitutions (Py = pyrimidine; N = any nucleotide; nucleotides consistent with consensus shaded; changes underlined). For a cryptic splice site, part of an exon might be lost from the mRNA or if the cryptic site lies within an intron then a segment of that intron will be retained in the mRNA. Human alternative isoform, cryptic and skipped splice sites depend on splice site strength, composition, GC content, location and binding site strength of polypyrimidine tract and branch site (Wang and Marin 2006). Cryptic splice sites have been reported for HIF-1α. A dicistronic reporter plasmid containing the 5′ UTRs from *HIF-1α* and *VEGF* UTRs generates downstream luciferase mainly due to cryptic promoter activity (Bert et al. 2006).

The T→A transversion missense substitution is interesting because it changes the codon from Phe→Tyr (F374Y) which is from a non-polar, hydrophobic to a polar, hydrophilic amino acid. This is a nonconservative codon change and hence is more likely to affect function (Chapter 1: Section 1.3.7.4). Long strings of hydrophobic amino acids may indicate secondary structural elements buried within protein complexes or lipid membranes. Runs with many polar residues are likely not to have the potential to form a hydrophobic core for a tertiary structure. Hydrophobic and

hydrophilic residues in larger sections might be informative regarding solubility and total charge. The G→A change did not reduce the string of four non-polar, hydrophobic amino acids. The T→A change affected a string of three hydrophobic amino acids by swapping a polar for a nonpolar but did not reduce the hydrophobic string size.

The i8/e9 and e9 waveforms clearly show the different DHPLC patterns (Figure 5-25, Figure 5-26, Figure 5-27, Figure 5-28 and Figure 5-29). A sample from each different waveform was sequenced and the variants were summarised (Table 5-9 and Table 5-10). There were no statistically significant differences found (Table 5-10). The differences between the cyclists and the controls for Waveform D is interesting but the sample size is too small (Table 5-10).

This i12/e13 A→C transversion substitution is near the i12/e13 junction but is too far away from e13 to be part of the 3′ splice site or even part of the branchpoint sequence.

This e15 G→A transition substitution is a synonymous codon: GTG and GTA both code for amino acid valine (Figure 5-24). It will not alter, therefore, charge or conformation and is unlikely to be functionally significant (Chapter 1: Section 1.3.7.4). It is possible that it could act as a cryptic splice site. The sequence change from AAGGTGTCAG→AAGGTATCAG is similar to the consensus sequence 3′ splice site 5′-PyPyPyPyPyPyNCAG↓-3′ (Py = pyrimidine; N = any nucleotide; nucleotides consistent with consensus shaded; changes underlined). This is unlikely, however, because there are only five out of ten nucleotides that fit the consensus sequence and the nucleotide change does not increase this proportion.

For e15, the control DHPLC waveforms showed six samples that have waveforms that were different from the normal single peak whereas the cyclists did not show any (Figure 5-30, Figure 5-31 and Figure 5-32). Since there were so few variant waveforms, no further samples were run, since it would not achieve statistical significance.

e15 contains the C-terminal activation domain (amino acids 820–870). Folding of the HIF-1α C-terminal transactivation domain is stabilised and transactivated by widespread hydrophobic and polar interactions (Freedman et al. 2002; Ruas et al. 2002). The C-terminal activation domain of the hypoxia-inducible transcription factors HIF-1α and HIF-2α binds domains of the transcriptional coactivators essential for hypoxia-responsive transcription (Kasper et al. 2005). Since the e15 variant does not change the amino acid, the EPAS1 C-terminal activation domain is unlikely to be affected by the e15 G→A transition substitution.

5.4.2.2 DNA Variants and Physiological Regression

This section of the study explored the relationship between the *EPAS1* i8/e9 C_7/C_5GC polymorphism and various physical, physiological and performance variables. Multiple regression analysis yielded one-, two- and three-variable models, composed of age, weight, Olympian and/or sport, which accounted for 35–70% of the variance in outcome performance variables (Table 5-11).

Weight and sport were the statistically significant predictors of $\dot{V}O_2$ max (z-score) ($p < 0.001$). *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of $\dot{V}O_2$ max when added to the model ($p = 0.964$) and approximately 35% of the variance in $\dot{V}O_2$ max was accounted for by the linear combination of these variables (Table 5-11). It was expected that weight (negative cooefficient) and sport would be predictors of $\dot{V}O_2$ max. It was not surprising that the *EPAS1* i8/e9 C_7/C_5GC genotype was not a predictor of $\dot{V}O_2$ max, since there was no association for the genetic data. This was, however, the same situation as the *ACE I/D* and MYBPC3 e6 polymorphisms. Perhaps the effect size is too small to be detected compared to the other predictive physical/phenotype variables.

Weight was the only statistically significant predictors of 2 km Row Time (z-score) ($p < 0.001$) and approximately 70% of the variance in 2 km Row Time was accounted for by this variable. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of 2 km Row Time when added to the model ($p = 0.809$) (Table 5-11). Weight (positive cooefficient) was once again expected to be a predictor of 2 km Row Time. Interestingly, height was not an explanatory variable for rowing. Height has traditionally been associated with success in rowing.

Age was the only statistically significant predictor of Ironman Time (z-score) ($p < 0.001$) and approximately 58% of the variance in Ironman Time was accounted for by this variable. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of Ironman Time when added to the model ($p = 0.402$) (Table 5-11). There was a much larger age range in the Ironman than in the other sports and hence it was expected that age (negative cooefficient) would be a significant predictor of Ironman Time.

The above three sports-specific measures of aerobic fitness were combined in a Fitness Z-score for analysis. Age, Olympian and sport were the statistically significant predictors of Fitness Z-score ($p < 0.001$) and approximately 44% of the variance in Fitness Z-score was accounted for by the linear combination of these variables. *EPAS1* i8/e9 C_7/C_5GC was not a significant predictor of Fitness Z-score when added to the model ($p = 0.795$) (Table 5-11). Age (negative cooefficient), Olympian (positive cooefficient) and sport were expected to be associated with the Fitness Z-score.

Age was a variable that appeared to be related to performance only in Ironman. One study showed that age was negatively correlated to *ACE* levels (Cambien 1988). This may have been a confounding factor in this study especially with the older athletes (i.e. Ironman) because it may have reduced the effect of the genotype.

The lack of significance of the *EPAS1* i8/e9 C_7/C_5GC genotype to the model may have been due to group sizes being too small or due to the incomplete physiological data set reducing the statistical power. It also could have been because whatever might have been contributing to a difference in the genotype distributions between some of the athlete groups and controls was not measured by the standard physiological tests used in the various sports.

There are many other factors that influence elite athletic performance and their phenotypes were not measured here. Some endurance sports require extreme levels of physical training which may have compromised the immune system of the athletes

(Gleeson et al. 1995; Gleeson et al. 2004) and this was not part of the physiological data set used for comparison used in this study.

Overall, this study of the *EPAS1* i8/e9 C_7/C_5GC polymorphism showed that it appeared to be associated with some sports but the underlying mechanism remains unclear.

The fact that no statistically significant differences were found does not mean that there are not statistically significant changes in the region bounded by i8/e9–e16. There could be important changes in the large intronic (non-coding) regions that were not scanned. Sequencing the whole gene may find the changes causing the significant association for this gene but this would be a significant undertaking.

It is likely that important variants in the gene have a small effect size, since the phenotype is of healthy athletes with improved cardiovascular performance and not seriously diseased people. The relatively small sample sizes involved in the study may, as a result, have precluded achieving statistically significant differences between the athletes and controls in this study. Future studies could investigate these changes using larger sample sizes or different groups of athletes or diseased phenotypes.

5.5 EPAS1 SNP Paper

Paper: The EPAS1 gene influences the aerobic-anaerobic contribution in elite endurance athletes.

For the purposes of Part 10, Division 4, 85(2) of the *University of Sydney (Amendment Act) Rule 1999 (as amended)*, a candidate may include in a thesis (whether in the body, or in one or more appendices) one or more published works of which the candidate is the sole or joint author.

Hum Genet (2005) 118: 416–423
DOI 10.1007/s00439-005-0066-0

ORIGINAL INVESTIGATION

Jennifer Henderson · Jason M. Withford-Cave
David L. Duffy · Stuart J. Cole · Nicole A. Sawyer
Jason P. Gulbin · Allan Hahn · Ronald J. Trent
Bing Yu

The *EPAS1* gene influences the aerobic–anaerobic contribution in elite endurance athletes

Received: 22 May 2005 / Accepted: 29 August 2005 / Published online: 6 October 2005

Absract *EPAS1* is a gene involved in complex oxygen sensing. It is expressed in microvascular endothelial cells, lung epithelial cells, cardiac myocytes and the brain. An association study was undertaken comparing elite endurance athletes classified into two groups according to a power–time model of performance intensity: power–time-maximum (PT-MAX; $N=242$, event duration 50 s to 10 min) and power–time–steady state (PT-SS; $N=151$, event duration ~2–10 h), with normal controls ($N=444$) using 12 SNPs across *EPAS1*. Ordinal regression analysis of allele frequencies revealed significant differences at SNPs 2 and 3 ($P=0.01$). Haplotype analysis revealed the presence of haplotypes involving SNPs 2–5 that significantly differentiated ($P<0.05$) the groups based on an ordinal ranking using the power–time classification. These same haplotypes differentiated the PT-MAX group in which a significant decrease in a haplotype (F: G-C-C-G; OR = 0.57, $P=0.02$, 95% CI 0.36–0.92) and increase in a second haplotype (G: A-T-G-G; OR = 1.75, $P=0.03$, 95% CI 1.05–2.91) was observed compared to controls. The PT-SS group was differentiated from the PT-MAX group by a third haplotype (H: A-T-G-A; OR = 0.46, $P=0.04$, 95% CI 0.22–0.96). Since *EPAS1* has a role as a sensor capable of integrating cardiovascular function, energetic demand, muscle activity and oxygen availability into physiological adaptation, we propose that DNA variants in *EPAS1* influence the relative contribution of aerobic and anaerobic metabolism and hence the maximum sustainable metabolic power for a given event duration.

J. Henderson · J. M. Withford-Cave
S. J. Cole · R. J. Trent (✉) · B. Yu
Department of Molecular & Clinical Genetics,
Royal Prince Alfred Hospital and Central Clinical School,
The University of Sydney (K25),
Camperdown 2050, Australia
E-mail: rtrent@med.usyd.edu.au
Tel.: +61-2-95157514
Fax: +61-2-95505412

D. L. Duffy
Queensland Institute of Medical Research,
300 Herston Road, Herston 4029, Australia

N. A. Sawyer · R. J. Trent · B. Yu
Sydney University Prince Alfred Macromolecular Analysis Centre (SUPAMAC), The University of Sydney (K25),
Sydney 2006, Australia

J. P. Gulbin · A. Hahn
Australian Institute of Sport,
Leverrier Crescent , Bruce 2616, Australia

Introduction

The relationship between endurance type sport and exercise intensity has long been regarded as hyperbolic (Fig. 1, Hill 1925). Accordingly, at high intensities exhaustion times are short, whereas, at a given low intensity, the hyperbole approaches an asymptote reflective of an intensity that theoretically could be sustained for indefinite periods. This intensity has been identified as critical power (Monod and Scherrer 1965) and has been taken to represent the maximum sustainable metabolic power.

The power–time relationship is determined by the contribution of available energy systems including: (1) alactic involving the energy substrates ATP and phosphocreatine stored within the muscle, (2) lactic which involves anaerobic glycolysis, and (3) aerobic which relies on oxidative phosphorylation. The first two pathways are dominant for short duration exercise (up to ~15 s), whereas for exercise of longer duration there exists an energetic flux between anaerobic and aerobic metabolism with the latter becoming the primary energy source as exercise duration increases.

The crossover to predominantly aerobic energy supply is reported to occur between 15 and 30 s (Spencer and Gastin 2001). At this turning-point, endurance-trained individuals derive a greater energy contribution

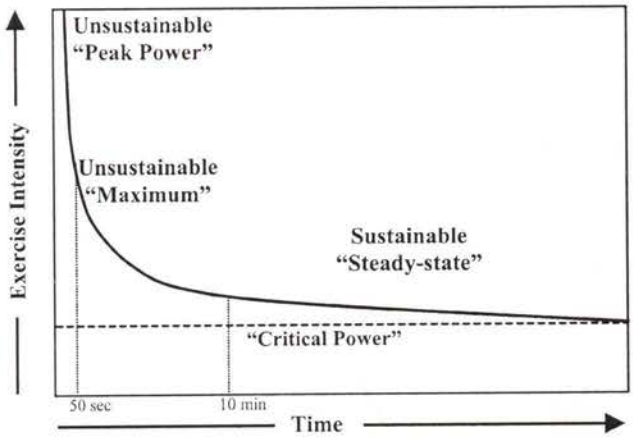


Fig. 1 Hyperbolic relationship between exercise intensity and time to exhaustion. Although exercise intensity can be represented by measures of power output or speed depending on the type of exercise, for the purposes of this paper, this relationship will be referred to as power–time to reflect the concept of maximum sustainable metabolic power. This relationship identifies an intensity known as *critical power* determined as the power asymptote of the power–time relationship. *Peak power* reflects the instantaneous work output resulting from finite energy stored within the muscle (ATP/phosphocreatine). *Maximum* reflects an intensity that fully exploits aerobic and anaerobic energy sources to maintain performance intensity. It is here that the power–time relationship is most meaningfully applied involving performances in the range of ~50 s to ~10 min (di Prampero 2003). *Steady-state* exercise reflects an intensity of exercise that is sustainable due to a dominance of aerobic energy supply that does not rely on a major contribution from anaerobic metabolism. Both peak power and maximum intensities are considered unsustainable due to finite energy supply and inhibitory byproducts

via aerobic metabolism compared to sprint-trained individuals who rely more heavily on anaerobic sources (Medbo and Sejersted 1985; Nummela and Rusko 1995). In the context of endurance, the determinants of energetics at this turning-point involve factors related to underlying cardiorespiratory function and metabolic characteristics in exercising muscles that determine the maximal sustainable performance intensity. Furthermore, the traits underlying the endurance phenotype may reflect evolutionary trade-offs that have enabled locomotive systems to evolve in a specialist versus generalist manner (Van Damme et al. 2002) suggesting a strong heritable component.

The physiological characteristics that contribute to the endurance phenotype involve measures such as maximal oxygen uptake (VO_2 max), cardiac output, aerobic enzyme capacity, glycogen stores (Kayser 2003) and maximal accumulated oxygen deficit (Medbo and Tabata 1989) or peak lactate as indicators of anaerobic capacity. However, these measures may not be sufficient for assessing and distinguishing elite and sub-elite endurance competitors and a more comprehensive interpretation of the relevant biological factors, training, recovery and competitive requirements has been suggested (Myburgh 2003). Recently, integrative paradigms have been proposed to explain endurance fatigue based on the possible involvement of complex sensing mechanisms that protect systems from myocardial ischaemia, metabolic stress and disruptions in homeostasis (Kayser 2003; Noakes et al. 2001; St Clair Gibson and Noakes 2004).

The genetic response to environmental factors such as oxygen deficiency, which is the stimulus for many adaptations including angiogenesis, erythropoiesis, vasomotor control, glucose and catecholamine metabolism, may provide a sensing mechanism capable of integrating the physiological signals that determine time to exhaustion. Hypoxia inducible factor (HIF) represents a major signalling pathway responsible for activating gene expression in response to oxygen levels through binding to a core response element in a growing repertoire of HIF-responsive genes including vascular endothelial growth factor, erythropoietin and adrenomedullin. These genes play a role in responding to oxygen deficiency, influencing cardiac output and vascular tone, and enhancing cellular oxygen utilisation (Semenza 1998).

The relative contribution of genes and the environment in determining performance in a wide range of endurance sporting situations promoted particular interest in genetics and exercise physiology. Indeed the cataloguing of the genetic basis of human variation in health and human performance phenotypes has been the topic of an annual review in recent years (Rankinen et al. 2004). While the functional significance of these genetic factors in determining elite endurance performance is unclear, there is increasing evidence to suggest that multiple variations in genetic make-up may modify gene expression through gene–environment interactions and so contribute to an individual's success in endurance type sports.

A genome-wide scan based on VO_2 max as the endurance phenotype, identified multiple loci (Bouchard et al. 2000). These included D2S2739 on chromosome 2p16.1 (SIBPAL linkage analysis $P < 0.01$). From an in silico search 5 Mb upstream and downstream of D2S2739 four plausible genes were identified as candidates involved in endurance traits. One of these genes was the endothelial PAS domain protein 1 (*EPAS1*) a gene involved in the HIF pathway. To date, no study has examined the impact of DNA variants within *EPAS1* in a physiological model such as elite endurance athletes. Therefore, we conducted a case–control association study of the *EPAS1* locus. Using a power–time model of endurance performance (Fig. 1), two athlete cohorts with different phenotypes (maximum intensity or steady-state intensity) were compared. From the association study, we found specific haplotypes that distinguished the two athlete cohorts. We propose that the DNA variants in *EPAS1* underlying these haplotypes confer a specific advantage in endurance performance, and may be particularly relevant to events and disciplines involving performances at maximal intensities between ~50 s and 10 min in which a substantial aerobic and anaerobic contribution is exploited.

Materials and methods

Subjects

The study cohort comprised male and female elite athletes participating in sport programs under the administration of the Australian Institute of Sport including 172 rowers (2,000 m), 42 swimmers (100–800 m), 28 middle distance runners (800–3,000 m), 24 Olympic distance triathletes, 41 sprinters (including 33 track-runners and eight track-cyclists), and 58 endurance cyclists (including road and mountain bike specialists). All had reached national or international level in their sport including a substantial number of Olympic and World Championship competitors. In addition, 127 competitors from the 2001 Australian Ironman Triathlon competition series were selected from a group of over 600 participants who agreed to be part of the study. Based on the age-corrected time percentage behind the overall winner, only those finishing in the top half of the field were included for genotyping resulting in a cohort of high-calibre Ironman athletes.

Using the power–time model (Fig. 1), with ~50 s as a cut-off for endurance, the sprinters consisting primarily of 100–400 m track runners and sprint cyclists were excluded from the association analysis. Endurance cyclists were also excluded, as this was a mixed group with insufficient data available for an appropriate classification. Those excluded were subsequently analysed when different sports groups were compared based on results obtained from the initial association study.

Endurance grouping

Endurance athletes were split into two groups based on the performance intensity as a function of event duration as in the power–time relationship outlined in Fig. 1. Under such a framework, those sporting events lasting ~50 s or longer but not more than 10 min were classified together as these events involve a performance effort requiring a maximum aerobic and anaerobic contribution that is unsustainable, leading to exhaustion in the order of seconds to minutes depending on the event. Examples included swimming, rowing and middle distance running, and this group was designated power–time-maximum (PT-MAX). In contrast, endurance sports involving performances lasting > 10 min require a sustainable performance intensity in the order of minutes to several hours, often referred to as steady-state, that involves a predominance of aerobic energy. Examples included Olympic and Ironman triathlon which typically last ~2 and ~9 h, respectively. This group was designated power–time-steady state (PT-SS). Table 1 outlines the sport groupings and provides world record times for associated events as a point of reference.

The control cohort comprised 444 healthy Caucasians of European decent. DNA samples were from a commercial source (European collection of cell cultures—ECACC, Wiltshire, UK) or from Australian nationals of European descent. Human Research Ethics Committees from the University of Sydney, Royal Prince Alfred Hospital and the Australian Institute of Sport approved the research proposal.

DNA preparation and SNP analysis

DNA was isolated from whole blood using QIAamp 96 DNA Blood Kit (Qiagen Inc., Hilden, Germany). DNA concentration was determined and an absorbance ratio of $A_{260}/A_{280} > 1.7$ was required to confirm the purity of the DNA. For SNP identification, 50 μL of DNA at a concentration of 5–10 ng μL^{-1} was pipetted into 96-well skirted plates. Plates were sealed and analysed at the Sydney University Prince Alfred Macromolecular Analysis Center (SUPAMAC: http://www.supamac.com.au). The *EPAS1* gene has 16 exons extending over 90 kb of genomic DNA. Twelve commercially available intronic SNPs (Table 2) were selected for study extending from exons 1–12 of the *EPAS1* gene. The first five SNPs were located within the large (~50 kb) intron 1 of

Table 1 Summary of sport grouping for association analysis based on the power–time model

PT-SS sport with events (time in h:min)	PT-MAX sport with events in metres (time in min:s)
Ironman: 3.8 km swim/180 km cycle/42.2 km run (>9:00)	Swimmers 100 m (0:47.84 WR[a]) 200 m (1:44.06 WR) 400 m (3:40.08 WR) 800 m (7:38.65 WR)
Olympic triathlon: 1.5 km swim/40 km cycle/10 km run (~1:50)	Middle distance runners 800 m (1:41.11 WR) 1,500 m (3:26.34 WR) 3,000 m (7:20.67 WR) Rowers[b] 2,000 m (5:19.85 WR)

[a] *WR* World Record. World Record Data obtained from: http://www.ausswim.telstra.com/au/records/details.cfm; http://www.athletics.org.au/content/records/P003fb00.pdf; http://www.worldrowing.com/results/besttimes.sps
[b] Men's 8 crew

this gene. SNPs were genotyped using a high-throughput system (ABI Prism 7900HT, Applied Biosystems Inc., Foster City, USA) using a pre-designed 5′ nuclease assay (TaqMan® SNP Genotyping Assay, Applied Biosystems Inc.) containing both forward and reverse primers and 6FAM™ and VIC® dye-MGB labelled probes. Following thermal cycling, genotype data were acquired automatically and analysed using sequence detection software (SDS v 2.1, Applied Biosystems Inc.).

Statistical analysis

Hardy–Wienberg equilibrium (HWE) and linkage disequilibrium (LD) blocks

Chi-square analysis of all subject data (cases and controls) was used to confirm HWE. Block structures were examined within the *EPAS1* locus using Haploview (v. 3.0) to calculate several pair-wise LD measures, which are then partitioned into block structures using common approaches to block definition (Barrett et al. 2005).

Association study

Ordinal regression analysis was conducted for all 12 SNPs to identify any association between group membership and specific SNPs using genotype data. Further analysis was performed to determine whether the combinations of SNPs improved the association result. This analysis determined if haplotypes contributed to the association result and was used to identify the contributing allelic combinations. Based on the ordinal groups, specific between-groups binomial regression analysis was also performed to determine which group contributed most to the overall association result and to identify the associated within-group haplotypes.

Haplotype analysis was performed using the *haplo.stats* computer software package (v1.1.1, http://www.mayo.edu/hsr/people/schaid.html). This program uses the expectation maximisation algorithm to compute the maximum likelihood estimate of haplotype probabilities and provides an iterative approach in situations with incomplete data sets or missing values. Three functions within the software were used in the statistical analysis: (1) *haplo.em* performs the initial ambiguous haplotype estimation, (2) *haplo.score* computes a score statistic to evaluate the association of trait with haplotypes and provides permutation as well as asymptotic *P*-values (Schaid et al. 2002), (3) *haplo.glm* performs a regression of trait on ambiguous haplotypes, and was used to calculate odds ratios and 95% confidence intervals.

Multiple SNP testing

We have dealt with the fact that multiple SNPs are being tested for association using both the most conservative Bonferroni correction for multiple testing and a correction based on a determination of the effective number of tests using principal components analysis (http://CRAN.R-project.org/). The latter was determined by the number of tests that explain at least 90% of the variance in the pair-wise linkage disequilibria (expressed as binary correlations) allowing the intercorrelation between the SNPs due to linkage disequilibrium.

Results

HWE and LD blocks

Regression analysis for the effect of age and sex on genotype revealed no significant correlation (data not shown). Furthermore, the association result was unchanged when the age and sex adjustment was removed (data not shown). Minor allele frequencies for all SNPs in the groups are summarised in Table 2 including data confirming HWE. Furthermore, strong LD was observed between several alleles in the study population. These resulted in block structures at SNPs 1–3, 6–7 and 9–12 based on the solid spline method using $D' > 0.80$ as a cut-off.

Table 2 Summary of SNPs used in the study

SNP[a] number	Celera SNP identification	Change	a/a	a/b	b/b	Minor allele		HWE exact *P*-value
1	hCV11639978	C/G	4	150	780	G	0.085	0.39
2	hCV11639984	A/G	306	456	174	A	0.429	0.89
3	hCV2148918	C/T	177	452	307	T	0.431	0.64
4	hCV2148915	C/G	290	451	187	C	0.445	0.64
5	hCV2162989	A/G	200	496	237	G	0.480	0.06
6	hCV2162974	A/T	20	194	721	T	0.125	0.13
7	hCV154424	G/T	19	185	725	T	0.120	0.09
8	hCV2162964	C/T	354	449	132	C	0.381	0.63
9	hCV7523424	A/G	267	485	181	A	0.454	0.15
10	hCV207915	C/T	55	366	514	T	0.255	0.39
11	hCV2162960	C/G	261	443	226	C	0.481	0.17
12	hCVF11158118	C/G	220	479	235	G	0.492	0.47

[a]SNP genotyping success rate > 99%

Association study

Based on ordinal data in which the control group = 0, PT-SS = 1, PT-MAX = 2, the initial regression against genotype distribution at all 12 SNPs revealed significant differences at SNPs 1, 2 and 3 with *P*-values of 0.04, 0.004 and 0.002, respectively. Prior to conducting further association analysis of haplotype data, an analysis of allele distribution was conducted using all sport subgroups, resulting in a significant difference with $P = 0.01$ for both SNPs 2 and 3. Figure 2 illustrates the observed trend in allele frequency for these SNPs.

Haplotype analysis was performed across the 12 SNPs to confirm the signals at SNPs 2 and 3 based on the ordinal ranking of groups, and the results indicated significant differences ($P < 0.05$) in the region of SNPs 1–6 (Table 3). Within this region of significance, score tests identified three contributing SNP combinations comprising SNPs 1–4, 2–4 and 2–5 ($P < 0.05$). Table 4 provides a summary of the score test results identifying specific contributing haplotypes and associated *P*-values. Binomial regression analysis of group assignment on haplotype was performed within this region to determine which groups accounted for the significant differences observed in the analysis of ordinal data, and to further evaluate the contributing haplotypes using odds ratios (OR). Significant differences were observed between the control and PT-MAX groups ($P = 0.01$). As in the ordinal analysis of SNPs 2–5, haplotype F: *G-C-C-G* and haplotype G: *A-T-G-G* (see Table 4) were identified as the significant allelic combinations with OR of 0.57 ($P = 0.02$; 95% confidence interval (CI) 0.36–0.92) and OR 1.75 ($P = 0.03$; 95% CI 1.05–2.91) respectively. A similar binomial analysis between the control and PT-SS groups revealed no significant difference in haplotype distribution ($P > 0.05$).

Further binomial analysis was conducted between the two athlete groups (PT-MAX and PT-SS) to identify potential differences in haplotype distributions between these groups. A significant difference was observed in haplotype distribution ($P = 0.03$) between the two groups. The contributing allelic combinations included haplotype F: *G-C-C-G* with an OR of 0.47 ($P = 0.01$; 95% CI 0.26–0.85). This analysis also revealed a variation of haplotype G: *A-T-G-G* (see Table 4) in which a base substitution at SNP 5 resulted in haplotype H: *A-T-G-A*, with an OR of 0.46 ($P = 0.04$; 95% CI 0.22–0.96). Figure 3 illustrates the frequency of allelic combinations contributing to the significant differences observed in both the ordinal and binomial regression of group membership on haplotype.

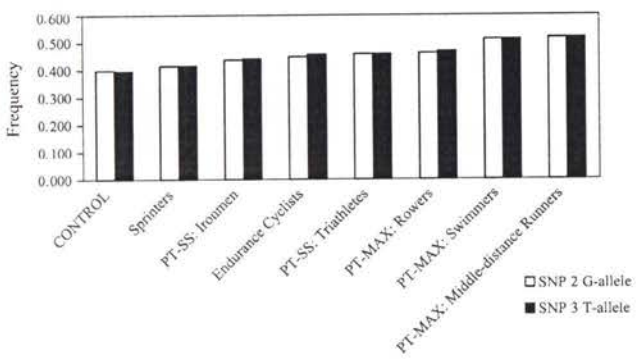


Fig. 2 Ordinal trend of allelic distribution for SNP 2 and SNP 3 between athlete sub-groups. Group membership is based on PT-MAX/PT-SS classification used in the association analysis (sprinters and endurance cyclists were not used for this analysis)

Multiple SNP testing

Due to the number of SNPs and groups being tested in this study, there is a need to account for a multiple testing effect. However, the extent of LD between the SNPs renders some redundancy in the informativeness of SNPs. Therefore, in accounting for the presence of LD between SNPs, the principal components analysis suggested that the first seven components (eigen vectors) explained 93% of the variance. Adjusting for the multiple comparisons, by applying a correction of seven, resulted in *P*-values of 0.03 and 0.01 for the ordinal group-wise regression analysis at SNPs 2 and 3, respectively. Furthermore, the most conservative Bonferroni correction (by a factor of 12) resulted in a continued significant difference for the ordinal group-wise regression analysis at SNPs 2 and 3 with *P*-values of 0.04 and 0.02 respectively. Given the continued significance by the Bonferroni correction regardless of the extent of LD, we are confident that our findings do not merely reflect a multiple testing effect. Since the remainder of the analysis was undertaken to understand the origin of the significant effect at SNPs 2 and 3 and to identify distinguishing haplotypes, multiple testing corrections were no longer applied.

Discussion

Sensing mechanisms are likely to be responsible for many of the adaptations induced in the cardiorespiratory system and skeletal muscle by exercise and changes in the environment (Fluck and Hoppeler 2003). In the

Table 3 Global comparison of sliding windows haplotypes in the region of SNPs 1–6[a]

Sliding windows haplotype					
2 SNP		3 SNP		4 SNP	
SNP	*P*-value	SNP	*P*-value	SNP	*P*-value
1–2	0.01	1–3	0.01	1–4	0.02
2–3	0.01	2–4	0.01	2–5	0.04
3–4	0.01	3–5	0.02	3–6	0.05
4–5	0.37	4–6	0.40		
5–6	0.26				

[a] *P*-values represent the levels of significance obtained by simulation

Table 4 Haplotypes in the region of SNPs 1–6

Score test of association[a]					
SNPs 1–4 ($P=0.01$)		SNPs 2–4 ($P=0.01$)		SNPs 2–5 ($P=0.04$)	
Haplotype	*P*-value	Haplotype	*P*-value	Haplotype	*P*-value
A) C-G-C-C	0.01	C) G-C-C	0.004	F) G-C-C-G	0.01
B) G-A-T-G	0.01	D) A-T-G	0.04	G) A-T-G-G	0.02
		E) A-T-C	0.04		

[a]This association was based on ordinal data, with the score statistic for SNP combinations as well as specific haplotypes. Global simulated *P*-values are shown in brackets. *P*-values represent the level of significance obtained by simulation

case of endurance exercise, these adaptations lead to an increased capacity for aerobic energy turnover induced by changes in oxygen delivery patterns. The extent of imbalances in cellular homeostasis that occur with exercise of different intensity and duration imply that both metabolic and mechanical factors can be sensed separately and integrated into complex transcriptional responses (Fluck and Hoppeler 2003). The consequence of such transcriptional regulation facilitates cellular capacity to tolerate further physiological stress.

Hypoxia inducible factor has been described as a sensor integrating muscle activity and oxygen availability into muscular remodelling (Fluck and Hoppeler 2003) and as a master regulator for oxygen homeostasis (Semenza 1998). As a transcription factor, HIF regulates a number of genes involved in the cellular and systemic responses to hypoxia including erythropoiesis, angiogenesis, vascular regulation and anaerobic metabolism. It exists as a dimer consisting of α and β subunits. The β-subunit is an aryl hydrocarbon receptor nuclear translocator (ARNT), while the α-subunit is a basic helix-loop-helix (bHLH)-PER-ARNT-SIM (PAS) protein that exists as two major isoforms: HIF-1α and EPAS1. Due to this link with HIF, *EPAS1* is also known as *HIF2A*.

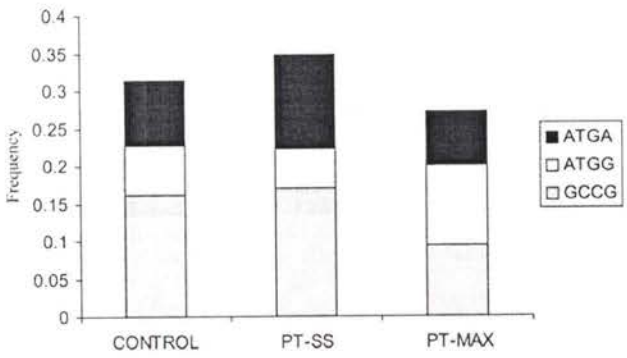


Fig. 3 Distribution of allelic combinations contributing to significant differences in haplotype distribution observed in the ordinal analysis based on group assignment in which *control* = 0, PT-SS = 1 and *PT-MAX* = 2

Hypoxia inducible factor activity is regulated at both transcriptional (Minet et al. 1999; Wang et al. 1995) and post-translational levels (Huang et al. 1996) and recent evidence has suggested a mechanism of transcription-dependent degradation via a feedback loop involving transcription inhibitors (Demidenko et al. 2005). Under the influence of hypoxia, regulation of the α-subunit appears to be the rate-limiting step. However, there is increasing evidence to suggest that HIF may be implicated in biological functions requiring its activation under normoxic conditions (Dery et al. 2005). Furthermore, although the EPAS1 and HIF-1α proteins share significant overall amino acid identity (48%) (Hu et al. 2003), several molecular, biochemical and physiologic observations suggest HIF-1α and EPAS1 represent distinct pathways. It has been shown that *EPAS1* may be the key modulator for the level of hypoxia likely to be present during endurance exercise since it has a slightly lower threshold for hypoxic gene activation (Wiesener et al. 1998). *EPAS1* also demonstrates a selective pattern of expression in microvascular endothelial cells, lung epithelial cells, cardiac myocytes and the brain (Hu et al. 2003; Wiesener et al. 2003).

Knockout mouse models have revealed a key physiological role for *EPAS1* demonstrating various phenotypes. In one case, *EPAS1*$^{-/-}$ mice had deficiencies in catecholamine sources from the organ of Zuckerland resulting in embryonic lethality due to bradycardia (Tian et al. 1998). In a different knockout model, *EPAS1* was found to play an important role in the remodelling of the primary vascular network (Peng et al. 2000). Other knockout models have revealed widespread multiple organ pathology, biochemical abnormalities, greater oxidative stress and an impaired response to oxidative stress in *EPAS1*$^{-/-}$ mice (Scortegagna et al. 2003a). Altered haematopoiesis resulting in a significant reduction in haematocrit levels and global depression in peripheral blood counts have also been observed in *EPAS1* null mice (Scortegagna et al. 2003b).

The present study has compared *EPAS1* SNPs in a cohort of endurance athletes and normal controls and has revealed a significant trend in which differences in allele frequency at SNPs 2 and 3 are dependent on an assignment of sport to a maximum or steady-state performance intensity group. Furthermore, when different sports groups were considered individually, a consistent significant trend was observed for allele frequencies at SNPs 2 and 3 resulting in those groups belonging to the PT-MAX group being clustered together and having the highest frequency of the distinguishing allele, while those in the PT-SS group were distributed in the middle, and controls the least (Fig. 2). This suggests that the groups in the association study are homogenous and that the classification system may reflect some physiological mechanism underlying the power–time model of endurance performance.

The sliding window haplotype analysis revealed a consistent signal across SNPs 1–6 suggesting this region as more relevant in distinguishing the association

groups. The ordinal analysis of haplotypes identified seven allelic combinations that contributed to the significant global result (Table 4). Of these, six contain the same two sets of allelic combinations at SNPs 2–4 suggesting this as a possible core area contributing to the association result. Further binomial analysis identified the same two specific allelic combinations (haplotype F and G) contributed to differences between the PT-MAX and control groups as was observed in the ordinal analysis. Interestingly, a comparison between the two athlete groups (PT-MAX and PT-SS) identified a third combination (haplotype H) in PT-SS athletes, which contributed to a significant difference between athlete groups.

Our work has identified three *EPAS1* haplotypes to be significantly associated with elite endurance athletes classified according to the power–time model of endurance. The presence of one (haplotype G) and the absence of another (haplotype F) at the same locus is observed in athletes involved in high intensity maximal exercise of a duration between 50 s and 10 min. In addition, athletes involved in a sustained steady-state effort (from ~2 to 10 h) demonstrate the increased presence of a third (haplotype H). This haplotype association may be indicative of underlying physiological factors determining the relative contribution of aerobic and anaerobic metabolism towards setting the maximum sustainable metabolic power for a given event duration.

We propose that the *EPAS1* haplotypes identified may be providing a more sensitive metabolic response in determining the aerobic and anaerobic contribution in endurance sport. Although a mechanism in which regulatory factors, including *EPAS1*, can integrate signals indicative of homeostatic disruption and metabolic stress into transcriptional adaptations is plausible, further work is required to determine the complex network by which this mechanism may be operational.

Acknowledgements Funding for this study was provided through a grant from the Australia Research Council. We thank the reviewers for their helpful feedback, and Prof K North for providing DNA from 33 track runners and 19 cyclists.

References

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

Bouchard C, Rankinen T, Chagnon YC, Rice T, Perusse L, Gagnon J, Borecki I, An P, Leon AS, Skinner JS, Wilmore JH, Province M, Rao DC (2000) Genomic scan for maximal oxygen uptake and its response to training in the HERITAGE Family Study. J Appl Physiol 88:551–559

Demidenko ZN, Rapisarda A, Garayoa M, Giannakakou P, Melillo G, Blagosklonny MV (2005) Accumulation of hypoxia-inducible factor-1alpha is limited by transcription-dependent depletion. Oncogene 24:4829–4838

Dery MA, Michaud MD, Richard DE (2005) Hypoxia-inducible factor 1: regulation by hypoxic and non-hypoxic activators. Int J Biochem Cell Biol 37:535–540

di Prampero PE (2003) Factors limiting maximal performance in humans. Eur J Appl Physiol 90:420–429

Fluck M, Hoppeler H (2003) Molecular basis of skeletal muscle plasticity—from gene to form and function. Rev Physiol Biochem Pharmacol 146:159–216

Hill AV (1925) The physiological basis of athletic records. Lancet 206:481–486

Hu CJ, Wang LY, Chodosh LA, Keith B, Simon MC (2003) Differential roles of hypoxia-inducible factor 1alpha (HIF-1alpha) and HIF-2alpha in hypoxic gene regulation. Mol Cell Biol 23:9361–9374

Huang LE, Arany Z, Livingston DM, Bunn HF (1996) Activation of hypoxia-inducible transcription factor depends primarily upon redox-sensitive stabilization of its alpha subunit. J Biol Chem 271:32253–32259

Kayser B (2003) Exercise starts and ends in the brain. Eur J Appl Physiol 90:411–419

Medbo JI, Sejersted OM (1985) Acid-base and electrolyte balance after exhausting exercise in endurance-trained and sprint-trained subjects. Acta Physiol Scand 125:97–109

Medbo JI, Tabata I (1989) Relative importance of aerobic and anaerobic energy release during short-lasting exhausting bicycle exercise. J Appl Physiol 67:1881–1886

Minet E, Ernest I, Michel G, Roland I, Remacle J, Raes M, Michiels C (1999) HIF1A gene transcription is dependent on a core promoter sequence encompassing activating and inhibiting sequences located upstream from the transcription initiation site and cis elements located within the 5'UTR. Biochem Biophys Res Commun 261:534–540

Monod H, Scherrer J (1965) The work capacity of synergistic muscular group. Ergonomics 8:329–338

Myburgh KH (2003) What makes an endurance athlete world-class? Not simply a physiological conundrum. Comp Biochem Physiol A Mol Integr Physiol 136:171–190

Noakes TD, Peltonen JE, Rusko HK (2001) Evidence that a central governor regulates exercise performance during acute hypoxia and hyperoxia. J Exp Biol 204:3225–3234

Nummela A, Rusko H (1995) Time course of anaerobic and aerobic energy expenditure during short-term exhaustive running in athletes. Int J Sports Med 16:522–527

Peng J, Zhang L, Drysdale L, Fong GH (2000) The transcription factor EPAS-1/hypoxia-inducible factor 2alpha plays an important role in vascular remodeling. Proc Natl Acad Sci USA 97:8386–8391

Rankinen T, Perusse L, Rauramaa R, Rivera MA, Wolfarth B, Bouchard C (2004) The human gene map for performance and health-related fitness phenotypes: the 2003 update. Med Sci Sports Exerc 36:1451–1469

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434

Scortegagna M, Ding K, Oktay Y, Gaur A, Thurmond F, Yan LJ, Marck BT, Matsumoto AM, Shelton JM, Richardson JA, Bennett MJ, Garcia JA (2003a) Multiple organ pathology, metabolic abnormalities and impaired homeostasis of reactive oxygen species in Epas1$^{-/-}$ mice. Nat Genet 35:331–340

Scortegagna M, Morris MA, Oktay Y, Bennett M, Garcia JA (2003b) The HIF family member EPAS1/HIF-2alpha is required for normal hematopoiesis in mice. Blood 102:1634–1640

Semenza GL (1998) Hypoxia-inducible factor 1: master regulator of O2 homeostasis. Curr Opin Genet Dev 8:588–594

Spencer MR, Gastin PB (2001) Energy system contribution during 200- to 1,500-m running in highly trained athletes. Med Sci Sports Exerc 33:157–162

St Clair Gibson A, Noakes TD (2004) Evidence for complex system integration and dynamic neural regulation of skeletal muscle recruitment during exercise in humans. Br J Sports Med 38:797–806

Tian H, Hammer RE, Matsumoto AM, Russell DW, McKnight SL (1998) The hypoxia-responsive transcription factor EPAS1 is essential for catecholamine homeostasis and protection against

heart failure during embryonic development. Genes Dev 12:3320–3324
Van Damme R, Wilson RS, Vanhooydonck B, Aerts P (2002) Performance constraints in decathletes. Nature 415:755–756
Wang GL, Jiang BH, Rue EA, Semenza GL (1995) Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. Proc Natl Acad Sci USA 92:5510–5514
Wiesener MS, Jurgensen JS, Rosenberger C, Scholze CK, Horstrup JH, Warnecke C, Mandriota S, Bechmann I, Frei UA, Pugh CW, Ratcliffe PJ, Bachmann S, Maxwell PH, Eckardt KU (2003) Widespread hypoxia-inducible expression of HIF-2alpha in distinct cell populations of different organs. Faseb J 17:271–273
Wiesener MS, Turley H, Allen WE, Willam C, Eckardt KU, Talks KL, Wood SM, Gatter KC, Harris AL, Pugh CW, Ratcliffe PJ, Maxwell PH (1998) Induction of endothelial PAS domain protein-1 by hypoxia: characterization and comparison with hypoxia-inducible factor-1alpha. Blood 92:2260–2268

Chapter 6
Summary and Conclusions

6.1 Summary and Conclusions

The aims of the present study were to identify genes related to human performance and to assess if these genes affect standard human athletic performance variables.

6.1.1 Results to Emerge from the Present Study

1. The *ACE I/D* polymorphism was associated with groups of athletes within a variety of sports, including the novel sport of rugby. Multiple regression analysis consistently showed weight, but not *ACE I/D*, to be a significant predictor of performance.
2. The *MYBPC3* e6 S236G variant was, for the first time, tested within a variety of sports and shown to be associated with rugby players and sprint runners. Multiple regression analysis consistently showed weight and elite, but not *MYBPC3* e6 S236G, to be significant predictors of performance.
3. The *EPAS1* gene was identified as being associated with elite endurance athletes.
4. A novel *EPAS1* i8/e9 C_7/C_5GC polymorphism was identified but did not appear to be associated with elite endurance athletes. Multiple regression analysis consistently showed weight, but not *EPAS1* i8/e9 C_7/C_5GC, to be a significant predictor of performance.
5. The established technologies of PCR, RFLP and gel electrophoresis testing can still be used to test, economically and reliably, a relatively large number of DNA samples for known polymorphisms (*ACE I/D* and *MYBPC3* e6 S236G).

6. A novel gene (*EPAS1*) related to endurance athletes, was identified using a combination of *in silico* search techniques, high-throughput SNP screening and association studies.
7. The comparatively new technology of high-throughput SNP detection was successfully used to test a large number of DNA samples for a SNP association study.
8. The newer technology of DHPLC DNA variant detection was successfully used for screening a relatively large number of DNA samples for significant novel DNA variants.

6.1.2 Implications of the Present Study

1. The knowledge of the association of the DNA variants *ACE I/D*, *MYBPC3* e6 S236G, *EPAS1* SNP and *EPAS1* i8/e9 C_7/C_5GC polymorphisms with a variety of sports will lead to further investigations of the associations found in the present study and of the mechanisms of association. This may, ultimately, lead to improvements in the selection and training of athletes for specific sports or events.
2. This knowledge will lead to further understanding of how some of these genes act as modifying genes in various diseases of the cardiovascular system and so result in the modification of the treatment of patients with these conditions.
3. The established technologies of PCR, RFLP and gel electrophoresis testing can still be used to economically and reliably test a large number of DNA samples for known polymorphisms as a starting point for investigation of novel DNA variants before larger studies or full-scale SNP studies are performed.

4. A wide variety of phenotypes and genes can be investigated using a combination of *in silico* search techniques, high-throughput SNP screening and association studies.
5. High-throughput SNP detection can be successfully used for testing a wide variety of phenotypes and genes. As the International HapMap Project reaches completion, SNP studies should become even more effective.
6. DHPLC DNA variant detection can be effectively used for screening a wide variety of phenotypes and genes for novel variants.

6.2 Aims of the Present Study

The aims of the present study were to identify genes related to human performance and assess if these genes affect standard human athletic performance variables. These genes were investigated using a combination of *in silico* search techniques, association studies and multiple regression analysis. These genes were screened for known polymorphisms and for new polymorphisms using a combination of PCR, DHPLC and SNP technology. The variants found in athletes were compared to control groups for differences and compared within athlete groups for correlation with phenotypic data.

6.2.1 Hypotheses of the Present Study

The hypotheses for the present study are that there are genes responsible for human performance, and variations in these genes explain, to a degree, differences in human performances. These variations were shown to be in greater frequencies in elite athletes compared to controls. The variations in these genes, however, did not explain the differences in the performance parameters used for the present study. The differences detected between elite athletes and controls were probably due to other factors not measured in the present study.

6.2.2 Limitations of the Present Study

6.2.2.1 Phenotype

Phenotype limitations probably would have impacted on the results of this study. For the majority of sports the type of sport and/or $\dot{V}O_2$ max was being used as a proxy for performance in that sport. In another study of Olympic rowers, $\dot{V}O_2$ max explained 72% of variability in rowing performance (Cosgrove et al. 1999). In cycling, peak power output explained 94% of the variance in $\dot{V}O_2$ max and 82% of the variability in a 20 km cycling time trial (Hawley and Noakes 1992). Another study suggested that $\dot{V}O_2$ max is only predictive of cycling performance when coupled with lactate, power, metabolic thresholds and efficiency measures (Faria et al. 2005). A swimmer can achieve a higher $\dot{V}O_2$ max during running than swimming. $\dot{V}O_2$ max is, therefore, probably not limiting for performance in swimming (Holmer 1992). Ironman $\dot{V}O_2$ max scores were found to be comparable to cyclists and distance runners in the literature (Hue et al. 2000). $\dot{V}O_2$ max would probably have a similar predictive ability for Ironman performance as it does in cycling and running.

From the reviewed literature, in the majority of the sports in the present study, there was a moderate correlation between $\dot{V}O_2$ max and performance. This made finding a significant association of the genetic variants with $\dot{V}O_2$ max or performance difficult. Recent research from another laboratory has investigated the association of several genes with Ironman performance. They were unable to find an association for uncoupling protein 3 with Ironman athletes or in a regression model with Ironman time (Hudson et al. 2004). The same laboratory found an association for *ACE I/D*

with Ironman athletes but did not report any investigation of a regression model for *ACE I/D* with Ironman time (Collins et al. 2004). They were, however, able to demonstrate that a combination of bradykinin β2 receptor and nitric oxide synthase genotypes together with body mass index and age in a regression model were the statistically significant predictors of Ironman time ($p = 0.002$), and approximately 15% of the variance in Ironman time was accounted for by the linear combination of these variables (Saunders et al. 2006). The aforementioned study demonstrates that the effect size of genotype on performance can be large enough to influence race outcome. When the margin for victory in the Olympic Games is usually less than 0.5% (Kearney 1999), the genetic effects of just a few genes may mean the difference between winning an Olympic Gold Medal and missing the starting line.

Mass scaling or power ratios were not used for the multiple regression modelling in the present study. This was because weight was a dependent variable in the all the models where weight data were available. If weight scaling had been used for the outcome fitness variables, this would be adjusting for weight twice.

6.2.2.2 Subjects

Subjects were selected because they were elite athletes. Many reports have, however, suggested that a significant proportion of elite athletes use performance enhancing drugs (Kennedy 2004). If this is accurate, this could mean that some of our case subjects may have been misclassified as being of the elite phenotype. Scientific studies suggest, nevertheless, that less than 6% of athletes take performance enhancing drugs and that anabolic steroid usage declines progressively from high school to the elite level (Berning et al. 2004). At worst, this would be a relatively

small proportion of misclassified subjects and would not affect the results of the present study.

It could be argued that the choice of rugby players as subjects was not optimal because of probable genetic trade-offs and elite athletes was not optimal because of their small variance in $\dot{V}O_2$ max. However, rugby players were chosen precisely because they were examples of athletes with mixed athletic abilities and because the majority of participants in sport participate in sports requiring mixed athletic abilities, e.g. soccer, football, racket sports, etc. Elite athletes may have a small variance in $\dot{V}O_2$ max. However, baseline $\dot{V}O_2$ max and response $\dot{V}O_2$ max are relatively independent of one another (p66: "age, sex, race, and initial fitness level have little influence on $\dot{V}O_2$ max response (Skinner et al. 2001)"), so there is no reason that the genetic variance in these traits should not be detectable.

6.2.2.3 Sample Sizes

Numbers of elite athletes are limited by definition. The United States Olympic Committee defines elite as in the top eight in the world in a given sport or event (Kearney 1999). The definition of elite in the present study was much broader than this (see p80) but achieving sufficient statistical power in most of the smaller athlete groups was still difficult. Caution should be exercised in the assumption of exclusion of association for the non-significant results in the present study as most of the statistical tests were significantly under power.

6.2.2.4 Physiological Testing

There were limitations to the accuracy and comparability of the physiological testing in the various sports at different times of the year from diverse locations and testers (Hagerman and Staron 1983). Even though much of the standardised fitness testing occurred at certified laboratories experienced in such testing, at the time of the present study it was uncertain that test results could be compared from one location to the next. This was especially true of the fitness testing of the non-elite rugby players. It was anticipated that the relatively large number of subjects involved in the present study would provide sufficient statistical power to overcome this limitation.

6.2.2.5 Genome

Considering the sheer size of the human genome, 25,000–30,000 genes, there would obviously be many dozens, if not hundreds, of genes involved in human performance. The effect size of each gene variant would, for that reason, be quite small.

6.2.2.6 Genetic Tradeoffs

Physical performance is considered to be inhibited by trade-offs between antagonistic pairs of evolutionarily significant traits and between incompatible specialist and generalist phenotypes (Van Damme et al. 2002; Wilson and James 2004). The traits of endurance and speed could be thought of in the same way. Endurance is negatively correlated with weight (all sports, except Ironman, which had no weight data) and age (Ironman). Speed is positively correlated with strength and negatively correlated with age (rugby, −0.07) and weight (rugby, −0.02).

Rugby requires high levels of the physiological traits of speed, strength and endurance. Reaching maximal fitness levels, however, is not the main goal of rugby training. The main goal of rugby training is to be able to effectively perform the individual and team, skills and tactics of rugby. Rugby players might only do just enough fitness training to reach an acceptable standard for the level at which they play. From that point they might concentrate their energy on playing skills and recovering from the stress and strain of weekly competition. It may, therefore, be a matter of good rugby playing ability with acceptable genetic profile achieving selection at a certain level of rugby.

Reaching the elite level, in many sports, often requires selection at a young age in representative teams, to gain experience playing in more intense competition, improving skills, developing new strategies and to gain exposure to better coaching. Rugby players might forsake improved physiological performance goals in favour of the goal of increasing lean body mass since rugby performance ranking has been significantly correlated in another study with the average mass of the teams (Olds 2001).

6.2.2.7 Systemic

Feed-forward is a term describing a kind of system which reacts to changes in its environment, usually to maintain some desired state of the system. A system which exhibits feed-forward behavior responds to a measured disturbance in a pre-defined way. Feed-forward control can respond more quickly to known and measurable kinds of disturbances, but cannot do much with novel disturbances. The feed-forward-loop, a network motif in genetic regulatory networks, involves two transcription factors:

one regulates the expression of the second, and both transcription factors regulate the expression of an effector gene (Wall et al. 2005). If the gene of interest is part of a feed-forward system, the effect size of variants may be small unless the other gene/s in the system also has important variants.

6.2.2.8 Performance Factors

Many factors of human performance were not part of the present study including: psychological, socioeconomic and surroundings (see p19). A complete understanding of gene-environment interaction in relation to human performance is not possible until these factors can be accounted for.

6.2.3 Significance of the Present Study

The identification of performance genes and understanding the function of these genes, will lead to exciting advances in sports science. The understanding of modifying genes of the cardiovascular system will lead to breakthroughs in understanding disease causation and also provide alternative targets for new therapies.

The present study covers a wide range of popular international sports which makes it valuable to the field of sports genetics. This is the first time the genetic contribution to performance in rugby players has been examined. This is a popular team ball sport which would allow the results of the present study to be generalisable to similar sports.

6.3 Future Directions

Future research to further these results might include: better quantification of phenotype; exploring disease-phenotype correlations; comprehensive study of gene structure and expression patterns; gene structure and transcript mapping studies; studying gene expression using cultured cells or cell extracts in microarrays; identifying regulatory sequences through the use of reporter genes and DNA-protein interactions; investigating gene function by identifying interactions between a protein and other macromolecules; altered gene expression studies; using knockout or knockin mouse models; or therapeutic intervention in patients based on a better understanding of the function of these gene variants.

Athlete training intervention studies could be used to better understand the function of these gene variants. Ethical considerations (athlete samples were required to be de-identified) prevented researchers in the present study from being able to perform functional studies or interventional training studies with the athletes. In the future, interventional training studies would be desirable. A study for the *EPAS1* genotypes and gene expression in athletes, using altitude training and/or high intensity training as the stimuli, would lead to a greater understanding of how these types of training work and of the genetic mechanisms involved.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular biology of the cell, 4th edn. Garland Science, New York

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-10

Alvarez R, Terrados N, Ortolano R, Iglesias-Cubero G, Reguero JR, Batalla A, Cortina A, Fernandez-Garcia B, Rodriguez C, Braga S, Alvarez V, Coto E (2000) Genetic variation in the renin-angiotensin system and athletic performance. Eur J Appl Physiol 82: 117-20

Andrikopoulos GK, Richter DJ, Needham EW, Tzeis SE, Zairis MN, Gialafos EJ, Vogiatzi PG, Papasteriadis EG, Kardaras FG, Foussas SG, Gialafos JE, Stefanadis CI, Toutouzas PK, Mattu RK (2004) The paradoxical association of common polymorphisms of the renin-angiotensin system genes with risk of myocardial infarction. Eur J Cardiovasc Prev Rehabil 11: 477-483

Astrand PO, Saltin B (1961) Maximal oxygen uptake and heart rate in various types of muscular activity. J Appl Physiol 16: 977-81

Atkins SJ (2004) Normalizing expressions of strength in elite rugby league players. J Strength Cond Res 18: 53-8

Atkinson G, Davison R, Passfield L, Nevill AM (2003) Could the correlation between maximal oxygen uptake and "ECONOMY" be spurious? Med Sci Sports Exerc 35: 1242-3; author reply 1244

Banister EW, Carter JB, Zarkadas PC (1999) Training theory and taper: validation in triathlon athletes. Eur J Appl Physiol Occup Physiol 79: 182-91

Banister EW, Morton RH, Fitz-Clarke J (1992) Dose/response effects of exercise modeled from training: physical and biochemical measures. Ann Physiol Anthropol 11: 345-56

Barley J, Blackwood A, Miller M, Markandu ND, Carter ND, Jeffery S, Cappuccio FP, MacGregor GA, Sagnella GA (1996) Angiotensin converting enzyme gene I/D polymorphism, blood pressure and the renin-angiotensin system in Caucasian and Afro-Caribbean peoples. J Hum Hypertens 10: 31-5

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-5

Barton GJ (1995) Protein secondary structure prediction. Curr Opin Struct Biol 5: 372-6

Bassett DR, Jr., Howley ET (1997) Maximal oxygen uptake: "classical" versus "contemporary" viewpoints. Med Sci Sports Exerc 29: 591-603

Bassett DR, Jr., Howley ET (2000) Limiting factors for maximum oxygen uptake and determinants of endurance performance. Med Sci Sports Exerc 32: 70-84

Batterham AM, Tolfrey K, George KP (1997) Nevill's explanation of Kleiber's 0.75 mass exponent: an artifact of collinearity problems in least squares models? J Appl Physiol 82: 693-7

Becker KG (2004) The common variants/multiple disease hypothesis of common complex genetic disorders. Med Hypotheses 62: 309-17

Berg K (2003) Endurance training and performance in runners: research limitations and unanswered questions. Sports Med 33: 59-73

Bergh U, Sjodin B, Forsberg A, Svedenhag J (1991) The relationship between body mass and oxygen uptake during running in humans. Med Sci Sports Exerc 23: 205-11

Berning JM, Adams KJ, Stamford BA (2004) Anabolic steroid usage in athletics: facts, fiction, and public relations. J Strength Cond Res 18: 908-17

Bert AG, Grepin R, Vadas MA, Goodall GJ (2006) Assessing IRES activity in the HIF-1{alpha} and other cellular 5' UTRs. Rna

Bilzon JL, Allsopp AJ, Tipton MJ (2001) Assessment of physical fitness for occupations encompassing load-carriage tasks. Occup Med (Lond) 51: 357-61

Binevski PV, Sizova EA, Pozdnev VF, Kost OA (2003) Evidence for the negative cooperativity of the two active sites within bovine somatic angiotensin-converting enzyme. FEBS Lett 550: 84-8

Bleumink GS, Schut AF, Sturkenboom MC, Deckers JW, van Duijn CM, Stricker BH (2004) Genetic polymorphisms and heart failure. Genet Med 6: 465-74

Bloem LJ, Manatunga AK, Pratt JH (1996) Racial difference in the relationship of an angiotensin I-converting enzyme gene polymorphism to serum angiotensin I-converting enzyme activity. Hypertension 27: 62-6

Bossone E, Vriz O, Bodini BD, Rubenfire M (2004) Cardiovascular response to exercise in elite ice hockey players. Can J Cardiol 20: 893-7

Bouchard C, An P, Rice T, Skinner JS, Wilmore JH, Gagnon J, Perusse L, Leon AS, Rao DC (1999) Familial aggregation of VO(2max) response to exercise training: results from the HERITAGE Family Study. J Appl Physiol 87: 1003-8

Bouchard C, Daw EW, Rice T, Perusse L, Gagnon J, Province MA, Leon AS, Rao DC, Skinner JS, Wilmore JH (1998) Familial resemblance for VO2max in the sedentary state: the HERITAGE family study. Med Sci Sports Exerc 30: 252-8

Bouchard C, Leon AS, Rao DC, Skinner JS, Wilmore JH, Gagnon J (1995) The HERITAGE family study. Aims, design, and measurement protocol. Med Sci Sports Exerc 27: 721-9

Bouchard C, Malina RM (1983) Genetics of physiological fitness and motor performance. Exerc Sport Sci Rev 11: 306-39

Bouchard C, Rankinen T (2001) Individual differences in response to regular physical activity. Med Sci Sports Exerc 33: S446-51; discussion S452-3

Bouchard C, Rankinen T, Chagnon YC, Rice T, Perusse L, Gagnon J, Borecki I, An P, Leon AS, Skinner JS, Wilmore JH, Province M, Rao DC (2000) Genomic scan for maximal oxygen uptake and its response to training in the HERITAGE Family Study. J Appl Physiol 88: 551-9

Bramble DM, Lieberman DE (2004) Endurance running and the evolution of Homo. Nature 432: 345-52

Brooks JH, Fuller CW, Kemp SP, Reddin DB (2005) A prospective study of injuries and training amongst the England 2003 Rugby World Cup squad. Br J Sports Med 39: 288-93

Brown NJ, Blais C, Jr., Gandhi SK, Adam A (1998) ACE insertion/deletion genotype affects bradykinin metabolism. J Cardiovasc Pharmacol 32: 373-7

Brown TA (2002) Genomes, 2nd ed. edn. BIOS Scientific Publishers, Ltd, Oxford, UK

Buresh R, Berg K (2002) Scaling oxygen uptake to body size and several practical applications. J Strength Cond Res 16: 461-5

Caceres JF, Kornblihtt AR (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet 18: 186-93

Capek P, Brdicka R (2006) Hypertrophic cardiomyopathy. Cas Lek Cesk 145: 93-6; discussion 96-7

Carbone V, Savaglio S (2001) Scaling laws and forecasting in athletic world records. J Sports Sci 19: 477-84

Carrier L, Bonne G, Bahrend E, Yu B, Richard P, Niel F, Hainque B, Cruaud C, Gary F, Labeit S, Bouhour JB, Dubourg O, Desnos M, Hagege AA, Trent RJ, Komajda M, Fiszman M, Schwartz K (1997) Organization and sequence of human cardiac myosin binding protein C gene (MYBPC3) and identification of mutations predicted to produce truncated proteins in familial hypertrophic cardiomyopathy. Circ Res 80: 427-34

Chamari K, Moussa-Chamari I, Boussaidi L, Hachana Y, Kaouech F, Wisloff U (2005) Appropriate interpretation of aerobic capacity: allometric scaling in adult and young soccer players. Br J Sports Med 39: 97-101

Charron P, Dubourg O, Desnos M, Bennaceur M, Carrier L, Camproux AC, Isnard R, Hagege A, Langlard JM, Bonne G, Richard P, Hainque B, Bouhour JB, Schwartz K, Komajda M (1998) Clinical features and prognostic implications of familial hypertrophic cardiomyopathy related to the cardiac myosin-binding protein C gene. Circulation 97: 2230-6

Chatard JC, Wilson B (2003) Drafting distance in swimming. Med Sci Sports Exerc 35: 1176-81

Chung MW, Tsoutsman T, Semsarian C (2003) Hypertrophic cardiomyopathy: from gene defect to clinical disease. Cell Res 13: 9-20

Claessens AL, Veer FM, Stijnen V, Lefevre J, Maes H, Steens G, Beunen G (1991) Anthropometric characteristics of outstanding male and female gymnasts. J Sports Sci 9: 53-74

Clarkson PM, Devaney JM, Gordish-Dressman H, Thompson PD, Hubal MJ, Urso M, Price TB, Angelopoulos TJ, Gordon PM, Moyna NM, Pescatello LS, Visich PS, Zoeller RF, Seip RL, Hoffman EP (2005a) ACTN3 genotype is associated with increases in muscle strength in response to resistance training in women. J Appl Physiol 99: 154-63

Clarkson PM, Hoffman EP, Zambraski E, Gordish-Dressman H, Kearns A, Hubal M, Harmon B, Devaney JM (2005b) ACTN3 and MLCK genotype associations with exertional muscle damage. J Appl Physiol 99: 564-9

Clifford PS, Hanel B, Secher NH (1994) Arterial blood pressure response to rowing. Med Sci Sports Exerc 26: 715-9

Collins DW, Jukes TH (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20: 386-96

Collins M, Xenophontos SL, Cariolou MA, Mokone GG, Hudson DE, Anastasiades L, Noakes TD (2004) The ACE gene and endurance performance during the South African Ironman Triathlons. Med Sci Sports Exerc 36: 1314-20

Cooper DN, Nussbaum RL, Krawczak M (2002) Proposed guidelines for papers describing DNA polymorphism-disease associations. Hum Genet 110: 207-8

Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. Hum Genet 78: 151-5

Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366: 1121-31

Cosgrove MJ, Wilson J, Watt D, Grant SF (1999) The relationship between selected physiological variables of rowers and rowing performance as determined by a 2000 m ergometer test. J Sports Sci 17: 845-52

Couture L, Chagnon M, Allard C, Bouchard C (1986) More on red blood cell genetic variation in Olympic athletes. Can J Appl Sport Sci 11: 16-8

Coyle EF, Feltner ME, Kautz SA, Hamilton MT, Montain SJ, Baylor AM, Abraham LD, Petrek GW (1991) Physiological and biomechanical factors associated with elite endurance cycling performance. Med Sci Sports Exerc 23: 93-107

Crilley JG, Boehm EA, Blair E, Rajagopalan B, Blamire AM, Styles P, McKenna WJ, Ostman-Smith I, Clarke K, Watkins H (2003) Hypertrophic cardiomyopathy due to sarcomeric gene mutations is characterized by impaired energy metabolism irrespective of the degree of hypertrophy. J Am Coll Cardiol 41: 1776-82

Czubryt MP, Olson EN (2004) Balancing contractility and energy production: the role of myocyte enhancer factor 2 (MEF2) in cardiac hypertrophy. Recent Prog Horm Res 59: 105-24

Danser AH, Schalekamp MA, Bax WA, van den Brink AM, Saxena PR, Riegger GA, Schunkert H (1995) Angiotensin-converting enzyme in the human heart. Effect of the deletion/insertion polymorphism. Circulation 92: 1387-8

De Feo P, Di Loreto C, Lucidi P, Murdolo G, Parlanti N, De Cicco A, Piccioni F, Santeusanio F (2003) Metabolic response to exercise. J Endocrinol Invest 26: 851-4

deGaray A, Levine L, Carter J (1974) Genetic and Anthropological Studies of Olympic Athletes. Academic Press Inc., New York

Deutsch MU, Maw GJ, Jenkins D, Reaburn P (1998) Heart rate, blood lactate and kinematic data of elite colts (under-19) rugby union players during competition. J Sports Sci 16: 561-70

Diet F, Graf C, Mahnke N, Wassmer G, Predel HG, Palma-Hohmann I, Rost R, Bohm M (2001) ACE and angiotensinogen gene genotypes and left ventricular mass in athletes. Eur J Clin Invest 31: 836-42

Doolan G, Nguyen L, Chung J, Ingles J, Semsarian C (2004) Progression of left ventricular hypertrophy and the angiotensin-converting enzyme gene polymorphism in hypertrophic cardiomyopathy. Int J Cardiol 96: 157-63

Duhaylongsod FG, Griebel JA, Bacon DS, Wolfe WG, Piantadosi CA (1993) Effects of muscle contraction on cytochrome a,a3 redox state. J Appl Physiol 75: 790-7

Duthie G, Pyne D, Hooper S (2003) Applied physiology and game analysis of rugby union. Sports Med 33: 973-91

Duthie G, Pyne D, Hooper S (2005) Time motion analysis of 2001 and 2002 super 12 rugby. J Sports Sci 23: 523-30

Eisenmann JC, Pivarnik JM, Malina RM (2001) Scaling peak VO2 to body mass in young male and female distance runners. J Appl Physiol 90: 2172-80

Evans AE, Poirier O, Kee F, Lecerf L, McCrum E, Falconer T, Crane J, O'Rourke DF, Cambien F (1994) Polymorphisms of the angiotensin-converting-enzyme gene in subjects who die from coronary heart disease. Q J Med 87: 211-4

Faria EW, Parker DL, Faria IE (2005) The science of cycling: physiology and training - part 1. Sports Med 35: 285-312

Fatini C, Guazzelli R, Manetti P, Battaglini B, Gensini F, Vono R, Toncelli L, Zilli P, Capalbo A, Abbate R, Gensini GF, Galanti G (2000) RAS genes influence exercise-induced left ventricular hypertrophy: an elite athletes study. Med Sci Sports Exerc 32: 1868-72

Fernhall B, Kohrt W (1990) The effect of training specificity on maximal and submaximal physiological responses to treadmill and cycle ergometry. J Sports Med Phys Fitness 30: 268-75

Firth JD, Ebert BL, Pugh CW, Ratcliffe PJ (1994) Oxygen-regulated control elements in the phosphoglycerate kinase 1 and lactate dehydrogenase A genes: similarities with the erythropoietin 3' enhancer. Proc Natl Acad Sci U S A 91: 6496-500

Flashman E, Redwood C, Moolman-Smook J, Watkins H (2004) Cardiac myosin binding protein C: its role in physiology and disease. Circ Res 94: 1279-89

Francks C, Fisher SE, Marlow AJ, MacPhie IL, Taylor KE, Richardson AJ, Stein JF, Monaco AP (2003) Familial and genetic effects on motor coordination, laterality, and reading-related cognition. Am J Psychiatry 160: 1970-7

Freedman SJ, Sun ZY, Poy F, Kung AL, Livingston DM, Wagner G, Eck MJ (2002) Structural basis for recruitment of CBP/p300 by hypoxia-inducible factor-1 alpha. Proc Natl Acad Sci U S A 99: 5367-72

Fukashiro S, Abe T, Shibayama A, Brechue WF (2002) Comparison of viscoelastic characteristics in triceps surae between Black and White athletes. Acta Physiol Scand 175: 183-7

Futterman LG, Myerburg R (1998) Sudden death in athletes: an update. Sports Med 26: 335-50

Gastin PB (2001) Energy system interaction and relative contribution during maximal exercise. Sports Med 31: 725-41

Gayagay G, Yu B, Hambly B, Boston T, Hahn A, Celermajer DS, Trent RJ (1998) Elite endurance athletes and the ACE I allele--the role of genes in athletic performance. Hum Genet 103: 48-50

Giaccia AJ, Simon MC, Johnson R (2004) The biology of hypoxia: the role of oxygen sensing in development, normal function, and disease. Genes Dev 18: 2183-94

Gleeson M, McDonald WA, Cripps AW, Pyne DB, Clancy RL, Fricker PA (1995) The effect on immunity of long-term intensive training in elite swimmers. Clin Exp Immunol 102: 210-6

Gleeson M, Nieman DC, Pedersen BK (2004) Exercise, nutrition and immune function. J Sports Sci 22: 115-25

Gonzalez AJ, Hernandez D, De Vera A, Barrios Y, Salido E, Torres A, Terrados N (2006) ACE gene polymorphism and erythropoietin in endurance athletes at moderate altitude. Med Sci Sports Exerc 38: 688-93

Goradia TM, Lange K, Miller PL, Nadkarni PM (1992) Fast computation of genetic likelihoods on human pedigree data. Hum Hered 42: 42-62

Gordon D, Finch SJ (2005) Factors affecting statistical power in the detection of genetic association. J Clin Invest 115: 1408-18

Gothie E, Richard DE, Berra E, Pages G, Pouyssegur J (2000) Identification of alternative spliced variants of human hypoxia-inducible factor-1alpha. J Biol Chem 275: 6922-7

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. Trends Genet 17: 100-7

Grivetti LE, Applegate EA (1997) From Olympia to Atlanta: a cultural-historical perspective on diet and athletic training. J Nutr 127: 860S-868S

Hagel B (2005) Hamstring injuries in Australian football. Clin J Sport Med 15: 400

Hagerman FC (1984) Applied physiology of rowing. Sports Med 1: 303-26

Hagerman FC, Staron RS (1983) Seasonal variables among physiological variables in elite oarsmen. Can J Appl Sport Sci 8: 143-8

Hanel B, Clifford PS, Secher NH (1994) Restricted postexercise pulmonary diffusion capacity does not impair maximal transport for O2. J Appl Physiol 77: 2408-12

Harris SP, Bartley CR, Hacker TA, McDonald KS, Douglas PS, Greaser ML, Powers PA, Moss RL (2002) Hypertrophic cardiomyopathy in cardiac myosin binding protein-C knockout mice. Circ Res 90: 594-601
Hattersley AT, McCarthy MI (2005) What makes a good genetic association study? Lancet 366: 1315-23
Hawley JA, Noakes TD (1992) Peak power output predicts maximal oxygen uptake and performance time in trained cyclists. Eur J Appl Physiol Occup Physiol 65: 79-83
Heil DP (1997) Body mass scaling of peak oxygen uptake in 20- to 79-yr-old adults. Med Sci Sports Exerc 29: 1602-8
Heiss HW, Wink K, Barmeyer J, Keul J, Reindell H (1977) Myocardial oxygen consumption and substrate uptake in man during physiological and pathological volume load. Basic Res Cardiol 72: 293-8
Henderson J, Withford-Cave JM, Duffy DL, Cole SJ, Sawyer NA, Gulbin JP, Hahn A, Trent RJ, Yu B (2005) The EPAS1 gene influences the aerobic-anaerobic contribution in elite endurance athletes. Hum Genet 118: 416-23
Hernandez D, de la Rosa A, Barragan A, Barrios Y, Salido E, Torres A, Martin B, Laynez I, Duque A, De Vera A, Lorenzo V, Gonzalez A (2003) The ACE/DD genotype is associated with the extent of exercise-induced left ventricular growth in endurance athletes. J Am Coll Cardiol 42: 527-32
Holden C (2004) Peering under the hood of Africa's runners. Science 305: 637-9
Holmer I (1992) Swimming physiology. Ann Physiol Anthropol 11: 269-76
Holmes DI, Zachary I (2005) The vascular endothelial growth factor (VEGF) family: angiogenic factors in health and disease. Genome Biol 6: 209
Hopkins WG, Hawley JA, Burke LM (1999) Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc 31: 472-85
Hudson DE, Mokone GG, Noakes TD, Collins M (2004) The -55 C/T polymorphism within the UCP3 gene and performance during the South African Ironman Triathlon. Int J Sports Med 25: 427-32
Hue O, Le Gallais D, Chollet D, Prefaut C (2000) Ventilatory threshold and maximal oxygen uptake in present triathletes. Can J Appl Physiol 25: 102-13
Iyer NV, Leung SW, Semenza GL (1998) The human hypoxia-inducible factor 1alpha gene: HIF1A structure and evolutionary conservation. Genomics 52: 159-65
Jaaskelainen P, Kuusisto J, Miettinen R, Karkkainen P, Karkkainen S, Heikkinen S, Peltola P, Pihlajamaki J, Vauhkonen I, Laakso M (2002) Mutations in the cardiac myosin-binding protein C gene are the predominant cause of familial hypertrophic cardiomyopathy in eastern Finland. J Mol Med 80: 412-22
Jardine MA, Wiggins TM, Myburgh KH, Noakes TD (1988) Physiological characteristics of rugby players including muscle glycogen content and muscle fibre composition. S Afr Med J 73: 529-32
Jensen K, Johansen L, Secher NH (2001) Influence of body mass on maximal oxygen uptake: effect of sample size. Eur J Appl Physiol 84: 201-5
Jiang BH, Semenza GL, Bauer C, Marti HH (1996) Hypoxia-inducible factor 1 levels vary exponentially over a physiologically relevant range of O2 tension. Am J Physiol 271: C1172-80
Jones AM, Carter H (2000) The effect of endurance training on parameters of aerobic fitness. Sports Med 29: 373-86
Jones JH, Lindstedt SL (1993) Limits to maximal performance. Annu Rev Physiol 55: 547-69

Kaplan NM, Kem DC, Holland OB, Kramer NJ, Higgins J, Gomez-Sanchez C (1976) The intravenous furosemide test: a simple way to evaluate renin responsiveness. Ann Intern Med 84: 639-45
Kasper LH, Boussouar F, Boyd K, Xu W, Biesen M, Rehg J, Baudino TA, Cleveland JL, Brindle PK (2005) Two transactivation mechanisms cooperate for the bulk of HIF-1-responsive gene expression. Embo J 24: 3846-58
Kearney JT (1999) Sport performance enhancement: design and analysis of research. Med Sci Sports Exerc 31: 755-7
Keller DI, Coirault C, Rau T, Cheav T, Weyand M, Amann K, Lecarpentier Y, Richard P, Eschenhagen T, Carrier L (2004) Human homozygous R403W mutant cardiac myosin presents disproportionate enhancement of mechanical and enzymatic properties. J Mol Cell Cardiol 36: 355-62
Kennedy MC (2004) Drugs, sport and the Olympics 2000-2004. Med J Aust 181: 227
Klissouras V (1971) Heritability of adaptive variation. J Appl Physiol 31: 338-44
Kohrt WM, Malley MT, Coggan AR, Spina RJ, Ogawa T, Ehsani AA, Bourey RE, Martin WH, 3rd, Holloszy JO (1991) Effects of gender, age, and fitness level on response of VO2max to training in 60-71 yr olds. J Appl Physiol 71: 2004-11
Korte FS, McDonald KS, Harris SP, Moss RL (2003) Loaded shortening, power output, and rate of force redevelopment are increased with knockout of cardiac myosin binding protein-C. Circ Res 93: 752-8
Kozak M (1996) Interpreting cDNA sequences: some insights from studies on translation. Mamm Genome 7: 563-74
Krawczak M, Schmidtke J (1998) DNA Fingerprinting, 2nd Ed. edn. BIOS Scientific Publishers, Oxford
Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57: 439-54
Kulikovskaya I, McClellan G, Flavigny J, Carrier L, Winegrad S (2003) Effect of MyBP-C binding to actin on contractility in heart muscle. J Gen Physiol 122: 761-74
Kunst G, Kress KR, Gruen M, Uttenweiler D, Gautel M, Fink RH (2000) Myosin binding protein C, a phosphorylation-dependent force regulator in muscle that controls the attachment of myosin heads by its interaction with myosin S2. Circ Res 86: 51-8
Lalouel JM (2001) From genetics to mechanism of disease liability. Adv Genet 42: 517-33
Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11: 241-7
Landers GJ, Blanksby BA, Ackland TR, Smith D (2000) Morphology and performance of world championship triathletes. Ann Hum Biol 27: 387-400
Latchman DS (2001) Transcription factors: bound to activate or repress. Trends Biochem Sci 26: 211-3
Laursen PB, Rhodes EC (2001) Factors affecting performance in an ultraendurance triathlon. Sports Med 31: 195-209
Lavoie JL, Sigmund CD (2003) Minireview: overview of the renin-angiotensin system--an endocrine and paracrine system. Endocrinology 144: 2179-83
Lechin M, Quinones MA, Omran A, Hill R, Yu QT, Rakowski H, Wigle D, Liew CC, Sole M, Roberts R, et al. (1995) Angiotensin-I converting enzyme genotypes and left ventricular hypertrophy in patients with hypertrophic cardiomyopathy. Circulation 92: 1808-12

Lortie G, Simoneau JA, Hamel P, Boulay MR, Landry F, Bouchard C (1984) Responses of maximal aerobic power and capacity to aerobic training. Int J Sports Med 5: 232-6

Lucia A, Gomez-Gallego F, Chicharro JL, Hoyos J, Celaya K, Cordova A, Villa G, Alonso JM, Barriopedro M, Perez M, Earnest CP (2005) Is there an association between ACE and CKMM polymorphisms and cycling performance status during 3-week races? Int J Sports Med 26: 442-7

Lucia A, Hoyos J, Perez M, Santalla A, Chicharro JL (2002) Inverse relationship between VO2max and economy/efficiency in world-class cyclists. Med Sci Sports Exerc 34: 2079-84

MacArthur DG, North KN (2004) A gene for speed? The evolution and function of alpha-actinin-3. Bioessays 26: 786-95

Macarthur DG, North KN (2005) Genes and human elite athletic performance. Hum Genet 116: 331-9

MacDougall JD, McKelvie RS, Moroz DE, Sale DG, McCartney N, Buick F (1992) Factors affecting blood pressure during heavy weight lifting and static contractions. J Appl Physiol 73: 1590-7

Maemura K, Hsieh CM, Jain MK, Fukumoto S, Layne MD, Liu Y, Kourembanas S, Yet SF, Perrella MA, Lee ME (1999) Generation of a dominant-negative mutant of endothelial PAS domain protein 1 by deletion of a potent C-terminal transactivation domain. J Biol Chem 274: 31565-70

Maron BJ, Thompson PD, Puffer JC, McGrew CA, Strong WB, Douglas PS, Clark LT, Mitten MJ, Crawford MH, Atkins DL, Driscoll DJ, Epstein AE (1996) Cardiovascular preparticipation screening of competitive athletes. A statement for health professionals from the Sudden Death Committee (clinical cardiology) and Congenital Cardiac Defects Committee (cardiovascular disease in the young), American Heart Association. Circulation 94: 850-6

Martino M, Gledhill N, Jamnik V (2002) High VO2max with no history of training is primarily due to high blood volume. Med Sci Sports Exerc 34: 966-71

Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6: 386-98

Maud PJ (1983) Physiological and anthropometric parameters that describe a rugby union team. Br J Sports Med 17: 16-23

Mayeux R (2005) Mapping the new frontier: complex genetic disorders. J Clin Invest 115: 1404-7

McLean DA (1992) Analysis of the physical demands of international rugby union. J Sports Sci 10: 285-96

Medbo JI, Tabata I (1989) Relative importance of aerobic and anaerobic energy release during short-lasting exhausting bicycle exercise. J Appl Physiol 67: 1881-6

Meurs KM, Sanchez X, David RM, Bowles NE, Towbin JA, Reiser PJ, Kittleson JA, Munro MJ, Dryburgh K, Macdonald KA, Kittleson MD (2005) A cardiac myosin binding protein C mutation in the maine coon cat with familial hypertrophic cardiomyopathy. Hum Mol Genet

Montgomery HE, Clarkson P, Dollery CM, Prasad K, Losi MA, Hemingway H, Statters D, Jubb M, Girvain M, Varnava A, World M, Deanfield J, Talmud P, McEwan JR, McKenna WJ, Humphries S (1997) Association of angiotensin-converting enzyme gene I/D polymorphism with change in left ventricular mass in response to physical training. Circulation 96: 741-7

Moolman-Smook J, Flashman E, de Lange W, Li Z, Corfield V, Redwood C, Watkins H (2002) Identification of novel interactions between domains of Myosin binding protein-C that are modulated by hypertrophic cardiomyopathy missense mutations. Circ Res 91: 704-11

Morgan DW, Pate R (2004) Could the correlation between maximal oxygen uptake and "economy" be spurious? Med Sci Sports Exerc 36: 345

Morton RH, Fitz-Clarke JR, Banister EW (1990) Modeling human performance in running. J Appl Physiol 69: 1171-7

Moseley L, Achten J, Martin JC, Jeukendrup AE (2004) No differences in cycling efficiency between world-class and recreational cyclists. Int J Sports Med 25: 374-9

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1: 263-73

Mullis KB (1990) Target amplification for DNA analysis by the polymerase chain reaction. Ann Biol Clin (Paris) 48: 579-82

Myerson S, Hemingway H, Budget R, Martin J, Humphries S, Montgomery H (1999) Human angiotensin I-converting enzyme gene and endurance performance. J Appl Physiol 87: 1313-6

Nagashima J, Musha H, Takada H, Awaya T, Oba H, Mori N, Ohmiya K, Nobuoka S, Murayama M (2000) Influence of angiotensin-converting enzyme gene polymorphism on development of athlete's heart. Clin Cardiol 23: 621-4

Nazarov IB, Woods DR, Montgomery HE, Shneider OV, Kazakov VI, Tomilin NV, Rogozkin VA (2001) The angiotensin converting enzyme I/D polymorphism in Russian athletes. Eur J Hum Genet 9: 797-801

Nevill A, Rowland T, Goff D, Martel L, Ferrone L (2004a) Scaling or normalising maximum oxygen uptake to predict 1-mile run time in boys. Eur J Appl Physiol 92: 285-8

Nevill AM, Holder RL (1995) Scaling, normalizing, and per ratio standards: an allometric modeling approach. J Appl Physiol 79: 1027-31

Nevill AM, Jobson SA, Palmer GS, Olds TS (2005) Scaling maximal oxygen uptake to predict cycling time-trial performance in the field: a non-linear approach. Eur J Appl Physiol 94: 705-10

Nevill AM, Stewart AD, Olds T, Holder R (2004b) Are adult physiques geometrically similar? The dangers of allometric scaling using body mass power laws. Am J Phys Anthropol 124: 177-82

Newton-Cheh C, Hirschhorn JN (2005) Genetic association studies of complex traits: design and analysis issues. Mutat Res 573: 54-69

Nicholas CW (1997) Anthropometric and physiological characteristics of rugby union football players. Sports Med 23: 375-96

Nicholl JP, Coleman P, Williams BT (1995) The epidemiology of sports and exercise related injury in the United Kingdom. Br J Sports Med 29: 232-8

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19: 233-40

Niemi AK, Majamaa K (2005) Mitochondrial DNA and ACTN3 genotypes in Finnish elite endurance and sprint athletes. Eur J Hum Genet 13: 965-9

Niimura H, Patton KK, McKenna WJ, Soults J, Maron BJ, Seidman JG, Seidman CE (2002) Sarcomere protein gene mutations in hypertrophic cardiomyopathy of the elderly. Circulation 105: 446-51
Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70: 157-69
Noakes TD, Tucker R (2004) Inverse relationship between VO2max and economy in world-class cyclists. Med Sci Sports Exerc 36: 1083-4; author reply 1085-6
Norton K, Olds T (2001) Morphological evolution of athletes over the 20th century: causes and consequences. Sports Med 31: 763-83
Nuutila P, Knuuti MJ, Heinonen OJ, Ruotsalainen U, Teras M, Bergman J, Solin O, Yki-Jarvinen H, Voipio-Pulkki LM, Wegelius U, et al. (1994) Different alterations in the insulin-stimulated glucose uptake in the athlete's heart and skeletal muscle. J Clin Invest 93: 2267-74
Oakley CE, Hambly BD, Curmi PM, Brown LJ (2004) Myosin binding protein C: structural abnormalities in familial hypertrophic cardiomyopathy. Cell Res 14: 95-110
Oakley D (2001) General cardiology: The athlete's heart. Heart 86: 722-6
Olds T (2001) The evolution of physique in male rugby union players in the twentieth century. J Sports Sci 19: 253-62
Olds TS, Norton KI, Lowe EL, Olive S, Reay F, Ly S (1995) Modeling road-cycling performance. J Appl Physiol 78: 1596-611
O'Rourke JF, Tian YM, Ratcliffe PJ, Pugh CW (1999) Oxygen-regulated and transactivating domains in endothelial PAS protein 1: comparison with hypoxia-inducible factor-1alpha. J Biol Chem 274: 2060-71
Ortlepp JR, Vosberg HP, Reith S, Ohme F, Mahon NG, Schroder D, Klues HG, Hanrath P, McKenna WJ (2002) Genetic polymorphisms in the renin-angiotensin-aldosterone system associated with expression of left ventricular hypertrophy in hypertrophic cardiomyopathy: a study of five polymorphic genes in a family with a disease causing mutation in the myosin binding protein C gene. Heart 87: 270-5
O'Toole ML, Hiller DB, Crosby LO, Douglas PS (1987) The ultraendurance triathlete: a physiological profile. Med Sci Sports Exerc 19: 45-50
Padilla S, Mujika I, Cuesta G, Goiriena JJ (1999) Level ground and uphill cycling ability in professional road cycling. Med Sci Sports Exerc 31: 878-85
Palmer BM, Georgakopoulos D, Janssen PM, Wang Y, Alpert NR, Belardi DF, Harris SP, Moss RL, Burgon PG, Seidman CE, Seidman JG, Maughan DW, Kass DA (2004a) Role of cardiac myosin binding protein C in sustaining left ventricular systolic stiffening. Circ Res 94: 1249-55
Palmer BM, Noguchi T, Wang Y, Heim JR, Alpert NR, Burgon PG, Seidman CE, Seidman JG, Maughan DW, LeWinter MM (2004b) Effect of cardiac myosin binding protein-C on mechanoenergetics in mouse myocardium. Circ Res 94: 1615-22
Pate RR, Macera CA, Bailey SP, Bartoli WP, Powell KE (1992) Physiological, anthropometric, and training correlates of running economy. Med Sci Sports Exerc 24: 1128-33
Pelliccia A, Maron BJ (2001) Athlete's heart electrocardiogram mimicking hypertrophic cardiomyopathy. Curr Cardiol Rep 3: 147-51
Pelliccia A, Maron BJ, Culasso F, Spataro A, Caselli G (1996) Athlete's heart in women. Echocardiographic characterization of highly trained elite female athletes. Jama 276: 211-5

Pelliccia A, Spataro A, Caselli G, Maron BJ (1993) Absence of left ventricular wall thickening in athletes engaged in intense power training. Am J Cardiol 72: 1048-54

Perkins MJ, Van Driest SL, Ellsworth EG, Will ML, Gersh BJ, Ommen SR, Ackerman MJ (2005) Gene-specific modifying effects of pro-LVH polymorphisms involving the renin-angiotensin-aldosterone system among 389 unrelated patients with hypertrophic cardiomyopathy. Eur Heart J 26: 2457-62

Perusse L, Rankinen T, Rauramaa R, Rivera MA, Wolfarth B, Bouchard C (2003) The human gene map for performance and health-related fitness phenotypes: the 2002 update. Med Sci Sports Exerc 35: 1248-64

Price DA, Fisher ND (2003) The renin-angiotensin system in blacks: active, passive, or what? Curr Hypertens Rep 5: 225-30

Prior SJ, Hagberg JM, Phares DA, Brown MD, Fairfull L, Ferrell RE, Roth SM (2003) Sequence variation in hypoxia-inducible factor 1alpha (HIF1A): association with maximal oxygen consumption. Physiol Genomics 15: 20-6

Province MA, Rice TK, Borecki IB, Gu C, Kraja A, Rao DC (2003) Multivariate and multilocus variance components method, based on structural relationships to assess quantitative trait linkage via SEGPATH. Genet Epidemiol 24: 128-38

Pugh CW, O'Rourke JF, Nagao M, Gleadle JM, Ratcliffe PJ (1997) Activation of hypoxia-inducible factor-1; definition of regulatory domains within the alpha subunit. J Biol Chem 272: 11205-14

Quarrie KL, Handcock P, Waller AE, Chalmers DJ, Toomey MJ, Wilson BD (1995) The New Zealand rugby injury and performance project. III. Anthropometric and physical performance characteristics of players. Br J Sports Med 29: 263-70

Ramirez CD, Padron R (2004) [Familial hypertrophic cardiomyopathy: genes, mutations and animal models. A review]. Invest Clin 45: 69-99

Ranade K, Chang MS, Ting CT, Pei D, Hsiao CF, Olivier M, Pesich R, Hebert J, Chen YD, Dzau VJ, Curb D, Olshen R, Risch N, Cox DR, Botstein D (2001) High-throughput genotyping with single nucleotide polymorphisms. Genome Res 11: 1262-8

Rankinen T, Perusse L, Gagnon J, Chagnon YC, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C (2000a) Angiotensin-converting enzyme ID polymorphism and fitness phenotype in the HERITAGE Family Study. J Appl Physiol 88: 1029-35

Rankinen T, Perusse L, Rauramaa R, Rivera MA, Wolfarth B, Bouchard C (2001) The human gene map for performance and health-related fitness phenotypes. Med Sci Sports Exerc 33: 855-67

Rankinen T, Wolfarth B, Simoneau JA, Maier-Lenz D, Rauramaa R, Rivera MA, Boulay MR, Chagnon YC, Perusse L, Keul J, Bouchard C (2000b) No association between the angiotensin-converting enzyme ID polymorphism and elite endurance athlete status. J Appl Physiol 88: 1571-5

Richard P, Charron P, Carrier L, Ledeuil C, Cheav T, Pichereau C, Benaiche A, Isnard R, Dubourg O, Burban M, Gueffet JP, Millaire A, Desnos M, Schwartz K, Hainque B, Komajda M (2003) Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. Circulation 107: 2227-32

Rico-Sanz J, Rankinen T, Joanisse DR, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C (2003a) Associations between cardiorespiratory responses to

exercise and the C34T AMPD1 gene polymorphism in the HERITAGE Family Study. Physiol Genomics 14: 161-6
Rico-Sanz J, Rankinen T, Joanisse DR, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C (2003b) Familial resemblance for muscle phenotypes in the HERITAGE Family Study. Med Sci Sports Exerc 35: 1360-6
Rico-Sanz J, Rankinen T, Rice T, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C (2004) Quantitative trait loci for maximal exercise capacity phenotypes and their responses to training in the HERITAGE Family Study. Physiol Genomics 16: 256-60
Riechman SE, Zoeller RF, Balasekaran G, Goss FL, Robertson RJ (2002) Prediction of 2000 m indoor rowing performance using a 30 s sprint and maximal oxygen uptake. J Sports Sci 20: 681-7
Rigat B, Hubert C, Alhenc-Gelas F, Cambien F, Corvol P, Soubrier F (1990) An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. J Clin Invest 86: 1343-6
Riordan JF (2003) Angiotensin-I-converting enzyme and its relatives. Genome Biol 4: 225
Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-7
Rizzo M, Gensini F, Fatini C, Manetti P, Pucci N, Capalbo A, Vono MC, Galanti G (2003) ACE I/D polymorphism and cardiac adaptations in adolescent athletes. Med Sci Sports Exerc 35: 1986-90
Rogers DM, Olson BL, Wilmore JH (1995) Scaling for the VO2-to-body size relationship among children and adults. J Appl Physiol 79: 958-67
Rossi GP, Narkiewicz K, Cesari M, Winnicki M, Bigda J, Chrostowska M, Szczech R, Pawlowski R, Pessina AC (1999) Genetic determinants of plasma ACE and renin activity in young normotensive twins. J Hypertens 17: 647-55
Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134: 204-18
Ruas JL, Poellinger L, Pereira T (2002) Functional analysis of hypoxia-inducible factor-1 alpha-mediated transactivation. Identification of amino acid residues critical for transcriptional activation and/or interaction with CREB-binding protein. J Biol Chem 277: 38723-30
Sabbah HN, Shimoyama H, Sharov VG, Kono T, Gupta RC, Lesch M, Levine TB, Goldstein S (1996) Effects of ACE inhibition and beta-blockade on skeletal muscle fiber types in dogs with moderate heart failure. Am J Physiol 270: H115-20
Sadayappan S, Gulick J, Osinska H, Martin LA, Hahn HS, Dorn Ii GW, Klevitsky R, Seidman CE, Seidman JG, Robbins J (2005) Cardiac Myosin-Binding Protein-C Phosphorylation and Cardiac Function. Circ Res
Sands WA, McNeal JR, Stone MH (2005) Plaudits and pitfalls in studying elite athletes. Percept Mot Skills 100: 22-4
Saunders CJ, Xenophontos SL, Cariolou MA, Anastassiades LC, Noakes TD, Collins M (2006) The bradykinin {beta}2 receptor (BDKRB2) and endothelial nitric oxide synthase 3 (NOS3) genes and endurance performance during Ironman Triathlons. Hum Mol Genet 15: 979-87
Saunders PU, Pyne DB, Telford RD, Hawley JA (2004) Factors affecting running economy in trained distance runners. Sports Med 34: 465-85

Sayed-Tabatabaei FA, Oostra BA, Isaacs A, van Duijn CM, Witteman JC (2006) ACE polymorphisms. Circ Res 98: 1123-33
Scharhag J, Schneider G, Urhausen A, Rochette V, Kramann B, Kindermann W (2002) Athlete's heart: right and left ventricular mass and function in male endurance athletes and untrained individuals determined by magnetic resonance imaging. J Am Coll Cardiol 40: 1856-63
Schunkert H (1997) Polymorphism of the angiotensin-converting enzyme gene and cardiovascular disease. J Mol Med 75: 867-75
Schut AF, Bleumink GS, Stricker BH, Hofman A, Witteman JC, Pols HA, Deckers JW, Deinum J, van Duijn CM (2004) Angiotensin converting enzyme insertion/deletion polymorphism and the risk of heart failure in hypertensive subjects. Eur Heart J 25: 2143-8
Scortegagna M, Ding K, Oktay Y, Gaur A, Thurmond F, Yan LJ, Marck BT, Matsumoto AM, Shelton JM, Richardson JA, Bennett MJ, Garcia JA (2003a) Multiple organ pathology, metabolic abnormalities and impaired homeostasis of reactive oxygen species in Epas1-/- mice. Nat Genet 35: 331-40
Scortegagna M, Morris MA, Oktay Y, Bennett M, Garcia JA (2003b) The HIF family member EPAS1/HIF-2alpha is required for normal hematopoiesis in mice. Blood 102: 1634-40
Scott AC, Roe N, Coats AJ, Piepoli MF (2003) Aerobic exercise physiology in a professional rugby union team. Int J Cardiol 87: 173-7
Secher NH (2003) Are the lungs built for rowing? Scand J Med Sci Sports 13: 337-8
Semenza GL (1994a) Regulation of erythropoietin production. New insights into molecular mechanisms of oxygen homeostasis. Hematol Oncol Clin North Am 8: 863-84
Semenza GL (1994b) Transcriptional regulation of gene expression: mechanisms and pathophysiology. Hum Mutat 3: 180-99
Semenza GL, Agani F, Booth G, Forsythe J, Iyer N, Jiang BH, Leung S, Roe R, Wiener C, Yu A (1997) Structural and functional analysis of hypoxia-inducible factor 1. Kidney Int 51: 553-5
Semenza GL, Jiang BH, Leung SW, Passantino R, Concordet JP, Maire P, Giallongo A (1996) Hypoxia response elements in the aldolase A, enolase 1, and lactate dehydrogenase A gene promoters contain essential binding sites for hypoxia-inducible factor 1. J Biol Chem 271: 32529-37
Semenza GL, Roth PH, Fang HM, Wang GL (1994) Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1. J Biol Chem 269: 23757-63
Sham PC, Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59 (Pt 1): 97-105
Sims JM, Patzer B, Kumudavalli-Reddy M, Martin AF, Rabinowitz M, Zak R (1976) The pathways of protein synthesis and degradation in normal heart and during development and regression of cardiac hypertrophy. Recent Adv Stud Cardiac Struct Metab 12: 19-28
Skinner JS, Jaskolski A, Jaskolska A, Krasnoff J, Gagnon J, Leon AS, Rao DC, Wilmore JH, Bouchard C (2001) Age, sex, race, initial fitness, and response to training: the HERITAGE Family Study. J Appl Physiol 90: 1770-6
Skinner JS, Wilmore KM, Krasnoff JB, Jaskolski A, Jaskolska A, Gagnon J, Province MA, Leon AS, Rao DC, Wilmore JH, Bouchard C (2000) Adaptation to a standardized training program and changes in fitness in a large, heterogeneous population: the HERITAGE Family Study. Med Sci Sports Exerc 32: 157-61

Smith DJ (2003) A framework for understanding the training process leading to elite performance. Sports Med 33: 1103-26
Soubrier F, Alhenc-Gelas F, Hubert C, Allegrini J, John M, Tregear G, Corvol P (1988) Two putative active centers in human angiotensin I-converting enzyme revealed by molecular cloning. Proc Natl Acad Sci U S A 85: 9386-90
Squire JM, Luther PK, Knupp C (2003) Structural evidence for the interaction of C-protein (MyBP-C) with actin and sequence identification of a possible actin-binding domain. J Mol Biol 331: 713-24
Staley JP, Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell 92: 315-26
Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H (2005) Function of alternative splicing. Gene 344: 1-20
Strachan T, Read AP (1999) Human Molecular Genetics 2, 2nd ed. edn. BIOS Scientific Publishers, Ltd, Oxford, UK
SUPAMAC (2005a) DHPLC.
SUPAMAC (2005b) SNPs/Real Time PCR.
Svedahl K, MacIntosh BR (2003) Anaerobic threshold: the concept and methods of measurement. Can J Appl Physiol 28: 299-323
Swain DP (1994) The influence of body mass in endurance bicycling. Med Sci Sports Exerc 26: 58-63
Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res 8: 748-54
Takahashi N, Hagaman JR, Kim HS, Smithies O (2003) Minireview: computer simulations of blood pressure regulation by the renin-angiotensin system. Endocrinology 144: 2184-90
Taylor RR, Mamotte CD, Fallon K, van Bockxmeer FM (1999) Elite athletes and the gene for angiotensin-converting enzyme. J Appl Physiol 87: 1035-7
Telford R (2003) Personal communication. Genetics in Sport Forum. Australian Institute of Sport, Canberra
Thompson PD, Tsongalis GJ, Ordovas JM, Seip RL, Bilbie C, Miles M, Zoeller R, Visich P, Gordon P, Angelopoulos TJ, Pescatello L, Moyna N (2006) Angiotensin-converting enzyme genotype and adherence to aerobic exercise training. Prev Cardiol 9: 21-4
Tian H, McKnight SL, Russell DW (1997) Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. Genes Dev 11: 72-82
Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F (1992) Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet 51: 197-205
Tong CW, Gaffin RD, Zawieja DC, Muthuchamy M (2004) Roles of phosphorylation of myosin binding protein-C and troponin I in mouse cardiac muscle twitch dynamics. J Physiol 558: 927-41
Toussaint HM, Hollander AP (1994) Energetics of competitive swimming. Implications for training programmes. Sports Med 18: 384-405
Tsianos G, Sanders J, Dhamrait S, Humphries S, Grant S, Montgomery H (2004) The ACE gene insertion/deletion polymorphism and elite endurance swimming. Eur J Appl Physiol 92: 360-2

Turgut G, Turgut S, Genc O, Atalay A, Atalay EO (2004) The angiotensin converting enzyme I/D polymorphism in Turkish athletes and sedentary controls. Acta Medica (Hradec Kralove) 47: 133-6
Van Damme R, Wilson RS, Vanhooydonck B, Aerts P (2002) Performance constraints in decathletes. Nature 415: 755-6
Wall ME, Dunlop MJ, Hlavacek WS (2005) Multiple functions of a feed-forward-loop gene circuit. J Mol Biol 349: 501-14
Wang GL, Semenza GL (1996) Molecular basis of hypoxia-induced erythropoietin expression. Curr Opin Hematol 3: 156-62
Wang M, Marin A (2005) Characterization and prediction of alternative splice sites. Gene
Wang M, Marin A (2006) Characterization and prediction of alternative splice sites. Gene 366: 219-27
Wang P, Zou Y, Fu C, Zhou X, Hui R (2005) MYBPC3 polymorphism is a modifier for expression of cardiac hypertrophy in patients with hypertrophic cardiomyopathy. Biochem Biophys Res Commun 329: 796-9
Weibel ER, Bacigalupe LD, Schmitt B, Hoppeler H (2004) Allometric scaling of maximal metabolic rate in mammals: muscle aerobic capacity as determinant factor. Respir Physiol Neurobiol 140: 115-32
Weisberg A, Winegrad S (1996) Alteration of myosin cross bridges by phosphorylation of myosin-binding protein C in cardiac muscle. Proc Natl Acad Sci U S A 93: 8999-9003
Weisberg A, Winegrad S (1998) Relation between crossbridge structure and actomyosin ATPase activity in rat heart. Circ Res 83: 60-72
Welsman JR, Armstrong N, Nevill AM, Winter EM, Kirby BJ (1996) Scaling peak VO2 for differences in body size. Med Sci Sports Exerc 28: 259-65
West JB, Mathieu-Costello O (1995) Stress failure of pulmonary capillaries as a limiting factor for maximal exercise. Eur J Appl Physiol Occup Physiol 70: 99-108
Whyte G, Lumley S, George K, Gates P, Sharma S, Prasad K, McKenna WJ (2000) Physiological profile and predictors of cycling performance in ultra-endurance triathletes. J Sports Med Phys Fitness 40: 103-9
Whyte GP, George K, Sharma S, Firoozi S, Stephens N, Senior R, McKenna WJ (2004) The upper limit of physiological cardiac hypertrophy in elite male and female athletes: the British experience. Eur J Appl Physiol 92: 592-7
Wickens M, Anderson P, Jackson RJ (1997) Life and death in the cytoplasm: messages from the 3' end. Curr Opin Genet Dev 7: 220-32
Wilson RS, James RS (2004) Constraints on muscular performance: trade-offs between power output and fatigue resistance. Proc Biol Sci 271 Suppl 4: S222-5
Winegrad S (2003) Myosin-binding protein C (MyBP-C) in cardiac muscle and contractility. Adv Exp Med Biol 538: 31-40; discussion 40-1
Winegrad S (2005) Cardiac myosin binding protein C: modulator of contractility. Adv Exp Med Biol 565: 269-81; discussion 281-2, 405-15
Winter P, Hickey G, Fletcher H (2002) Instant Notes: Genetics, 2nd Ed. edn. BIOS Scientific Publishers, Ltd, Oxford, UK
Wolfarth B, Bray MS, Hagberg JM, Perusse L, Rauramaa R, Rivera MA, Roth SM, Rankinen T, Bouchard C (2005) The human gene map for performance and health-related fitness phenotypes: the 2004 update. Med Sci Sports Exerc 37: 881-903

Woods D, Hickman M, Jamshidi Y, Brull D, Vassiliou V, Jones A, Humphries S, Montgomery H (2001) Elite swimmers and the D allele of the ACE I/D polymorphism. Hum Genet 108: 230-2

Woods DR, Brull D, Montgomery HE (2000) Endurance and the ACE I/D polymorphism. Sci Prog 83: 317-36

Xiao W, Oefner PJ (2001) Denaturing high-performance liquid chromatography: A review. Hum Mutat 17: 439-74

Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60: 1513-31

Yamamoto F, Clausen H, White T, Marken J, Hakomori S (1990) Molecular genetic basis of the histo-blood group ABO system. Nature 345: 229-33

Yang N, MacArthur DG, Gulbin JP, Hahn AG, Beggs AH, Easteal S, North K (2003) ACTN3 genotype is associated with human elite athletic performance. Am J Hum Genet 73: 627-31

Yang Q, Osinska H, Klevitsky R, Robbins J (2001) Phenotypic deficits in mice expressing a myosin binding protein C lacking the titin and myosin binding domains. J Mol Cell Cardiol 33: 1649-58

Yoshiga CC, Higuchi M (2003) Rowing performance of female and male rowers. Scand J Med Sci Sports 13: 317-21

Yu B, French JA, Carrier L, Jeremy RW, McTaggart DR, Nicholson MR, Hambly B, Semsarian C, Richmond DR, Schwartz K, Trent RJ (1998) Molecular pathology of familial hypertrophic cardiomyopathy caused by mutations in the cardiac myosin binding protein C gene. J Med Genet 35: 205-10

Zak R (1977) Metabolism of myofibrillar proteins in the normal and hypertrophic heart. Basic Res Cardiol 72: 235-40

Zhang B, Tanaka H, Shono N, Miura S, Kiyonaga A, Shindo M, Saku K (2003) The I allele of the angiotensin-converting enzyme gene is associated with an increased percentage of slow-twitch type I fibers in human skeletal muscle. Clin Genet 63: 139-44

Zhu X, Bouzekri N, Southam L, Cooper RS, Adeyemo A, McKenzie CA, Luke A, Chen G, Elston RC, Ward R (2001) Linkage and association analysis of angiotensin I-converting enzyme (ACE)-gene polymorphisms with ACE concentration and blood pressure. Am J Hum Genet 68: 1139-48

Appendix 1

Nomenclature for describing variants

Amino acid substitutions

Use the one-letter codes: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine; X means a stop codon. 3-letter codes are also acceptable.

R117H or **Arg117His** - replace arginine 117 by histidine (the initiator methionine is codon 1).

G542X or **Gly542Stop** - glycine 542 replaced by a stop codon.

Nucleotide substitution

The A of the initiator ATG codon is +1; the immediately preceding base is -1. There is no zero. Give the nucleotide number followed by the change. For changes within introns, when only the cDNA sequence is known in full, specify the intron number by IVSn or the number of the nearest exon position.

1162G→A - replace guanine at position 1162 by adenine.

621+1G→T or **IVS4+1G→T** - replace G by T at the first base of intron 4; exon 4 ends at nt 621.

Deletions and insertions

Use Δ for deletions and ins for insertions. As above, for DNA changes the nucleotide position or interval comes first, for amino acid changes the amino acid symbol comes first.

ΔF508 - delete phenylalanine 508

6232-6236Δ or **6232-6236ΔATAAG** - delete 5 nucleotides (which can be specified) starting with nt 6232.

409-410insC - insert C between nt 409 and 410.