UNIVERSITY OF SOUTHAMPTON

VARIATION IN THE GENE FOR INSULIN-LIKE GROWTH FACTOR II AND ITS RELATIONSHIP WITH ANTHROPOMETRIC TRAITS

Thomas Richard Gaunt, B.Sc

A thesis presented for the degree of

DOCTOR OF PHILOSOPHY

in the

HUMAN GENETICS DIVISION
FACULTY OF MEDICINE, HEALTH AND BIOLOGICAL SCIENCES

July 2002

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE, HEALTH AND BIOLOGICAL SCIENCES
HUMAN GENETICS DIVISION

Doctor of Philosophy

VARIATION IN THE GENE FOR INSULIN-LIKE GROWTH FACTOR II AND ITS RELATIONSHIP WITH ANTHROPOMETRIC TRAITS

by Thomas Richard Gaunt

Obesity is a complex disorder caused by a combination of genetic and environmental factors. One candidate is the 30kb IGF2 gene coding for insulin-like growth factor II (IGF-II) on chromosome 11p15.5. Previous work identified an association between the IGF2 gene 3' untranslated region (3'-UTR) ApaI polymorphism and body mass index (BMI) in over 2500 middle-aged Caucasoid males from the Northwick Park Heart Study (NPHSII). A further single nucleotide polymorphism (SNP) in the P1 promoter of IGF2 was also found to be significantly associated with BMI. This study identified a further eleven novel polymorphisms and confirmed four published polymorphisms in the IGF2 gene by use of single-strand conformation polymorphism analysis (SSCP) and denaturing high performance liquid chromatography (DHPLC). Nine of the eleven novel polymorphisms were single nucleotide polymorphisms (SNPs) and two were homopolymeric tract length polymorphisms. Eight SNPs were genotyped in 2743 samples from the same cohort as the ApaI and P1 promoter polymorphisms and, combined with these data, yield four SNPs significantly associated with BMI in middle-aged men. Regression analysis indicates that three of these associations are significantly independently associated with BMI.

Haplotype analysis of NPHSII results identified significant differences in haplotype frequencies between BMI quartiles. The data indicated an association between a haplotype containing 'light' alleles for each of the four significantly associated SNPs and BMI in middle-aged men, supporting the individual association data.

Genotyping these SNPs in a second cohort (containing 626 men and 428 women) did not identify significant associations in men, although BMI trends for two of the SNPs were similar to those observed in the NPHSII cohort. Several associations were found in women suggesting a role for IGF2 in weight determination. No evidence was found to support the hypothesis that IGF2 influences foetal and early life development as well as adult weight.

The high-throughput genotyping in this project required more than 75,000 PCR reactions and electrophoresis tracks. This necessitated the development of novel high-throughput methodology to achieve results within a restricted time-scale and budget. The use of robotically aliquoted long PCR template, 384-well microplate array diagonal gel electrophoresis and gel image analysis software generated data rapidly, efficiently and economically, with minimum impact on DNA bank resources.

A quantitative-competitive RT-PCR assay was developed to enable an investigation of the functional role of polymorphisms in the IGF2 gene when appropriate samples become available. This will allow a comparison of gene expression levels in individuals of different genotypes.

The discovery of significant associations between IGF2 SNPs and BMI will be important to future investigations into the role of this gene in weight determination.

Table of contents

Table of figures

Table of tables

Preface

THESIS DECLARATON: This thesis is the result of work done while I was in registered postgraduate candidature. The project is based on work done jointly with others, but a substantial part is my original work, with work carried out by others outlined below.

This part-time PhD project was carried out within the Human Genetics Division at the University of Southampton under a British Heart Foundation Grant (PG/98158) and latterly an MRC programme grant. The project as a whole benefited from the contributions of several people, and this thesis therefore contains elements of their work. Here I will note the aspects of this project that involved other people for the purposes of defining my own role (a separate 'acknowledgements' section expresses my personal gratitude to those who contributed).

Microplate array diagonal gel electrophoresis (MADGE) and high-density variants were invented by Professor Ian Day (Day, US patent 6,071,396 and Day, GB patent GB2284484). The first use for high-throughput genotyping was during this project, and consequently I carried out the trouble-shooting and development work in conjunction with our laboratory manager, Lesley Hinks.

MADGE image analysis (manual and automatic) and data handling were developed for this project by me using software provided by Phoretix International.

Methods for long PCR and DOP-PCR were developed by Lesley Hinks. They were first used for high-throughput genotyping in this project, and aspects of their use (including robotic aliquoting etc.) were developed by me in conjunction with Lesley. Three-primer ARMS assays were already in use by others, but were integrated into the high-throughput MADGE system for the first time during this project. Long PCR's were optimised by me, then performed on bank DNA by Sylvia Diaper and Kristy Newman. Kelly Wilkinson carried out DOP-PCR on bank DNA.

I developed the quantitative-competitive RT-PCR assay for IGF2.

Sandra O'Dell designed SSCP primers prior to my joining the project. I optimised SSCP PCR reactions, while Sylvia Diaper carried out the SSCP PCRs and gels.

I designed DHPLC assays, optimised and performed PCR and ran the DHPLC assays.

Sequencing assays were designed, optimised and performed by me, and I carried out the sequence and SNP map construction.

Genotyping assays were designed and optimised by me. Sylvia Diaper and Kristy Newman carried out the PCR, electrophoresis and calling of DNA banks. I carried out data analysis, verification and handling. Prior to starting the project three SNPs has been genotyped in NPHSII and one in Hertfordshire (Sandra O'Dell and Sylvia Diaper). These have been included in the analyses, and are indicated in the text.

I performed Hardy-Weinberg, linkage disequilibrium and haplotype analyses on the data, and some preliminary one-way ANOVAs on NPHSII were carried out by Sandra O'Dell and myself. Further phenotype analysis was not possible due to retention of phenotype data by project directors, so NPHSII data analysis was performed by Jacqueline Cooper at the University of London, and Hertfordshire data analysis was carried out by Holly Syddall and Faiza Tabassum at the MRC Environmental Epidemiology Unit in Southampton. Haplotype analysis involved discussion of methodology with Santiago Rodriguez.

Acknowledgments

Many thanks to my supervisors Dr. Sandra O'Dell and Professor Ian Day who have supported and encouraged me in my work on this project, and provided much advice and help. My wife Lindsey has also provided endless support and advice throughout my PhD, especially during the difficult bits, and for this I am very grateful.

I thank Sylvia Diaper and Kristy Newman for their expert technical assistance with the high-throughput parts of this project. Sylvia was also involved in the initial SSCP scan of *IGF2*. I am very grateful to Lesley Hinks for her excellent help, support, friendship and advice throughout this project, and for her role in the development of high throughput genotyping methods. Tricia Briggs provided much technical support, and prepared high-quality equalised DNA samples from the NPHSII and Hertfordshire DNA banks. Kelly Wilkinson prepared DOP-PCR template that was used for genotyping three of the *IGF2* SNPs in Hertfordshire.

My thanks also go to Jacqueline Cooper (MRC Epidemiology and Medical Care Unit), Holly Syddall and Faiza Tabassum (MRC Environmental Epidemiology) for prompt analysis of lots of data at short notice, and to all at the MRC who contributed. This project would not have been possible without all the people who participated in the Northwick Park Heart Study and the Hertfordshire Study. The work was supported by grants from the British Heart Foundation (PG/98158) and the Medical Research Council.

I thank Diane Brown and Carolyn Wallis who have provided secretarial support during my project. I also thank Drs. Ros Ganderton, John Holloway, Manolis Spanakis, Shaoli Zhang, Santiago Rodriguez and Derek Mann for their contributions.

My thanks go to all the people in the division, particularly in Lab 18, who have helped to make this project an enjoyable and rewarding experience.

Finally, I would like to thank my family for all their support and encouragement during this and previous studies.

Abbreviations

Term	Definition
α-MSH	α-melanocortin stimulating hormone
A	Adenine
ANOVA	Analysis of Variance
ARMS	Amplification Refractory Mutation System
BMI	Body Mass Index
bp	Base pairs (of nucleic acid)
C	Cytosine
CGAP	Cancer Genome Anatomy Project
CHD	Coronary Heart Disease
CNTF	Ciliary Neurotrophic Factor
CpG	Cytosine-phosphate-Guanine dinucleotide
CV	Coefficient of Variation
dbSNP	SNP database - NCBI
DEPC	Di-ethyl Pyrocarbonate
DEXA	Dual-Energy X-ray Absorptiometry
DHPLC	Denaturing High Performance Liquid Chromatography
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotide Tri-Phosphates
DOP-PCR	Degenerate Oligonucleotide Primer PCR
G	Guanine
GHRH	Growth Hormone Releasing Hormone
IDDM	Insulin Dependent Diabetes Mellitus (type I diabetes)
IGF1	The gene for Insulin-like Growth Factor I
IGF2	The gene for Insulin-like Growth Factor II
IGFBP	Insulin-like growth factor binding protein
IGF-I	Insulin-like Growth Factor I, the gene product of IGF1
IGF-II	Insulin-like Growth Factor II, the gene product of IGF2
INS	The gene for Insulin
kb	Kilobases (of nucleic acid)
kDa	KiloDaltons
LD	Linkage Disequilibrium
MADGE	Microplate Array Diagonal Gel Electrophoresis
Mb	Megabases (of nucleic acid)
MC4R/MC5R	Melanocortin Receptors
MODY	Maturity Onset Diabetes of the Young
mRNA	Messenger Ribonucleic Acid
NCBI	National Centre for Biotechnology Information
NIDDM	Non-Insulin Dependent Diabetes Mellitus (type II diabetes)
NPHSII	Northwick Park Heart Study II
NPY	Neuropeptide Y
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
POP	Performance Optimised Polymer
RNA	Ribonucleic Acid
rNTPs	Ribonucleotide Tri-Phosphates
RT-PCR	Reverse transcriptase-PCR

Term	Definition
SNP	Single Nucleotide Polymorphism
SSCP	Single Strand Conformation Polymorphism
T	Thymine
TH	The gene for Tyrosine Hydroxylase
U	Uracil
UCP	Uncoupling Protein
UTR	Untranslated Region - 5'UTR is transcribed sequence before start of translated sequence, 3'UTR is transcribed sequence after end of translated sequence
VNTR	Variable Number of Tandem Repeat polymorphism
WHR	Waist to Hip Ratio

Publications arising from this project

Gaunt TR, Cooper JA, Miller GJ, Day IN, O'Dell SD.

Positive associations between single nucleotide polymorphisms in the IGF2 gene region and body mass index in adult males.

Human Molecular Genetics, 2001 Jul 1;10(14):1491-501.

Xiao-He Chen, Sandra D. O'Dell, Lesley J. Hinks, Emmanuel Spanakis, **Tom R. Gaunt**, Rosalind H. Ganderton and Ian N.M. Day

High-resolution MADGE

Technical Tips Online, 2001, 1:116:T02156

Tom R. Gaunt, Lesley J. Hinks, Xiao-he Chen, Sandra D. O'Dell, Emmanuel Spanakis, Rosalind H. Ganderton and Ian N.M. Day

384-well MADGE for high-throughput DNA-bank studies

Technical Tips Online, 2000, 1:108:P02069

Rosalind H. Ganderton, Sandra D. O'Dell, **Tom R. Gaunt**, Xiao-he Chen, Lesley J. Hinks, Emmanuel Spanakis and Ian N.M. Day

Microplate-array diagonal-gel electrophoresis (MADGE) systems for high-throughput electrophoresis

Technical Tips Online, 2000, 1:108:P02068

O'Dell SD, **Gaunt TR**, Day IN.

SNP genotyping by combination of 192-well MADGE, ARMS and computerized gel image analysis.

Biotechniques, 2000 Sep;29(3):500-4, 505-6.

Chapter 1 : Introduction

1.1 Complex traits

While monogenic (Mendelian) inheritance can explain the manner in which certain characteristics are transmitted from one generation to the next, the majority of phenotypes are inherited in a more complex manner. Polygenic inheritance describes a situation in which inheritance of a condition or phenotype is the result of the interactions and contributions of several genes. The term "complex trait" refers to a phenotype (or trait) which does not follow Mendelian inheritance, but is influenced by multiple genes and environmental factors (Lander & Schork 1994). A complex trait can be defined as a trait for which there is "no single locus that contains alleles that are necessary or sufficient for disease" (Pritchard 2001).

Major genes have a large effect on phenotype. Oligogenes contribute less but are more frequent and polygenes have a very minor role, and are very frequent (Morton 1998). The genes underlying a complex trait may therefore vary in their contribution to that trait, with the number of genes increasing as the effect per gene decreases.

While it is clear that complex traits are the result of the effects of multiple genes, the frequency of contributing variants and the effect of allelic heterogeneity is less clear. Are complex diseases caused by common variants or rare variants (Pritchard 2001)? The success of a disease-mapping strategy may be very dependent on the answer to this question. The number of disease-causing alleles varies significantly between loci (Reich & Lander 2001), and is not as straightforward as suggested by the "common disease, common variants

hypothesis" (Lander 1996). The greater the allelic heterogeneity at a locus, the more difficult it is to identify and the harder it is to clinically test (Reich & Lander 2001).

Complex phenotypes are distinct from complex traits. A simple phenotype is a dichotomy in which disease is scored as 'normal' or 'affected'. A complex phenotype is either a polychotomy or a continuum of expression and/or severity (irrespective of complexity of inheritance) (Morton *et al.* 1991).

1.2 Complex disease analysis

The term "complex disease" refers to complex traits that affect health or survival. There are several different approaches to the analysis of complex diseases: linkage analysis, allele-sharing analysis, association analysis and animal models (Lander & Schork 1994). Each of these methods has different advantages and different requirements in terms of experimental materials and methods. The first two are pedigree-based analyses involving comparison of the inheritance of chromosomal regions with inheritance of disease. These methods are limited to a resolution of about 1 megabase of DNA (due to insufficient recombination within a pedigree) (Devlin & Risch 1995). Association analysis allows higher-resolution mapping by searching for marker alleles that are in linkage disequilibrium (LD) with a disease-causing allele (assessed by the association between the marker genotype and disease phenotype). Association analysis involves population rather than family studies (Lander & Schork 1994).

1.2.1 Linkage analysis

Linkage analysis tests for a genetic relationship between loci to enable mapping of a disease-causing locus. The extent to which recombination (Figure 1-1) separates two loci increases as genetic distance increases. If two loci are far apart (suggesting that they are completely unlinked) then the recombination fraction between them will be 0.5 (i.e. 50% of meiotic events will be recombinant by chance). If the loci are linked the recombination fraction will be θ (where $\theta<0.5$) (Strachan & Read 1999).

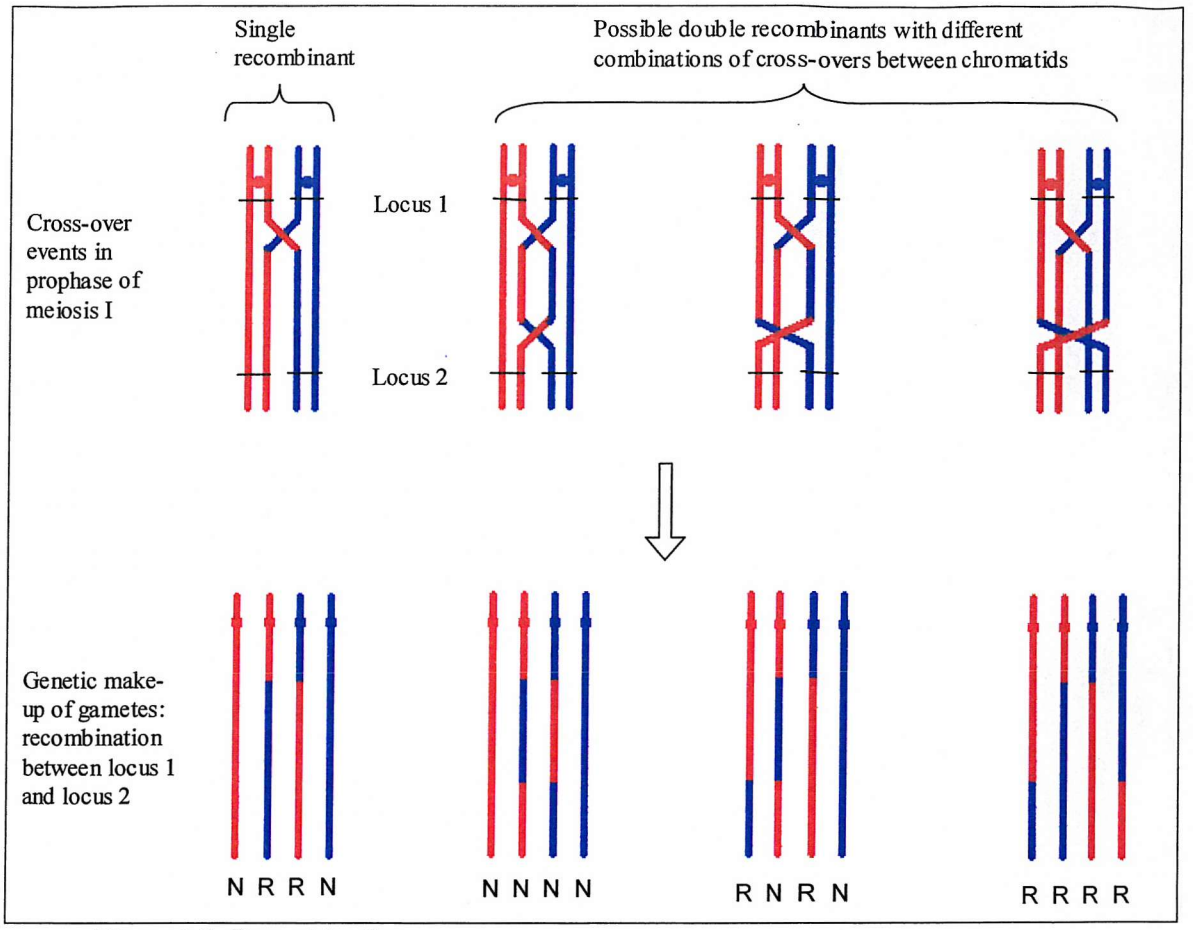


Figure 1-1: Recombination
Adapted from Strachan & Read (1999). N=non-recombinant, R=recombinant. N/(N+R)=0.5 (as double crossovers occur in random proportions) so average effect is 50% recombinants.

Parametric linkage analysis can use a direct calculation of the recombination fraction in pedigrees where parental phase is known (recombination fraction = R/(N+R), where R = recombinants and N = non-recombinants). However, in many cases parental phase is not known, so linkage analysis uses the ratio of (H_1) the likelihood of the pedigree occurring if loci linked (recombination fraction = θ) to (H_0) the likelihood of the pedigree occurring if loci are not linked (recombination fraction = 0.5). This ratio is the odds ratio of linkage (Strachan & Read 1999). A commonly used measure of linkage is the logarithm of odds (lod score), a function of θ, with the most likely recombination fraction being that which gives the maximum lod score (Morton 1955). The value at which a lod score becomes significant is affected by the overall improbability of two loci being linked (Strachan & Read 1999), although a value of ≥ 3 is considered necessary (although not sufficient) to indicate linkage (Morton 1998).

Multipoint linkage analysis utilises computer programs to analyse a framework of multiple markers – the disease locus is tested in each marker interval, and the likelihood of

this explaining the pedigree data calculated. This gives a map with the maximum peak indicating the most likely location for the disease locus. This method depends on accurate marker mapping as the size of marker intervals will affect the lod score (Strachan & Read 1999).

1.2.2 Allele-sharing methods

Parametric linkage analysis relies on the specification of a precise and valid model to produce useful results. This becomes a problem with diseases which do not follow Mendelian inheritance or do not have straightforward and reliable diagnostic criteria (Strachan & Read 1999). Non-parametric methods are more appropriate for many common diseases. Allele-sharing analysis is a non-parametric alternative of which the simplest form is the affected sib-pair analysis (Lander & Schork 1994). In affected sib-pair analysis chromosomal regions can be defined as identical by descent (IBD – copies of the same ancestral allele) or identical by state (IBS – the same allele from different ancestors). If there is no linkage between a chromosomal region and disease, then siblings would be expected to share IBD alleles in the following ratio IBD(0) 25%:IBD(1) 50%:IBD(2):25% (Strachan & Read 1999). If linkage exists then this ratio should distort to indicate the greater than random co-segregation of that region with disease (Figure 1-2) – this can be tested with a χ^2 test (Lander & Schork 1994). This method is more robust than linkage analysis, but less powerful than a correctly specified linkage model. It allows excess allele-sharing to be detected in the presence of genetic heterogeneity, phenocopy and incomplete penetrance (Lander & Schork 1994).

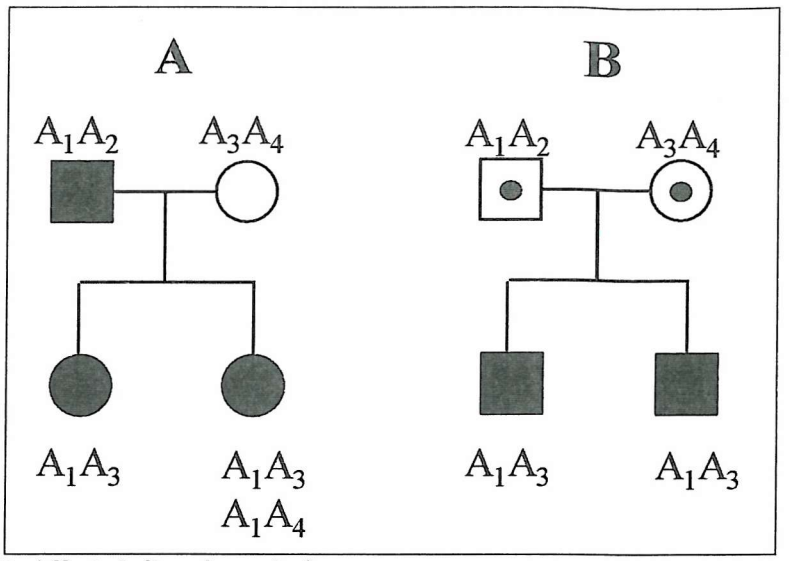


Figure 1-2: Affected sib-pair analysis
Adapted from Strachan & Read (1999). Square=male, circle=female. Closed shape indicates diseased state, open indicates non-diseased and open with dot indicates carrier of disease allele. A normal unlinked locus would have a 1:2:1 sharing ratio for 0, 1 or 2 alleles. A shows a dominant condition in which the ratio is 0:1:1 for 0, 1 or 2 alleles. B shows a recessive condition in which alleles are shared at a ratio of 0:0:1 for 0, 1 or 2 alleles.

Risch et al present an example of this approach in a study of autism (Risch *et al.* 1999): 97 independent affected sib-pairs (ASPs) and 51 discordant sib-pairs (DSPs) were used for linkage analysis with 362 markers. Sharing was 51.6% in ASPs and 50.8% in DSPs, which in this case was likely to involve multiple loci (≥15) and did not identify any particular chromosomal region with significant linkage. While in many cases both parents were available, the genotype of missing parents could often be reconstructed using the genotypes of children. With multiple affected sibs, independent pairs were selected as pairs comprising the first sib and one other (pairings not involving the first sib would not be independent of those including the first sib). Independent pairs within one family numbered n-1 where n was the number of affected sibs. Initial analysis tested the significance of deviation from the expected 50% identical by descent allele-sharing using a χ^2 test. A multipoint sib-pair analysis was then performed to obtain the maximum information.

1.2.3 Association studies

Association studies need not use family data at all (although the transmission disequilibrium test does (Schaid 1998)). Association refers to the greater than random co-occurrence of an allele (or phenotype) at locus 1 with an allele (or phenotype) at locus 2

(Strachan & Read 1999). This is distinct from linkage, which examines the co-inheritance of loci (not alleles).

1.2.3.1 Linkage disequilibrium

Linkage disequilibrium (LD) is the measure of association, and is defined as the magnitude of the difference between the product of the allele frequencies and the haplotype frequency (Terwilliger & Weiss 1998). For allele *X* at locus 1, allele *Y* at locus 2 and linkage disequilibrium δ:

$$P_{XY} = P_X \times P_Y + \delta \qquad \text{(Terwilliger \& Weiss 1998).}$$

Several measures of LD are used, based on estimation of haplotype frequencies and observation of allele and genotype frequencies – these tend to give results with similar trends but differing magnitudes (Devlin & Risch 1995).

Recombination tends to cause LD to decrease, thus both time and physical distance influence the magnitude of LD. For genetic mapping the distance over which useful linkage disequilibrium extends varies throughout the genome, but estimates vary from 3kb (Kruglyak 1999) to 100kb (Collins *et al.* 1999). Overestimating the extent of linkage disequilibrium could result in disease loci not being detected in a mapping project, while underestimating could vastly increase the cost of a mapping project. The application of a Bonferroni correction for multiple testing can result in the loss of significance of association if many markers are tested, and even without correction, some single-nucleotide polymorphisms (SNPs) <10kb from the Apo-E polymorphism (positively associated with Alzheimer's Disease) are not themselves significantly associated (Martin *et al.* 2000). It seems likely that accurate mapping of disease loci in complex genetic disorders requires a high density of markers, particularly if SNPs are used as these markers are generally less informative than microsatellites (Martin *et al.* 2000).

1.2.3.2 Population association analysis

Association analysis is the mapping of disease alleles by the association of markers with disease. A marker allele is associated with disease if it occurs at a significantly higher frequency in cases than controls (Terwilliger & Weiss 1998), or is associated with a difference in mean value for a particular phenotype (quantitative traits). A positive

association can arise if: (a) the marker is disease-causing, (b) the marker is in linkage disequilibrium with the disease-causing allele or (c) population admixture has caused an artefact (Lander & Schork 1994). Note that in complex diseases a 'disease-causing allele' may be an allele that only contributes to a proportion of the disease phenotype, or has a susceptibility effect. Alternative alleles at the same locus that associate with a 'non-disease' state may be said to be 'protective'. Population stratification (admixture) occurs when a population has distinct subsets that may co-incidentally have both a higher frequency of a phenotype and a higher frequency of a particular allele – an extreme example of this is the association between *HLA-A1* and the ability to eat with chopsticks in the San Francisco population. This is explained by the fact that *HLA-A1* is more common among Asians than Caucasians, rather than any real genetic effect (Lander & Schork 1994). Stratification may be avoided by the use of homogeneous populations, and ideally by verification with family-based association studies (Lander & Schork 1994).

1.2.3.3 Family-based association analysis

The transmission-disequilibrium test (TDT) (Spielman *et al.* 1993) is commonly used for association analysis of family-based studies. Variants exist which enable analysis with only one parent (1-TDT (Sun *et al.* 1999)) and with no parents (sib-TDT or S-TDT (Spielman & Ewens 1998)). The TDT statistic is calculated as:

$$TDT = (x - y)^2/(x+y)$$

where x is the number of times allele X is transmitted from a heterozygous parent to an affected child, and y is the number of times allele Y is transmitted from a heterozygous parent to an affected child (Schaid 1998). The TDT test is "a test for linkage in the presence of linkage disequilibrium" (Schaid 1998). The basis of the test is that if no linkage disequilibrium exists then the frequency of transmission of an allele from a heterozygous parent to an affected child should be 50% ($TDT = (0.5n\text{-}0.5n)^2/n = 0$). A deviation from this 50% transmission frequency indicates both linkage (between marker locus and disease locus) and association (between the more frequent allele and the disease-causing allele).

1.2.4 Animal studies

Animal studies allow many of the problems of genetic heterogeneity in human populations to be removed. Artificially selected populations can be created, and phenotypes measured very accurately. However, the results of animal experiments are often not mirrored in humans, and their usefulness may be limited (Lander & Schork 1994).

1.3 Genetic variation

1.3.1 Types of marker

Table 1-1 shows the types of human genetic markers that have been developed for the mapping and diagnosis of human disease presented in chronological order of development. Those involving direct genotyping of DNA include restriction fragment length polymorphisms (RFLPs), minisatellites, microsatellites and single nucleotide polymorphisms (SNPs) (Strachan & Read 1999). RFLPs and SNPs are less informative (maximum heterozygosity = $2pq = 0.5$ for a SNP with allele frequencies $p = 0.5$ $q = 0.5$) than microsatellites and minisatellites, but are cheaper and easier to genotype and more abundant in the genome.

Table 1-1: Types of human genetic marker
Adapted from Strachan & Read (1999)

Marker type	Number of loci	Features of marker
Blood Groups	Tens	Difficult to localise Problems with dominance affect genotyping
Serum proteins	Tens	Variants identified by electrophoresis Complex assays, difficult to localise
HLA tissue types	1 (haplotype)	Very informative, but very restricted distribution Only tests for linkage to HLA region
RFLPs	Hundreds of thousands	Diallelic markers. Easy to genotype Easy to localise. Many are SNPs
Minisatellites	Tens of thousands	Multi-allelic, so very informative Not always suited to PCR – Southern blot often used Easy to localise, but not even distribution
Microsatellites	Hundreds of thousands	Multi-allelic, very informative Suited to PCR and multiplexing Easy to localise, well distributed
SNPs	Millions	Diallelic, so less informative Suitable for automation and multiplexing Easy to localise, very well distributed in genome

1.3.2 Single-nucleotide polymorphisms

1.3.2.1 Nature of SNPs

Single nucleotide polymorphisms (SNPs) account for 90% of sequence variation in humans (Collins *et al.* 1998). SNPs can be defined as single-base pair loci in genomic DNA at which alternative nucleotides exist in some chromosomes (with a frequency between 1% and 99%) in one or more populations (Brookes 1999). The frequency of SNPs (the rate at which they occur per unit sequence length) varies between estimates, depending on the chromosomal region examined, the type of sequence (intronic, exonic, promoter, coding etc) and the number of samples examined. A frequency of 1 SNP per 1000bp when two chromosomes are compared is generally accepted (Brookes 1999). A study testing for SNPs in 2.3 megabases of human genomic DNA sequence tagged sites (STS) confirms this with SNPs occurring at frequencies of 1/657 to 1/1159 (depending on sequence type and detection method) (Wang *et al.* 1998). In another study non-coding SNPs were found to occur at a rate of 1 per 354bp, while coding SNPs (cSNPs) were found at 1 per 346bp with synonymous cSNPs occurring at 1 per 656bp and non-synonymous cSNPs occurring at 1 per 734bp (560

SNPs in 106 genes) (Cargill *et al.* 1999). The SNP Consortium and the Human Genome Project have together identified 1.42 million unique SNPs, an average of 1 per 1.9kb (Sachidanandam *et al.* 2001), confirming that SNPs are at least as frequent as 1 per 1900bp.

SNPs can be subdivided into two categories: (1) transition SNPs which involve a purine-pyrimidine base-pair changing to a purine-pyrimidine base-pair and (2) transversion SNPs which involve a purine-pyrimidine base-pair changing to a pyrimidine-purine base-pair (or vice versa) (Figure 1-3) (Brookes 1999).

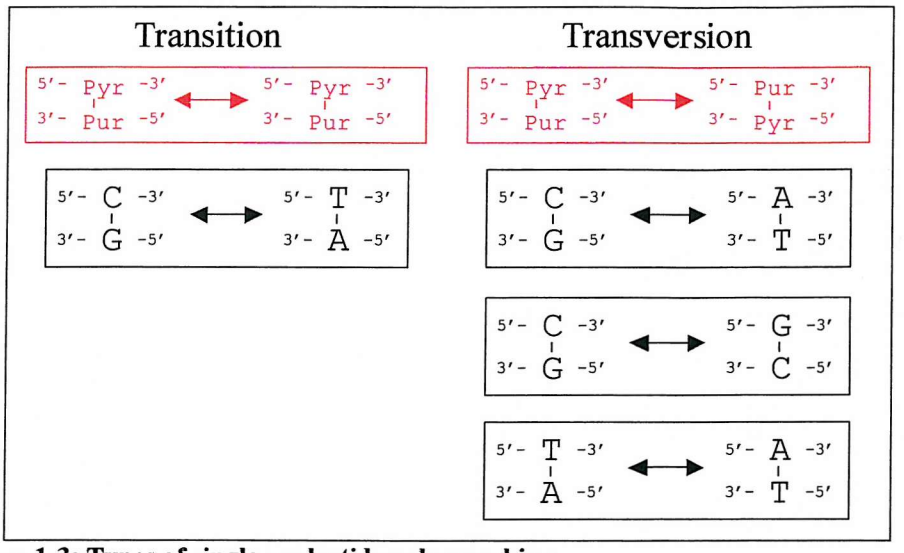


Figure 1-3: Types of single nucleotide polymorphism

About 2/3 of SNPs involve the transition C(G)↔T(A) (Wang *et al.* 1998) while the other three types occur at similar levels to each other; the high frequency of C↔T transitions may be due in part to the common deamination of methylated cytosines at CpG dinucleotides (Brookes 1999). About 24% of SNPs occur within a CpG dinucleotide (Wang *et al.* 1998).

Most SNPs are diallelic. This means that at a particular SNP locus two alternative nucleotides exist at a frequency of between 1% and 99% in at least one population when multiple chromosomes are tested. However, theoretically triallelic SNPs (SNP loci with three alternative nucleotides) and even tetraallelic SNPs (SNP loci with four alternative nucleotides) could occur through subsequent mutation events affecting one of the alleles of an existing SNP (Brookes 1999). It seems reasonable to assume that if a diallelic SNP occurs a conservative once per 1000bp, then triallelic SNPs arising independently would occur at a frequency of 10^{-3}x10^{-3} =10^{-6}, or once per million bases. Similarly tetra-allelic SNPs might occur at a rate of 10^{-9}, or about 3 times in the human genome of 3x10^{9} bp. An example of a triallelic SNP is a T/A/G SNP in the coding region of the Fcγ receptor type IIIA gene

(*FcγRIIIA*) with allele frequencies of 86% (T^{230}), 8% (A^{230}) and 6% (G^{230}) (de Haas *et al.* 1996). Although the authors do not discuss it, it is likely that both the A and G alleles arose independently from the T allele (rather than A arising from G or vice versa) – this would account for the relatively high allele frequencies of both. While this type of SNP is rare, it is possible for genotyping experiments to miss triallelic SNPs (Figure 1-4).

Triallelic SNP allele frequencies:	p=0.75, q=0.24 and r=0.01
Under Hardy-Weinberg equilibrium:	$p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1$

Frequencies and number per 1000 samples

	p^2	q^2	r^2	2pq	2pr	2qr
Freq	0.5625	0.0576	0.0001	0.3600	0.0150	0.0048
Number	562.5	57.6	0.1	360	15	4.8

If ARMS assay or RFLP assumes biallelic, r will not be detected (dropout) and 2pr and 2qr heterozygotes will add to p^2 and q^2 homozygote numbers respectively.

$p^2 + q^2 + 2pq = 1$ where $p = 0.76$ and $q = 0.24$

Frequencies and number per 1000 samples

	p^2	q^2	2pq
Observed number	577.5	360	62.4
Expected number (HW equilibrium)	573.9	367.3	58.8

$\chi^2 = 0.4$ (not significant at 1 degree of freedom)

Figure 1-4: Misinterpretation of tri-allelic SNPs
Hypothetical figures for a typical population study of 1000 samples are shown. If the least common of the three alleles is 1% (the minimum allele frequency to be considered a SNP) then this allele can be 'missed' by a typical genotyping assay without affecting the population statistics significantly.

1.3.2.2 Mapping of disease loci with SNPs

Over the last few years much has been published on the potential use of SNPs for high-resolution mapping of human disease loci ((Collins *et al.* 1999; Brookes 1999; Schork *et al.* 2000; Martin *et al.* 2000; Wang *et al.* 1998)). These typically diallelic markers are less potentially informative than multiallelic markers (Table 1-2), and therefore more are required to map a disease. However, as they are the most common class of sequence variation (Collins

et al. 1998) and more straightforward to genotype, they are likely to be of much use in the mapping of disease loci.

Table 1-2: Maximum heterozygosity for different numbers of alleles
(actual heterozygosity = $1-(p_1^2+p_2^2+\ldots+p_n^2)$ where p_n is frequency of allele n.
Actual heterozygosity is lower with low/high frequency alleles).

Alleles	1	2	3	4	5	6	7	8	9	10	n
Maximum heterozygosity	0.00	0.50	0.67	0.75	0.80	0.83	0.86	0.88	0.89	0.90	$\frac{n-1}{n}$

Most SNPs must be effectively neutral (to account for their frequency), but some contribute to disease susceptibility and resistance (Collins *et al.* 1999). The proximity hypothesis suggests that most disease-causing variants will not be included in a panel of markers, and therefore localisation of these variants will depend on linkage disequilibrium (LD) between marker alleles and disease causing alleles (Kruglyak 1999; Collins *et al.* 1999). Detection of alleles of small effect would require a very high density of markers under the proximity hypothesis (Collins *et al.* 1999). The magnitude and range of LD is of critical importance in the mapping of human disease loci. If LD extends over a long range, then mapping will be low resolution, but will require few markers. If LD only extends over a short range, then mapping will be much more precise, but a high density of markers will be required. The extent of LD varies significantly across the genome (Taillon-Miller *et al.* 2000), depending on the frequency of recombination within particular regions (recombination "hot-spots" and "cold-spots") and the age of haplotypes (Maniatis *et al.* 2002).

1.3.2.3 SNP nomenclature

Nomenclature of SNPs is a complex issue, and central to the dissemination of information on variation in the human genome. The 'Nomenclature Working Group' (Antonarakis 1998) suggested a system for the description of variations in both DNA and protein sequences. The details of this system have been updated to include more types of variation as these have been suggested (den Dunnen & Antonarakis 2000; den Dunnen & Antonarakis 2001), and the system is still evolving, with updates at http://www.dmd.nl/mutnomen.html.

The recommended nomenclature for SNPs is as follows (from den Dunnen & Antonarakis 2001):

1. Nucleotides are A (adenine), C (cytosine), G (guanine) and T (thymine) (upper case)

2. Nucleotides are numbered relative to the A of the ATG translation initiation codon, with this nucleotide designated +1, and the preceding nucleotide (5') –1 (no 0).
3. Non-coding regions: SNPs within introns are designated by the last nucleotide of the preceding exon, a plus sign and the position 3' of the exon end *or* the first nucleotide of the next exon, a minus sign, and the position 5' of the exon start. In the 3'UTR the first nucleotide after the translation stop codon is designated *1.
4. The nucleotide substitution is indicated by a ">" character.

An example would therefore ~~by:~~ be 85+7C>T indicates a cytosine to thymine transition at the 7th nucleotide in the intron between cDNA nucleotides 85 and 86 (relative to ATG).

This system (den Dunnen & Antonarakis 2001) is relatively robust, widely used, and covers the majority of possibilities, allowing unambiguous description of SNPs in human genes. However, it has certain limitations:

1. In alternatively spliced genes the position of a SNP relative to the ATG translation initiation codon will vary according to the exon combination in the transcript.
2. The system adopts a gene-orientated approach to SNP nomenclature – this ignores the possibility of naming SNPs in regions where gene structure is not defined.
3. The system is vulnerable to the possibility of sequencing errors or insertion/deletion polymorphisms moving the position of the SNP relative to the ATG codon in different versions of the sequence.
4. The use of a (preferably genomic) reference sequence has been suggested as a method to resolve some of these problems (http://www.dmd.nl/mutnomen.html). However, this website still recommends numbering relative to a translation initiation codon.

We have adopted the reference sequence approach in this project (Gaunt *et al.* 2001) to describe the position of SNPs relative to a defined GenBank accession stored in the 'Entrez Nucleotides' database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). However, we describe the SNP by nucleotide number within that sequence, plus a description of the nucleotide change (e.g. L15440-6815A/T describes an adenine to thymine transversion at nucleotide 6815 in GenBank sequence L15440). While this system does not have the benefit of indicating the position of a SNP within a gene (e.g. intronic, exonic or promoter), it is unambiguous and allows anyone to identify the SNP in the context of a unique sequence that can then be related back to its position in the context of a gene.

1.3.3 Deleterious genetic variation

A large proportion of the genome does not code for proteins, although that doesn't necessarily mean it is "junk DNA" (Wong *et al.* 2000). The effect of a polymorphism on survival of an individual will therefore depend on its location and nature. Non-synonymous mutations (those causing an amino acid change and/or a frame shift) in the coding region of a gene are the most likely to have a deleterious (or advantageous) effect because of their potential effect on protein structure and function. Synonymous mutations are less likely to have an effect. Mutations in gene promoters may affect the binding of transcription factors and therefore expression levels of the gene. These may therefore also be deleterious. Other regulatory elements such as enhancers and imprinting control elements may also influence expression levels or imprinting and therefore be deleterious if altered by a mutation. The least significant mutations are probably those within non-regulatory introns and intergenic regions.

Polymorphism rates give a basic estimation of the relative importance of different sites. Cargill *et al* characterised 560 SNPs and estimated polymorphic rates for different types (Cargill *et al.* 1999): non-coding and coding SNPs both occurred at a rate of around 3 per kilobase. Non-synonymous variants were observed at a rate of 38% of the rate of synonymous variants, non-conservative non-synonymous SNPs occurred at half the rate of conservative non-synonymous SNPs (Cargill *et al.* 1999). This is consistent with the hypothesis that SNPs affecting protein structure and function are less frequent than those that do not. This lower frequency is due to reduced survival of individuals with deleterious mutations (natural selection) rather than a lower mutation rate (which is affected by sequence context, for example CpG dinucleotides (Wang *et al.* 1998), but not functional significance of the site).

Eyre-Walker *et al* investigated the rate of deleterious mutations by comparing gene sequences for chimpanzees, gorillas and humans (Eyre-Walker & Keightley 1999). In their sample of 46 genes they found 143 predicted non-synonymous mutations (0.0034 per nucleotide) compared to an expected 231 (0.0056 per nucleotide) if non-synonymous mutations were neutral (non-deleterious and non-advantageous). Their conservative estimates of amino-acid altering mutation rate (M) and deleterious mutation rate (U) are M=4.2(±0.5) and U=1.6(±0.8) mutations per diploid genome per generation (Eyre-Walker & Keightley 1999).

Mutation type is probably also very important in the effect that a mutation has on organism survival. A single-nucleotide polymorphism affects only one nucleotide, while

microsatellites affect many nucleotides, insertion/deletion variants can affect many genes, while significant chromosomal abnormalities have the largest effect.

1.4 Population Obesity

1.4.1 Obesity

Obesity is a complex disorder comprising a group of physiological conditions influenced by both genetic and environmental factors (Jebb 1997). Adiposity is an important factor in health, and obesity may lead to significant health problems in affected individuals. About 15% of people in Britain are obese (Body Mass Index (BMI) >30kgm^{-2}) and this figure seems to be rising (Jeffcoate 1998). Obesity results from a prolonged period of energy intake exceeding energy expenditure (Jebb 1997). Industrialisation has led to decreased activity and increased fat intake in western societies, and may be influential in the increase in mean BMI in these populations (James 1996).

Obesity is a major risk factor for coronary heart disease (CHD) (Bertrais *et al.* 1999; Morricone *et al.* 1999) and non-insulin-dependent diabetes mellitus (NIDDM) (Björntorp 1996), with fat distribution being important in both. Central abdominal obesity is considered particularly important in the onset of NIDDM (Björntorp 1996). However, obesity itself may not be the cause of these diseases, but another symptom of mutual underlying risk factors, including hypertension, cholesterol intake and reduced activity (Shaper 1996).

General obesity in adults is a strong predictor of hypertension (Shaper 1996), while central abdominal obesity is a strong predictor of cardiovascular events (Bertrais *et al.* 1999). General obesity is associated with risk factors for coronary heart disease, although this relationship is complicated by smoking, which is a risk factor for CHD, but tends to result in lower body weight (Shaper 1996). Central abdominal obesity is a more common fat distribution pattern in men than women, and may account for the higher risk of cardiovascular events in men (Bertrais *et al.* 1999).

Measurements of obesity include Body Mass Index (BMI = weight·height^{-2}), Waist to Hip Ratio (WHR – ratio of circumference measurements), waist circumference, percentage fat and fat mass. BMI, WHR and waist circumference are the simplest to measure. Of these, BMI gives a good indication of general obesity, while WHR and waist circumference both indicate levels of central obesity (James 1996). WHR may be influenced by other factors,

such as hip bone size and muscle mass, and so is not a perfect measure of visceral fat (Björntorp 1996). BMI is usually used as a statistic for estimating levels of obesity in populations. Definitions vary, but it is generally accepted that normal BMI falls within the 20-25kgm^{-2} range, while Grade I (overweight) is 25-29kgm^{-2}, Grade II (obesity) is 30-39kgm^{-2} and Grade III (extreme obesity) is over 40kgm^{-2} (Jeffcoate 1998). However it is important to note that BMI is not height-independent at the same power of height in all populations; body shape varies in different populations with some being naturally taller than others (James 1996).

1.4.2 Genetics and mechanisms of obesity

There is little doubt that obesity has a strong genetic component. Obesity aggregates in families, although this could also be accounted for by environmental and cultural factors (Bouchard 1996). More significant evidence is obtained from twin studies in which genetic variation may account for up to 70% (Bouchard 1996) of obesity depending on gender and measure of weight used (Nelson *et al.* 1999), although estimates vary from as little as 30% to as much as 80% (Echwald 1999). Adoption studies tend to give lower estimates (around 30%) (Bouchard 1996), and family studies (which allow more complex segregational analyses) also suggest a genetic component of around 30% (Echwald 1999). Figure 1-5 shows how genes and environment interact to influence weight.

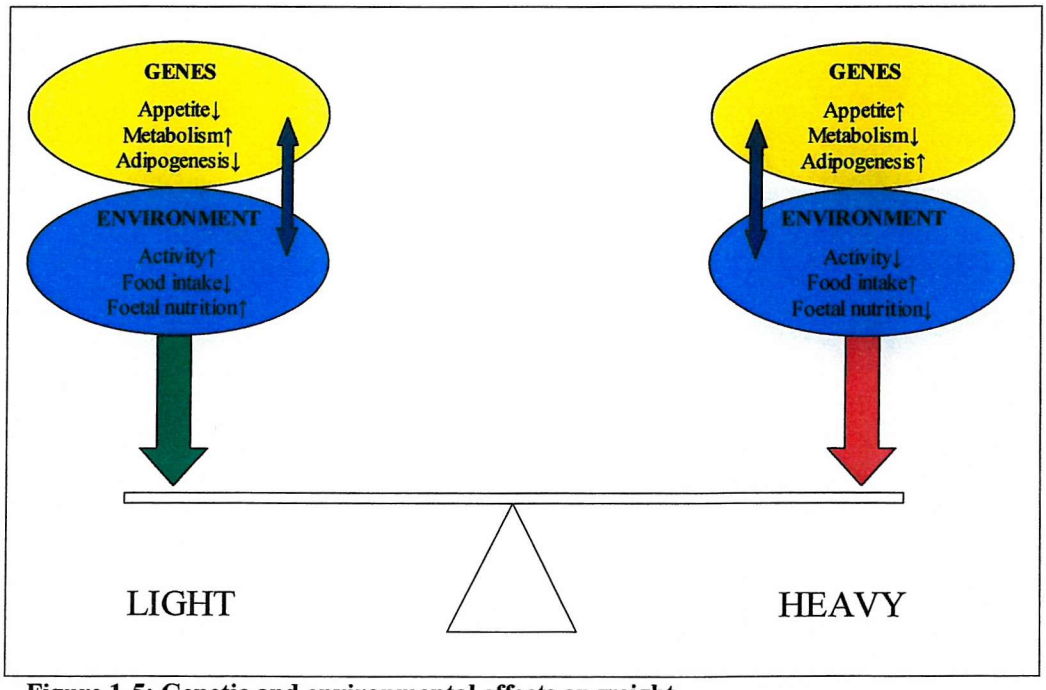


Figure 1-5: Genetic and environmental effects on weight

The genetic component of obesity has been studied extensively in mouse models. The discovery of a gene expressed in adipose tissue in mice, a deficiency of which results in morbid obesity and type II diabetes (Zhang *et al.* 1994), caused much interest. This 'obese' gene was found to suppress food intake and reduce body weight by inhibition of neuropeptide-Y levels (which stimulates food uptake and decreases thermogenesis) in the hypothalamus (Stephens *et al.* 1995). The 'obese' gene encodes leptin, a 16kDa adipocyte-secreted protein (Schalling *et al.* 1999). Leptin was found to suppress food intake and decrease body weight in normal mice and ob/ob (leptin deficient) mice, but not in db/db mice (which have no leptin receptor) (Stephens *et al.* 1995; Campfield *et al.* 1995). Administration of leptin also causes a decrease in hypothalamic galanin (GAL), melanin-concentrating hormone (MCH), pro-opiomelanocortin (POMC) and neuropeptide Y (NPY) gene expression (Sahu 1998). However, although mutations in the human leptin gene do occur (Montague *et al.* 1997), they are rare (Echwald 1999; Schalling *et al.* 1999). In humans circulating leptin levels are correlated to body fat content (Caro *et al.* 1996), and the higher levels observed in obese individuals suggest a leptin resistance mechanism of obesity involving the leptin receptor or other downstream mediators which control appetite (Echwald 1999; Kopelman 1999). Despite the rarity of leptin mutations, the sensitivity of the human energy homeostasis system to leptin may make leptin augmentation an appropriate treatment for some common forms of obesity in which leptin levels are lower (Farooqi *et al.* 2001).

A 106nt insertion mutation in the leptin receptor gene (*LepR*) causing a truncated intracellular domain is responsible for the severely obese phenotype of the diabetes mouse (db/db) (Chen *et al.* 1996). Less severely affected are KK mice, which have three nucleotide polymorphisms in *LepR*, one of which results in an aspartate to asparagine conversion in part of the second extracellular cytokine-receptor homology module – these mice have an apparently multigenic syndrome of moderate obesity and hyperinsulinaemia (Igel *et al.* 1998). Activation of the leptin receptor in the hypothalamus activates signal transducers and activators of transcription STAT-3, STAT-5 and STAT-6 (Ghilardi *et al.* 1996), suppresses neuropeptide-Y synthesis and release (Stephens *et al.* 1995) and induces pro-opiomelanocortin (POMC), which is cleaved to produce αMSH (Cummings & Schwartz 2000). A major functional mutation in *LepR* causing loss of transmembrane and intracellular domains was found to cause early-onset morbid obesity in one family, with lack of pubertal development and reduction in secretion of thyrotropin and growth hormone (Clement *et al.* 1998). Another polymorphism in the leptin receptor gene was found to be associated with change in body weight, fat mass and body mass index in Australian women (de Silva *et al.*

2001). A polymorphism in the 3'UTR of *LepR* was found to be associated with serum insulin levels and risk of type 2 diabetes in non-diabetic men, indicating that less critical polymorphisms may also affect the function of the leptin receptor (Lakka *et al.* 2000). Potential mediators of leptin activity are suggested by the presence of leptin receptors in neurons containing: neuropeptide Y (NPY), agouti-related peptide (AgRP), pro-opiomelanocortin (POMC), cocaine- and amphetamine-regulated transcript (CART), melanin-concentrating hormone (MCH) and orexin (Meister 2000).

Neuropeptide Y is an appetite stimulator (Miner *et al.* 1989) which is up-regulated by fasting (Boswell *et al.* 1999) and down-regulated by leptin (Mercer *et al.* 1997). A study in a Mexican population of 914 individuals found associations between a polymorphism in the *NPY* promoter and waist-to-hip ratio, suggesting that this is a potential contributing factor in obesity (Bray *et al.* 2000).

Agouti-related protein is a melanocortin receptor antagonist which compensates for a lack of NPY in NPY-/- mice, up regulates appetite and appears to have targets other than melanocortin receptor 4 (MC4R) in the hypothalamus (Marsh *et al.* 1999b). MC4R inhibits feeding when activated by α-melanocyte stimulating hormone (αMSH), although MC4R-/- mice still respond to other anorectic factors such as ciliary neurotrophic factor (CNTF) and corticotropin releasing factor (CRF) (Marsh *et al.* 1999a).

Expression of the neuropeptide melanin-concentrating hormone (MCH) is up-regulated in ob/ob (leptin deficient) mice, and up-regulated in both normal and obese mice under fasting conditions (Qu *et al.* 1996). Administration of leptin decreases production of hypothalamic melanin-concentrating hormone (Sahu 1998). Administration of MCH stimulates feeding in rats (Qu *et al.* 1996). This indicates a role for this hormone in weight, food intake and regulation.

Leptin regulates the secretion of growth hormone (GH) by increasing the expression of growth hormone releasing hormone (GHRH) and decreasing the expression of somatostatin (SMS) in rats (Cocchi *et al.* 1999), while GH upregulates IGF-II expression in liver (von Horn *et al.* 2002). Leptin also appears to have a direct role in down-regulating insulin-like growth factor I (IGF-I) production in fasting rats, although the mechanism for this is unclear (LaPaglia *et al.* 1998). The GH-IGF axis is implicated in obesity by several studies. Obesity correlates with decreased GH, increased free IGF-I and increased total IGF-II in prepubertal children (Argente *et al.* 1997). These results agree with another study on obese children, with lower leptin, higher IGF-II, unchanged total IGF-I and a higher IGF to IGF binding protein ratio (Radetti *et al.* 1998). Total IGF-I is negatively correlated with

visceral fat mass in middle-aged men (Mårin *et al.* 1993). Frystyk *et al* investigated obesity in men and women, and found suppressed levels of GH and IGFBP-1, unaltered levels of total IGF-I and increased levels of free IGF-I, IGFBP-3, insulin and IGF-II (Frystyk *et al.* 1995). Infusion of IGF-I suppresses GH and IGF-II levels, the former potentially being a feedback mechanism (Guler *et al.* 1989). Obesity therefore appears to involve a decrease in growth hormone, decreased or unaltered levels of total IGF-I, but an increase in IGF-II and free IGF-I, the latter potentially due to a decrease in IGF binding protein rather than an increase in production of IGF-I.

The insulin gene (*INS*) variable number tandem repeat (VNTR) polymorphism is associated with both juvenile (Le Stunff *et al.* 2000) and adult (O'Dell *et al.* 1999) obesity.

The uncoupling proteins 2 and 3 (UCP2 and UCP3) are involved in thermogenesis (Fleury *et al.* 1997; Gong *et al.* 1997). The *UCP2* gene is widely expressed in human tissues, and is up regulated in white fat in response to fat intake (Fleury *et al.* 1997). The *UCP3* gene is highly skeletal-muscle specific in humans (Boss *et al.* 1997), but also expressed in rodent brown adipose tissue (Gong *et al.* 1997). *UCP2* maps to chromosomal regions linked to obesity and its role in thermogenesis makes it a potential candidate obesity gene (Fleury *et al.* 1997). Regulation of these genes by leptin (Scarpace *et al.* 1998) also indicates a potential role in weight regulation by increased energy expenditure. However, while *UCP2* appears to influence energy metabolism (Walder *et al.* 1998) neither *UCP2* nor *UCP3* polymorphisms associate with body mass index or fat mass (Walder *et al.* 1998; Schrauwen *et al.* 1999).

The β-2 and –3 adrenergic receptors have a role in lipolysis (Mori *et al.* 1999). The Trp64Arg mutation in the β-3 adrenergic receptor is associated with visceral obesity, presumably due to reduced lipolysis in visceral adipose tissue (Kim-Motoyama *et al.* 1997). The Gln27Glu mutation in the β-2 adrenergic receptor is associated with subcutaneous fat accumulation, but not visceral fat (Mori *et al.* 1999). The Arg16Gly mutation in the β-2 adrenergic receptor appears to enhance lipolysis, with greater weight loss in dieting women with the polymorphism than without (Sakane *et al.* 1999).

Orexins are neuropeptides which increase food intake in rats and are up-regulated during fasting, indicating a potential role in regulation of feeding behaviour (Sakurai *et al.* 1998). Orexins act through two G protein-coupled receptors, type 1 and type 2 orexin receptors (Sakurai *et al.* 1998).

A comprehensive microarray-based study on gene expression in rodents demonstrated significant differences in levels of expression in 214 transcripts of 11,000 tested (Nadler *et al.* 2000). Genes that are involved in adipogenesis (defined as having increased expression

during adipogenesis) tended to have significantly lower expression in obesity. Many genes were involved in either obesity or diabetes, and a smaller number were involved in both, including the β-3 adrenergic receptor (Nadler *et al.* 2000).

The relationships of some of these genes and pathways are shown in Figure 1-6. This figure is based on the relationships indicated in the publications cited above.

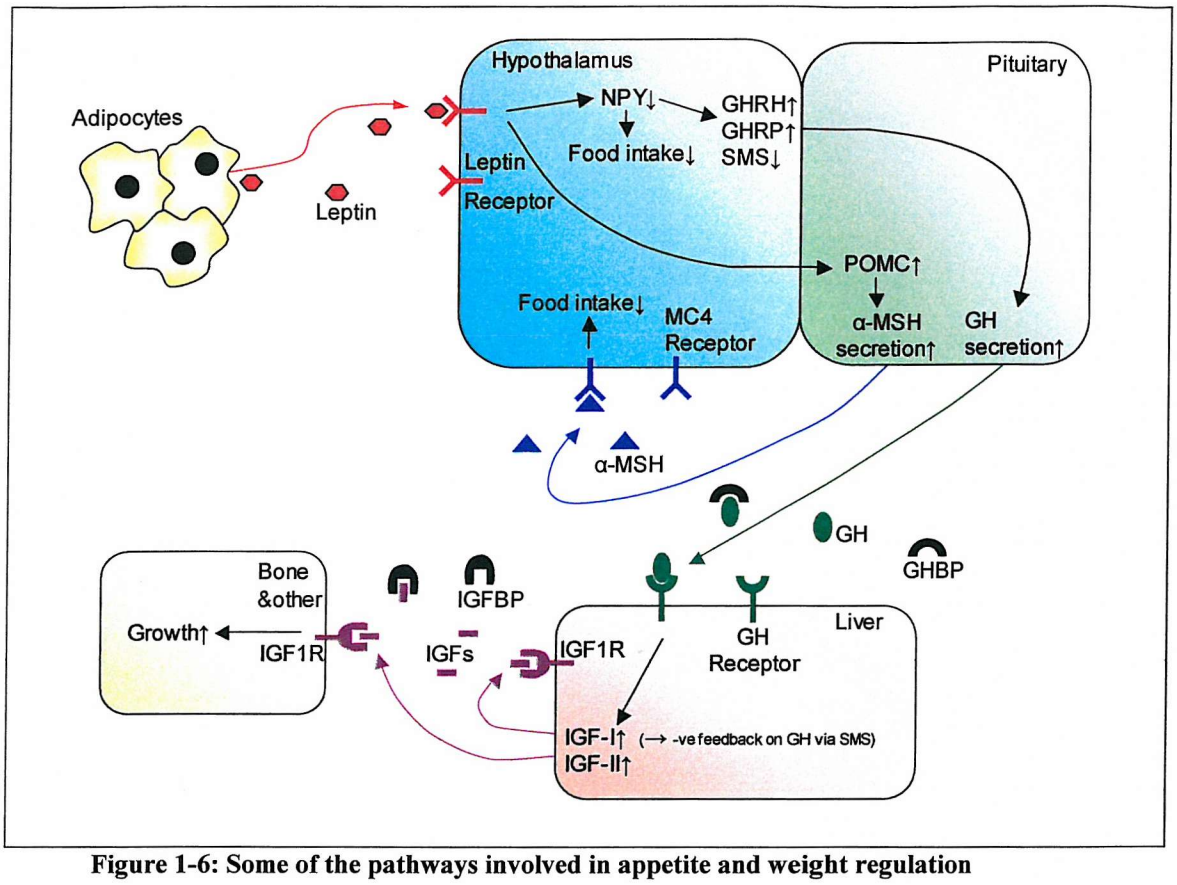


Figure 1-6: Some of the pathways involved in appetite and weight regulation

1.4.3 Role of *IGF2* in obesity

Insulin-like growth factors I and II (IGF-I and IGF-II) are growth factors that exert both metabolic and mitogenic effects through the type 1 IGF receptor (Nielsen 1992). They are therefore potential candidates for a role in weight determination.

A previous association study investigating a potential role for the *IGF2* gene in weight (O'Dell *et al.* 1997) used the *ApaI* polymorphism in the 3'UTR of *IGF2* (Tadokoro *et al.* 1991) as a linkage disequilibrium marker. AA homozygotes (non-cutting with *ApaI*) were found to have a higher mean serum IGF-II concentration in 92 middle-aged men and a 4 kilogram lower mean weight than GG homozygotes in 1474 healthy middle-aged men (O'Dell *et al.* 1997). This association suggests that the *IGF2* gene may be involved in

determination of weight. A study on IGF-I found a similar effect with this protein, with serum IGF-I negatively correlated with visceral fat mass (Mårin *et al.* 1993).

A recent study investigating the relationship between the *IGF2 ApaI* polymorphism and weight sampled 500 individuals of both sexes and with an age range of 19 to 90 years (Roth *et al.* 2002). While no association between genotype and BMI was found, the authors did report a higher fat mass (determined by dual-energy X-ray absorptiometry; DEXA) with *ApaI* AA genotype ($P<0.05$) in 427 Caucasian individuals (Roth *et al.* 2002), apparently opposite to the association between weight and *ApaI* GG genotype reported by O'Dell *et al* (1997). However, the cohort of Roth *et al* was significantly different from that of O'Dell *et al*. In the former study a small heterogeneous cohort was tested (men and women within a broad age-range), while in the latter selection criteria included only males within a narrow age-range.

Ukkola *et al* studied over-feeding (1000kcal surplus per day for 100 days) in twelve pairs of male twins (age 21 years ± 2 years) (Ukkola *et al.* 2001). In these individuals fat mass was significantly higher in *ApaI* GG homozygotes both before and after overfeeding than AA/AG individuals (combined). Subcutaneous fat was also significantly thicker after overfeeding (Ukkola *et al.* 2001). This study corresponds with that of O'Dell *et al* (1997), despite the lower age range and small sample number. This suggests that the apparent contradiction between Roth *et al* (2002) and the other two studies is due to the inclusion of women.

Other studies have investigated the levels of IGF-II protein in individuals of different weights. One study demonstrated that serum total IGF-II levels were higher in obese men ($25<BMI<30$ and $BMI>30$, $p<0.05$) than controls ($BMI<25$), and higher in severely obese women ($BMI>30$) than moderately obese women ($25<BMI<30$, $p<0.05$) (Frystyk *et al.* 1995). Raised serum free IGF-I, IGFBP-3 and insulin levels were found in obese individuals, while total IGF-I levels were normal and fasting serum growth hormone levels were suppressed in these subjects (Frystyk *et al.* 1995).

In children another study showed IGF-II raised in obesity, while IGF-I was not significantly different between controls and obese children and growth hormone was found to be lower in obese individuals (Radetti *et al.* 1998). Another study in children investigated the effects of weight reduction on IGFs and growth hormone. IGF-II and free IGF-I were significantly elevated in the serum of obese prepubertal children, even after a period of weight reduction, and IGF effects were largely growth hormone independent (Argente *et al.* 1997).

An animal-based study investigated the concentration of insulin-like growth factors in the tissues of pig foetuses; in pigs selective breeding has led to obese and lean animals differentiated by their back-fat thickness (Hausman *et al.* 1991). In lean foetuses liver and muscle IGF-II concentrations were higher than in pre-obese counterparts; it should be noted that lean foetuses are larger than pre-obese foetuses at the same stage of development (Hausman *et al.* 1991). A study in mice investigating the effect of a 12kb deletion downstream of *IGF2* (5' of the *IGF2-H19* imprinting control region – see 1.6.2) found that there was a decrease in *IGF2* expression (consistent with the deleted region containing a positive regulatory element (Jones *et al.* 2001). This reduction in *IGF2* expression was accompanied by increased fat deposition and obesity, but with hypophagia in obese animals suggesting that IGF-II alters fat metabolism rather than appetite (Jones *et al.* 2001).

A study in adults from three ethnic groups in Manchester (UK) found no relationship between IGF-II and BMI or WHR; these groups had high mean BMIs, but were not selected for obesity (Cruickshank *et al.* 2001). IGF-II showed strong ethnic differences, but was unrelated to other variables. IGF-I was inversely related to age, while IGFBP-1 levels were independently related to fasting insulin and BMI (Cruickshank *et al.* 2001). However, combined analysis of the relationship between BMI and IGF-II levels in men and women may be inappropriate.

The insulin-like growth factors appear to have a role in weight determination. The exact mechanisms are complicated by the relationship of the two insulin-like growth factors with growth hormone, IGF binding proteins and the type 1 IGF receptor. Alterations in the levels of any of these proteins may influence the levels and activity of the others. Elevation or suppression of the levels of an element in the pathway may contribute to weight determination, or merely be symptomatic of the activity of another element in the pathway. One hypothesis is that elevated free IGF-I may result from decreased availability of binding proteins (in part due to increased IGF-II), and may down-regulate growth hormone production, while maintaining normal growth in obese individuals (Argente *et al.* 1997). *IGF2* is therefore a good candidate gene for weight determination.

1.5 Foetal origins of adult disease

The 'foetal origins' hypothesis suggests that predisposition to cardiovascular and metabolic disease in later life is determined by adaptations to *in utero* endocrine status and

nutrition (Barker 1995). The term 'programming' describes a "general process whereby a stimulus or insult at a critical period of development has lasting or lifelong significance" (Lucas 1991). 'Metabolic imprinting' describes "basic biological phenomena that putatively underlie relations among nutritional experiences of early life and later diseases" (Waterland & Garza 1999), and is an alternative name for programming intended to reflect the irreversible nature of the effect (Waterland & Garza 2000). Whichever name is used, the fundamental reasoning of these theories is that environmental effects *in utero* influence health in later life.

The relationship between foetal development and adult consequences has been noted for coronary heart disease (Barker 1995; Eriksson *et al.* 2001), insulin resistance (Eriksson *et al.* 2002), obesity (Jackson *et al.* 1996; Ravelli *et al.* 1999) and even marital status (Phillips *et al.* 2001). The 'thrifty phenotype' hypothesis proposes that poor intra-uterine or infant growth plays a major role in the development of type II diabetes (Hales & Barker 1992; Hales & Barker 2001). The development of a thrifty phenotype is suggested to be beneficial for short-term survival in a restrictive environment, but detrimental to health in later (post-reproductive) life (Godfrey & Barker 2001). Poor growth in early life may be the result of maternal malnutrition, malnutrition in infancy (Hales & Barker 2001) and/or genetic factors (Frayling & Hattersley 2001; Dunger *et al.* 1998; Ong *et al.* 2000; Ong *et al.* 1999). The "fetal insulin hypothesis" proposes that genetically determined insulin resistance results in both poor growth *in utero* and insulin resistance in adult life (Hattersley & Tooke 1999). Hattersley *et al* found that a glucokinase mutation in the mother resulting in hyperglycaemia caused higher birth weight, while a glucokinase mutation in the foetus resulted in lower birth weight (Hattersley *et al.* 1998). A further study by Velho *et al* found no association between maternal and foetal glucokinase mutations and adult height, weight or BMI in maturity onset diabetes of the young (MODY2) kindreds (Velho *et al.* 2000).

A study comparing obesity in 50 year old men and women exposed to famine in late, mid or early gestation with men and women not exposed to famine found significantly higher BMI in women exposed to famine in early gestation (Ravelli *et al.* 1999). However, men showed no significant difference in BMI between exposed and non-exposed groups, and neither did women exposed in mid or late gestation (Ravelli *et al.* 1999). An earlier study on 300,000 19 year old men exposed *in utero* to the Dutch famine (1944-45) reported that famine during the last trimester of development and the first few months of life resulted in lower obesity rates, while exposure to famine in the first half of pregnancy resulted in significantly higher obesity rates (Ravelli *et al.* 1976). This suggested that poor nutrition in

the early stages of pregnancy affected development of the systems regulating food intake and growth resulting in an excess of fat accumulation with increased food availability later in life (Ravelli *et al.* 1976).

These results suggest that both environmental and genetic factors are involved in the observed relationships between foetal growth and adult disease. The relative effects of these factors in determining early growth are therefore important in determining the potential significance of genetic factors in metabolic and cardiovascular disease in later life.

1.6 The IGF2 gene and its product

1.6.1 The genetics of *IGF2*

Until recently the published gene sequence for *IGF2* consisted of several overlapping GenBank sequences with several gaps (Figure 1-7). Now the complete sequence of the gene has been published in a comparison with the orthologous domain on mouse chromosome 7 (Onyango *et al.* 2000). The gene is ~35kb in length, and is located at the telomeric end of the short arm of chromosome 11 (11p15.5) (Brissenden *et al.* 1984) ~5.6kb downstream of the tyrosine hydroxylase (TH) gene and ~1.5kb downstream of the insulin (*INS*) gene (Lucassen *et al.* 1993).

The gene contains 10 exons (Ikejiri *et al.* 1991; Mineo *et al.* 2000) and has four promoters (Pagter-Holthuizen *et al.* 1988; Hyun *et al.* 1993). The preproIGF-II peptide is encoded by exons 7, 8 and the first part of exon 9 (following conventional numbering, which assumes nine exons), while exons 1 to 6 (including 4b) constitute different 5'UTRs depending on promoter (Nielsen 1992).

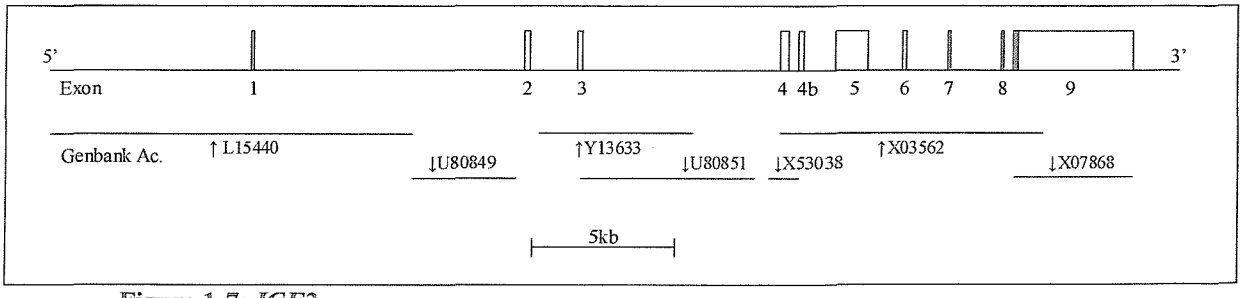


Figure 1-7: *IGF2* sequence

The structure of the *IGF2* gene has led to some confusion over the past decade. The gene has four promoters, and is commonly assumed to have nine exons (von Horn *et al.*

2002; van Dijk *et al.* 2001; Gaunt *et al.* 2001; Sakatani *et al.* 2001; Wu *et al.* 1997b; Vu & Hoffman 1996; Vu & Hoffman 1994). However, a ten exon structure has recently been confirmed (Figure 1-8) (Mineo *et al.* 2000). The presence of a 10^{th} exon (designated 4b) between exons 4 and 5 had previously been reported (Ikejiri *et al.* 1991), but was assumed to be a unique event in a human histocytoma cell line (Mineo *et al.* 2000). Exon 4b falls under the control of the P2 promoter, and is alternatively spliced, with P2 transcripts consisting either of exons 4, 7, 8 and 9 or exons 4, 4b, 7, 8 and 9 (Mineo *et al.* 2000).

The relative level of expression of *IGF2* from the four promoters varies depending on life stage and tissue. In liver, only promoters 2, 3 and 4 are used in the foetus, while promoter 1 becomes active from about two months after birth and promoter 3 activity is frequently low or non-existent in adults (Li *et al.* 1996). The imprinting of both *IGF2* and *H19* are influenced by an imprinting control region upstream of *H19* (see section 1.6.2 (Reed *et al.* 2001; Frevel *et al.* 1999)). In mice a positive regulatory element for *IGF2* expression exists upstream of this region (Jones *et al.* 2001). Two more regulatory elements exist in promoter 1, and form an inverted repeat, which is bound by inverted repeat binding protein (IRBP) and suppresses P1 activity (Rodenburg *et al.* 1996).

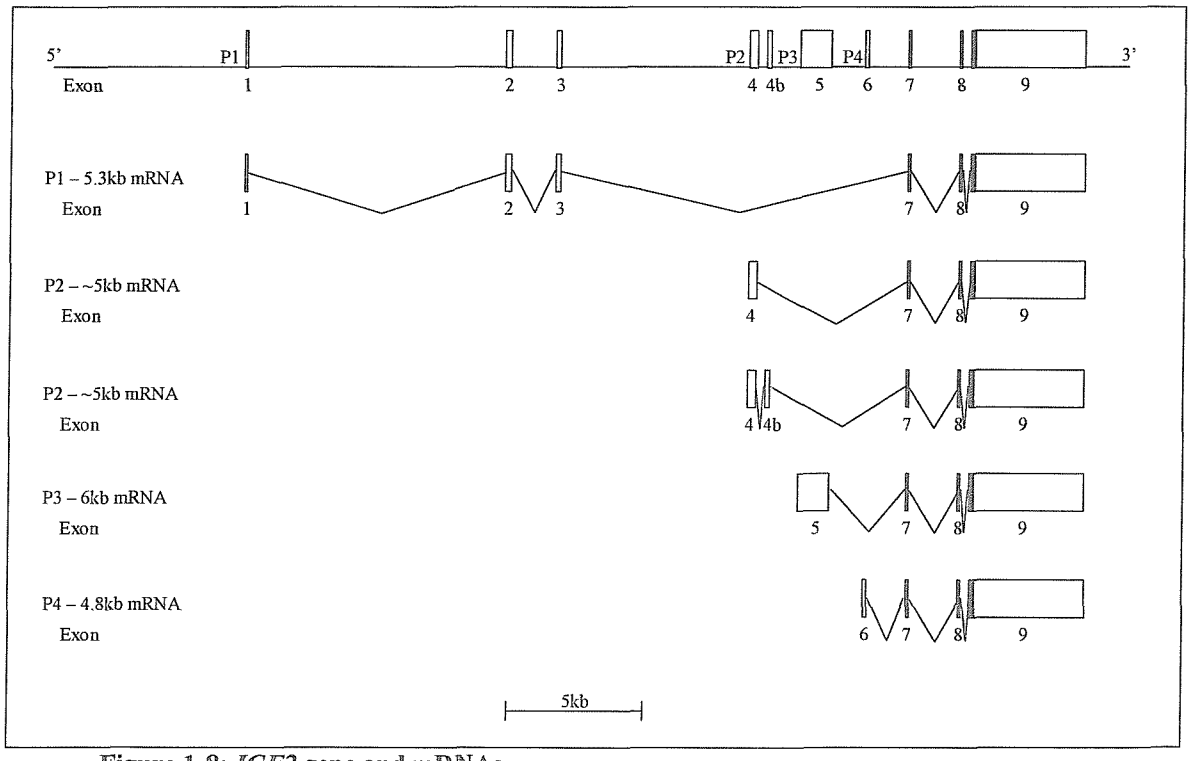


Figure 1-8: *IGF2* gene and mRNAs
(adapted from Nielsen, 1992 and Mineo *et al.*, 2000)

IGF2 mRNA levels are regulated by an endonucleolytic cleavage process involving a site in the 3'UTR of the gene. Within the 3'UTR there is a conserved sequence with a predicted strong higher order structure (Figure 1-9) (Christiansen *et al.* 1994). The cleavage

event occurs in a conserved area between a potential double-hairpin structure and a potential guanosine quadruplex. The cleavage is rare and rate-limiting (Nielsen 1992), and results in an upstream transcript with no polyadenylated tail, which has a short half-life, and a relatively stable 1.8kb 3' cleavage product (which may be protected by the quadruplex structure) (Christiansen *et al.* 1994). It is suggested that the higher order structure in the surrounding sequence is important in protecting the purine-rich cleavage site from involvement in other secondary structure, thus making it available for cleavage, or alternatively that the secondary structures may provide binding sites for trans-acting factors (Nielsen 1992).

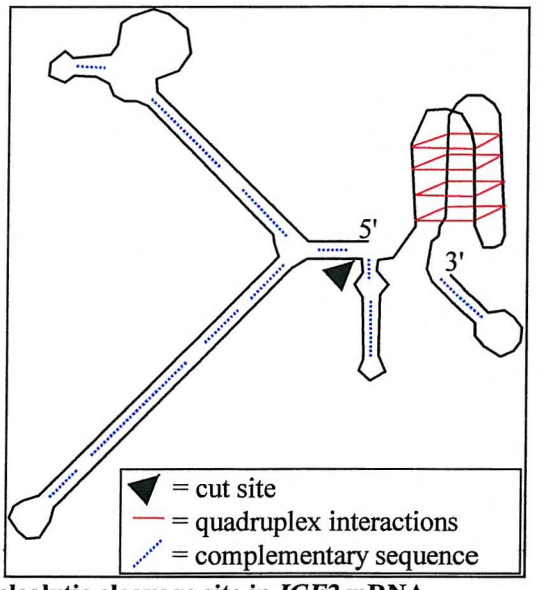


Figure 1-9: Endonucleolytic cleavage site in *IGF2* mRNA
(adapted from Christiansen *et al.*, 1994)

The cleavage of IGF-II mRNAs involves two elements in the 3'UTR, separated by almost 2kb, which act in *cis* (Figure 1-10) (Meinsma *et al.* 1992; Scheper *et al.* 1995; Scheper *et al.* 1996b; Scheper *et al.* 1996a). These elements in the transcript interact to form a stem structure, which is bound by IGF-II cleavage unit binding protein (ICU-BP) (Scheper *et al.* 1996b). Presence of the stem-loop structure stabilises the two additional stem-loops in element II (Scheper *et al.* 1995). Binding by ICU-BP requires both elements I and II, and appears to be necessary but not sufficient for cleavage to occur as a first step in an IGF-II mRNA degradation model (Scheper *et al.* 1996b). Two IGF-II cleavage unit RNA-protein complexes (ICU-RPC1 and ICU-RPC2) are found to exist in Hep3B cells (Scheper *et al.* 1996a). IGF-II activates a signalling pathway involving $p70^{S6k}$, which results in a transition from ICU-RPC2 to ICU-RPC1, rendering the IGF-II mRNA susceptible to cleavage – this is a potential negative-feedback system for IGF-II (Scheper *et al.* 1996a).

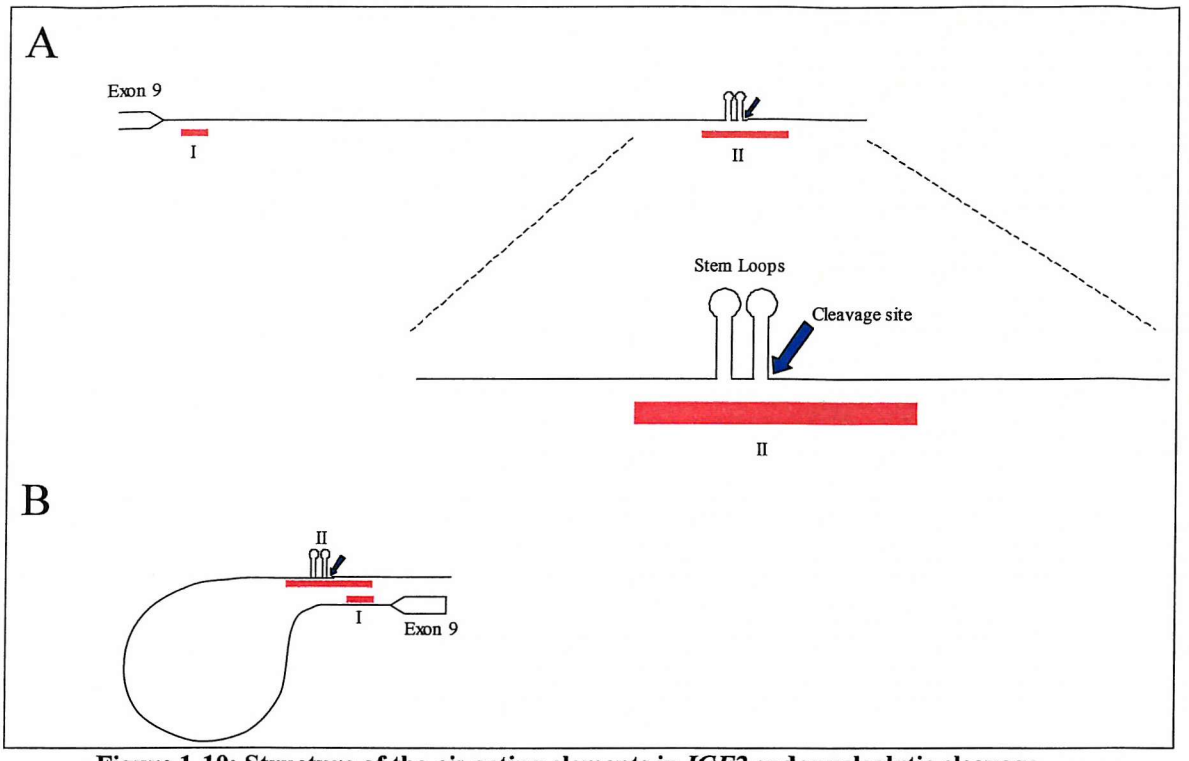


Figure 1-10: Structure of the *cis*-acting elements in *IGF2* endonucleolytic cleavage
Adapted From Scheper *et al.*, 1996b. (A) Element I is located just downstream of the *IGF2* stop codon, while element II is located a further approximately 2kb downstream. There are two stem-loop structures in element II and the cleavage site is marked with a blue arrow. (B) An interaction between elements I and II results in a stem structure, with the two stem loops in element II and the cleavage site shown.

The critical elements for endonucleolytic cleavage in *IGF2* appear to be:

1. two elements in the 3'UTR of the gene, ~2kb apart, the second of which contains the cleavage site (Scheper *et al.* 1995; Scheper *et al.* 1996b)
2. a double hairpin/stem-loop structure immediately 5' of the cleavage site (Scheper *et al.* 1996b; Christiansen *et al.* 1994)
3. a guanosine-rich region immediately 3' of the cleavage site with a putative quadruplex structure (Christiansen *et al.* 1994)
4. the formation of an IGF-II cleavage unit RNA-protein complex (ICU-RPC1) (Scheper *et al.* 1996a)

This type of regulation of mRNA levels has been observed in several other genes. In avian apolipoprotein II, mRNA is degraded following an endonucleolytic cleavage event at 5'-AAU-3'/5'-UAA-3' elements in the 3'UTR of the gene (Binder *et al.* 1989). This mechanism differs from the *IGF2* cleavage mechanism in that it appears to occur at several different sites, and cleavage occurs at a different sequence motif. However, cleavage in both genes allows degradation of message without deadenylation (Binder *et al.* 1989).

The protooncogene c-*myc* appears to be subject to regulation of RNA stability by endonucleolytic cleavage resulting in generation of 3' and 5' truncated species in the decay process (Ioannidis *et al.* 1996). The exact mechanism of this cleavage is not clear, but there is evidence for multiple-site cleavage in this system (Ioannidis *et al.* 1996).

Endonucleolytic cleavage in the cytokine *groα* results in removal of a 130-nucleotide sequence from the 3'UTR of the gene leaving a 0.9 kilobase (poly-A$^-$) product. This process is regulated by interleukin-1 (Stoeckle 1992).

Albumin mRNA in *Xenopus* liver is cleaved at multiple sites by an oestrogen-regulated polysomal ribonuclease that appears to recognise the sequence APyrUGA (Chernokalskaya *et al.* 1997). This sequence differs from both the avian apolipoprotein II and the human *IGF2* cleavage sequences, and is regulated by oestrogen (unlike the *groα* cleavage system).

While endonucleolytic cleavage of RNA appears to occur in a number of genes, the mechanism and regulation differs between genes. The result in all cases is likely to be degradation by exonuclease activity due to the absence of a protective cap at the 5' ends, and polyadenylation at the 3' ends of the cleaved RNAs. In human *IGF2* the 3' fragment is stable, which may be due to the protective effect of the guanosine quadruplex against 5'-3' exonuclease activity (Christiansen *et al.* 1994).

1.6.2 Imprinting in the *IGF2* region

Genomic imprinting is the epigenetic modification of genes within oocytes or sperm that results in parentally determined differential expression of those genes post-fertilisation (Reed *et al.* 2001). *IGF2* and *H19* (downstream of *IGF2*) are both parentally imprinted, with *IGF2* preferentially expressed from the paternal allele, and *H19* preferentially expressed from the maternal allele (Rainier *et al.* 1993). These two loci are closely linked on chromosome 11, and are of interest because of their reciprocal imprinting pattern (Ohlsson *et al.* 1994). An imprinting control region (ICR) 2kb upstream of *H19* and ~70kb downstream of *IGF2* is critical in regulation of the imprinting and expression of *IGF2* (Reed *et al.* 2001; Frevel *et al.* 1999). Imprinting of *IGF2* in blood cells has been shown to be dependent on expression of *H19*, with monoallelic expression occurring only in those individuals that express *H19* (Giannoukakis *et al.* 1996). Expression of the *IGF2* gene varies by promoter: promoter 1 is responsible for bi-allelic expression of *IGF2*, while promoters 2-4 initiate expression only

from the paternal allele (Vu & Hoffman 1994). Differential methylation at CpG islands is believed to be responsible for imprinting (Li *et al.* 1993). DNA-methyltransferase deficient mice show expression from the paternal (normally unexpressed) allele of *H19*, while the normally expressed *IGF2* paternal and *IGF2R* maternal alleles are repressed (Li *et al.* 1993). The role of *H19* in the regulation of *IGF2* in *trans* (Li *et al.* 1998) fits with this pattern – *H19* suppresses *IGF2* expression, and is regulated normally by maternal imprinting. *IGF2* itself is regulated by paternal imprinting and suppressed by *H19* (Li *et al.* 1998).

The common control of *IGF2* and *H19* is assumed to arise from a system of methylation and enhancers: on the maternal allele cis-acting elements (enhancers) preferentially activate transcription of *H19* (at the expense of *IGF2*), while on the paternal allele methylation of regulatory sequences prevents this interaction, allowing the enhancers to activate *IGF2* (Leighton *et al.* 1995). An evolutionary theory may explain why this mechanism exists. A gene for a protein that increases the demand for resources from the mother (e.g. a foetal growth factor such as IGF-II) will be transcribed less from the maternal allele if a mechanism exists for this to occur (the evolutionary stable state is no transcription from the maternal allele), whereas it is in the interests of the father that his allele is expressed as much as possible (to benefit his offspring at the expense of half-sibs) (Haig & Westoby 1989). If this theory is correct, then one would expect paternal imprinting of loci that increase foetal growth and nutrition (e.g. *IGF2*) and maternal imprinting of loci that regulate those loci (e.g. *H19*) or proteins (e.g. *IGF2R*). In mouse the *IGF2* and *IGF2R* (type 2 IGF receptor) genes are oppositely imprinted, which agrees with the hypothesis that the balance between the maternal imprinting of the receptor and the paternal imprinting of *IGF2* regulates foetal growth – maternally derived receptor responsible for degradation of IGF-II counteracts and regulates the paternally derived growth-promoting IGF-II (Haig & Graham 1991).

1.6.3 The structure and function of IGF-II

The mature insulin-like growth factor II protein is a 67 amino acid polypeptide which has a 62% amino acid sequence homology with insulin-like growth factor I (Rinderknecht & Humbel 1978). IGF-II has a molecular weight of 7471 (Rinderknecht & Humbel 1978), and is a single polypeptide chain with three disulphide bridges (Smith *et al.* 1989) (which are at corresponding positions in insulin, IGF-I and IGF-II (Lowe 1996)). In IGF-II the disulphide

bridges are between cysteines 9→47 (peptide surface), 21→60 (core) and 46→51 (core) (Terasawa *et al.* 1994).

PreproIGF-II is a 180 amino acid polypeptide comprising six domains: a 24 amino acid signal peptide at the N-terminal, domains B, C, A and D of the mature peptide and an 89 amino acid carboxy-terminal E domain which is removed during processing (Figure 1-11) (Lowe 1996). The prepro-peptides for all the insulin-like peptides have similar structural organisation (Nielsen 1992). In proinsulin, IGF-I and IGF-II there are B, C (connecting peptide) and A domains, while the C-terminal D domain is present only in IGF-I (8 residues) and IGF-II (6 residues) (Rinderknecht & Humbel 1978). IGF-II contains three α-helices, one at Glutamic acid 12 – Cysteine 21 in the B domain, one at Glutamic acid 44 – Arginine 49 in the A domain and the third at Leucine 53 – Tyrosine 59 in the A domain (Terasawa *et al.* 1994). The two A domain α-helices lie parallel (reverse orientation), and form a hydrophobic core in conjunction with the B domain α-helix (Terasawa *et al.* 1994).

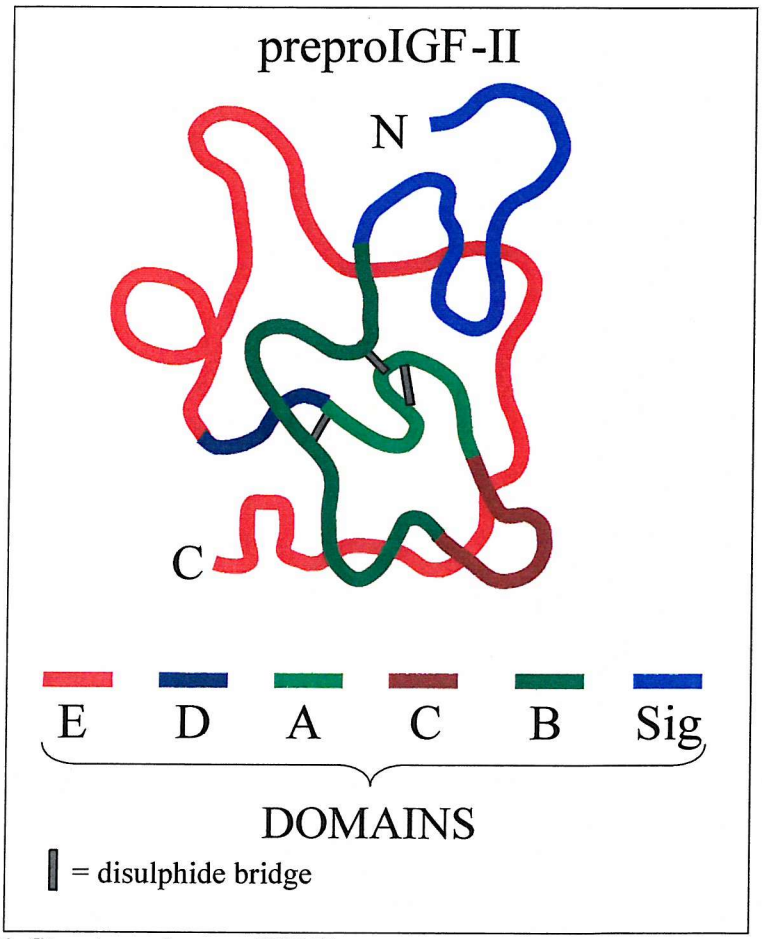


Figure 1-11: Structure of preproIGF-II
(adapted from Lowe, 1996)

IGF-II primarily acts through the type 1 IGF receptor (IGF1R), which consists of two ligand-binding α-subunits and two membrane-spanning β-subunits (similar to the insulin receptor - see Figure 1-12) (Nielsen 1992). The critical IGF-II residues in binding with the type 1 IGF receptor are Tyrosine 27 and residues 1-6 (Forbes *et al.* 2002). Binding of IGF-II to the type 1 receptor results in a countering effect of translocation of the type 1 receptor to the cell membrane (Nielsen 1992). Activation of the type 1 receptor (a tyrosine kinase) results in both autophosphorylation of the receptor (Jacobs *et al.* 1983) and phosphorylation of other proteins, including IRS-1 (insulin receptor substrate 1) (Nielsen 1992). This initiates a signalling cascade to the nucleus, resulting in transcription factor phosphorylation and DNA synthesis (Lowe 1996). Both IGF-I and IGF-II bind the type 1 receptor with high affinity, while insulin binds with low affinity (Casella *et al.* 1986).

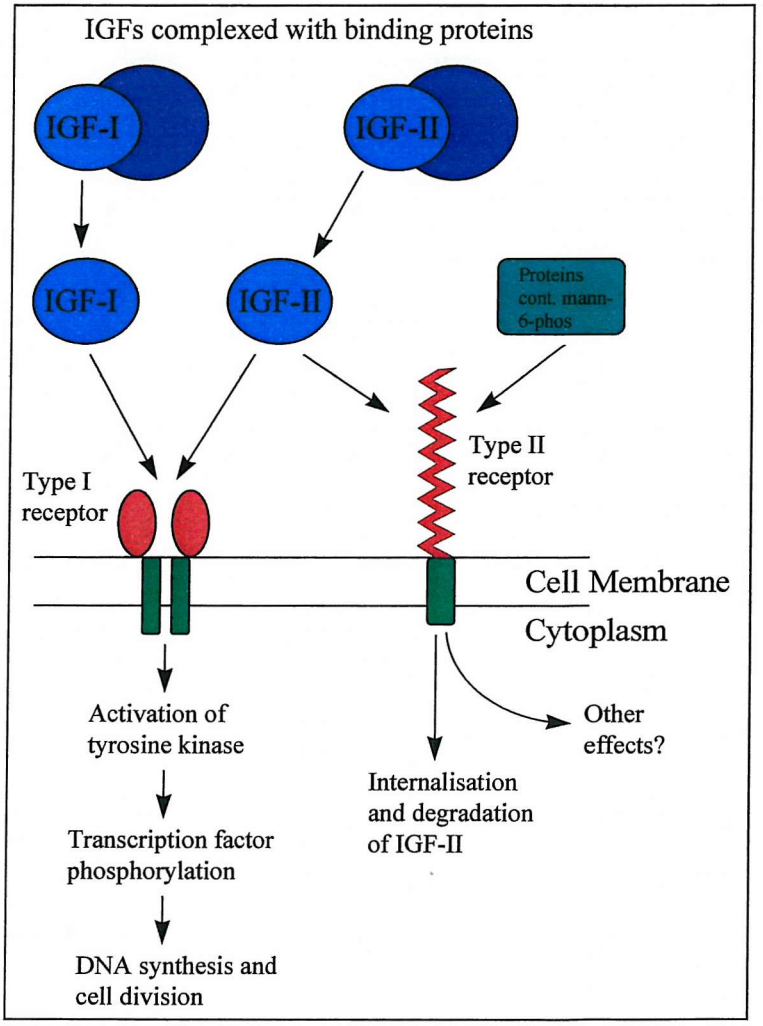


Figure 1-12: IGF-II function
(adapted from Nielsen, 1992 and Lowe, 1996)

IGF-II can also act through the insulin receptor, and has a growth-promoting action through this receptor in mouse embryogenesis (Louvi *et al.* 1997). IGF-II and insulin both

stimulate cell proliferation and DNA synthesis by binding this receptor, while IGF-I does not act through the insulin receptor (Morrione *et al.* 1997). The difference in receptor affinity between the IGFs is proposed to be due to side-chain differences rather than major structural differences (Torres *et al.* 1995). In humans IGF-II acts through isoform-A of the insulin receptor in stimulation of the growth of thyroid cancer (Vella *et al.* 2002).

The type 2 IGF receptor (IGF2R, also known as the mannose-6-phosphate receptor) is specific to IGF-II, with minimal binding of IGF-I and no binding of insulin (Terasawa *et al.* 1994; Linnell *et al.* 2001). The critical residues for binding differ from those for the IGF1R: Phenylalanine 48, Arginine 49, Serine 50, Alanine 54 and Leucine 55 (Terasawa *et al.* 1994). Binding of IGF-II to the type 2 receptor is not responsible for stimulation of DNA and glycogen synthesis (Sakano *et al.* 1991). The IGF2R has a negative regulatory effect on the activity of IGF-II, resulting in reduced organ growth (Zaina & Squire 1998). Soluble IGF2R also plays a role in the transport of circulating IGF-II (Valenzano *et al.* 1995).

IGF binding proteins (IGFBP) interact with both IGF-I and IGF-II. More than 95% of *IGF*s are complexed with IGFBPs in plasma (Nielsen 1992). Total serum levels of IGFBP-3 and IGF-II are elevated in obesity, while serum free IGF-II is constant (Frystyk *et al.* 1995), suggesting that the IGF binding proteins are very important in the regulation of serum free IGF-II. Some of the binding proteins are carrier proteins that cannot cross the endothelium in complex with IGF-II (e.g. IGFBP-3), while others can cross the endothelium with IGF-II and may influence interaction with the type 1 receptor (Lowe 1996). IGFBP-3 is the major binding protein for IGFs in serum, increasing the half-life of these proteins and creating an intra-vascular storage reservoir (Lowe 1996). IGFBPs 1 to 6 bind IGF-II with high affinity, while all except IGFBP-6 bind IGF-I (albeit with only 20 to 74% the affinity of IGF-II) (Bach *et al.* 1993). IGFBP-7 and 8 are members of a putative family of lower-affinity IGF binding proteins (Kim *et al.* 1997).

A number of high-molecular weight forms of IGF-II occur in cattle (at least 12) of 12 to 20kDa, with amino-terminal sequence matching that of the mature 7.5kDa IGF-II (Valenzano *et al.* 1995). They are associated with circulating soluble IGF2R (SIGF2R) (Valenzano *et al.* 1995) and display a mitogenic activity that is not inhibited by IGF binding proteins in the same way as IGF-II (Blahovec *et al.* 2001). However, the affinity of high-molecular weight forms for IGF binding proteins appears similar to that of IGF-II, and in fibroblasts the biological activity of the different forms appears similar (Valenzano *et al.* 1997). High-molecular weight forms have also been found in humans (Blahovec *et al.* 2001).

1.7 Objectives of the study

The hypothesis of this study is that one or more polymorphisms or haplotypes within or near the *IGF2* gene influences weight in middle-aged men. There is evidence that the *ApaI* polymorphism in the 3'UTR of the *IGF2* gene is associated with BMI in middle-aged men (O'Dell *et al.* 1997). The aim of this project is to investigate *IGF2* thoroughly to find polymorphisms, test these for association with weight phenotypes and clarify the role of *IGF2* in weight determination. There are five main objectives of this study:

The first objective was to identify single-nucleotide polymorphisms (SNPs) in the *IGF2* gene to build a set of potential markers for linkage disequilibrium mapping of a possible aetiological site. To identify relatively rare polymorphisms while still allowing efficient screening, a subset of 40 individuals from the Northwick Park Heart Study II (NPHSII) cohort were used. 40 individuals should allow detection of polymorphisms with a heterozygote frequency of 2.5% or more (allele frequency of 1.25%). Single-strand conformation polymorphism (SSCP) analysis and denaturing high performance liquid chromatography (DHPLC) were used for this part of the project, and regions of the gene were selected for serial scanning due to the limited throughput of these systems.

The second objective of the study was the development of high-throughput genotyping methods to enable the genotyping of several SNPs in the 2743 individuals of NPHSII. To genotype 10 SNPs in 2743 samples a total of 54,560 PCR reactions and gel tracks are required (by dual-reaction ARMS assay). Existing technologies were either too expensive or too inefficient for this. In addition, the large quantity of DNA required necessitated the adoption of a conservative system for use of DNA banks, preferably by DNA amplification.

The third objective was the genotyping of 10 SNPs in the NPHSII cohort and analysis with respect to weight phenotypes in order to identify any positive associations with weight in middle-aged men.

The fourth objective was the genotyping of 10 SNPs in the smaller Hertfordshire cohort, for which more extensive phenotype data is available to enable an analysis of associations between genotype and both early growth and adult weight phenotypes. This would enable an investigation of any potential role of the (foetal) growth factor *IGF2* in determining birth weight, early growth and adult disease. This cohort also contains both men and women, allowing an investigation of the effect of gender on any observed association between genotype and weight.

The fifth objective was to develop an RNA quantitation assay for the measurement of *IGF2* expression levels in individuals with different genotypes. This should help determine how genotype influences downstream levels of IGF-II, and therefore establish a link between gene expression, genotype and weight phenotypes.

Chapter 2 : General Methods

2.1 Materials

2.1.1 DNA samples

2.1.1.1 Northwick Park Heart Study II

The largest set of DNA samples used in this project was collected for the Northwick Park Heart Study II (NPHSII, a prospective study of coronary heart disease - Director George Miller). The NPHSII cohort contains 2743 samples of DNA from healthy men aged 51-62 years (at time of collection). Samples were supplied as standardised 1/5 dilutions of original extracted DNA. Initially 1/40 dilutions were used for the SSCP scan and genotyping. Later the DNA bank was equalised to 10ng/μl by Tricia Briggs. This equalised DNA was used for later genotyping and long PCR. Stocks are stored at -70°C, while working arrays are stored at -20°C.

2.1.1.2 Hertfordshire DNA Bank

This DNA bank comprises two sub-sets: North Hertfordshire and East Hertfordshire. The total number of samples is 1108, of which 448 are women and 660 are men (unlike the Northwick Park Heart Study DNA bank, this DNA bank contains samples from both men and women). The age range of this cohort was 59 to 72 years at the time of DNA collection and

adult phenotype recording. Data for this DNA bank also included early life measures of birth weight and weight at one year.

2.1.2 Chemicals and Reagents

2.1.2.1 PCR reagents

Recombinant Taq DNA polymerase (Life Technologies, Paisley, UK and Promega, Southampton, UK) was used at a concentration of 0.02 Units per microlitre of reaction. $MgCl_2$ (Life Technologies) was used in reactions at concentrations of between 1.0mM and 2.5mM depending on optimisation results. Detergent W1 (Life Technologies) was used in some reactions to enhance Taq activity.

10x polymerase mix was prepared using 0.5M KCl (BDH, Poole, UK), 100mM Tris-HCl (BDH) pH 8.3, 0.01% Gelatin and 2mM dNTPs (Amersham Life Sciences, Bucks, UK and Life Technologies). This was filter sterilised and stored in aliquots at -20°C for up to six months. Polymerase mix was used at a 1x concentration in all standard PCRs. An alternative PCR buffer containing 0.5M NaCl (BDH) instead of KCl was used for amplification of the sequence including the mRNA processing site in the 3'UTR. Some PCRs were performed using the standard 10x polymerase buffer provided with the enzyme (Life Technologies or Promega), with dNTPs (Amersham Life Sciences, Bucks, UK and Life Technologies) added separately.

Long PCR buffer was prepared using 140mM Ammonium acetate (Sigma-Aldrich Company Ltd., Poole, UK), 500mM Tris-HCl (BDH) pH 8.9. Pwo DNA polymerase (Roche Diagnostics, Lewes, UK) was added to long PCR reactions at a concentration of 0.004Units per 10µl reaction.

A stock solution of Betaine (Sigma-Aldrich Company Ltd.) (a natural metabolite of choline (Haubrich *et al.* 1975)) was prepared at 5M, and stored at 4°C for up to six months. Betaine was used at a working concentration of 1.3M in PCRs of GC rich regions (i.e. 75% or greater) (Henke *et al.* 1997), and in long PCRs (Barnes unpublished http://barnes1.wustl.edu/~wayne/faq.wpa/).

PCR primers (MWG Biotech, Milton Keynes, UK) were rehydrated in water or 10mM Tris-HCl (pH 8) and stored as a 100µM solution at -20°C. Final concentration of oligonucleotide primers in the PCR reaction varied, but was usually 0.4µM.

Water was purified to approximately 16.4MΩ resistance by reverse osmosis, then de-ionisation. Water was autoclaved to sterilise before use in PCR reactions.

2.1.2.2 Electrophoresis reagents

10x Tris Borate EDTA (T.B.E) buffer was prepared using 0.89M Tris (BDH), 0.89M Boric Acid (Sigma-Aldrich Company Ltd.) and 0.02M Ethylenediaminetetra-acetic acid (EDTA) (BDH).

30% Acrylamide:Bis (19:1) (Severn Biotech, Kidderminster, UK) was stored at 4°C and polymerised to form gel matrices at concentrations of 4% to 10% depending on resolution required. N,N,N'-Tetramethylethylenediamide (TEMED) was obtained from Sigma-Aldrich Company Ltd. Ammonium persulphate (APS) was obtained from Fisher Scientific UK Ltd, Loughborough, UK. These reagents were used to catalyse polymerisation of acrylamide gel solutions.

Agarose (Life Technologies) was used to prepare agarose gels. 0.7g of agarose was melted in 100mls of 1x T.B.E buffer to create a 0.7% gel.

Ethidium bromide ($10mgml^{-1}$) (Life Technologies) was stored in lightproof containers at room temperature and used at a working concentration of $1mgml^{-1}$ as a pre-electrophoresis stain to detect double-stranded DNA.

Formamide dye mix was prepared using 98% formamide (Sigma-Aldrich Company Ltd), 10mM EDTA (pH 8) (BDH) and 0.025% Xylene Cyanol (Sigma-Aldrich Company Ltd). This was used as a 3x mix (i.e. one volume of formamide dye mix to two volumes of sample).

"Sticky silane" was prepared using 90% Ethanol, 0.5% glacial acetic acid (BDH) and 0.5% γ-methacryloxypropyltrimethoxysilane (Sigma-Aldrich Company Ltd).

2.1.2.3 Single-Strand Conformation Polymorphism Analysis (SSCP)

Two different SSCP gels were used (Hinks *et al.* 1995). Gels to be run at 4°C were prepared with 0.5x T.B.E, 5.7% acrylamide (BDH) and 0.152% bis-acrylamide (Promega, Southampton, UK). A volume of 100mls of gel mix was then polymerised with 71µl TEMED and 210µl 25% ammonium persulphate.

Gels to be run at 25°C were made as above, but with the addition of 5% glycerol (Sigma-Aldrich Company Ltd). Running buffer for both gels was 0.5x T.B.E, with 5% glycerol added for gels run at 25°C.

Loading buffer for SSCP was prepared with 100mls formamide, 0.1g Xylene Cyanol, 0.1g bromophenol blue (Sigma-Aldrich Company Ltd) and 4mls 0.5M Na_2EDTA. Medical X-Ray Film was obtained from Genetic Research Instrumentation Ltd., Braintree, UK.

2.1.2.4 Denaturing High Performance Liquid Chromatography (DHPLC)

Pre-made buffers "A" and "B" containing 100mM Triethylammonium acetate (TEAA) and 0.1mM EDTA plus 25% acetonitrile (Aldrich, Poole, UK) in buffer B were obtained from Varian Ltd., Walton-on-Thames, UK. Alternatively buffers were made from TEAA concentrate (Transgenomic, Crewe, UK): Buffer "A" contained 100mM TEAA and buffer "B" contained 100mM TEAA and 25% acetonitrile.

Diluted acetonitrile wash solutions were prepared from HPLC grade acetonitrile by dilution with 18.2MΩ HPLC grade water (Sigma-Aldrich Company Ltd).

Samples for DHPLC were unmodified PCR reactions. Quality control standards were initially obtained from Transgenomic. Later controls were a heteroduplex formed from a PCR of marker DYS271 (sequence obtained from GenBank) and a HaeIII digest of pUC18 (Sigma-Aldrich Company Ltd).

2.1.2.5 Sequencing

Sequencing reagents obtained from Applied Biosystems, Foster City, CA, USA included: POP-6 polymer, BigDye Terminator cycle sequencing kit, Template Suppression Reagent (T.S.R) and 10x Genetic Analyzer Buffer. A 2.5x dilution buffer for the BigDye Terminator kit was prepared using 200mM Tris-HCl (BDH) pH 9.0 and 5mM $MgCl_2$. This was used to replace half of the recommended quantity of Big Dye ready reaction mix.

Shrimp Alkaline Phosphatase and Exonuclease I were obtained from Amersham Life Sciences, Bucks, UK. Primers used for sequencing (also used in PCR) were obtained from MWG Biotech, Milton Keynes, UK. Ethanol and 3M Sodium acetate pH 5.2 (Sigma-Aldrich Company Ltd) were used for precipitation of sequencing reactions to remove excess salts and dye-labelled ddNTPs.

2.1.2.6 Quantitative RT-PCR

RNA extraction was performed from blood samples (anonymised, from volunteers within the division) or from HepG2 and HuH7 liver cells (provided by Derek Mann, School of Medicine, Southampton University) using the SV 96 Total RNA Isolation System

(Promega, Southampton, UK). Sodium acetate and RNase-free ethanol were obtained from Sigma-Aldrich Company Ltd., Poole, UK. Trypsin-EDTA (Life Technologies, Paisley, UK) was used to loosen cells from culture flasks. Following extraction, samples were treated with DNase using DNA-*free*™ (Ambion, Huntingdon, UK). Reverse transcription (RT) was performed using M-MLV Reverse Transcriptase, while RNase activity was inhibited with RNasin® Ribonuclease Inhibitor (Promega, Southampton, UK). Water was treated with DEPC (Sigma-Aldrich Company Ltd) to destroy RNase. Oligonucleotide primers were supplied by MWG-Biotech, Milton Keynes, UK. T7 RNA Polymerase, RNasin® and rNTPs (Promega, Southampton, UK) were used to generate RNA standard from a PCR product. PCR of RT reactions used reagents as described in sections 2.1.2.1 and 2.1.3.1. *ApaI* restriction enzyme was from Promega, Southampton, UK.

2.1.2.7 $(CA)_n$ repeat genotyping

5' 6-FAM and HEX labelled oligonucleotides were obtained from MWG Biotech, Milton Keynes, UK. Standard PCR was performed using reagents as described in sections 2.1.2.1 and 2.1.3.1. PCR products were digested with BstUI (New England Biolabs UK Ltd, Herts, UK).

Samples were prepared for electrophoresis by combining 1µl of PCR product with 12µl of a 24:1 mix of formamide (Sigma-Aldrich Company Ltd) and Tamra 500 size standard (Applied Biosystems, Foster City, CA, USA). Samples were through POP4 polymer using 10x Genetic Analyzer buffer on an ABI310 genetic analyser (Applied Biosystems, Foster City, CA, USA).

2.1.3 Equipment

2.1.3.1 Polymerase Chain Reaction (PCR)

Polypropylene 96-well PCR plates were obtained from Advanced Biotechnologies, Epsom, UK. Rubber plate sealing mats were obtained from Hybaid Ltd., Ashford, UK. Polypropylene 384-well plates and sealing mats were obtained from Genetic Research Instrumentation Ltd. PCRs were performed on PTC-225 TETRAD DNA Engines (MJ Research Inc., Waltham, MA).

2.1.3.2 Microplate Array Diagonal Gel Electrophoresis (MADGE)

MADGE (Day & Humphries 1994) gel formers were supplied by MADGEBio, Grantham, UK. Several variants were used: 96-well MADGE (Day & Humphries 1994) with 71.6° orientation, $2mm^3$ wells and 26mm track length, 192-well MADGE (O'Dell *et al.* 2000) with 71.6° orientation, $2mm^3$ wells and 12mm track length and 384-well MADGE (Gaunt *et al.* 2000) with 78.7° orientation, $1.5mm^3$ wells and 10mm track length.

GelBond Film (Sigma-Aldrich Company Ltd) was used as a support for agarose MADGE gels.

Horizontal gel electrophoresis tanks were obtained from Sigma-Aldrich Laboratory Equipment, Poole, UK, and custom-built dry gel electrophoresis systems were provided by Medical Electronics and Engineering at Southampton General Hospital. Power Packs for electrophoresis were obtained from Bio-Rad, Hemel Hempstead, UK.

Glass plates were cut from standard float glass to 110mm x 170mm were obtained from a local glass merchant (Shirley Glass, Shirley, Southampton).

Samples were loaded either by multi-channel pipette, or by 96-pin passive replicator (MADGEBio). Coloured loading templates were prepared to place under gels and aid positioning of the replicator in higher-density gels.

Images were scanned using a Fluorimager 595 (Molecular Dynamics Inc., California, USA). Image analysis was performed with Phoretix 1D Advanced (Phoretix International, Newcastle-upon-Tyne, UK).

2.1.3.3 Single-Strand Conformation Polymorphism (SSCP)

PCRs were performed using PTC-225 TETRAD DNA Engines (MJ Research). Gels were run using a Macrophor sequencing system (Pharmacia, Bucks, UK), comprising an electrophoresis chamber with thermostatic circulator, a gel casting unit and glass plates and spacers. A Bio-Rad power-pack was used for electrophoresis. A Bio-Rad Model 583 Gel Dryer was used to dry gels. A FujiFilm FPM800A developer (Fuji Photo Film UK Ltd., London, UK) was used to develop the autoradiography film after exposure.

2.1.3.4 Denaturing High Performance Liquid Chromatography (DHPLC)

DHPLC was performed initially using a Varian Pro-Star Helix DHPLC system (Varian Ltd.). Later DHPLC scanning was performed with a Transgenomic Wave DHPLC

system (Transgenomic). Columns for DHPLC were provided by the system suppliers, as were all replacement components (guard columns, high pressure tubing, filters etc).

2.1.3.5 Sequencing and $(CA)_n$ repeat genotyping

Sequencing and genescan reactions were performed on a TETRAD DNA Engine (MJ Research). Sequencing and genescan electrophoresis and detection were performed on an ABI PRISM 310 capillary sequencer (Applied Biosystems). Sample tubes, septa, racks and capillaries were also obtained from Applied Biosystems.

2.1.3.6 Quantitative-Competitive RT-PCR

RNA samples were quantified using a Genequant spectrophotometer (Amersham Life Sciences, Bucks, UK).

2.2 Methods

2.2.1 PCR

Primers for PCR were designed manually, or with Primer3 (Whitehead Institute, http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primer sequences are shown in Appendix A. DNA was dried onto plastic PCR plates by incubating at 80°C for 15 minutes. A standard 10x PCR buffer (10x polymerase mix) was used in conjunction with detergent W1, Taq and $MgCl_2$ to create a 1x Master Mix, which was then added directly to dried DNA in plastic PCR plates. PCR plates were briefly centrifuged to remove air bubbles, and then placed on a TETRAD DNA Engine thermal cycler. Programs varied depending on experiment, but were based on a 2 minute denaturation at 94°C followed by 20 to 35 cycles of three 30 second steps: denaturation at 94°C, annealing at optimal anneal temperature and extension at 72°C. A final 10 minute extension at 72°C was added.

2.2.1.1 SSCP PCR

Primers for SSCP were designed manually to give overlapping amplicons of 150-200bp in length (Appendix A). Primers were optimised using a $MgCl_2$ titration at an annealing temperature of 5°C below the lowest T_m of the two primers. Primer pairs that failed to optimise under these conditions were tested at lower temperatures if no product was seen,

or higher temperatures if non-specific products were seen. If the region was very GC rich, 1.3M Betaine was added to the PCR to improve yield.

Following optimisation, PCR was performed on a set of 40 individuals selected from the Northwick Park Heart Study DNA bank. The 10x PCR buffer was modified to include 0.2mM dCTP and 2mM of the other three dNTPs. 0.2μCi of ^{33}P-dCTP was included per 10μl reaction, thus producing ^{33}P radiolabelled products.

2.2.1.2 DHPLC PCR

Primers for DHPLC were designed using Primer3 (Whitehead Institute). Primers were designed to have a T_m as close to 60°C as possible (Appendix A). Initial testing was then performed using 2mM $MgCl_2$, with a touchdown PCR protocol: 94°C for 2 minutes, followed by 20 cycles of 94°C for 30 seconds, anneal for 30 seconds, 72°C for 30 seconds, with the annealing temperature reducing from 65°C to 55°C by 0.5°C per cycle. This was followed by 20 cycles of 94°C for 30 seconds, 50°C for 30 seconds and 72°C for 30 seconds. A 10 minute extension at 72°C was added, then a touchdown of 35 cycles of 1 minute each with a single temperature reducing from 95°C to 60°C by 1°C per cycle (i.e. cooling by 1°C per minute) to maximise heteroduplex formation. The majority of primer pairs worked efficiently and specifically with these conditions. Any remaining primer pairs were optimised to their individual requirements using a $MgCl_2$ titration and a gradient thermal-cycler to determine optimum annealing temperature and $MgCl_2$ concentrations. Any which were still problematic due to a high GC content were tested with 1.3M betaine in the reaction.

2.2.1.3 Sequencing PCR

Primers for sequencing were selected from those used for DHPLC or SSCP (Appendix A). Pairs were selected that allowed the maximum sequence to be read over the segment of interest. Primers were optimised for annealing temperature and $MgCl_2$ concentration, with special care taken to ensure no non-specific products were detectable. PCRs were performed in 25μl volumes to ensure sufficient quantity for multiple sequencing reactions. Thermal cycling was: 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, anneal for 30 seconds and 72°C for 30 seconds to 1 minute (1 minute per kb, minimum of 30 seconds). A final 72°C extension of 10 minutes was added.

2.2.1.4 Long PCR

Long PCR (Barnes 1994; Cheng *et al.* 1994) was used to generate gene-specific template for some of the genotyping assays. Primers for long PCR were selected using Primer3 (Whitehead Institute). Primers were selected for length (optimum 28nt) and theoretical T_m (optimum 70°C) (Appendix A). A combination of Pwo DNA polymerase (with 3'-5' exonuclease activity) and standard Taq DNA polymerase was used to maximise yield and accuracy. A specific long PCR buffer was used containing ammonium acetate instead of potassium chloride.

2.2.1.5 ARMS PCR

Primers for amplification refractory mutation system (ARMS) PCRs were selected by combination of manual selection (particularly for the allele-specific primers which can only be selected manually), and Primer3 (Whitehead Institute). ARMS primers were designed with additional mismatches at the -3 base (Little 1997) (Appendix A). ARMS primers were initially optimised for temperature and $MgCl_2$ concentration on their own. When an optimal temperature was determined, selections of potential control primers taken from stocks were added, and the reaction checked. Control primers that gave a clear separate band from the allele-specific reaction, without interacting excessively with the ARMS primers were selected to act as controls in the ARMS assay. The detergent W1 was not added to ARMS reactions to avoid any potential reduction in specificity (in house observation). PCRs on bank DNA or long PCR templates were performed in 384-well PCR plates to maximise throughput. Cycling conditions were 94°C for 2 minutes, followed by 35 (genomic DNA) or 17-25 cycles (Long-PCR) of 94°C for 30 seconds, anneal for 30 seconds and 72°C for 30 seconds. A final 72°C step of 10 minutes was added.

2.2.1.6 PCR of RT reactions

Reverse Transcriptase (RT) reactions were diluted 1:2 in RNase-free water (DEPC treated overnight, then autoclaved). Primers were P1d (5' – CCC ACA ACA ACC CTC TTA AAA C – 3') and P5d (5' - GAG AAG GGA GAT GGC GGT A – 3') (designed using Primer3, Whitehead Institute). PCR mix contained 0.4µM each primer, 0.025units/µl Taq DNA polymerase, 0.2mM dNTPs, 1mM $MgCl_2$ and 1x PCR buffer. 2 to 5µl of RT reaction were added to each sample tube (depending on expected yield of cDNA), and then PCR mix

added to a final volume of 15 to 25µl. Cycling conditions were 94°C for 2 minutes, followed by 25-35 cycles of 94°C for 30 seconds, anneal for 30 seconds and 72°C for 1 minute. A final 72°C step of 10 minutes was added. Electrophoresis was performed on H-PAGE or MADGE gels containing 1xT.B.E and 1µgml^{-1} Ethidium bromide at 150V for 20 minutes (section 2.2.2).

2.2.1.7 PCR for $(CA)_n$ repeat genotyping

The CA repeat in the 3' UTR of the *IGF2* gene was genotyped using a method modified from (Rainier *et al.* 1993). Long-PCR products spanning the 3'UTR of the *IGF2* gene (prepared for SNP genotyping: section 2.2.1.4) for the 40 heaviest and 40 lightest samples in Northwick Park Heart Study II were arrayed in one 96-well PCR plate, and amplified with the primers IGF2-DR-AF-FAM (5'-(6-FAM)-ACT TTA TGC ATC CCC GCA GCT ACA-3') and IGF2-DR-BR-HEX (5'-(HEX)-AGA ATC TTA GCG GGA CTT TGG CCT-3') (Rainier *et al.* 1993). $MgCl_2$ concentration was 1.5mM, with 0.02units/µl Taq DNA polymerase and 0.4µM of each primer. PCR cycling was 94°C for 2 minutes, followed by 25 cycles of 94°C for 1 minute and 72°C for 1 minute. A final step of 72°C for 10 minutes was added.

2.2.1.8 Degenerate-oligonucleotide primer PCR

Degenerate oligonucleotide primer (DOP) PCR was performed using 2µl (20ng) of genomic DNA. 50µl reactions contained 5µl 10x Long PCR buffer, 1.25µl 10mM dNTPs, 0.5µl of DOP primer (5' – CCGACTCGAGNNNNNNATGTGG – 3') at 100pmol/µl, 2.5mM $MgCl_2$, 13µl 5M betaine, 0.5µl Gibco Taq (5U/µl), 1µl Pwo polymerase (diluted to 0.02units/µl in water) and 26.25µl water. PCR thermal cycling was performed as follows: one cycle of 94°C for 5 minutes, then 8 cycles of 94°C for 1 minute, 30°C for 1 minute, 72°C for 3 minutes, then 28 cycles of 94°C for 1 minute, 60°C for 1 minute and 72°C for 3 minutes.

2.2.2 MADGE

Open-faced acrylamide MADGE gels were prepared using plastic MADGE formers. Alternatively an "H-PAGE" gel former comprising 6 horizontal rows of 16 2mm^3 teeth was used for certain applications. A 50ml (per gel) acrylamide gel mix was prepared using 1x T.B.E and between 4% and 10% acrylamide (depending on the resolution required). A glass

plate was coated with "sticky silane" by spraying the solution onto the surface of the plate and wiping off the excess. Ammonium persulphate and TEMED were then added to the gel solution, and the gel mixed. The gel was poured immediately into the open-faced MADGE former, and then the glass plate placed on top, silanised side downwards, ensuring that no bubbles were introduced. The gel was allowed to polymerise for 10 minutes, then removed from the former (attached to the glass plate). The gel was stained before use in a solution of 1x T.B.E containing 1μgml^{-1} Ethidium bromide for 20 minutes. Gels were stored until use at 4°C.

Agarose gels were prepared using the same formers, but without the glass plate. Agarose gel mix was prepared by melting 0.7g agarose in 100ml 1x T.B.E (for a 0.7% gel) in a microwave oven. The mix was allowed to cool to about 50°C, and then poured into the MADGE former. GelBond film was then floated on the agarose gel mix (hydrophilic side down). The gel was allowed to set for 30 minutes, and then removed from the former attached to the GelBond. The gel was stained in Ethidium Bromide (1μgml^{-1} in 1x T.B.E) before use, or by adding Ethidium Bromide at 1μgml^{-1} to the gel running buffer.

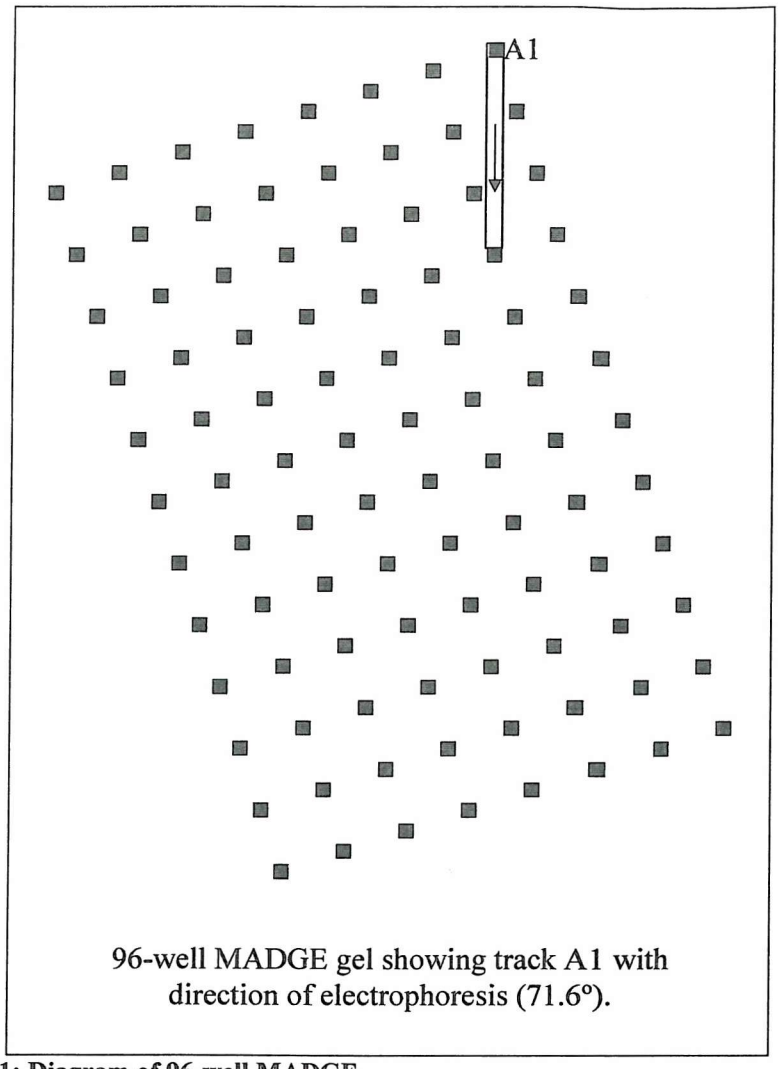
Formamide loading dye was added to PCR products, and then mixed and loaded onto the gel with a multi-channel pipette or a 96-pin passive replicator (dry gel system only).

Agarose and acrylamide gels were both electrophoresed in standard horizontal electrophoresis tanks, using a running buffer of 1x T.B.E., at 150V. Alternatively, acrylamide gels were run in a dry electrophoresis apparatus with platinum electrodes making direct contact with the gel (with no running buffer) at 150V. Typical electrophoresis times were 10 to 30 minutes depending on gel matrix type and product size. PCR products of around 200 to 400 bp resolve on a 5% acrylamide gel in 15 minutes.

Following electrophoresis gels were visualised using a Fluorimager 595 with 514nm excitation wavelength and 610RG long pass emission filter.

2.2.2.1 96-well MADGE

The standard MADGE system provides 96 2mm^3 wells arrayed in an 8x12 microplate format with 9mm well-to-well spacing (Day & Humphries 1994). The array is rotated to 71.6° to create a 26mm track length. The wells take up to 5μl of sample, and can be loaded with a standard multi-channel pipette. Figure 2-1 shows the layout of a standard MADGE former. 96-well MADGE was used for analysing all small-scale PCRs and PCR optimisations.



96-well MADGE gel showing track A1 with direction of electrophoresis (71.6°).

Figure 2-1: Diagram of 96-well MADGE

2.2.2.2 192 and 384-well MADGE

Higher-density variants of the MADGE system allow for higher throughput electrophoresis of large-scale genotyping projects. 192-well MADGE gels (O'Dell *et al.* 2000) integrate two 96-well microplate arrays in such a way that the corresponding wells from each array are vertically aligned (i.e. track A1 of the second array is positioned at the end of track A1 of the first array) (Figure 2-2). 384-well MADGE gels (Gaunt *et al.* 2000) are designed along the same principle, with four arrays, all aligned vertically (Figure 2-2). In the 384-well format the angle of the array is greater (78.7°), with 1.5mm^3 wells, enabling a still adequate 10mm track length.

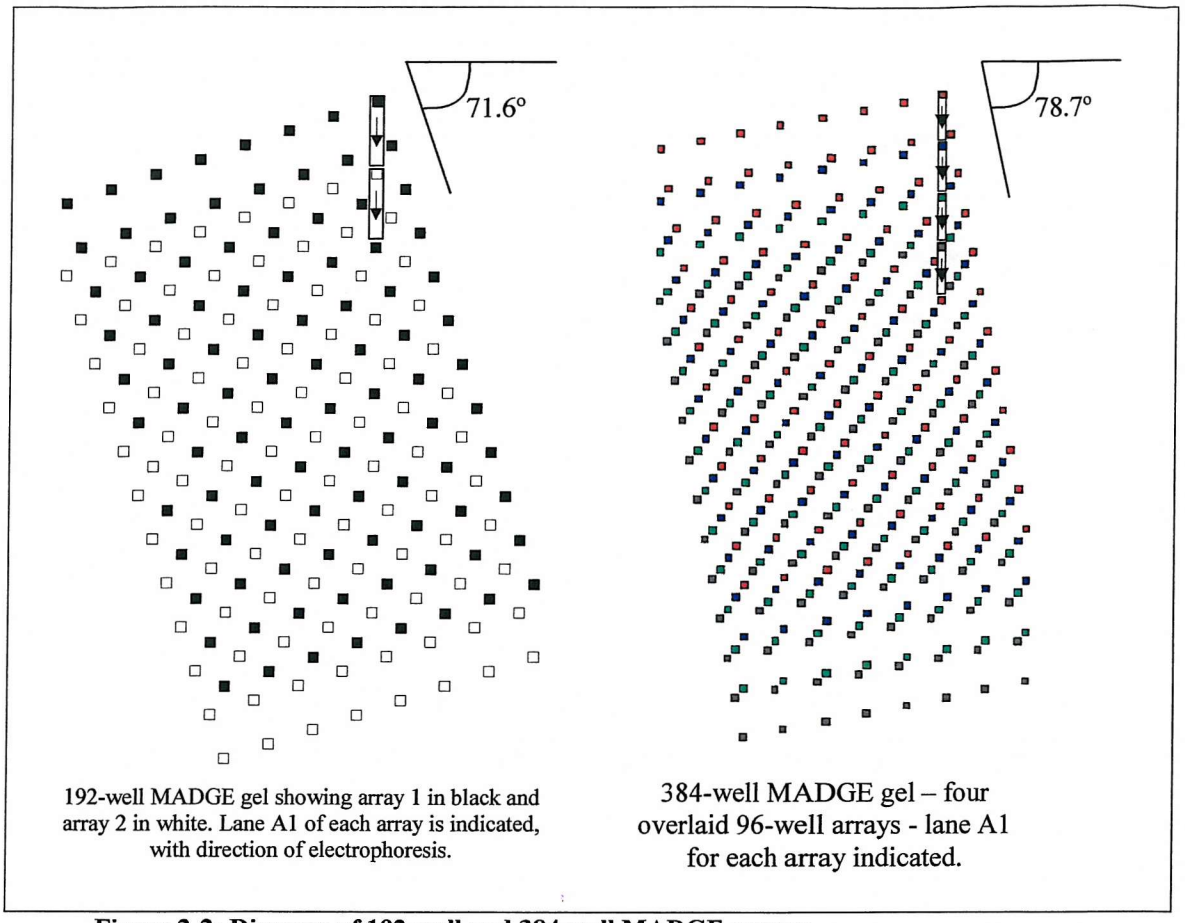


Figure 2-2: Diagram of 192-well and 384-well MADGE arrays

384-well gels were loaded using a 96-pin passive replicator. The split-pins on this device hold approximately 2µl of sample. Two loading actions per 96-well array were performed, and a coloured template was used underneath the gel to ensure correct location of the pins. Gels loaded in this manner cannot be electrophoresed in a submerged system, and so a semi-dry electrophoresis system was used in which electrodes make direct contact with the surface of the gel. The relatively short (8 to 10 minute) runs ensured that the buffer within the gel was not exhausted during electrophoresis.

Analysis of 192-well and 384-well MADGE images was performed using Phoretix 1D Advanced software (O'Dell *et al.* 2000). This allowed pairs of corresponding lanes from two arrays to be analysed as one concatenated lane. Two approaches were used for genotyping with Phoretix. The first was by direct manual calling. For every lane of every gel a genotype was assigned by keystroke. ARMS™ assays using a pair of reactions on each sample were thus rapidly analysed, with genotype calls exported directly to spreadsheet along with well identifier. The second approach was to use the software to measure band intensity for each of the potential four bands in a concatenated pair of lanes. These data were then exported to a spreadsheet containing formulae and scatter plots to allow immediate automatic

assignment of genotype according to the ratio of allele-specific bands to control bands. This allowed very high throughput sample analysis with minimal risk of sample identification error.

Genotype calls were independently verified to ensure accuracy, and negative controls included to ensure no contamination.

2.2.3 SSCP

SSCP loading buffer (7µl) was added to 3µl ^{33}P-labelled PCR product. The samples were denatured at 95°C for 5 minutes, and then cooled on ice. Samples were loaded on two gels, one containing glycerol in the gel and running buffer, and the other not. The gel containing glycerol was run at 2000 Volts, 25°C for 3 hours. The gel with no glycerol was run at 650 Volts, 4°C for 16 hours.

Gels were vacuum-dried onto 3MM paper, and then placed in a cassette with autoradiography film for 48 to 72 hours. The film was then removed from the cassette and developed using a Fujifilm developer.

Autoradiography films were analysed manually on a light-box. For each set of run conditions all the samples were compared to each other. Samples that differed significantly in the number or pattern of bands were determined to be polymorphic. SSCP depends on the distinct resolution of single-stranded DNA of different sequence content, in which alternative conformations cause differences in the rate of migration through a gel. Thus homozygotes would normally show two bands and heterozygotes would show four (although some may co-run and be indistinguishable). Homozygous mutants can be distinguished from homozygous wild-types. Polymorphic samples were selected for sequencing to identify the exact position and nature of the polymorphism.

2.2.4 Denaturing High Performance Liquid Chromatography (DHPLC)

2.2.4.1 Experimental considerations for DHPLC

Prior to analysis of a particular fragment, optimal temperature for analysis was determined by using software to predict the temperature ("DHPLC melt", Stanford University, http://insertion.stanford.edu, or "Wavemaker", Transgenomic). Test samples were

then run at the predicted temperatures and 1°C above and below those temperatures. The optimal temperature for analysing each fragment was determined as the highest temperature at which the sample could be run without the product eluting earlier than at lower temperatures (i.e. raising the temperature by 1°C would cause significantly earlier elution).

Some fragments contained multiple predicted melting domains that melted at different temperatures. For these fragments separate analyses for each melting domain are necessary. The temperatures used for analysis of lower melting temperature domains were those predicted by the software (raising the temperature by 1°C for these domains would not usually result in earlier elution). If two melting domains differed by less than 3°C, the samples were only run at the higher temperature. DHPLC is based on the principal that close to the melting temperature of a duplex the duplex is less helical if there is a mismatch. A duplex with a mismatch will then act like a shorter fragment and elute from a dsDNA column at a lower acetonitrile concentration than a full duplex of the same length.

2.2.4.2 Operation of the DHPLC

The DHPLC was programmed to run all the samples at the temperatures determined to be optimal for that particular fragment. PCR products for DHPLC were placed in the autosampler of the DHPLC system. The instrument automatically loaded the samples, and then applied a gradient of acetonitrile calculated by the software to result in maximal elution differential for heteroduplexes and homoduplexes in a fragment of a particular length and theoretical T_m at a particular temperature. Gradients were applied over a time period of 7 minutes. Samples were detected using a UV detector to measure the absorbency at 260nm of the eluate. Presence of DNA in the eluate resulted in higher absorbency values, represented as a peak on a graph of time vs. absorbency. Figure 2-3 shows a schematic of the DHPLC procedure.

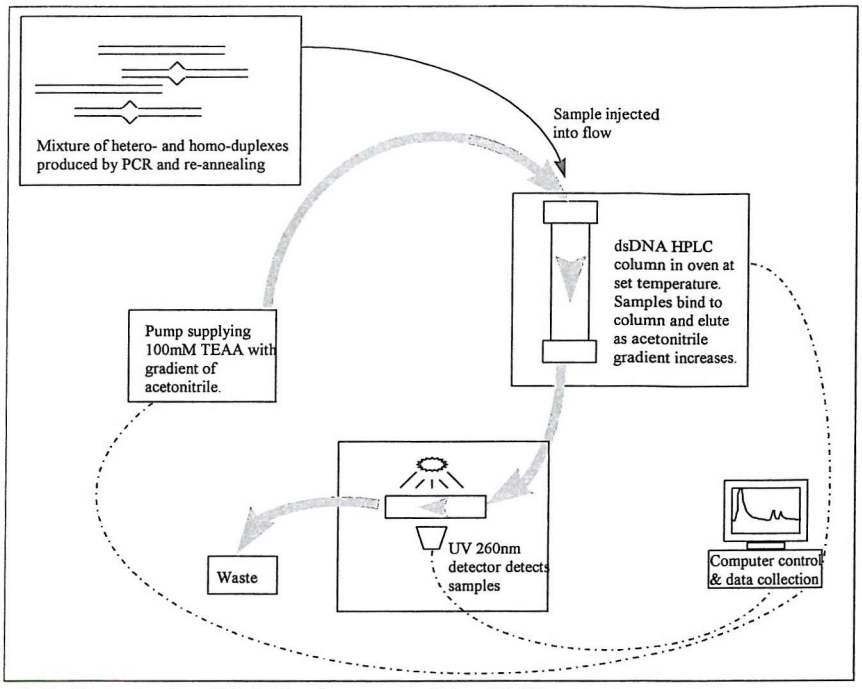


Figure 2-3: Denaturing High Performance Liquid Chromatography

2.2.4.3 Analysis of DHPLC data

Following DHPLC of a set of samples, the graphs for all samples were compared. Standard homozygous samples showed a large primer peak at 0.5 to 1 minutes. A second smaller peak was seen at around 3 to 5 minutes (depending on the acetonitrile gradient). Heterozygotes were identified by the presence of at least one additional peak eluting before the main homoduplex peak seen in homozygotes. The elution of heteroduplex peaks was typically around 0.5 minutes earlier than homoduplex peaks. Homozygous mutants were not distinguished from homozygous wild-types.

The position and nature of the mutations identified by this approach were not ascertained, although in fragments with multiple melting domains it was sometimes possible to identify which domain contained the mutation. Heterozygous samples were tested by sequencing to identify the polymorphism.

2.2.5 Sequencing

PCR products for sequencing were first treated with shrimp alkaline phosphatase and exonuclease I to deactivate PCR dNTPs and primers. 5µl PCR product (containing approximately 30 to 90ng DNA) was incubated with 2 Units shrimp alkaline phosphatase and

2 Units exonuclease I at 37°C for 45 minutes. The enzymes were deactivated by incubating at 80°C for 15 minutes.

5µl of this treated PCR product was then used for the sequencing reaction. 4µl BigDye Terminator Cycle Sequencing Ready Reaction Mix, 3.2 picomoles of one PCR primer and 4µl of a sequencing buffer (200mM Tris pH 9.0, 5mM $MgCl_2$) were added to each sample, and the volume made up to 20µl. The samples were then placed in a PTC-225 TETRAD DNA Engine Thermal cycler. Cycling conditions were 25 cycles of 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes, with 1°Csec^{-1} ramp times.

Sequencing samples were ethanol precipitated using 2µl 3M sodium acetate pH 5.2 and 50µl absolute ethanol for each 20µl reaction. The samples were precipitated on ice for 15 minutes, and pelleted at 21,000xg in a microcentrifuge for 30 minutes. The supernatant was discarded, and the pellet washed with 250µl of 70% ethanol. The samples were then centrifuged at 21,000xg for 5 minutes, and the supernatant carefully removed and discarded. Samples were dried with the tubes open at 95°C for 1 minute.

Sample pellets were resuspended in 25µl Template Suppression Reagent. Samples were transferred to sample tubes for the ABI310 and sealed with rubber septa, then denatured at 95°C for 4 minutes and quenched on ice for 4 minutes. The samples were then loaded into the sample tray of an ABI310 capillary DNA sequencer, and electrophoresed through a 61cm capillary containing POP-6 polymer for 2 hours. Fluorescence was detected by the ABI-310, and the results analysed automatically to determine the sequence. The shortest fragments (primer + dye-labelled ddNTP) pass the detector first, with the dye indicating this first base, followed by successively longer fragments (primer + $(dNTP)_n$ + dye-labelled ddNTP). A coloured "ladder" of fragments is thus detected with order and fluorescence identifying the sequence.

2.2.6 $(CA)_n$ repeat genotyping

After amplification (section 2.2.1.7), PCR products were digested with 2 units of BstUI in a total volume of 10µl for 1 hour at 60°C. After digestion the PCR products were checked on a 5% MADGE gel (section 2.2.2) to ensure successful amplification and confirm product intensity.

Samples were then prepared for analysis on the ABI310 genetic analyser. A loading buffer of 24:1 deionised formamide to Genescan 500-TAMRA marker was prepared. 12µl of

buffer was added to 1µl of each sample in sample tubes and mixed. The tubes were sealed with rubber septa, and heated to 95°C for 5 minutes, then placed on ice for 5 minutes. Tubes were then placed in the ABI310 autosampler, and electrophoresed for 35 minutes per sample through POP-4 polymer in a 50cm capillary using virtual filter set C. Injection time was varied from 5 seconds for high yield samples (as determined from MADGE) to 20 seconds for low yield samples.

Following electrophoresis the samples were analysed and sized using PE Biosystems Genescan software (version 3.1), and Genotyper (version 2.1). Samples were assigned to allele bins on the basis of electrophoretic mobility, with separate sizing for each half of the repeat region (as determined by dye colour).

2.2.7 Quantitative Competitive RT-PCR

2.2.7.1 RNA purification

RNA was extracted from peripheral blood and cultured cells using Promega's "SV Total RNA Isolation System". *For blood*: 5mls of blood were pelleted in a centrifuge at 400x*g* for 5 minutes. Red blood cells were lysed by gentle mixing with 15mls of "SV RNA Red Blood Cell Lysis Solution" in two batches, followed by centrifugation at 400x*g* for 5 minutes and the supernatant was discarded. *For cell lines*: cells were loosened from the flask with 3mls Trypsin-EDTA at 37°C. The cells were then pelleted at 400x*g* for 5 minutes, washed three times with 10 mls phosphate buffered saline and pelleted at 400x*g* for 5 minutes. *For blood cells and cell lines*: the pelleted cells were lysed with 175µl of "SV RNA Lysis Buffer" by pipetting. 350µl of "SV RNA Dilution Buffer" was added, and mixed by inverting 3-4 times. The samples were incubated in a hot block at 70°C for 5 minutes, and then pelleted at 14000x*g* for 10 minutes in a microcentrifuge. The lysate was mixed with 200µl 95% ethanol, and then spun into an SV RNA spin column at 14000x*g* for 1 minute. The eluate was discarded. The column was then washed with 600µl "SV RNA Wash Solution" (containing ethanol) at 14000x*g* for 1 minute. 5µl of DNase I combined with 5µl of 0.09M $MnCl_2$ and 40µl of "Yellow Core Buffer" was added to the column and incubated at room temperature for 15 minutes. 200µl of "SV DNase Stop Solution" was then added, and spun at 14000x*g* for 1 minute. The column was then washed with 600µl "SV RNA Wash Solution" at 14000x*g* for 1 minute, then 250µl of "SV RNA Wash Solution" at 21000x*g* for 2

minutes. RNA was eluted from the column in 100µl of RNase-free water at 14000x*g* for 1 minute.

Finally, RNA from cells, blood cells and prepared RNA standard was treated with Ambion's "DNA-free" kit. 0.1 volume of 10x DNase I Buffer and 3µl (6 units) of DNase I were added to the sample and incubated at 37°C for 2 hours (total RNA samples) or overnight (RNA standard). 0.2 volumes of "DNase Inactivation Reagent" was then added and mixed. Finally, the samples were centrifuged at 10000x*g* for one minute, and the supernatant removed to a clean tube. RNA was stored at –70°C.

2.2.7.2 RNA standard generation

Standard RNA was generated from PCR product containing a T7 recognition site in the upstream primer using T7 RNA polymerase. A mix was prepared containing: 40µl of PCR product, 20µl of 5x T7 buffer, 10µl of 100mM DTT, 2.5µl of 40U/µl RNasin, 20µl of 2.5mM rNTPs, 2.7µl of 15U/µl T7 RNA polymerase and 4.83µl of water. This mix was incubated at 37°C for 2 hours.

Reactions were cleaned by sodium-acetate/ethanol precipitation: 100µl of sample mix plus 11µl of 3M sodium acetate (pH 5.2) and 330µl of 95% ethanol were incubated on ice for 30 minutes, then pelleted at 21000x*g* for 15 minutes. The supernatant was discarded, and the pellet washed with 500µl of 70% ethanol. The sample was then pelleted at 21000x*g* for 2 minutes, the supernatant discarded and the pellet air-dried. Finally, the pellet was resuspended in 100µl of RNase-free water, and treated with Ambion's DNA-free kit (see section 2.2.7.1). RNA was quantified using a Genequant spectrophotometer, and stored at –70°C.

2.2.7.3 RT-PCR

Reverse transcription from total RNA/standard RNA mixes was performed using M-MLV Reverse Transcriptase (RNase H minus). 1µg of total RNA (plus the required volume of RNA standard) was mixed with 500ng of primer 5d (5' – GAG AAG GGA GAT GGC GGT A – 3') in a total volume of 15µl and incubated at 70°C for 5 minutes, then cooled on ice. 10µl of a mix containing 5µl of 5x M-MLV-RT buffer, 1.25µl of 10mM dNTPs, 1µl of M-MLV Reverse Transcriptase (200 units), 0.625µl of Rnasin (25 units) and 2.125µl of water was then added to the RNA/primer mix. Samples were incubated at 42°C for 1 hour, then 95°C for 5 minutes and 80°C for 5 minutes.

PCR of RT products is described in section 2.2.1.6.

2.2.7.4 *ApaI* digestion of RT-PCR products

RT-PCR reaction products and controls were digested with *ApaI* using the following mix: 1µl of 10x buffer, 0.1µl of BSA (both supplied with enzyme), 1.5µl of ApaI (15 units), 4.4µl of water and 3µl of sample. The mix was incubated at 37°C for 16 hours, then 65°C for 15 minutes.

2.2.8 Statistical Analysis

2.2.8.1 Test of Hardy-Weinberg Equilibrium

Genotyping results were tested for Hardy-Weinberg equilibrium using a χ^2 test. Observed values of genotype numbers were compared to expected numbers, and non-significant p-values were used as an indication that genotyping was reliable. Expected values for genotype frequencies were determined by deriving allele frequencies from the observed genotypes:

$$p = \frac{n_{11} + \frac{1}{2}n_{12}}{n_{total}} \quad \text{and} \quad q = \frac{n_{22} + \frac{1}{2}n_{12}}{n_{total}}$$

where n_{11}, n_{12}, and n_{22} were the numbers of each genotype, n_{total} was the total number of samples, p was the frequency of allele 1 and q was the frequency of allele 2. Hardy-Weinberg equilibrium requires:

$$p^2 + 2pq + q^2 = 1$$

Expected genotype frequencies were therefore calculated as:

$$P_{11} = p^2 \;,\; P_{12} = 2pq \text{ and } P_{22} = q^2$$

Expected genotype numbers were calculated by multiplying the expected frequency by the number of samples. χ^2 values were calculated as:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Non-significant χ^2 values (1 degree of freedom) indicated that the sample was in Hardy-Weinberg equilibrium, significant values indicated that the sample was in disequilibrium. This did not distinguish between a sample being in disequilibrium due to selective genotyping, genotyping errors or population disequilibrium.

All χ^2 tests were performed in a spreadsheet (Microsoft Excel) using the following functions: *countif(cell-range, "=genotype")* to count each genotype, and *chidist(χ^2, 1)* to determine the *P*-value (1 degree of freedom).

2.2.8.2 Test of genotype-phenotype association

2.2.8.2.1 Northwick Park Heart Study II

Genotype data was placed in a spreadsheet with phenotype data including Body Mass Index (BMI), weight and height. Association between genotype and each of these phenotypes was tested separately. One-way Analysis of Variance (ANOVA) was used to test for a statistically significant difference between the mean phenotype for each genotype group. The null hypothesis is that samples are from normally distributed populations with equal means and variances. ANOVA tests the within-group and between-group variances. If these variances are significantly different it is assumed that samples are from populations with different means. SPSS and Microsoft Excel were both used to perform one-way ANOVA on the genotype data.

To determine the independence of positively associated SNPs a stepwise multi-way ANOVA model was fitted using forward selection. The model was initiated with the most significant SNP, and at each step the most significant SNP was added into the model. Sequential R^2 values indicated the proportion of variance accounted for by each SNP as it entered the model (above that already explained by the previous term in the model), and partial R^2 values indicated the proportion of variance explained by each term above that explained by all other terms in the model. Partial R^2 values therefore indicated the independent significance of the SNPs.

2.2.8.2.2 Hertfordshire cohort

The relationship between each phenotype (weight, height, body mass index, birth weight and weight at one) and each SNP was tested using both one-way analysis of variance (ANOVA) and linear regression. These were performed by the MRC Environmental Epidemiology Unit using the statistical package "STATA" (release 6). ANOVA tests for the difference in mean of a continuously distributed phenotype variable between different genotype groups (2 degrees of freedom). Regression analysis tests for a trend in the means of a continuously distributed variable across the genotype groups which effectively counts

alleles and tests for a relationship between number of alleles and phenotype. Relationship between gender and genotype was tested using cross-tabulation and χ^2 distribution.

2.2.8.3 Calculation of linkage disequilibrium between markers

Pairwise linkage disequilibrium (LD) between SNPs in *IGF2* was calculated using "2LD", a program created by JH Zhao (King's College London - http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.stm), which implements a gene-counting algorithm to estimate haplotype frequencies. To enable pairwise LD analysis of multiple markers in an efficient way a script (Appendix B, Script 1) was used to automate the processes of feeding data for pairs of SNPs to 2LD and extracting the results from the output files. The data were then imported into Microsoft Excel for graphical analysis. 2LD calculates D, D_{max}, D' and |D'| (Lewontin 1964) and variances (Zapata *et al.* 2001):

$$D = p_{AB}p_{ab} - p_{Ab}p_{aB}$$

and

$$D' = D/D_{max} = \begin{cases} \dfrac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{\min(p_A p_b, p_a p_B)} & \text{If } D > 0 \\ \dfrac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{\min(p_A p_B, p_a p_b)} & \text{If } D < 0 \end{cases}$$

where A & a are alleles at locus 1, B & b are alleles at locus 2, p_{AB} is estimated frequency of haplotype AB and p_A is frequency of allele A *etc.* (formulae from (Weiss & Clark 2002)).

2.2.8.4 Haplotype analysis

Multi-locus haplotype estimation used "EH", a program created by Xiaoli Xie and Jurg Ott (Rockefeller University, New York, USA - http://linkage.rockefeller.edu/ott/eh.htm) (Xie & Ott 1993). This program uses an expectation-maximisation (EM) algorithm. Contingency tables were generated using "Conting", a linkage utility program created by Jurg Ott.

2.2.9 Construction of a gene map of *IGF2*

2.2.9.1 Contig construction

IGF2 sequence was obtained from GenBank (http://www.ncbi.nlm.nih.gov/Entrez) . Initially the available sequence consisted of a number of overlapping GenBank sequences with some gaps. During the project a 78505bp contig containing the entire *IGF2* gene (antisense) became available. This was then aligned with the other available, informative *IGF2* sequences using "SeqMan" from the "Lasergene" package (DNAStar Inc., WI, USA). This allowed contig construction and generation of a consensus sequence. The consensus sequence was imported into "MapDraw" (also from the "Lasergene" package") for annotation using the information contained in the other GenBank sequences.

2.2.9.2 Calculation of GC and CpG content of *IGF2*

%GC was calculated using Script 2 (Appendix B) which cycles through a sequence calculating the %GC as:

$$\frac{(G+C)}{(A+T+G+C)} \text{ x100}$$

These data were recorded in an output file which was then imported into Microsoft Excel for graphical presentation.

The observed to expected CpG ratio of the sequence was calculated using Scripts 3 and 4 (Appendix B), which cycle through a sequence calculating CpG ratio as:

$$\text{CpG ratio} = \frac{f(\text{CpG})}{f(\text{C}) \times f(\text{G})}$$

where f(CpG) is the frequency of the dinucleotide 5'-CG-3' and f(C) and f(G) are the frequencies of C and G nucleotides respectively. CpG ratio is calculated in each user-defined window, at each user-defined step. Script 3 was used to calculate the CpG ratio across sequence for graphical plotting (using Microsoft Excel to open the output file and plot the data). Script 4 was used to determine the position of CpG islands (defined as a 200 base pair window of sequence with a GC content of ≥50% and a CpG ratio of ≥0.6 (Gardiner-Garden & Frommer 1987)).

The data from all three scripts were collated with information from other sources to generate a map using the "scatter graph" facility in Microsoft Excel.

2.2.9.3 *In silico* identification of SNPs

Prior to lab-based SNP scanning of *IGF2* an *in silico* approach was attempted to identify potential SNPs in the gene. Three main approaches were attempted:

1. SNP databases were interrogated, including the Cancer Genome Anatomy Project database (http://cgap.nci.nih.gov/), dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp) and the Whitehead Institute SNP database (http://www-genome.wi.mit.edu/snp/human/). Typically SNP sequences were matched to sequence using the BLAST alignment tool (http://www.ncbi.nlm.nih.gov/BLAST/), as SNP databases contained limited location information.
2. GenBank sequences for *IGF2* were aligned using "Gap4" from the "Staden Package" (Medical Research Council, Cambridge, UK http://www.mrc-lmb.cam.ac.uk/pubseq/). Sequences were compared for differences.
3. GenBank sequences were searched against the GenBank database (http://www.ncbi.nlm.nih.gov/entrez) using the BLAST alignment tool (http://www.ncbi.nlm.nih.gov/BLAST/) to identify any potential sequence differences (Forsberg *et al.* 1998).

Chapter 3 : Methods developed

3.1 Preface

This chapter presents a number of methods that were developed and adopted during the course of this project. In each section a method is described with illustrative results where appropriate.

This project involved the genotyping of nearly four thousand DNA samples for 10 SNPs. Two DNA banks were used, and approximately 75,000 PCR reactions and gel tracks were necessary to obtain and analyse the data. This scale was beyond the scope of standard genotyping methods in the context of a PhD project, and required a high-throughput, array-based system with efficient assay design, optimisation and semi-automated analysis. Existing array-based gel technology was improved for higher-throughput genotyping, while software systems for analysis were developed.

The quantity of DNA required also necessitated some methods development. DNA banks are a limited, non-renewable resource, and as such require conservation to maximise the benefit:cost ratio. Techniques to preserve and amplify the DNA banks were developed, and used as the basis for the majority of the genotyping presented here. In conjunction with this a modified ARMS assay protocol was developed for use on long PCR product.

Finally, an *IGF2* mRNA quantitation protocol was developed. A lack of sample availability meant that this could not reach the results stage in the current project. However, the method is established and ready for sample analysis, and so is presented in this section.

Note that while this is a discussion of methods developed in this project, the details of the methods are discussed in Chapter 2 (General Methods). Development of several of these methods involved other people: their work is acknowledged in the preface to this thesis. The complete system of high-throughput genotyping methods was first used in this project, although some components had been used before. High-throughput 384-well MADGE was first used in this project, as was computer-based analysis of these gels. The use of long PCR and DOP-PCR was already established in the laboratory, but was combined with 384-well MADGE here for the first time. Three-primer ARMS assays were based on existing methods, and already in limited use in the laboratory; their use with long PCR and 384-well MADGE to improve design and optimisation efficiency was initiated here. The quantitative-competitive RT-PCR of *IGF2* described here was developed specifically for this project.

3.2 384-well MADGE

3.2.1 Introduction

The standard microplate-array diagonal gel electrophoresis (MADGE) format provides 96 2mm^3 wells in the industry standard microplate format, with track lengths of 26mm (Day & Humphries 1994; Day 1995; Day 2000). This system has several advantages over alternative electrophoresis systems:

- The closed-casting, open-electrophoresis format allows both acrylamide and agarose to be used in horizontal gels
- The microplate format allows multiple (8, 12 or 96) loading for speed and accuracy
- The microplate-based format allows arrays to be handled as one, reducing the labelling required (compared to sample-orientated systems where each sample's location must be identified). MADGE well locations A1 to H12 correspond to the wells in the original sample storage and PCR plates
- The flexibility in matrix type allows rapid low resolution electrophoresis, and slower high resolution electrophoresis in the same system in a range of sizes from less than 100 base pairs up to several thousand base pairs

However, the limitations of this system are experienced when attempting to run very high-throughput genotyping projects using dual-reaction amplification reaction mutation system (ARMS) assays:

- Equipment (electrophoresis tanks) limit the number of reactions that can be run – ideally people-time should limit the number of reactions for maximum efficiency (note that gels can be stacked in one tank – but this is not very convenient)
- In a dual-reaction assay the two reactions must be run on separate gels, resulting in more potential variability, and making analysis more complex and less efficient

The early ARMS assays in this project were carried out using 192-well MADGE, which had been developed and adopted prior to the start of the project (O'Dell *et al.* 2000). This system resolves the issue of running dual-reaction assays, and also helps to improve throughput – only one electrophoresis tank is required per array (when arrays are being assayed in parallel). The reduced track length of 12mm is still adequate for most well designed ARMS assays (easy resolution of a 25% size difference).

3.2.2 High-throughput MADGE developments

In an attempt to further increase the efficiency of electrophoresis the number of wells was increased to 384 (four 96-well arrays), with a different angle of electrophoresis (78.7°) and a 10mm track length from 1.5mm^3 wells (Gaunt *et al.* 2000). This system was first applied in the high-throughput genotyping of Northwick Park Heart Study II (NPHSII) in this project. A semi-dry electrophoresis system and passive replicator loading system had previously been designed, and these were also adopted for this project.

- Gels are cast in the same way as 96-well gels:
- Acrylamide gel mix is poured into a 384-well gel former
- A silanised glass plate is overlaid, and the gel allowed to set.
- The glass plate with gel attached is then prised away from the former.

Gels are stained prior to electrophoresis in electrophoresis buffer (T.B.E.) containing ethidium bromide for about 20 minutes. Excess buffer is then wiped from the surface of the gel using a glass rod, and the gel laid on a horizontal surface with a coloured template (Figure 3-1) underneath.

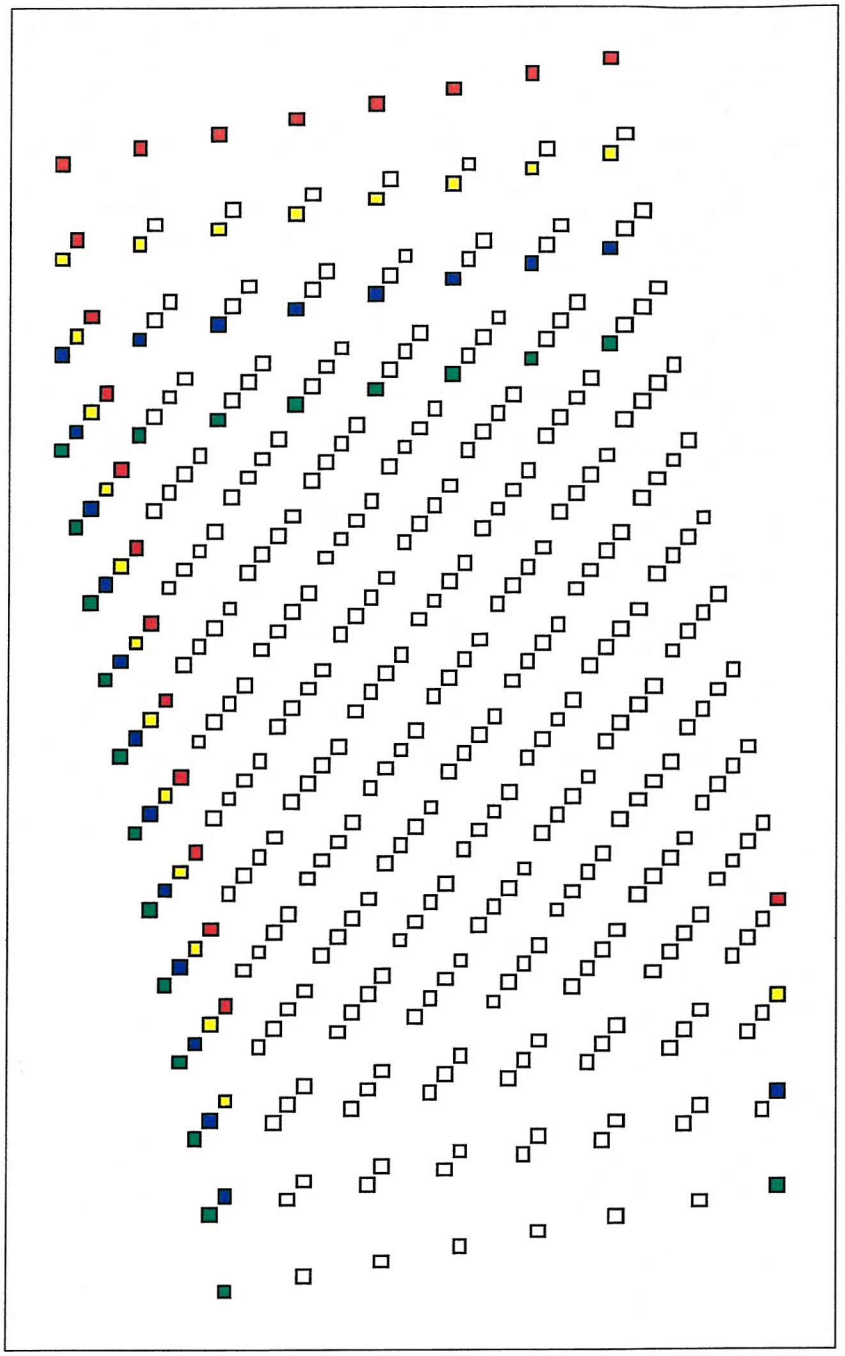


Figure 3-1: MADGE loading template.
A coloured template underneath the 384-well gel allows visualisation of the four arrays for easy location of a 96-pin replicator. Different arrangements of colours suit different people, so several formats exist, and are laminated.

Samples from 384-well PCR plates are loaded onto 384-well gels using a 96-split-pin passive replicator (designed for dot-blot applications). The coloured template assists visualisation of the wells for each array. The normal loading procedure is to locate the pins along one long edge with the replicator tilted towards the operator. The replicator is then lowered so that the remaining 84 wells make contact with the appropriate wells in the gel. A

close-up diagram of one pin is shown in Figure 3-2. These pins hold and transfer approximately 1 to 2µl of sample/dye mix, and often require two loadings for sufficient band intensity. The pins are rinsed in deionised water and dried by blotting between loadings to prevent cross-contamination of sample wells. This system allows rapid loading, and greatly reduces the risk of sample loading errors by rotation or mis-location (as can occur with an 8-channel pipette).

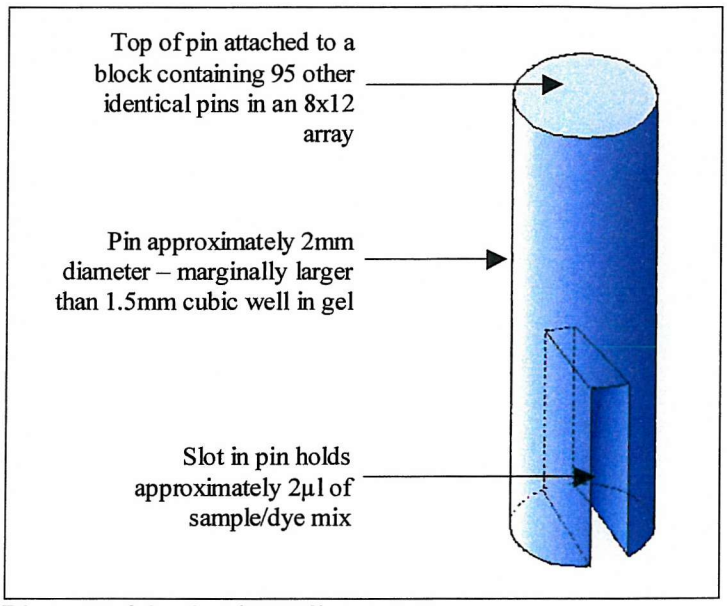


Figure 3-2: Diagram of the tip of a replicator pin.
Surface tension holds the sample in the slot at the end of the pin. Contact with residual buffer in the well of the gel then draws sample out by diffusion. Slot still contains liquid after loading, which is removed by washing before the next loading.
Pin diameter is slightly larger than width of well, but elasticity of polyacrylamide and shape of well allow pin to enter.

Electrophoresis is performed in a semi-dry electrophoresis system (Day & Hamid 2000) (Figure 3-3). This system uses direct electrode contact with the gel matrix (suitable for acrylamide, but not agarose), which allows electrophoresis to occur with no additional buffer. The gel matrix contains 1xT.B.E., which is sufficient for typical run times of 8 to 30 minutes. The major advantages of this system are:

- Loading is performed dry – essential for passive replicator use
- Loading can be performed at the user's bench, ensuring that electrophoresis apparatus is not obstructed by people loading gels
- The system is clean and tidy – no buffer spillage
- The system is modular – electrophoresis boxes are slotted into a tower, and connected to a power pack in parallel

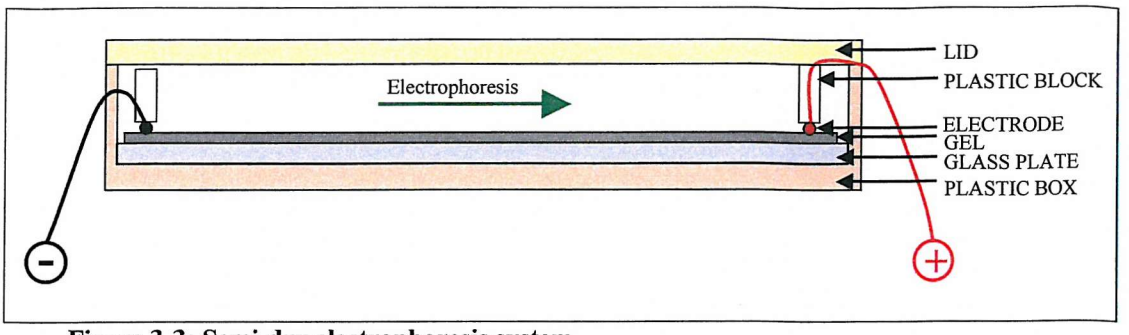


Figure 3-3: Semi-dry electrophoresis system.
Gel is placed inside a box with no additional buffer. Platinum electrodes are pressed into the gel matrix, and a current applied across them.

Following electrophoresis 384-well MADGE gels are scanned using a Molecular Dynamics Fluorimager 595, and images stored in digital format for later analysis. Any type of gel imaging system may be used, but digital systems are more useful as they allow computer-based image analysis.

3.2.3 Conclusions

The 384-well MADGE gel quadruples throughput relative to the original 96-well MADGE. The linear arrangement of arrays allows one sample to be represented in two concatenated lanes where a dual reaction is necessary. This accelerates and simplifies analysis compared to 96-well MADGE. The MADGE system in general allows array-based rather than sample-based assaying, greatly decreasing the risk of error and increasing the efficiency.

Sample loading by passive replicator is consistent with the array-based approach to sample-handling, and allows only two types of error: the (unlikely) 180° rotation error (tested by asymmetrically positioned negative controls) and mis-positioning of pins (observed by user). By comparison 8-channel pipettes allow up to 11 alternative mis-locations in both sample array and gel in addition to these errors.

The semi-dry electrophoresis system was an essential requirement for passive sample loading on 384-well gels, but is also suitable for any horizontal polyacrylamide gel electrophoresis. The modular system removes the requirement for additional buffers and allows gel-loading at the user's bench. The system can also be used for dry loading and "running-in" of agarose gels prior to submerged electrophoresis, although the high temperatures generated are unsuitable for prolonged agarose gel electrophoresis.

3.3 Computerised analysis of MADGE gel images

96-well MADGE gels can be analysed 'by eye' from a printout of the gel image. However, the development of higher-density gel formats made the development of *in silico* image analysis very desirable. Phoretix International incorporated a MADGE module into their Phoretix 1D Advanced software, which allows the selection of two-dimensional arrays of lanes. There are three limitations to this system:

- 192- and 384-well gels are linearly overlaid arrays of 96-lanes. These cannot be analysed by a standard rectangular matrix
- The software does not detect lanes automatically – this requires user input
- The software is often unsuccessful at band detection

However, the software enables straightforward, accurate manual analysis of high-density gels by aligning labelled lanes in a row for manual analysis of genotype. When band detection is used (requiring some manual correction) band intensity can be used for semi-automated genotyping.

3.3.1 Manual calling

Manual calling is the assignment of genotype by user recognition of band pattern for each sample. For ARMS assays the lanes for two arrays (the allele 1 reaction and the allele 2 reaction for a single original array) are treated as one (O'Dell *et al.* 2000). A 384-well gel therefore has two separate double-arrays, each of which comprises an array of 96 lanes for allele 1 and an array of 96 lanes for allele 2 (Figure 3-4). The gel is analysed twice, once for each double-array. A 96-node grid is drawn over the gel, using the 'Create Lanes' function in Phoretix 1D Advanced, in such a way that each node falls in a well of the first array of one double-array. A lane is then drawn from one node down across the two lanes for one sample using the 'Lane Creation' tool. The software automatically creates the remaining 95 lanes relative to the other 95 nodes.

The image can be adjusted with respect to contrast and brightness to ensure maximum ease of visualisation. The 'Matching' function is then used. The user assigns a genotype by a single key-press for each double-lane. The presence of control bands in both lanes of the double-lane is required for a valid call. The presence of an allele-specific band indicates the presence of that allele, and therefore the genotype. After assignment of genotypes, the data

for the full array is exported to Microsoft Excel for incorporation into a database of all genotypes.

This method is rapid and simple. A double-array representing 96 samples can be genotyped in around 3 minutes. However, the method is susceptible to occasional typing errors or genotype assignment errors caused by failure to check both halves of the double-lane (typically with a low frequency polymorphism a rare homozygote may be mis-called as a heterozygote). The latter problem is usually resolved by a second visual scan of the lanes checking the upper half of the double-lane for single bands, while both problems are addressed by a system of thorough independent checking (re-calling of genotypes and comparison). However, the ideal approach would be an automated system that minimises user error, requiring the user to check only those samples that the software could not resolve.

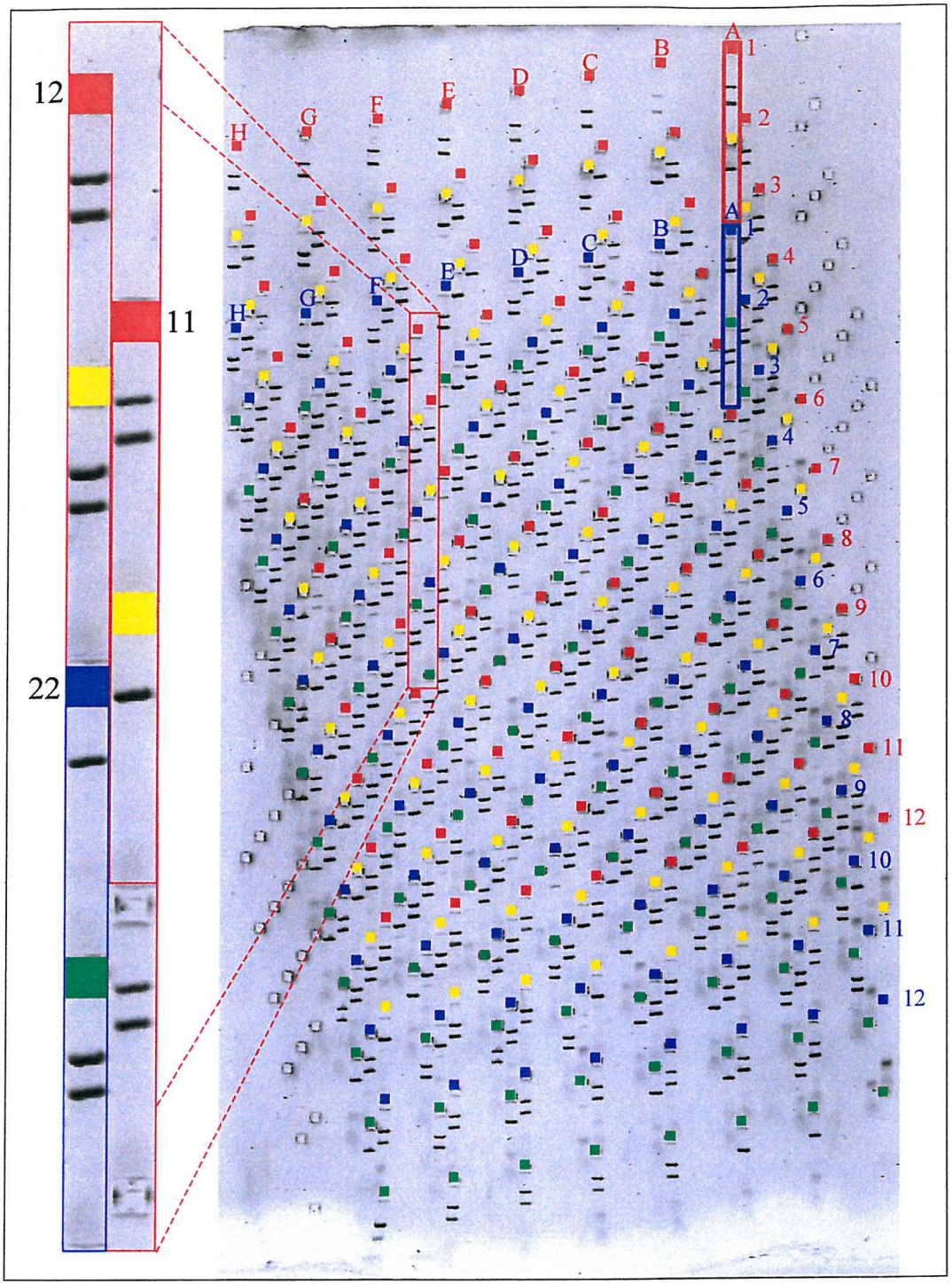


Figure 3-4: 384-well MADGE gel.
Each array of 96 coloured squares identifies the well positions into which one array of 96 samples is loaded. For dual reaction assays the red and yellow arrays represent allele 1 and allele 2 reactions for array 1, while the blue and green arrays represent the corresponding reactions for array 2. The double-lanes for sample A1 for both arrays are indicated by rectangles on the main image. The enlarged section shows each of the three genotypes represented on the gel.

3.3.2 Cluster analysis

The first step towards automation was the development of cluster analysis of genotype based on the ratio of intensity of allele-specific bands. Band intensity is measured after background subtraction (band volume = sum of pixel values for the selected area) using Phoretix 1D Advanced. Band volumes for the four potential bands (including the volume from the position of absent allele-specific bands) are exported from Phoretix to Microsoft Excel, and all subsequent calculations are performed in a custom-designed spreadsheet using Microsoft Visual Basic for Applications (VBA) macros.

The two allele-specific reactions are performed in separate PCR wells, and therefore cannot be compared directly. However, the inclusion of a control reaction enables comparison of allele-specific band intensity to control band intensity, thus giving a measure of amplification success for that allele. By comparing the ratio of allele-specific band intensity to control for one reaction with the other, allele bias can be determined – if this is towards allele 1 (allele 1 homozygote), allele 2 (allele 2 homozygote) or whether amplification efficiency is equal (heterozygous). In reality the difference in 3' primer/template duplex sequence between the two alleles often results in different relative efficiencies of the two allele-specific reactions. The integral control PCR also often differs in efficiency from the allele-specific PCR, so the ideal situation of allele:control ratios of 1 = present and 0 = absent is rarely observed.

The following calculations are used in cluster analysis:

```
VOLUME ALLELE 1      = A1
VOLUME CONTROL 1     = C1
VOLUME ALLELE 2      = A2
VOLUME CONTROL 2     = C2
A1 RATIO             = A1/C1
A2 RATIO             = A2/C2
ALLELE RATIO (AR)    = ((A1/C1)/(A2/C2))
                     = (A1xC2)/(A2xC1)
CLUSTER INTERVAL     = AR(N)/AR(N+1)
CLUSTER BOUNDARIES   = MAXIMUM 2 VALUES FOR CLUSTER INTERVAL
```

The calculation of cluster boundaries is critical to this method. Manual determination can be used, and a spreadsheet has been set up which enables the user to select cluster boundaries with scroll buttons (not shown). However, an automated method is preferable because it is less arbitrary and requires less user time. Fairly sophisticated clustering methods exist, and k-means clustering, neural nets and decision trees have all been tested with SNP genotype data in the context of the TaqMan genotyping system (Heil *et al.* 2002). This group, however, found that these methods were not ideally suited to genotyping, and developed a

novel algorithm which utilised genetic qualities of the data to enhance accuracy – this method was not available in detail as Celera Genomics are currently seeking a patent to cover it (Heil *et al.* 2002).

For the purposes of this project a very simple clustering algorithm was used to demonstrate the functionality of this approach. The method uses allele ratio (AR), which is the single-value measure of relative intensities of the two bands. If genotypes are distinct (and PCR failures excluded from analysis) then there are three groups of AR values with a gap between each group. By ranking the AR values for an array in increasing order, and then calculating the intervals between successive AR values (cluster intervals) these large gaps (cluster boundaries) are identified. Figure 3-5 shows the ranked AR values (red line) and corresponding cluster intervals (blue line) for a real array. Two distinct peaks are observed – these indicate the boundaries between genotype clusters.

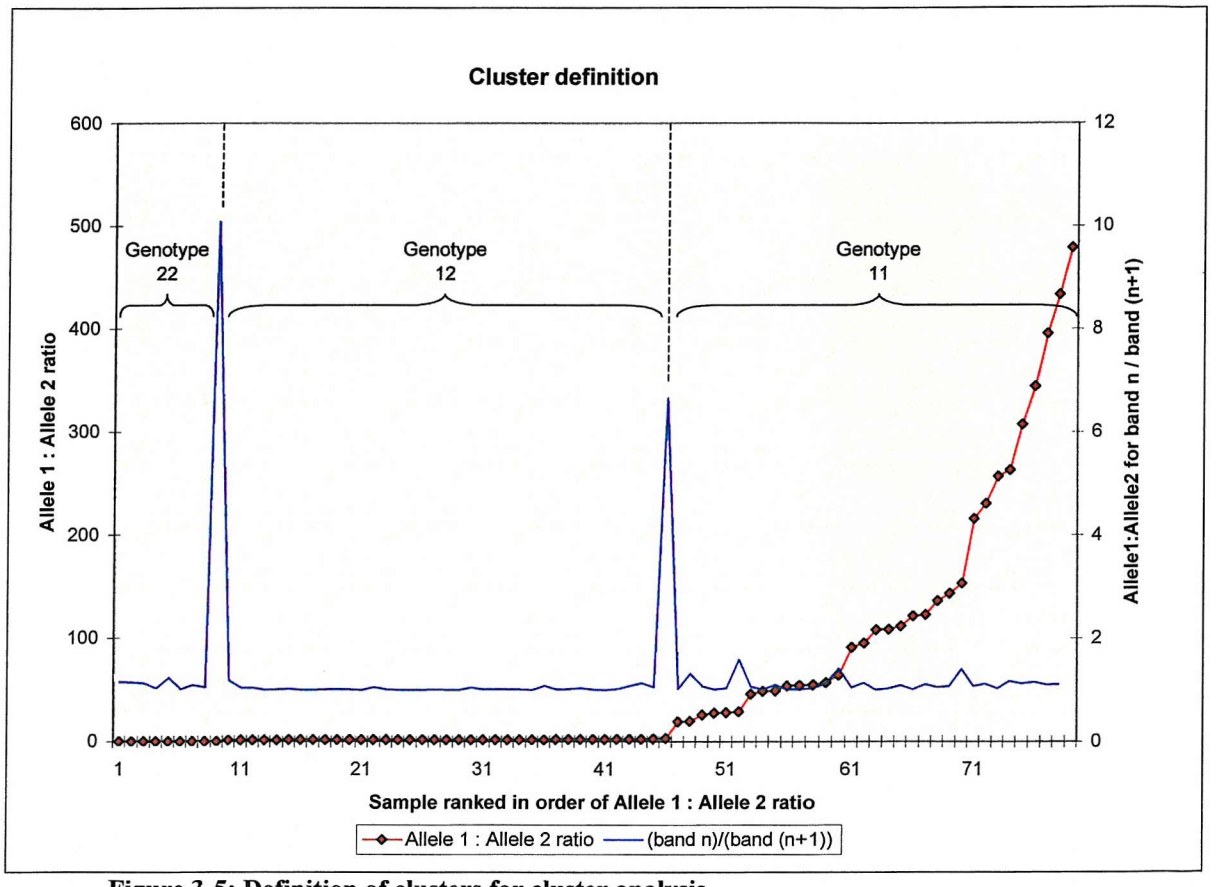


Figure 3-5: Definition of clusters for cluster analysis.
The ratio (A1*C2)/(A2*C1) (= (A1/C1)/(A2/C2)) gives a single value for each sample (allele ratio). This is plotted in numerical order as the red line with red data-points. By dividing each allele ratio by the next allele ratio in the sequence a cluster interval can be determined (distance between points) – this is shown as a blue line. The cluster boundaries occur where the two maxima are indicated. The allele ratios of the data points either side of this boundary define the value a in the line y = ax, which defines the edge of the cluster.

Once cluster boundaries are determined and defined, these values are used to assign genotype in the following way:

BOUNDARY B_{22} = AR_X WHERE X IS SAMPLE TO IMMEDIATE LEFT OF PEAK 1
BOUNDARY B_{12} = AR_Y WHERE Y IS SAMPLE TO IMMEDIATE LEFT OF PEAK 2
FOR SAMPLE *N*, GENOTYPE = 11 **IF** $AR_N > AR_Y$
12 **IF** $AR_X < AR_N \leq AR_Y$
22 **IF** $AR_N \leq AR_X$

The spreadsheet assigns genotypes according to the formulae above, and these can then be transferred to a database. A scatter plot of A1 ratio versus A2 ratio is created in Excel to allow user-visualisation of the clusters, to ensure that these are distinct and genotypes have been called correctly a scatter plot of A1 ratio versus A2 ratio is created in Excel (Figure 3-6).

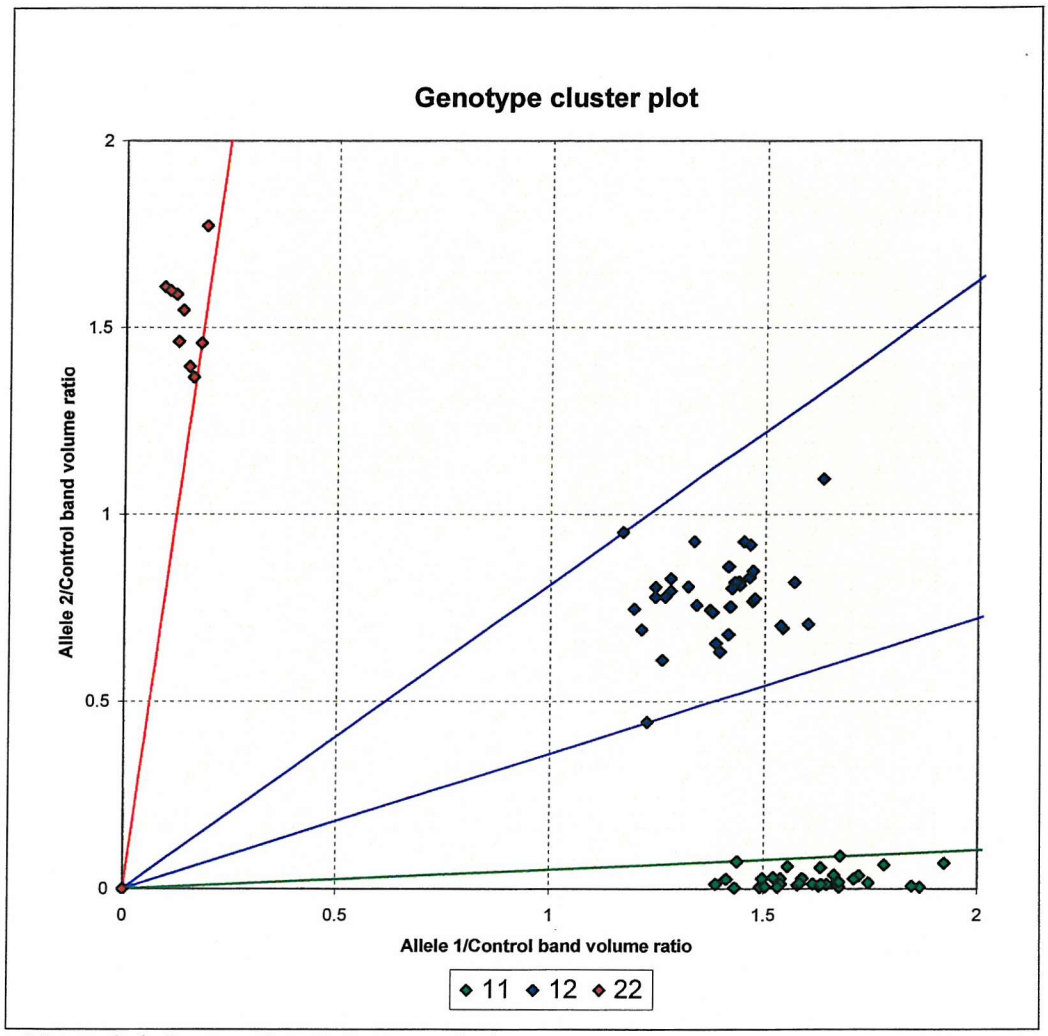


Figure 3-6: Scatter plot of cluster analysis.
Boundary lines are drawn in for clarity. Each point represents one sample. Genotypes are shown in different colours.

Using this method the only user input required is during the initial band detection phase, and in verifying that the scatter plot is correct. Band detection is a complex software

issue, and will hopefully be improved in future software versions. However, with the current version of Phoretix this method is too labour-intensive to be used routinely. Typically one 96-lane array will take 10 to 15 minutes to analyse compared to about 3 minutes using the manual calling method. This time is >95% band detection and <5% spreadsheet analysis of the band volumes, so with improvement in band detection it would be reasonable to expect 96-well array analysis to take under 30 seconds with virtually no user input.

This clustering algorithm requires relatively high quality data, with discrete, tight clusters and no intermediate samples. Also, PCR failures are excluded at the band detection stage, although this could be incorporated into the spreadsheet if band detection were automated. A more sophisticated algorithm would ideally allow samples to fall between clusters, and categorise these as "unknowns". Also, a useful feature would be the automated identification of poor quality arrays for manual checking – this is the approach used by Celera Genomics with their automated genotyping software, which achieves >99% accuracy on good arrays if poor quality arrays are excluded (Heil *et al.* 2002).

Another limitation of this clustering algorithm (and others (Heil *et al.* 2002)) is the problem of differing numbers of cluster. If a SNP with a rare allele is genotyped, then the rare homozygote may occur in some arrays but not others. Therefore some arrays will have four clusters (PCR failure, common homozygote, heterozygote and rare homozygote), while others will have only three.

This method has several limitations, but the potential to improve genotyping speed and efficiency makes further development important. The use of controls to normalise allele-specific band intensity data from polyacrylamide gels works effectively to provide scatter plots similar to those produced by assays where allele intensities are compared directly in liquid-phase (Mein *et al.* 2000; Heil *et al.* 2002).

The method presented here will convert band intensity data to genotype data with minimal user input. It provides a graphical representation of the data in a two-dimensional scatter plot along with a statistical verification of Hardy-Weinberg equilibrium.

3.4 Use of long PCR and DOP-PCR templates

Another potential problem with high-throughput genotyping is the excessive use of DNA resources. At a fairly conservative 2x10ng per ARMS assay, the typical 100 to 500µg of DNA obtained from a 10ml blood sample would last for 5000 to 25000 assays assuming

100% efficiency. However, wastage, the requirement to use more DNA for some assays, the lower yield for some samples, the requirement to share with other laboratories and extraction from smaller samples or buccal cells drastically reduce the number of assays that can be performed. The most effective approach to maximising these resources is to amplify the DNA bank using PCR. Two main methods exist: whole genome amplification using degenerate oligonucleotide primers (DOP-PCR, (Cheung & Nelson 1996)) and long PCR (Barnes 1994). The former is a non-specific amplification representing much of the genome that is suitable for many uses, while the latter is a specific amplification of one region of the genome that is particularly well suited to projects studying one region in depth (Figure 3-7). Long PCR creates sufficient product for at least 500 reactions per 10µl PCR from 20ng genomic DNA. DOP-PCR creates sufficient product for 25 reactions per 50µl PCR from 20ng genomic DNA.

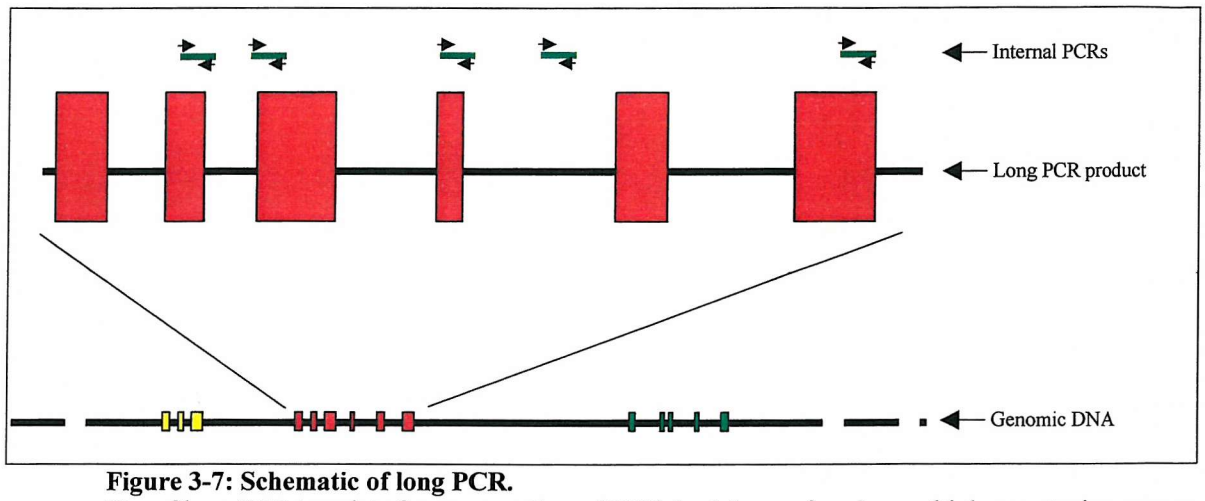


Figure 3-7: Schematic of long PCR.
Use of long PCR template for conservation of DNA bank by performing multiple genotyping assays on the product of a single PCR reaction from the original bank. Typically a gene or part of a gene (up to 10kb) is amplified, then diluted 1/100. 2µl is used as template for subsequent internal (100 to 500bp) reactions.

A method for long PCR and its use in genotyping was already established in our laboratory (Gu *et al.* 2000), and this was adopted for the generation of template for high-throughput genotyping in this project. Long PCR product was diluted and aliquoted using a Robbins Hydra 96-channel robotic pipettor. Aliquots were dried on 384-well PCR plates, which were stored at 4°C until used. This system suits the aliquoting of many duplicate plates at one time. Subsequent PCRs were then carried out by adding the appropriate PCR mix to a dried plate and subjecting to thermal cycling. This system greatly improved the speed and efficiency of high-throughput genotyping. In addition to the time saved by batch template aliquoting, the use of long PCR reduced the quantity of mis-priming in subsequent PCRs,

speeding up the assay optimisation time. The purity of long PCR template also reduced the number of PCR cycles required for sufficient amplification from 30 to 20 on average, reducing PCR cycling times by a third.

DOP-PCR (whole genome amplification) (Cheung & Nelson 1996) utilises degenerate oligonucleotide primers to randomly amplify many parts of the human genome. This allowed genotyping of SNPs that were isolated from others in a more efficient manner than by generating a specific long PCR product just for that assay. A potential criticism of DOP-PCR is that Taq polymerase may introduce false mutations during the whole genome amplification that affect the accuracy of genotyping. To minimise this risk we used a modified protocol in which a "proof-reading" polymerase (Pwo) with 3'→5' exonuclease activity was added at a low level to correct any mutations. This is essentially the same principle as that used in long PCR (Barnes 1994) to reduce mutation rates. In addition, both starting genomic DNA and DOP-PCR template were not used at extremely low concentrations, as small copy numbers would result in more variability in the percentage of false mutation copies between samples.

Long PCR template was used for the genotyping of seven SNPs in both Northwick Park Heart Study II (NPHSII) and Hertfordshire DNA banks. Genomic DNA was used to genotype one SNP in Hertfordshire and NPHSII, and DOP-PCR was used as template for the genotyping of two SNPs in Hertfordshire.

Table 3-1 shows the results of conservative approaches to DNA bank use. Using both long PCR and DOP-PCR to generate templates for genotyping PCRs we used approximately 10% of the DNA we would have used without applying these methods. The importance of preserving valuable DNA resources justifies the time taken in generating these templates.

Table 3-1: Use of genomic DNA resources
Total nanograms of DNA used for each phase and the potential that would have been used without conservative use (based on 20-50ng per reaction)
* 20ng generated enough DOP-PCR template for 12 dual-reaction ARMS assays – only two were needed and the rest was available for other users

DNA bank	long PCR	DOP-PCR	Direct genotyping	Total used	Potential use
Northwick Park Heart Study II (NPHSII)	40ng	0ng	0ng	40ng	400ng – 1000ng
Hertfordshire	40ng	<20ng*	20ng	<80ng	400ng – 1000ng

Long PCR is ideally suited to projects concentrating intensively on one gene. The greater template purity reduces the quantity of mis-priming in the genotyping assay and reduces the number of PCR cycles required, increasing efficiency. Up to 500 PCRs can be performed on the long PCR product of 20ng of genomic DNA. DOP-PCR is less productive, with 20ng genomic DNA generating enough DOP-PCR template for 25 PCR reactions. However, the whole-genome nature of this template makes it well suited to projects where markers are too far apart for long PCR.

3.5 Three-primer ARMS assays

The conventional amplification refractory mutation system (ARMS) assay (Newton *et al.* 1989) uses a four-primer system, in which there is one reaction for each allele. Each reaction contains an allele-specific primer, a 'common' primer and a pair of independent control primers. This system ensures that a PCR failure is distinguished from an absent allele because if the allele is absent (homozygous for the other allele) then there will still be a control band. If the allele is present, the allele-specific band (and control band) will also be present. If there is no band, then the PCR has failed and the genotype cannot be determined. The control primers are usually sited well away from the allele-specific primers, and usually generate a larger product.

The use of long PCR as template for ARMS assays introduced a problem for this system. Control primers must also be sited within the long PCR product, and sufficiently far away from the allele-specific and common primers to prevent any cross-reaction. This limits the available choice of primers and makes ARMS assay design more difficult. For this project a three-primer ARMS assay was used, based on the tetraprimer method (Ye *et al.* 1992; Ye *et al.* 2001) which uses two oppositely orientated allele-specific primers in a single reaction with a pair of external control primers to generate a large control product, and two allele-specific products of different sizes. The tetra-primer method is difficult to optimise because of the difference in sequence of the two allele-specific primers, and may not be suitable where the sequence to one side of the SNP is of low complexity or high GC content. The three-primer method (Figure 3-8) uses an allele-specific primer (which is orientated depending on the sequence context of the SNP), and a pair of outer control primers. Two reactions are performed (one for each allele). This system is easier to design and optimise than either of the other two, and is suited to long PCR template. It was therefore the most

appropriate choice for the high-throughput genotyping part of this project where multiple genotyping assays had to be developed rapidly, especially as conservative use of DNA banks (section 3.4) reduced the advantage of tetra-primer requiring only a single reaction.

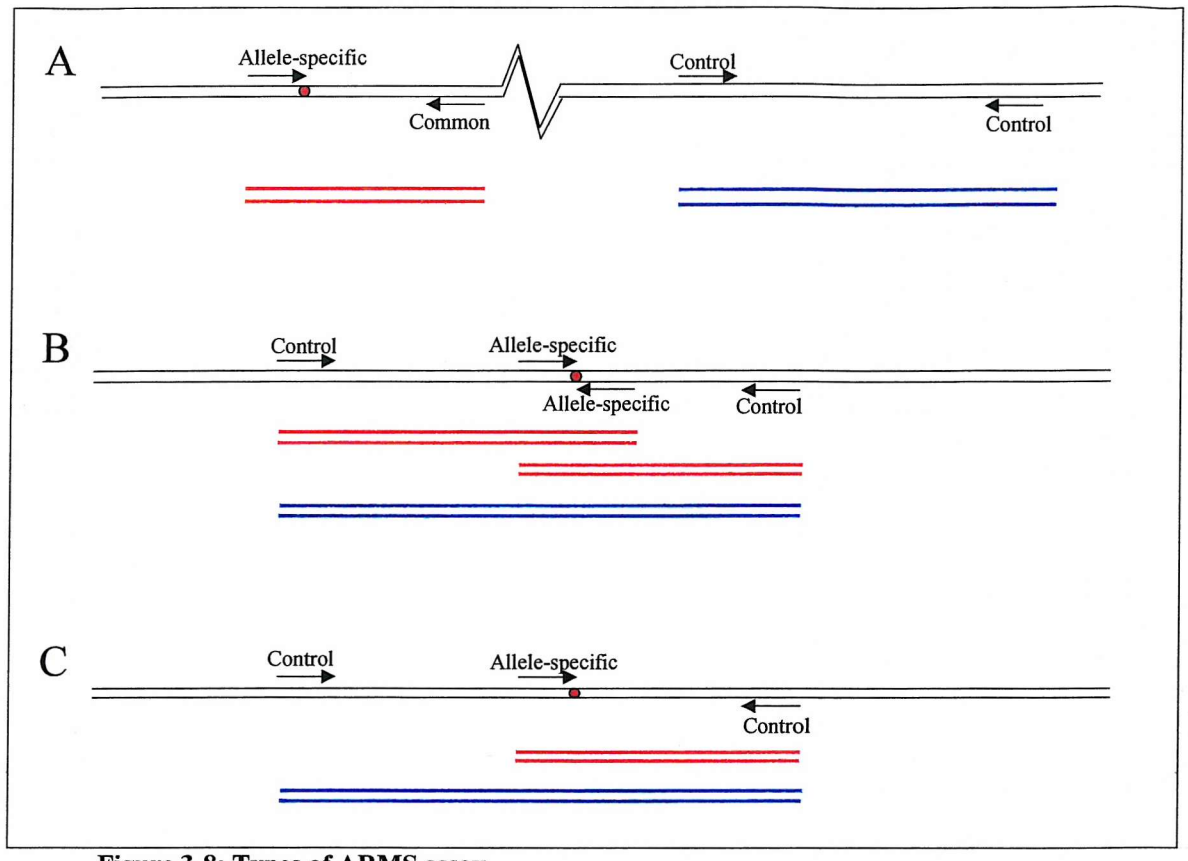


Figure 3-8: Types of ARMS assay.
Templates are black lines (broken to show unspecified distance in A), primers are black arrow, allele-specific products are red lines, and control products are blue lines. (A) the conventional ARMS assay is a duplex reaction with an independent control (Newton *et al.* 1989). Two separate reactions are performed, one for each allele. (B) The tetraprimer ARMS assay combines the two reactions and control to generate a long control band in all reactions, and either one or both of the two allele-specific bands (Ye *et al.* 1992; Ye *et al.* 2001). (C) The three-primer ARMS assay is based on tetraprimer, but with two separate reactions, one for each allele.

3.6 Quantitative-competitive RT-PCR

3.6.1 Introduction

Quantitation of RNA levels is important in determining the activity of a gene at different stages of development, in different tissues and under different environmental, genetic and biological influences. In the context of this project quantitative RT-PCR is necessary to allow us to determine whether polymorphisms in *IGF2* influence gene

transcription or protein function. Different methods for RNA quantitation exist, based around the reverse-transcription polymerase chain reaction (RT-PCR) procedure. Two main categories exist: real-time RT-PCR and end-point RT-PCR.

Real-time RT-PCR utilises fluorescent detection of PCR product during the polymerase chain reaction to quantify the product during the logarithmic phase of amplification, when product quantity most accurately reflects initial template quantity (Heid *et al.* 1996). Comparison to known standards enables highly accurate quantitation, but it is an expensive and relatively low-throughput method.

End-point RT-PCR quantifies initial template on the basis of final product quantity, as compared to a known standard. Non-competitive RT-PCRs use co-amplification of a standard house-keeping gene to allow comparison of final quantity and determination of initial concentration – this method is not very accurate because of variation in the sequences and primers of the two products, and potentially different expression between individuals (Bustin 2000). Quantitative-competitive reverse-transcription polymerase chain reaction (QC-RT-PCR) (Wang *et al.* 1989) is more accurate than non-competitive RT-PCR. This method utilises an internal standard that matches the test sequence, but is either shorter, longer or contains a restriction site to allow differentiation from the test sequence (Bustin 2000). Because the sequence and primer binding sites match, the amplification efficiency of standard and test samples are very similar, resulting in better accuracy than non-competitive RT-PCR (Bustin 2000).

A standard can be created by the use of an internal primer with a 5' tail that matches one of the outer primers. When used with an outer primer with a 5' T7 tail this generates a shorter PCR product which still has primer-binding sites for the two external primers (Zhang & Byrne 1997). This PCR product is then transcribed to RNA using T7 RNA polymerase. Although competitor could be added as DNA at the PCR stage, ideally it should be added as RNA at the RT stage (Bustin 2000). Additional improvements in accuracy of QC-RT-PCR can be made by the use of titration of standard, or by variation of PCR cycle numbers (Zhang *et al.* 1997).

3.6.2 Quantitative-competitive RT-PCR of *IGF2*

3.6.2.1 QC-RT-PCR assay design

A QC-RT-PCR was designed to incorporate the polymorphic *ApaI* restriction site in the 3'UTR of *IGF2* to enable the possibility of differentiating between the expression levels of the two alleles. This is strategy B in Figure 3-9. There were several alternative strategies available, but this one was selected for its ability to distinguish between alleles of the *ApaI* restriction site polymorphism. This polymorphism was of interest because genotype at this locus is significantly associated with body mass index in middle-aged men (O'Dell *et al.* 1997).

Four primers were designed: P1d and P5d are shown in Figure 3-9, with P1d (sense) upstream of the *ApaI* site and P5d (antisense) downstream. Primer P3 (antisense) was sited immediately upstream of the *ApaI* site such that a P1d→P3d PCR product would not contain that site, but a P1d→P5d PCR product would. Primer P3d also had a 5' tail matching the sequence of P5d. A fourth primer matched P1d, but had a T7 tail. The sequences of these primers were (5' tails underlined):

```
P1d         5'-CCCACAACAACCCTCTTAAAAC-3'
P1d-T7      5'-TAATACGACTCACTATAGGGCCCACAACAACCCTCTTAAAAC-3'
P3d         5'-GAGAAGGGAGATGGCGGTAGGGAGTGCCAAATTCCTTTA-3'
P5d         5'-GAGAAGGGAGATGGCGGTA-3'
```

The design of this assay is summarised in Figure 3-10. A PCR with primers P1d-T7 and P3d generated a PCR product which was transcribed into RNA using T7 RNA polymerase. The resulting RNA, after removal of DNA with DNase, matched the sequence of a P1d→P5d PCR product for 192 nucleotides, but did not include 115 nucleotides of sequence immediately adjacent to the P5d binding site. This could then be distinguished from *IGF2* mRNA by length. The *ApaI* restriction site was positioned such that digestion of a sample containing that restriction site (common homozygote or heterozygote) would produce products of 257bp and 51bp, also distinct in size from both the standard and the full-length product.

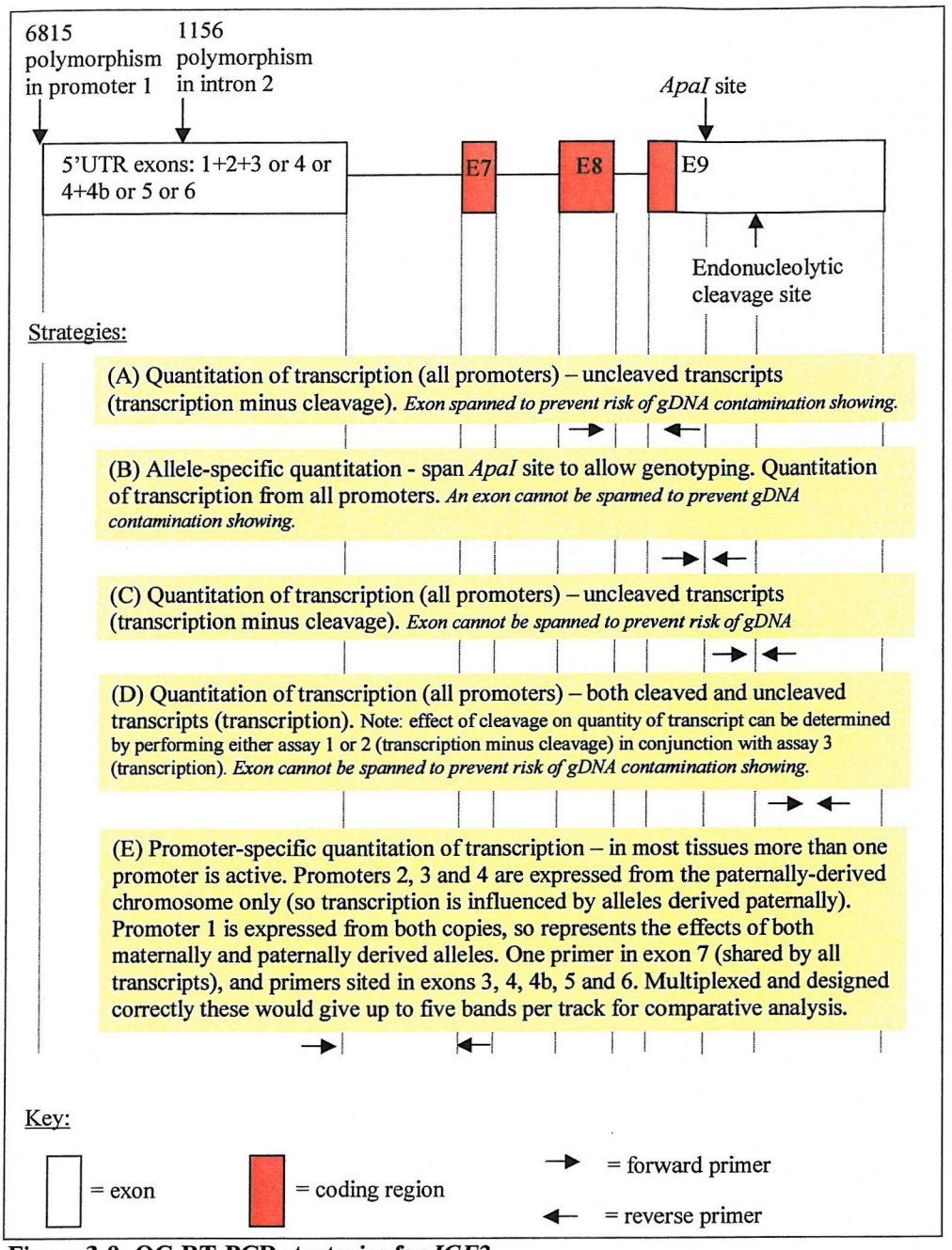


Figure 3-9: QC-RT-PCR strategies for *IGF2*
Different primer locations within the *IGF2* enable different aspects of transcription and RNA regulation to be investigated. Effects of individual promoters or of the endonucleolytic cleavage site can be investigated. Each yellow box represents a strategy, with primer positions below representing their position on the map.

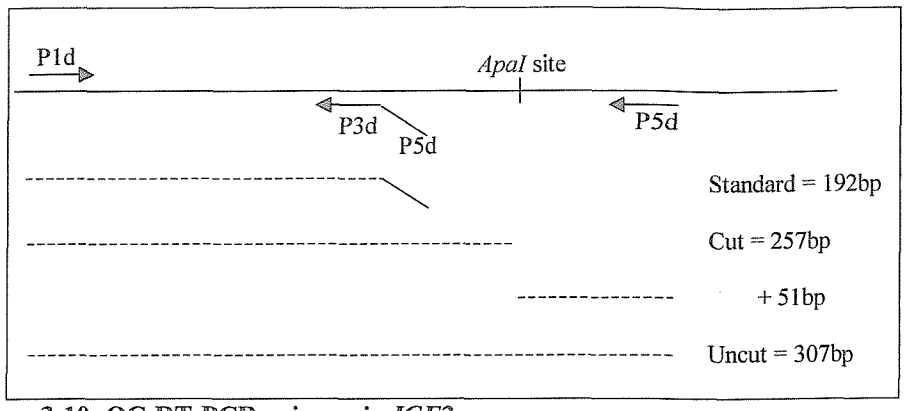


Figure 3-10: QC-RT-PCR primers in *IGF2*
QC-RT-PCR primers were designed to generate an RNA standard of 192nucleotides. RT-PCR of a mixture of this standard and total RNA would therefore result in PCR products of 307bp (from *IGF2* RNA) and 192bp (from the RNA standard). A subsequent *ApaI* digest would cleave the 307bp product to 257bp and 51bp in samples containing that restriction site.

A lack of sample availability meant that this assay had to be set up using RNA extracted from HepG2 cells (a liver cell line known to express *IGF2*). RNA standard was introduced at a range of concentrations to HepG2 RNA, and the mixture then reverse-transcribed to cDNA using primer 5d. At this stage the cDNA present comprised both cDNA standard generated from the RNA standard and *IGF2* cDNA. The ratio of these two cDNAs was expected to reflect the initial molar ratio of the RNA templates from which they were derived. In the case of the standard RNA this was a known quantity.

PCR of the cDNAs for 30 cycles was then used to amplify the cDNA up to levels that could be measured using ethidium bromide on an acrylamide gel. The samples were electrophoresed on horizontal 5% polyacrylamide gels for 20 minutes at 150V to allow adequate separation of the bands. Ethidium bromide staining was used, and gels were imaged on a Molecular Dynamics Fluorimager 595. Phoretix 1D Advanced software was then used to measure band intensity after background subtraction. Figure 3-11 shows a gel of four replicates of QC-RT-PCR. From right to left on the gel the quantity of RNA standard added to the original total RNA decreases (halving each time). As a consequence the ratio of the 307bp to 192bp bands changes, with the 307bp *IGF2* RNA out-competing the standard where the standard is at a quantity less than 0.81femtomoles.

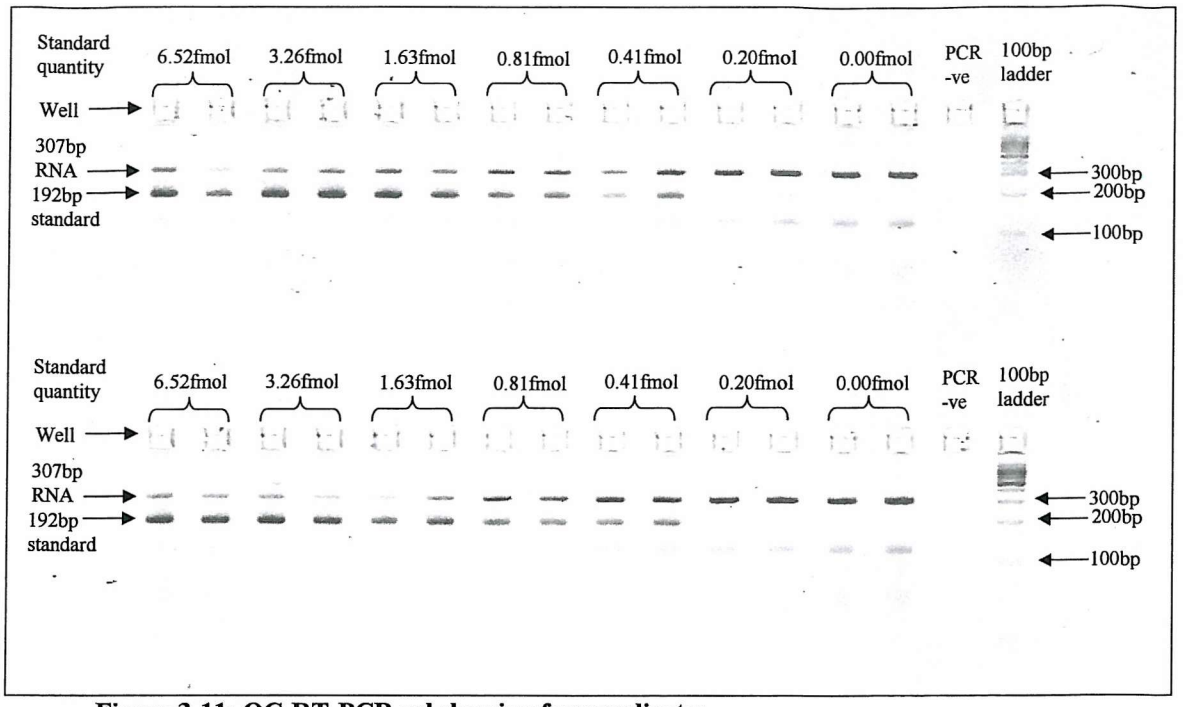


Figure 3-11: QC-RT-PCR gel showing four replicates.
This gel shows four replicate QC-RT-PCR experiments. Each row contains two replicates, with alternate samples from each replicate. Primer-dimers are visible below the standard at lower standard concentrations.

3.6.2.2 Accuracy and reproducibility of QC-RT-PCR

To compare the results from different replicates the band intensities were all entered into a spreadsheet. Band intensity cannot be compared between gels because variability between gels is too high. However, ratios between standard RNA product and *IGF2* RNA product were calculated for twenty replicates, which were five PCRs on each of four RT reactions. These are plotted in Figure 3-12. The results show high variability at high standard concentrations, but appear to be relatively accurate at a ratio of 1 (the standard concentration at which we assume standard and test RNAs are equimolar). The ratios show a decrease between 3.26 and 6.52 femtomoles of standard in contrast to the increase in ratio observed in the rest of the graph as standard concentration increases. This is probably due to intensity limits of the imaging equipment (i.e. the standard intensity is recorded as lower than it should be, while the less intense test sample is recorded accurately).

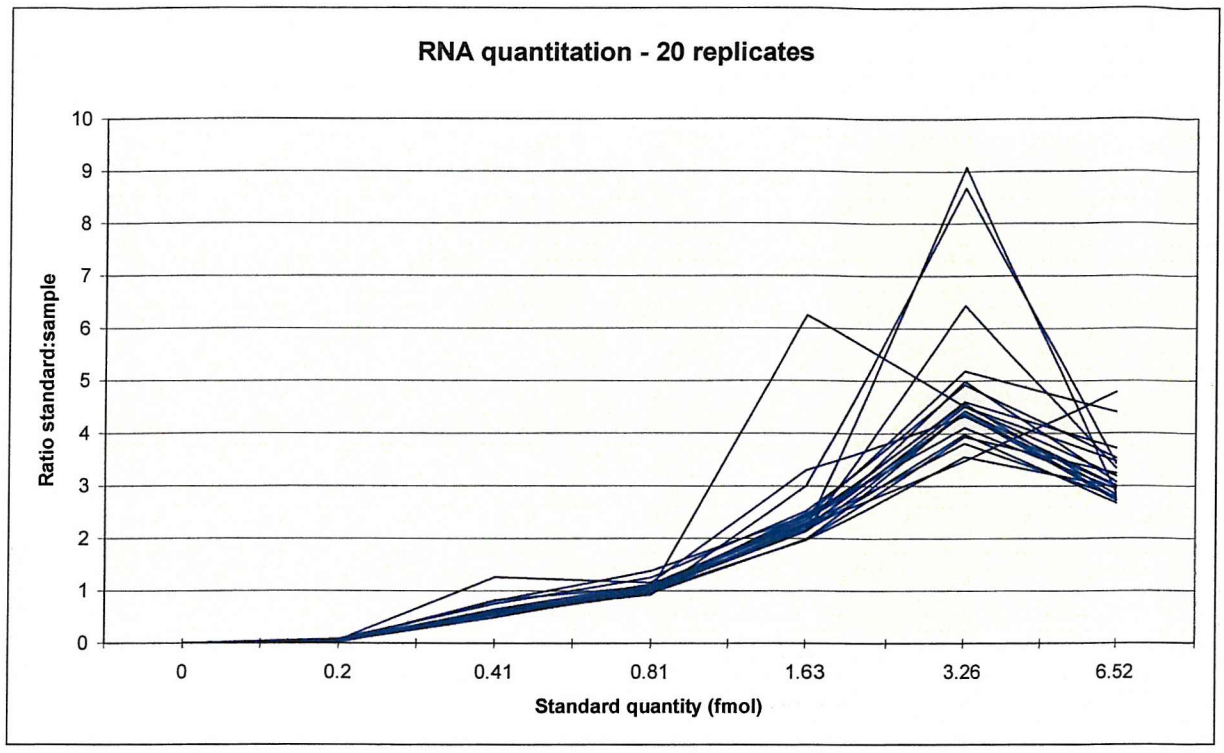


Figure 3-12: RNA quantitation using QC-RT-PCR.
Samples are replicates from the same original HepG2 RNA. Four replicate RT reactions are shown for each standard RNA quantity, and five PCR reactions on each of the four RT reactions (total = 20 replicates).

To determine the consistency of this approach, the mean ratios and standard deviations were plotted (Figure 3-13). The small standard deviation at the 1:1 ratio intersection on the gel indicates that this method is fairly consistent between experiments. The inter-assay coefficient of variation (CV) in band intensity ratio at the 0.81 femtomole quantity was 9.96%. The higher standard deviation (and CV) at other points (both higher and lower) demonstrates that measuring the ratio at a single concentration of standard RNA would be unreliable. A titration of standard RNA is therefore necessary to allow accurate RNA quantitation by this method.

Negative RT and negative PCR controls were included in all experiments to identify any genomic contamination.

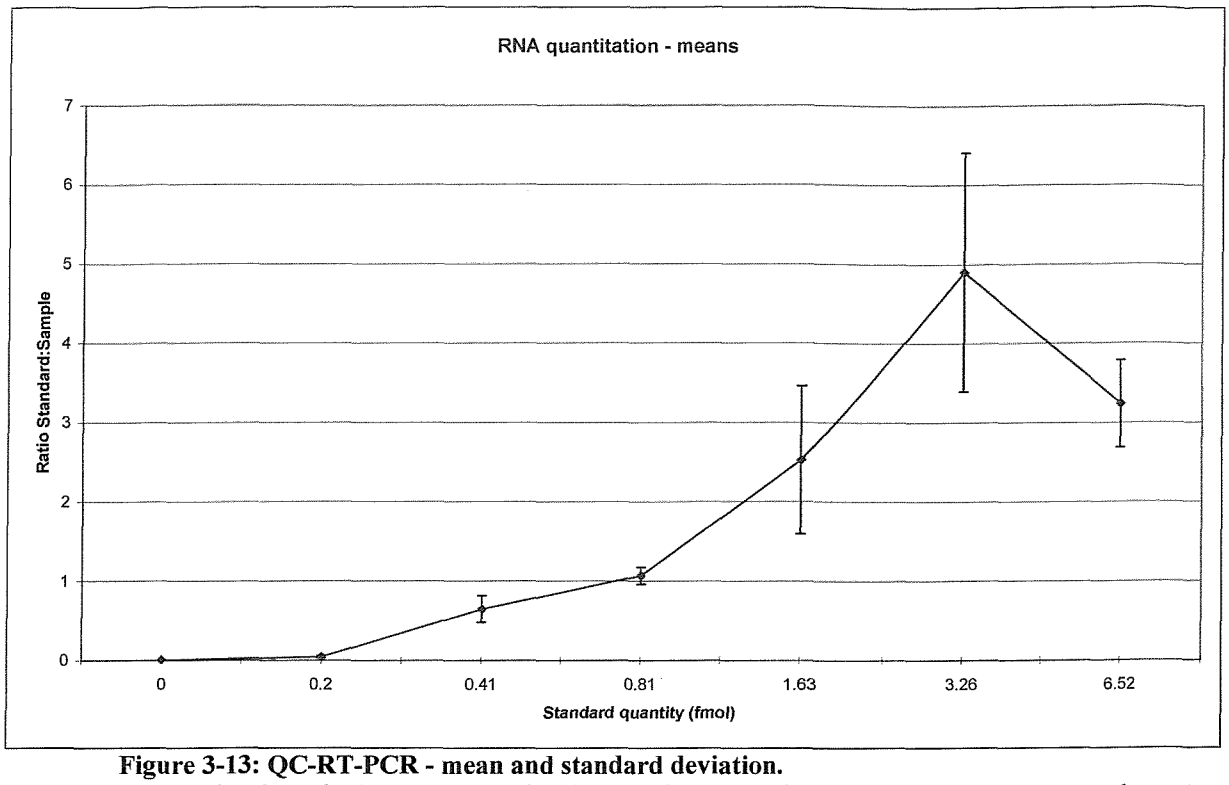


Figure 3-13: QC-RT-PCR - mean and standard deviation.
Mean ratio of standard to test sample band intensities. The point at which the line intersects the y=1 line should indicate the standard quantity (x) at which standard and *IGF2* RNAs were equimolar in the original mixture. Error bars show standard deviation.

3.6.2.3 Restriction digestion of QC-RT-PCR products

The use of *ApaI* to genotype QC-RT-PCR products, and to compare expression of each allele was tested (Figure 3-14). Digestion of QC-RT-PCR products appears to be nearly complete – the residual 307bp band may be due to inhibition of digestion by components of the RT reaction. The genomic DNA control shows that this sample only contains the cutting allele of the *ApaI* polymorphism. Visualisation of relative band intensities indicates that this method should still allow quantitation of individual alleles, with the *IGF2* and standard QC-RT-PCR bands being approximately equal intensity at 0.81 femtomoles of standard RNA. However, this experiment should be tested with other RNA samples to ensure that digestion is representative of genotype (i.e. to verify whether the apparently partial digest here is a consequence of residual RT components).

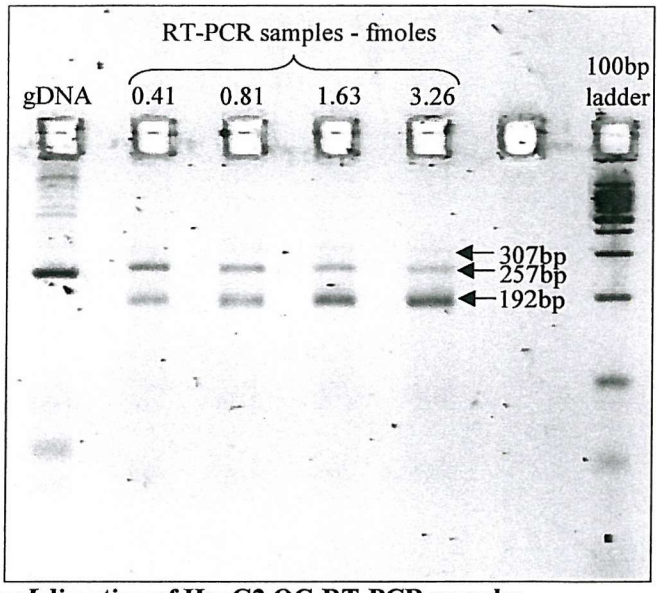


Figure 3-14: *ApaI* digestion of HepG2 QC-RT-PCR samples
Four different standard concentrations are shown. The genomic DNA control was generated using different primers from the same cell line, and shows that HepG2 cells contain only the cutting allele at this polymorphic site.

3.6.3 Discussion

The quantitative-competitive RT-PCR assay presented here is intended to allow an investigation of the relationship between genotypes at polymorphic loci and expression of the *IGF2* gene. This will help in the understanding of the observation of statistically significant association between genotype and anthropometric phenotypes. This approach should indicate whether an associated polymorphism marks a polymorphism that influences gene expression, or whether it marks a polymorphisms that influences something further downstream (e.g. protein function). It cannot necessarily identify the mechanism of alteration of gene expression.

The assay is a relatively accurate approach to determining the expression level of *IGF2*. When a titration of standard RNA is used the method is accurate to within 10%, which should be sufficient to detect marked differences in gene transcription between genotypes. Replication of assays would be necessary to ensure representative results, but the use of MADGE gels to improve throughput (Zhang *et al.* 2002) would make this straightforward.

A complicating issue for this assay is the variable tissue-specific imprinting of *IGF2* (Wu *et al.* 1997b). Promoter 1 is expressed biallelically, while promoters 2→4 are expressed monoallelically (Vu & Hoffman 1994). However, the levels of expression from each of the promoters varies between tissues, meaning that variable levels of expression of each of the parental alleles are observed in different tissues (Li *et al.* 1996; Wu *et al.* 1997b; Ekstrom *et*

al. 1995). This would require an approach to quantitative RT-PCR in which homozygous samples were used to prevent the lower levels of expression of the maternal allele influencing the result. In addition, promoter-specific RT-PCRs could be used to determine the ratio of expression from each of the four promoters in each sample.

3.7 Conclusions

3.7.1 High-throughput genotyping methods

A system has been developed which utilizes 384-well PCR, electrophoresis and analysis, and is probably the maximum density achievable without robotic automation. As a result throughput is limited by the time required for experimental design and analysis of results, with DNA resources, equipment and time in the laboratory no longer being rate-limiting factors (Figure 3-15). Improvements in experimental design strategies (three-primer ARMS assays) and analysis (Phoretix and cluster analysis), and potential future improvements (mainly software based) are required to reduce the extent to which office-based work limits throughput in the lab.

Modern molecular population genetics is based on rapidly evolving technology, with methods becoming out of date within a few years. The high-throughput methods presented here were essential for this project and will be used in other projects because they present the most cost-effective approach available for this level of throughput. However, continual development will be necessary to maintain their effectiveness.

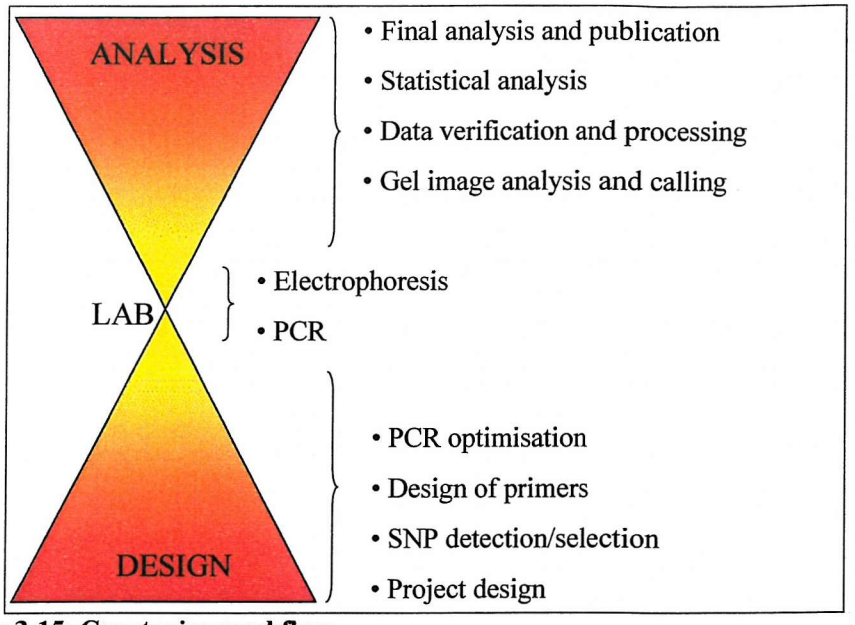


Figure 3-15: Genotyping workflow.
The majority of work is in the design and analysis phases of genotyping. The laboratory-based part of the project is efficient, and limited by the office-based parts.

3.7.2 Quantitative-competitive RT-PCR

A quantitative-competitive RT-PCR for *IGF2* has been established, and will enable an investigation of the relationship between genotypes at different loci and expression levels of the gene. The limiting factor for this project is availability of samples, particularly as a relatively large number of samples will be required to provide statistically useful data for all genotypes at all loci.

Chapter 4 : A map of polymorphism in *IGF2*

4.1 Introduction

4.1.1 Detection of unknown mutations

Detection of unknown mutations is more complex than the detection of known mutations (i.e. genotyping) because determination of false negatives (i.e. samples which incorrectly appear to contain no mutation) is difficult. The ideal method is therefore one with a high reliability, sensitivity and specificity (although this can only be determined by comparison with other methods). Thorough bi-directional sequencing is considered the 'gold standard' for mutation detection (and the only method for identification) (Cotton 1998), although even with modern chemistries the detection of heterozygous samples can be difficult.

Single-strand conformation polymorphism (SSCP) analysis (Orita *et al.* 1989a; Orita *et al.* 1989b) is a well-established technique which distinguishes between single-stranded PCR amplicons on the basis of their electrophoretic mobility under non-denaturing conditions. Secondary structures within the single-stranded molecules differ depending on sequence context, and thus mobility during electrophoresis differs depending on sequence. Wild-type and mutant alleles are thus distinguished on the basis of how far they run in a gel.

Several techniques for mutation detection utilise the existence of heteroduplexes in heterozygous samples. A heteroduplex is a double-stranded DNA duplex comprising one wild-type strand and one mutant strand. The disruption of base-pairings between the two strands at the position of the mutation results in physical characteristics that can be distinguished by several methods: (1) difference in electrophoretic mobility (White *et al.* 1992), (2) cleavage at the point of mismatch using chemicals (Cotton *et al.* 1988) or enzymes (Myers *et al.* 1985a), (3) difference in duplex denaturation with denaturing gradient gel electrophoresis (Myers *et al.* 1985b) or denaturing high performance liquid chromatography (Oefner & Underhill 1995). Many variants of these methods exist, but in general the more sensitive methods are more complex, time-consuming and expensive, while the more straightforward methods are less sensitive (O'Donovan *et al.* 1998).

In silico methods compare sequence data to identify mismatches using computer programs. Some of these utilise automated sequence trace data, while others utilise large sequence text databases. "Fluorescent sequence trace subtraction" compares fluorescent chromatograms from automated sequencers, generating a "difference trace" by subtracting each sample from a consensus chromatogram (i.e. plotting the difference in intensity of fluorescence between sample and consensus) (Bonfield *et al.* 1998). The "difference trace" shows peaks only where the sequences differ, and thus indicates potential mutations. Another approach is to align simple text sequences from public databases and identify differences between them. One approach is to use the public expressed sequence tag (EST) databases that have many duplicates of the same expressed sequences. Sequence assembly of many sequences from the same region (with the assumption that they originate from different individuals) allows the detection of differences between those sequences (Picoult-Newberg *et al.* 1999). A more sophisticated approach combines the method of chromatogram comparison with that of multiple sequence alignment from public databases. POLY-BAYES (a Bayesian inference engine for polymorphism detection) aligns EST chromatograms (where available in dbEST) and performs quality assessment on the sequence data prior to assembly and comparison (Marth *et al.* 1999). The advantage of this approach is that sequence quality is assessed, reducing the chance of false positives due to sequencing errors.

In silico methods allow very large-scale scanning of the entire genome, but are limited by the availability of sequence data, and tend to produce poor quality results. Lab-based methods are more reliable, but are best suited to relatively small regions of the genome.

4.1.2 DHPLC and SSCP

Single-strand conformation polymorphism analysis (SSCP) (Orita *et al.* 1989b) is an established and commonly used technique for the detection of unknown mutations, requiring only standard laboratory equipment, while denaturing high performance liquid chromatography (DHPLC) (Oefner & Underhill 1995) is a more recent development usually performed on a specialist DHPLC instrument. Important considerations in choosing a mutation detection technology are: sensitivity, specificity, throughput and cost.

SSCP with autoradiography produces detection rates of approximately 65% (Eng *et al.* 2001) to 80% (when two temperatures are used, as in this project) (Hinks *et al.* 1995), compared to a detection rate of 95% (Dobson-Stone *et al.* 2000) to 100% (Eng *et al.* 2001; O'Donovan *et al.* 1998) for DHPLC. Fluorescent SSCP appears to improve the detection rate of SSCP to 93% (Dobson-Stone *et al.* 2000). However, while DHPLC maintains nearly 100% efficiency with product sizes of around 500bp, and can still work at 1500bp, SSCP is less efficient with product sizes over 200bp (O'Donovan *et al.* 1998). Direct comparisons of SSCP and DHPLC have shown DHPLC to be superior in sensitivity (Choy *et al.* 1999; Jones *et al.* 1999), except in the case of fluorescent SSCP, where the sensitivity was similar to that of DHPLC (Dobson-Stone *et al.* 2000).

The sensitivity of DHPLC with larger PCR products (O'Donovan *et al.* 1998) allows more rapid scanning of long stretches of sequence. However, with short exons of around 200bp or less DHPLC no longer has this advantage over SSCP. The automation of DHPLC instruments allows many more samples to be scanned per operator than is practical with SSCP.

The cost of DHPLC apparatus exceeds that of SSCP apparatus. However, DHPLC detects PCR products using light absorbance at 260nm, whereas SSCP usually utilises fluorescence or radio-labelling, both of which are relatively expensive per sample, and require equipment for gel-scanning or film developing.

Two advantages of SSCP over DHPLC are: its detection of different sequence variants within the same fragment (Dobson-Stone *et al.* 2000), and the ability to distinguish between wild-type and mutant homozygotes. For SNP detection DHPLC is largely dependent on heteroduplex formation for mutation detection, and thus cannot distinguish between homozygotes of different sequence. However, addition of a known wild-type sample to an unknown homozygote will generate heteroduplex only if the unknown sample is a mutant, so this approach may be used.

4.1.3 Published polymorphisms in *IGF2*

At the start of the project several polymorphisms in the *IGF2* gene had been identified and characterised by others. Two studies into the role of the insulin gene (*INS*) in type 1 diabetes identified a number of polymorphisms in the *INS* gene and surrounding region, which includes the 5' end of the *IGF2* gene (Owerbach & Gabbay 1993; Lucassen *et al.* 1993). The furthest 3' of these polymorphisms was an *AluI* polymorphism (Y13633-1252T→C) in exon 3 (Lucassen *et al.* 1993).

At the 3' end of the gene three polymorphisms were identified in the literature: an *ApaI* polymorphism (X07868-820G→A) (Tadokoro *et al.* 1991), a *HinfI* polymorphism (X07868 – approximately 25bp upstream of the *ApaI* site (Vu & Hoffman 1994) and a complex repeat polymorphism of CA dinucleotides periodically interspersed with other sequences (Rainier *et al.* 1993). These had been used in projects investigating imprinting of the *IGF2* gene (Vu & Hoffman 1994; Wu *et al.* 1997a; Giannoukakis *et al.* 1996; Rainier *et al.* 1993), comparing expression of the two alleles in heterozygous samples. The *ApaI* site had also been used in an investigation of population obesity (O'Dell *et al.* 1997).

4.2 Results

4.2.1 Polymorphisms identified by SSCP

Single-strand conformation polymorphism analysis (SSCP) was used to scan the exons and promoter regions of the *IGF2* gene, and in addition to confirm the presence of several published polymorphisms in intron 1. Exons 2 and 4b were not scanned because their locations were not known at the time of the scan. Exon 5 was only partially scanned because the high GC content of this exon (Figure 4-1) prevented amplification of several fragments. The entire coding region (consisting of exons 7, 8 and the first part of exon 9) and the remaining exons were all scanned with overlapping amplicons initiating in the surrounding intronic sequence and spanning the entire exon. The following published polymorphisms were excluded from the SSCP scan: Y13633-1252T→C (Lucassen *et al.* 1993), X07868-820G→A (Tadokoro *et al.* 1991) and the 3'UTR $(CA)_n$ repeat (Rainier *et al.* 1993). The extent of the SSCP scan of *IGF2* is shown in Figure 4-2.

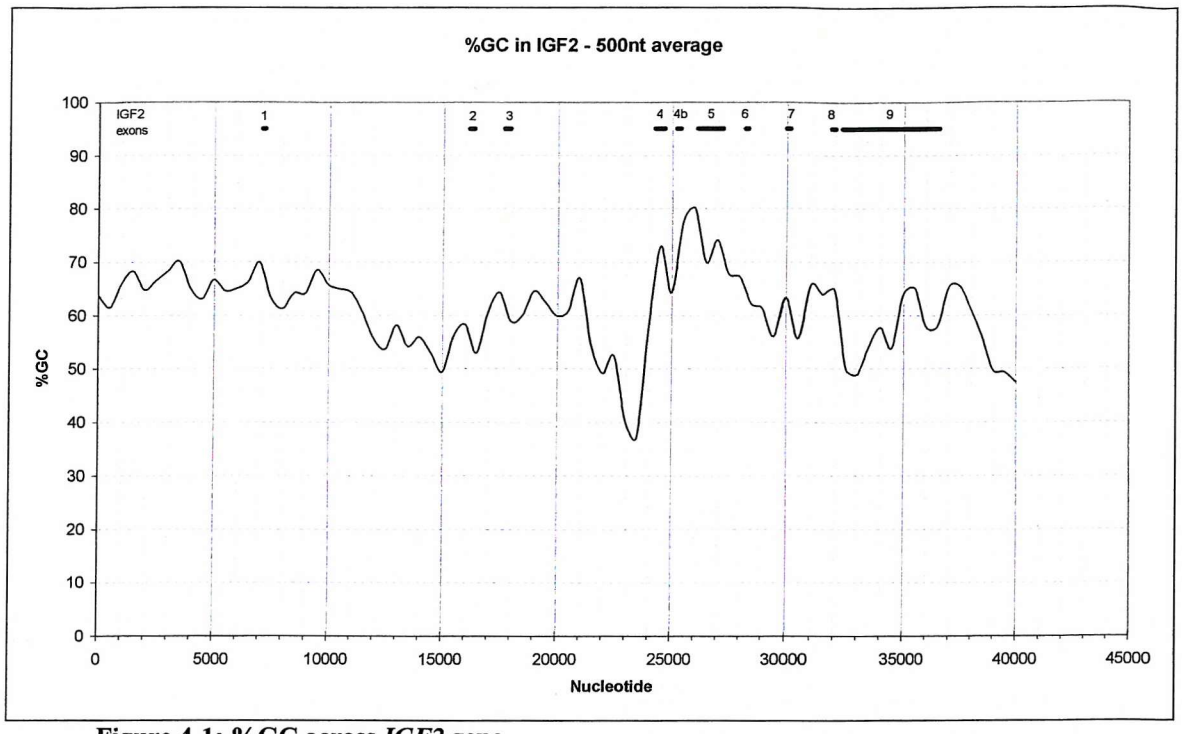


Figure 4-1: %GC across *IGF2* gene.
Peaks around exon 5 interfere with amplification. %GC is shown by thin line. Thick lines indicate the positions of *IGF2* exons relative to the plot.

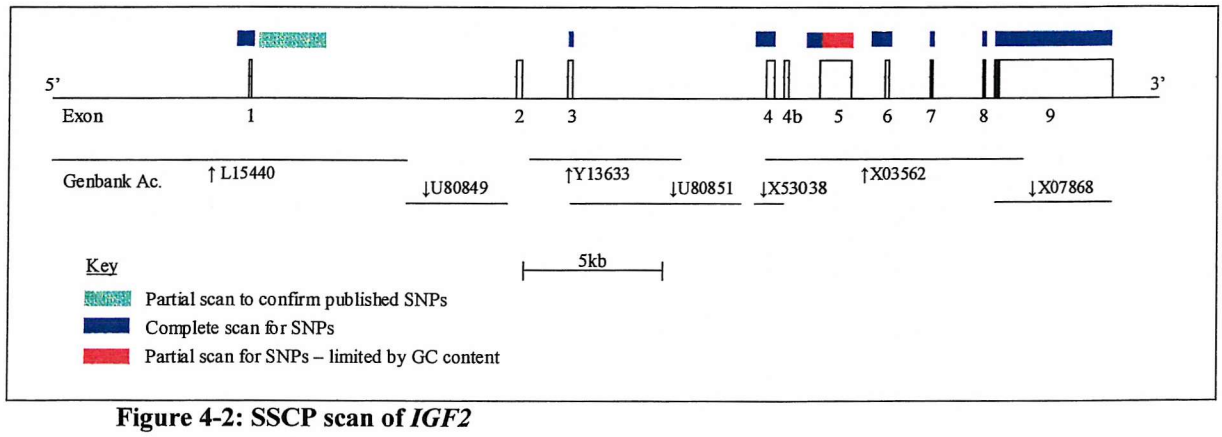


Figure 4-2: SSCP scan of *IGF2*

The lengths of the different SSCP amplicons in *IGF2* exons are shown in Figure 4-3. The average SSCP amplicon length was 212bp, with a minimum of 115bp and a maximum of 290bp. Amplicon length was determined by the location of primers, which was dependent on sequence content. Amplicons were overlapped by an average of 60bp, and a minimum of 40bp within each exon. This ensured that maximum coverage was achieved; 40bp is the minimum overlap with primers of 20 nucleotides to ensure that polymorphisms are not missed within the primer sequences.

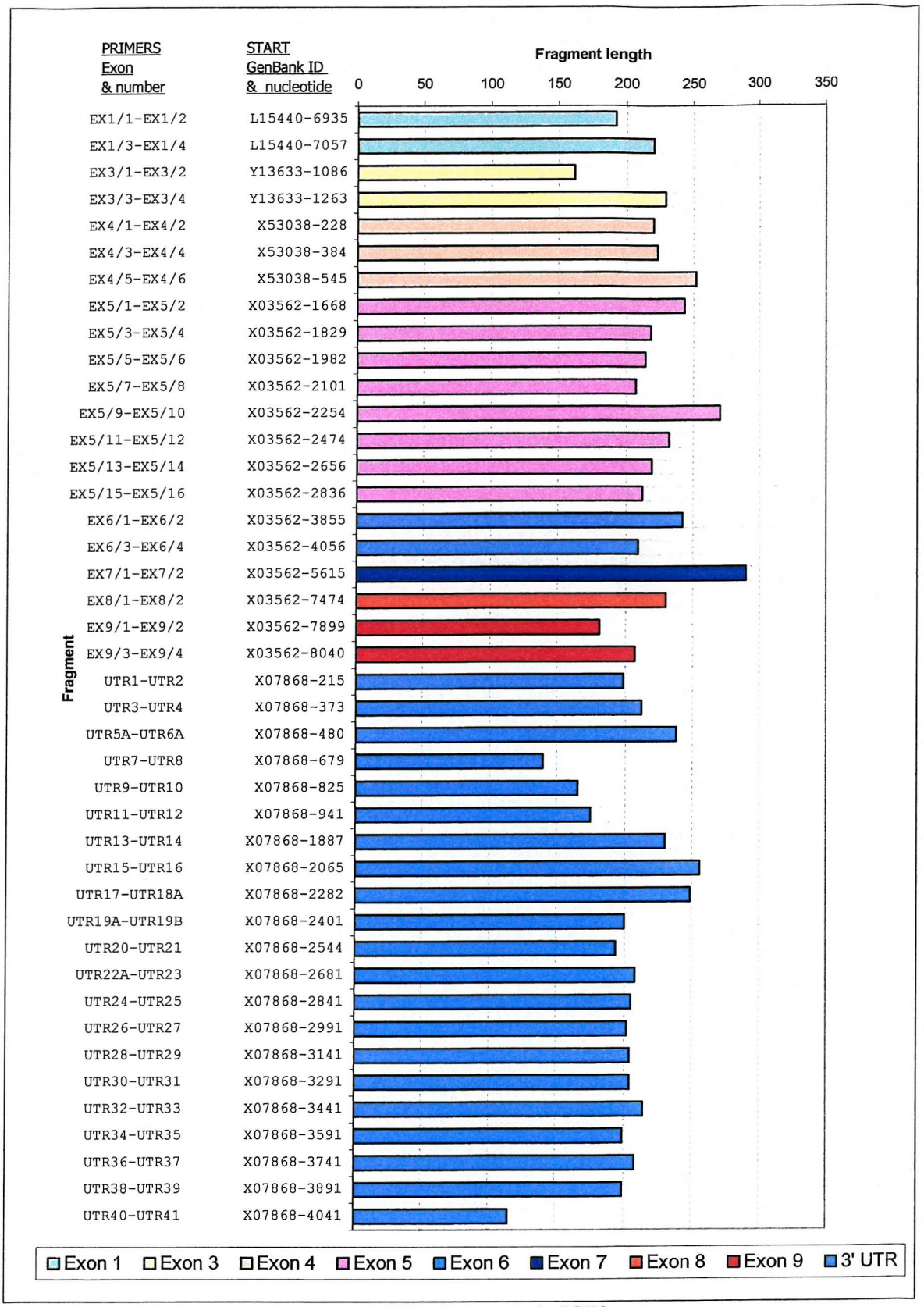


Figure 4-3: Length and number of SSCP fragments in *IGF2* exons.
EX1/1 refers to primer 1 in exon 1. UTR1 refers to primer 1 in the 3' UTR. GenBank sequence and amplicons start are also shown.

For a diallelic polymorphism SSCP allows discrimination between the three genotypes. Usually two bands are seen for a common homozygote, with two bands of different mobility indicating a rare homozygote. Four bands indicate heterozygotes. Some bands may co-migrate resulting in a reduction in the number of bands seen. The heterozygote pattern should, however, always be a combination of the patterns for the two homozygotes. Figure 4-4 shows a sample SSCP result. The figure shows two gels, each split into two sets of ten samples that are homozygous for one of the two *ApaI* alleles (X07868-820G→A in the 3'UTR of *IGF2*). This enabled a preliminary estimation of the extent of linkage disequilibrium (LD) between the newly identified SNP and the previously published polymorphism. If a SNP was in strong LD with the *ApaI* site then the common homozygotes tended to segregate into one half of the gel and the rare homozygotes into the other. If LD was weak between *ApaI* and a new SNP then there was no such segregation.

In Figure 4-4 genotypes for the newly identified SNP are shown as 11 (common homozygote), 12 (heterozygote) and 22 (rare homozygote). Heterozygotes appear to have three bands due to similar migration of the lowest two bands, although these are separate in the less intense samples (lower right panel, samples 7, 8 and 9). This polymorphism does not appear to be in strong linkage disequilibrium with the *ApaI* site.

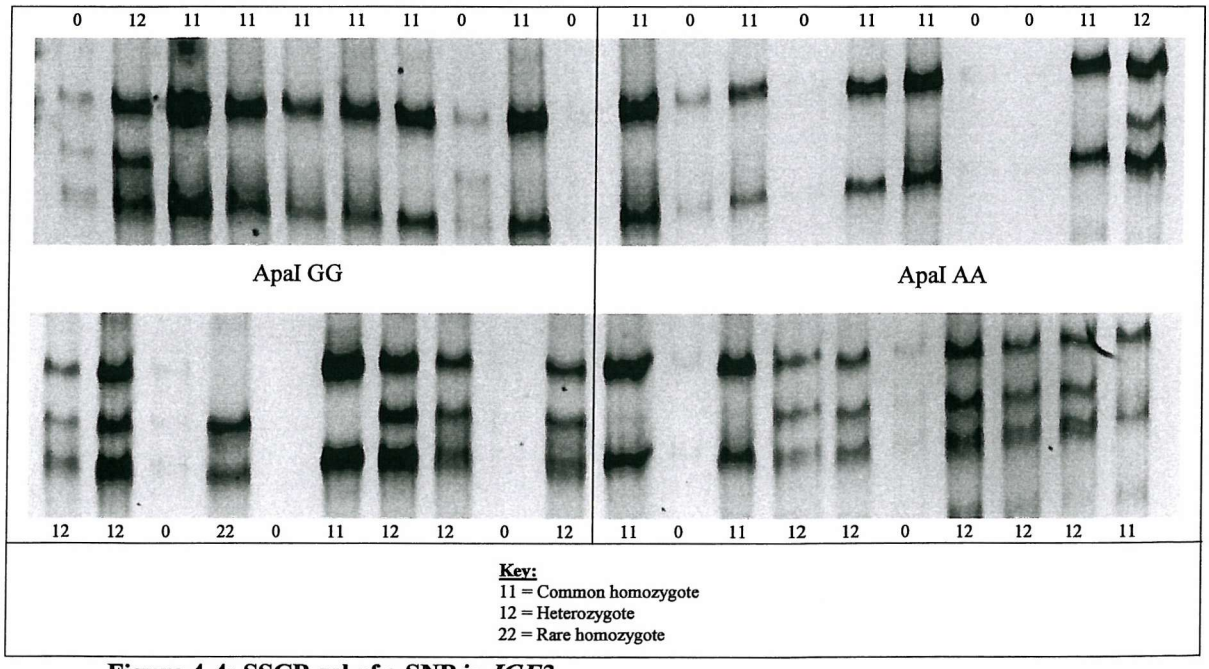


Figure 4-4: SSCP gel of a SNP in *IGF2*.
Four panels of 10 individuals from 2 separate gels are shown with genotypes. Three/four bands are observed in heterozygotes and two in homozygotes. 11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.

Seven polymorphisms were identified by SSCP in 7.5kb of scanned sequence, and four published polymorphisms were confirmed in additional sequence. These are shown in Table 4-1.

Table 4-1: Polymorphisms discovered by SSCP.
Polymorphisms are identified by a GenBank reference of the sequence and a nucleotide number. These are all in previously published sequence with the exception of the A→G SNP in missing sequence in X07868 between nucleotides 3750 and 3751.
* Published polymorphism confirmed by SSCP

GenBank ID	Nucleotide	Polymorphism
L15440	7368	A→T * (Lucassen *et al.* 1993)
L15440	7686	G→A * (Lucassen *et al.* 1993)
L15440	8065	G→A * (Lucassen *et al.* 1993)
L15440	8173	C→T * (Lucassen *et al.* 1993)
Y13633	1156	T→C
X07868	266	C→T
X07868	556	Poly C
X07868	1926	C→G
X07868	In unpublished sequence between 3750 and 3751	A→G
X07868	3944	G→A
X07868	4085	G→A

4.2.2 Polymorphisms identified by DHPLC

Denaturing high performance liquid chromatography (DHPLC) was used to scan some intronic regions in the *IGF2* gene for unknown polymorphisms. DHPLC was also used to scan some of the exonic sequence not covered by SSCP. The larger fragment size of DHPLC allowed more rapid scanning of these wider regions. The regions scanned are shown in Figure 4-5.

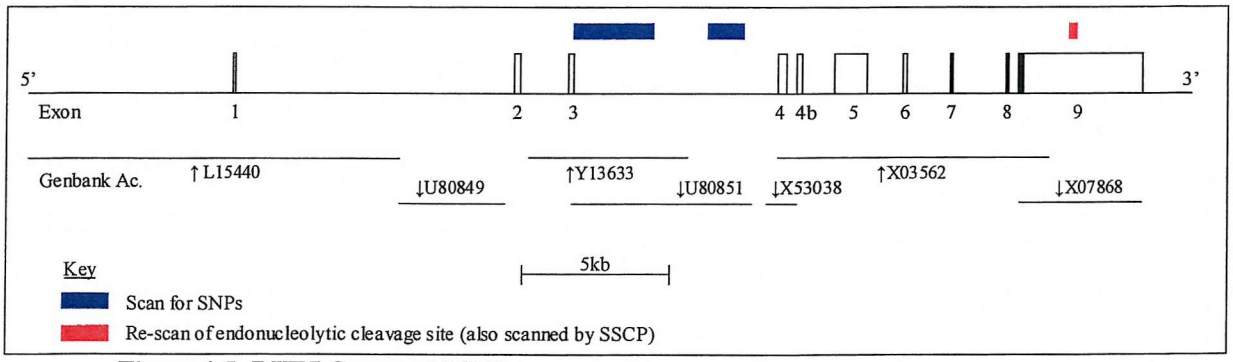


Figure 4-5: DHPLC scan of *IGF2*.

The DHPLC system is a high-performance liquid chromatography system with a double-strand DNA specific column. Acetonitrile interferes with the binding of DNA to the column, with the percentage required to cause elution dependent on the length of the DNA duplex. By maintaining the column at a temperature close to the melting temperature of the sequence heteroduplexes are less double-stranded than homoduplexes, and therefore elute at a lower concentration of acetonitrile. Typical DHPLC results are shown in Figure 4-6. Homozygous samples produce only homoduplex PCR product, which all elutes at the same concentration of acetonitrile, thus producing just one peak. Heterozygotes produce both heteroduplex and homoduplex PCR products, which elute at different acetonitrile concentrations. Panel B in Figure 4-6 shows a more complex pattern. This was later confirmed to be a homopolymeric tract length polymorphism rather than a SNP. More complex patterns may indicate the presence of more than one SNP in a sample, or just a more complex type of polymorphism – polymorphisms that alter the length of the homoduplex, such as that seen in panel B may produce two homoduplex bands (the shorter one eluting earlier) and one or two heteroduplex bands.

With DHPLC there is usually no distinction between wild-type and mutant homozygotes, so the method is dependent on the presence of heterozygotes for mutation detection. The numbers of each homozygote can only be determined by sequencing, or by combining with a reference sample (known homozygote) to generate artificial heterozygotes.

Four polymorphisms were identified within the scanned sequence using DHPLC. These are shown in Table 4-2. The identities of the polymorphisms were confirmed by sequencing.

Table 4-2: Polymorphisms discovered by DHPLC.
Polymorphisms are identified by a GenBank reference of the sequence and a nucleotide number.

GenBank ID	Nucleotide	Polymorphism
Y13633	2482	A/C
Y13633	2722	C/T
U80851	5345	Poly T
X07868	2207	C/T

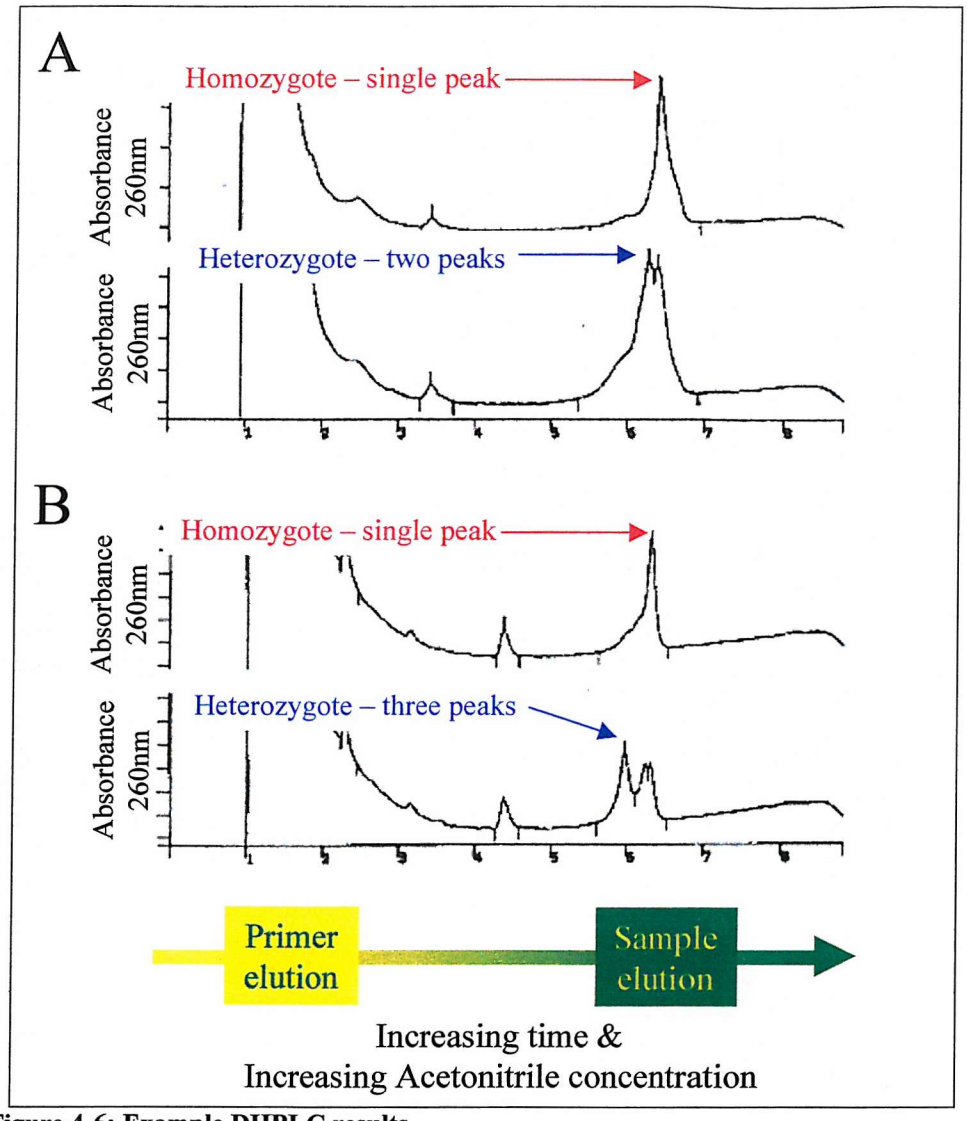


Figure 4-6: Example DHPLC results.
(A) typical results with the upper chromatogram showing a single peak in the sample elution phase and the lower chromatogram showing a split (double) peak in the sample elution phase. (B) more complex pattern, with the lower chromatogram showing a split peak and a third more separated peak in the sample elution phase. A (large) primer peak is always present and non-specific PCR products are seen to elute before the sample.

4.2.3 Sequencing of SNPs

Both SSCP and DHPLC identify the presence of a sequence variant in a fragment. However, the position and type of polymorphism are not determined by these methods. Fluorescence-based automated sequencing was used to confirm and identify the variants. The ABI310 capillary sequencer was ideally suited to this application, designed for low throughput sequencing with high quality results. Samples were sequenced using BigDye Terminator chemistry, which generates relatively even peak heights (compared to non-BigDye chemistry), and POP-6 polymer with a high-resolution run module (rather than high-

speed). Fragments were generated using primers external to the original SSCP or DHPLC fragment, and then sequenced twice, once with the forward primer and once with the reverse. Example SNP chromatograms are shown in Figure 4-7. Homozygotes showed a clear sequence of single peaks, while heterozygotes showed an overlaid double peak at the site of the SNP. Sequence alignment software identified the difference between homozygotes, while a visual scan of the sequence identified the location of heterozygous SNPs.

More complex sequencing results were generated for homopolymeric tract length polymorphisms. This type of polymorphism is a run of several identical nucleotides that is polymorphic in length rather than sequence content. The example shown in Figure 4-8 has three alleles identified in the chromatograms shown. The 10/11C heterozygote has a complex sequence with the upstream sequence being clear, but the downstream sequence appearing heterozygous at many positions due to the frame shift present in half the PCR product caused by the tract expansion. The 9C homozygote has clear sequence either side of the tract.

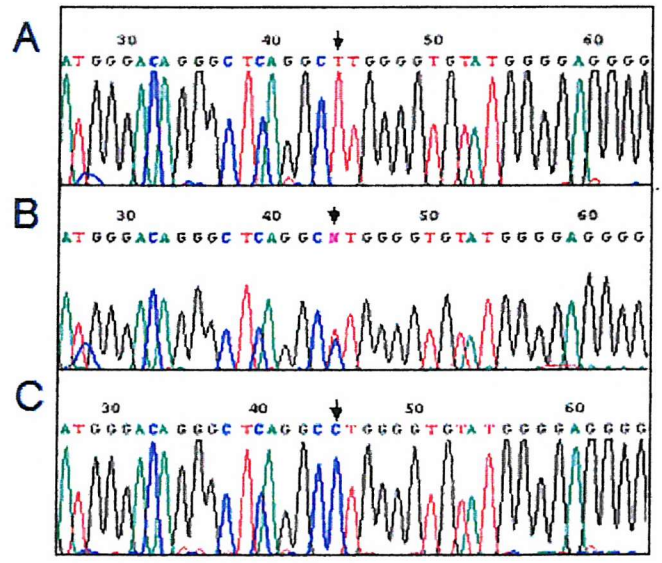


Figure 4-7: Sequencing result for Y13633-1156T/C.
(A) T/T homozygote (T arrowed); (B) T/C heterozygote showing overlaid peaks (arrowed) – typically heterozygote peaks are smaller than adjacent peaks due to division of template between the two alleles; (C) C/C homozygote (C arrowed)

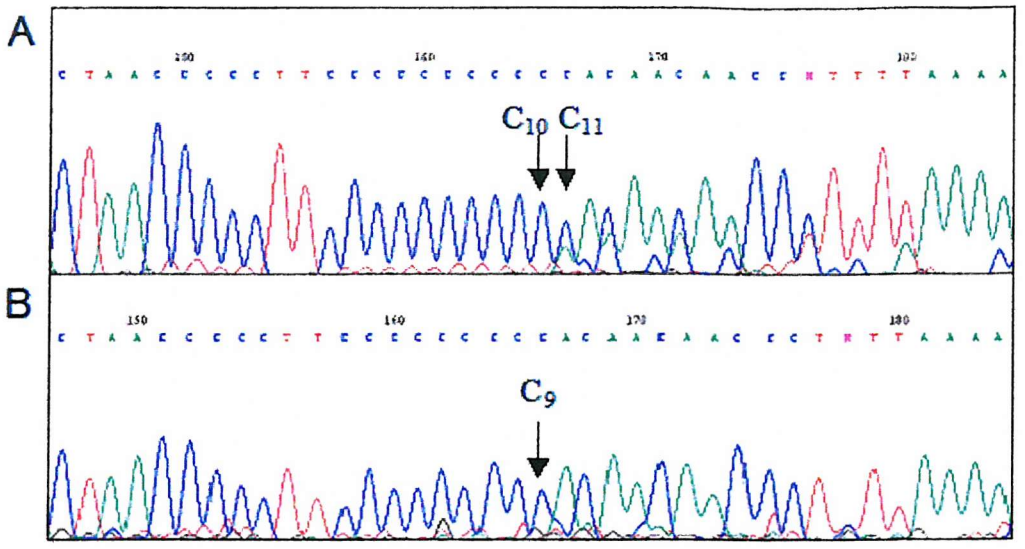


Figure 4-8: Sequencing result for X07868-556polyC.
(A) C_{10}/C_{11} heterozygote appears as a homopolymeric tract with a downstream heterozygous 1-base shift indicating the presence of two length alleles in the tract; (B) C_9 homozygote has a shorter homopolymeric tract with no apparent downstream shift.

Figure 4-9 shows a selection of sequencing results for different SNPs. Two transitions are shown (equivalent to each other), and two of the three possible transversions are shown. The other transversion (A→T) was not seen in any of the newly identified SNPs. This selection indicates the ease with which polymorphisms can be identified, even in heterozygous sequence showing some noisy signal and uneven peak heights (X07868-4085). The lack of suitable software for heterozygote detection meant that all sequence chromatograms had to be manually checked for heterozygotes. However, the use of SSCP and DHPLC meant that the number of samples that had to be sequenced was small. There were no other types of polymorphism identified by this part of the project. Of 11 polymorphisms identified, nine were SNPs and 2 were homopolymeric tract expansions.

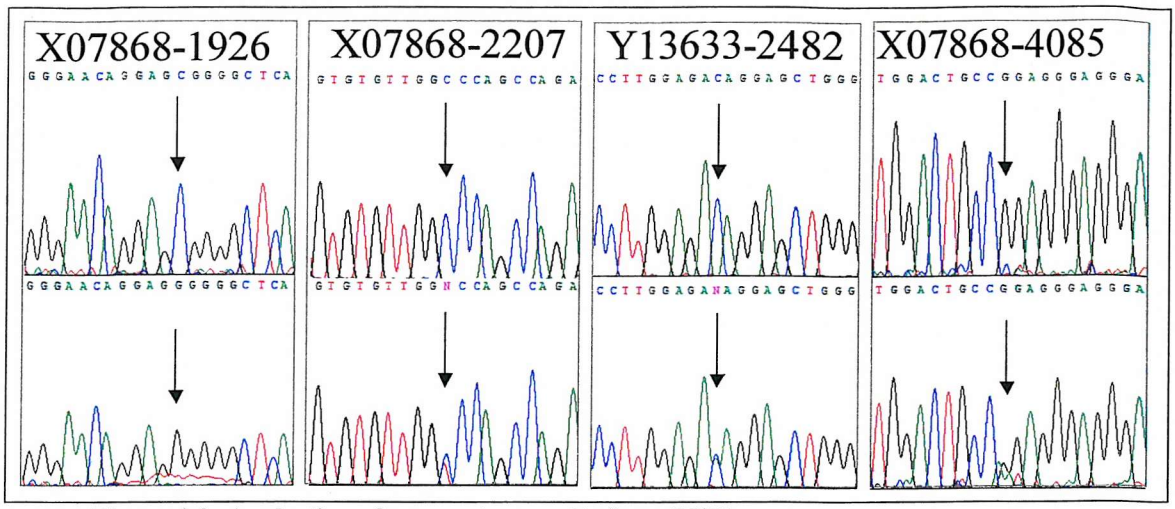


Figure 4-9: A selection of sequencing results from *IGF2*.
Two transitions (C→T and G→A) and two transversions (C→G and C→A) are shown.

X07868-3750exA→G was found in sequence not included in the GenBank sequence X07868. An additional 74 nucleotides of sequence were found between X07868-3750 and X07868-3751 (Figure 4-10). This sequence was found in three of three samples sequenced, and also in the draft genomic sequence AC006408, indicating that this may be the wild-type sequence, and X07868 the result of a sequencing error.

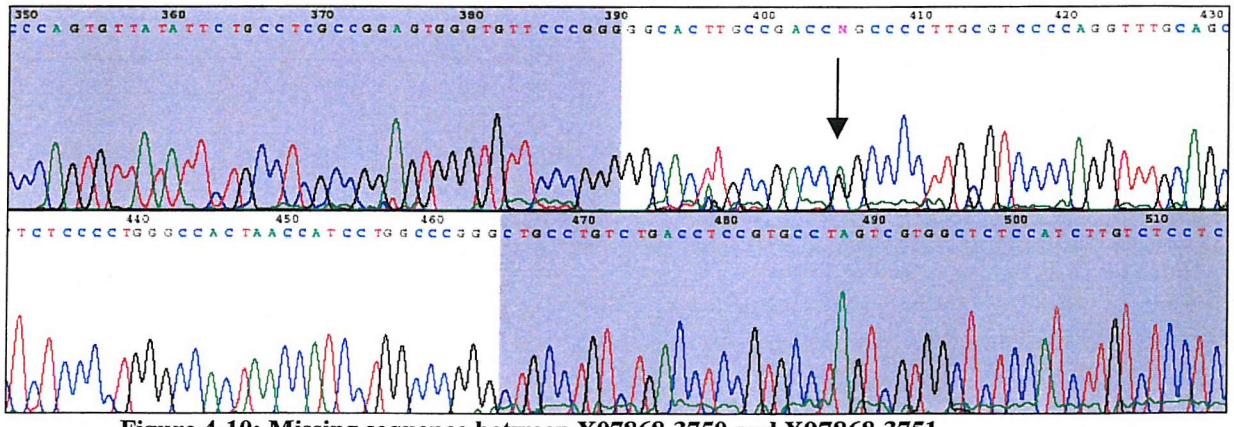


Figure 4-10: Missing sequence between X07868-3750 and X07868-3751
A G→A heterozygote is shown at position 405 (arrowed). The greyed regions show the flanking sequence from GenBank accession X07868. 390 on this sequence corresponds to 3750 on X07868, and 465 corresponds to 3751 on X07868.

4.2.4 In silico mutation detection

In silico mutation detection was attempted by three different methods. The first method was assembly of the previously published *IGF2* sequences using "Gap4" from the "Staden Package", and comparison for mismatches. The second was BLAST alignment of representative *IGF2* sequences against the dbEST database at the NCBI website

(http://www.ncbi.nlm.nih.gov/BLAST/) and comparison for mismatches. The third method was the searching of online databases for identified SNPs. These included SNP databases the Cancer Genome Anatomy Project database (http://cgap.nci.nih.gov/), dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp) and the Whitehead Institute SNP database (http://www-genome.wi.mit.edu/snp/human/).

The first two methods produced no useable data. Many sequence differences were observed between small numbers of overlapping sequences, indicating probable sequencing errors. Because chromatogram quality data was not available no assessment could be made of the likelihood of different potential polymorphisms.

The third method identified several SNPs, all in the 3'UTR (probably due to the 3'UTR being more represented in expressed sequence tag sequencing data than other regions of the gene). Table 4-3 shows the results of this search. These SNPs all fall within the region scanned by SSCP, but the only one that could be confirmed was X07868-820G→A (Tadokoro *et al.* 1991). The CGAP SNPs were identified from 2 chromosomes according to the database (i.e. one heterozygous individual) and therefore may not be reliable.

Table 4-3: SNPs identified by searching online databases.
CGAP indicates a Cancer Gene Anatomy Project SNP, WIAF indicates a Whitehead Institute SNP. Where indicated the SNPs were found in dbSNP (with origin identified).

SNP identification	GenBank ID and nucleotide
CGAP-C-62511 (dbSNP)	X07868-54G/T
CGAP-C-60205 (dbSNP)	X07868-234G/T
WIAF-10657 (dbSNP)	X07868-417A/G
CGAP-C-57610 (dbSNP)	X07868-444T/A
CGAP-C-64673 (dbSNP)	X07868-751C/T
WIAF-1160	X07868-820G/A
CGAP-C-15529	X07868-3545C/T
WIAF-1245	L15440-4462A/T

4.2.5 Summary of identified polymorphisms in *IGF2*

The confirmed sequence variants in *IGF2* are shown in Table 4-4. Normally about 2/3 of SNPs involve the transition C(G)↔T(A) (Wang *et al.* 1998). In this set of SNPs 81% (13/16) involve this transition. The three transversions (T(A)↔A(T), A(T)↔C(G) and C(G)↔G(C)) normally occur at similar levels to each other, and here one of each is observed. About 24% of SNPs usually occur within a CpG dinucleotide (Wang *et al.* 1998); here 31% of SNPs are within CpG dinucleotides.

SNPs published by others have been renamed here according to the SNP naming conventions of this project. Using the identifiers of Lucassen *et al* and Owerbach *et al*, which are numbered according to the insulin gene (*INS*) transcription start site (with no reference sequence), the SNPs could not be identified on GenBank sequence L15440. Primer sequences from their respective papers were used to identify the locations of some SNPs, and the locations of the others were calculated relative to those. In addition, the locations of the polymorphisms on a consensus sequence (see section 4.3.5) are shown in Table 4-4 to allow comparison of the distances between loci. Figure 4-14 in section 4.3.5 shows these locations graphically.

Table 4-4: Confirmed sequence variants in the *IGF2* gene.
Published polymorphisms either confirmed by SSCP or by genotyping (ARMS assay)
[a] SNP ID in publication
[b] Location in a single consensus sequence based on GenBank sequence AC006408

GenBank Sequence ID	Nucleotide	Polymorphism	Source	Consensus location [b]
L15440	6815	A→T	+2331[a] (Lucassen *et al.* 1993)	6836
L15440	7368	G→A	+3123[a] (Owerbach & Gabbay 1993)	7392
L15440	7686	G→A	+3201[a] (Lucassen *et al.* 1993)	7710
L15440	8065	C→T	+3580[a] (Lucassen *et al.* 1993)	8090
L15440	8173	C→T	+3688[a] (Lucassen *et al.* 1993)	8198
Y13633	1156	T→C	SSCP scan	17614
Y13633	1252	T→C	~+11000(*AluI*)[a](Lucassen *et al.* 1993)	17710
Y13633	2482	A→C	DHPLC scan	19181
Y13633	2722	C→T	DHPLC scan	19421
U80851	5345	Poly T	DHPLC scan	23275→23288
X07868	266	C→T	SSCP scan	32564
X07868	556	Poly C	SSCP scan	32854→32864
X07868	820	G→A	*ApaI*[a] (Tadokoro *et al.* 1991)	33118
X07868	1120-1822	Complex $(CA)_n$ repeat polymorphism	(Rainier *et al.* 1993)	33418→34128
X07868	1926	C→G	SSCP scan	34232
X07868	2207	C→T	DHPLC scan	34514
X07868	In sequence between 3750&3751	A→G	SSCP scan	36077
X07868	3944	G→A	SSCP scan	36330
X07868	4085	G→A	SSCP scan	36470

The allele frequencies for each of the SNPs as estimated from either SSCP or ARMS assay results are shown in Table 4-5. Note that SSCP results were from individuals selected for X07868-820G→A genotype, and are probably not therefore representative unless not in

linkage disequilibrium with X07868-820G→A. For example X07868-4085G→A was in apparently very strong linkage disequilibrium with X07868-820G→A with no rare homozygotes (11 common) in the X07868-820G group and only two common homozygotes (9 rare) in the X07868-820A group.

Table 4-5: Frequencies of SNPs in the *IGF2* gene
Frequencies of *IGF2* SNPs calculated from [a] SSCP samples (<40 samples) or [b] genotyping (>1400 samples). p = common allele frequency, q = rare allele frequency.

Polymorphism	p	q	Number
L15440-6815 A→T	0.76	0.24	2394 [b]
L15440-7368 G→A	0.51	0.49	35 [a]
L15440-7686 G→A	0.95	0.05	37 [a]
L15440-8065 C→T	0.79	0.21	31 [a]
L15440-8173 C→T	0.66	0.34	1458 [b]
Y13633-1156 T→C	0.52	0.48	1567 [b]
Y13633-1252 T→C	0.53	0.47	2209 [b]
Y13633-2482A→C	0.52	0.48	2101 [b]
Y13633-2722 C→T	0.72	0.28	2037 [b]
X07868-266 C→T	0.90	0.10	2213 [b]
X07868-820 G→A	0.73	0.27	2560 [b]
X07868-1926 C→G	0.71	0.29	1872 [b]
X07868-2207 C→T	0.95	0.05	2134 [b]
X07868-3750ex A→G	0.89	0.11	2154 [b]
X07868-3944 G→A	0.64	0.36	28 [a]
X07868-4085 G→A	0.50	0.50	26 [a]

4.2.6 Mapping *IGF2*

To enable the mapping of newly identified SNPs in the context of the sequence of *IGF2* it was necessary to assemble the available sequence into a unique consensus, and characterise the components of the gene. The draft sequence for this region of chromosome 11 recently became available as GenBank accession AC006408 (an antisense sequence). This was used as the root sequence against which to assemble informative *IGF2* sequences previously available in GenBank (Figure 4-11).

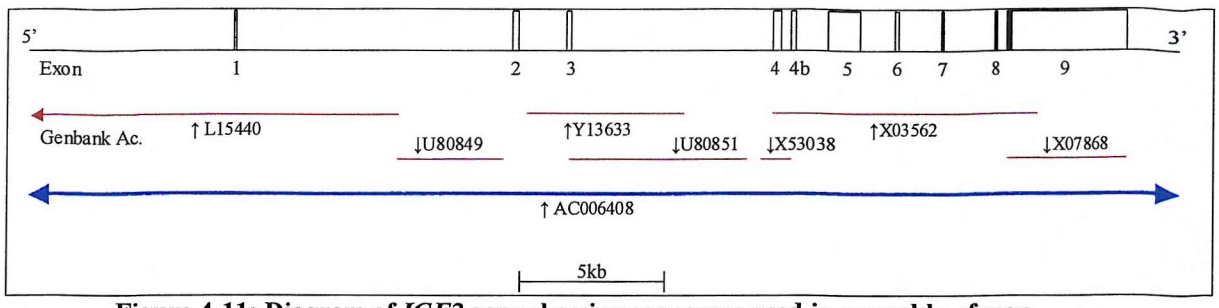


Figure 4-11: Diagram of *IGF2* gene showing sequences used in assembly of map

AC006408 extended upstream about 6kb (nearly as far as L15440, not shown in Figure 4-11) and downstream about 45kb. Information on exon and feature location from each of the GenBank sequences was entered into the consensus using "MapDraw" from the "DNAStar" package. The consensus sequence and the locations of exons were then exported for further analysis.

Custom designed scripts were used to determine GC content and the CpG ratio in 200 and 500bp windows. Smaller windows provide more information but also more noise, so two window sizes were used to optimise the information. Microsoft Excel was used to plot these data on graphs. In addition the locations of the features were entered into Microsoft Excel and plotted on the graphs. Figure 4-12 shows these data for *IGF2* and the region of chromosome 11 immediately upstream. Noticeably high GC contents and CpG ratios are observed between nucleotides 20000 and 28000. The thick red lines at the top of the graph show putative CpG islands calculated from the CpG ratio and GC content. A large CpG island covers promoters 2 to 4 of *IGF2* (indicated by downward pointing arrows), including exons 4, 4b and 5. The same data for the entire consensus sequence is shown in Figure 4-13.

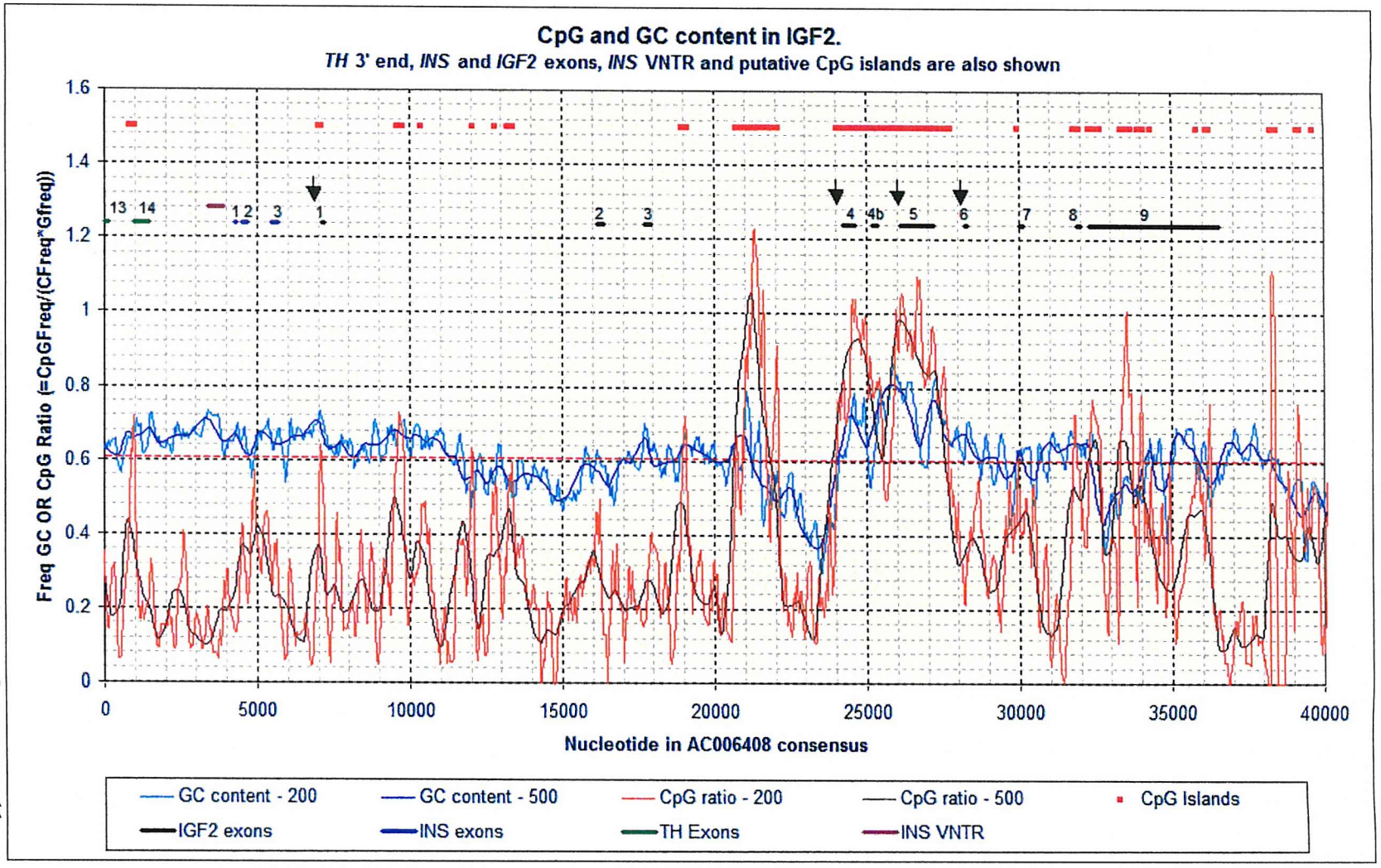


Figure 4-12: Map of the *IGF2* gene and upstream region on chromosome 11

Map of the *IGF2* gene and upstream region on chromosome 11
Shown are location of *INS*, the 3' end of *TH* and all of *IGF2*. The plots are CpG ratio and GC content in both 200 and 500bp windows. Black arrows indicate *IGF2* promoters.

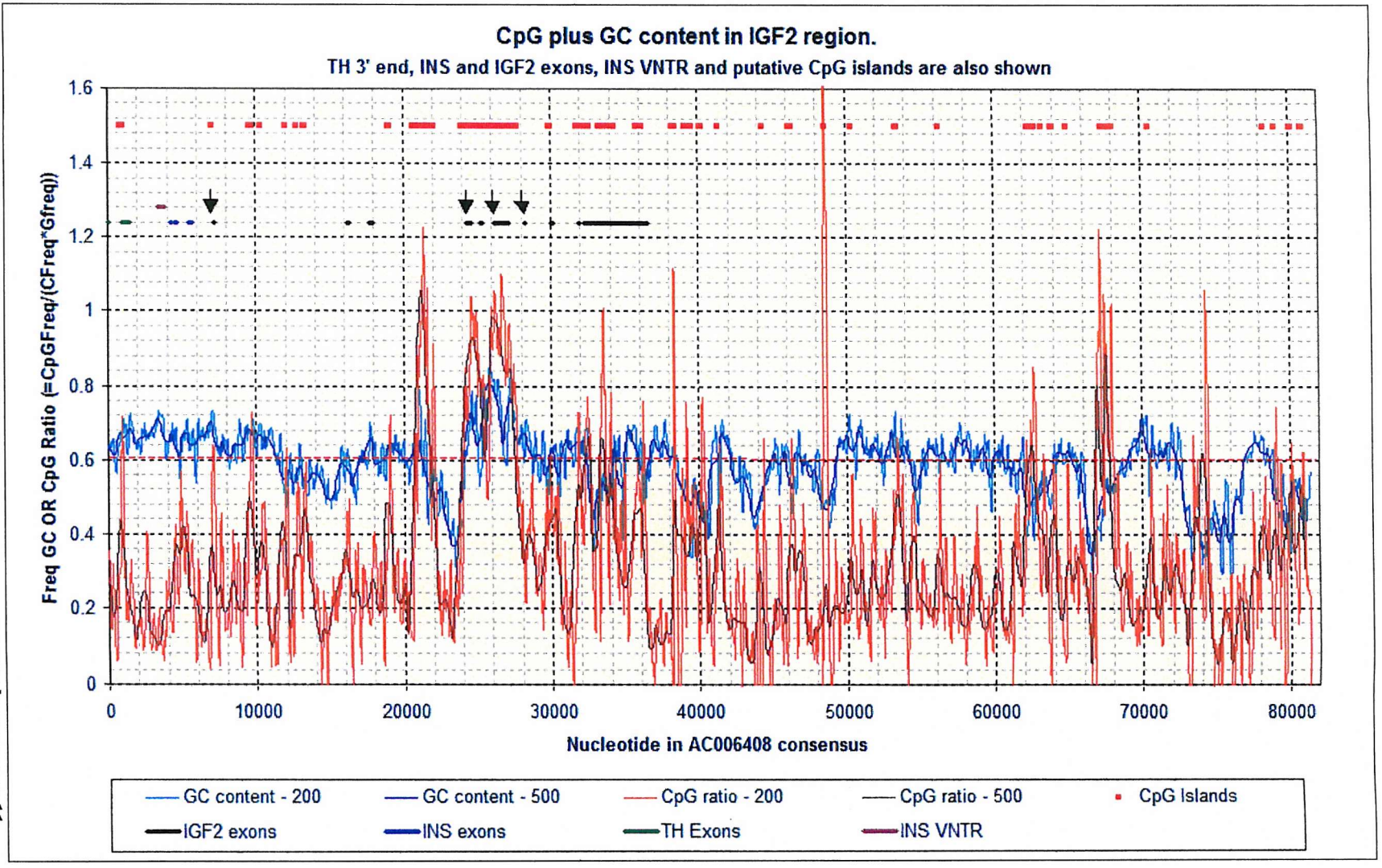


Figure 4-13: Map of the full AC006408 consensus sequence on chromosome 11.

Map of the full AC006408 consensus sequence on chromosome 11.
Shown are location of *INS*, the 3' end of *TH* and all of *IGF2*. The plots are CpG ratio and GC content in both 200 and 500bp windows. Black arrows indicate *IGF2* promoters.

A map showing the location of polymorphisms in *IGF2* based on the consensus generated from GenBank AC006408 is shown in Figure 4-14. The spacing of loci across the gene is clustered, reflecting the regions scanned.

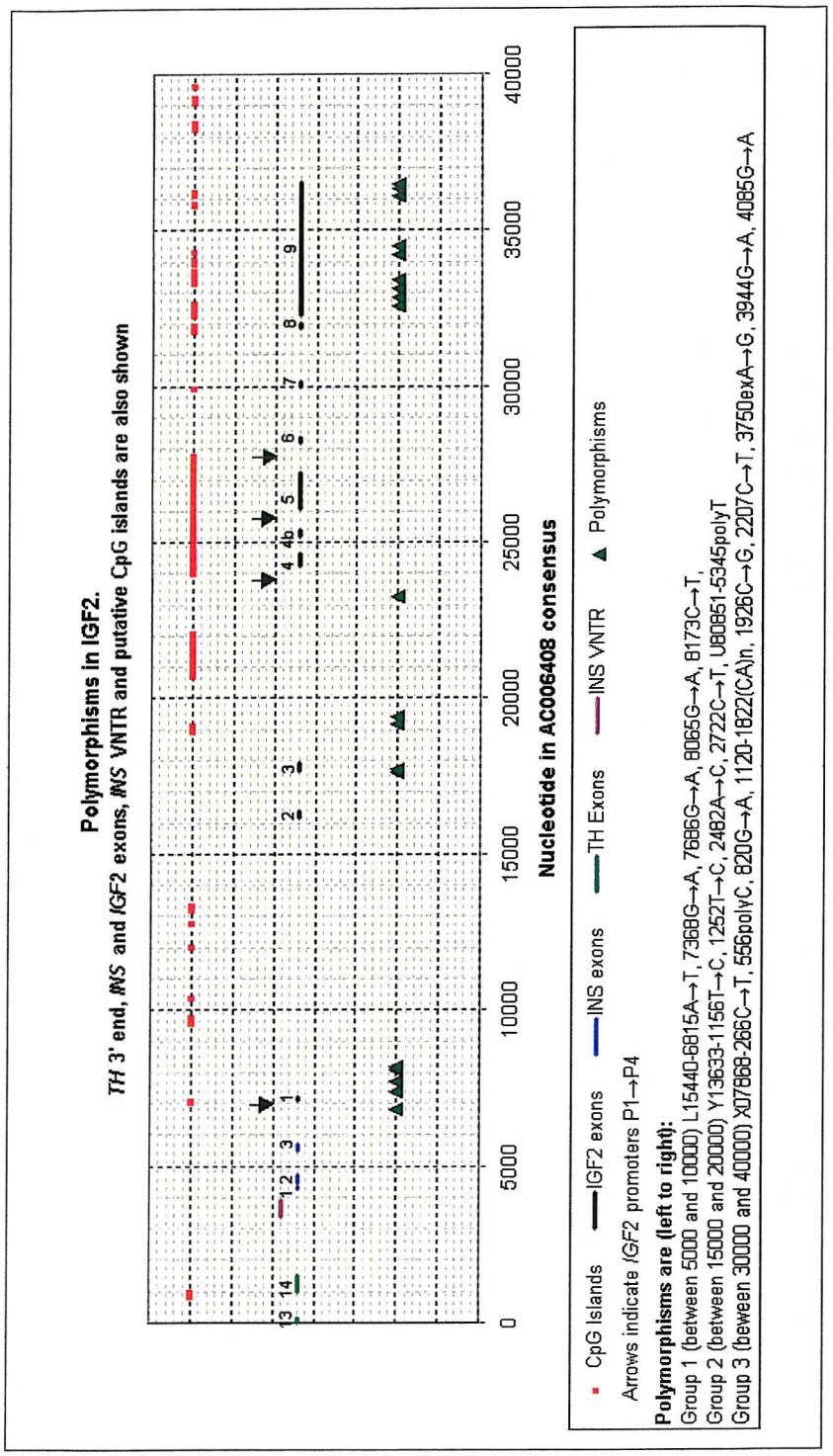


Figure 4-14: Polymorphisms in *IGF2*

Polymorphisms in *IGF2*

SNPs are shown as closed green triangles relative to *IGF2* exons and CpG islands. Polymorphisms are listed in Table 4-4, including references for polymorphisms not identified by this project (all SNPs in group 1, Y13633-1252T→C, X07868-820G→A and X07868-1822(CA)$_n$).

4.3 Discussion

4.3.1 SSCP mutation detection

A number of polymorphisms were identified by the use of single-strand conformation polymorphism (SSCP) analysis. No polymorphisms were found in the coding region of the gene (exons 7, 8 and the first part of exon 9), ruling out the possibility of an amino acid change affecting protein function. However, a number of previously unknown polymorphisms were identified in the 3' untranslated region (exon 9) and within intronic sequence. In addition several previously published polymorphisms were confirmed using this method. No assessment was made of the sensitivity of this method as this has been investigated by others previously (Choy *et al.* 1999; Jones *et al.* 1999; Eng *et al.* 2001), and this would have made the process too time-consuming. However, 7 previously unidentified polymorphisms were identified in 7.5kb of scanned sequence, a frequency of one polymorphism per 1.07kb (one SNP per 1.25kb) which is consistent with the frequency of SNPs in other parts of the genome (Brookes 1999; Wang *et al.* 1998). The use of two temperatures (4°C and 25°C) for SSCP improved sensitivity, as we detected some polymorphisms at only one of those temperatures (i.e. if we had scanned at only one temperature we would have missed some SNPs).

In addition to single nucleotide polymorphisms (SNPs) a homopolymeric tract length polymorphism was identified in the 3'UTR of the gene. Three different alleles for this polymorphism were observed making it potentially more informative than a diallelic SNP. However while this type of polymorphism can be genotyped with a modified ARMS assay protocol (Graack & Kress 1999) this procedure is very difficult and not suitable for high-throughput analysis, limiting its usefulness for this study. Potentially SSCP or DHPLC would be more suitable for genotyping this type of polymorphism, although the risk exists that the pattern for a rare genotype may match that of a common genotype and be mis-called.

Testing 40 individuals with known genotypes for a SNP of interest (X07868-820G→A, which had previously shown an association with BMI (O'Dell *et al.* 1997)) proved fairly effective at indicating the degree of linkage disequilibrium (LD) present between the new polymorphism and X07868-820G→A. However, the numbers were small and not always very informative, and this method gave only a poor indication of the LD between pairs of newly discovered SNPs (tested by a χ^2 test). Scanning a larger number of randomly

selected individuals would have allowed a more useful analysis of pairwise LD between samples, and would have been useful in excluding less informative SNPs (i.e. those that would duplicate the information provided by other SNPs). This would potentially reduce genotyping costs in DNA, time and resources, while still maximising information.

Regions excluded from the SSCP scan included known polymorphisms: Y13633-1252T→C (Lucassen *et al.* 1993), X07868-820G→A(Tadokoro *et al.* 1991) and the 3'UTR $(CA)_n$ repeat (Rainier *et al.* 1993). It was considered that scanning these regions would be a waste of resources. However, SSCP was carried out with primers immediately adjacent to the polymorphic sites, and as such the regions not scanned were very limited. Exon 5 contained very GC rich sequence that was not amenable to polymerase chain reaction, even with the addition of betaine (Henke *et al.* 1997). Whilst scanning exons increased the chances of finding sequence variants that may have a direct effect on expression or protein function, the aim of the scan was to find a reasonable number of SNPs that could be used in linkage disequilibrium mapping. The lack of SSCP coverage of certain regions was therefore not considered a major problem, as linkage disequilibrium between identified SNPs and any functionally important sequence variants should allow that association to be detected.

4.3.2 DHPLC mutation detection

Denaturing high performance liquid chromatography was chosen as the method for mutation scanning in some of the intronic regions of *IGF2*. Four polymorphisms were identified and characterised. Of these, three (two SNPs and one homopolymeric tract length polymorphism) were found in 4.5 kb of scanned sequence in intron 3, and one was found in close proximity to the endonucleolytic cleavage site in the 3' UTR of *IGF2*. This had not been detected by SSCP because of PCR failure, but was repeated by DHPLC using a sodium chloride based PCR buffer, as potassium ions have been observed to enhance secondary structure in this sequence (Christiansen *et al.* 1994). The intron 3 polymorphism frequency was one polymorphism per 1.5kb (or one SNP per 2.25kb), which is lower than the frequency identified in the SSCP scan of introns, but not inconsistent with the frequency in other regions (Brookes 1999; Wang *et al.* 1998).

4.3.3 *In silico* mutation detection

The *in silico* mutation detection approach proved relatively ineffective. The sequence alignment methods generated a large number of false positives (as confirmed by the SSCP and sequencing results). While this approach was potentially faster than lab-based techniques, the time and resources wasted in checking false positives would make it inefficient.

At the time the online databases were searched the information available seemed to be variable in quality and very limited in scope. Most SNPs were identified in databases by surrounding sequence, rather than by an alignment with a GenBank sequence. This meant that searching by gene name or GenBank accession number was not a viable option. As more genome sequence becomes available and more of it is annotated, SNP data is likely to become easier to find, more reliable and this approach will therefore become more useful. The most effective use of online SNP databases is likely to involve custom programming by researchers to enable them to access the data relevant to them and keep it up to date as new data become available.

4.3.4 Summary of confirmed sequence variants

A fairly comprehensive set of polymorphisms in the *IGF2* gene has been compiled using both published data and polymorphisms identified in the laboratory. The distribution of these SNPs across the *IGF2* gene is not even (Figure 4-14). This is the result of the exclusion of some areas of the gene from mutation scanning. Estimates of linkage disequilibrium vary from 3kb (Kruglyak 1999) to 100kb (Collins *et al.* 1999). It was therefore considered unnecessary to scan the entire gene, and regions were selected on the basis of the likelihood of a SNP within that region being functional (e.g. protein coding sequence). The detection rate for polymorphisms was comparable with other published frequencies (Brookes 1999; Wang *et al.* 1998).

The compilation of data from different sources provided a potentially informative resource for other projects. The time-consuming process of identifying the actual location of SNPs identified by restriction endonuclease name and/or location relative to a gene feature emphasised the importance of consistent and unambiguous nomenclature. This has been attempted by the system of using sequences in GenBank as reference sequences and referring to polymorphism locations relative to those sequences. In addition, Appendix C presents the

polymorphisms in sequence context with 250bp of sequence each side of the locus. This is another consistent, unambiguous system, allowing direct access to the sequence for assay design.

The number of SNPs identified was considered ample for an association study, and the most relevant regions had been scanned for potentially functional polymorphisms. Further mutation detection was therefore not carried out.

4.3.5 Mapping of *IGF2*

The *IGF2* gene is 29440bp in length from the start of exon 1 to the end of exon 9 (excluding promoter 1). Overall the GC content is relatively high, and there are a number of potential CpG islands across the gene, consistent with the imprinting of *IGF2*, believed to be caused by differential methylation at CpG islands (Li *et al.* 1993). The density of CpG rich sequence is highest around the promoter 2→4 region of *IGF2*, covering exons 4, 4b, 5 and 6 (Figure 4-12). This fits with the pattern of expression of *IGF2* from the different promoters: promoter 1 is responsible for bi-allelic expression of *IGF2*, while promoters 2-4 initiate expression only from the paternal allele (Vu & Hoffman 1994), suggesting a role for methylation of the maternal allele around the CpG rich P2→P4 region.

The lack of overlapping contigs around the AC006408 draft sequence prevented further extension of the map to include *H19* and the imprinting control region (ICR) 2kb upstream of *H19* and ~70kb downstream of *IGF2* (Reed *et al.* 2001; Frevel *et al.* 1999). However, this sequence provided the first complete sequence for *IGF2* on which to base a map of the gene (other *IGF2* sequences in GenBank leave several gaps in the gene sequence).

The map provided a useful framework for the polymorphism data, allowing the relative positions of polymorphism to be visualised in the context of features of the gene (Figure 4-14). Actual positions allow relatively precise measures of distance (Table 4-4), although the consensus is unlikely to be completely error-free, and therefore exact distances cannot be assumed.

4.4 Conclusions

Eleven new polymorphisms have been identified: six SNPs and one homopolymeric tract length polymorphism by SSCP and three SNPs and one homopolymeric tract length

polymorphism by DHPLC. *In silico* mutation scanning proved ineffective, but a further seven SNPs and one complex $(CA)_n$ repeat polymorphism were identified in the literature. The total number of polymorphisms in *IGF2* collated from other sources and this project is therefore 19 (Table 4-4). A consensus sequence of the gene has been generated, and a map created based on that sequence allowing all the polymorphisms to be accurately mapped in the context of the sequence for the first time (Figure 4-14).

Chapter 5 : Genetic Epidemiology of NPHSII

5.1 Introduction

5.1.1 The Northwick Park Heart Study II

The DNA samples used in this part of the project were collected as part of the Northwick Park Heart Study II (NPHSII, a prospective study of coronary heart disease - Director George Miller). The NPHSII DNA bank contains 2743 DNA samples that were extracted from healthy men aged 51-62 years (with no selection for disease status). Phenotype data in the database included weight, height, BMI, serum triglycerides, cholesterol and diabetes status.

5.1.2 BMI distribution in Northwick Park Heart Study II

A scatter plot of height versus weight for the NPHSII population (Figure 5-1) shows the range of BMIs within this population. Most individuals fall in the range 20 to 30kgm^{-2}. The 'ideal' range of 20 to 25kgm^{-2} is shown as green data points. Extremely obese individuals fall below the BMI=35 line (red points). The statistics for this population are shown in Table 5-1. The generally accepted definitions of weight are: normal BMI=20→25kgm^{-2}, overweight BMI=25→29kgm^{-2}, obese BMI=30→35kgm^{-2}, severely obese BMI=35→40kgm^{-2} and morbidly obese BMI>40kgm^{-2} (Jeffcoate 1998), although

exact cut-offs and names vary between studies. For the purposes of this study the above cut-offs were used.

Table 5-1: Statistics on NPHSII population

	Weight	Height	BMI
Mean	80.51kg	1.74m	26.44kg/m²
Median	79.80kg	1.75m	26.13kg/m²
Mode	70.00kg	1.74m	24.88kg/m²
Min	44.50kg	1.48m	14.92kg/m²
Max	137.40kg	2.08m	49.48kg/m²
Range	92.90kg	0.60m	34.56kg/m²
St.Deviation	11.73kg	0.07m	3.53kg/m²

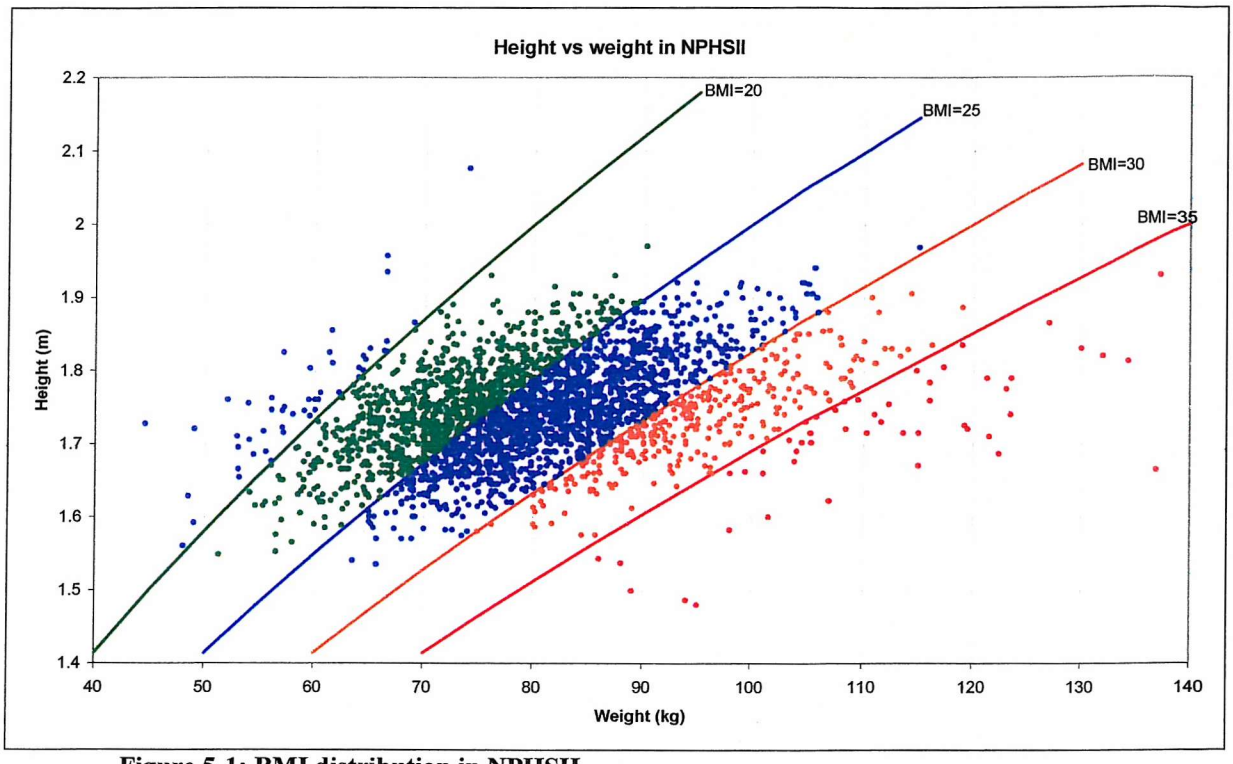


Figure 5-1: BMI distribution in NPHSII.
BMI = weight(kg)/height(m)^2

5.1.3 Polymorphisms

The polymorphisms selected for genotyping are shown in Table 5-2. In addition to these polymorphisms three others had been previously genotyped in the NPHSII population (O'Dell *et al.* 1997; Gaunt *et al.* 2001; Gu *et al.* 2002), and these were included in the data analysis for completeness. These were L15440-6815A→T (Lucassen *et al.* 1993), Y13633-1252T→C (*AluI* restriction site) (Lucassen *et al.* 1993) and X07868-820G→A (*ApaI* restriction site) (Tadokoro *et al.* 1991).

The majority of the polymorphisms were single-nucleotide substitutions (8 SNPs, plus 3 previously genotyped). However, there was also a repeat polymorphism located in the 3'UTR of the *IGF2* gene (Rainier *et al.* 1993), which can be split into upstream and downstream fragments (by restriction digest). There are three common and two rare alleles in the upstream fragment, and two common and one rare alleles in the downstream fragment (Rainier *et al.* 1993).

Table 5-2: Polymorphisms tested in NPHSII

Reference sequence (GenBank ID)	Nucleotide number(s) and sequence	Type of polymorphism	Reference
L15440	8173 C→T	SNP	(Lucassen *et al.* 1993)
Y13633	1156 T→C	SNP	Chapter 3
Y13633	2482 A→C	SNP	Chapter 3
Y13633	2722 C→T	SNP	Chapter 3
X07868	266 C→T	SNP	Chapter 3
X07868	1122-1876 $(CA)_n$ (complex repeat)	Repeat polymorphism	(Rainier *et al.* 1993)
X07868	1926 C→G	SNP	Chapter 3
X07868	2207 C→T	SNP	Chapter 3
X07868	Sequence between 3750 and 3751 A→G	SNP	Chapter 3

Polymorphisms were selected for genotyping on a number of criteria, including location and suitability for ARMS assay design. One previously published SNP from the 5' end of the gene (L15440-8173C→T) was selected, all the SNPs identified in the central part of the gene (introns 2 and 3) and 5 polymorphisms from the 3' UTR (including a $(CA)_n$ repeat polymorphism). Two SNPs in the 3'UTR identified in this project were excluded from genotyping because of the large number already selected in this region.

5.1.4 Objectives

The primary aim of this part of the project was to test the hypothesis that one or more polymorphisms in the *IGF2* gene influence adiposity (as indicated by body mass index) in middle-aged men. The *ApaI* polymorphism in the 3'UTR of the *IGF2* gene is associated with BMI in middle-aged men (O'Dell *et al.* 1997), and potentially marks another site which influences gene expression or protein function. The objectives of this part of the project were to genotype a selection of the SNPs identified in the SNP scan of *IGF2* in a population of middle-aged men. The results should determine the extent to which *IGF2* is involved in weight determination.

Northwick Park Heart Study II (NPHSII) is a large cohort of unrelated individuals, well suited to population association studies of polygenic traits. The nature of obesity prevents case-control analysis unless arbitrary cut-offs are used. The plan was therefore to carry out a large-scale genotyping project, comparing mean BMI between genotype groups (rather than genotype or allele frequencies between phenotype groups).

5.2 Results

5.2.1 *IGF2* $(CA)_n$ repeat polymorphism

This polymorphism required the use of low-throughput ABI genotyping technology, and thus only a selection of samples from NPHSII was genotyped. Forty samples were selected from the upper end of the BMI range, and forty from the lower end. This generated arbitrarily defined "case" and "control" samples.

5.2.1.1 Hardy-Weinberg equilibrium test

A test of conformity to Hardy-Weinberg equilibrium may be used to indicate whether a population is randomly selected, out-breeding and not stratified. It is also often used to indicate the quality of genotyping data. As a test of data quality the assumption is made that the population is in Hardy-Weinberg equilibrium, and if the genotype frequencies for a particular SNP are not in equilibrium, then there are potentially data errors. Table 5-3 shows the Hardy-Weinberg equilibrium test for the genotypes of two fragments of the $(CA)_n$ repeat polymorphism in the 3'UTR of *IGF2*. Neither set of genotype frequencies deviates significantly from Hardy-Weinberg equilibrium.

Table 5-3: Test of Hardy-Weinberg equilibrium in $(CA)_n$ repeat.
Expected numbers were calculated from allele frequencies, then observed numbers compared to expected with a χ^2 test. A *P*-value <0.05 would indicate deviation from Hardy-Weinberg equilibrium

$(CA)_n$ repeat 5' fragment Allele 1=383, 2=398 and 3=402			$(CA)_n$ repeat 3' fragment Allele 1=399, 2=404		
Observed (O)	11	7	Observed (O)	11	30
	12	31		12	30
	13	3		22	10
	22	30			
	23	1	p		0.64
	33	0	q		0.36
p		0.33	Expected (E)	11	28.93
q		0.64		12	32.14
r		0.03		22	8.93
Expected (E)	11	8.00	$\chi2 = \frac{(O-E)^2}{E}$	11	0.04
	12	30.67		12	0.14
	13	1.33		22	0.13
	22	29.39			
	23	2.56	$\Sigma\chi2$		0.31
	33	0.06	*P*-value		0.86
$\chi2 = \frac{(O-E)^2}{E}$	11	0.13			
	12	0.00			
	13	2.08			
	22	0.01			
	23	0.95			
	33	0.06			
$\Sigma\chi2$		2.22			
P-value		0.33			

5.2.1.2 $(CA)_n$ repeat genotypes

The $(CA)_n$ repeat polymorphism in the 3'UTR of the *IGF2* gene was genotyped by PCR and restriction digest (Figure 5-2), resulting in two distinctly fluorescent-labelled fragments. These were resolved on an ABI310 automated genetic analyser (see Methods, Chapter 2). The 6-FAM labelled 5' fragment yielded three alleles within the set of samples analysed, while the HEX labelled 3' fragment yielded only two alleles. Within the sample set analysed, 5 of the 6 possible 5' genotypes were observed and all 3 of the possible 3' genotypes were observed (Table 5-4). 7 of 15 possible combinations of genotypes (between the two fragments) were observed.

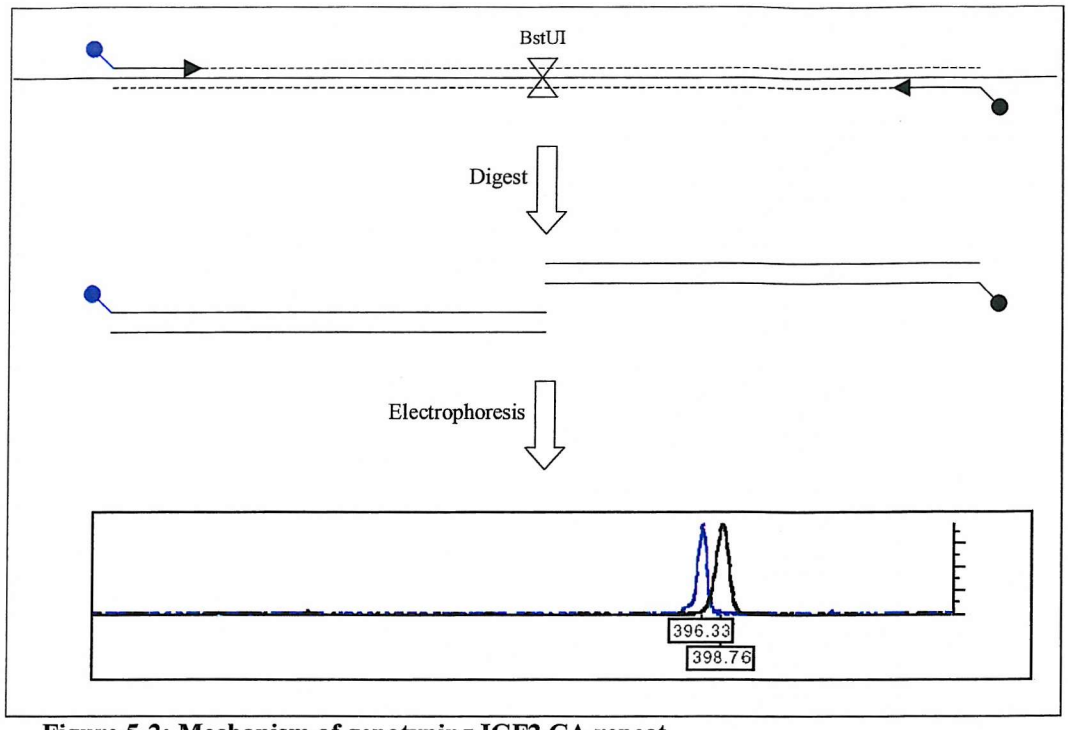


Figure 5-2: Mechanism of genotyping IGF2 CA repeat

Table 5-4: Genotype counts for CA repeat

5' end		3' end		Combined genotypes	
Genotype	Count	Genotype	Count	Genotype	Count
383-383	6	399-399	30	383-383-404-404	6
383-398	31	399-404	33	383-398-399-399	1
383-402	3	404-404	10	383-398-399-404	28
398-398	29	Total	73	383-402-404-404	3
398-402	1			398-398-399-399	28
Total	70			398-398-399-404	1
				398-402-399-404	1
				Total	68

Figure 5-3 shows example ABI chromatograms of each of the five observed 5' CA repeat genotypes. The chromatograms shown are a sub-section of the full data-range recorded by the machine. The x-axis represents scan number (implying size) and the y-axis indicates fluorescence intensity. Internal sizing was achieved by use of a Tamra-labelled size standard (not shown). There are at least six possible genotypes, but the 402-402 homozygote was not observed in this sample set. Figure 5-4 shows the example chromatograms of the three 3' CA repeat genotypes. The relative intensity of the HEX-labelled 3' fragment was greater than that of the 6-FAM labelled 5' fragment, but both were well within the detectable range. Peak sizing and binning was straightforward, as the smallest size difference to be resolved was 4 nucleotides (alleles 398 and 402 in the 5' fragment). Allele sizes were assigned according to the sizes identified by Ranier *et al* (1993) rather than the actual size identified on the

ABI310. This is for consistency – absolute sizing is fairly arbitrary and tends to vary between methods depending on the standard used.

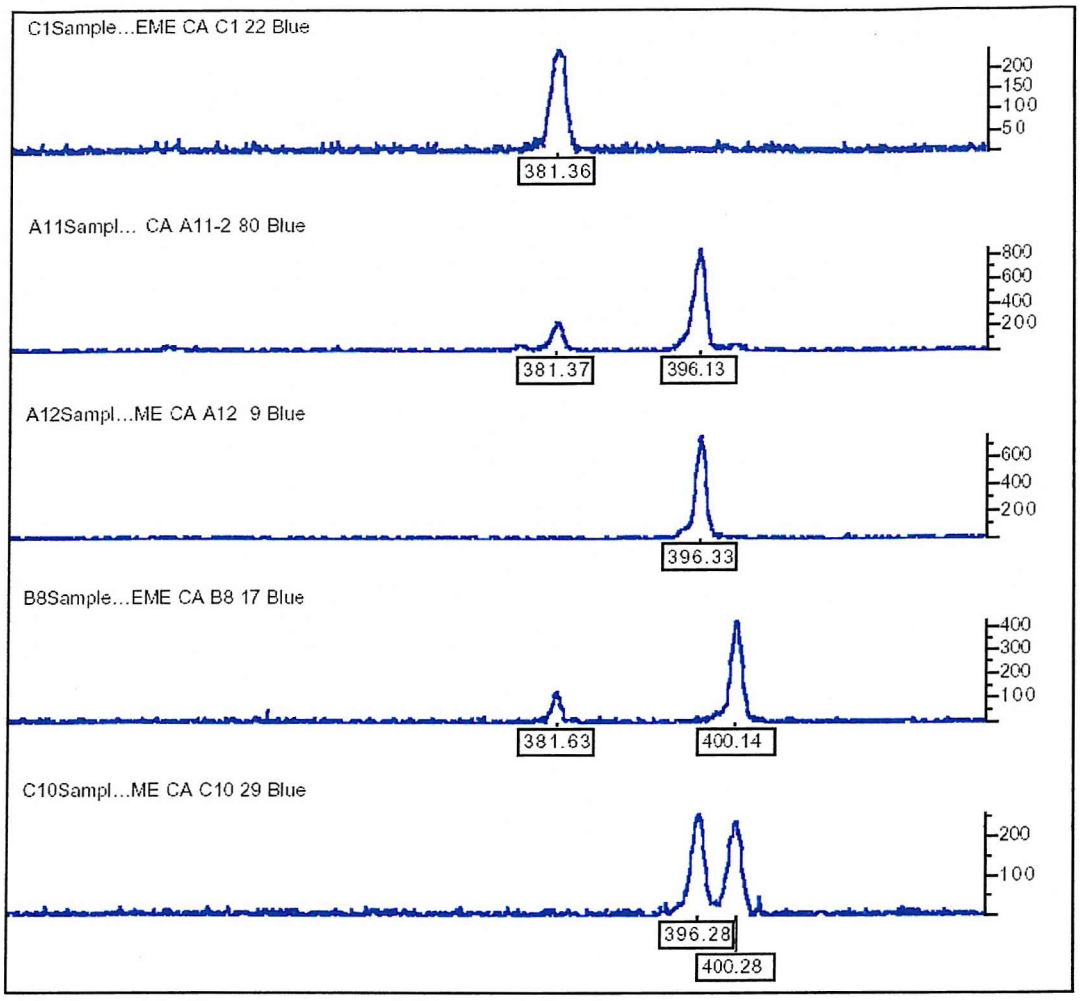


Figure 5-3: Chromatograms of CA repeat 5' end polymorphism

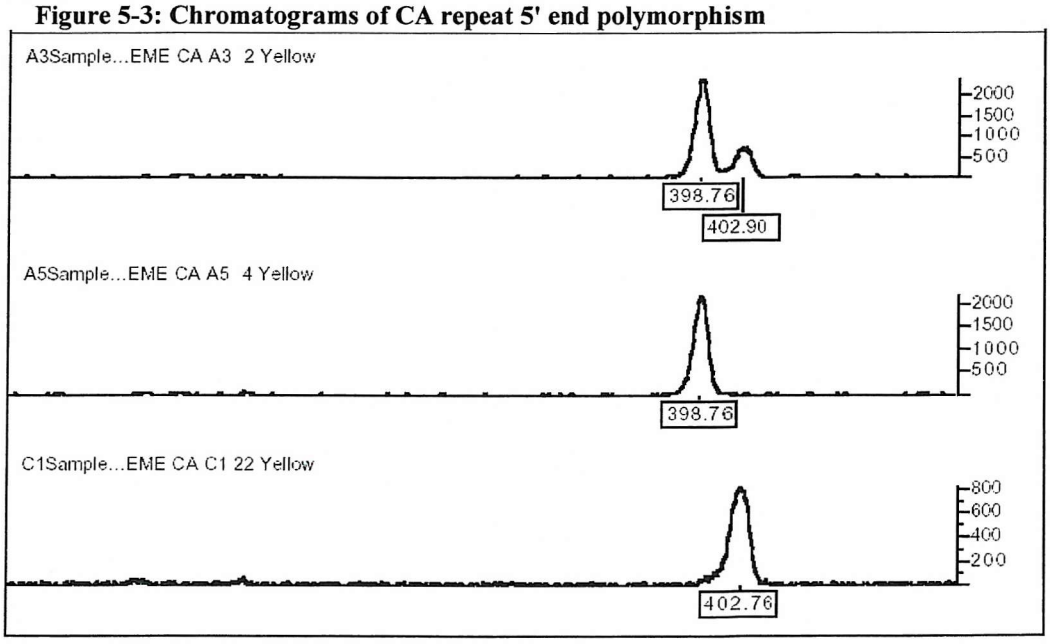


Figure 5-4: Chromatograms of CA repeat 3' end polymorphism

Comparison of mean BMI for the different genotype groups in each of the two parts of this repeat polymorphism reveals trends in both fragments. Table 5-5 shows the mean BMI's for categories of genotype at the 5' end of the CA repeat. The 383 homozygotes showed a normal BMI of 22kgm^{-2}, with both 398 and 402 alleles tending to associate with increased BMI (both as heterozygotes and homozygotes). However, this was not significant. Table 5-6 shows the results of an analysis of variance (ANOVA) of these data.

Table 5-5: CA repeat 5' mean BMIs

Genotype	Number	Mean BMI	Standard Deviation
383-383	7	22.05	7.41
383-398	31	27.54	10.23
383-402	3	30.85	10.07
398-398	30	29.76	10.12
398-402	1	39.16	-
Total	72	28.23	10.02

Table 5-6: CA repeat 5' genotyping ANOVA

	Sum of Squares	df	Mean Square	F	Significance
Between Groups	491.97	4	122.99	1.241	0.302
Within Groups	6642.00	67	99.134		
Total	7133.97	71			

Table 5-7 shows the mean BMI's for categories of genotype at the 3' end of the CA repeat. 404 homozygotes have a high normal BMI of 24.23, while the 399 allele tends to associate with increased BMI. This is also not significant by ANOVA (Table 5-8).

Table 5-7: CA repeat 3' mean BMIs

Genotype	Number	Mean BMI	Standard Deviation
399-399	31	29.52	10.39
399-404	33	28.73	9.92
404-404	11	24.23	8.48
Total	75	28.40	9.96

Table 5-8: CA repeat 3' genotyping ANOVA

	Sum of Squares	df	Mean Square	F	Significance
Between Groups	233.41	2	116.70	1.182	0.312
Within Groups	7106.90	72	98.71		
Total	7340.31	74			

Figure 5-5 shows the numbers of each genotype for both ends of the CA repeat polymorphism, separated by BMI as "heavy" (BMI>35) and "light" (BMI<19). The deviations from the expected values of equal numbers in both groups are shown, but these are

not statistically significant by χ^2 test. Allele frequencies in each group were also calculated (Table 5-9), and were not significantly different by χ^2 test (p=0.12).

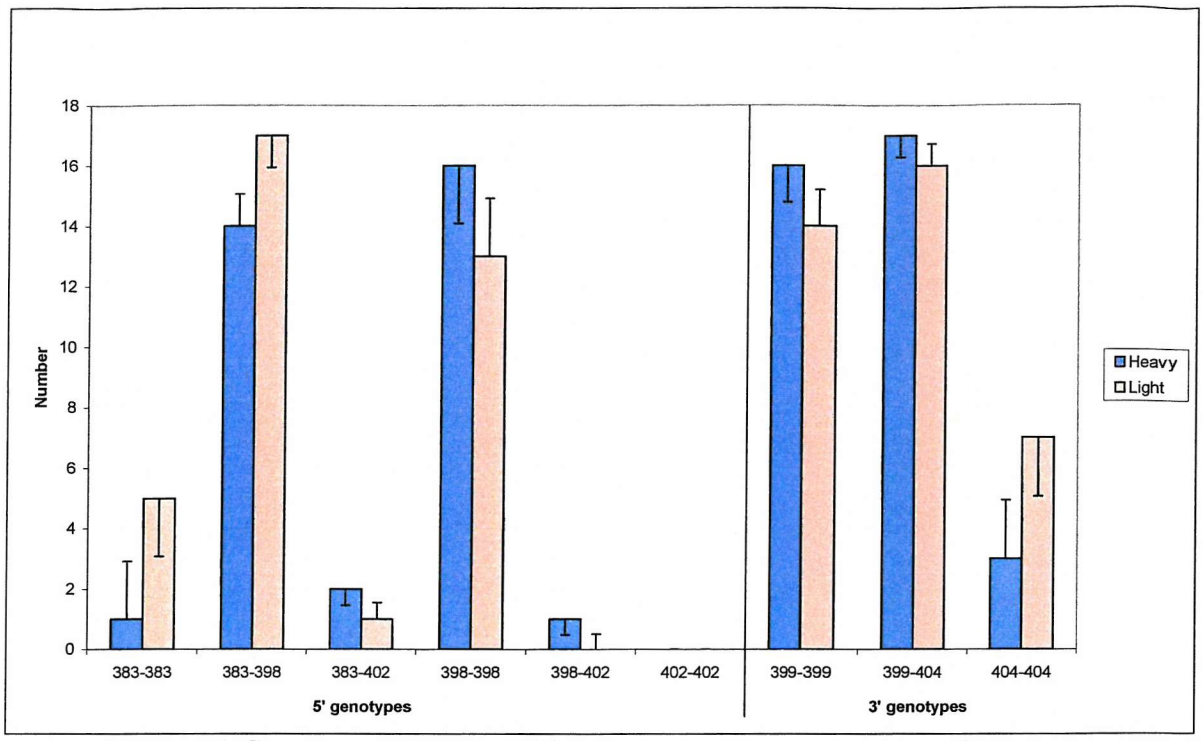


Figure 5-5: Genotype numbers by BMI group for the CA repeat polymorphism
Deviation bars show deviation of observed from expected number (assuming no genetic effect) - this is not significant. Note that numbers in heavy and light groups are not exactly equal.

Table 5-9: Allele frequencies in CA repeat
Allele frequencies distinguished by lean (BMI<20kgm^2) or obese (BMI>35kgm^2)

	Lean group n=40	Heavy group n=40
5' - 383	0.41	0.26
5' - 398	0.58	0.70
5' - 402	0.01	0.04
3' - 399	0.58	0.69
3' - 404	0.42	0.31

The $(CA)_n$ repeat polymorphism is in close physical proximity to the X07868-820G→A SNP (*ApaI*) (302bp). The positive association between *ApaI* and BMI (O'Dell *et al.* 1997) is therefore likely to be marked by this site. Table 5-10 shows haplotype frequencies estimated from the data by "EH" (Xie & Ott 1993). Estimated haplotype frequencies are significantly different from those expected under the null hypothesis of no allelic association χ^2=187.17 (11df, p<0.001). 88.75% of haplotypes are either 1-397-399 (*ApaI*-5'$(CA)_n$-3'$(CA)_n$) or 2-382-404. Only 31.75% of haplotypes would fall into those two categories if no linkage disequilibrium existed. Figure 5-6 is a graphical representation of these data, showing the haplotype frequencies estimated from the genotype data as blue bars, and the haplotype

frequencies that would be expected under the null hypothesis of no allelic association as red bars.

Table 5-10: Estimated haplotype frequencies for *ApaI* and $(CA)_n$ repeat
Estimated haplotype frequencies for ApaI, 5' (CA)n and 3' (CA)n repeat fragments calculated using "EH" (Xie & Ott 1993). Allele frequencies in this set: *ApaI* [1 = 0.6838, 2 = 0.3162], 5' $(CA)_n$ [382 = 0.3309, 397 = 0.6397, 401 = 0.0294], 3' $(CA)_n$ [399 = 0.6397, 404 = 0.3603]. χ^2 for hypothesis of allelic association = 187.17 (11df, $p<0.001$).

ApaI	5' $(CA)_n$	3' $(CA)_n$	Expected frequency under null hypothesis (no LD)	Estimated frequency
1	382	399	0.144743	0.000000
1	382	404	0.081522	0.045200
1	397	399	0.279837	0.609212
1	397	404	0.157609	0.000000
1	401	399	0.012866	0.000000
1	401	404	0.007246	0.029411
2	382	399	0.066924	0.007353
2	382	404	0.037693	0.278329
2	397	399	0.129387	0.023141
2	397	404	0.072873	0.007353
2	401	399	0.005949	0.000000
2	401	404	0.003350	0.000000

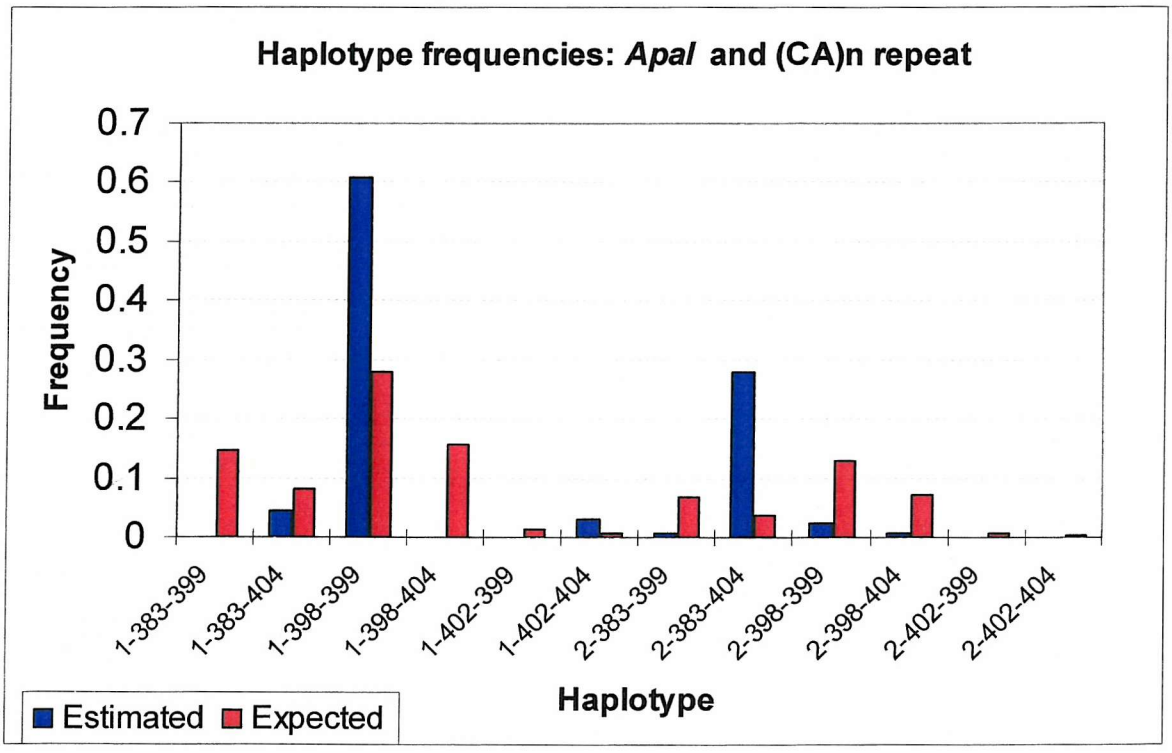


Figure 5-6: Graph of haplotype frequencies: ApaI and $(CA)_n$ repeat
Estimated and expected (under no association) haplotype frequencies for *ApaI*, 5' $(CA)_n$ and 3' $(CA)_n$ repeat fragments calculated using "EH" (Xie & Ott 1993). The most common haplotype is 1-397-399 where 1 is the *ApaI* allele, 397 is the 5' $(CA)_n$ repeat allele and 399 is the 3' $(CA)_n$ repeat allele.

5.2.2 SNP genotypes

5.2.2.1 Hardy-Weinberg equilibrium test

As in section 5.2.1.1, a test of Hardy-Weinberg equilibrium was carried out on all the SNP genotypes as a test of data quality. The assumption is made that the population is in Hardy-Weinberg equilibrium, and if the genotype frequencies for a particular SNP are not in equilibrium, then there are potentially data errors. Table 5-11 shows these data. All eight SNPs are in Hardy-Weinberg equilibrium.

Table 5-11: Test of Hardy-Weinberg equilibrium in genotyped SNPs.
Expected numbers calculated from allele frequencies, then observed numbers compared to expected with a χ^2 test. A *P*-value <0.05 would indicate deviation from Hardy-Weinberg equilibrium

		L15440 8173 C→T	Y13633 1156 T→C	Y13633 2482 A→C	Y13633 2722 C→T	X07868 266 C→T	X07868 1926 C→G	X07868 2207 C→T	X07868 3750 A→G
Observed (O)	11	621	403	537	1041	1806	933	1916	1696
	12	689	812	1091	850	381	786	214	439
	22	148	352	473	146	26	153	4	19
p (freq 1)		0.662	0.516	0.515	0.720	0.902	0.708	0.948	0.889
q (freq 2)		0.338	0.484	0.485	0.280	0.098	0.292	0.052	0.111
Expected (E)	11	639.3	417.6	557.7	1055.0	1801.1	939.2	1917.7	1703.4
	12	652.2	782.6	1049.5	821.8	390.6	773.5	210.4	424.1
	22	166.3	366.6	493.7	160.0	21.1	159.2	5.7	26.4
$\chi2 = \frac{(O-E)^2}{E}$	11	0.527	0.515	0.771	0.187	0.013	0.042	0.002	0.032
	12	2.068	1.099	1.639	0.962	0.238	0.202	0.060	0.517
	22	2.027	0.587	0.871	1.235	1.097	0.245	0.545	2.078
$\Sigma\chi2$		4.622	2.201	3.281	2.384	1.347	0.489	0.606	2.628
P-value	2df	0.099	0.333	0.194	0.304	0.510	0.783	0.738	0.269

5.2.2.2 Association analysis by one-way ANOVA

The genotype data for *IGF2* SNPs were placed in a database with phenotype data. Relevant phenotype data for the Northwick Park Heart Study were: height (metres), weight (kilograms) and Body Mass Index (BMI, weight x height^{-2} in kgm^{-2}). A one-way analysis of variance (ANOVA) was used to assess the significance of the differences in mean phenotype measurements with respect to genotype. Mean BMI for homozygous wild-types, heterozygotes and homozygous mutants were determined, and the difference between the means tested for significance. These data were combined with data for X07868-820 (*ApaI*) (O'Dell *et al.* 1997), Y13633-1252 and L15440-6815 (Gaunt *et al.* 2001; Gu *et al.* 2002) which pre-dated this project. The ANOVA results are shown in Table 5-12.

Table 5-12: Association between *IGF2* SNPs and phenotypes in NPHSII.
* indicates statistically significant *P*-value. 11 refers to common homozygotes, 12 to heterozygotes and 22 to rare homozygotes. GenBank accessions are shown for the sequences in which these SNPs were found. Each genotype column shows number (n), mean BMI (mean) and standard deviation (s.d.)

SNP ID = GenBank ref + nt number	11 n, mean, s.d.	12 n, mean, s.d.	22 n, mean, s.d.	*P*-value of ANOVA	Number
L15440-6815	1377, 26.61, 3.50	906, 26.61, 3.42	111, 25.28, 3.60	0.00012 *	2394
L15440-8173	621, 26.37, 3.67	689, 26.44, 3.52	148, 26.55, 3.38	0.840	1458
Y13633-1156	403, 26.15, 3.43	812, 26.30, 3.46	352, 26.83, 3.58	0.017 *	1567
Y13633-1252	629, 26.55, 3.29	1069, 26.41, 3.63	511, 26.36, 3.49	0.620637	2209
Y13633-2482	537, 26.37, 3.41	1091, 26.40, 3.50	473, 26.58, 3.36	0.560297	2101
Y13633-2722	1041, 26.7, 3.48	850, 26.42, 3.44	146, 26.33, 3.48	0.931	2037
X07868-266	1806, 26.43, 3.57	381, 26.61, 3.23	26, 26.47, 3.00	0.653244	2213
X07868-820	1362, 26.40, 3.30	1019, 26.20, 3.40	179, 25.30, 3.20	0.0004 *	2560
X07868-1926	933, 26.56, 3.44	786, 26.44, 3.63	153, 25.58, 3.16	0.0062 *	1872
X07868-2207	1916, 26.44, 3.52	214, 26.25, 3.52	4, 29.15, 3.00	0.228781	2134
X07868-3750ex	1696, 26.41, 3.55	439, 26.65, 3.42	19, 25.96, 3.16	0.358891	2154

The first column shows an identifier for each SNP, which comprises a GenBank accession number for a reference sequence and a nucleotide number. Columns two, three and four contain genotype numbers, mean BMI and standard deviation for each of the three genotypes at each SNP. Column five contains the *P*-value of the ANOVA. Column six indicates the total number of samples for which both genotype and BMI data were available. Four SNPs showed positive associations (i.e. significant difference between the mean BMI between genotype groups). These were L15440-6815A→T (Gaunt *et al.* 2001; Gu *et al.* 2002), X07868-820G→A (O'Dell *et al.* 1997), Y13633-1156T→C and X07868-1926C→G. The latter two were discovered and genotyped in this project.

5.2.2.3 BMI distributions by genotype

The BMI distributions for each of the significantly associated SNPs in *IGF2* are shown in Figure 5-7. For each SNP there are two graphs. The first graph shows the number

of individuals in each BMI category (2kgm^{-2} divisions) separated into the three genotypes. The distributions are typical skewed BMI distributions with slight variations in the position of the peak (e.g. for L15440-6815 in panel A, the 11 peak is at 28kgm^{-2}, while the 22 peak is at 26kgm^{-2}). The second graph for each SNP shows the contribution of each genotype to the total number in each BMI category. In L15440-6815 (A), X07868-820 (C) and X07868-1926 (D) the trend is for decreased contribution of rare homozygotes (designated 22) to higher BMI categories relative to common homozygotes (designated 11). The opposite trend is seen in Y13633-1156 (B) where rare homozygotes (22) contribute more to the higher BMI categories than to the lower categories.

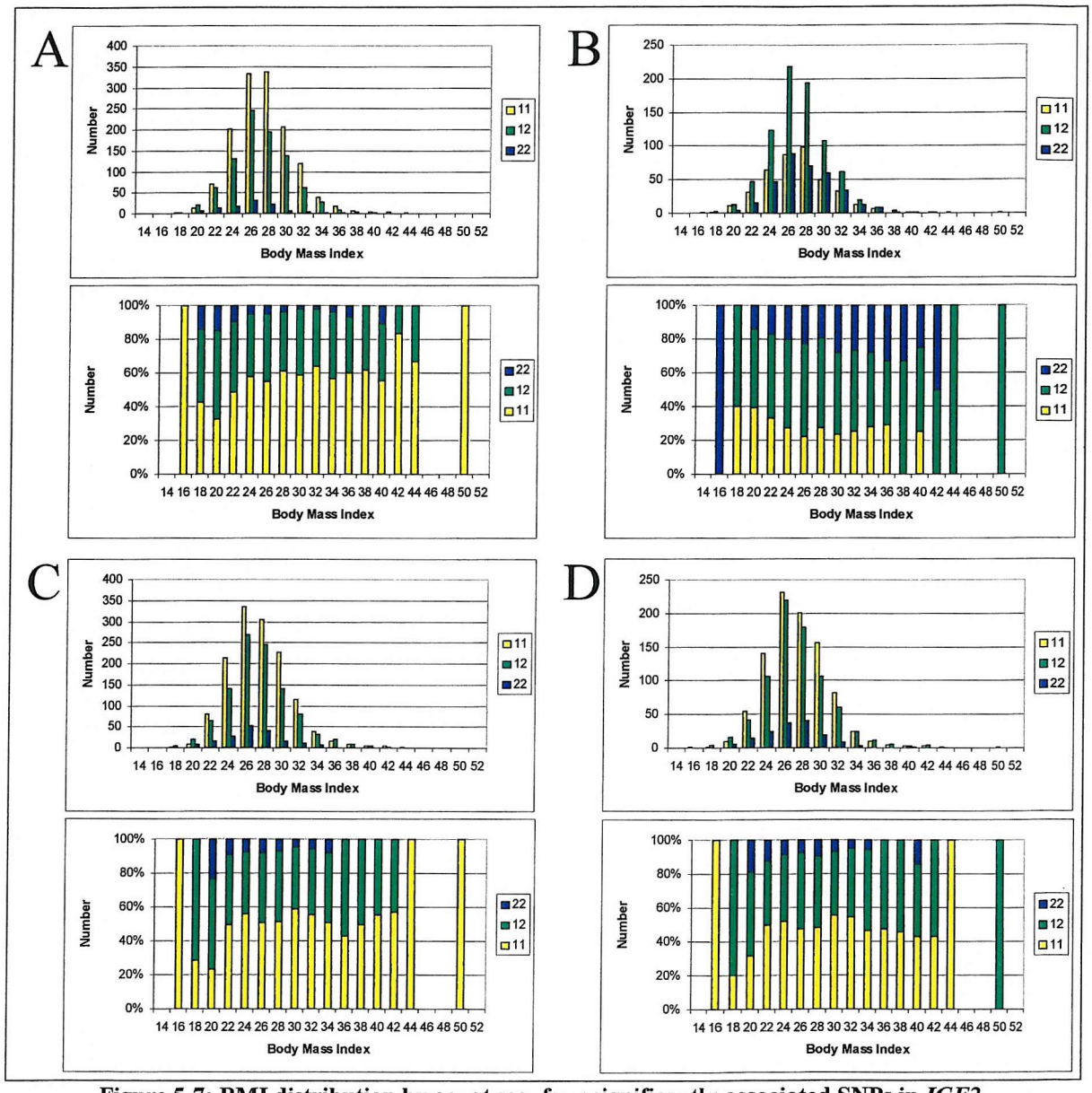


Figure 5-7: BMI distribution by genotype- four significantly associated SNPs in *IGF2*.
For each SNP the upper graph shows absolute numbers in each category, while the lower graph shows the percentage of individuals of each genotype in each category.
(A) L15440-6815, (B) Y13633-1156, (C) X07868-820 (*ApaI*) and (D) X07868-1926.

5.2.2.4 Regression analysis of BMI on significantly associated SNPs

The independence of the four significantly associated SNPs was tested by stepwise multiple regression ANOVA. Figure 5-8 shows a graphical representation of the results in Table 5-13. The sequential R^2 values (where R^2 is the regression sum of squares) indicate the proportion of variance accounted for by each SNP above that previously entered into the model (L15440-6815*A/T* entered the model first as it was the most significant term). L15440-6815*A/T* accounted for 1.03% of variance in BMI. Addition of Y13633-1156*T/C* into the model accounted for another 0.78% of variance in BMI. Addition of X07868-820*G/A* contributed another 0.44%. X07868-1926*C/G* did not enter the model, as it did not account for any variance above the terms already in the model.

The partial R^2 values indicate the independent contributions of each of the SNPs to variance in BMI. Y13633-1156*T/C* accounts for the most variance of the three SNPs (0.91%), with L15440-6815*A/T* accounting for 0.48% of variance and X07868-820*G/A* accounting for 0.44% of variance in BMI.

Table 5-13: Stepwise regression model for BMI.
The model was fitted using forward (step-up) selection.
[a] The β-coefficient (with standard error) is the difference in BMI between the common (11) homozygotes and each of the other groups (12 and 22)
[b] *P* values indicate the association of each term with BMI after adjustment of all other terms in the model
[c] The partial R^2 indicates the proportion of variance explained by each term above that already explained by the other terms in the model
[d] The sequential R^2 indicates the proportion of variance explained by each term above that already explained by the previous term entered in the model

SNP	Genotype	β(SE)[a]	*P* value[b]	Partial R^{2c}	Sequential R^{2d}
L15440-6815 *A/T*	12	-0.53(0.22)			
	22	-0.64(0.46)	0.040	0.48	1.03
Y13633-1156 *T/C*	12	-0.21(0.25)			
	22	0.64(0.30)	0.002	0.91	0.78
X07868-820 *G/A*	12	-0.25(0.21)			
	22	-0.92(0.39)	0.050	0.44	0.44

(Modified from Gaunt *et al.* 2001, Human Molecular Genetics 10(14):1491-1501 with permission from Oxford University Press.)

The total variance in BMI accounted for by the three SNPs is 1.83%. While L15440-6815*A/T* initially appears the most significantly associated SNP, this significance diminishes as the effect of each of the other SNPs is taken into account. The three SNPs all independently contribute to variance in BMI, with Y13633-1156*T/C* having the largest effect.

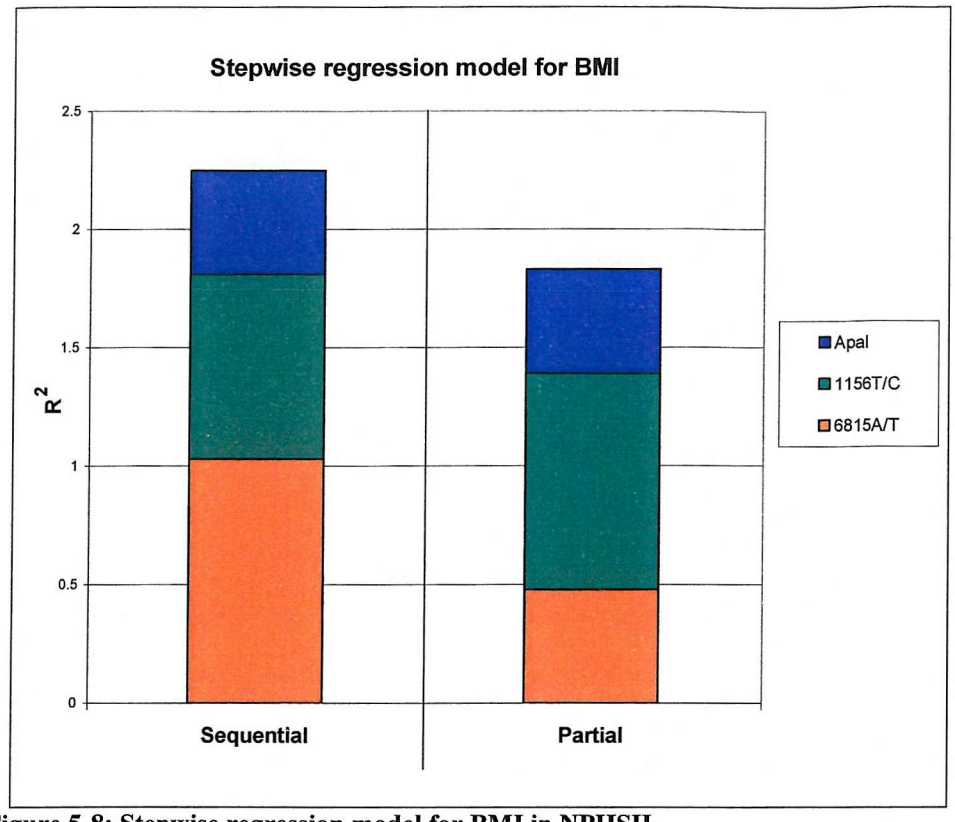
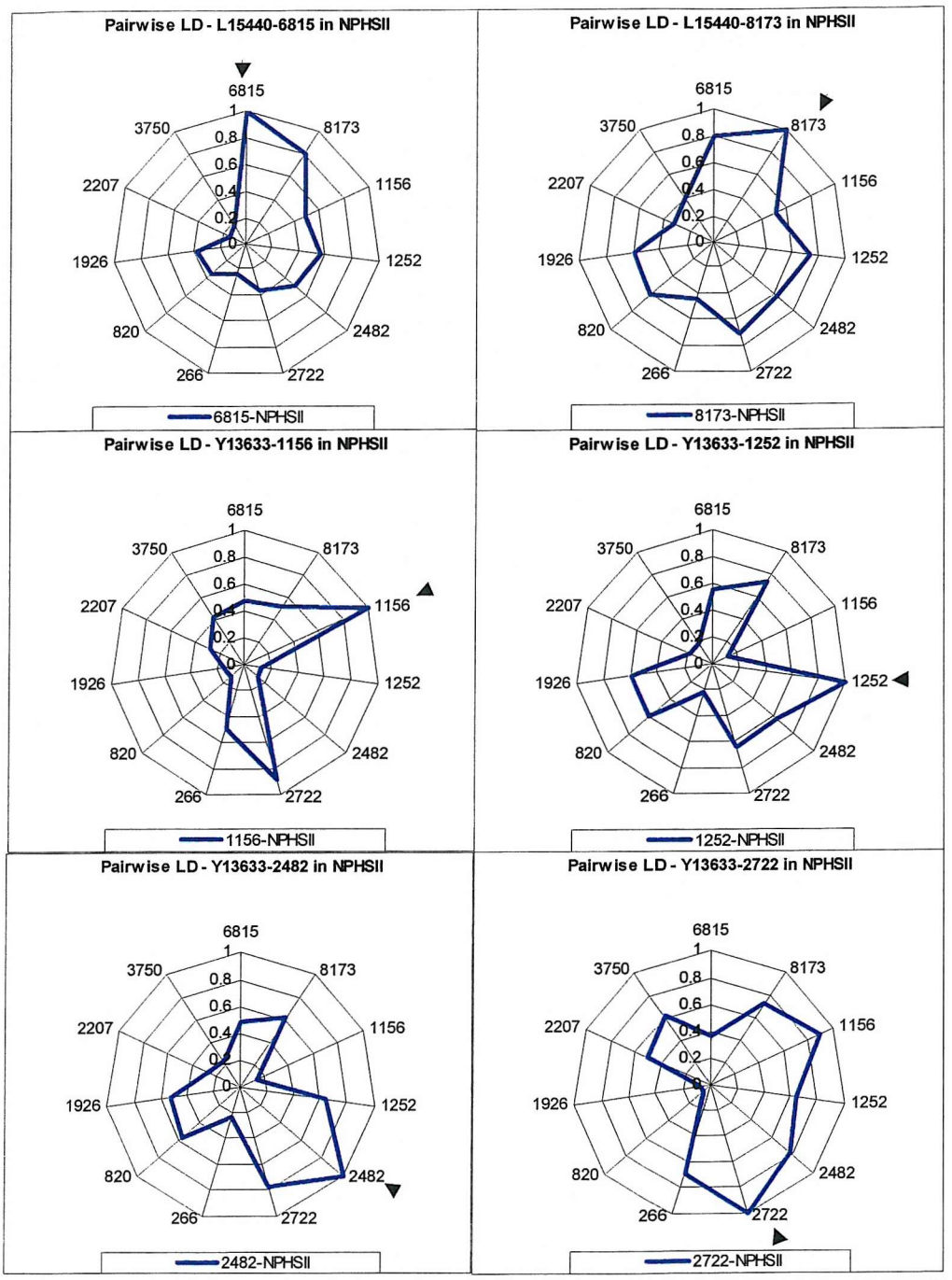


Figure 5-8: Stepwise regression model for BMI in NPHSII.
The sequential R^2 values show the additive contribution to variance in BMI as each SNP was added into the model. The partial R^2 values show the independent contribution of each SNP to the total variance in BMI.

5.2.2.5 Linkage disequilibrium analysis of *IGF2* SNPs

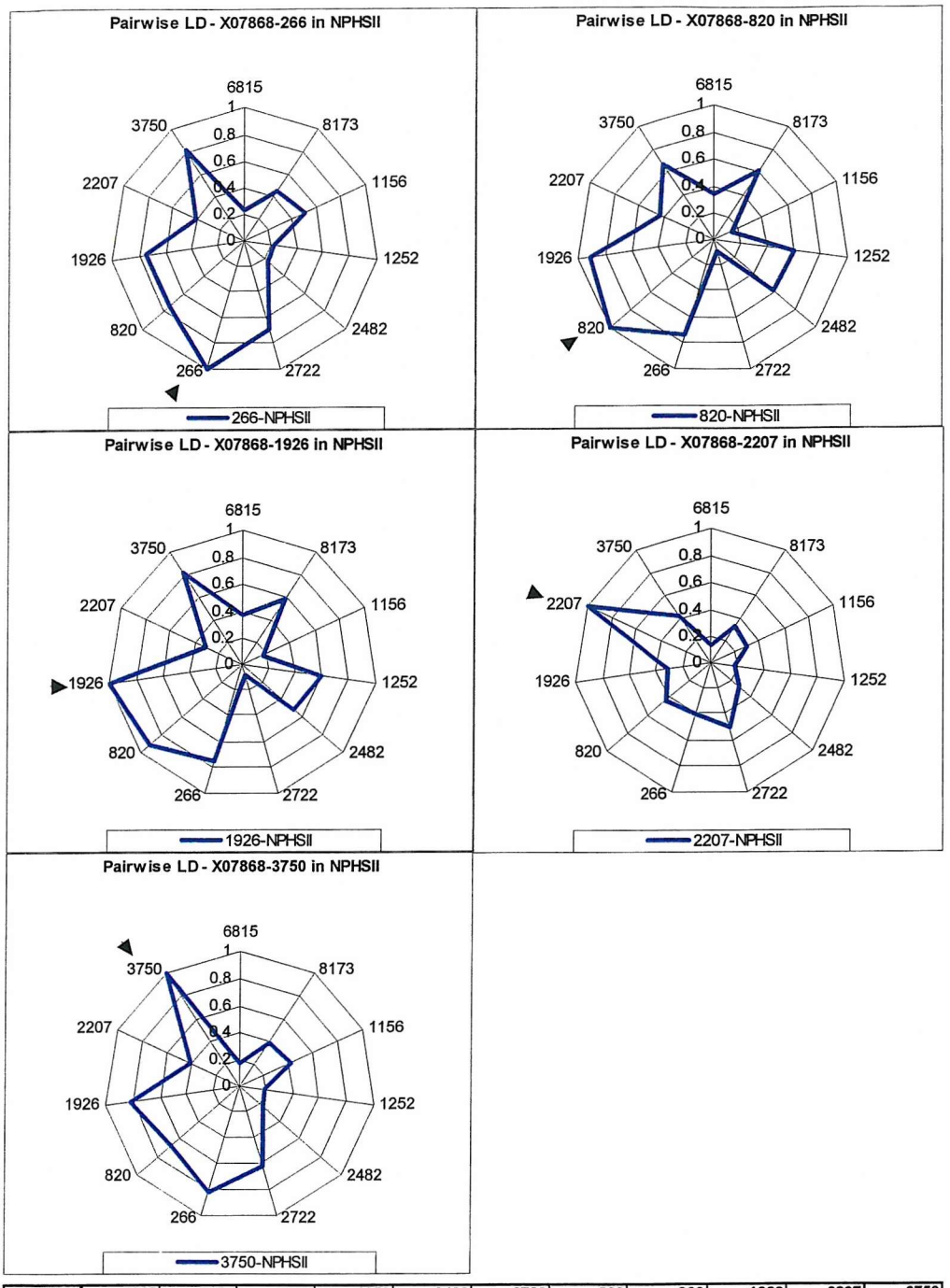
Pairwise calculations of |D'| (Lewontin 1964) were calculated using "2LD", which handles single pairs of markers. To carry out pairwise analysis on 11 SNPs this required 55 executions of "2LD" (10+9+8…+1), each with a separate input file containing only the two SNPs being tested. To achieve this a script was written (Appendix B, Script 1) which extracted each pair from a data file of multiple SNPs, and executed 2LD 55 times, collating the results in a single data file.

The data for the pairwise |D'| between SNPs in *IGF2* are plotted in Figure 5-9 and Figure 5-10. The shapes of the plots allow direct comparison of patterns of LD between SNPS, indicating which SNPs will tend to mirror the results of others, and which are more independent. |D'|=1 for SNPs in complete linkage disequilibrium and |D'|=0 for SNPs in complete equilibrium, therefore higher values indicate greater linkage disequilibrium.



NPHSII |D'|	6815	8173	1156	1252	2482	2722	266	820	1926	2207	3750
6815	1	0.80	0.48	0.55	0.48	0.36	0.23	0.34	0.37	0.13	0.17
8173	0.80	1	0.51	0.74	0.62	0.71	0.45	0.60	0.59	0.32	0.39
1156	0.48	0.51	1	0.12	0.13	0.89	0.50	0.14	0.16	0.28	0.42
1252	0.55	0.74	0.12	1	0.64	0.64	0.22	0.59	0.60	0.18	0.18
2482	0.48	0.62	0.13	0.64	1	0.77	0.24	0.58	0.51	0.27	0.23
2722	0.36	0.71	0.89	0.64	0.77	1	0.69	0.08	0.07	0.50	0.62
266	0.23	0.45	0.50	0.22	0.24	0.69	1	0.74	0.75	0.40	0.82
820	0.34	0.60	0.14	0.59	0.58	0.08	0.74	1	0.92	0.43	0.67
1926	0.37	0.59	0.16	0.60	0.51	0.07	0.75	0.92	1	0.30	0.82
2207	0.13	0.32	0.28	0.18	0.27	0.50	0.40	0.43	0.30	1	0.41
3750	0.17	0.39	0.42	0.18	0.23	0.62	0.82	0.67	0.82	0.41	1

Figure 5-9: Pairwise |D'| between *IGF2* markers in NPHSII.
Here the first six SNPs (5'→3') within the gene are plotted. Each plot represents one SNP (arrowed), and the line intersects at the corresponding value of |D'| on the lines for each of the other SNPs. The table shows actual values of |D'|.



NPHSII |D'|	6815	8173	1156	1252	2482	2722	266	820	1926	2207	3750
6815	1	0.80	0.48	0.55	0.48	0.36	0.23	0.34	0.37	0.13	0.17
8173	0.80	1	0.51	0.74	0.62	0.71	0.45	0.60	0.59	0.32	0.39
1156	0.48	0.51	1	0.12	0.13	0.89	0.50	0.14	0.16	0.28	0.42
1252	0.55	0.74	0.12	1	0.64	0.64	0.22	0.59	0.60	0.18	0.18
2482	0.48	0.62	0.13	0.64	1	0.77	0.24	0.58	0.51	0.27	0.23
2722	0.36	0.71	0.89	0.64	0.77	1	0.69	0.08	0.07	0.50	0.62
266	0.23	0.45	0.50	0.22	0.24	0.69	1	0.74	0.75	0.40	0.82
820	0.34	0.60	0.14	0.59	0.58	0.08	0.74	1	0.92	0.43	0.67
1926	0.37	0.59	0.16	0.60	0.51	0.07	0.75	0.92	1	0.30	0.82
2207	0.13	0.32	0.28	0.18	0.27	0.50	0.40	0.43	0.30	1	0.41
3750	0.17	0.39	0.42	0.18	0.23	0.62	0.82	0.67	0.82	0.41	1

Figure 5-10: Pairwise |D'| between IGF2 markers in NPHSII.
Here the last five SNPs (5'→3') within the gene are plotted. Each plot represents one SNP (arrowed), and the line intersects at the corresponding value of |D'| on the lines for each of the other SNPs. The table shows actual values of |D'|.

The data for Y13633-1252 and Y13633-2482 (Figure 5-9) indicate that these two SNPs share similar linkage disequilibrium relationships with the other *IGF2* SNPs (while not in very strong linkage disequilibrium with each other |D'|=0.64). In contrast, Y13633-1156 and Y13633-2722 (Figure 5-9) are in strong linkage disequilibrium (|D'|=0.89), but have different patterns of LD with other SNPs. Y13633-1156 appears to be relatively independent of all SNPs except Y13633-2722.

In the 3'UTR of the gene the two SNPs significantly associated with BMI (X07868-820 and X07868-1926) are in strong linkage disequilibrium (|D'|=0.92) and share a very similar pattern of LD relationships with other *IGF2* SNPs. This is consistent with the stepwise regression results (section 5.2.2.4) that show X07868-1926 not independently contributing to the variance in BMI.

The full data for pairwise |D'| and their variance and χ^2 values are shown in Appendix D. The χ^2 test compares estimated haplotype frequencies as 'observed' values to the expected haplotype frequencies under the null hypothesis of no linkage disequilibrium (calculated from allele frequencies). A significant χ^2 value therefore indicates a significant deviation from the null hypothesis of no linkage disequilibrium between the SNPs. The following pairs of SNPs did not significantly deviate from linkage equilibrium: L15440-6815 & X07868-2207, Y13633-1252 & X07868-2207, X07868-266 & X07868-2207, X07868-1926 & X07868-2207, X07868-2207 & X07868-3750 according to a χ^2 test.

5.2.2.6 Haplotype analysis of associated *IGF2* SNPs

The haplotype frequencies for the four significantly associated SNPs (1: L15440-6815A→T, 2: Y13633-1156T→C, 3: X07868-820G→A, 4: X07868-1926C→G) were estimated in order to identify the most common haplotypes, and those that correspond to differences in BMI. For this purpose the "EH" program was used (Xie & Ott 1993). This program estimates the most likely haplotype frequencies based on the genotype data entered into the program, and compares them to the haplotype frequencies that would be expected under the null hypothesis of no allelic association. The haplotype frequencies for each BMI quartile are shown in Figure 5-11.

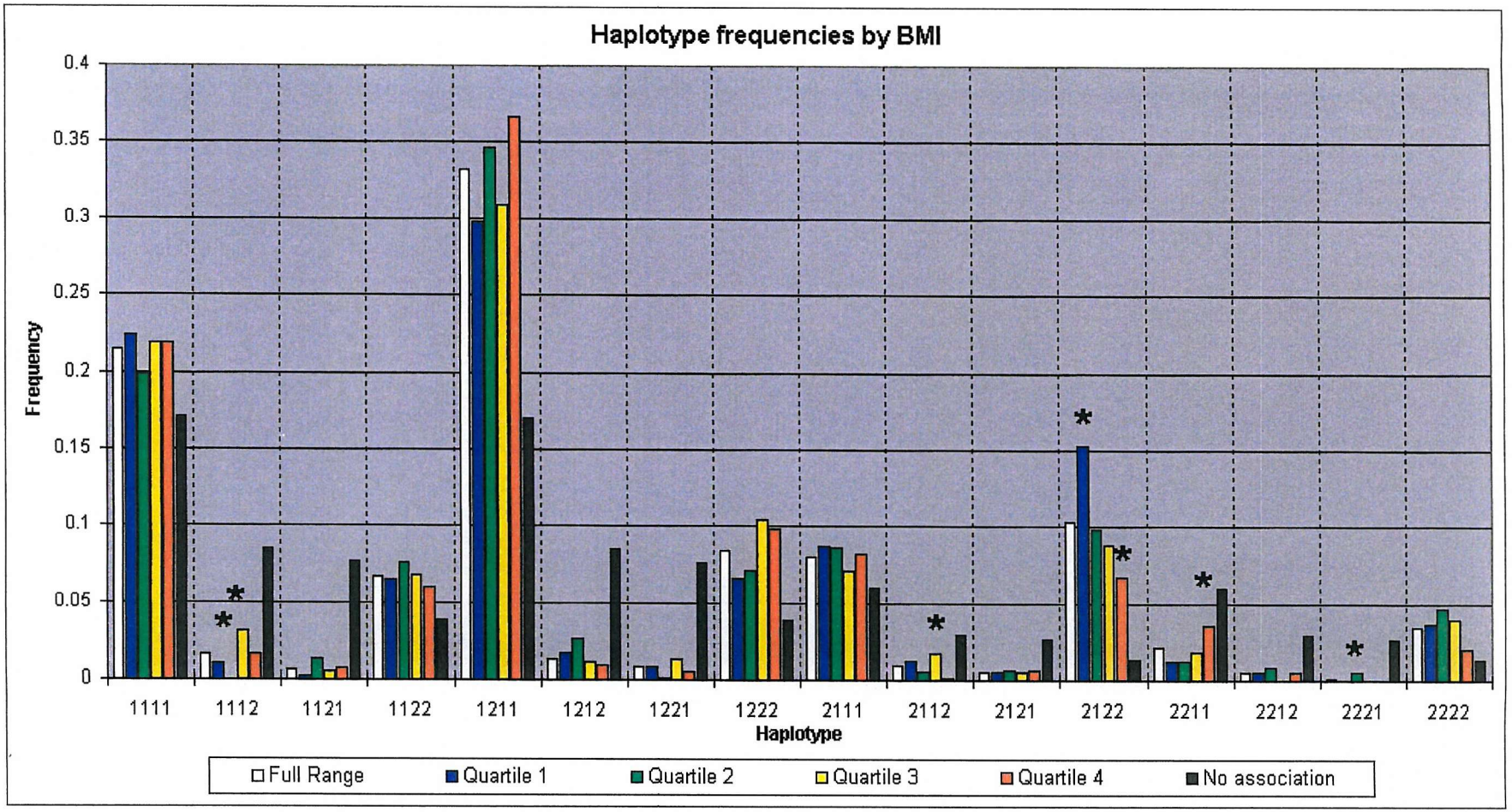


Figure 5-11: Haplotype frequencies for each BMI category
Quartiles: quartile 1 (light): blue, quartile 2: green, quartile 3: yellow, quartile 4 (heavy): orange. Also shown are the estimated haplotype frequencies for the full data set (white), and the expected haplotype frequencies under the null hypothesis of no association (black). Order of loci in haplotype is 1: L15440-6815A→T, 2: Y13633-1156T→C, 3: X07868-820G→A, 4: X07868-1926C→G
* = Contributes to χ^2 significance ($P<0.05$)

Figure 5-11 shows the trends for estimated haplotype frequencies in each of the quartiles of body mass index. Also shown are the estimated haplotype frequencies for the entire BMI distribution (white bars) and the haplotype frequencies expected for the entire BMI distribution under the null hypothesis of no linkage disequilibrium (black bars). Estimated haplotype frequencies for the entire BMI distribution were significantly different from the expected haplotype frequencies under the null hypothesis of no linkage disequilibrium ($\chi^2 = 1235.46$, 15d.f., $p<0.001$).

A χ^2 test comparing estimated haplotype numbers (Table 5-14) for each quartile against those that would be expected under the null hypothesis of no association between haplotype frequency and BMI (Table 5-15) showed that five haplotypes made major contributions to an overall χ^2 of 97.6 ($p = 0.000009$, 45d.f.) (Table 5-16). These were 1112 (quartiles 2 and 3), 2122 (quartiles 1 and 4), 2112 (quartile 3), 2211 (quartile 4) and 2221 (quartile 2). The most frequent of these is the 2122 haplotype, indicating that these alleles of the four SNPs act on a single haplotype to significantly influence BMI. This is consistent with the observation that homozygotes for each of these alleles are significantly lighter than the homozygotes for the alterative allele at each locus (Table 5-12).

Table 5-14: Table of estimated haplotype numbers in each BMI quartile
Calculated as haplotype frequency x number of informative chromosomes per quartile

	Quartile 1 (Light)	Quartile 2	Quartile 3	Quartile 4 (Heavy)	
1111	108	96	105	105	414
1112	5	0	15	8	28
1121	1	7	3	4	14
1122	31	36	33	29	129
1211	143	167	149	176	635
1212	8	12	5	4	30
1221	4	0	6	3	13
1222	31	34	50	47	163
2111	42	41	34	39	157
2112	6	2	8	0	17
2121	2	3	2	3	10
2122	74	47	42	32	195
2211	6	6	9	17	38
2212	2	4	0	3	9
2221	0	2	0	0	2
2222	17	22	19	10	69
	481	481	481	481	1924

Table 5-15: Haplotype numbers expected under independence
Calculated from column and row totals of Table 5-14: (column total*row total)/table total

	Quartile 1 (Light)	Quartile 2	Quartile 3	Quartile 4 (Heavy)
1111	103.43	103.43	103.43	103.43
1112	7.02	7.02	7.02	7.02
1121	3.47	3.47	3.47	3.47
1122	32.29	32.29	32.29	32.29
1211	158.71	158.71	158.71	158.70
1212	7.60	7.60	7.60	7.60
1221	3.30	3.30	3.30	3.30
1222	40.77	40.77	40.77	40.77
2111	39.24	39.24	39.24	39.24
2112	4.15	4.16	4.15	4.15
2121	2.61	2.62	2.61	2.61
2122	48.84	48.84	48.84	48.84
2211	9.47	9.47	9.47	9.47
2212	2.29	2.30	2.29	2.29
2221	0.60	0.61	0.60	0.60
2222	17.19	17.19	17.19	17.19

Table 5-16: χ^2 between estimated haplotype and haplotypes under independence
Major contributions to the overall χ^2 value are bold, with haplotypes highlighted (yellow)
(i.e χ^2>3.84 at 1d.f. is significant P<0.05, χ^2>6.63 at 1d.f. is very significant P<0.01)
Overall χ^2 = 97.60 (45 degrees of freedom). P = 0.000009

	Quartile 1 (Light)	Quartile 2	Quartile 3	Quartile 4 (Heavy)
1111	0.17	0.59	0.03	0.04
1112	0.68	**6.88**	**9.55**	0.13
1121	1.66	2.78	0.23	0.01
1122	0.03	0.54	0.00	0.39
1211	1.50	0.41	0.63	1.91
1212	0.05	3.14	0.66	1.39
1221	0.14	2.48	2.64	0.18
1222	2.11	1.12	2.18	1.07
2111	0.19	0.12	0.67	0.00
2112	0.55	0.68	**3.94**	3.63
2121	0.06	0.07	0.03	0.03
2122	**12.48**	0.07	0.83	**5.57**
2211	1.35	1.30	0.04	**6.28**
2212	0.01	1.45	2.29	0.04
2221	0.60	**4.31**	0.59	0.28
2222	0.00	1.58	0.21	3.06

The haplotype estimated to be most common is 1211. Note that Y13633-1156 is close to 0.5 allele frequency, and so the labelling of alleles as 1 and 2 may not indicate wild-type and mutant alleles in this case. The 1211 haplotype showed a trend toward higher frequency in higher BMI quartiles, which was not significant. The second most common was the 1111 haplotype. This showed no strong trend with BMI.

The third most common haplotype was 2122, which is the "opposite" haplotype to 1211 (i.e. the opposite allele is observed at each locus). This haplotype showed the opposite trend to the 1211 haplotype, with higher frequencies in the lower BMI categories. This trend

contributed significantly to the overall χ^2 statistic, with quartile 1 (lowest BMI) contributing a χ^2 of 12.48 (P<0.001, 1d.f.). The trend extended across all four quartiles. This indicates that the haplotype 2122 confers protection against weight gain in middle-aged men.

5.3 Discussion

5.3.1 $(CA)_n$ repeat genotype

The $(CA)_n$ repeat genotyping project did not give statistically significant results. This may have been due to the small number of samples (<80). These individuals were selected from the tails of the BMI distribution (BMI<20 and BMI>35), resulting in a "case-control" type of analysis rather than a population-wide analysis. When the same type of analysis was applied to X07868-820G/A (*ApaI*) the *P*-value became more significant (0.0004→0.0002). With L15440-6815A/T and Y13633-1156T/C the *P*-values increased slightly (0.0012→ 0.0077 and 0.017→0.0455 respectively) but still remained significant. This indicates that genotyping the extreme tails of the BMI distribution in the case of these particular SNPs provides equivalent results to genotyping the entire distribution, although this may vary between polymorphisms. Some polymorphisms may associate with BMI only within the "normal" BMI range, being less important in the extremes. These are of the most interest in common disease. Genotyping of extremes would be expected to be most effective with rare polymorphisms of large effect rather than common polymorphisms of small effect, and so is not appropriate for an investigation of a complex polygenic disease such as obesity.

The 5' component of the marker had a heterozygosity of 50% (35/70), which is equal to the maximum possible heterozygosity of a diallelic SNP. The existence of three alleles potentially allows further subdivision of haplotypes. The 3' half of the repeat marker (diallelic) had a lower heterozygosity at 45% (33/73). The two components of this marker therefore provided little more information than a SNP could provide, and required more expense and time to genotype per sample. Two rare 5' alleles and one rare 3' allele were not observed in our sample, although they have been reported by others (Rainier *et al.* 1993). Strong LD between the two components means that genotyping both ends is unnecessary, but in this case is easily achieved in one assay.

Both components of this marker showed a trend towards different mean BMIs in different genotypes, and the level of allelic association with X07868-820G→A suggests that

this trend would be expected in the $(CA)_n$ repeat marker as a reflection of the association between BMI and genotype in X07868-820G→A (O'Dell *et al.* 1997). The 5' fragment trends are for higher BMI with allele 397, and lower with 382 (Table 5-5). The 3' fragment trends are for higher BMI with allele 399 and lower with 404 (Table 5-7). The genotype numbers in each weight group confirm these trends (Figure 5-5). Higher than expected haplotype frequencies for 1-397-399 and 2-382-404 (Figure 5-6) are observed, where these haplotypes are "high-high-high" and "low-low-low" with respect to BMI at each locus. This observation fits with the hypothesis that alleles of this marker reside on a haplotype with X07868-820G and a variant predisposing individuals to weight gain (or X07868-820A and a variant protecting individuals from weight gain).

5.3.2 SNP genotyping

The genotyping of SNPs in the *IGF2* gene demonstrates the effectiveness and the limitations of this strategy for mapping of a causal site in a complex disease. Three SNPs in the gene have been found to be independently significantly associated with BMI in middle-aged men. The locations of these positive associations are: P1 promoter at the 5' end of the gene (L15440-6815) (Gu *et al.* 2002; Gaunt *et al.* 2001), intron 2 in the middle of the gene (Y13633-1156) and the 3' untranslated region (X07868-820) (O'Dell *et al.* 1997). While one other site (X07868-1926) was found to be significantly associated with BMI, it was in strong LD with X07868-820, and therefore not independent.

An interesting observation for the use of SNPs in mapping complex disease genes is that a number of SNPs were not significantly associated with BMI, despite close physical proximity to other markers that were. This is not surprising, since linkage disequilibrium depends not just on physical distance, but also on chronological distance. It seems likely therefore that despite the distances over which LD is known to exist (3kb (Kruglyak 1999) to 100kb (Collins *et al.* 1999)), the use of SNPs with this spacing for LD mapping may be insufficient to detect a biologically important site. The thorough approach would be to genotype every available SNP within a gene when using the methods of analysis used here.

5.3.3 Independent SNP associations with BMI

The three independent positive associations between SNPs in the *IGF2* gene and body mass index observed in this study indicate that *IGF2* has a role in determination of weight. While the mechanism of this is unclear, it is consistent with observations that IGF-II levels are altered in obesity (Argente *et al.* 1997; Frystyk *et al.* 1995). The relatively low contribution to overall variance in BMI is consistent with the nature of complex genetic disease. The *IGF2* gene is one of potentially many genetic factors contributing to a phenotype that is also influenced by environmental factors.

The most significant independent association is that of the SNP Y13633-1156 with body mass index. The stepwise regression analysis showed that this contributed nearly 50% of the influence of *IGF2* SNPs on BMI (partial $R^2 = 0.91$) compared to L15440-6815 and X07868-820 which each contributed about 25% (partial $R^2 = 0.48$ and 0.44 respectively). The data for Y13633-1156 were less statistically significant than that for the other two SNPs ($P = 0.017$), as the number of genotypes obtained for this SNP (1567) were less than were obtained for the other two SNPs (2394 and 2560). Also the difference in mean BMI between common and rare homozygotes for Y13633-1156 was less ($0.6 kgm^{-2}$) than L15440-6815 ($1.3 kgm^{-2}$) and X07868-820 ($1.1 kgm^{-2}$). This may be because the Y13633-1156 SNP marks a causal site with less effect on BMI, or it may be because linkage disequilibrium between Y13633-1156 and a causal site is less than LD between the other significantly associated SNPs and the causal sites that they mark. The latter hypothesis is consistent with the observation that Y13633-1156 has a high rare allele frequency, indicating that it is an old SNP for which LD with other SNPs may have decayed due to recombination.

The difference in mean BMI between genotype groups for each of these SNPs was not great. This is not unexpected for a complex genetic disease that is influenced by multiple genes and environmental conditions. None of the SNPs were in a coding sequence or an obvious regulatory sequence, and so a direct extreme effect was not expected. However, the independent positive associations between SNPs in the *IGF2* gene and body mass index in middle-aged men suggest that this gene or an adjacent gene (*INS*, *H19* or *TH*) contain one or more variants that influence either gene expression levels or activity of the gene product, and thus influence predisposition to weight gain.

Another group has recently published a study that found no association between X07868-820G→A genotypes and BMI, and a significant opposite association between genotype and fat mass (Roth *et al.* 2002). They did not replicate the other significantly

associated SNPs in their study, but the use of a smaller population of mixed sex and broad age range would not be expected to produce equivalent results. This indicates a limitation of the study of NPHSII: no indication of the role of *IGF2* in weight determination in women or different age categories can be determined. This requires a study of different cohorts.

5.3.4 Multiple testing

No correction for multiple testing has been applied here. Eleven SNPs were tested against three phenotypes (height, weight and BMI), a total of 33 tests. However, height, weight and BMI are not independent, and therefore do not constitute independent multiple tests. Also, the SNPs are not completely independent as indicated by the relatively high linkage disequilibrium between them. A conventional Bonferroni correction of 33 would therefore be excessively stringent (although both L15440-6815A→T and X07868-820G→A would still remain significant at p=0.00396 and 0.0132 respectively). A correction factor of 10 may also still be conservative due to the lack of independence of SNPs, so although a correction factor of 10 would make Y13633-1156T→C and X07868-1926C→G non-significant, the uncorrected p-values are provided for information and on the basis that application of an incorrect correction factor could be more misleading than application of no correction factor.

5.3.5 Haplotype analysis in *IGF2*

The association of the haplotype 2122 (order: L15440-6815A→T, Y13633-1156T→C, X07868-820G→A, X07868-1926C→G) with lower BMI fits with the observed association data (section 5.2.2.2) that associates lower BMI with 22 genotypes at SNPs L15440-6815, X07868-820 and X07868-1926, and with 11 genotypes at Y13633-1156. This indicates that these alleles are part of a haplotype that contains one or more polymorphisms that protect against weight gain.

The haplotype 1211 shows a non-significant trend for higher frequency in higher BMI quartiles, consistent with the association data. The lack of a BMI trend with the common 1111 haplotype probably reflects the opposing association of the different SNPs: the association of the Y13633-1156 11 genotype with lower BMI and the association of the 11 genotypes at the three other sites with higher BMI. Table 5-13 shows the results of a stepwise

regression analysis of the SNPs in *IGF2*. In this analysis X07868-1926 is not independent of X07868-820. The relative contributions to variance in BMI of the other three SNPs are indicated by partial R^2 values of 0.48 (L15440-6815), 0.91 (Y13633-1156) and 0.44 (X07868-820). This indicates that Y13633-1156 contributes to nearly half the variance in BMI, with the other SNPs accounting for the other half. This is consistent with the hypothesis of the 1111 haplotype being neutral with respect to BMI as a result of the opposing associations of the SNPs on that haplotype. The haplotype data therefore supports the regression analysis data in that the most independent SNP appears to be Y13633-1156T→C.

With a common allele frequency of 0.52 the Y13633-1156 single-nucleotide polymorphism is probably the oldest of the four significantly associated SNPs (i.e. the earliest to occur in history). L15440-6815, X07868-820 and X07868-1926 have common allele frequencies of 0.76, 0.73 and 0.71 respectively, indicating a similar age for these three SNPs. Assuming the common alleles of the latter three SNPs are wild-type, then the mutant allele at each of these loci is on a haplotype containing genetic variant(s) that result in lower body mass index in adult men.

A limitation of this analysis is the use of estimated haplotypes (derived using "EH" (Xie & Ott 1993)). Any genotypes that are heterozygous at more than one locus are not completely informative. In the simplest case of a two-locus haplotype that is heterozygous at both loci (Aa and Bb) there are two possible haplotype combinations (AB + ab OR Ab + aB). Haplotype frequencies in ambiguous data are therefore estimated based on frequencies observed in the unambiguous data. Haplotype frequency estimation has been demonstrated to be accurate for common haplotypes but not rare haplotypes (where direct molecular haplotyping would be necessary) (Tishkoff *et al.* 2000). The 2122 haplotype is relatively common at a frequency of >0.1 (in the entire BMI distribution). Estimation of haplotypes assumes Hardy-Weinberg equilibrium (Tishkoff *et al.* 2000); the independent genotype data for all the *IGF2* SNPs are in Hardy-Weinberg equilibrium.

5.3.6 Conclusions

Four significant associations between single-nucleotide polymorphisms in the *IGF2* gene and body mass index (BMI) in middle-aged men have been identified. Of these, three are independently significantly associated with BMI, and contribute to 1.83% of variance in body mass index in this population. One is in the P1 promoter, one is in intron 2 and the third

is in the 3'UTR close to an endonucleolytic cleavage site (as is the non-independent SNP). Haplotype analysis confirms an association between a haplotype containing alleles at all four loci and BMI.

Chapter 6 : Genetic Epidemiology: Hertfordshire Cohort

6.1 Introduction

6.1.1 The Hertfordshire cohort

This resource was developed for a project on the study of foetal and infant growth, adult lifestyle and their relationship with chronic disease. Between 1911 and 1948 midwives attending births in Hertfordshire recorded birth weights in ledgers that have survived to the present day. Routine follow-ups were also recorded, with data on weight at one year, breast and/or bottle-feeding and whether the child was weaned by that time (Fall *et al.* 1995). These valuable data were the basis of a project involving the collection of phenotype data and DNA that was run by the MRC Environmental Epidemiology Unit in Southampton. In 1994 the collection of these data and samples from Hertfordshire residents born in Hertfordshire between 1931→1939 began. The ages at that time were 55 to 63 years. This initial phase of the project ended in 1996 with 408 men and 303 women from North Hertfordshire and 255 men and 146 women from East Hertfordshire. The collection of further samples began in 1999 and is still in progress.

The data available from these individuals are very extensive, and include early life phenotypes. This makes it a very useful resource for the investigation of the foetal origins of adult disease, including cardiovascular disease (Fall *et al.* 1995), obesity (Law *et al.* 1992) and type 2 diabetes (Hales *et al.* 1991) (see section 1.5). The availability of DNA also provides an opportunity to investigate any potential influence of genotype on both adult and

early life phenotypes. Genes are known to influence birth weight (Hattersley *et al.* 1998; Frayling & Hattersley 2001), and therefore an association between foetal growth and adult phenotype may be a consequence of genetic factors.

6.1.2 Polymorphisms

The polymorphisms selected for genotyping are shown in Table 5-2. In addition to these polymorphisms one other had been previously genotyped in the Hertfordshire population (Aihie-Sayer *et al.* 2002). This polymorphism was X07868-820G→A (*ApaI* restriction site) (Tadokoro *et al.* 1991). The polymorphisms were single-nucleotide substitutions (10 genotyped here, plus 1 previously genotyped). Polymorphisms were selected for genotyping on the basis of replication of the NPHSII genotyping project.

Table 6-1: Polymorphisms tested in Hertfordshire cohort

Reference sequence (GenBank ID)	Nucleotide number(s) and sequence	Type of polymorphism	Reference
L15440	6815 A→T	SNP	(Lucassen *et al.* 1993)
L15440	8173 C→T	SNP	(Lucassen *et al.* 1993)
Y13633	1156 T→C	SNP	Chapter 3
Y13633	1252 T→C	SNP	(Lucassen *et al.* 1993)
Y13633	2482 A→C	SNP	Chapter 3
Y13633	2722 C→T	SNP	Chapter 3
X07868	266 C→T	SNP	Chapter 3
X07868	1926 C→G	SNP	Chapter 3
X07868	2207 C→T	SNP	Chapter 3
X07868	Sequence between 3750 and 3751 A→G	SNP	Chapter 3

6.1.3 Objectives

The objectives of this part of the project were to investigate the associations observed in the NPHSII population. The primary hypothesis was that polymorphisms in the insulin-like growth factor II (*IGF2*) gene are associated with body mass index in adults. Chapter 5 demonstrated this association in middle-aged men, and this chapter expands on that by including women and older men. The secondary hypothesis was that *IGF2* polymorphisms influence foetal or early growth in addition to adult body mass index. The objectives for this

part of the project were:

1. Genotype single-nucleotide polymorphisms in the *IGF2* gene in the Hertfordshire cohort.
2. Analyse genotypes at different loci for associations with body mass index in men and women from the Hertfordshire cohort (independently).
3. Investigate association between birth weight, weight at one and genotype to determine whether polymorphisms in *IGF2* influence foetal or early growth.

6.2 Results

6.2.1 SNP genotypes

6.2.1.1 Data quality

A test of Hardy-Weinberg equilibrium was carried out on all the SNP genotypes as a test of data quality. The assumption is made that the population is in Hardy-Weinberg equilibrium, and if the genotype frequencies for a particular SNP are not in equilibrium, then there are potentially either data errors or stratification within the population. All ten SNPs are in Hardy-Weinberg equilibrium within the Hertfordshire cohort (Table 6-2).

Table 6-2: Test of Hardy-Weinberg equilibrium in genotyped SNPs.
Expected numbers calculated from allele frequencies, then observed numbers compared to expected with a χ^2 test. A *P*-value <0.05 would indicate deviation from Hardy-Weinberg equilibrium

		L15440-6815 A/T	L15440-8173 C/T	Y13633-1156 T/C	Y13633-1252 T/C	Y13633-2482 A/C	Y13633-2722 C/T	X07868-266 C/T	X07868-1926 C/G	X07868-2207 C/T	X07868-3750ex A/G
Observed	11	501	350	225	201	229	455	645	492	788	755
	12	340	400	410	473	420	368	143	389	103	162
	22	54	115	143	228	185	56	5	65	4	13
	p (freq 1)	0.750	0.636	0.553	0.485	0.526	0.727	0.904	0.726	0.938	0.899
	q (freq 2)	0.250	0.364	0.447	0.515	0.474	0.273	0.096	0.274	0.062	0.101
Expected	11	503.06	349.71	237.66	212.20	231.08	464.53	647.38	498.18	787.44	751.50
	12	335.87	400.58	384.68	450.60	415.84	348.94	138.24	376.63	104.12	169.00
	22	56.06	114.71	155.66	239.20	187.08	65.53	7.38	71.18	3.44	9.50
$\chi^2=(O-E)^2/E$	11	0.008	0.000	0.674	0.591	0.019	0.195	0.009	0.077	0.000	0.016
	12	0.051	0.001	1.667	1.114	0.042	1.041	0.164	0.406	0.012	0.290
	22	0.076	0.001	1.030	0.525	0.023	1.386	0.767	0.537	0.091	1.289
χ^2		0.135	0.002	3.371	2.230	0.083	2.622	0.940	1.020	0.103	1.595
P-value		0.935	0.999	0.185	0.328	0.959	0.270	0.625	0.600	0.950	0.451

The frequencies of the ten SNPs within the Hertfordshire men and women was tested by χ^2 (Table 6-3). All ten SNPs showed the same distribution in men and women, which acts

as a test of data quality and indicates that no selective effects are influencing genotype between genders. The low frequency SNPs (X07868-266C→T, X07868-2207C→T and X07868-3750exA→G) have been analysed in two categories: common homozygote and combined heterozygote/rare homozygote. This is because these SNPs have 5, 4 and 13 rare homozygotes respectively, which would be statistically very weak as categories on their own.

Table 6-3: Relationship between gender and genotype frequencies
Tested by χ^2. *P*-value indicates the probability of men and women having the same genotype frequencies. Note that 12 and 22 numbers are combined for rare SNPs: X07868-266C→T, X07868-2207C→T and X07868-3750exA→G

SNP	Men			Women			Pearson	*P*-
	11	12	22	11	12	22	χ^2	value
L15440-6815 A/T	301	196	34	200	144	20	0.81	0.67
L15440-8173 C/T	195	218	68	155	182	47	0.78	0.68
Y13633-1156 T/C	117	236	83	108	174	60	2.11	0.35
Y13633-1252 T/C	127	267	140	74	206	88	3.26	0.20
Y13633-2482 A/C	149	260	123	80	160	62	1.39	0.50
Y13633-2722 C/T	252	209	37	203	159	19	2.32	0.31
X07868-266 C/T	363	75		282	73		1.53	0.22
X07868-1926 C/G	283	226	46	209	163	19	4.24	0.12
X07868-2207 C/T	446	65		342	42		0.66	0.42
X07868-3750ex A/G	454	96		301	79		1.64	0.20

6.2.1.2 Association analysis by one-way ANOVA

Associations between genotype and five different phenotypes for each of the ten SNPs plus X07868-820G→A (*ApaI* – genotyped previously) were tested by one-way analysis of variance. The phenotypes tested were weight, height, body mass index (BMI), birth weight and weight at one. Weight and height have been excluded from the tabulated results, as BMI is a more useful indicator of obesity and adiposity. Weight and height are incorporated into the BMI measure (weight x height^{-2}). In addition to ANOVA, a linear regression approach was used to test for trends in the means of each phenotype across genotype groups. This is effectively an allele count approach (i.e. 0, 1 or 2 rare alleles).

Table 6-4 shows the analyses for body mass index (BMI) in men. No significant *P*-values were found with either the ANOVA or regression analyses, indicating no association between genotype and body mass index for any of the 11 single nucleotide polymorphisms in Hertfordshire men. However trends were observed for L15440-6815A→T and X07868-820G→A that are consistent with those observed in NPHSII (magnitudes of 0.6 and 0.63kgm^{-2} in Herts and 1.33 and 1.1kgm^{-2} in NPHSII respectively). For both SNPs the rare (22) genotype group have a lower mean body weight than the common (11) genotype group.

A small trend for Y13633-1156T→C is in the opposite direction to that observed in NPHSII, but of small magnitude (0.24kgm^{-2}).

Table 6-5 shows the analyses for body mass index (BMI) in women. Significant *P*-values were found for Y13633-2482A→C (ANOVA *P*=0.02, regression *P*=0.02), Y13633-2722C→T (ANOVA *P*=0.02) and X07868-266C→T (ANOVA *P*=0.002, regression *P*=0.002). Y13633-2482A→C and X07868-266C→T also showed significant associations between genotype and weight (ANOVA *P*=0.03 and *P*=0.02 respectively, and regression *P*=0.02 for the latter). No other significant associations were observed in women for weight or BMI, and none for height, although trends for L15440-6815A→T and X07868-820G→A were opposite to those observed in men for both the Hertfordshire cohort and NPHSII.

Table 6-4: Association between genotype and body mass index (BMI) in men
11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.
n = number
Trend = direction of change in phenotype from 11 to 22
Mag. = Magnitude (of difference between 11 and 22, only shown if greater than 0.5)
† = One-way ANOVA (difference in mean between genotype; 2 d.f.)
‡ = Linear regression analysis (trend in means- allele count approach; 1 d.f.)
* = significant (*P*<0.05)
** = highly significant (*P*<0.01)
Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.

BMI-MEN	11:mean, sd, *n*	12:mean, sd, *n*	22:mean, sd, *n*	† *P*-value	*n*	Trend	Mag.	‡ Beta	‡ *P*-value
L15440-6815A/T	26.9, 3.43, 301	26.92, 3.68, 194	26.3, 3.45, 34	0.62	529	DOWN	>0.5	-0.15	0.55
L15440-8173C/T	26.98, 3.66, 194	26.62, 3.41, 217	27.15, 3.57, 67	0.43	478	UP		-0.03	0.91
Y13633-1156T/C	27.22, 3.68, 116	26.96, 3.42, 236	26.98, 3.35, 83	0.79	435	DOWN		-0.13	0.58
Y13633-1252T/C	27.03, 3.42, 126	26.72, 3.48, 266	27.07, 3.51, 140	0.55	532	UP		-0.002	0.63
Y13633-2482A/C	27.12, 3.65, 149	26.99, 3.64, 259	26.89, 3.14, 122	0.86	530	DOWN		-0.11	0.59
Y13633-2722C/T	26.97, 3.66, 250	26.87, 3.44, 209	27.69, 3.31, 37	0.43	496	UP	>0.5	0.15	0.55
X07868-266C/T	27.07, 3.39, 363	26.95, 3.58, 73		0.79	436	DOWN	>0.5	-0.12	0.79
X07868-820G/A	27.14, 3.63, 313	26.69, 3.46, 240	26.52, 3.0, 46	0.25	599	DOWN	>0.5	-0.05	0.31
X07868-1926C/G	26.99, 3.7, 281	26.97, 3.5, 226	26.95, 2.99, 46	0.99	553	DOWN		-0.16	0.94
X07868-2207C/T	27.11, 3.52, 445	26.49, 2.77, 65		0.18	510	DOWN	>0.5	-0.62	0.18
X07868-3750exA/G	26.78, 3.33, 452	27.27, 4.3, 96		0.21	548	UP	>0.5	0.49	0.21

Table 6-5: Association between genotype and body mass index (BMI) in women
11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.
n = number
Trend = direction of change in phenotype from 11 to 22
Mag. = Magnitude (of difference between 11 and 22, only shown if greater than 0.5)
† = One-way ANOVA (difference in mean between genotype; 2 d.f.)
‡ = Linear regression analysis (trend in means- allele count approach; 1 d.f.)
* = significant ($P<0.05$)
** = highly significant ($P<0.01$)
Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.
Yellow highlight indicates significant results.

BMI-WOMEN	11:mean, sd, *n*	12:mean, sd, *n*	22:mean, sd, *n*	† *P*-value	*n*	Trend	Mag.	‡ Beta	‡ *P*-value
L15440-6815A/T	26.67, 4.2, 200	27.45, 4.75, 144	27.87, 3.6, 20	0.19	364	UP	>0.5	0.69	0.07
L15440-8173C/T	27.3, 4.67, 154	26.52, 3.97, 182	27.1, 4.62, 47	0.24	383	DOWN		-0.3	0.36
Y13633-1156T/C	26.78, 4.28, 108	27.14, 4.17, 173	27.01, 4.02, 60	0.78	341	UP		0.15	0.64
Y13633-1252T/C	26.8, 4.16, 74	26.52, 4.32, 206	27.52, 4.4, 88	0.19	368	UP	>0.5	0.39	0.25
Y13633-2482A/C	28.3, 4.61, 80	26.63, 4.43, 159	26.64, 4.34, 62	0.02 *	301	DOWN	>0.5	-0.88	0.02 *
Y13633-2722C/T	26.86, 4.4, 203	26.67, 4.19, 158	29.52, 5.04, 19	0.02 *	380	UP	>0.5	0.45	0.23
X07868-266C/T	26.43, 4.12, 282	28.18, 4.64, 73		0.002 **	355	UP	>0.5	1.75	0.002 **
X07868-820G/A	26.66, 3.99, 233	27.01, 1.77, 160	27.65, 4.53, 23	0.49	416	UP	>0.5	0.09	0.3
X07868-1926C/G	26.8, 4.08, 209	26.82, 4.59, 163	27.97, 4.5, 19	0.52	391	UP	>0.5	0.25	0.49
X07868-2207C/T	26.96, 4.29, 341	26.66, 4.7, 42		0.67	383	DOWN	>0.5	-0.3	0.67
X07868-3750exA/G	26.61, 4.15, 301	27.49, 4.78, 79		0.1	380	UP	>0.5	0.88	0.1

The analyses of birth weight and genotype for Hertfordshire men are shown in Table 6-6. No significant associations were observed in either one-way analysis of variance (ANOVA) or regression analysis.

Table 6-7 shows the analyses of birth weight and genotype in Hertfordshire women. The only SNP showing a significant association between genotype and birth weight is X07868-3750exA→G (ANOVA $P=0.009$ and regression $P=0.009$). This SNP showed no association between genotype and either BMI (Table 6-5) or weight. None of the other SNPs showed association between genotype and birth weight.

Table 6-6: Association between genotype and birth weight (ounces) in men
11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.
n = number
Trend = direction of change in phenotype from 11 to 22
Mag. = Magnitude (of difference between 11 and 22, only shown if greater than 2.5 ounces)
† = One-way ANOVA (difference in mean between genotype; 2 d.f.)
‡ = Linear regression analysis (trend in means- allele count approach; 1 d.f.)
* = significant ($P<0.05$)
** = highly significant ($P<0.01$)
Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.

B.Wt-MEN	11:mean, sd, *n*	12:mean, sd, *n*	22:mean, sd, *n*	† *P*-value	*n*	Trend	Mag.	‡ Beta	‡ *P*-value
L15440-6815A/T	125, 17.98, 301	124.33, 19.1, 196	120.71, 22.51, 34	0.44	531	DOWN	>2.5	-1.42	0.28
L15440-8173C/T	123.44, 18.8, 195	126, 17.75, 218	126.29, 18.97, 68	0.3	481	UP	>2.5	1.71	0.16
Y13633-1156T/C	124.62, 21.52, 117	124.25, 17.7, 236	127.57, 16.59, 83	0.36	436	UP	>2.5	1.3	0.32
Y13633-1252T/C	123.83, 19.32, 127	125.07, 19.3, 267	124.01, 18.18, 140	0.78	534	UP		0.06	0.95
Y13633-2482A/C	124.68, 17.68, 149	124.76, 18.95, 260	126.68, 20.04, 123	0.6	532	UP		0.96	0.4
Y13633-2722C/T	124.83, 20.25, 252	125.35, 17.08, 209	126.86, 16.48, 37	0.81	498	UP		0.79	0.55
X07868-266C/T	125.63, 18.65, 363	124.35, 18.99, 75		0.59	438	DOWN	>2.5	-1.28	0.59
X07868-820G/A	124.79, 18.69, 315	125.37, 18.72, 240	122.43, 18.13, 47	0.61	602	DOWN		-0.22	0.39
X07868-1926C/G	124.73, 18.63, 283	124.1, 18.55, 226	122.83, 18.04, 46	0.79	555	DOWN		-0.81	0.51
X07868-2207C/T	124.76, 18.51, 446	124.77, 16.81, 65		0.99	511	UP	>2.5	0.005	0.99
X07868-3750exA/G	124.3, 18.74, 454	124.88, 18.89, 96		0.78	550	UP	>2.5	0.57	0.78

Table 6-7: Association between genotype and birth weight (ounces) in women
11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.
n = number
Trend = direction of change in phenotype from 11 to 22
Mag. = Magnitude (of difference between 11 and 22, only shown if greater than 2.5 ounces)
† = One-way ANOVA (difference in mean between genotype; 2 d.f.)
‡ = Linear regression analysis (trend in means- allele count approach; 1 d.f.)
* = significant ($P<0.05$)
** = highly significant ($P<0.01$)
Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.
Yellow highlight indicates significant results.

B.Wt-WOMEN	11:mean, sd, *n*	12:mean, sd, *n*	22:mean, sd, *n*	† *P*-value	*n*	Trend	Mag.	‡ Beta	‡ *P*-value
L15440-6815A/T	120.42, 19.22, 200	121.75, 16.78, 144	125, 13.97, 20	0.49	364	UP	>2.5	1.77	0.26
L15440-8173C/T	119.97, 16.72, 155	119.86, 17.91, 182	123.23, 19.98, 47	0.48	384	UP	>2.5	1.11	0.41
Y13633-1156T/C	122.27, 16.49, 108	121.16, 18.88, 174	121.6, 15.53, 60	0.87	342	DOWN		-0.45	0.74
Y13633-1252T/C	122.7, 18.69, 74	120.34, 18.95, 206	121.16, 15.11, 88	0.63	368	DOWN		-0.69	0.62
Y13633-2482A/C	121.73, 14.59, 80	119.89, 18.63, 160	121.87, 19.09, 62	0.65	302	UP		-0.05	0.97
Y13633-2722C/T	121.93, 18.38, 203	118.78, 17.19, 159	126.95, 13.72, 19	0.07	381	UP	>2.5	-0.75	0.63
X07868-266C/T	120.03, 18, 282	123.03, 15.59, 73		0.19	355	UP	>2.5	3	0.19
X07868-820G/A	121.85, 17.69, 234	119.26, 17.88, 160	119.83, 16.48, 23	0.35	417	DOWN	>2.5	-0.18	0.6
X07868-1926C/G	122.15, 17.97, 209	119.73, 17.77, 163	124.32, 16.11, 19	0.32	391	UP		-0.96	0.53
X07868-2207C/T	120.39, 17.62, 342	121.88, 16.88, 42		0.6	384	UP	>2.5	1.49	0.6
X07868-3750exA/G	119.91, 17.76, 301	125.78, 16.97, 79		0.009 **	380	UP	>2.5	5.87	0.009 **

The analyses of weight at one and genotype for Hertfordshire men are shown in Table 6-8. No significant associations were observed in either one-way analysis of variance (ANOVA) or regression analysis.

Table 6-9 shows the analyses of weight at one and genotype in Hertfordshire women. No significant associations were observed in either one-way analysis of variance (ANOVA) or regression analysis.

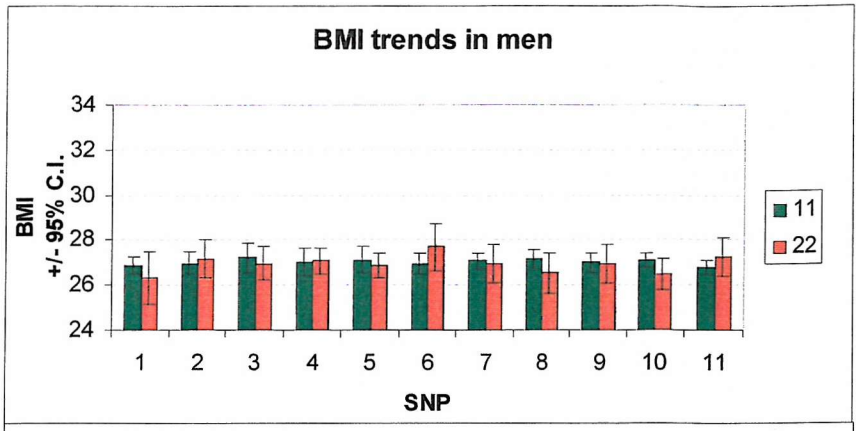
Table 6-8: Association between genotype and weight at one year (ounces) in men
11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.
n = number
Trend = direction of change in phenotype from 11 to 22
Mag. = Magnitude (of difference between 11 and 22, only shown if greater than 5 ounces)
† = One-way ANOVA (difference in mean between genotype; 2 d.f.)
‡ = Linear regression analysis (trend in means- allele count approach; 1 d.f.)
* = significant ($P<0.05$)
** = highly significant ($P<0.01$)
Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.

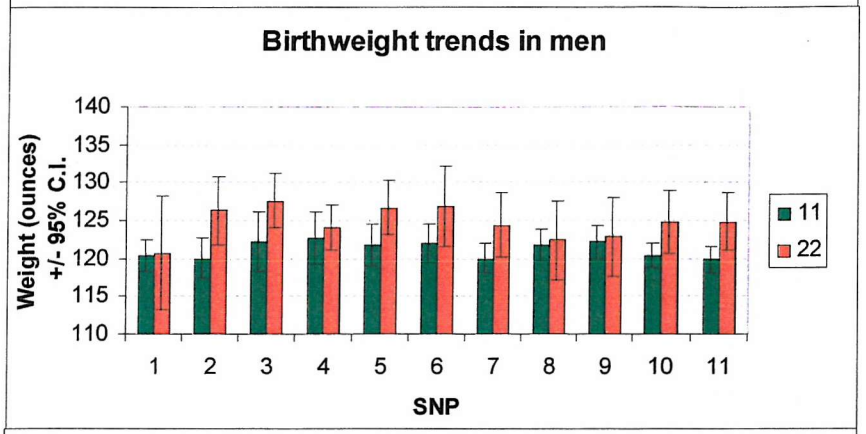
Wt:1yr-MEN	11:mean, sd, *n*	12:mean, sd, *n*	22:mean, sd, *n*	† *P*-value	*n*	Trend	Mag.	‡ Beta	‡ *P*-value
L15440-6815A/T	365.01, 40.35, 301	362.42, 42.43, 196	355.62, 47.94, 34	0.42	531	DOWN	>5	-3.66	0.213
L15440-8173C/T	362.07, 45.15, 195	365.45, 41.38, 218	357.57, 36.38, 68	0.38	481	DOWN		-0.83	0.76
Y13633-1156T/C	366.79, 44.24, 117	361.85, 40.87, 236	363.89, 44.55, 83	0.59	436	DOWN		-1.78	0.56
Y13633-1252T/C	360.1, 39.34, 127	364.2, 42.07, 267	362.52, 41.22, 140	0.65	534	UP		1.14	0.65
Y13633-2482A/C	359.97, 44.84, 149	364.38, 41.94, 260	364.59, 41.68, 123	0.55	532	UP		2.42	0.35
Y13633-2722C/T	362.45, 42.6, 252	362.81, 40.8, 209	370.14, 47.41, 37	0.58	498	UP	>5	2.24	0.46
X07868-266C/T	364.58, 43.35, 363	359.75, 38.07, 75		0.37	438	DOWN	>5	-4.83	0.37
X07868-820G/A	359.90, 40.57, 315	364.79, 43.99, 240	358.53, 42.81, 47	0.34	602	UP		-0.19	0.75
X07868-1926C/G	360.44, 40.6, 283	363.82, 43.18, 226	360.26, 41.97, 46	0.64	555	DOWN		1.38	0.62
X07868-2207C/T	362.6, 42.15, 446	364.91, 41.46, 65		0.68	511	UP	>5	2.3	0.68
X07868-3750exA/G	361.73, 42.62, 454	360.66, 42.58, 96		0.82	550	DOWN	>5	-1.07	0.82

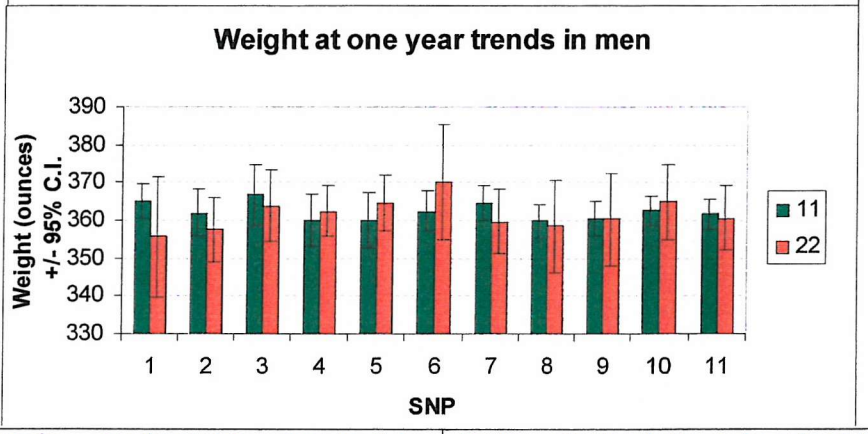
Table 6-9: Association between genotype and weight at one year (ounces) in women
11 = common homozygote, 12 = heterozygote, 22 = rare homozygote.
n = number
Trend = direction of change in phenotype from 11 to 22
Mag. = Magnitude (of difference between 11 and 22, only shown if greater than 5 ounces)
† = One-way ANOVA (difference in mean between genotype; 2 d.f.)
‡ = Linear regression analysis (trend in means- allele count approach; 1 d.f.)
* = significant ($P<0.05$)
** = highly significant ($P<0.01$)
Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.

Wt:1yr-WOMEN	11:mean, sd, *n*	12:mean, sd, *n*	22:mean, sd, *n*	† *P*-value	*n*	Trend	Mag.	‡ Beta	‡ *P*-value
L15440-6815A/T	341.58, 32.45, 200	341.49, 39.76, 144	345.1, 36.05, 20	0.91	364	UP		0.76	0.81
L15440-8173C/T	339.44, 37.39, 155	342.39, 37.11, 182	345.74, 31.85, 47	0.54	384	UP	>5	3.09	0.27
Y13633-1156T/C	337.38, 35.52, 108	345.28, 38.54, 174	335.9, 29.98, 60	0.09	342	DOWN		0.57	0.84
Y13633-1252T/C	343.62, 33.84, 74	342.93, 37.62, 206	336.68, 35.21, 88	0.35	368	DOWN	>5	-3.6	0.21
Y13633-2482A/C	343.1, 36.32, 70	341.46, 37.42, 132	346.29, 34.75, 58	0.7	260	UP		1.44	0.66
Y13633-2722C/T	340.73, 37.18, 203	344.51, 35.97, 159	338.63, 32.46, 19	0.56	381	DOWN		1.73	0.58
X07868-266C/T	342.65, 32.79, 282	337.7, 38.5, 73		0.27	355	DOWN	>5	-4.95	0.27
X07868-820G/A	343.46, 35.23, 234	338.76, 38.64, 160	333.39, 30.0, 23	0.26	417	DOWN	>5	-0.92	0.2
X07868-1926C/G	344.59, 35.64, 209	337.71, 37.7, 163	336.74, 31.11, 19	0.16	391	DOWN	>5	-5.65	0.07
X07868-2207C/T	342.17, 36.15, 342	339.67, 35.07, 42		0.67	384	DOWN	>5	-2.51	0.67
X07868-3750exA/G	339.86, 34.84, 301	345.63, 42.16, 79		0.21	380	UP	>5	5.77	0.21

Trends for each SNP and phenotype are shown in Figure 6-1 (men) and Figure 6-2 (women). These are intended to indicate the magnitude and direction of trends, not significance (although significant results are indicated).

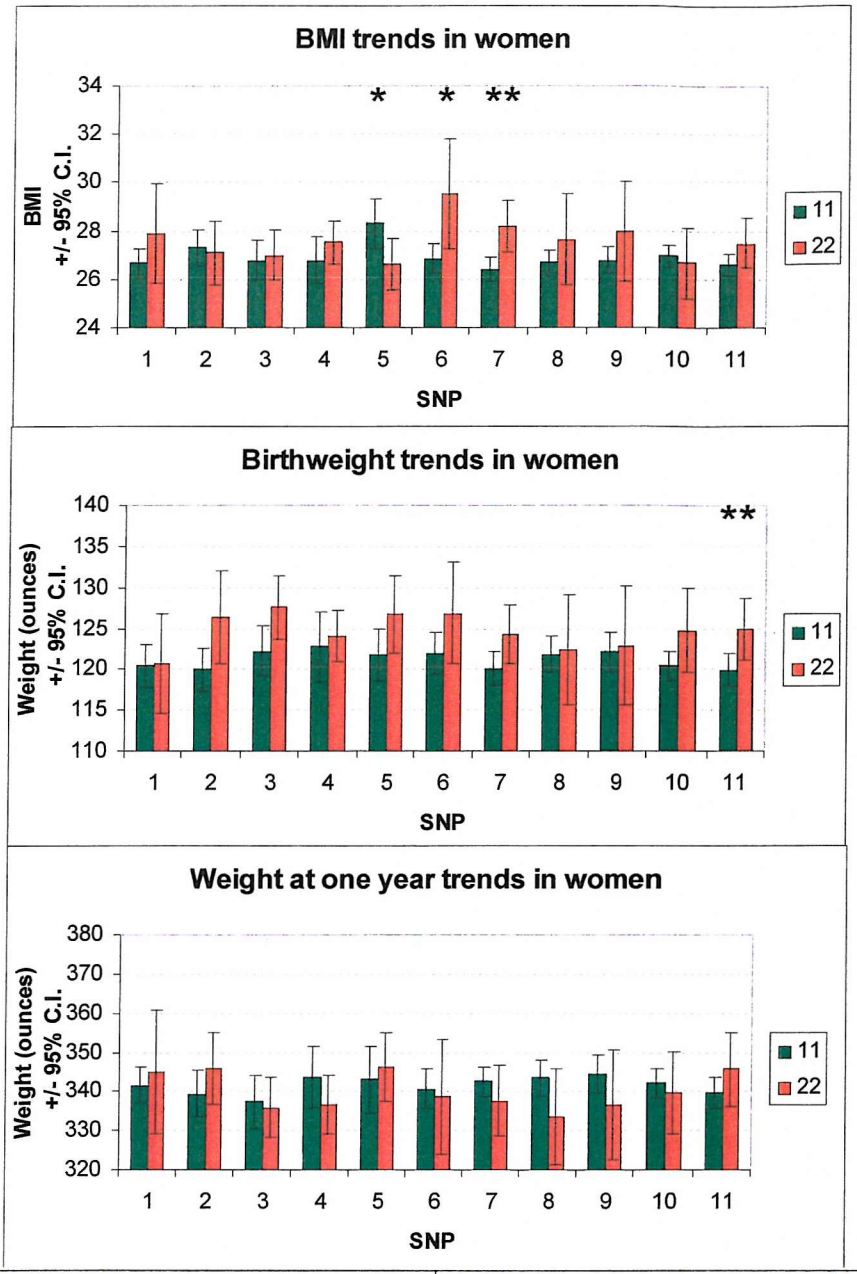






Code	Genbank ID and nucleotide	Code	Genbank ID and nucleotide
1	L15440-6815*A/T*	7	X07868-266*C/T*
2	L15440-8173*C/T*	8	X07868-820*G/A*
3	Y13633-1156*T/C*	9	X07868-1926*C/G*
4	Y13633-1252*T/C*	10	X07868-2207*C/T*
5	Y13633-2482*A/C*	11	X07868-3750ex*A/G*
6	Y13633-2722*C/T*		

Figure 6-1: Trends in BMI, birth weight and weight at one in men
Bars show mean value in each homozygote group with 95% confidence intervals. Note that the *x*-axis does not intercept the *y*-axis at 0. Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.



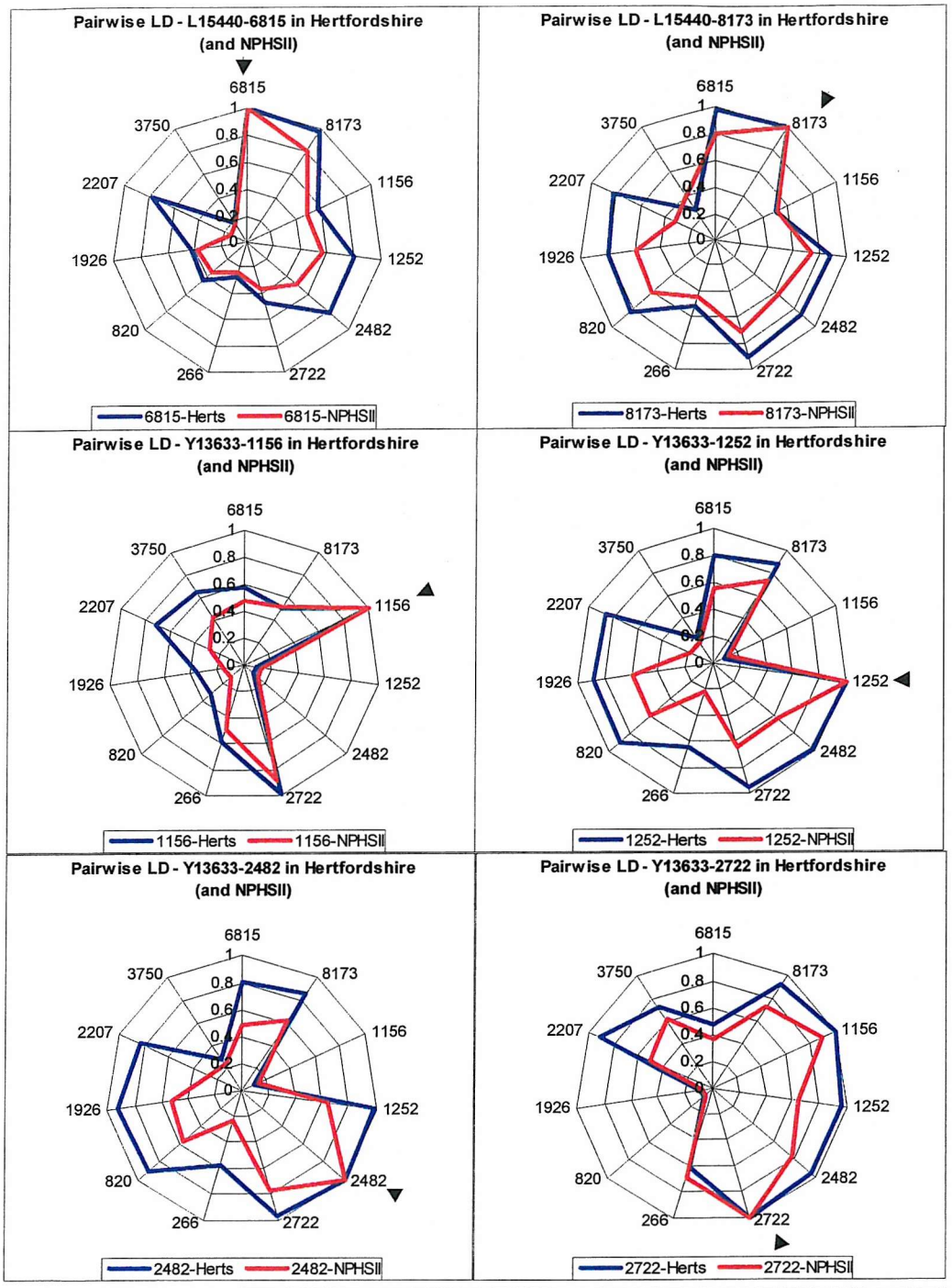
Code	Genbank ID and nucleotide	Code	Genbank ID and nucleotide
1	L15440-6815*A/T*	7	X07868-266*C/T*
2	L15440-8173*C/T*	8	X07868-820*G/A*
3	Y13633-1156*T/C*	9	X07868-1926*C/G*
4	Y13633-1252*T/C*	10	X07868-2207*C/T*
5	Y13633-2482*A/C*	11	X07868-3750ex*A/G*
6	Y13633-2722*C/T*		

Figure 6-2: Trends in BMI, birth weight and weight at one in women
Bars show mean value in each homozygote group with 95% confidence intervals. Note that the *x*-axis does not intercept the *y*-axis at 0. Note that 12 and 22 are combined for X07868-266C→T, X07868-2207C→T and X07868-3750exA→G.
* significant at $P<0.05$, ** significant at $P<0.01$

6.2.1.3 Linkage disequilibrium analysis of *IGF2* SNPs

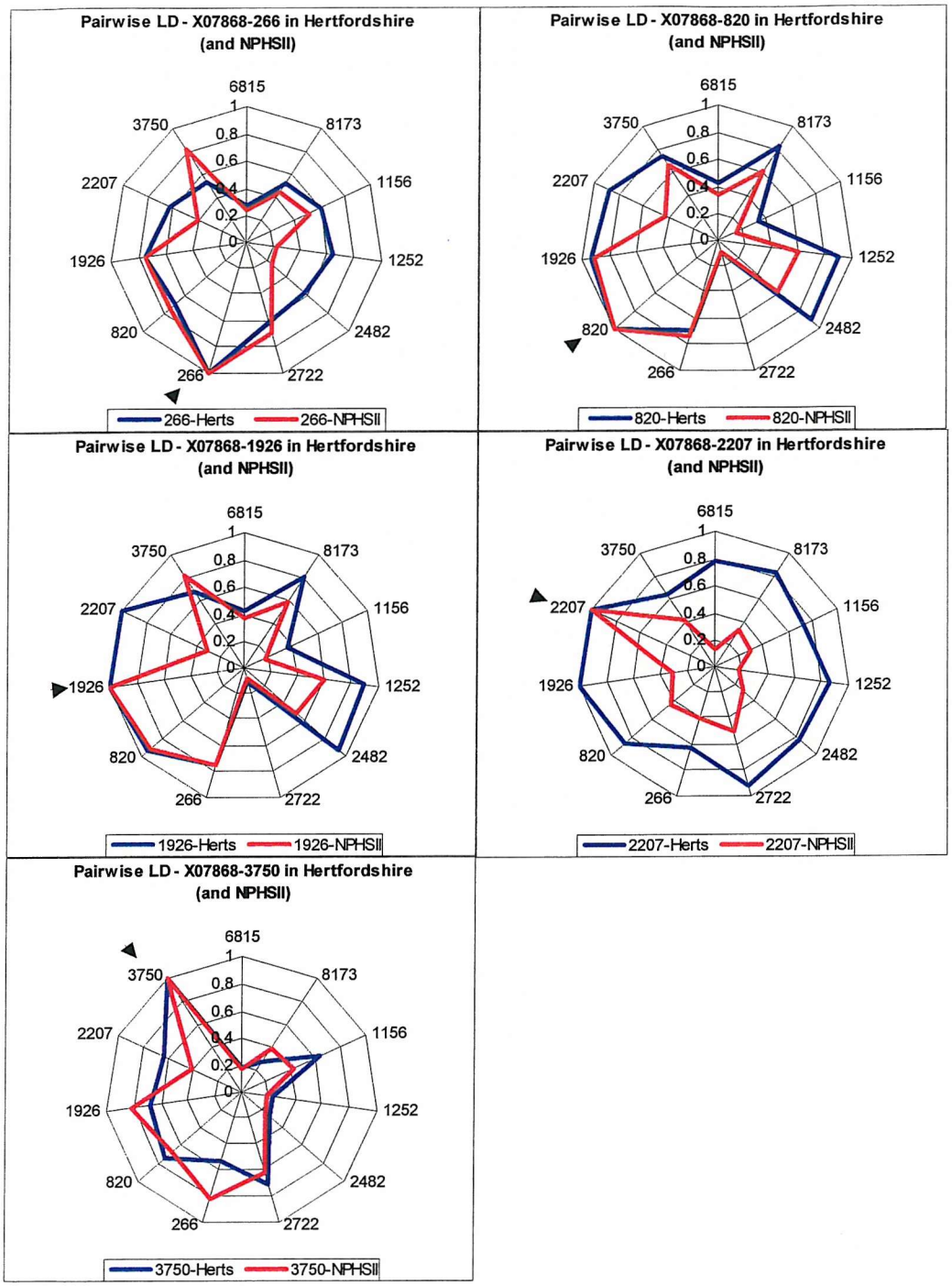
Pairwise calculations of |D'| (Lewontin 1964) were calculated using "2LD", which handles single pairs of markers using the same scripted automation as was used for NPHSII (Chapter 5). To carry out pairwise analysis on 11 SNPs this required 55 executions of "2LD": to achieve this a script was written (Appendix B, Script 1) which extracted each pair from a data file of multiple SNPs, and executed 2LD 55 times, collating the results in a single data file.

The data for the pairwise |D'| between SNPs in *IGF2* are plotted in Figure 6-3 and Figure 6-4 (full data are in Appendix E). The shapes of the plots allow direct comparison of patterns of LD between SNPS, indicating which SNPs tend to mirror the results of others, and which are more independent. |D'|=1 for SNPs in complete linkage disequilibrium and |D'|=0 for SNPs in complete equilibrium, therefore higher values indicate greater linkage disequilibrium. To enable direct comparison, the |D'| data for NPHSII (Chapter 5) have been plotted on the same graphs (red lines). In general similar patterns of LD for each of the SNPs are seen between the two cohorts. The magnitude of |D'| is generally greater in the Hertfordshire cohort than in NPHSII. The primary difference between the two data sets appears to be in the X07868-2207C→T SNP, which displays much higher LD in the Hertfordshire cohort than in NPHSII (observed in all graphs, and particularly in the graph for that SNP). This is a low frequency SNP with a rare homozygote count of 4 in the Hertfordshire cohort (rare allele frequency 0.062), and this may therefore skew the statistics.



HERTS |D'|	6815	8173	1156	1252	2482	2722	266	820	1926	2207	3750
6815	1	0.97	0.57	0.80	0.80	0.47	0.27	0.42	0.42	0.79	0.18
8173	0.97	1	0.50	0.87	0.85	0.91	0.52	0.82	0.80	0.83	0.27
1156	0.57	0.50	1	0.09	0.09	0.98	0.59	0.33	0.35	0.72	0.64
1252	0.80	0.87	0.09	1	0.97	0.96	0.64	0.90	0.89	0.86	0.23
2482	0.80	0.85	0.09	0.97	1	0.96	0.57	0.91	0.92	0.83	0.27
2722	0.47	0.91	0.98	0.96	0.96	1	0.61	0.09	0.09	0.92	0.71
266	0.27	0.52	0.59	0.64	0.57	0.61	1	0.68	0.75	0.62	0.53
820	0.42	0.82	0.33	0.90	0.91	0.09	0.68	1	0.94	0.86	0.74
1926	0.42	0.80	0.35	0.89	0.92	0.09	0.75	0.94	1	1.00	0.68
2207	0.79	0.83	0.72	0.86	0.83	0.92	0.62	0.86	1.00	1	0.63
3750	0.18	0.27	0.64	0.23	0.27	0.71	0.53	0.74	0.68	0.63	1

Figure 6-3: Pairwise |D'| between *IGF2* SNPs in the Hertfordshire cohort and NPHSII
SNPs in P1, intron 1, intron 2, exon 3 and intron 3 of *IGF2*.
Blue line: Hertfordshire cohort pairwise |D'|
Red line: NPHSII pairwise |D'|
Arrow – indicates SNP against which others are tested.



HERTS |D'|	6815	8173	1156	1252	2482	2722	266	820	1926	2207	3750
6815	1	0.97	0.57	0.80	0.80	0.47	0.27	0.42	0.42	0.79	0.18
8173	0.97	1	0.50	0.87	0.85	0.91	0.52	0.82	0.80	0.83	0.27
1156	0.57	0.50	1	0.09	0.09	0.98	0.59	0.33	0.35	0.72	0.64
1252	0.80	0.87	0.09	1	0.97	0.96	0.64	0.90	0.89	0.86	0.23
2482	0.80	0.85	0.09	0.97	1	0.96	0.57	0.91	0.92	0.83	0.27
2722	0.47	0.91	0.98	0.96	0.96	1	0.61	0.09	0.09	0.92	0.71
266	0.27	0.52	0.59	0.64	0.57	0.61	1	0.68	0.75	0.62	0.53
820	0.42	0.82	0.33	0.90	0.91	0.09	0.68	1	0.94	0.86	0.74
1926	0.42	0.80	0.35	0.89	0.92	0.09	0.75	0.94	1	1.00	0.68
2207	0.79	0.83	0.72	0.86	0.83	0.92	0.62	0.86	1.00	1	0.63
3750	0.18	0.27	0.64	0.23	0.27	0.71	0.53	0.74	0.68	0.63	1

Figure 6-4: Pairwise |D'| between *IGF2* SNPs in the Hertfordshire cohort and NPHSII
SNPs in 3'UTR of *IGF2*
Blue line: Hertfordshire cohort pairwise |D'|
Red line: NPHSII pairwise |D'|
Arrow – indicates SNP against which others are tested.

6.3 Discussion

6.3.1 Genotype/phenotype associations

6.3.1.1 One-way ANOVA

The association between genotype and phenotype was tested for eleven SNPs and five phenotypes. No correction for multiple testing has been applied (see section 6.3.1.3), and therefore the results should be treated with caution.

No significant associations between genotype and phenotype were observed in the Hertfordshire men. The direction of trends with BMI agreed with NPHSII results for L15440-6815A→T and X07868-820G→A, but there was no trend for X07868-1926C→G, and Y13633-1156T→C showed a small opposite trend (Figure 6-1). The magnitudes of trends were smaller than those in NPHSII (approximately half the BMI difference).

In Hertfordshire women several significant (non-corrected) associations were observed. Y13633-2482A→C was significantly associated with BMI (p=0.02). Y13633-2722C→T was significantly associated with BMI (p=0.02). X07868-266C→T was highly significantly associated with BMI (p=0.002) and significantly associated with weight (p=0.02). X07868-3750exA→G was highly significantly associated with birth weight (p=0.009). However, the latter of these two SNPs were low frequency SNPs for which the heterozygote and rare homozygote samples had been grouped. This was for statistical reasons and is not necessarily biologically appropriate, as no hypothesis can be assumed about the effect of each genotype. However, as the results indicate either a massive effect for the rare homozygote component, a mild effect for the heterozygote component or a combination of both the result is still relevant, but would need further work to determine whether the effect is allele-dose dependent, or whether it occurs in only one genotype.

6.3.1.2 Linear regression

As with the one-way ANOVA analyses, the association between genotype and phenotype was tested for eleven SNPs and five phenotypes. Again no correction for multiple testing was applied, and therefore the results should be treated with caution. This analysis

tests for an allele-dose effect, determining whether there is a correlation between phenotype and number of alleles.

No significant associations between genotype and phenotype were observed in the Hertfordshire men. In Hertfordshire women several significant (non-corrected) associations were observed. Y13633-2482A→C was significantly associated with BMI (p=0.02) and weight (p=0.03). X07868-266C→T was highly significantly associated with BMI (p=0.002) and significantly associated with weight (p=0.02). X07868-3750exA→G was highly significantly associated with birth weight (p=0.009). Apart from the association between Y13633-2482A→C and weight, and the lack of association between Y13633-2722C→T and BMI, these results agree with those from the one-way analysis of variance. This type of analysis requires a trend across all three genotypes, while the one-way ANOVA tests only for a significant difference between any two.

6.3.1.3 Multiple testing

Eleven SNPs and five phenotypes were tested by two methods, a total of 110 tests. Applying a Bonferroni correction this would make the minimum required *p*-value 0.00045 (0.05/110). The lowest *p*-value is that between body mass index and X07868-266C/T in women (p=0.002, Table 6-5). However, the Bonferroni correction tends to greatly over-correct in the case of associated markers (Cardon & Bell 2001). In addition, the relationship between phenotypes and between analysis types means that the tests are not all truly independent (e.g. BMI is derived from weight and height). The correction value that should be applied is somewhere between 0 and 110, although the exact value cannot easily be defined. The data presented here are therefore not corrected (to avoid conservative analysis leading to false negatives), and should be treated with caution as potentially containing false positives. A correction value of ≤25 would be required for the most significant association to remain significant (p=0.002 → p=0.05).

6.3.2 LD analyses in the Hertfordshire cohort

The linkage disequilibrium (LD) analyses in the Hertfordshire cohort largely agree with those in NPHSII. Disagreements appear to be a trend for larger |D'| values in the Hertfordshire cohort, with the biggest differences occurring where allele frequencies are low.

The overall pattern is for fairly high linkage disequilibrium across the gene. This is consistent with the expectation that LD extends over distances up to 100kb (Collins *et al.* 1999).

The high levels of linkage disequilibrium between Y13633-2482A→C and Y13633-2722C→T (|D'| = 0.96) suggest that the observed association with BMI in women at these two sites may be a consequence of them marking a single causal allele on the same haplotype. X07868-266C→T is in comparatively less LD with these two SNPs (|D'| = 0.57 and 0.61 respectively), and therefore may mark an independent causal site.

The high levels of LD across these SNPs also help to indicate that these tests are not truly independent in the multiple testing sense (section 6.3.1.3). Clearly this cannot be considered as a single test as each SNP contributes both shared information and additional, unique information.

6.3.3 Conclusions

This study was not a replication of the NPHSII study (Chapter 5), because the Hertfordshire cohort contains men and women of a different age range. However, it provides additional data that may contribute to a future meta-analysis. The male component of the Hertfordshire cohort (n=626) is less than a quarter the number of the NPHSII cohort (n=2743). This may explain the lack of significant associations: a quick check of this number of men from NPHSII shows that there is no significant association between BMI and genotype at the L15440-6815A→T locus if only 626 NPHSII men are analysed (P=0.778 compared to 0.00012 when 2743 men are tested). However, the genotyping of the Hertfordshire cohort enabled an investigation of the influence of *IGF2* SNP genotypes on BMI in women, and also their effects on early life phenotypes in both men and women.

Interestingly, although the number of women tested is small (n=428), several significant associations were observed. Whether these remain significant after a correction for multiple testing depends on the correction used (section 6.3.1.3). While the Hertfordshire results on their own may not provide significant supporting evidence for a role for *IGF2* in predisposition to weight gain in men, they do suggest that *IGF2* SNP genotypes may influence weight in women. However, the associated SNPs and trends are unique to the Hertfordshire women.

In summary, the analysis of genotype/phenotype associations in the Hertfordshire cohort provided some evidence for a role for *IGF2* in predisposition to weight gain. The problem of correction for multiple testing needs to be addressed for this type of study to

enable meaningful, repeatable results to be achieved. An additional 1292 Hertfordshire men and 1250 Hertfordshire women will be available for analysis in the near future which will enable the power of studies to be strengthened, and the data presented here will contribute to that analysis. A haplotype analysis of this data is planned (depending on the release of data from the MRC Environmental Epidemiology Unit, Southampton), which will provide greater sensitivity to prove or disprove these trends.

Chapter 7 : General Discussion

7.1 New methods developed

7.1.1 High-throughput genotyping

A system has been developed which enables efficient, conservative high-throughput genotyping at minimal cost. High-throughput genotyping methodology using 384-well liquid-phase detection of fluorescent oligonucleotide probes in several variations is a method of choice for many groups, but the high associated costs of equipment and probes are a limiting factor. One such method is the 5'-nuclease TaqMan assay, which distinguishes between alleles on the basis of allele-specific probe degradation, and consequently de-quenching of fluorescence, by Taq polymerase (Lee *et al.* 1993). A second method, called Molecular Beacons, uses self-complementary allele-specific probes that self-quench when not hybridised and fluoresce if they match the allele (Tyagi *et al.* 1998). The reporter-quencher nature of these probes requires two fluorescence labels, and the requirement for two probes doubles that number again. This is a cost per assay of several hundred pounds, and there is also the cost of the detection system, currently around £63,000 for an ABI9700 from PE Applied Biosystems.

The system presented here has a low cost per assay of less than £50 for primers. The cost of Taq polymerase is similar for this system and for fluorescence-based systems.

Equipment costs are also roughly equivalent for the two approaches if a Fluorimager (Molecular Dynamics) is used, but significantly cheaper digital systems can be obtained for a few thousand pounds and are sufficient for this system. The minimum costs for both equipment and assay reagents are therefore lower for this system than for alternative high-throughput genotyping approaches. A significant advantage of this system is that all reagents are easily available from many suppliers, which ensures no limitation to supply and minimises the costs.

The future requirements for greater throughput can be achieved by *in silico* automation of assay design and gel image analysis. The semi-automated image analysis system presented here using allele ratio clustering provides a promising step towards full automation of the analysis side, which will require some development from external programmers for completion. Specifically, the system presented here will analyse data from the point at which band intensity data is available. The requirements for full automation are in the lane and band detection software, currently provided by Phoretix International.

7.1.2 Quantitative-competitive RT-PCR

The quantitative-competitive RT-PCR assay presented here will require a substantial number of samples in order to provide sufficient examples of each genotype for all associated SNPs in the *IGF2* gene. Currently these are not available – the main source of biallelic *IGF2* expression in the adult is in the liver (Wu *et al.* 1997b), and obtaining this material is difficult. Monoallelic expression exists in other tissues, including peripheral blood leukocytes, although tests of RNA extracted from blood showed only very low levels of *IGF2* expression (data not shown). This assay is therefore awaiting the availability of samples.

7.2 Map of polymorphism in IGF2

The map of polymorphism in the *IGF2* gene presented here is currently the most comprehensive available, and includes compiled sequence and polymorphism data from published sources and sequence and polymorphism data identified here for the first time. A set of nineteen polymorphisms across the 30kb gene include two homopolymeric tract length

polymorphisms, one complex $(CA)_n$ repeat polymorphism and 16 single-nucleotide polymorphisms. For the first time a complete, annotated sequence of the *IGF2* gene has been generated, enabling the mapping of data in context.

In silico mutation detection was significantly less effective than the lab-based approaches of single-strand conformation polymorphism analysis and denaturing high performance liquid chromatography. However, the number of sequences in GenBank is increasing exponentially and currently exceeds 20,000,000,000 (http://www.ncbi.nlm.nih.gov/entrez/). *In silico* sequence alignment approaches to the detection of sequence variants may therefore become the most efficient method in the near future. In addition, single-nucleotide polymorphism (SNP) databases will become more useful with time as the number of accessions increases.

7.3 Genetic Epidemiology of NPHSII

The genotyping of eleven polymorphisms (including three which pre-date this project) identified four significant associations between genotype and BMI. An over-conservative Bonferroni correction still leaves two of these associations significant, and so there is strong evidence to support the hypothesis that single-nucleotide polymorphisms in the *IGF2* gene influence body mass index in middle-aged men. Multiple regression analysis indicates that three of the SNPs are independently significantly associated with body mass index, while the fourth SNP is in strong linkage disequilibrium with one of the three and does not account for any variation in BMI above that accounted for by the other SNPs.

Linkage disequilibrium (LD) within *IGF2* was observed to be fairly high, which fits with estimates of LD extending approximately 100kb (Collins *et al.* 1999). However, the assumption that markers at that spacing would detect associations with disease-causing sites would appear to be false. Seven of the eleven SNPs genotyped in *IGF2* showed no significant association with BMI in NPHSII men, indicating that to detect an association all available SNPs should be genotyped. This, however, introduces the difficult issue of multiple testing. The more SNPs tested, the greater the chance of both true positives and false positives. Conversely, if a minimal set of SNPs is genotyped to prevent this problem, the risk of a false negative is high. It seems logical, though, that if many sites are tested, then a truly important gene will be detected by multiple significant associations (as has occurred here), while false

positives will not cluster. The excessively stringent nature of the Bonferroni correction in the case of associated markers, and the lack of availability of a generally applicable alternative (Cardon & Bell 2001) means that either studies are published uncorrected (often simply reporting the one significant association with no indication of other polymorphisms tested), or useful data is wasted.

A simple haplotype analysis confirmed the relative independence of the intron 2 SNP (Y13633-1156T→C) as an older SNP frequently occurring on haplotypes both with and without the rare alleles of the other three significantly associated SNPs. This indicates that this polymorphism arose first, and the other three occurred later, subdividing two major haplotypes into three. The fact that this SNP is older and probably non-functional means that the significance of the association is expected to be less, as LD between this SNP and a functional site is likely to have decayed with time. The younger SNPs will be more strongly associated with the causal sites that they mark – this is borne out in the association data.

The haplotype analysis also confirmed that haplotype frequencies are associated with body mass index, with the "light" alleles at each of the four significantly associated SNPs present on a common haplotype that is significantly more frequent in the lowest body mass index quartile and significantly less frequent in the highest body mass quartile than would be expected under independence.

Subsequent haplotype analyses on this data set (Rodriguez, Gaunt and Day, unpublished) has demonstrated that this type of analysis applied to random sets of five or six SNPs shows very significant differences in haplotype frequency distributions between different BMI quartiles. The analysis of all eleven SNPs generates too many rare haplotypes for useful analysis, but by random subdivision of the full data set into two parts and exclusion of rare haplotypes (which reduce power), significance is observed in both parts ([L15440-8173C→T, Y13633-2722C→T, X07868-820G→A, X07868-2207C→T, X07868-3750exA→G]: P=0.000055 and [L15440-6815A→T, Y13633-1156T→C, Y13633-1252T→C, Y13633-2482A→C, X07868-266C→T and X07868-1926C→G]: P=0.000016). This approach avoids the multiple testing issues by asking the question: are haplotypes in *IGF2* associated with body mass index in middle-aged men? The significant associations between haplotype frequencies and body mass index confirm the role of *IGF2* in weight determination.

7.4 Genetic Epidemiology of the Hertfordshire cohort

The same eleven single-nucleotide polymorphisms (including one which pre-dates this project) were genotyped in the Hertfordshire cohort, a smaller population, but one that was potentially useful because of the inclusion of women and the availability of early life phenotype data. However, the data did not provide any additional evidence for the role of *IGF2* in weight determination in men (although some trends agreed with those in NPHSII). Some evidence for a role of *IGF2* in body mass index determination in women was found for three SNPs (different from those identified to have an association with body mass index in NPHSII men). However, correction for multiple testing is likely to make these associations non-significant (depending on the correction factor used). The over-conservative nature of the Bonferroni correction for multiple testing seems inappropriate, but if applied to these data then none of the associations are significant.

Three other studies have investigated one of these SNPs in relation to weight phenotypes. Roth *et al* (2002) found no association between *ApaI* (X07868-820G→A) genotype and body mass index in a mixed population of 500 men and women aged 19 to 90 years. However, the evidence presented here suggests that different SNPs associate in men and women, and also that these effects are only significant in large populations. The use of such a broad age range may also confound detection of associations, as the importance of this gene in weight determination may change throughput an individual's life.

Another study in a population of women identified significant associations between *ApaI* AA genotype and higher BMI, fat mass and percentage fat (Sun *et al.* 1998). This is consistent with our observations in *IGF2* SNPs. In Hertfordshire women, X07868-820G→A (*ApaI*) showed a trend for higher BMI with the rare (2 = A) allele. This is opposite to the significant association between the rare allele and lower BMI seen in NPHSII men (Chapter 5).

Ukkola *et al.* (2001) found a lower fat mass in X07868-820A allele carriers than G allele carriers before and after overfeeding in male monozygotic twins (19-23 years old). This agrees with the NPHSII results, which indicate an association between X07868-820AA genotype and lower body mass index. This twin study confirms the direction of trend between genotype and BMI in men, despite the much lower age range.

Linkage disequilibrium in the Hertfordshire cohort was similar to that observed in NPHSII, with some of the SNPs showing higher levels of |D'| in the Hertfordshire data. This

is probably the result of the smaller numbers in Hertfordshire reducing the power of the statistic, which requires estimation of haplotype frequencies, and is therefore likely to be less accurate with smaller numbers from which to estimate.

The evidence arising from the genotyping of the Hertfordshire cohort is supportive of a role for *IGF2* in weight determination, but not significant with this number of samples. The trends are consistent with those observed in NPHSII men and in women by Sun *et al.* (1998).

7.5 Relationship to other studies

The studies investigating a role for *IGF2* in weight are summarised in Table 7-1. No other studies are on the scale of the NPHSII study presented here, and many include broad age ranges or mixed genders. Several studies show an association between IGF-II levels in the body and body mass index or other weight-related phenotypes. The data of Sun *et al.* (1998) and Ukkola *et al.* (2001) support the findings presented here, while the data of Roth *et al.* (2002) is contradictory. However, direct comparison of studies is not usually possible due to differences in ethnicity, age, gender, study size and selection criteria. The weight of the evidence in the literature is therefore consistent with the finding of this study that the *IGF2* gene has a role in weight determination.

Table 7-1: Studies investigating the role of *IGF2* in weight

Study	Observations
Hausman *et al.*, (1991)	**Pig foetuses:** Liver and muscle IGF-II concentrations lower in pre-obese (smaller) pig foetuses than lean (larger) foetuses.
Frystyk *et al.* (1995)	**Men and women – 31 controls, 33 obese and 28 severely obese:** Serum total IGF-II levels significantly higher in obese men than normal men and higher in severely obese women than in moderately obese women.
O'Dell *et al.* (1997)	**2605 men 51-62 years old:** Significant associations between higher BMI and X07868-820GG. Also association between lower serum IGF-II and X07868-820GG in 92 individuals.
Radetti *et al.* (1998)	**21 obese and 32 control children:** IGF-II levels raised in obesity.
Argente *et al.* (1997)	**65 obese prepubertal children and 174 age-matched controls:** IGF-II levels raised in obesity.
Sun *et al.* (1998)	**286 obese and 82 non-obese women:** Significant association between X07868-820AA genotype and higher BMI in normal weight women. The opposite direction of trends to those observed in men in NPHSII agrees with trends observed in Hertfordshire.
Cruickshank *et al.* (2001)	**465 adults from three ethnic groups:** No significant associations between IGF-II levels and BMI or WHR in normal weight individuals
Ukkola *et al.* (2001)	**Twelve pairs of monozygotic male twins 19-23 years old:** Fat mass significantly higher in X07868-820GG homozygotes both before and after over-feeding.
Jones *et al.* (2001)	**Mice:** Decreased *IGF2* expression associated with increased fat deposition and obesity.
NPHSII cohort (Gaunt *et al.*, 2001) and presented here	**2743 men 51-62 years old:** Independent significant associations between higher BMI and L15440-6815AA, Y13633-1156CC, X07868-820GG. Non-independent significant association between BMI and genotype for X07868-1926CC.
Hertfordshire cohort, presented here	**626 men and 428 women 55-63 years old:** Some significant associations between genotype and BMI for Y13633-2482A→C, Y13633-2722C→T and X07868-266C→T may disappear after correction for multiple testing. Trends tend to agree with NPHSII for men, opposite trends are observed for women.
Roth *et al.* (2002)	**500 men and women 19-90 years old:** Significant associations between higher fat mass and X07868-820AA in 427 Caucasians, no association with BMI.

7.6 Conclusions

The *IGF2* gene has been thoroughly investigated as a candidate gene for weight determination. Novel single-nucleotide polymorphisms (SNPs) have been identified, and data collected from other sources to provide a comprehensive map of genetic variability in the

gene. This variability has been used to test for association with body mass index and other phenotypes in two cohorts: Northwick Park Heart Study II (NPHSII), containing 2743 men aged 51 to 62 years old, and the Hertfordshire cohort containing 626 men and 428 women aged 57 to 65 years old. Strong evidence for a role of *IGF2* in determination of BMI in men was obtained from NPHSII, with three independently significantly associated SNPs identified. Haplotype analysis confirmed a role for haplotypes in the *IGF2* region influencing body mass index.

Results from the Hertfordshire cohort provided no additional significant evidence, but suggested a potential role of *IGF2* in BMI determination in women, with different SNPs showing some association with that phenotype. Trends for BMI with genotype in men were largely consistent between NPHSII and Hertfordshire cohorts. The project was achieved by the development and adoption of a system of high-throughput genotyping methodology, which was implemented here for the first time.

The difficulty of detecting minor (~1%) polygene effects in small populations are indicated by the loss of statistical power when investigating a set of *IGF2* SNPs in a population of 626 Hertfordshire men compared to 2743 NPHSII men. Many projects use relatively small cohorts to investigate small genetic effects in complex disease, with a major risk of false positives from multiple testing. The UK Biobank of up to 50,000 UK individuals (http://www.mrc.ac.uk/index/public_interest/public-consultation/public-biobank_consult.htm) is likely to be of major significance in enabling powerful association studies of complex polygenic diseases in the future.

7.7 Further work

The results presented here indicate two primary directions in which the project should continue. The first is an investigation of the role of *IGF2* in determination of BMI in women, as suggested by the data from the (relatively) small Hertfordshire cohort. A resource of 3900 women collected as part of the British Women's Heart and Health Study (BWHHS) will soon become available, and should provide ample statistical power to detect any true associations. This will resolve the question of whether the associations observed here were genuine or an artefact of multiple testing. In addition to BWHHS at least a further 2400 samples are due to

become available from Hertfordshire as an addition to the existing DNA bank. These samples may also be used to provide further evidence for these associations (or disprove them).

The second obvious route for this project to follow is an investigation of functional aspects of *IGF2*. L15440-6815A→T is within the P1 (adult biallelic) promoter, and may potentially mark an effect there. Promoter studies would help to identify potential mechanisms for transcriptional regulation of *IGF2* that may impact upon body mass index. The quantitative-competitive RT-PCR assay that has been developed here will also provide a means for investigating the biological mechanisms underlying the associations observed. By comparing transcription levels in individuals of different genotype, it may be possible to identify which polymorphisms are relevant. X07868-820G→A and X07868-1926C→G are both close to (but not within) elements of the *IGF2* 3'UTR endonucleolytic cleavage site. If these mark a functional polymorphism which alters cleavage, then they are likely to associate with different levels of uncleaved transcript detectable by QC-RT-PCR. It is possible that one of these two SNPs may influence RNA folding and therefore directly influence cleavage.

The final area of further work would leave this project, and attempt to apply the same high-throughput genotyping approaches to other candidate genes in an attempt to identify combinations of genes that are important. A multi-gene analysis may be the most powerful way of identifying the underlying genetic basis of complex diseases.

Appendices

Appendix A: primer sequences

Table A.1: SSCP primers
Primer name contains either "EX*x*" referring to exon *x* of the *IGF2* gene or "UTR" referring to the 3'UTR. GenBank sequence and start nucleotides are shown and pairs of primers are highlighted as each pair of rows in the table.

Primer name	Primer sequence	Theor. Tm	GenBank ID	Primer start	For/Rev	Product length
IG2EX1/1	gccctgttcctgaagctctg	61.4	L15440	6935	F	194
IG2EX1/2	ccatggacggctgctgccga	65.5	L15440	7128	R	
IG2EX1/3	gggtggacggccggacactg	67.6	L15440	7057	F	221
IG2EX1/4	ggcaacagcttggccgatgg	63.5	L15440	7277	R	
IG2EX3/1	aagccgctgcccagatcctg	63.5	Y13633	1086	F	163
IG2EX3/2	cggtggtgactcttcggccc	65.5	Y13633	1248	R	
IG2EX3/3	gaggtggattccagccccca	63.5	Y13633	1263	F	230
IG2EX3/4	tcagggtgcctgagacactc	61.4	Y13633	1492	R	
IG2EX4/1	gtttagtcattaatcacggt	51.2	X53038	228	F	221
IG2EX4/2	gctctggggacttcgtagga	61.4	X53038	448	R	
IG2EX4/3	ccccaaccccgcgcacagcg	69.6	X53038	384	F	224
IG2EX4/4	gaagacccggggacaatgcc	63.5	X53038	607	R	
IG2EX4/5	caggaaagcgaccgggcatt	61.4	X53038	545	F	253
IG2EX4/6	aggggcgcagaggcggaggg	69.6	X53038	797	R	
IG2EX5/1	ccccgctcttggctcgggtt	65.5	X03562	1668	F	244
IG2EX5/2	gagcgcgggcaggcgtgggc	71.7	X03562	1911	R	
IG2EX5/3	ccggcggagctgcgtgaggc	69.6	X03562	1829	F	219
IG2EX5/4	aagaggaggcggcggggaat	63.5	X03562	2047	R	
IG2EX5/5	ccgtcccgggcgccgtccgc	73.7	X03562	1982	F	215
IG2EX5/6	aaggggagcggcccgaggct	73.7	X03562	2196	R	
IG2EX5/7	gcgggcgcccagctcggttt	67.6	X03562	2101	F	208
IG2EX5/8	gcgggcgcccagctcggttt	67.6	X03562	2308	R	
IG2EX5/9	cccgggctcgcgacggcaga	69.6	X03562	2254	F	271
IG2EX5/10	aggcgagagggcgggcgtga	67.6	X03562	2524	R	
IG2EX5/11	ccagctcctagcctccgact	63.5	X03562	2474	F	233
IG2EX5/12	gctgttgtatcaaggataga	53.2	X03562	2706	R	
IG2EX5/13	cactctgtctctcccactat	57.3	X03562	2656	F	220
IG2EX5/14	ccgagtcgcgggggccgaat	67.6	X03562	2875	R	
IG2EX5/15	cccgctctgccccgtcgcac	69.6	X03562	2836	F	213
IG2EX5/16	ggccgggcgtcgccgcaagc	71.7	X03562	3048	R	
IG2EX6/1	cataaaactgaggcactgac	55.3	X03562	3855	F	243
IG2EX6/2	gcgggctggcggctgcaggg	71.7	X03562	4097	R	
IG2EX6/3	cggcccacgcctgggcctcg	71.7	X03562	4056	F	210
IG2EX6/4	agcatgcagcggtgcggagc	65.5	X03562	4265	R	
IG2UTR1	agatggccagcaatcggaag	59.4	X07868	215	F	200

Primer name	Primer sequence	Theor. Tm	GenBank ID	Primer start	For/Rev	Product length
IG2UTR2	aggggccgaggagagtagc	65.5	X07868	414	R	
IG2UTR3	gtgccccgcctccccgaaac	67.6	X07868	373	F	213
IG2UTR4	aaacattaaactaacccccct	51.2	X07868	585	R	
IG2UTR5	gttttaagagggttgttgtg	53.2	X07868	535	F	193
IG2UTR6	caaattcctttattttgcca	49.1	X07868	727	R	
IG2UTR7	tattaaaaacgaattggctg	49.1	X07868	679	F	141
IG2UTR8	ccttcttttctctttgctgg	55.3	X07868	819	R	
IG2UTR9	agaaatcacaggtgggcacg	59.4	X07868	825	F	167
IG2UTR10	tcctttggtcttactgggtc	57.3	X07868	991	R	
IG2UTR11	ccatcactaaaaatcacaga	51.2	X07868	941	F	177
IG2UTR12	gctgcggggatgcataaagt	59.4	X07868	1117	R	
IG2UTR13	aggccaaagtcccgctaaga	59.4	X07868	1887	F	231
IG2UTR14	aaggaggccagcctcacaag	61.4	X07868	2117	R	
IG2UTR15	ctaattccatctttccacca	53.2	X07868	2065	F	257
IG2UTR16	ccccaagatcttccttccag	59.4	X07868	2321	R	
IG2UTR17	cctgactccctggtgtgctc	63.5	X07868	2282	F	216
IG2UTR18	acttcctaccccagaactcc	59.4	X07868	2497	R	
IG2UTR20	acttgcagaattacatagag	44.5	X07868	2544	F	195
IG2UTR21	cttgcttttgtcactgcccc	60.5	X07868	2738	R	
IG2UTR22	ctcagaaaccaaattaaacc	49.5	X07868	2681	F	209
IG2UTR23	acttctccaaggggctcat	60	X07868	2889	R	
IG2UTR24	gagggtggagcctcctgggg	67.5	X07868	2841	F	206
IG2UTR25	gcctcaggccagccaggagc	69	X07868	3046	R	
IG2UTR26	aggtgtcaggagggtgctcg	62	X07868	2991	F	203
IG2UTR27	cctggagacaaggcagggtg	62.5	X07868	3193	R	
IG2UTR28	ctctccctcggtgacatctt	56	X07868	3141	F	205
IG2UTR29	tcaccagacccgtgggctcc	67.5	X07868	3345	R	
IG2UTR30	ccaggtctgcccatcgacca	66	X07868	3291	F	205
IG2UTR31	accaggacccagaagcctca	61.5	X07868	3495	R	
IG2UTR32	cccctgcacgcagcccgact	72	X07868	3441	F	215
IG2UTR33	gccaatgtgggttccacaat	60	X07868	3655	R	
IG2UTR34	cattggacagaagcccaaag	58	X07868	3591	F	200
IG2UTR35	gatggagagccacgactagg	57	X07868	3790	R	
IG2UTR36	tgttcccgggctgcctgtct	67	X07868	3741	F	209
IG2UTR37	aggagctgagttgagtcaaa	52	X07868	3949	R	
IG2UTR38	cacctgtgcctgccgcctcg	71.5	X07868	3891	F	200
IG2UTR39	ccctccggcagtccagtacc	64	X07868	4090	R	
IG2UTR40	atcttctgaggtgttccact	50.5	X07868	4041	F	116
IG2UTR41	aggccttccttcccttccc	65.5	X07868	4156	R	

Table A.2: DHPLC primers for Intron 3
Data from Primer3 desgin (http://genome.wi.mit.edu)

IGF2-INT3/1+2 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any    3'    rep seq
LEFT PRIMER       1454   18    59.23    61.11  5.00  0.00 11.00 AACTGCGAGGCAGAGAGG
RIGHT PRIMER      1850   22    59.82    36.36  3.00  1.00 12.00 TTTTTAGGTTCTTCCCCAATGA
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 397, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00
       TARGETS (start, len)*: 1472,350
```

IGF2-INT3/3+4 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any    3'    rep seq
LEFT PRIMER       1792   20    59.28    55.00  2.00  0.00 12.00 GGGCAGAAAGTGGGTAGAGA
RIGHT PRIMER      2200   20    60.44    50.00  5.00  3.00 11.00 TGCACACTCCAAAGGACTGA
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 409, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 1828,350
```

IGF2-INT3/5+6 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any    3'    rep seq
LEFT PRIMER       2107   20    60.25    50.00  4.00  0.00 11.00 AGAAGCCGGGAGTGTTCTTT
RIGHT PRIMER      2579   20    60.27    45.00  4.00  0.00 12.00 GGCCTTTCTCATTCCCATTT
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 473, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 2180,350
```

IGF2-INT3/7+8 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any    3'    rep seq
LEFT PRIMER       2496   20    59.57    50.00  2.00  1.00 12.00 GTAAAATGCTGGGCTTGGTC
RIGHT PRIMER      2943   19    59.95    63.16  4.00  3.00 11.00 CGGCTCCCCTCTAGTCAAG
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 448, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00
       TARGETS (start, len)*: 2559,350
```

IGF2-INT3/9+10 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       2889   20   59.53   55.00  3.00  1.00 12.00 GAGGAGCCAGTCTTGAGGAA
RIGHT PRIMER      3316   20   60.55   55.00  3.00  0.00 12.00 CAGTTCCGAGTTGTGGGTTC
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 428, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 3.00
       TARGETS (start, len)*: 2924,350
```

IGF2-INT3/11+12 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       3262   20   60.24   55.00  4.00  1.00 10.00 GGTGTAATTGTCCCCAGCAC
RIGHT PRIMER      3679   20   60.07   50.00  5.00  3.00 12.00 CAAGAACATCTGGCCCATCT
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 418, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 3296,350
```

IGF2-INT3/13+14 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       3621   18   60.19   61.11  3.00  0.00 12.00 AGAAGGCTGGTGGGAAGG
RIGHT PRIMER      4046   18   60.31   66.67  7.00  3.00 11.00 CCCTAAGGGGACCTGTGG
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 426, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 3658,350
```

IGF2-INT3/15+16 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       3950   20   59.99   55.00  4.00  2.00 12.00 CACTGTAGTTGCCCCAGGAT
RIGHT PRIMER      4444   18   62.40   61.11  4.00  0.00 10.00 AGGCTATTCCCCGGCTTG
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 495, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 4028,350
```

IGF2-INT3/17+18 Y13633

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       4326   20   60.25   45.00  3.00  0.00 12.00 TGGTTTCACCCTTGTTTGGT
RIGHT PRIMER      4754   20   59.96   45.00  7.00  3.00 12.00 GGCTGAAATTCTGCTTTTCG
SEQUENCE SIZE: 4992
INCLUDED REGION SIZE: 4992

PRODUCT SIZE: 429, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 4426,300
```

IGF2-INT3/19+20 U80851

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       3378   19   60.43   52.63  4.00  0.00 12.00 GCCCTTGGGTTTTCTTCCT
RIGHT PRIMER      3851   20   59.93   45.00  6.00  1.00 11.00 CCGGAAATTAACGTCCAAGA
SEQUENCE SIZE: 5952
INCLUDED REGION SIZE: 5952

PRODUCT SIZE: 474, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00
       TARGETS (start, len)*: 3455,350
```

IGF2-INT3/21+22 U80851

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       3808   21   60.74   42.86  6.00  2.00 11.00 CATTTCGACGGGTCACAATAA
RIGHT PRIMER      4256   20   60.53   45.00  4.00  3.00 11.00 AACAGCAAATGTCCCATGCT
SEQUENCE SIZE: 5952
INCLUDED REGION SIZE: 5952

PRODUCT SIZE: 449, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00
       TARGETS (start, len)*: 3831,350
```

IGF2-INT3/23+24 U80851

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       4161   20   60.70   55.00  3.00  3.00 12.00 GCTTACTGTTTCCCGCACAC
RIGHT PRIMER      4625   20   60.03   60.00  5.00  3.00 10.00 ACCTGGGTGGCACTGTCTAC
SEQUENCE SIZE: 5952
INCLUDED REGION SIZE: 5952

PRODUCT SIZE: 465, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00
       TARGETS (start, len)*: 4236,350
```

IGF2-INT3/25+26 U80851

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       4579   18   60.21   61.11  6.00  1.00 12.00 CTGGCCACGAATCAGGTC
RIGHT PRIMER      4989   20   59.73   50.00  3.00  2.00 12.00 AGCGTCCAGAACATCCAAGT
SEQUENCE SIZE: 5952
INCLUDED REGION SIZE: 5952

PRODUCT SIZE: 411, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00
       TARGETS (start, len)*: 4605,350
```

IGF2-INT3/27+28 U80851

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       4940   20   59.55   50.00  3.00  0.00 11.00 TCACACTGGACCCTTTTCCT
RIGHT PRIMER      5395   20   59.55   55.00  4.00  0.00 12.00 GCAGGGACCATAAAGTCAGG
SEQUENCE SIZE: 5952
INCLUDED REGION SIZE: 5952

PRODUCT SIZE: 456, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 3.00
       TARGETS (start, len)*: 4969,350
```

IGF2-INT3/29+30 U80851

```
Using mispriming library humrep_and_simple.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER       5287   20   60.04   50.00  3.00  3.00 11.00 TGTCTCCCATCCTTCCAAAG
RIGHT PRIMER      5754   21   59.98   42.86  4.00  0.00 12.00 CAGTACATTTGGGGAAAGCAA
SEQUENCE SIZE: 5952
INCLUDED REGION SIZE: 5952

PRODUCT SIZE: 468, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00
       TARGETS (start, len)*: 5375,350
```

Table A.3: Sequencing primers

These are selected primer pairs from the SSCP (Table A.1) and DHPLC (Table A.2) primers.

SNP	Location	Sequencing primers	Product size
1156 T/C	GenBank Y13633 nt. 1156 T/C	EX3/1 5′-AAGCCGCTGCCCAGATCCTG-3′ AluC 5'-CCTGGGCCACAGGCCACAGCAGCTCACC-3'	315
2482 A/C	GenBank Y13633 nt. 2482 A/C	INT3/5 5′-AGAAGCCGGGAGTGTTCTTT-3′ INT3/6 5′-GGCCTTTCTCATTCCCATTT-3′	473
2722 C/T	GenBank Y13633 nt. 2722 C/T	INT3/7 5′-GTAAAATGCTGGGCTTGGTC-3′ INT3/8 5′-CGGCTCCCCTCTAGTCAAG-3′	448
5345 polyT	GenBank U80851 nt. 5345 polyT	INT3/29 5′- TGTCTCCCATCCTTCCAAAG-3′ INT3/30 5′- CAGTACATTTGGGGAAAGCAA-3′	468
266 C/T	GenBank X07868 nt. 266 C/T	EX9/2 5′-CTGCGTGCCCGCCGGGGTCA-3′ UTR/2 5′-AGGGGGCCGAGGAGAGTAGC-3′	320
556 polyC	GenBank X07868 nt. 556 polyC	UTR/3 5′-GTGCCCCGCCTCCCCGAAAC-3′ UTR/8 5′-CCTTCTTTTCTCTTTGCTGG-3′	447
1926 C/G	GenBank X07868 nt. 1926 C/G	UTR/13 5′-AGGCCAAAGTCCCGCTAAGA-3′ UTR/16 5′-CCCCAAGATCTTCCTTCCAG-3′	434

SNP	Location	Sequencing primers	Product size
2207 C/T	GenBank X07868 nt. 2207 C/T	UTR/13 5′-AGGCCAAAGTCCCGCTAAGA-3′ UTR/16 5′-CCCCAAGATCTTCCTTCCAG-3′	434
3750ex A/G	GenBank X07868 nt. 3750 A/G	*TGTTCCCGGGGGCACTTGCCGACCAGCCCCTTGCGTCCC* *CAGGTTTGCAGCTCTCCCCTGGGCCACTAACCATCCTGG* *CCCGGGCTGCCTGTCT* *UNDERLINED SEQUENCE IS PUBLISHED X07868* UTR/32 5′-CCCCTGCACGCAGCCCGACT-3′ UTR/39 5′-CCCTCCGGCAGTCCAGTACC-3′	724
3944 G/A	GenBank X07868 nt. 3944 G/A	UTR/36 5′-TGTTCCCGGGCTGCCTGTCT-3′ UTR/41 5′-AGGCCTTCCTTCCCCTTCCC-3′	416
4085 G/A	GenBank X07868 nt. 4085 G/A	UTR/36 5′-TGTTCCCGGGCTGCCTGTCT-3′ UTR/41 5′-AGGCCTTCCTTCCCCTTCCC-3′	416

Table A.4: ARMS primers

SNP	Location	ARMS Primers	Anneal temp., $MgCl_2$, cycle number	Control Primers (see SSCP primers)
8173 *C/T*	GenBank L15440 nt8173	**8173FA1** CAGCTGCTCCCCAAGTCTCCCAGAC **8173FA2** CAGCTGCTCCCCAAGTCTCCCAGAT **8173RC** CCTTGGCTCCAGGCTTCAGTCTCTGC	65°C, 1.25mM, 35 cycles	IGF2UTR24 and IGF2UTR25
1156 *T/C*	GenBank Y13633 nt1156	**1156RA1** CCCCCTCCCCATACACCCGAA **1156RA2** CCCCCTCCCCATACACCCGAG **1156FC1** TCTCTCCATCAGCACCGGGAAGC **1156RC2** AAGGAATCCACCTCCTCCCACACAAG	72°C, 1.5mM, 20 cycles	None - integrated control
2482 *A/C*	GenBank Y13633 nt. 2422	**Y13633-2482FA1** GGCGCTCTGAGGACCCTTGGATAA **Y13633-2482FA2** GGCGCTCTGAGGACCCTTGGATAC **Y13633-2482RC1** AGACTCACCCCATGGAGCCTGGC **Y13633-2482FC2** GGGAGTGAGCGCAAGGTCAGCG	72°C, 1.5mM, 20 cycles	None - integrated control
2722 *C/T*	GenBank Y13633 nt. 2722	**Y13633-2722RA1** GCACCTCACCAGGGCAACGATG **Y13633-2722RA2** AAGCACCTCACCAGGGCAACGATA **Y13633-2722RC2** TGGCTCCTCCCCCACTACTTCCA **Y13633-2482FC2** GGGAGTGAGCGCAAGGTCAGCG	72°C, 1.5mM, 20 cycles	None - integrated contol
266 *C/T*	GenBank X07868 nt266	**266C** CGTTCCAATATGACACCTGGAAGCAGT **266A1** TTATGAGGCTGCAGGATGGTGTCG **266A2** TTATGAGGCTGCAGGATGGTGTCA	66°C, 1.5mM, 20 cycles	IGF2UTR36 and IGF2UTR39
1926 C/G	GenBank X07868 nt1926	**1926C** ACACCCCCACAAAATTAGATGAAAACAA **1926A1** TATGTGCAGTGGAGATGGGAACAGGTGG **1926A2** TATGTGCAGTGGAGATGGGAACAGGTGC	60°C, 1.5mM, 20 cycles	IGF2UTR24 and IGF2UTR25
2207 C/T	GenBank X07868 nt 2207	**2207A1** CCCGATACACCTTACTTACTGTGTGTTTGC **2207A2** CCCGATACACCTTACTTACTGTGTGTTTGT **2207C** CGCAAAGATGATCCCTAGGTGTGCT	68°C, 1.5mM, 17 cycles	IGF2UTR38 and IGF2UTR41
3750ex A/G	GenBank X07868 (within unpublished sequence at nt3750)	**3750exC** GCCATTCGGAACATTGGACAGAAGC **3750exA1** TTTAAACCTGGGGACGCAAGGGTCT **3750exA2** TTTAAACCTGGGGACGCAAGGGTCC	60°C, 1.5mM, 17 cycles	IGF2UTR24 and IGF2UTR27

Table A.5: SSCP primer conditions for IGF2 exons
Oligo name is "IG2" (short for IGF2) followed by primer location (eg EX2 refers to Exon 2, UTR refers to 3'UTR) followed by primer number.

Exon	Fragment	Start	Length	Anneal	MgCl2	Additives
EX1	EX1/1-EX1/2	L15440-6935	193	55	1.5	
EX1	EX1/3-EX1/4	L15440-7057	220	55	1.5	1.3M Betaine
EX3	EX3/1-EX3/2	Y13633-1086	162	60	1.5	
EX3	EX3/3-EX3/4	Y13633-1263	229	55	1.5	
EX4	EX4/1-EX4/2	X53038-228	220	45	1.5	
EX4	EX4/3-EX4/4	X53038-384	223	60	1.5	
EX4	EX4/5-EX4/6	X53038-545	252	55	1	
EX5	EX5/1-EX5/2	X03562-1668	243	58	2	1.3M Betaine
EX5	EX5/3-EX5/4	X03562-1829	218	58	1	1.3M Betaine
EX5	EX5/5-EX5/6	X03562-1982	214	55	1	1.3M Betaine
EX5	EX5/7-EX5/8	X03562-2101	207	Not optimised		
EX5	EX5/9-EX5/10	X03562-2254	270	55	1	1.3M Betaine
EX5	EX5/11-EX5/12	X03562-2474	232	Not optimised		
EX5	EX5/13-EX5/14	X03562-2656	219	Not optimised		
EX5	EX5/15-EX5/16	X03562-2836	212	Not optimised		
EX6	EX6/1-EX6/2	X03562-3855	242	45	1	
EX6	EX6/3-EX6/4	X03562-4056	209	60	1.5	
EX7	EX7/1-EX7/2	X03562-5615	290			
EX8	EX8/1-EX8/2	X03562-7474	230			
EX9	EX9/1-EX9/2	X03562-7899	182			
EX9	EX9/3-EX9/4	X03562-8040	207			
UTR	UTR1-UTR2	X07868-215	199	55	1.5	
UTR	UTR3-UTR4	X07868-373	212	55	1.5	
UTR	UTR5A-UTR6A	X07868-480	238	45	1.5	
UTR	UTR7-UTR8	X07868-679	140	55	1.5	
UTR	UTR9-UTR10	X07868-825	166	50	1	
UTR	UTR11-UTR12	X07868-941	176	50	1.5	
UTR	UTR13-UTR14	X07868-1887	230	55	1.5	
UTR	UTR15-UTR16	X07868-2065	256	50	1.5	

Exon	Fragment	Start	Length	Anneal	MgCl2	Additives
UTR	UTR17-UTR18A	X07868-2282	249	55	2.5	
UTR	UTR19A-UTR19B	X07868-2401	200	Not optimised		
UTR	UTR20-UTR21	X07868-2544	194	45	2.5	
UTR	UTR22A-UTR23	X07868-2681	208	Not optimised		
UTR	UTR24-UTR25	X07868-2841	205	65	1.5	
UTR	UTR26-UTR27	X07868-2991	202	60	1	
UTR	UTR28-UTR29	X07868-3141	204	45	1.5	
UTR	UTR30-UTR31	X07868-3291	204	55	1.5	
UTR	UTR32-UTR33	X07868-3441	214	50	1.5	
UTR	UTR34-UTR35	X07868-3591	199	50	1	
UTR	UTR36-UTR37	X07868-3741	208	50	1.5	
UTR	UTR38-UTR39	X07868-3891	199	60	1.5	
UTR	UTR40-UTR41	X07868-4041	115	45	2.5	

Appendix B: scripts

These scripts were created using Microsoft VBScript which runs under Windows Scripting Host (included in Microsoft Windows 98 or later and Microsoft Internet Explorer 4 or later). These scripts can be executed under Windows by double-clicking the file icon if it has a .vbs extension. Alternatively, with minimal modification they could be used on a web page.

Script 1: VBScript to preprocess multiple SNP data files for 2LD.exe

This script is intended to enable easy pairwise LD analysis of multiple SNPs using a program designed for pairwise analysis of single pairs. The script takes a file of genotypes for multiple SNPs (1 to n) and creates data files with pairs (for $x = 1$ to n and $y = (x + 1)$ to n the pairs are x, y. eg for 3 SNPs the pairs are [1, 2], [1, 3] and [2, 3]). The files are passed to 2LD.exe, a program created by JH Zhao of the Institute of Psychiatry, Kings College London. Zhao's program analyses the data to estimate haplotype frequencies, and calculate D, D_{max}, D', |D'| and variance. This script then extracts these values from the output file and collates them into a single file for all the pairwise calculations.

```
' ----------------------------------------------------------------------
' LD-prep
' This script processes multiple genotype files for pairwise LD analysis
' Author: T. R. Gaunt
' ----------------------------------------------------------------------
' INPUT FORMAT - Tab delimited text
'
' Remove data labels - column one must not have IDs - must be data
'
' First row: column label row, no spaces, 8 chars
' Subsequent rows: genotypes in format 11, 12 or 22
' Null values as 0
'
' Script will then take each pair and analyse with 2LD.exe by JH Zhao from
'   Institute of Psychiatry
'   Division of Psychological Medicine
'   Section of Genetic Epidemiology and Biostatistics
'   Kings College London
'   http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.stm
'
' ---
'
' This script does not contain 2LD - this must be obtained separately from
' the above site. This script is freely distributable and is distinct from
' the 2LD program which was created by Jing Hua Zhao.
'
' ----------------------------------------------------------------------
```

```
Option Explicit
  ' ---------
  ' Variables
  ' ---------
Dim infilename, inputfile, inputarray, inputdata, header, labels, columns, rows
Dim summaryname, countera, counterb, SNP1, SNP2, label2
Dim LDinput, LDoutput, outputfile, datarow, data, data1, data2, columncount
Dim a, b, c, d, counterc
Public WinShell
Public LDSummary
Set WinShell = CreateObject("WScript.Shell")
Public filesys
set filesys=CreateObject("Scripting.FileSystemObject")


sub LDexec(LDinput, LDoutput)
  ' ---------
  ' Variables
  ' ---------
  Dim run2LDa
  ' ------------------------------------------
  ' Run LDrun.bat which contains the following
  ' commands to run 2LD.exe with input & output
  ' filenames as parameters:
  '
  '       2LD %1 > %2
  '
  '------------------------------------------
  run2LDa = "LDrun.bat " & LDinput & " " & LDoutput
  WinShell.Run run2LDa,0 ,1
end sub


sub writeresults(SNP1b, SNP2b, LDoutputb)
  ' ---------
  ' Variables
  ' ---------
  Dim inputfileb, run2LDb, LDlineb, LDlinearrayb, Dprimefullb
  Dim Dprimeb, resultsb, resultsarrayb, Varfull, SD, Var, OtherLDlineb
  Dim OtherLDlinearrayb, SignedDprimefull, SignedDprime, samplenumli
  Dim samplenumber, Chisqline, Chisq, degfree, pvalue, Chisqlinesplit
  ' -----------
  ' Read results
  ' -----------
  set inputfileb=filesys.OpenTextFile(LDoutputb)
  resultsb = inputfileb.ReadAll
  inputfileb.Close
  ' -------------
  ' Create arrays
  ' -------------
  resultsarrayb = split(resultsb, VbCrLf)
  LDlineb = resultsarrayb(92)
  OtherLDlineb = resultsarrayb(15)
  samplenumli = resultsarrayb(5)
  Chisqline = resultsarrayb(9)
  Chisqlinesplit = split(Chisqline, " ")
  LDlinearrayb = split(LDlineb, " ")
  OtherLDlinearrayb = split(OtherLDlineb, VbTab)
  samplenumli = split(samplenumli," ")
  ' --------------
  ' Extract values
  ' --------------
  Chisq = Chisqlinesplit(7)
  degfree = Chisqlinesplit(8)
  pvalue = Chisqlinesplit(9)
  Dprimefullb = LDlinearrayb(3)
```

```
    SignedDprimefull = OtherLDlinearrayb(1)
    SignedDprime = left(SignedDprimefull,9)
    SD = LDlinearrayb(6)
    Varfull = LDlinearrayb(9)
    Dprimeb = left(Dprimefullb,8)
    Var = left(Varfull, 8)
    samplenumber = samplenumli(0)
    ' ----------------------
    ' Output to summary file
    ' ----------------------
    LDSummary.WriteLine(SNP1b & VbTab & SNP2b & VbTab & samplenumber & VbTab &
SignedDprime & VbTab & Dprimeb & VbTab & SD & VbTab & Var & VbTab & Chisq & VbTab &
degfree & VbTab & pvalue)
end sub


    ' ----------------------
    ' Main program
    ' ----------------------

infilename=inputbox("Enter filename to open","Open file...","Test.txt")
set inputfile=filesys.OpenTextFile(infilename)
inputdata = inputfile.ReadAll
inputfile.Close
inputarray = split(inputdata, VbCrLf)
header = inputarray(0)
labels = split(header, VbTab)
columns = UBound(labels)
rows = UBound(inputarray)
summaryname = "summary.xls"
set LDsummary = filesys.CreateTextFile(summaryname, True)
LDSummary.WriteLine("SNP 1" & VbTab & "SNP 2" & VbTab & "Sample no" & VbTab & "D'"
& VbTab & "D' coeff" & VbTab & "StDev" & VbTab & "Variance" & VbTab & "Chi.sq" &
VbTab & "deg.free." & VbTab & "p-value")
countera = 0
counterb = 1
counterc = countera + 1
columncount = columns + 1

    ' ----------------------
    ' Loop first SNP of pair
    ' ----------------------

do while countera < columncount

    ' ------------------------
    ' Loop second SNP of pair
    ' ------------------------

do while counterc < columncount
   SNP1 = labels(countera)
   label2 = countera + 1
   SNP2 = labels(counterc)
   LDinput = SNP1 & SNP2 & ".dat"
   set outputfile=filesys.CreateTextFile(LDinput, True)
   LDoutput = SNP1 & SNP2 & ".txt"

    ' ----------------------
    ' Loop through samples
    ' ----------------------

   do while counterb < rows
     datarow = inputarray(counterb)
     data = split(datarow, VbTab)
     data1 = data(countera)
     data2 = data(counterc)
```

```
    ' ----------------------
    ' Discard null values
    ' ----------------------

      If data1 = "0" then
        Else
        If data2 = "0" then
        Else
            a = left(data1,1)
            b = right(data1,1)
            c = left(data2,1)
            d = right(data2,1)
           outputfile.WriteLine(counterb & "    " & a & "  " & b & "    " & c & "  "
& d)
       End If
      End If
      counterb = counterb + 1
   loop
   counterb = 1
   outputfile.Close

   ' ----------------------
   ' Call sub-routines
   ' ----------------------

   Call LDexec(LDinput, LDoutput)
   Call writeresults(SNP1, SNP2, LDoutput)
   counterc = counterc + 1
loop
countera = countera + 1
counterc = countera + 1
loop

LDsummary.Close
WinShell.Popup "LDcalc Finished", 3,"Finished!",0
```

Script 2: VBScript to calculate %GC for graphing

This script calculates the %GC of a sequence for graphing. The user selects a sequence window size, and the script analyses each window for %GC content. The script uses a step size equal to the window size, although this could easily be changed. %GC is output to a results file for graph plotting.

```
'-------------------------------------------------------------------
'GC Counter
'Input file = sequence A,C,G,T,a,c,g,t with no numbers, carriage
'returns, spaces or header. Other characters ignored.
'Author: T. R. Gaunt
'-------------------------------------------------------------------

infilename=inputbox("%GC Counter" & VbCr & "~~~~~~~~~~~" & VbCr & "Enter filename
to open","Open file...","Test.txt")
outfilename=inputbox("Enter filename to save","Save file...","Results.xls")
average=inputbox("Enter the number of nucleotides over which you want to average
%GC" &VbCr & VbCr & "Note: 50 is recommended minimum, and 500 recommended
maximum","Average over...","100")
averaging=CInt(average)
```

```
set filesys=CreateObject("Scripting.FileSystemObject")
set inputfile=filesys.OpenTextFile(infilename)
set outputfile=filesys.CreateTextFile(outfilename, True)
x=inputfile.ReadAll
y=split(x, VbCrLf)
x=join(y,"")
y=len(x)
n=1
z=len(x)-averaging
do while n<z
  seq1=mid(x,n,averaging)
  seq2=lcase(seq1)
  seq1n=1
  GC=0
  GCount=0
  CCount=0
  TCount=0
  ACount=0
  do while seq1n<averaging
    base=mid(seq2,seq1n,1)
       If base = "g" Then
          GCount = GCount + 1
       ElseIf base = "c" Then
          CCount = CCount + 1
       ElseIf base = "a" Then
          ACount = ACount + 1
       ElseIf base = "t" Then
          TCount = TCount + 1
       Else bad = bad + 1
       End If
   seq1n=seq1n+1
   loop
     GC=(GCount+CCount)/(GCount+ACount+CCount+TCount)*100
    m=n-1
    outputfile.Write(m & VbTab & GC & VbCrLf)
     GC=0
     GCount=0
     CCount=0
     TCount=0
     ACount=0
  n=n+averaging
loop

outputfile.Close
inputfile.Close
msgbox("Finished!")
```

Script 3: VBScript to calculate CpG content for graphing

This script calculates the CpG content of a sequence. A step size and a window size can be defined by the user, and the script then cycles through the sequence calculating the CpG content within the sequence window at each step. The results are written to an output file for graph plotting.

```
'--------------------------------------------------------------------------
' CpG Counter
' Identifies CpG content of DNA - continuous data for graphing
' Copyright Tom Gaunt 2002
```

```
' Input file = sequence A,C,G,T,a,c,g,t with no numbers, carriage
' returns, spaces or header. Other characters ignored.
'-------------------------------------------------------------------------------
' ******************************************************************************
' This section asks for input on filenames and window/step values for calculation
' ******************************************************************************

infilename=inputbox("%GC Counter" & VbCr & "~~~~~~~~~~~" & VbCr & "Enter filename
to open","Open file...","test.txt")
outfilename=inputbox("Enter filename to save","Save file...","Results.xls")
average=inputbox("Enter the window size for CpG calculation" &VbCr & VbCr & "Note:
50 is recommended minimum, and 500 recommended maximum","Average over...","200")
step=inputbox("Enter the step size for CpG calculation" &VbCr & VbCr & "Note: half
window size is recommended maximum","Average over...","1")

' ******************************************************************************
' This section sets some variables, then reads in and formats the sequence
' ******************************************************************************

averaging=CInt(average)
steps=CInt(step)
set filesys=CreateObject("Scripting.FileSystemObject")
set inputfile=filesys.OpenTextFile(infilename)
set outputfile=filesys.CreateTextFile(outfilename, True)
x=inputfile.ReadAll
outputfile.Write("Base" & VbTab & "CpG Ratio" & VbTab & "GC content" & VbCrLf)
y=split(x, VbCrLf)
x=join(y,"")
x=lcase(x)
y=len(x)
n=1
z=len(x)-averaging

' ******************************************************************************
' This is the start of the main loop - number of cycles determined by var steps,
' window size and sequence length
' ******************************************************************************

do while n<z
  seq2=mid(x,n,averaging)
  seq1n=1
  GC=0
  GCount=0
  CCount=0
  bad=1
  CpGCount=0
  nonCpG=0
     CFreq=0
     GFreq=0
     OECpG=0
  TCount=0
  ACount=0

' ******************************************************************************
' This is the start of the CpG loop - number of cycles determined by window size.
' Steps through each dinucleotide and counts CpG, C, G, A, T
' ******************************************************************************

  do while seq1n<averaging
    base=mid(seq2,seq1n,2)
       If base = "cg" Then
          CpGCount = CpGCount + 1
       Else nonCpG = nonCpG + 1
       End If
    base1=mid(base,1,1)
    base2=mid(base,2,1)
```

```
        If base1 = "g" Then
           GCount = GCount + 1
        ElseIf base1 = "c" Then
           CCount = CCount + 1
        ElseIf base1 = "a" Then
           ACount = ACount + 1
        ElseIf base1 = "t" Then
           TCount = TCount + 1
        Else bad = 1
        End If
        If base2 = "g" Then
           GCount = GCount + 1
        ElseIf base2 = "c" Then
           CCount = CCount + 1
        ElseIf base2 = "a" Then
           ACount = ACount + 1
        ElseIf base2 = "t" Then
           TCount = TCount + 1
        Else bad = 1
        End If

   seqln=seqln+1
   loop

' ********************************************************************************
' This section calculates frequencies and CpG ratio for current window.
' ********************************************************************************

     GC=(GCount+CCount)/(ACount+CCount+GCount+TCount)
     CFreq=CCount/(ACount+CCount+GCount+TCount)
     GFreq=GCount/(ACount+CCount+GCount+TCount)
     CpG=CpGCount/(CpGCount+nonCpG)
    If (CFreq*GFreq)=0 Then
       OECpG=0
    Else OECpG=CpG/(CFreq * GFreq)
    End If
    m=n-1


' ********************************************************************************
' This section writes data out to outputfile.
' ********************************************************************************

    outputfile.Write(m & VbTab & OECpG & VbTab & GC & VbCrLf)
     GC=0
     GCount=0
     CCount=0
     CFreq=0
     GFreq=0
     OECpG=0
     CpGCount=0
     nonCpG=0
     TCount=0
     ACount=0
     m=0
     base1=""
     base2=""
     base=""
  n=n+steps
loop

outputfile.Close
inputfile.Close
msgbox("Finished!")
```

Script 4: VBScript to calculate CpG island locations

This script calculates the location of CpG islands in a sequence. A step size and a window size can be defined by the user, and the script then cycles through the sequence calculating the CpG content within the sequence window at each step. The location of CpG islands (CpG>0.6, GC>0.5 for 200 nucleotides) is written to an output file.

```
'--------------------------------------------------------------------------------
' CpG Counter
' Identifies CpG content of DNA - identifies CpG island locations
' Copyright Tom Gaunt 2002
' Input file = sequence A,C,G,T,a,c,g,t with no numbers, carriage
' returns, spaces or header. Other characters ignored.
'--------------------------------------------------------------------------------

infilename=inputbox("%GC Counter" & VbCr & "~~~~~~~~~~~" & VbCr & "Enter filename
to open","Open file...","test.txt")
outfilename=inputbox("Enter filename to save","Save file...","Results.xls")
average=inputbox("Enter the window size for CpG calculation" &VbCr & VbCr & "Note:
50 is recommended minimum, and 500 recommended maximum","Average over...","200")
step=inputbox("Enter the step size for CpG calculation" &VbCr & VbCr & "Note: half
window size is recommended maximum","Average over...","1")
averaging=CInt(average)
steps=CInt(step)
set filesys=CreateObject("Scripting.FileSystemObject")
set inputfile=filesys.OpenTextFile(infilename)
set outputfile=filesys.CreateTextFile(outfilename, True)
x=inputfile.ReadAll
outputfile.Write("CpG Islands" & VbCrLf & "~~~~~~~~~~~" & VbCrLf & VbCrLf)
y=split(x, VbCrLf)
x=join(y,"")
x=lcase(x)
y=len(x)
n=1
cpgend=-1
gap=10
cpgstart=-1
z=len(x)-averaging
do while n<z
  seq2=mid(x,n,averaging)
  seqln=1
  GC=0
  GCount=0
  CCount=0
  bad=1
  CpGCount=0
  nonCpG=0
     CFreq=0
     GFreq=0
     OECpG=0
  TCount=0
  ACount=0
  do while seqln<averaging
    base=mid(seq2,seqln,2)
       If base = "cg" Then
          CpGCount = CpGCount + 1
       Else nonCpG = nonCpG + 1
       End If
```

```
 base1=mid(base,1,1)
 base2=mid(base,2,1)
    If base1 = "g" Then
       GCount = GCount + 1
    ElseIf base1 = "c" Then
       CCount = CCount + 1
    ElseIf base1 = "a" Then
       ACount = ACount + 1
    ElseIf base1 = "t" Then
       TCount = TCount + 1
    Else bad = 1
    End If
    If base2 = "g" Then
       GCount = GCount + 1
    ElseIf base2 = "c" Then
       CCount = CCount + 1
    ElseIf base2 = "a" Then
       ACount = ACount + 1
    ElseIf base2 = "t" Then
       TCount = TCount + 1
    Else bad = 1
    End If

seq1n=seq1n+1
loop
  GC=(GCount+CCount)/(ACount+CCount+GCount+TCount)
  CFreq=CCount/(ACount+CCount+GCount+TCount)
  GFreq=GCount/(ACount+CCount+GCount+TCount)
  CpG=CpGCount/(CpGCount+nonCpG)
 If (CFreq*GFreq)=0 Then
    OECpG=0
 Else OECpG=CpG/(CFreq * GFreq)
 End If
 m=n-1
 If OECpG>0.6 AND GC>0.5 Then
   if gap>9 Then
     cpgstart=m
     cpgend=m+averaging
   else
     cpgstart=cpgstart
     cpgend=m+averaging
   end if
 Else
   If cpgend>0 AND gap>9 then
     outputfile.Write("CpG Island: " &"    " & cpgstart & ".." & cpgend &VbCrLf)
     cpgend=-1
     gap=0
   Else
     cpgend=-1
     gap=gap+1
   End If
 End If

 GC=0
 GCount=0
 CCount=0
 CFreq=0
 GFreq=0
 OECpG=0
 CpGCount=0
 nonCpG=0
 TCount=0
 ACount=0
 m=0
 base1=""
 base2=""
```

```
      base=""
   n=n+steps
loop

outputfile.Close
inputfile.Close
msgbox("Finished!")
```

Appendix C: SNPs in IGF2

SNPs are shown with 250bp flanking sequence on each side. These are all SNPs identified in the current project. Polymorphic sequence is shown in square brackets ([]) between the two segments of flanking sequence. Numbering is from a consensus sequence for the entire gene (based on GenBank AC006408) so that relative positions of SNPs can be compared easily.

Y13633-1156T/C

17364 CCACCCCCAG TGCCAGGCAG AAGCCCATCC TCACCCAGGA ACAGGGCAGC
17414 CTGTCCAACA GAAGGGTCTC GGCCTCTCCA TCAGCACCGG GAAGCCCTTT
17464 CTAGGCAAAC TTCTCACCAC TTCTTCCCTC CCTTATACTT TGAAAGAGGG
17514 AGCTCTAGGC AGGGGAGGGG CTAGAGGGGG AAGCCGCTGC CCAGATCCTG
17564 ACAAGGTGAC CTGAAGGAAC CCGGGGAGGG GGATGGGACA GGGCTCAGGC
[T/C]
17615 TGGGGTGTAT GGGGAGGGGG GCTTTGCTTT TAAAAGAGGT CATCTCAGCA
17665 ATATCTTTTT GTTTTTCCCC AGGGGCCGAA GAGTCACCAC CGAGCTTGTG
17715 TGGGAGGAGG TGGATTCCAG CCCCCAGCCC CAGGGCTCTG AATCGCTGCC
17765 AGCTCAGCCC CCTGCCCAGC CTGCCCCACA GCCTGAGCCC CAGCAGGCCA
17815 GAGAGCCCAG TCCTGAGGTG AGCTGCTGTG GCCTGTGGCC CAGGCGACCC

Y13633-2482A/C

18931 CAAGCACCCC AACCTGGACC CATATCCCCC ACGTACTTTT GGCTTTGGGC
18981 AGATTGAGCA GCCTTGGGGT GGTCTGTGCT GTCTGGTGTG GAGGGTTGCA
19031 GTTCGGGTCC TTAGTCCTAC TTCCCAGGCC GGCCGGGCTG ACGCCAGCGA
19081 gTGTGTCCTT CCCCAGCGAG GGGAGTGAGC GCAAGGTCAG CGCCTCGTCT
19131 GCGGCGCCCT GCAGGGGGTG ACGGAGGGGC GCTCTGAGGA CCCTTGGAGA
[A/C]
19182 AGGAGCTGGG TTTGTAAAAT GCTGGGCTTG GTCCCACGGA CGGCGGAGCG
19232 GTGAGCTCAG AGCCAGAGCT GGGGAGGAAA TGGGAATGAG AAAGGCCCAC
19282 TTCAGGGCTG GTGAGCGAGG GGATGGGGAG CAGCCACAGG CCGAGGCTGG
19332 GGCATGGGCC AGGCTCCATG GGGTGAGTCT GAGTCCTTGA GGGGATGTTC
19382 ATCCTCTGTG GAATGTGGGT TTGCCAGTGG AGAGGAGACC AGCGTTGCCC

Y13633-2722C/T

19171 CCCTTGGAGA AAGGAGCTGG GTTTGTAAAA TGCTGGGCTT GGTCCCACGG
19221 ACGGCGGAGC GGTGAGCTCA GAGCCAGAGC TGGGGAGGAA ATGGGAATGA
19271 GAAAGGCCCA CTTCAGGGCT GGTGAGCGAG GGGATGGGGA GCAGCCACAG
19321 GCCGAGGCTG GGGCATGGGC CAGGCTCCAT GGGGTGAGTC TGAGTCCTTG
19371 AGGGGATGTT CATCCTCTGT GGAATGTGGG TTTGCCAGTG GAGAGGAGAC
[C/T]
19422 AGCGTTGCCC TGGTGAGGTG CTGGTTCAGG GCTGGGGGGC GGACGCTGCT
19472 TGGGGCTAAA GTTCCTGCCG GCCAAGCTCT GGGTGGGAGG AGACCCTGGC
19522 CCCCTCCCAA CACCCTTGGA CTGCTGGCGG GACCCTTCCT ACCTCCGGGG
19572 GCTGGAAgTA GTGGGGGAGG AGCCAGTCTT GAGGAAGAAC CCCGATGCTG
19622 GTCTTGACTA GAGGGGAGCC GGTGTGCTTT TCGAGCCTCA GGGTGACCCG

U80851-5345polyT

23025 TATTTCCTTC AAACACACTG CACACTCCCT ATCACWAAAT CTGAAAATTC
23075 CTAAGTCCTA AGACCTAGGA ATTCTGAATC CCCTTTCTTT AAAATGTACA
23125 TATGGACCCC CAAGTCCTCC AAGGACTCTG AGCAACTTCC CTAGATCTTT
23175 ARATTCAAAA ACGATTTTCC TGGAGCCCCC AAATTGCGGT ATTGTCTCCC
23225 AKCCTTCCAA AGCAAATTGA GATTTTTTTC CCTTCACAAA ACAATTGAGG
[TTTTTTTTTTTTTT]
23289 AATACTGATT TATGAGTCTC CTGACTTTAT GGTCCCTGCC CTGGGTCCCC
23339 CTACATTTAG AAAATGTTCC ATGGACCCCC AAAGCACACT AAAAAATGTC
23389 CCTGGGTCCC AAGAAATCCC AGGCATgGAA AAACCTGCGA CCTATAAGTT
23439 TCCTAGCTAC TAACTAGGTT TCCAGAAATT TAGATATCAA ATCTCCATTG
23489 GGTAATTTCC ATGTGTCCCA AAAACTTGAA ATGTGTTTCA CTGGGGCTCC

X07868-266C/T

32314 TACCCCgTGG GCAAGTTCTT CCAATATGAC ACCTGGAAGC AGTCCACCCA
32364 GCGCCTGCGC AGGGGCCTGC CTGCCCTCCT GCGTGCCCGC CGGGGTCACG
32414 TGCTCGCCAA GGAGCTCGAG GCGTTCAGGG AGGCCAAACG TCACCGTCCC
32464 CTGATTGCTC TACCCACCCA AGACCCCGCC CACGGGGGCG CCCCCCCAGA
32514 GATGGCCAGC AATCGGAAGT GAGCAAAACT GCCGCAAGTC TGCAGCCCGG
[C/T]
32565 GCCACCATCC TGCAGCCTCC TCCTGACCAC GGACGTTTCC ATCAGGTTCC
32615 ATCCCGAAAA TCTCTCGGTT CCACGTCCCC cTGGGGCTTC TCCTGACCCA
32665 GTCCCCGTGC CCCGCCTCCC CGAAACAGGC TACTCTCCTC GGCCCCCTCC
32715 ATCGGGCTGA GGAAGCACAG CAGCATCTTC AAACATGTAC AAAATCGATT
32765 GGCTTTAAAC ACCCTTCACA TACCCTCCCC CCAAATTATC CCCAATTATC

X07868-556polyC

32604 CATCAGGTTC CATCCCGAAA ATCTCTCGGT TCCACGTCCC CcTGGGGCTT
32654 CTCCTGACCC AGTCCCCGTG CCCCGCCTCC CCGAAACAGG CTACTCTCCT
32704 CGGCCCCCTC CATCGGGCTG AGGAAGCACA GCAGCATCTT CAAACATGTA
32754 CAAAATCGAT TGGCTTTAAA CACCCTTCAC ATACCCTCCC CCCAAATTAT
32804 CCCCAATTAT CCCCACACAT AAAAAATCAA AACATTAAAC TAACCCCCTT
[CCCCCCCCCCC]
32865 ACAACAACCC TCTTAAAACT AATTGGCTTT TTAGAAACAC CCCACAAAAG
32915 CTCAGAAATT GGCTTTAAAA AAAACAACCA CCAAAAAAAA TCAATTGGCT
32965 AAAAAAAAAA AGTATTAAAA ACGAATTGGC TGAGAAACAA TTGGCAAAAT
33015 AAAGGAATTT GGCACTCCCC ACCCCCCTCT TTCTCTTCTC CCTTGGACTT
33065 TGAGTCAAAT TGGCCTGGAC TTGAGTCCCT GAACCAGCAA AGAGAAAAGA

X07868-1926C/G

33982 CAGCACACAC ATGCACACAC AGCACACACA CTCATGCGCA GCACATACAT
34032 GAACACAGCT CACAGCACAC AAACACGCAG CACACACGTT GCACACGCAA
34082 GCACCCACCT GCACACACAC ATGCGCACAC ACACGCACAC CCCCACAAAA
34132 TTRGATGAAA ACAATAAGCA TATCTAAGCA ACTACGATAT CTGTATGGAT
34182 CAGGCCAAAG TCCCGCTAAG ATTCTCCAAT GTTTTCATGG TCTGAGCCCC
[C/G]
34233 CTCCTGTTCC CATCTCCACT GCCCCTCGGC CCTGTCTGTG CCCTGCCTCT
34283 CAGAGGAGGG GGCTCAGATG GTGCGGCCTG AGTGTGCGGC CGGCGGCATT
34333 TGGGATACAC CCGTAgGGTG GGCGGGGTGT GTCCCAGGCC TAATTCCATC
34383 TTTCCACCAT GACAGAGATG CCCTTGTGAG GCTGGCCTCC TTGGCGCCTG
34433 TCCCCACGGC CCCCGCAGCG TGAGCCACGA TGCTCCCCAT ACCCCACCCA

X07868-2207C/T

34264 CTGTCTGTGC CCTGCCTCTC AGAGGAGGGG GCTCAGATGG TGCGGCCTGA
34314 GTGTGCGGCC GGCGGCATTT GGGATACACC CGTAgGGTGG GCGGGGTGTG
34364 TCCCAGGCCT AATTCCATCT TTCCACCATG ACAGAGATGC CCTTGTGAGG
34414 CTGGCCTCCT TGGCGCCTGT CCCCACGGCC CCCGCAGCGT GAGCCACGAT
34464 GCTCCCCATA CCCCACCCAT TCCCGATACA CCTTACTTAC TGTGTGTTGG
[C/T]
34515 CCAGCCAGAG TGAGGAAGGA GTTTGGCCAC ATTGGAGATG GcCGGTAGCT
34565 GAGCAGACAT GCCCCCACGA GTAGCCTGAC TCCCTGGTGT GCTCCTGGAA
34615 GGAAGATCTT GGGGACCCCC CCACCGGAGC ACACCTAGGG ATCATCTTTG
34665 CCCGTCTCCT GGGGACCCCC CAAGAAATGT GGAGTCCTCG GGGGCCGTGC
34715 ACTGATGCGG GGAGTGTGGG AAGTCTGGCG GTTGGAGGGG TGGGTGGGGG

X07868-3750exA/G

35827 CGGGGGTCGT CCATGCCAGT CCGCCTCAGT CGCAGAGGGT CCCTCGGCAA
35877 GCGCCCTGTG AGTGGGCCAT TCGGAACATT GGACAGAAGC CCAAAGAGCC
35927 AAATTGTCAC AATTGTGGAA CCCACATTGG CCTGAGATCC AAAACGCTTC
35977 GAGGCACCCC AAATTACCTG CCCATTCGTC AGGACACCCA CCCACCCAGT
36027 GTTATATTCT GCCTCGCCGG AGTGGGTGTT CCCGGGggca cttgccgacc
[A/G]
36078 gcccccttgcg tccccaggtt tgcagctctc ccctgggcca ctaaccatcc
36128 tggcccgggC TGCCTGTCTG ACCTCCGTGC CTAGTCGTGG CTCTCCATCT
36178 TGTCTCCTCC CCGTGTCCCC AATGTCTTCA GTGGGGGGCC CCCTCTTGGG
36228 TCCCCTCCTC TGCCATCACC TGAAGACCCC CACGCCAAAC ACTGAATGTC
36278 ACCTGTGCCT GCCGCCTCGG TCCACCTTGC GGCCCGTGTT TGACTCAACT

X07868-3944G/A

36080 cccttgcgtc cccaggtttg cagctctccc ctgggccact aaccatcctg
36130 gcccgggCTG CCTGTCTGAC CTCCGTGCCT AGTCGTGGCT CTCCATCTTG
36180 TCTCCTCCCC GTGTCCCCAA TGTCTTCAGT GGGGGGCCCC CTCTTGGGTC
36230 CCCTCCTCTG CCATCACCTG AAGACCCCCA CGCCAAACAC TGAATGTCAC
36280 CTGTGCCTGC CGCCTCGGTC CACCTTGCGG CCCGTGTTTG ACTCAACTCA
[G/A]
36331 CTCCTTTAAC GCTAATATTT CCGGCAAAAT CCCATGCTTG GGTTTTGTCT
36381 TTAACCTTGT AACGCTTGCA ATCCCAATAA AGCATTAAAA GTCATGATCT
36431 TCTGAGTGTT CCACTCTCTG ACTTGGGTAC TGGACTGCCG GAGGGAGGGA
36481 AGGGGCTGAG CACCTGGAAG CAGGCAGAGG GGGATAGAAG AGGGAAGGGG
36531 AAGGAAGGCC TTAGGGGTGT GGACACCTCT CTCCGTCCCC TGATCACATA

X07868-4085G/A

36220 CTCTTGGGTC CCCTCCTCTG CCATCACCTG AAGACCCCCA CGCCAAACAC
36270 TGAATGTCAC CTGTGCCTGC CGCCTCGGTC CACCTTGCGG CCCGTGTTTG
36320 ACTCAACTCA GCTCCTTTAA CGCTAATATT TCCGGCAAAA TCCCATGCTT
36370 GGGTTTTGTC TTTAACCTTG TAACGCTTGC AATCCCAATA AAGCATTAAA
36420 AGTCATGATC TTCTGAGTGT TCCACTCTCT GACTTGGGTA CTGGACTGCC
[G/A]
36471 GAGGGAGGGA AGGGGCTGAG CACCTGGAAG CAGGCAGAGG GGGATAGAAG
36521 AGGGAAGGGG AAGGAAGGCC TTAGGGGTGT GGACACCTCT CTCCGTCCCC
36571 TGATCACATA CATGGAGAAA TGAGAGAGCT GGAAGCCAGA CTCTCAGACT
36621 CACTGTCGTG CACCTGAAGC CAGGGGGTCT GGGACAGTGT CAGGCACCAA
36671 GTTCTCAAAG ATGGGGGTGC CACGAAGGGT AGGAGCCTGG GGGGCTTTTT

Appendix D: pairwise |D'| between IGF2 SNPs in NPHSII

SNP 1	SNP 2	Sample no	|D'|	Variance	χ^2	*P*-value
L15440-6815	L15440-8173	1312	0.796826	0.000883	132.07	p<0.001
L15440-6815	Y13633-1156	1396	0.478253	0.000956	120.49	p<0.001
L15440-6815	Y13633-1252	1968	0.552797	0.000629	208.94	p<0.001
L15440-6815	Y13633-2482	1855	0.482149	0.00076	125.67	p<0.001
L15440-6815	Y13633-2722	1801	0.359433	0.001816	26.89	p<0.001
L15440-6815	X07868-266	1944	0.234987	0.000977	35.89	p<0.001
L15440-6815	X07868-820	2268	0.340591	0.00034	209.95	p<0.001
L15440-6815	X07868-1926	1664	0.371756	0.000481	172.43	p<0.001
L15440-6815	X07868-2207	1886	0.134438	0.013963	0.55	Not sig
L15440-6815	X07868-3750	1888	0.17397	0.000851	20.86	p<0.001
L15440-8173	Y13633-1156	959	0.508591	0.00108	87.1	p<0.001
L15440-8173	Y13633-1252	1204	0.735425	0.000508	376.61	p<0.001
L15440-8173	Y13633-2482	1160	0.618182	0.00062	261.4	p<0.001
L15440-8173	Y13633-2722	1081	0.7121	0.001096	125.91	p<0.001
L15440-8173	X07868-266	1195	0.450841	0.005458	11.76	p<0.001
L15440-8173	X07868-820	1373	0.603833	0.001137	111.87	p<0.001
L15440-8173	X07868-1926	1042	0.590744	0.001454	86.15	p<0.001
L15440-8173	X07868-2207	1141	0.321027	0.003865	13.94	p<0.001
L15440-8173	X07868-3750	1142	0.390293	0.005126	10.36	p<0.01
Y13633-1156	Y13633-1252	1314	0.118231	0.000441	16.15	p<0.001
Y13633-1156	Y13633-2482	1356	0.130967	0.000518	15.07	p<0.001
Y13633-1156	Y13633-2722	1249	0.89051	0.000236	676.12	p<0.001
Y13633-1156	X07868-266	1315	0.496504	0.002631	41.59	p<0.001
Y13633-1156	X07868-820	1465	0.13728	0.000852	11.56	p<0.001
Y13633-1156	X07868-1926	1173	0.160077	0.000952	14.32	p<0.001
Y13633-1156	X07868-2207	1268	0.280588	0.009884	3.98	p<0.05
Y13633-1156	X07868-3750	1251	0.417802	0.00274	30.89	p<0.001
Y13633-1252	Y13633-2482	1720	0.636938	0.00028	728.58	p<0.001
Y13633-1252	Y13633-2722	1653	0.641471	0.000519	334.51	p<0.001
Y13633-1252	X07868-266	1803	0.218886	0.00235	11.39	p<0.001
Y13633-1252	X07868-820	2099	0.594832	0.000446	351.53	p<0.001
Y13633-1252	X07868-1926	1525	0.602075	0.000577	276.86	p<0.001
Y13633-1252	X07868-2207	1747	0.176898	0.005911	2.27	Not sig
Y13633-1252	X07868-3750	1742	0.184478	0.002151	7.84	p<0.01
Y13633-2482	Y13633-2722	1758	0.768681	0.000402	421.02	p<0.001
Y13633-2482	X07868-266	1822	0.235279	0.002389	11.3	p<0.001
Y13633-2482	X07868-820	1972	0.575469	0.000555	233.09	p<0.001
Y13633-2482	X07868-1926	1616	0.514302	0.000663	166.73	p<0.001
Y13633-2482	X07868-2207	1776	0.268246	0.004617	6.34	p<0.05
Y13633-2482	X07868-3750	1763	0.227923	0.002192	10.91	p<0.01

SNP 1	SNP 2	Sample no	|D'|	Variance	χ^2	*P*-value
Y13633-2722	X07868-266	1712	0.68625	0.002872	41.73	p<0.001
Y13633-2722	X07868-820	1927	0.081611	0.000298	11.62	p<0.001
Y13633-2722	X07868-1926	1486	0.073588	0.000389	7.39	p<0.01
Y13633-2722	X07868-2207	1673	0.496767	0.008674	8.88	p<0.01
Y13633-2722	X07868-3750	1648	0.617305	0.003232	35.62	p<0.001
X07868-266	X07868-820	2083	0.739579	0.00224	57.99	p<0.001
X07868-266	X07868-1926	1760	0.74667	0.002377	51.61	p<0.001
X07868-266	X07868-2207	2004	0.396504	0.028735	1.52	Not sig
X07868-266	X07868-3750	2026	0.817924	0.000441	981.07	p<0.001
X07868-820	X07868-1926	1764	0.917317	0.000112	1892.95	p<0.001
X07868-820	X07868-2207	2024	0.425296	0.008419	7.36	p<0.01
X07868-820	X07868-3750	2024	0.666916	0.002416	49.11	p<0.001
X07868-1926	X07868-2207	1775	0.300159	0.010056	2.87	Not sig
X07868-1926	X07868-3750	1727	0.819421	0.001517	65.27	p<0.001
X07868-2207	X07868-3750	1973	0.413254	0.02265	2	Not sig

Appendix E: pairwise |D'| between IGF2 SNPs in Hertfordshire cohort

SNP 1	SNP 2	Sample no	|D'|	Variance	χ^2	*P*-value
L15440-6815	L15440-8173	740	0.973708	0.000198	176.56	P<0.01
L15440-6815	Y13633-1156	671	0.57478	0.001787	85.4	P<0.01
L15440-6815	Y13633-1252	790	0.801732	0.000904	177.55	P<0.01
L15440-6815	Y13633-2482	709	0.799646	0.001017	163.16	P<0.01
L15440-6815	Y13633-2722	756	0.4731	0.003716	18.96	P<0.01
L15440-6815	X07868-266	689	0.267292	0.002841	13.76	P<0.01
L15440-6815	X07868-820	854	0.422787	0.000825	148.17	P<0.01
L15440-6815	X07868-1926	811	0.419715	0.000893	126.79	P<0.01
L15440-6815	X07868-2207	774	0.787753	0.008657	14.13	P<0.01
L15440-6815	X07868-3750	795	0.18178	0.002279	8.6	P<0.01
L15440-8173	Y13633-1156	668	0.498602	0.001569	56.29	P<0.01
L15440-8173	Y13633-1252	748	0.872959	0.000395	460.42	P<0.01
L15440-8173	Y13633-2482	713	0.85123	0.000471	432.05	P<0.01
L15440-8173	Y13633-2722	782	0.910403	0.00055	166.36	P<0.01
L15440-8173	X07868-266	673	0.516452	0.007891	11.7	P<0.01
L15440-8173	X07868-820	832	0.821957	0.000972	139.07	P<0.01
L15440-8173	X07868-1926	777	0.80168	0.001098	122.41	P<0.01
L15440-8173	X07868-2207	743	0.826778	0.002757	69.44	P<0.01
L15440-8173	X07868-3750	762	0.270367	0.00833	4.06	P<0.05
Y13633-1156	Y13633-1252	686	0.091362	0.001139	2.89	Not sig
Y13633-1156	Y13633-2482	646	0.093428	0.001243	2.66	Not sig
Y13633-1156	Y13633-2722	705	0.984214	0.000069	528.21	P<0.01
Y13633-1156	X07868-266	603	0.592909	0.005564	25.42	P<0.01
Y13633-1156	X07868-820	747	0.325764	0.001884	28.66	P<0.01
Y13633-1156	X07868-1926	727	0.352678	0.001839	33.72	P<0.01
Y13633-1156	X07868-2207	704	0.722549	0.006351	12.43	P<0.01
Y13633-1156	X07868-3750	706	0.639342	0.004382	33.7	P<0.01
Y13633-1252	Y13633-2482	727	0.974958	0.000067	1293.75	P<0.01
Y13633-1252	Y13633-2722	768	0.958787	0.000197	327.93	P<0.01
Y13633-1252	X07868-266	689	0.641661	0.004694	31.88	P<0.01
Y13633-1252	X07868-820	865	0.898194	0.000424	296.98	P<0.01
Y13633-1252	X07868-1926	824	0.89118	0.000455	288.38	P<0.01
Y13633-1252	X07868-2207	778	0.861972	0.002538	49.97	P<0.01
Y13633-1252	X07868-3750	800	0.225743	0.005877	4.21	P<0.05
Y13633-2482	Y13633-2722	739	0.957182	0.000213	316.19	P<0.01
Y13633-2482	X07868-266	630	0.569183	0.005552	23.63	P<0.01
Y13633-2482	X07868-820	799	0.905885	0.000424	292.96	P<0.01
Y13633-2482	X07868-1926	760	0.920111	0.000367	310.93	P<0.01
Y13633-2482	X07868-2207	724	0.827047	0.003304	45.75	P<0.01
Y13633-2482	X07868-3750	739	0.270378	0.006555	5.65	P<0.05
Y13633-2722	X07868-266	691	0.6056	0.009201	8.44	P<0.01
Y13633-2722	X07868-820	843	0.0891	0.000672	5.45	P<0.05

SNP 1	SNP 2	Sample no	|D'|	Variance	χ^2	*P*-value
Y13633-2722	X07868-1926	809	0.094704	0.000687	6.1	P<0.05
Y13633-2722	X07868-2207	769	0.917454	0.003366	25.01	P<0.01
Y13633-2722	X07868-3750	788	0.708945	0.005902	17.58	P<0.01
X07868-266	X07868-820	764	0.682945	0.006876	10.7	P<0.01
X07868-266	X07868-1926	744	0.74692	0.005685	11.3	P<0.01
X07868-266	X07868-2207	708	0.615555	0.041915	2.62	Not sig
X07868-266	X07868-3750	713	0.530543	0.001952	159.64	P<0.01
X07868-820	X07868-1926	898	0.9358	0.000171	1181.66	P<0.01
X07868-820	X07868-2207	849	0.864136	0.004517	16.9	P<0.01
X07868-820	X07868-3750	883	0.736948	0.00493	21.22	P<0.01
X07868-1926	X07868-2207	845	0.995566	0.000156	24.35	P<0.01
X07868-1926	X07868-3750	864	0.677337	0.005729	17.86	P<0.01
X07868-2207	X07868-3750	814	0.628859	0.034603	3.08	Not sig

Glossary

Term	Definition
Adenine	One of the four "bases" of DNA
Adipose	Of fat
Adiposity	Fatness
Admixture	Addition of a minor ingredient - in this context addition of genes by an immigrant minority
Aetiological	A causal factor
Algorithm	A set of rules for problem-solving - e.g. computer programming
Alignment	Comparison of two or more sequences for the maximum match
Aliquot	A fraction of a whole; aliquoting: the act of dispensing multiple fractions of an original
Alleles	One of two or more alternative sequence variants
Allele-specific PCR	A polymerase chain reaction in which a primer matches only one allele
Amino acid	A component of a protein
Amplicons	A sequence amplified by polymerase chain reaction
Amplification	Multiple replication of a target sequence by polymerase chain reaction
Amplification refractory mutation system	A system using allele-specific PCR to distinguish between individuals of different genotypes
Annealing	Association of complementary nucleic acid to form a duplex
Antisense	Reverse-complement of a sequence (reverse order with G&C swapped and A&T swapped)
Autoradiography	Detection of radio-label using sensitive film
Autosome	Non-sex chromosome
Bonferroni correction	A correction for multiple testing: p-value is multiplied by the number of tests performed
Buccal	Of the mouth
Capillary	Ultra-fine tubing through which surface tension rather than gravity draws liquid
Centrifuge	Rotation of a sample-tube at high-speed to separate materials on the basis of density
Chromatograms	Line graphs (usually multi-coloured) of fluorescence against time
Codon	A nucleotide triplet that specifies an amino acid
Coefficient of Variation	A statistical test of variation between tests
Cohort	A group of people with a common statistical characteristic
Concatenated	Joined or linked together
Contig	A list of overlapping cloned sequence fragments that collectively contain the sequence of a continuous DNA strand
Cytosine	One of the four "bases" of DNA
Deadenylation	Removal of polyadenylated tail from mRNA
Deletion	Loss of a segment of sequence from within a continuous sequence
Denaturation	Dissociation of complementary strands of a nucleic acid
Diploid	Having two copies of each autosome
Dominant	A trait that is expressed in a heterozygote
Downstream	In the 3' direction: left to right when the sequence is written in the conventional way
Duplex	Pair of associated complementary strands
Electrophoresis	Separation of molecules on the basis of size and charge through a polymer by use of electrical current
Elute	Dissociation of a molecule from a fixed column matrix into solution
Endonuclease	An enzyme cleaving within a nucleic acid at a specific sequence
Enhancers	Elements stimulating transcription of a gene that may not be in the immediate vicinity of that gene
Epidemiology	Study of incidence and distribution of diseases

Term	Definition
Exon	A segment of gene that is represented in the mature RNA product. May contain coding or non-coding sequence
Exonuclease	An enzyme that cleaves nucleic acid sequence in a non-sequence specific manner
Fluorescence	Emission of light at a specific wavelength requiring an input of energy at a specific wavelength
Genbank	NCBI database containing the majority of publically available sequence data
Gene	A collection of sequence elements which together result in the production of an RNA transcript and usually a protein
Genome	The total genetic complement of an organism
Genotype	The genetic constitution of an individual - usually referring to a specific locus and therefore comprising two alleles
Genotyping	The process of identifying the genotype of an individual
Guanine	One of the four "bases" of DNA
Hairpin loops	Partially self-complementary sequences which form a loop
Haplotype	A series of alleles found at different loci on the same parentally derived chromosome
Haplotyping	The process of identifying the haplotype of an individual. Sometimes achieved by estimation
Heteroduplex	A nucleic acid duplex comprising two non-complementary sequences (may represent sense and anti-sense strands from two different alleles
Heterogeneity	Non-uniform
Heterozygous	An individual with two alleles at a particular locus.
Homoduplexes	A nucleic acid duplex comprising two complementary sequences
Homopolymeric tract	A sequence run of multiple identical nucleotides
Homozygous	An individual with two copies of the same allele at a particular locus
Hyperglycaemia	An excess of glucose in the blood stream
Hypertension	High blood pressure
Hypophagia	Suppressed food intake
Hypothalamus	A region of the brain controlling temperature, appetite etc.
Imprinting	Determination of gene expression by parent of origin
Incomplete penetrance	A situation in which a genotype does not always cause a phenotype
Inheritance	The transmission of genetic material (and consequently genetically determined characteristics) from parent to child
Insertion	Addition of sequence within a continuous sequence interrupting the original sequence
Intergenic	Between genes
Intra-uterine	Within the uterus - during gestation
Intron	A segment of gene that is not represented in the mature RNA product. Separates exons
Kilobase	One thousand bases of nucleic acid
Linkage	The tendency of genes to be co-inherited as a consequence of physical proximity
Linkage disequilibrium	A situation where two alleles co-occur more or less frequently than would be expected from their allele frequencies
Locus	A unique chromosomal location
Megabase	One million bases of nucleic acid
Meiosis	Cell division resulting in two daughter cells with half the numer of chromosomes of the parent
Mendelian	A pedigree pattern resulting from determination at a single chromosome location: Dominant or recessive, autosomal or sex-linked
Metabolic	Chemical processes resulting in energy production
Methylation	Cytosines in CpG dinucleotides are liable to be methylated - in promoters this can result in lack of expression
Microarray	Very high density arrays of oligonucleotides to allow genotyping or expression analysis

Term	Definition
Microplate	Industry-standard sample-handling vessel comprising 96, 384 or 1536 wells.
Microsatellites	Small run of tandem repeats of a very simple (1-4bp) DNA sequence
Minisatellites	Tandemly repeated DNA of longer sequences than a microsatellite
Mis-priming	Non-sequence specific polymerase chain reaction resulting from poor primer design or optimisation
Mitogenic	Processes resulting in mitotic cell division
Mitosis	Cell division resulting in two daughter cells with the same chromosome number as the parent cell
Mole	SI unit of an amount of substance: 1 mole = molecular weight in grammes
Monogenic	Caused by a single gene
Non-insulin-dependent diabetes mellitus	Type II diabetes - characterised by insulin resistance
Non-parametric	Statistical methods not based upon stringent assumptions - more appropriate for much biological data
Nucleotides	A purine or pyrimidine base linked to a sugar and a phosphate group. The building block of nucleic acids
Obese	Very overweight. Usually defined as body mass index > 30
Oligonucleotide	A sequence of several nucleotides. Synthetic oligonucleotides are used to prime (initiate) polymerase chain reaction
Pedigree	A family. Graphical representation of a family and their disease and genetic statistics
Pellet	Material collecting at the bottom of a tube during centrifugation: the most dense material
Phenocopy	The same phenotype caused independently by more than one genetic factor
Phenotype	The observable characteristics of an organism
Pixel	Picture element in digital imaging
Polyadenylated	Addition of multiple (~200) adenine residues to the 3' end of a mRNA - stabilises RNA
Polygenic	Caused by multiple genes
Polymerase	An enzyme involved in the formation of a polymer - required in polymerase chain reaction
Polymorphic	Existing in more than one form
Polymorphism	In genetics a locus at which more than one allele exists with a minimum allele frequency of 1%
Polypeptide	A chain of amino acids constituting either a protein or a subunit of a protein
Precipitate	Remove from solution
Primer	Oligonucleotide used to initiate polymerase chain reaction
Programming	The permanent modification of an individuals metabolism by intra-uterine environment resulting in effects in later life
Promoter	A combination of sequence elements to which an RNA polymerase binds in order to initiate gene transcription
Quantitative trait	A trait which is a continuous variable (such as weight or height)
Recessive	A character manifest only in the homozygote
Recombination	The exchange of genetic material between homologous chromosomes during meiosis
RNase	An enzyme degrading RNA
Sequencing	The determination of the sequence of nucleotides for a particular fragment of nucleic acid
Silanised	A glass plate treated with a chemical to ensure that a gel matrix adheres to it
Single-nucleotide polymorphisms	Polymorphisms involving a change from one nucleotide to another
Stratification	A population that is genetically non-homogeneous as a result of insufficient mixing of sub-groups
Supernatant	Liquid at the top of a sample tube after centrifugation
Synonymous substitution	A single-base change that alters a codon sequence but not the resulting amino acid

Term	Definition
Template	A sequence which acts as a blueprint for polymerase chain reaction
Tetraprimer	A type of allele-specific PCR involving three possible fragments primed by four oligonucleotides
Thymine	One of the four "bases" of DNA
Titration	Ascertainment of the amount of a substance required by testing of dilutions
Transcript	RNA transcribed from a gene
Transcription	Production of RNA from a DNA sequence - requires RNA polymerase interaction with promoter
Transition	A type of SNP involving purine to purine or pyrimidine to pyrimidine conversion
Translation	Production of protein from an RNA transcript
Transversion	A type of SNP involving purine to pyrimidine or pyrimidine to purine conversion
Untranslated region	Regions present in the RNA transcript that are not translated to protein
Upstream	In the 5' direction: right to left when the sequence is written in the conventional way
VBA macros	A system of scripted automation within certain software packages
Whole genome amplification	Amplification of genomic DNA with degenerate primers that amplify many random sequences in the genome

References

Aihie-Sayer A, Syddal H, O'Dell SD, Day INM & Cooper C 2002 Polymorphism of the *IGF2* gene, birth weight and adult grip strength. *Age and Ageing* (in press).

Antonarakis SE 1998 Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Human Mutation* **11** 1-3.

Argente J, Caballo N, Barrios V, Pozo J, Munoz MT, Chowen JA & Hernandez M 1997 Multiple endocrine abnormalities of the growth hormone and insulin-like growth factor axis in prepubertal children with exogeneous obesity: effect of short- and long-term weight reduction. *Journal of Clinical Endocrinology and Metabolism* **82** 2076-2083.

Bach LA, Hsieh S, Sakano K, Fujiwara H, Perdue JF & Rechler MM 1993 Binding of mutants of human insulin-like growth factor II to insulin-like growth factor binding proteins 1-6. *Journal of Biological Chemistry* **268** 9246-9254.

Barker DJ 1995 Fetal origins of coronary heart disease. *British Medical Journal* **311** 171-174.

Barnes WM 1994 PCR amplification of up to 35-kb DNA with high-fidelity and high yield from -bacteriophage templates. *Proceedings of the National Academy of Sciences* **91** 2216-2220.

Bertrais S, Balkau B, Vol S, Forhan A, Calvet C, Marre M, Eschwege E & Study ESIR 1999 Relationships between abdominal body fat distribution and cardiovascular risk factors: an explanation for women's healthier cardiovascular risk profile. The D.E.S.I.R Study. *International Journal of Obesity* **23** 1085-1094.

Binder R, Hwang SP, Ratnasabapathy R & Williams DL 1989 Degradation of apolipoprotein II mRNA occurs via endonucleolytic cleavage at 5'-AAU-3'/5'-UAA-3' elements in single-stranded loop domains of the 3'-noncoding region. *Journal of Biological Chemistry* **264** 16910-16918.

Björntorp P 1996 Diabetes. In *Ciba Foundation Symposium 201 (1996): The Origins and Consequences of Obesity*, Eds DJ Chadwick & G Cardew. Chichester: John Wiley and Sons.

Blahovec J, Kostecka Z, Lacroix MC, Cabanie L, Godeau F, Mester J & Cavaille F 2001 Mitogenic activity of high molecular weight forms of insulin-like growth factor-II in amniotic fluid. *Journal of Endocrinology* **169** 563-572.

Bonfield JK, Rada C & Staden R 1998 Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Research* **26** 3404-3409.

Boss O, Samec S, Paoloni-Giacobino A, Rossier C, Dulloo A, Seydoux J, Muzzin P & Giacobino JP 1997 Uncoupling protein-3: a new member of the mitochondrial carrier family with tissue-specific expression. *FEBS Letters* **408** 39-42.

Boswell T, Nicholson MA & Bunger L 1999 Neuropeptide Y gene expression in lines of mice subjected to long-term divergent selection on fat content. *Journal of Molecular Endocrinology* **23** 77-83.

Bouchard C 1996 Genetics of obesity in humans: current issues. In *Ciba Foundation Symposium 201 (1996): The Origins and Consequences of Obesity*, Eds DJ Chadwick & G Cardew. Chichester: John Wiley and Sons.

Bray MS, Boerwinkle E & Hanis CL 2000 Sequence variation within the neuropeptide Y gene and obesity in Mexican Americans. *Obesity Research* **8** 219-226.

Brissenden JE, Ullrich A & Francke U 1984 Human chromosomal mapping of genes for insulin-like growth factors I and II and epidermal growth factor. *Nature* **310** 781-784.

Brookes AJ 1999 The essence of SNPs. *Gene* **234** 177-186.

Bustin SA 2000 Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology* **25** 169-193.

Campfield LA, Smith FJ, Guisez Y, Devos R & Burn P 1995 Recombinant mouse OB protein: evidence for a peripheral signal linking adiposity and central neural networks. *Science* **269** 546-549.

Cardon LR & Bell JI 2001 Association study designs for complex diseases. *Nature Reviews.Genetics* 2 91-99.

Cargill M, Atshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ & Lander ES 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* 22 231-238.

Caro JF, Kolaczynski JW, Nyce MR, Ohannesian JP, Opentanova I, Goldman WH, Lynn RB, Zhang PL, Sinha MK & Considine RV 1996 Decreased cerebrospinal-fluid/serum leptin ratio in obesity: a possible mechanism for leptin resistance. *Lancet* 348 159-161.

Casella SJ, Han VK, D'Ercole AJ, Svoboda ME & Van Wyk JJ 1986 Insulin-like growth factor II binding to the type I somatomedin receptor. Evidence for two high affinity binding sites. *Journal of Biological Chemistry* 261 9268-9273.

Chen H, Charlat O, Tartaglia LA, Woolf EA, Weng X, Ellis SJ, Lakey ND, Culpepper J, Moore KJ, Breitbart RE, Duyk GM, Tepper RI & Morgenstern JP 1996 Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in db/db mice. *Cell* 84 491-495.

Cheng S, Chang SY, Gravitt P & Respess R 1994 Long PCR. *Nature* 369 684-685.

Chernokalskaya E, Dompenciel R & Schoenberg DR 1997 Cleavage properties of an estrogen-regulated polysomal ribonuclease involved in the destabilization of albumin mRNA. *Nucleic Acids Research* 25 735-742.

Cheung VG & Nelson SF 1996 Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proceedings of the National Academy of Sciences* 93 14676-14679.

Choy YS, Dabora SL, Hall F, Ramesh V, Niida Y, Franz D, Kasprzyk-Obara J, Reeve MP & Kwiatkowski DJ 1999 Superiority of denaturing high performance liquid chromatography over single-stranded conformation and conformation-sensitive gel electrophoresis for mutation detection in TSC2. *Annals of Human Genetics* 63 (Pt 5) 383-391.

Christiansen J, Kofod M & Nielsen FC 1994 A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA. *Nucleic Acids Research* 22 5709-5716.

Clement K, Vaisse C, Lahlou N, Cabrol S, Pelloux V, Cassuto D, Gourmelen M, Dina C, Chambaz J, Lacorte JM, Basdevant A, Bougneres P, Lebouc Y, Froguel P & Guy-Grand B 1998 A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* 392 398-401.

Cocchi D, De GC, V, Bagnasco M, Bonacci D & Muller EE 1999 Leptin regulates GH secretion in the rat by acting on GHRH and somatostatinergic functions. *Journal of Endocrinology* 162 95-99.

Collins A, Lonjou C & Morton NE 1999 Genetic epidemiology of single-nucleotide polymorphisms. *Proceedings of the National Academy of Sciences* 96 15173-15177.

Collins FS, Brooks LD & Chakravati A 1998 A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Research* 1229-1231.

Cotton RG 1998 Mutation detection and mutation databases. *Clinical Chemistry and Laboratory Medicine* 36 519-522.

Cotton RG, Rodrigues NR & Campbell RD 1988 Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. *Proceedings of the National Academy of Sciences* 85 4397-4401.

Cruickshank JK, Heald AH, Anderson S, Cade JE, Sampayo J, Riste LK, Greenhalgh A, Taylor W, Fraser W, White A & Gibson JM 2001 Epidemiology of the insulin-like growth factor system in three ethnic groups. *American Journal of Epidemiology* 154 504-513.

Cummings DE & Schwartz MW 2000 Melanocortins and body weight: a tale of two receptors. *Nature Genetics* 26 8-9.

Day, I. N. M. Electrophoresis gel-matrix layer. UNIV LONDON. Patent GB19930024901 19931203 [GB2284484]. 1995. GB.

Day, I. N. M. Gel-matrix electrophoresis. University College London. Patent 849697 [*6,071,396*]. 2000. London,GB . 1997.

Day, I. N. M. and Hamid, R. GEL ELECTROPHORESIS . Day, I. N. M. and Univ Southampton. Patent WO2000GB01739 20000505 [WO0068677]. 2000. GB.

Day INM & Humphries SE 1994 Electrophoresis for genotyping: microplate array diagonal gel electrophoresis on horizontal polyacrylamide gels, Hydrolink or agarose. *Analytical Biochemistry* **222** 391-394.

de Haas M, Koene HR, Kleijer M, de Vries E, Simsek S, van Tol MJ, Roos D & dem Borne AE 1996 A triallelic Fc gamma receptor type IIIA polymorphism influences the binding of human IgG by NK cell Fc gamma RIIIa. *Journal of Immunology* **156** 3948-3955.

de Silva AM, Walder KR, Boyko EJ, Whitecross KF, Nicholson G, Kotowicz M, Pasco J & Collier GR 2001 Genetic variation and obesity in Australian women: a prospective study. *Obesity Research* **9** 733-740.

den Dunnen JT & Antonarakis SE 2000 Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Human Mutation* **15** 7-12.

den Dunnen JT & Antonarakis SE 2001 Nomenclature for the description of human sequence variations. *Human Genetics* **109** 121-124.

Devlin B & Risch N 1995 A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics* **29** 311-322.

Dobson-Stone C, Cox RD, Lonie L, Southam L, Fraser M, Wise C, Bernier F, Hodgson S, Porter DE, Simpson AH & Monaco AP 2000 Comparison of fluorescent single-strand conformation polymorphism analysis and denaturing high-performance liquid chromatography for detection of EXT1 and EXT2 mutations in hereditary multiple exostoses. *European Journal of Human Genetics* **8** 24-32.

Dunger DB, Ong KK, Huxtable SJ, Sherriff A, Woods KA, Ahmed ML, Golding J, Pembrey ME, Ring S, Bennett ST & Todd JA 1998 Association of the INS VNTR with size at birth.

ALSPAC Study Team. Avon Longitudinal Study of Pregnancy and Childhood. *Nature Genetics* 19 98-100.

Echwald SM 1999 Genetics of human obesity: lessons from mouse models and candidate genes. *Journal of Internal Medicine* 245 653-666.

Ekstrom TJ, Cui H, Li X & Ohlsson R 1995 Promoter-specific IGF2 imprinting status and its plasticity during human liver development. *Development* 121 309-316.

Eng C, Brody LC, Wagner TM, Devilee P, Vijg J, Szabo C, Tavtigian SV, Nathanson KL, Ostrander E & Frank TS 2001 Interpreting epidemiological research: blinded comparison of methods used to estimate the prevalence of inherited mutations in BRCA1. *Journal of Medical Genetics* 38 824-833.

Eriksson JG, Forsen T, Tuomilehto J, Jaddoe VW, Osmond C & Barker DJ 2002 Effects of size at birth and childhood growth on the insulin resistance syndrome in elderly individuals. *Diabetologia* 45 342-348.

Eriksson JG, Forsen T, Tuomilehto J, Osmond C & Barker DJ 2001 Early growth and coronary heart disease in later life: longitudinal study. *British Medical Journal* 322 949-953.

Eyre-Walker A & Keightley PD 1999 High genomic deleterious mutation rates in hominids. *Nature* 397 344-347.

Fall CH, Osmond C, Barker DJ, Clark PM, Hales CN, Stirling Y & Meade TW 1995 Fetal and infant growth and cardiovascular risk factors in women. *British Medical Journal* 310 428-432.

Farooqi IS, Keogh JM, Kamath S, Jones S, Gibson WT, Trussell R, Jebb SA, Lip GY & O'Rahilly S 2001 Partial leptin deficiency and human adiposity. *Nature* 414 34-35.

Fleury C, Neverova M, Collins S, Raimbault S, Champigny O, Levi-Meyrueis C, Bouillaud F, Seldin MF, Surwit RS, Ricquier D & Warden CH 1997 Uncoupling protein-2: a novel gene linked to obesity and hyperinsulinemia. *Nature Genetics* 15 269-272.

Forbes BE, Hartfield PJ, McNeil KA, Surinya KH, Milner SJ, Cosgrove LJ & Wallace JC 2002 Characteristics of binding of insulin-like growth factor (IGF)-I and IGF-II analogues to

the type 1 IGF receptor determined by BIAcore analysis. *European Journal of Biochemistry* 269 961-968.

Forsberg L, de Faire U & Morgenstern R 1998 Identification of genetic polymorphisms in the 'expressed sequence tag' (EST) database. *Technical Tips Online* 1 T01440.

Frayling TM & Hattersley AT 2001 The role of genetic susceptibility in the association of low birth weight with type 2 diabetes. *British Medical Bulletin* 60 89-101.

Frevel MAE, Hornberg JJ & Reeve AE 1999 A potential imprint control element: identification of a conserved 42bp sequence upstream of H19. *Trends in Genetics* 15 216-218.

Frystyk J, Vestbo E, Skjærbæk C, Mogensen CE & Orskov H 1995 The Insulin-Like Growth Factors in Human Obesity. *Metabolism* 44 37-44.

Gardiner-Garden M & Frommer M 1987 CpG islands in vertebrate genomes. *Journal of Molecular Biology* 196 261-282.

Gaunt TR, Cooper JA, Miller GJ, Day IN & O'Dell SD 2001 Positive associations between single nucleotide polymorphisms in the IGF2 gene region and body mass index in adult males. *Human Molecular Genetics* 10 1491-1501.

Gaunt TR, Hinks LJ, Chen Xh, O'Dell SD, Spanakis E, Ganderton RH & Day INM 2000 384-well MADGE for high-throughput DNA bank studies. *Technical Tips Online* 1 02069.

Ghilardi N, Ziegler S, Wiestner A, Stoffel R, Heim MH & Skoda RC 1996 Defective STAT signaling by the leptin receptor in diabetic mice. *Proceedings of the National Academy of Sciences* 93 6231-6235.

Giannoukakis N, Deal C, Paquette J, Kukuvitis A & Polychronakos C 1996 Polymorhic functional imprinting of the human *IGF2* gene among individuals, in blood cells, is associated with *H19* expression. *Biochemical and Biophysical Research Communications* 220 1014-1019.

Godfrey KM & Barker DJ 2001 Fetal programming and adult health. *Public Health Nutrition* 4 611-624.

Gong DW, He Y, Karas M & Reitman M 1997 Uncoupling protein-3 is a mediator of thermogenesis regulated by thyroid hormone, beta3-adrenergic agonists, and leptin. *Journal of Biological Chemistry* 272 24129-24132.

Graack HR & Kress H 1999 Detection of frame-shifts within homopolymeric DNA tracts using the amplification refractory mutation system (ARMS). *Biotechniques* 27 662-4, 666.

Gu D, O'Dell SD, Chen XH, Miller GJ & Day IN 2002 Evidence of multiple causal sites affecting weight in the IGF2-INS-TH region of human chromosome 11. *Human Genetics* 110 173-181.

Gu DF, Hinks LJ, Morton NE & Day IN 2000 The use of long PCR to confirm three common alleles at the CYP2A6 locus and the relationship between genotype and smoking habit. *Annals of Human Genetics* 64 383-390.

Guler HP, Schmid C, Zapf J & Froesch ER 1989 Effects of recombinant insulin-like growth factor I on insulin secretion and renal function in normal human subjects. *Proceedings of the National Academy of Sciences* 86 2868-2872.

Haig D & Graham C 1991 Genomic imprinting and the strange case of the insulin-like growth factor II receptor. *Cell* 64 1045-1046.

Haig D & Westoby M 1989 Parent-specific gene expression and the triploid endosperm. *The American Naturalist* 134 147-155.

Hales CN & Barker DJ 1992 Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia* 35 595-601.

Hales CN & Barker DJ 2001 The thrifty phenotype hypothesis. *British Medical Bulletin* 60 5-20.

Hales CN, Barker DJ, Clark PM, Cox LJ, Fall C, Osmond C & Winter PD 1991 Fetal and infant growth and impaired glucose tolerance at age 64. *British Medical Journal* 303 1019-1022.

Hattersley AT, Beards F, Ballantyne E, Appleton M, Harvey R & Ellard S 1998 Mutations in the glucokinase gene of the fetus result in reduced birth weight. *Nature Genetics* 19 268-270.

Hattersley AT & Tooke JE 1999 The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. *Lancet* 353 1789-1792.

Haubrich DR, Wang PF & Wedeking PW 1975 Distribution and metabolism of intravenously administered choline[methyl- 3-H] and synthesis in vivo of acetylcholine in various tissues of guinea pigs. *The Journal of Pharmacology and Experimental Therapeutics* 193 246-255.

Hausman GJ, Campion DR & Buonomo FC 1991 Concentration of insulin-like growth factors (IGF-I and IGF-II) in tissues of lean and obese pig fetuses. *Growth, Development & Aging* 55 43-52.

Heid CA, Stevens J, Livak KJ & Williams PM 1996 Real time quantitative PCR. *Genome Research* 6 986-994.

Heil J, Glanowski S, Scott J, Winn-Deen E, McMullen I, Wu L, Gire C & Sprague A 2002 An automated computer system to support ultra high throughput SNP genotyping. *Pacific Symposium on Biocomputing* 30-40.

Henke W, Herdel K, Jung K, Schnorr D & Loening SA 1997 Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Research* 25 3957-3958.

Hinks LJ, Price SE, Mason CR & Thompson RJ 1995 Single strand conformation analysis of two genes contained within the first intron of the neurofibromatosis type I gene in patients with multiple sclerosis. *Neuropathology and Applied Neurobiology* 21 201-207.

Hyun SW, Kim SJ, Park K, Rho HM & Lee YI 1993 Characterization of the P4 promoter region of the human insulin-like growth factor II gene. *FEBS Letters* 332 153-158.

Igel M, Taylor BA, Phillips SJ, Becker W, Herberg L & Joost HG 1998 Hyperleptinemia and leptin receptor variant Asp600Asn in the obese, hyperinsulinemic KK mouse strain. *Journal of Molecular Endocrinology* 21 337-345.

Ikejiri K, Wasada T, Haruki K, Hizuka N, Hirata Y & Yamamoto M 1991 Identification of a novel transcription unit in the human insulin-like growth factor-II gene. *Biochemical Journal* 280 439-444.

Ioannidis P, Havredaki M, Courtis N & Trangas T 1996 In vivo generation of 3' and 5' truncated species in the process of c-myc mRNA decay. *Nucleic Acids Research* **24** 4969-4977.

Jackson AA, Langley-Evans SC & McCarthy HD 1996 Nutritional influences in early life upon obesity and body proportions. In *Ciba Foundation Symposium 201 (1996): The Origins and Consequences of Obesity*, Eds DJ Chadwick & G Cardew. Chichester: John Wiley and Sons.

Jacobs S, Kull FC, Jr., Earp HS, Svoboda ME, Van Wyk JJ & Cuatrecasas P 1983 Somatomedin-C stimulates the phosphorylation of the beta-subunit of its own receptor. *Journal of Biological Chemistry* **258** 9581-9584.

James WPT 1996 The epidemiology of obesity. In *Ciba Foundation Symposium 201 (1996): The Origins and Consequences of Obesity*, Eds DJ Chadwick & G Cardew. Chichester: John Wiley and Sons.

Jebb SA 1997 Aetiology of obesity. *British Medical Bulletin* **53**.

Jeffcoate W 1998 Obesity is a disease: food for thought. *Lancet* **351** 903-904.

Jones AC, Austin J, Hansen N, Hoogendoorn B, Oefner PJ, Cheadle JP & O'Donovan MC 1999 Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis. *Clinical Chemistry* **45** 1133-1140.

Jones BK, Levorse J & Tilghman SM 2001 Deletion of a nuclease-sensitive region between the Igf2 and H19 genes leads to Igf2 misregulation and increased adiposity. *Human Molecular Genetics* **10** 807-814.

Kim HS, Nagalla SR, Oh Y, Wilson E, Roberts CT, Jr. & Rosenfeld RG 1997 Identification of a family of low-affinity insulin-like growth factor binding proteins (IGFBPs): characterization of connective tissue growth factor as a member of the IGFBP superfamily. *Proceedings of the National Academy of Sciences* **94** 12981-12986.

Kim-Motoyama H, Yasuda K, Yamaguchi T, Yamada N, Katakura T, Shuldiner AR, Akanuma Y, Ohashi Y, Yazaki Y & Kadowaki T 1997 A mutation of the beta 3-adrenergic receptor is associated with visceral obesity but decreased serum triglyceride. *Diabetologia* **40** 469-472.

Kopelman P 1999 Aetiology of Obesity II: Genetics. In *Obesity: The Report of the British Nutrition Foundation Task Force (1999)*, Blackwell Science Ltd.

Kruglyak L 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22** 139-144.

Lakka HM, Oksanen L, Tuomainen TP, Kontula K & Salonen JT 2000 The common pentanucleotide polymorphism of the 3'-untranslated region of the leptin receptor gene is associated with serum insulin levels and the risk of type 2 diabetes in non-diabetic men: a prospective case-control study. *Journal of Internal Medicine* **248** 77-83.

Lander ES 1996 The new genomics: global views of biology. *Science* **274** 536-539.

Lander ES & Schork NJ 1994 Genetic Dissection of Complex Traits. *Science* **265** 2037-2045.

LaPaglia N, Steiner J, Kirsteins L, Emanuele M & Emanuele N 1998 Leptin alters the response of the growth hormone releasing factor- growth hormone--insulin-like growth factor-I axis to fasting. *Journal of Endocrinology* **159** 79-83.

Law CM, Barker DJ, Osmond C, Fall CH & Simmonds SJ 1992 Early growth and abdominal fatness in adult life. *Journal of Epidemiology and Community Health* **46** 184-186.

Le Stunff C, Fallin D, Schork NJ & Bougneres P 2000 The insulin gene VNTR is associated with fasting insulin levels and development of juvenile obesity. *Nature Genetics* **26** 444-446.

Lee LG, Connell CR & Bloch W 1993 Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Research* **21** 3761-3766.

Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A & Tilghman SM 1995 Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* **375** 34-39.

Lewontin RC 1964 The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49 49-67.

Li E, Beard C & Jaenisch R 1993 Role for DNA methylation in genomic imprinting. *Nature* 366 362-365.

Li X, Cui H, Sandstedt B, Nordlinder H, Larsson E & Ekstrom TJ 1996 Expression levels of the insulin-like growth factor-II gene (IGF2) in the human liver: developmental relationships of the four promoters. *Journal of Endocrinology* 149 117-124.

Li YM, Franklin G, Cui HM, Svensson K, He XB, Adam G, Ohlsson R & Pfeifer S 1998 The H19 transcript is associated with polysomes and may regulate IGF2 expression in trans. *Journal of Biological Chemistry* 273 28247-28252.

Linnell J, Groeger G & Hassan AB 2001 Real time kinetics of insulin-like growth factor II (IGF-II) interaction with the IGF-II/mannose 6-phosphate receptor: the effects of domain 13 and pH. *Journal of Biological Chemistry* 276 23986-23991.

Little S 1997 ARMS analysis of point mutations. In *Laboratory methods for the detection of mutations and polymorphisms in DNA*, pp 45-51. Ed GR Taylor. Boca Raton, FLA.: CRC Press.

Louvi A, Accili D & Efstratiadis A 1997 Growth-promoting interaction of IGF-II with the insulin receptor during mouse embryonic development. *Developmental Biology* 189 33-48.

Lowe J 1996 Insulin-like Growth Factors. *Scientific American* 62-71.

Lucas A 1991 Programming by early nutrition in man. *Ciba Foundation Symposium* 156 38-50.

Lucassen AM, Julier C, Beressi JP, Boitard C, Froguel P, Lathrop M & Bell JI 1993 Susceptibility to insulin-dependent diabetes mellitus maps to a 4.1kb segment of DNA spanning the insulin gene and associated VNTR. *Nature Genetics* 4 305-310.

Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X & Morton NE 2002 The first linkage disequilibrium (LD) maps: delineation of hot and cold

blocks by diplotype analysis. *Proceedings of the National Academy of Sciences* 99 2228-2233.

Mårin P, Kvist H, Lindstedt G, Sjöström L & Björntorp P 1993 Low concentrations of insulin-like growth factor-I in abdominal obesity. *International Journal of Obesity* 17 83-89.

Marsh DJ, Hollopeter G, Huszar D, Laufer R, Yagaloff KA, Fisher SL, Burn P & Palmiter RD 1999a Response of melanocortin-4 receptor-deficient mice to anorectic and orexigenic peptides. *Nature Genetics* 21 119-122.

Marsh DJ, Miura GI, Yagaloff KA, Schwartz MW, Barsh GS & Palmiter RD 1999b Effects of neuropeptide Y deficiency on hypothalamic agouti-related protein expression and responsiveness to melanocortin analogues. *Brain Research* 848 66-77.

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY & Gish WR 1999 A general approach to single-nucleotide polymorphism discovery. *Nature Genetics* 23 452-456.

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schemechel DE, Purvis I, Pericak-Vance MA, Roses AD & Vance JM 2000 SNPing Away at Complex Diseases: Analysis of Single-Nucleotide Polymorphisms around ApoE in Alzheimer Disease. *American Journal of Human Genetics* 67 383-394.

Mein CA, Barratt BJ, Dunn MG, Siegmund T, Smith AN, Esposito L, Nutland S, Stevens HE, Wilson AJ, Phillips MS, Jarvis N, Law S, de Arruda M & Todd JA 2000 Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Research* 10 330-343.

Meinsma D, Scheper W, Holthuizen PE, Van den Brande JL & Sussenbach JS 1992 Site-specific cleavage of IGF-II mRNAs requires sequence elements from two distinct regions of the IGF-II gene. *Nucleic Acids Research* 20 5003-5009.

Meister B 2000 Control of food intake via leptin receptors in the hypothalamus. *Vitamins and Hormones* 59 265-304.

Mercer JG, Moar KM, Rayner DV, Trayhurn P & Hoggard N 1997 Regulation of leptin receptor and NPY gene expression in hypothalamus of leptin-treated obese (ob/ob) and cold-exposed lean mice. *FEBS Letters* **402** 185-188.

Mineo R, Fichera E, Liang SJ & Fujita-Yamaguchi Y 2000 Promoter usage for insulin-like growth factor-II in cancerous and benign human breast, prostate, and bladder tissues, and confirmation of a 10th exon. *Biochemical and Biophysical Research Communications* **268** 886-892.

Miner JL, Della-Fera MA, Paterson JA & Baile CA 1989 Lateral cerebroventricular injection of neuropeptide Y stimulates feeding in sheep. *American Journal of Physiology* **257** R383-R387.

Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, Sewter CP, Digby JE, Mohammed SN, Hurst JA, Cheetham CH, Earley AR, Barnett AH, Prins JB & O'Rahilly S 1997 Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387** 903-908.

Mori Y, Kim-Motoyama H, Ito Y, Katakura T, Yasuda K, Ishiyama-Shigemoto S, Yamada K, Akanuma Y, Ohashi Y, Kimura S, Yazaki Y & Kadowaki T 1999 The Gln27Glu beta2-adrenergic receptor variant is associated with obesity due to subcutaneous fat accumulation in Japanese men. *Biochemical and Biophysical Research Communications* **258** 138-140.

Morricone L, Ferrari M, Emrini R, Inglese L, Garancini P & Caviezel F 1999 The role of central fat distribution in coronary artery disease in obesity: comparison of nondiabetic obese, diabetic obese, and normal weight subjects. *International Journal of Obesity* **23** 1129-1135.

Morrione A, Valentinis B, Xu SQ, Yumet G, Louvi A, Efstratiadis A & Baserga R 1997 Insulin-like growth factor II stimulates cell proliferation through the insulin receptor. *Proceedings of the National Academy of Sciences* **94** 3777-3782.

Morton NE 1955 Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7** 277-318.

Morton NE 1998 Significance levels in complex inheritance. *American Journal of Human Genetics* 62 690-697.

Morton NE, Shields DC & Collins A 1991 Genetic epidemiology of complex phenotypes. *Annals of Human Genetics* 55 (Pt 4) 301-314.

Myers RM, Larin Z & Maniatis T 1985a Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes. *Science* 230 1242-1246.

Myers RM, Lumelsky N, Lerman LS & Maniatis T 1985b Detection of single base substitutions in total genomic DNA. *Nature* 313 495-498.

Nadler ST, Stoehr JP, Schueler KL, Tanimoto G, Yandell BS & Attie AD 2000 The expression of adipogenic genes is decreased in obesity and diabetes mellitus. *Proceedings of the National Academy of Sciences* 97 11371-11376.

Nelson TL, Vogler GP, Pederson NL & Miles TP 1999 Genetic and environmental influences on waist-to-hip ratio and waist circumference in an older Swedish twin population. *International Journal of Obesity* 23 449-455.

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N & Smith JC 1989 Analysis of any point mutation in DNA: the amplification refractory mutation system (ARMS). *Nucleic Acids Research* 17 2503-2516.

Nielsen FC 1992 The molecular and cellular biology of insulin-like growth factor II. *Progress in Growth Factor Research* 4 257-290.

O'Dell SD, Miller GJ, Cooper JA, Hindmarsh PC, Pringle PJ, Ford H, Humphries SE & Day INM 1997 ApaI polymorphism in insulin-like growth factor II (IGF2) gene and weight in middle-aged males. *International Journal of Obesity* 21 822-825.

O'Dell SD, Bujac SR, Miller GJ & Day INM 1999 Association of IGF2 RFLP and INS VNTR class I allele size with obesity. *European Journal of Human Genetics* 7 821-827.

O'Dell SD, Gaunt TR & Day INM 2000 SNP genotyping by combination of 192-well MADGE, ARMS and computerized gel image analysis. *Biotechniques* 29 500-506.

O'Donovan MC, Oefner PJ, Roberts SC, Austin J, Hoogendoorn B, Guy C, Speight G, Upadhyaya M, Sommer SS & McGuffin P 1998 Blind Analysis of Denaturing High-Performance Liquid Chromatography as a Tool for Mutation Detection. *Genomics* 52 44-49.

Oefner PJ & Underhill PA 1995 Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC). *American Journal of Human Genetics* 57 A266.

Ohlsson R, Hedborg F, Holmgren L, Walsh C & Ekstrom TJ 1994 Overlapping patterns of IGF2 and H19 expression during human development: biallelic IGF2 expression correlates with a lack of H19 expression. *Development* 120 361-368.

Ong K, Kratzsch J, Kiess W, Costello M, Scott C & Dunger D 2000 Size at birth and cord blood levels of insulin, insulin-like growth factor I (IGF-I), IGF-II, IGF-binding protein-1 (IGFBP-1), IGFBP-3, and the soluble IGF-II/mannose-6-phosphate receptor in term human infants. The ALSPAC Study Team. Avon Longitudinal Study of Pregnancy and Childhood. *Journal of Clinical Endocrinology and Metabolism* 85 4266-4269.

Ong KL, Phillips DI, Fall C, Pouton J, Bennett ST, Golding J, Todd JA & Dunger DB 1999 The insulin gene VNTR, type 2 diabetes and birth weight. *Nature Genetics* 21 262-263.

Onyango P, Miller W, Lehoczky J, Leung CT, Birren B, Wheelan S, Dewar K & Feinberg AP 2000 Sequence and Comparative Analysis of the Mouse 1-Megabase Region Orthologous to the Human 11p15 Imprinted Domain. *Genome Research* 10 1697-1710.

Orita M, Suzuki Y, Sekiyu T & Hayashi K 1989a Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* 5 874-879.

Orita M, Iwahana H, Kanazawa H, Hayashi K & Seika T 1989b Detection of polymorphisms in human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences* 86 2766-2770.

Owerbach D & Gabbay KH 1993 Localization of a type I diabetes susceptibility locus to the variable tandem repeat region flanking the insulin gene. *Diabetes* 42 1708-1714.

Pagter-Holthuizen P, Jansen M, van der Kammen RA, van Schaik FM & Sussenbach JS 1988 Differential expression of the human insulin-like growth factor II gene. Characterization of the IGF-II mRNAs and an mRNA encoding a putative IGF-II-associated protein. *Biochimica et Biophysica Acta* **950** 282-295.

Phillips DI, Handelsman DJ, Eriksson JG, Forsen T, Osmond C & Barker DJ 2001 Prenatal growth and subsequent marital status: longitudinal study. *British Medical Journal* **322** 771.

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA & Boyce-Jacino M 1999 Mining SNPs from EST databases. *Genome Research* **9** 167-174.

Pritchard JK 2001 Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* **69** 124-137.

Qu D, Ludwig DS, Gammeltoft S, Piper M, Pelleymounter MA, Cullen MJ, Mathes WF, Przypek R, Kanarek R & Maratos-Flier E 1996 A role for melanin-concentrating hormone in the central regulation of feeding behaviour. *Nature* **380** 243-247.

Radetti G, Bozzola M, Pasquino B, Paganini C, Aglialoro A, Livieri C & Barreca A 1998 Growth hormone bioactivity, insulin-like growth factors (IGFs) and IGF binding proteins in obese children. *Metabolism* **47** 1490-1493.

Rainier S, Johnson LA, Dobry GJ, Ping AJ, Grundy PE & Feinberg AP 1993 Relaxation of imprinted genes in human cancer. *Nature* **362** 747-749.

Ravelli AC, Der Meulen JH, Osmond C, Barker DJ & Bleker OP 1999 Obesity at the age of 50 y in men and women exposed to famine prenatally. *The American Journal of Clinical Nutrition* **70** 811-816.

Ravelli GP, Stein ZA & Susser MW 1976 Obesity in young men after famine exposure in utero and early infancy. *New England Journal of Medicine* **295** 349-353.

Reed MR, Huang CF, Riggs AD & Mann JR 2001 A complex duplication created by gene targeting at the imprinted H19 locus results in two classes of methylation and correlated Igf2 expression phenotypes. *Genomics* **74** 186-196.

Reich DE & Lander ES 2001 On the allelic spectrum of human disease. *Trends in Genetics* 17 502-510.

Rinderknecht E & Humbel RE 1978 Primary structure of human insulin-like growth factor II. *FEBS Letters* 89 283-286.

Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J, Kalaydjieva L, McCague P, Dimiceli S, Pitts T, Nguyen L, Yang J, Harper C, Thorpe D, Vermeer S, Young H, Hebert J, Lin A, Ferguson J, Chiotti C, Wiese-Slater S, Rogers T, Salmon B, Nicholas P & Myers RM 1999 A genomic screen of autism: evidence for a multilocus etiology. *American Journal of Human Genetics* 65 493-507.

Rodenburg RJT, Krijger JJ, Holthuizen PE & Sussenbach JS 1996 The liver-specific promoter of the human insulin-like growth factor-II gene contains two negative regulatory elements. *FEBS Letters* 394 25-30.

Roth SM, Schrager MA, Metter.E.J., Riechmann SE, Fleg JL, Hurley BF & Ferrell RE 2002 *IGF2* genotype and obesity in men and women across the adult age span. *International Journal of Obesity* 26 585-587.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning ZM, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann NS, Zody MC, Linton L, Lander ES & Altshuler D 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409 928-933.

Sahu A 1998 Evidence suggesting that galanin (GAL), melanin-concentrating hormone (MCH), neurotensin (NT), proopiomelanocortin (POMC) and neuropeptide Y (NPY) are targets of leptin signaling in the hypothalamus. *Endocrinology* 139 795-798.

Sakane N, Yoshida T, Umekawa T, Kogure A & Kondo M 1999 Beta2-adrenoceptor gene polymorphism and obesity. *Lancet* 353 1976.

Sakano K, Enjoh T, Numata F, Fujiwara H, Marumoto Y, Higashihashi N, Sato Y, Perdue JF & Fujita-Yamaguchi Y 1991 The design, expression, and characterization of human insulin-like growth factor II (IGF-II) mutants specific for either the IGF-II/cation-independent mannose 6-phosphate receptor or IGF-I receptor. *Journal of Biological Chemistry* **266** 20626-20635.

Sakatani T, Wei M, Katoh M, Okita C, Wada D, Mitsuya K, Meguro M, Ikeguchi M, Ito H, Tycko B & Oshimura M 2001 Epigenetic heterogeneity at imprinted loci in normal populations. *Biochemical and Biophysical Research Communications* **283** 1124-1130.

Sakurai T, Amemiya A, Ishii M, Matsuzaki I, Chemelli RM, Tanaka H, Williams SC, Richardson JA, Kozlowski GP, Wilson S, Arch JR, Buckingham RE, Haynes AC, Carr SA, Annan RS, McNulty DE, Liu WS, Terrett JA, Elshourbagy NA, Bergsma DJ & Yanagisawa M 1998 Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92** 573-585.

Scarpace PJ, Nicolson M & Matheny M 1998 UCP2, UCP3 and leptin gene expression: modulation by food restriction and leptin. *Journal of Endocrinology* **159** 349-357.

Schaid DJ 1998 Transmission disequilibrium, family controls and great expectations. *American Journal of Human Genetics* **63** 935-941.

Schalling M, Johansen J, Nordfors L & Lonnqvist F 1999 Genes involved in animal models of obesity and anorexia. *Journal of Internal Medicine* **245** 613-619.

Scheper W, Holthuizen PE & Sussenbach JS 1996a Growth-condition-dependent regulation of insulin-like growth factor II mRNA stability. *The Biochemical Journal* **318 (Pt 1)** 195-201.

Scheper W, Holthuizen PE & Sussenbach JS 1996b The cis-acting elements involved in endonucleolytic cleavage of the 3'UTR of human IGF-II mRNAs bind a 50kDa protein. *Nucleic Acids Research* **24** 1000-1007.

Scheper W, Meinsma D, Holthuizen PE & Sussenbach JS 1995 Long-Range RNA Interaction of Two Sequence Elements Required for Endonucleolytic Cleavage of Human Insulin-Like Growth Factor II mRNAs. *Molecular and Cellular Biology* **15** 235-245.

Schork NJ, Nath SK, Fallin D & Chakravati A 2000 Linkage Disequilibrium Analysis of Biallelic DNA Markers, Human Quantitative Trait Loci, and Threshold-Defined Case and Control Subjects. *American Journal of Human Genetics* 67 1208-1218.

Schrauwen P, Xia J, Walder K, Snitker S & Ravussin E 1999 A novel polymorphism in the proximal UCP3 promoter region: effect on skeletal muscle UCP3 mRNA expression and obesity in male non-diabetic Pima Indians. *International Journal of Obesity* 23 1242-1245.

Shaper AG 1996 Obesity and Cardiovascular Disease. In *Ciba Foundation Symposium 201 (1996): The Origins and Consequences of Obesity*, Eds DJ Chadwick & G Cardew. Chichester: John Wiley and Sons.

Smith MC, Cook JA, Furman TC & Occolowitz JL 1989 Structure and activity dependence of recombinant human insulin-like growth factor II on disulfide bond pairing. *Journal of Biological Chemistry* 264 9314-9321.

Spielman RS & Ewens WJ 1998 A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* 62 450-458.

Spielman RS, McGinnis RE & Ewens WJ 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 52 506-516.

Stephens TW, Basinski M, Bristow PK, Bue-Valleskey JM, Burgett SG, Craft L, Hale J, Hoffmann J, Hsiung HM, Kriauciunas A & . 1995 The role of neuropeptide Y in the antiobesity action of the obese gene product. *Nature* 377 530-532.

Stoeckle MY 1992 Removal of a 3' non-coding sequence is an initial step in degradation of gro alpha mRNA and is regulated by interleukin-1. *Nucleic Acids Research* 20 1123-1127.

Strachan T & Read AP 1999 *Human Molecular Genetics*. BIOS Scientific Publishers Ltd.

Sun F, Flanders WD, Yang Q & Khoury MJ 1999 Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *American Journal of Epidemiology* 150 97-104.

Sun, G., Chagnon, Y. C., Gagnon, J., Perusse, L., Rao, D. C., Rivera, M. A., Sjöström, L., and Bouchard, C. *ApaI* polymorphism in IGF-II gene associated with higher BMI and fat mass (FM) in women from the Swedish Obese Subjects (SOS) and the Quebec Family Study Cohorts. Genetic and Molecular Basis of Obesity, Satellite Symposium of the 8th International Congress on Obesity , 18. 1998.

Tadokoro K, Fujii H, Inoue T & Yamada M 1991 Polymerase chain reaction (PCR) for detection of Apa1 polymorphism at the insulin-like growth factor II gene (IGF2). *Nucleic Acids Research* **19** 6967.

Taillon-Miller P, Bauer-Sardiña I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP & Kwok PY 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics* **25** 324-328.

Terasawa H, Kohda D, Hatanaka H, Nagata K, Higashihashi N, Fujiwara H, Sakano K & Inagaki F 1994 Solution structure of human insulin-like growth factor II; recognition sites for receptors and binding proteins. *The EMBO Journal* **13** 5590-5597.

Terwilliger JD & Weiss KM 1998 Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **9** 578-594.

Tishkoff SA, Pakstis AJ, Ruano G & Kidd KK 2000 The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *American Journal of Human Genetics* **67** 518-522.

Torres AM, Forbes BE, Aplin SE, Wallace JC, Francis GL & Norton RS 1995 Solution structure of human insulin-like growth factor II. Relationship to receptor and binding protein interactions. *Journal of Molecular Biology* **248** 385-401.

Tyagi S, Bratu DP & Kramer FR 1998 Multicolor molecular beacons for allele discrimination. *Nature Biotechnology* **16** 49-53.

Ukkola O, Sun G & Bouchard C 2001 Insulin-like growth factor 2 (IGF2) and IGF-binding protein 1 (IGFBP1) gene variants are associated with overfeeding-induced metabolic changes. *Diabetologia* **44** 2231-2236.

Valenzano KJ, Heath-Monnig E, Tollefsen SE, Lake M & Lobel P 1997 Biophysical and biological properties of naturally occurring high molecular weight insulin-like growth factor II variants. *Journal of Biological Chemistry* 272 4804-4813.

Valenzano KJ, Remmler J & Lobel P 1995 Soluble insulin-like growth factor II/mannose 6-phosphate receptor carries multiple high molecular weight forms of insulin-like growth factor II in fetal bovine serum. *Journal of Biological Chemistry* 270 16441-16448.

van Dijk EL, Sussenbach JS & Holthuizen PE 2001 Kinetics and regulation of site-specific endonucleolytic cleavage of human IGF-II mRNAs. *Nucleic Acids Research* 29 3477-3486.

Velho G, Hattersley AT & Froguel P 2000 Maternal diabetes alters birth weight in glucokinase-deficient (MODY2) kindred but has no influence on adult weight, height, insulin secretion or insulin sensitivity. *Diabetologia* 43 1060-1063.

Vella V, Pandini G, Sciacca L, Mineo R, Vigneri R, Pezzino V & Belfiore A 2002 A novel autocrine loop involving IGF-II and the insulin receptor isoform-A stimulates growth of thyroid cancer. *Journal of Clinical Endocrinology and Metabolism* 87 245-254.

von Horn H, Ekstrom C, Ellis E, Olivecrona H, Einarsson C, Tally M & Ekstrom TJ 2002 GH is a regulator of IGF2 promoter-specific transcription in human liver. *Journal of Endocrinology* 172 457-465.

Vu TH & Hoffman A 1996 Alterations in the promoter-specific imprinting of the insulin-like growth factor-II gene in Wilms' tumor. *Journal of Biological Chemistry* 271 9014-9023.

Vu TH & Hoffman AR 1994 Promoter-specific imprinting of the human insulin-like growth factor-II gene. *Nature* 371 714-717.

Walder K, Norman RA, Hanson RL, Schrauwen P, Neverova M, Jenkinson CP, Easlick J, Warden CH, Pecqueur C, Raimbault S, Ricquier D, Silver MH, Shuldiner AR, Solanes G, Lowell BB, Chung WK, Leibel RL, Pratley R & Ravussin E 1998 Association between uncoupling protein polymorphisms (UCP2-UCP3) and energy metabolism/obesity in Pima indians. *Human Molecular Genetics* 7 1431-1435.

Wang AM, Doyle MV & Mark DF 1989 Quantitation of mRNA by the polymerase chain reaction. *Proceedings of the National Academy of Sciences* **86** 9717-9721.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES & . 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280** 1077-1082.

Waterland RA & Garza C 1999 Potential mechanisms of metabolic imprinting that lead to chronic disease. *The American Journal of Clinical Nutrition* **69** 179-197.

Waterland RA & Garza C 2000 Reply to A lucas. *The American Journal of Clinical Nutrition* **71** 602-603.

Weiss KM & Clark AG 2002 Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* **18** 19-24.

White MB, Carvalho M, Derse D, O'Brien SJ & Dean M 1992 Detecting single base substitutions as heteroduplex polymorphisms. *Genomics* **12** 301-306.

Wong GK, Passey DA, Huang Y, Yang Z & Yu J 2000 Is "junk" DNA mostly intron DNA? *Genome Research* **10** 1672-1678.

Wu HK, Squire JA, Catzavelos CG & Weksberg R 1997a Relaxation of imprinting of human insulin-like growth factor II gene, IGF2, in sporadic breast carcinomas. *Biochemical and Biophysical Research Communications* **235** 123-129.

Wu HK, Squire JA, Song Q & Weksberg R 1997b Promoter-dependent tissue-specific expressive nature of imprinting gene, insulin-like growth factor II, in human tissues. *Biochemical and Biophysical Research Communications* **233** 221-226.

Xie X & Ott J 1993 Testing linkage disequilibrium between a disease gene and marker loci. *American Journal of Human Genetics* **53** 1107.

Ye S, Dhillon S, Ke X, Collins AR & Day IN 2001 An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Research* **29** E88.

Ye S, Humphires SE & Green F 1992 Allele-specific amplification by tetra-primer PCR. *Nucleic Acids Research* **20** 1152.

Zaina S & Squire S 1998 The Soluble Type 2 Insulin-like Growth Factor (IGF-II) Receptor Reduces Organ Size by IGF-II-mediated and IGF-II-independent Mechanisms. *The Journal of Biological Chemistry* **273** 28610-28616.

Zapata C, Carollo C & Rodriguez S 2001 Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. *Annals of Human Genetics* **65** 395-406.

Zhang J, Day IN & Byrne CD 2002 A novel medium throughput quantitative competitive PCR technology to simultaneously measure mRNA levels from multiple genes. *Nucleic Acids Research* **30** e20.

Zhang J, Desai M, Ozanne SE, Doherty C, Hales CN & Byrne CD 1997 Two variants of quantitative reverse transcriptase PCR used to show differential expression of alpha-, beta- and gamma-fibrinogen genes in rat liver lobes. *The Biochemical Journal* **321 (Pt 3)** 769-775.

Zhang J & Byrne C 1997 A Novel Highly Reproducible Quantitative Competitive RT PCR System. *Journal of Molecular Biology* **274** 338-352.

Zhang Y, Proenca R, Maffei M, Barone M, Leopold L & Friedman JM 1994 Positional cloning of the mouse obese gene and its human homologue. *Nature* **372** 425-432.