Genetic Analyses of Quantitative Traits in Human Twins

Beben Benyamin

Thesis submitted for the degree
of
Doctor of Philosophy

University of Edinburgh

2006

Acknowledgments

This PhD thesis is the results of three-years work conducted at the Institute of Evolutionary Biology, the University of Edinburgh, Scotland (October 2003 - February 2005) and the Genetic Epidemiology Group, the Queensland Institute of Medical Research (QIMR), Australia (March 2005 - October 2006). My PhD was funded by a generous scholarship from the School of Biological Sciences, the University of Edinburgh and an Overseas Research Scholarship from the Department of Education and Skills, UK. I am very thankful to the two institutions that make my PhD studies possible.

It is not an exaggeration that this thesis will not exist without help, work and support of many wonderful people. Now, with my pleasure, it is time for me to thank them all.

I have been very fortunate to have Professor Peter M. Visscher as my principal supervisor. Peter has been very supportive throughout my PhD. Not only he is very supportive with my research, but also with my (family) well-being. Peter is always happy to spend his time to answer all my questions and to derive and explain mathematical formulas patiently. Even before I started my PhD, Peter has helped me in obtaining the scholarship from the School of Biological Sciences. His constant motivation from the very beginning of my PhD made me realize that three years are indeed very short time. *Be precise!* is his favourite words that mark most of my drafts. This has forced me to write in more logical, understandable and scientifically sound sentences. I would like also to thank Peter and Naomi for always inviting us to their house, both in Edinburgh and Brisbane for delicious meals and BBQs. *Terima kasih, Peter!*

My sincere gratitude goes to Professor William (Bill) G. Hill (University of Edinburgh) as my second supervisor. Bill's sharp and critical comments have shaped my thesis. I was so impressed by how quickly Bill's response to comment on all my drafts. I would like also to thank Professor Nicholas G. Martin for the hospitality provided in his Genetic Epidemiology Lab at the QIMR. Nick has also kindly allowed me to use his large dataset on adult height and provided invaluable comments on Chapter 6.

All work presented in this thesis, except those on computer simulations, were based on data provided by collaborators. I am very grateful for their confidence in allowing me to analyse their data. For this, I would like to thank Professor Ian J. Deary (School of Psychology, University of Edinburgh) for providing unique twin data on intelligence and anthropometric measures from the Scottish Mental Surveys of 1932 and 1947 (Chapter 2). Ian has also provided me with invaluable comments to the draft of paper (Benyamin *et al.*, 2005), which became Chapter 2. I would like also to thank Professor Robert Plomin (Institute of Psychiatry (IOP), London) for kindly allowing me to use the U.K. Twins' Early

Developments Study (TEDS) data and for helpful comments on the draft of paper (Benyamin *et al.*, 2006), which became Chapter 3. Thanks also to Ms Nicole Harlaar (also from IOP) for providing information related to the TEDS data. I am indebted to Professor Kirsten Ohm Kyvik and Dr Karoline Schousboe (both from the Danish Twin Registry, University of Southern Denmark, Odense) for allowing me to use the GEMINAKAR data on the metabolic syndrome and also providing comments to Chapter 5. Expert comments and suggestions from Professor Thorkild Soerensen and Dr Mogens Fenger (both from Copenhagen University Hospitals, Denmark) are also very helpful in shaping Chapter 5.

I am indebted to Professor Michael C. Neale of Virginia Commonwealth University, USA. Mike helped me with the mixture distribution model, which is implemented in his statistical package Mx. In particular, Mike has spent time to run additional simulation that was used for part of the discussion in Chapter 2. I would like also to thank Mr Ian M. White (University of Edinburgh) for helping me with the use of a mixed linear model and ASReml statistical package (Chapter 5). Many thanks also go to Dr Arthur Gilmour (Orange Agricultural Institute, Australia) for providing me with a beta version of ASReml for calculating the standard errors of (co)variance components of multivariate analysis with more than 10 traits.

My special thanks go to Ms Carole Ferrier (secretary) and Ms Jayne Glendinning (finance officer), both at the Institute of Evolutionary Biology, University of Edinburgh. By helping me with administrative- and welfare-related issues, Carole and Jayne have both made my PhD studies easier. A special thank also for Ms Coral Pink (Secretary at the Genetic Epidemiology, QIMR). Coral has been very helpful in finding an accommodation in Brisbane and made my move to Brisbane easier. In addition, Coral has also been very helpful with the day to day administrative-related issues at the QIMR. Many thanks also to the Genetic Epidemiology IT staffs: David Smyth, Scott Gordon and Harry Beeby. They are very helpful with most of my computer-related problems. In particular, Scott has spent a great deal of his time to help me with the genome scan data.

During my studies, I am very lucky to have many wonderful friends that always available to help and cheer me. Among many that I cannot mention individually, they are: Bernardo, Dan Gaffney, Dan Halligan, Dario and Pau (in Edinburgh), Allan, Belinda, Brendan, Carl, Enda, Gu, Jonathan, Kate, Lannie, Manuel, Mark, Michelle, Narelle, Penelope, Sarah, Sri, Stuart and Will (in Brisbane). Their senses of humour, which exploded during lunch and afternoon tea time have made my studies enjoyable. Thanks a lot guys! Playing tennis with Brendan is *a stressful* moment. I mean, I am still unable to claim for 100-bucks prize money for beating him! A special thank to Allan who patiently answered my questions on genetics/statistics and helped me with computer simulations. In particular, Allan kindly allowed me to use his C++ library to support my simulations and gone through most of the drafts of my thesis.

My sincere gratitude goes to my teachers, Dr Muladno (Institut Pertanian Bogor (IPB), Indonesia) and Professor Chris Moran (University of Sydney, Australia) for teaching me genetics while I was studying at IPB and the University of Sydney. Both Muladno and Chris were also very supportive academic supervisors.

In Brisbane, I develop a friendship with Indonesian friends from various Indonesian communities. I am particularly very grateful to join the Tennis-Dong Club, where Pak Hendry Baiquni and *the Dongers* taught me how to play tennis. *Salam Donger!*

During my PhD, my family in Tasikmalaya, Mojokerto and Bandung have been very supportive. *Hatur nuhun* to my parents, *Ema* and *Apa* in Tasikmalaya.

Finally, this thesis is dedicated to my wife, Ike Herwidi and our daughter, Tiara Kinanti. Ike, thanks for your love, understanding and constant supports. *Tiara anakku, kehadiranmu adalah kebahagiaan. Semoga suatu hari nanti, thesis ini bisa menjadi sumber inspirasi dan motivasi di dalam hidupmu!*

Brisbane, 21 November 2006

Publications

The following publications are a direct outcome of the research presented in the thesis:

Benyamin, B., V. Wilson, L. J. Whalley, P. M. Visscher and I. J. Deary. 2005. Large, consistent estimates of the heritability of cognitive ability in two entire populations of 11-year-olds twins from Scottish Mental Surveys of 1932 and 1947. *Behavior Genetics* 35: 525 - 534 (Chapter 2).

Benyamin, B., I. J. Deary and P. M. Visscher. 2006. Precision and bias of a normal finite mixture distribution model to analyze twin data when zygosity is unknown: simulations and application to IQ phenotypes on a large sample of twin pairs. *Behavior Genetics* 36: 935 - 946 (Chapter 3).

Benyamin, B., T. I. A. Sørensen, K. Schousboe, M. Fenger, P. M. Visscher and K. O. Kyvik. Are there common genetic and environmental factors behind the endophenotypes associated with the metabolic syndrome? *Diabetologia (In Press)* (Chapter 5).

Benyamin, B., M. Perola, B. K. Cornes, P. A. F. Madden, A. Palotie, G. W. Montgomery, L. Peltonen, N. G. Martin and P. M. Visscher. Within-family outliers: segregating alleles or environmental effects? A linkage analysis of height from 5,815 sibling pairs. *Submitted* (Chapter 6).

The following publications are associated with the research presented in the thesis:

Visscher, P. M., **B. Benyamin** and I. White. 2004. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin Research* 7: 670 - 674.

Fenger, M., K. Schousboe, T. I. A. Sørensen and K. O. Kyvik. Variance decomposition of apolipoproteins and lipids in Danish twins. *Atherosclerosis.* Published Online 24 May 2006[1].

[1] The published paper mistakenly left out the name of B. Benyamin as the second author. A corrigendum has been approved by the editor and will be printed in the next available issue of the journal.

Abstract

One of the major goals in human genetics is to identify gene(s) underlying variation in quantitative traits. Determining whether the phenotype is heritable is necessary before embarking on gene identification. In humans, populations of twin pairs provide an elegant natural experiment to partition phenotypic variance of a trait into genetic and environmental factors. By comparing the resemblance of genetically identical (monozygotic) twin pairs to that of non-identical (dizygotic) twin pairs, the (classical) twin design has been widely used to estimate the proportion of phenotypic variance due genetic (heritability) and environmental factors. In addition, collections of dizygotic twins have been suggested as an important design for performing genetic linkage and association studies. Dizygotic twins are sibling pairs controlled for age and shared environmental factors.

The aim of the research presented in this thesis is to understand the genetic basis of the variation of human quantitative traits using data from twins and (to some extent) their families. Traits investigated include behavioural traits (intelligence), clinical traits (the metabolic syndrome) and anthropometric measures (height). The focus of the thesis is mainly on heritability estimation using both conventional and novel statistical methods. The identification of gene(s) underlying complex phenotypes by means of linkage analysis is also presented.

The importance of human twins for understanding genetic variation in human quantitative traits is reviewed. This includes the use of twins for estimating the heritability and identifying gene(s) underlying complex quantitative traits using genetic linkage and association studies (Chapter 1). The use of a novel finite mixture distribution model to partition phenotypic co(variance) of a trait into underlying genetic and environmental factors from twins of unknown zygosity is presented (Chapter 2-4). The Scottish Mental Surveys of 1st June 1932 and 4th June 1947, respectively, administered the same validated verbal reasoning test (the Moray House Test) to almost everyone born in 1921 or 1936 and attending school in Scotland. Information on zygosity was unavailable. A novel application of a finite mixture distribution model estimated a large and consistent heritability of cognitive ability of about ~0.7. This study is the first to estimate genetic and environmental components of cognitive ability in entire school-attending populations and implies that large (national) data collections can provide sufficient information on twin pairs to estimate genetic parameters, even without known zygosity (Chapter 2). The precision and bias of the finite mixture distribution model were assessed using computer simulations and application to IQ measures from a large sample of known zygosity twins (twins from the U.K. Twins' Early Developments Study). It is shown that the mixture distribution is unbiased provided that the twins' trait values are (bivariate) normally distributed and the sample size is large.

However, if the bivariate normality assumption is violated, then the mixture distribution provides biased estimates (Chapter 3). The extension of the model to multivariate analysis is discussed. The simulations show that multivariate analysis decreases the standard error of the variance component estimates (Chapter 4). Another statistical model, a mixed linear model is used to partition the phenotypic (co)variances of traits into genetic and environmental factors from twins of known zygosity (twins from the Danish Twin Registry). Its application to understand the underlying genetic and environmental aetiology of endophenotypes associated with the metabolic syndrome (the cluster of obesity, insulin resistance, dyslipidaemia and hypertension) showed that endophenotypes associated with the metabolic syndrome do not appear to share a substantial common genetic or familial environmental background (Chapter 5). Finally, a genome-wide linkage analysis to identify gene/chromosomal regions associated with adult height reveals several chromosomal regions that showed a modest linkage to adult height. This analysis provides further evidence for the polygenic nature of body height (Chapter 6).

In conclusion, populations of twins are elegant natural experiments that provide means for understanding the genetic architecture of human quantitative traits. A finite mixture distribution model has been shown to be reliable and very useful in decomposing phenotypic (co)variance of quantitative traits collected from twins of unknown zygosity into genetic and environmental components. Twins have also been shown to be useful for understanding genetic and environmental aetiology of multiple phenotypes syndrome, and for identifying gene/chromosomal regions underlying variation of human quantitative traits (Chapter 7).

Table of Contents

1 Introduction

1.1 Variation in Human Phenotypes

The ultimate aim of research in human genetics is to understand the genetic basis of human phenotypes. That is, to identify and characterize gene(s) underlying variation in human phenotypes. In genetics, human phenotypes can be categorized into those which show simple (Mendelian) patterns of inheritance and those which show complex (non-Mendelian) patterns of inheritance (e.g. Botstein and Risch, 2003; Risch, 2000). In simple Mendelian inheritance the phenotype can be explained by single genes, which are sufficient (and necessary) to determine the phenotypes. This includes phenotypes such as albinism, colour blindness, Huntington's disease, phenylketonuria (PKU) and sickle cell anemia. On the other hand, many other phenotypes that are evolutionary and medically important show a complex pattern of inheritance, in which many genes and environmental factors influence the phenotype. In other words, the relationship between genotype and phenotype is not simple (Lander and Schork, 1994). These phenotypes can form either a discrete (e.g. the absence or presence of breast cancer) or continuous/quantitative distribution (e.g. anthropometric measures, such height and weight). In the literature, the complex phenotypes are sometimes referred to as complex traits, quantitative traits or multifactorial traits. Throughout the thesis, "phenotype" will be used interchangeably with "trait".

In the last two decades, huge efforts have been invested into identification and characterization of gene(s) influencing both Mendelian and complex traits. While success stories have commonly been reported for Mendelian traits, for which about 1,700 genes have been identified (Antonarakis and

Beckmann, 2006; O'Connor and Crystal, 2006; Online Mendelian Inheritance in Man, 2006), the identification of genes underlying complex traits has been slow and difficult (Altmuller *et al.*, 2001). Various methods have been explored and applied to identify gene(s) underlying complex traits. Linkage analysis and association studies are the two most commonly used methods for gene identification. While linkage analysis follows the (co)segregation of markers and traits in a pedigreed population, association studies correlate genotypes of an individual with its phenotype in the population. In addition, with the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), the availability of single nucleotide polymorphisms (SNPs) in the public databases (The International SNP Map Working Group, 2001) and the completion of the first stage of the International HapMap Project (The International HapMap Consortium, 2005), there has been a huge interest in genome-wide association analysis, where hundreds of thousands of SNPs covering the genome are tested for association with complex traits (Hirschhorn and Daly, 2005).

Before embarking on gene identification, it is important to have some knowledge whether the trait is heritable (Martin *et al.*, 1997). For quantitative traits, the most important parameter to serve this purpose is the heritability, which quantifies the proportion of phenotypic variance of a trait that is due to genetic factors. Although heritability for many different traits have been reported, this is not a fixed quantity. It is a function of allele frequency and specific to a population at any one time. Any changes in allele frequency (e.g. due to selection) or environment will change the heritability (Falconer and Mackay, 1996). Therefore, it is still important for heritability to be estimated for a specific trait for a given population at a specific time.

In human genetics, many heritability estimations are performed by utilising twins. In fact, it is the "workhorse" of research on understanding the genetic basis of individual differences in human quantitative traits (Plomin *et al.*, 2001). As a consequence of twin studies, it is now widely accepted that genetic factors are important sources of the variation in most human traits. Heritability is estimated by exploiting the fact that monozygotic (MZ) twin pairs share all of their genes and that dizygotic (DZ) twin pairs share on average half of their segregating genes. By comparing the resemblance of MZ pairs with DZ pairs reared together, the so-called classical twin design enables one to disentangle and quantify the proportion of phenotypic variance due to genetic (heritability) and common environmental factors shared by twin pairs. By assuming that MZ pairs share the same common environmental experiences with that of DZ pairs and all genetic variation is additive, the heritability of a trait can be inferred from the excess of similarity of MZ pairs compared to DZ pairs.

Besides the classical twin design, the heritability of quantitative traits has also been estimated using adoption designs. This is a very powerful design in separating the influence of genetics and environment on the resemblance between relatives. This design compares the resemblance between biologically related people who have been reared apart. These can be the comparison between offspring with their biological parents or monozygotic or dizygotic twins with their co-twins that have been adopted into different families. The similarities in their traits indicated the importance of genetic factors in the absence of a correlation between the environments of the biological and adopted parents. Currently, adoption studies are becoming very limited due to the declining number of adoptions as the consequences of contraception and the increased number of abortions (Plomin *et al.*, 2001).

1.2 Twins in Human Genetics

The idea of using twins in genetic studies has been usually attributed to the work of Francis Galton (1875) in his article "The history of twins, as a criterion of the relative powers of nature and nurture" published in *Fraser's Magazine* (1875) and the *Journal of the Anthropological Institute* (1876). While Galton was credited for proposing that twins can be used to study the effect of non-shared environmental factors on the resemblance of twins, he did not discover what is now called the twin design (Rende *et al.*, 1990). The use of the classical twin design is possible by the recognition that there are two types of twins by a Scottish obstetrician, J. Matthews Duncan in the 19^{th} century (Hall, 2003). The detailed method of the twin design for estimating heritability by comparing the resemblance between MZ and DZ twin pairs was described independently 50 years later by Curtis Merriman and Hermann Siemens, both in 1924 (Rende *et al.*, 1990). However, a recent paper by Liew *et al.* (2005) argued that the twin design was first described two years earlier by Walter Jablonski, an opthalmologist in Frankfurt, Germany. Jablonski compared the within pair difference of MZ and DZ twins to assess the contribution of heredity to the refraction in human eyes (Liew *et al.*, 2005).

In this era of molecular genetics, where the research interest has shifted from only understanding the relative importance of genetic and environmental factors influencing variation of human phenotypes toward dissecting the underlying gene(s), populations of twins are still important (Lyons and Bar, 2001; MacGregor *et al.*, 2000; Martin *et al.*, 1997). For genetic linkage and association studies, DZ twins are sibling pairs controlled for the age and common environmental effects. Compared to the ordinary sibling pairs, DZ twins have additional advantages. These include a decreased probability of non-paternity and the higher tendency of twins to participate in research (Martin *et al.*, 1997). MZ twins that are discordant for various traits are also useful for the study of epigenetics, gene expressions (MacGregor *et al.*, 2000) and variability genes (Berg, 1988). These advantages are coupled with the availability of twin registries worldwide that register and maintain contact with a large number of twins, providing enormous resources for a wide range of phenotypes (Boomsma *et al.*, 2002; Busjahn and Hur, 2006).

1.2.1 Biology of Twins

Monozygotic twins are derived from a single fertilized ovum and account for about a third of all spontaneous twinning (Hall, 2003). The reason for MZ twinning is not clear, but it is suggested as a result of a delay in timing of normal development, which can be caused by impaired transport through the fallopian tube, conception in close proximity to oral contraceptive use, and minor trauma to the blastocyst (Hankins and Saade, 2005). MZ twinning is virtually independent of race, genetic factors, maternal age and parity (Hankins and Saade, 2005).

Depending on whether twins share the same placenta (chorion) and membranes (amnion), MZ twins can be classified into three different types: dichorionic-diamniotic, monochorionic-diamniotic and monochorionic-monoamniotic (Hall, 2003; Phillips, 1993). About one third of MZ twins are of dichorionic-diamniotic type, where they share different placentas and membranes. The other two thirds of MZ twins share the same placenta but have different membranes (monochorionic-diamniotic). In addition, a small proportion of MZ twins (1-2%) have one set of placenta and membranes (monochorionic-monoamniotic) (Hall, 2003). The differences in placental and membranes types are highly related to the timing when a fertilised age divided into two separate zygotes, i.e. dichorionic-diamniotic MZ twins (formed between days 0-4); monochorionic-diamniotic MZ twins (formed between days 4-7); monochorionic-monoamniotic MZ twins (formed up to day 14) (Hall, 2003; Phillips, 1993) (see Figure 1.1). MZ twins are two genetically identical individuals (clones), except for very rare cases, such as heterokaryotypical MZ (MZ twins with different chromosomal composition) and MZ with chromosomal mosaicism (Gringas and Chen, 2001).

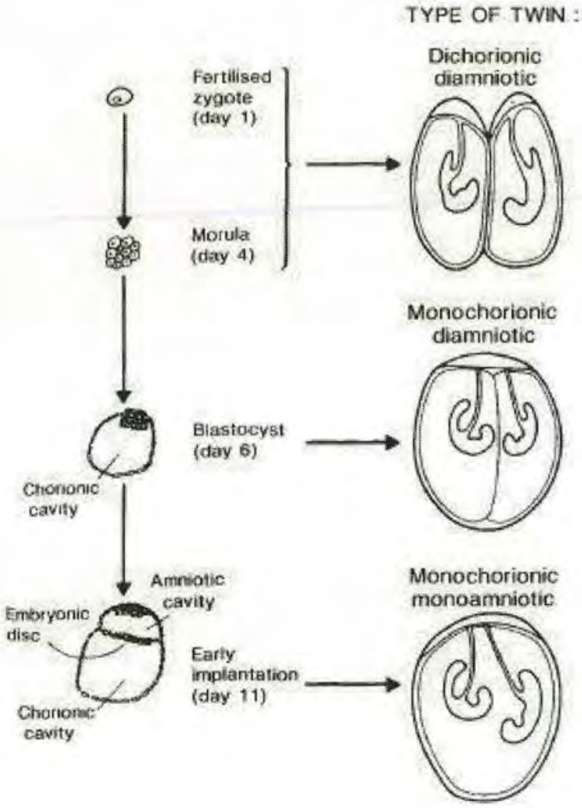

Figure 1.1: Three types of monozygotic twins [source: Phillips (1993)]

On the other hand, dizygotic twins are formed from two ova fertilized by two spermatozoa. Genetically, DZ twins are the same as regular sibling-pair, who share on average half of their segregating genes. DZ twins carry separate set of placentas and membranes (dichorionic-diamniotic). Dizygotic twinning happens because more than one dominant ovarian follicle was matured in the same menstrual cycle (Hall, 2003). Spontaneous DZ twinning is under genetic control and the combined risk to the first degree female relatives is more than 2 (Duffy *et al.*, 2001b). A region on chromosome 3 has been linked to DZ twinning (Busjahn *et al.*, 2000), but not replicated (Duffy *et al.*, 2001a). DZ twinning is associated with follicle stimulating hormone (FSH) (Hall, 2003) and influenced by race, parity, maternal age and nutrition (Hankins and Saade, 2005).

Spontaneous twinning rates vary across different countries and races (Hall, 2003). In Asian countries, about 6 in 1,000 livebirths are twins. In European and African countries, the prevalences are about 10-20 and 40 in 1,000 livebirths, respectively. While MZ twinning rate is fairly constant around the world (about 4 in 1000) (Hankins and Saade, 2005), DZ twinning varies across countries, races (Tong *et al.*, 1997) and time. Thus, the MZ/DZ ratio is different between countries and races.

One of the early methods designed to diagnose the zygosity of twins (whether twins are MZ or DZ) is to use physical similarities questionnaires, which include questions such as eye and hair colour, hair texture and facial appearance (Cederlof *et al.*, 1961; Sarna *et al.*, 1978). The questionnaires can determine the zygosity of twins with an accuracy of about 95% (Goldsmith, 1991). With the availability of molecular markers, zygosity can now be determined by using a set of DNA microsatellite markers, which can provide an accuracy of 98-99% (Becker *et al.*, 1997). Which methods to be used will depend on the type of research and resource availability. In large epidemiological studies, questionnaires can be easy and cheap to use, whereas biological assessment might be only feasible in smaller sampled studies due to a higher cost (Jackson *et al.*, 2001).

1.2.2 Twins and Heritability Estimation

The earliest use of twins in human genetics has been to resolve the 'nature-nurture' debate regarding the sources of individual differences for various human physical and mental characteristics (Galton, 1875; Thorndike, 1905). Both studies suggested the importance of nature or hereditary factors on the variation of human characteristics. Since then, twins have been and are being used widely to understand the genetic and environmental aetiology underlying variation in most human phenotypes, ranging from physical to behavioral characteristics (Boomsma *et al.*, 2002).

There are several designs which utilise twins for estimating the genetic and environmental sources of individual differences in human phenotypes, including the (classical) twin method, the adoption design, and the extended twin design. Among these designs, the classical twin method, which compares the similarity of MZ pairs to that of DZ pairs, is the most widely used. The reasons for its popularity include the relatively simple design, the availability of simple statistical methods for the analysis and the easy collection of MZ and DZ twin samples from available twin registries.

From standard quantitative genetic theory (see Falconer and Mackay, 1996; Lynch and Walsh, 1998), phenotypic variance (V_P) of a trait can be partitioned into variances due to additive genetic effects (V_A), common environmental effects shared by family members (V_C), dominance genetic effects (V_D), interaction between additive genetic effect (V_{AA}), interaction between additive and dominance genetic effects (V_{AD}), interaction between dominance genetic effects (V_{DD}), other non-additive genetic effects and specific individual environmental effects (V_E). For each of these effects, their covariance in MZ and DZ twin pairs is presented in Table 1.1.

Table 1.1: Covariances of MZ and DZ twin pairs

Covariance	$\mathbf{V}_A$	$\mathbf{V}_D$	$\mathbf{V}_{AA}$	$\mathbf{V}_{AD}$	$\mathbf{V}_{DD}$	$\mathbf{V}_C$	$\mathbf{V}_E$
Monozygotic (MZ) twins	1	1	1	1	1	1	0
Dizygotic (DZ) twins	1/2	1/4	1/4	1/8	1/16	1	0

In the classical twin design, the heritability of a trait is estimated by contrasting the covariance of MZ twin pairs to that of DZ twin pairs. Since MZ twins are genetically identical, the only source for their difference is the specific individual environmental effect. Thus, at first glance, it seems that MZ twins alone can provide an estimate for genetic variance. However, MZ pairs reared together also share a common environment from conception to birth and also during period they were reared together. Therefore, the genetic variance is confounded with the common environmental variance. With the availability of DZ twins, this problem may be partly overcome. DZ twins are genetically like full sibs, but share a common environment similar with that of MZ twins. Thus, by contrasting the resemblance between MZ to that of DZ pairs, the heritability of a trait can be estimated (Falconer and Mackay, 1996). The problem with the classical twin design is that there are many possible effects, but only two variance components can be estimated in addition to the environmental variance. To estimate heritability, further assumptions have to be made. Usually, in particular if the observed MZ correlation is less than twice the DZ correlation, variances due to epistatic effects (V_{AA}, V_{AD} and V_{DD}) are ignored. A more complicated design, which includes additional family relationships is required to estimate those variances.

From the classical twin design, there are only three observed statistics [variance of the trait (V_P)], MZ correlation (r_{MZ}) and DZ correlation (r_{DZ})] and four variance components to be estimated, V_C and V_D are confounded and cannot be estimated simultaneously. Dominance genetic variance increases the correlation between MZ pairs compared to DZ pairs, while common environmental variance increases the correlation between DZ pairs compared to MZ pairs. Therefore, in the classical twin design, these variances are usually modelled separately using an **ACE** (**A**dditive genetic, **C**ommon environment, **E**rror variances) or **ADE** (**A**dditive genetic, **D**ominant genetic, **E**rror variances) model depending on the observed correlations of MZ and DZ pairs. If MZ correlation is less than twice the DZ correlation, then an ACE model is appropriate, by assuming that no dominant genetic variance is involved. However, if the MZ correlation is more than twice DZ correlation, then an ADE model is usually applied with the assumption that common environmental variance is zero (see Evans *et al.*, 2002).

The variance component estimation in the next section will consider the commonly used ACE model. Components of variance can be estimated by comparing the similarity of MZ and DZ pairs. MZ twin pairs are genetically identical, so any pair difference is due to individual specific environmental effect. On the other hand, pair difference in DZ twins is caused by half of additive genetic variance and common environmental variance.

Let the phenotypic variance be standardised to 1 and a^2, c^2 and e^2 be the standardised variance components for V_A, V_C, and V_E, respectively. Then:

$$a^2 + c^2 + e^2 = 1 \tag{1.1}$$

$$r_{MZ} = a^2 + c^2 \tag{1.2}$$

$$r_{DZ} = \frac{a^2}{2} + c^2 \tag{1.3}$$

By assuming that both type of twins experienced the same common environmental exposures, i.e. the V_C component is the same for MZ and DZ pairs, a^2 and c^2 can be estimated as:

$$\hat{a^2} = 2(\hat{r}_{MZ} - \hat{r}_{DZ}) \tag{1.4}$$

and

$$\hat{c^2} = 2\hat{r}_{DZ} - \hat{r}_{MZ} \tag{1.5}$$

In an ADE model, the corresponding variance components can be obtained in a similar fashion by considering that MZ pair shared all of their dominance variance, whereas a quarter is shared by DZ pairs. These simple models can be extended into more complicated models, including sex-specific variance components (Chapter 5) and multiple-traits analysis (Chapters 4 and 5) and a maximum likelihood-based method is usually used.

The classical twin design is based on several important assumptions. These include the assumptions that twins are representative of the general population, that MZ twins share all of their genotype, and that the common environmental experiences shared by both type of twins are the same. While there are still criticisms about the reliability of the assumptions (e.g. Joseph, 2002; Phillips, 1993), these assumptions are testable and can be justified

(Evans and Martin, 2000; Martin *et al.*, 1997). Provided that these assumptions are satisfied, the twin design is a very useful tool in human genetics for understanding the genetic and environmental causes of individual difference in human phenotypes.

Another important assumption in heritability estimation or population genetic studies in general, is that mating occurs at random. However, it is known that for some characteristics, mating is not random. For example, Silventoinen *et al.* (2003) reported that for body height and body mass index (BMI), there was evidence for assortative mating. Positive phenotypic assortment was also found in other human phenotypes, including physical characteristics, education, religion, personality, socioeconomic status and cognitive traits (see e.g. Cavalli-Sforza and Bodmer, 1999; Merikangas, 1982). In the classical twin design, the effect of assortative mating cannot be estimated because phenotypic information on the parents are not available. A more complicated design, which includes the parents of the twins is needed. However, when estimating heritability using the classical twin design, it is still important to remember the possible effect of assortative mating on heritability estimates. In the presence of assortative mating, the phenotypic correlation between relatives increases. In the classical twin design, this tends to increase the covariance of DZ pairs. Since the heritability was estimated by contrasting the covariance of MZ to that of DZ pairs, an increase in DZ covariance will produce a biased downward heritability estimate (Neale and Maes, 2004).

The classical twin design relies on the availability of zygosity information. While zygosity status of the twins can now be easily and economically obtained, it is not always known. The zygosity status of twins identified from large population-based surveys in the fields of social sciences, economics or education (Deary *et al.*, 2004; Scottish Council for Research in Education, 1933; Scottish Council for Research in Education, 1949; Scarr-Salapatek, 1971; Webbink *et al.*, 2006) are usually unknown. Genetic analysis for these data is still possible, but it requires different statistical methods. A finite mixture distribution model has been proposed to analyse data from twins of unknown zygosity (Neale, 2003). In Chapter 2, this model is applied to analyse IQ data from twins of unknown zygosity from the Scottish Mental Surveys 1932 (Scottish Council for Research in Education, 1933) and 1947 (Scottish Council for Research in Education, 1949). The precision and bias of the model for single trait (Chapter 3) and multiple traits (Chapter 4) are assessed by computer simulation.

1.2.3 Twins and Gene Identification

Two of the most commonly used methods to map and identify genes underlying human quantitative traits are linkage and association studies. While linkage is considered ideal for detecting genes with large effects over a larger genetic distance from a marker, association is very good for detecting genes with small effects and closer to a marker. Hence, these methods are said to be complementary (Plomin *et al.*, 2001).

Populations of twins are not only ideal for estimating the heritability of human phenotypes, they are also useful for identifying genes through linkage and association studies. Sibling-pairs can be used for linkage analysis and family-based association test. Compared to ordinary siblings, DZ twins have additional advantages. DZ twins are matched for age, which is important for traits whose expression are age specific (Martin *et al.*, 1997). DZ twins are also matched for measured and unmeasured shared environmental effects. This makes the separation of genetic and environmental factors from the phenotypic difference between twins an easier task (MacGregor *et al.*, 2000).

Genetic Linkage:

Linkage analysis tests for the (co)segregation of a quantitative trait loci (QTL) and a marker allele in a pedigreed population. In a sibling-pair design, the evidence for linkage between a marker and a trait is provided by the excess of alleles shared identical by descent (IBD) at a marker for phenotypically similar pairs. In other words, linkage analysis correlates the genetic similarity of sibling-pairs, which is expressed as the proportion of allele shared IBD ($\hat{\pi}$), with their phenotypic similarity.

In a sibling-pair design, the earliest method for detecting linkage between a marker and a QTL is the Haseman-Elston regression method (Haseman and Elston, 1972). This method regresses the phenotypic similarity of sibling-pairs on their genotypic similarity. If X and Y are the standardised phenotypic values of Sib_1 and Sib_2, their phenotypic similarity is expressed as $(X-Y)^2$ and their genotypic similarity is expressed as $\hat{\pi}$. Therefore, the regression equation as described in Sham and Purcell (2001) is

$$(X-Y)^2 = 2(1-r) - 2Q(\hat{\pi} - 0.5) + \epsilon \tag{1.6}$$

r and Q are the sibling correlation and the proportion of phenotypic variance explained by the additive effects of the QTL, respectively. The presence of linkage between a marker and a QTL is tested by comparing the null hypothesis that the regression coefficient $(-2Q)$ is zero against the alternative hypothesis that the regression coefficient is negative.

Another popular method for linkage analysis is the variance component approach, which is considered to be more powerful than the Haseman-Elston method (e.g. Almasy and Blangero, 1998; Amos, 1994; Sham and Purcell, 2001; Visscher and Hopper, 2001). This method is basically an extension of variance components model used for heritability estimation (Plomin *et al.*, 2001), in which Q is added into the sibling-pair variance-covariance structure. The evidence of linkage is provided if Q is significantly greater than zero. In Chapter 6, a variance component linkage analysis to map QTL for body height is presented.

Genetic Association:

While linkage analysis correlates the phenotypic similarity with the genotypic similarity of individuals in a family, genetic association tests for the correlation between a genetic marker and a trait in a population. The association between a marker and a quantitative trait is tested by comparing the phenotypic values of individuals with different genotypes. The evidence for association is provided if there is a significant phenotypic difference between the genotype classes.

Association studies can be subject to spurious associations (Lander and Schork, 1994). That is, the association between a marker allele and a trait can be caused by other factors such as ethnicity, admixture and population stratification. Consider the example of testing for an association between a marker and adult height in a mixture of Chinese and European populations, where due to population structure only, the frequency of the A allele is higher in the European population. As Europeans are on average taller than Chinese, an association between height and the A allele will be observed. However, this is entirely due to population structure.

The association conducted within a family, family-based association test, is robust to admixture and population stratification (Laird and Lange, 2006) because ethnicity or population substructure do not vary within families. In a sibling pair design, the association between a marker and a quantitative trait can be partitioned into within and between pair associations (Fulker *et al.*, 1999). Because within pair association accounts for shared genetic and environmental effects, it reflects only the true association. On the other hand, the between family association reflects both the true and possible spurious associations (Posthuma *et al.*, 2004).

1.3 The Aim and Organization of The Thesis

The aim of the research presented in this thesis is to enhance our understanding of the genetic basis of the variation of quantitatively distributed human complex phenotypes using twin and (to some extent) family data. These include the variation of behavioural traits (intelligence), clinical traits (the metabolic syndrome) and anthropometric measures (body height). The focus in this thesis is on heritability estimation using both conventional and novel statistical methods. The identification of gene(s) underlying complex phenotypes (body height) by means of linkage analysis is also presented.

The thesis is organised into 7 chapters. The focus in Chapters 2-4 is on a novel finite mixture distribution method for estimating heritability from twins of unknown zygosity. The estimation of heritability of IQ score from population-based surveys, the Scottish Mental Surveys of 1932 and 1947, where zygosity is unknown is discussed in Chapter 2. Discussions on the precision and bias of the finite mixture distribution model using computer simulations and its application to IQ measures from a large sample of zygosity known twins [twins from the U.K. Twins' Early Developments Study (TEDS)] are presented in Chapter 3. The extension of the finite mixture distribution to multivariate analysis is discussed in Chapter 4. The application of a mixed linear model for estimating heritability from twins of known zygosity for the traits associated with the metabolic syndrome is presented in Chapter 5. Moving from heritability estimation, the focus of Chapter 6 is on linkage analysis for identification of gene/chromosomal regions associated with body height. In addition to twin data, the linkage analysis also utilised the data from other family members. Finally, the thesis is concluded in Chapter 7 (General Discussion).

2 A Mixture Distribution Model to Estimate Variance Components from Twins of Unknown Zygosity: Application to IQ Measures from The Scottish Mental Surveys of 1932 and 1947

2.1 Abstract

Twin studies provide estimates of genetic and environmental contributions to cognitive ability differences, but could be based on biased samples. This chapter presents whole-population estimates using twins from unique mental surveys in Scotland. The Scottish Mental Surveys of 1st June 1932 (SMS1932) and 4th June 1947 (SMS1947) administered the same validated verbal reasoning test to almost everyone born in 1921 or 1936, respectively, and attending school in Scotland. There were 572 twin pairs from the SMS1932, and 517 pairs from the SMS1947. Information on zygosity was unavailable. A novel application of a mixture distribution was used to estimate genetic and environmental components of verbal reasoning variation by maximum likelihood. The study found consistent heritability (~0.70) and shared environment (~0.21) estimates. The estimates decreased slightly when additional quantitative traits (height and weight) were added in a multivariate analysis. More generally for studies in genetics, the methodological innovation developed here implies that large (national) data collections can provide sufficient information on twin pairs to estimate genetic parameters, even without zygosity.

2.2 Introduction

Intelligence differences in humans have a well-understood phenotypic structure (Carroll, 1993) , strong predictive validity for health, education and occupational outcomes (Gottfredson and Deary, 2004; Neisser *et al.*, 1996), and correlate with brain structure and function (Gray and Thompson, 2004). The genetic and environmental contributions to variation in intelligence at different ages are of considerable interest, but are not fully understood (Plomin and Spinath, 2004). To date, twin, family and adoption studies suggest that, including studies at all ages, about 50% of the variation in human intelligence, as measured by psychometric tests, is attributable to additive genetic effects (Bouchard and McGue, 2003; Bouchard *et al.*, 1990; Plomin and Spinath, 2004; Plomin *et al.*, 2001). These sources also indicate that: the bulk of the substantial environmental effect is from sources not shared by siblings in the same rearing family; the genetic influence is especially strong on the general intelligence factor; the genetic contribution is stronger in adulthood than in childhood; and the effect of the shared rearing environment decreases almost to zero in early adulthood.

A potentially serious and unanswered problem is the representativeness of the samples on which genetic and environmental contributions to variation in cognitive ability scores are based (Bouchard and McGue, 2003; Joseph, 2003). Volunteer samples could be especially prone to such effects and attrition from population-referenced samples could have biasing effects. Therefore, it is highly desirable to provide estimates of environmental and genetic contributions to intelligence variation based upon complete populations.

The present study analysed data from unique, whole populations of 11-year-olds tested in the Scottish Mental Surveys of 1932 (SMS1932) (Scottish Council for Research in Education, 1933) and 1947 (SMS1947) (Deary *et al.*, 2004; Scottish Council for Research in Education, 1949). On June 1st 1932 and June 4th 1947, the Scottish Council for Research in Education organised the mental testing of all children attending Scottish schools and born in 1921 or 1936, respectively. In the Scottish Mental Survey of 1932 (SMS1932) 87,498 children were tested (Scottish Council for Research in Education, 1933), and 70,805 were tested in the Scottish Mental Survey of 1947 (SMS1947) (Scottish Council for Research in Education, 1949). The mental test used was a version of the Moray House Test No. 12 (Scottish Council for Research in Education, 1933).

In addition to the novelty of using data from an entire population, a finite mixture distribution model that does not require zygosity of twin pairs to be known (Neale, 2003) was applied. This method provides a unique opportunity to perform genetic analysis on data collected (for non-biological purposes) from large population cohorts in, for example, the social sciences, as long as twin pairs can be identified from local identifiers such as family, school and date of birth.

2.3 Subjects and Methods

2.3.1 Study Populations

Twin pairs were explicitly ascertained in the SMS1947, and some extra demographic and social information was collected from them (Mehrota and Maxwell, 1949; Scottish Council for Research in Education, 1949; Scottish Council for Research in Education, 1953). No attempt was made to establish zygosity. A total of 517 pairs were identified, 320 same-sex (SS) pairs and 197 opposite-sex (OS) pairs. In the SMS1932, twin pairs were not explicitly ascertained (Deary *et al.*, 2004; Scottish Council for Research in Education, 1933). They were identified for the present study by matching pairs of subjects for: surname, date of birth and school identifier. A total of 572 pairs were ascertained, 382 SS and 190 OS pairs. The zygosity status of these twin pairs is not known. The number of twin pairs as a proportion of the entire population was 0.64% and 0.70% for the 1932 and 1947 populations, respectively, slightly lower than the current rate of twinning in Caucasian populations (Imaizumi, 2003). In total, the two populations contained 1089 twin pairs, 702 SS and 387 OS twin pairs.

Twins who attended different schools, including brothers and sisters attending single sex schools (rarer at the primary level than at the secondary level) would not be identified. The computerised SMS1932 database is not complete. The ledgers containing the data from Fife, Angus and Wigtown have not been traced. There is no obvious bias entailed by these omissions as there is no reason to believe that genetic and environmental contributions to intelligence would be any different in those areas. In one area of Scotland, 10 twin pairs were from birth years 1922 and 1923. These were retained because all children were tested in the area from these birth years. Another two of the 'twin' pairs are from triplets. Two pairs had no information on sex and were omitted.

2.3.2 Measures

The version of the Moray House Test No. 12 (MHT) was very like that used in the United Kingdom for selection from primary to secondary school education when children were about 11 years old (Deary *et al.*, 2004). Based on 71 questions, the maximum possible score on this test is 76. The MHT is a group-administered test with a time limit of 45 minutes, and contains a preponderance of verbal reasoning items, but also other material including numerical and spatial items. It was validated against the Stanford revision of the Binet test (r ~0.8) (Scottish Council for Research in Education, 1933). The MHT has high stability of individual differences over more than 60 years, with a correlation coefficient between MHT score between age 11 and 80 of 0.66 (Deary *et al.*, 2000; Deary *et al.*, 2004).

Additional phenotypic data, including height, weight and hence body mass index (BMI, body mass in kg divided by the square of the height in m) were available for the SMS1947 twins. Height and weight were measured by a nurse either in the school or at home (Scottish Council for Research in Education, 1949). Since these traits have well-known heritabilities, including these indices afforded a check on heritability estimates against other studies. Moreover, some studies have reported a moderate phenotypic correlation between IQ and height (Humphreys *et al.*, 1985; Johnson, 1991; Teasdale *et al.*, 1989). By performing multivariate genetic analyses, the present study allowed an examination of the environmental and genetic correlations between cognitive ability and height. Lastly, in the method that was used to estimate genetic parameters, multiple traits provide more information on zygosity than data on a single trait.

2.4 Statistical Methods

Zygosity status, i.e. whether a twin pair is either monozygotic (MZ) or dizygotic (DZ), was not available for the same-sex twin pairs. Opposite-sex pairs are DZ. The number of MZ twins as a fraction of all twin pairs (or as a fraction of SS pairs) can be estimated assuming that the probability that a DZ pair is same-sex is 0.5. From each of the populations, the proportion of MZ twin pairs was estimated using Weinberg's differential rule as 1 - 2 × (proportion of OS twin pairs) (Weinberg, 1902).

Intraclass correlations of SS and OS were obtained by partitioning the total variance into a between (σ_b^2) and within (σ_w^2) pair variance using ANOVA, and fitting sex and cohort (for the combined MHT score of the SMS 1932 and 1947) as covariates.

To partition the observed intraclass correlation of OS and SS twin pair phenotypic similarity into possible underlying causes, a model was fitted that partitions the covariance between twin pairs into an additive genetic (A), a common environmental (C), and a residual environmental (E) component of variance (Neale and Maes, 2004). The variance components of MHT score and other traits were estimated by fitting a finite mixture distribution method (Neale, 2003) using the Mx statistical package (Neale *et al.*, 2002).

The basic principle of the mixture model is illustrated in Figure 2.1, where two normal distributions are fitted to the distribution of same-sex pair difference. The combined MHT score, after adjustment for a sex and a cohort effect was used as an example. There is strong statistical evidence for the hypothesis that there are two distributions against the hypothesis that there is a single normal distribution (likelihood-ratio-test, 2 degrees of freedom, $P - value = 6.1 \times 10^{-10}$). From the data, the estimated proportion of MZ pairs among SS pairs was 0.45 and used to weight the analysis. The estimates of the variances for the two underlying distributions, assumed to correspond to MZ and DZ pairs, are 48.4 and 236.1, respectively, assuming that the smaller within-pair variance corresponds to MZ twins. These estimates of the within-pair variance correspond to $2 \times (1 - r)$ times the phenotypic variance, where r is the MZ or DZ correlation. Given the estimate of the total phenotypic variance of 243.4 (Table 2.4), the inferred estimates of the MZ and DZ intra-class correlations are 0.90 and 0.51, respectively, which correspond well to the results from the full mixture model, presented later. In the full model, both the within and between same-sex variances are partitioned into two groups simultaneously, and appropriate weights are given to all sources of information by maximum likelihood.

Sex and cohort effects (for combined MHT score) were fitted as covariates for all traits. The mixture proportion, in this case the proportion of MZ pairs among SS pairs, is assumed to be known, and the variance components are estimated using maximum likelihood.

Male and female specific variance components were estimated and a likelihood-ratio-test was used to test the hypothesis that these components were the same. The covariance between same-sex DZ pairs and opposite-sex DZ pairs is not necessarily equal; for example, if there are sex-specific genetic effects or if the shared environmental covariance differs between the two types of DZ pairs. This difference was estimated by allowing the genetic correlation between males and females to differ from unity, and tested using a likelihood-ratio-test.

Since additional phenotypic data were available from the SMS1947 population, multivariate genetic analyses were performed on the 1947 data, with height, weight, BMI and MHT score as phenotypes. Given that additional phenotypes provide additional zygosity information, the mixture model is likely to give more precise estimates of parameters with multiple traits (Neale, 2003).

2.5 Results

Descriptive statistics of the MHT score and additional traits are presented in Table 2.1. As is known, girls scored higher in the MHT in the SMS1947, and there was an increase in the mean MHT score from the 1932 population to the 1947 population (Scottish Council for Research in Education, 1949). In

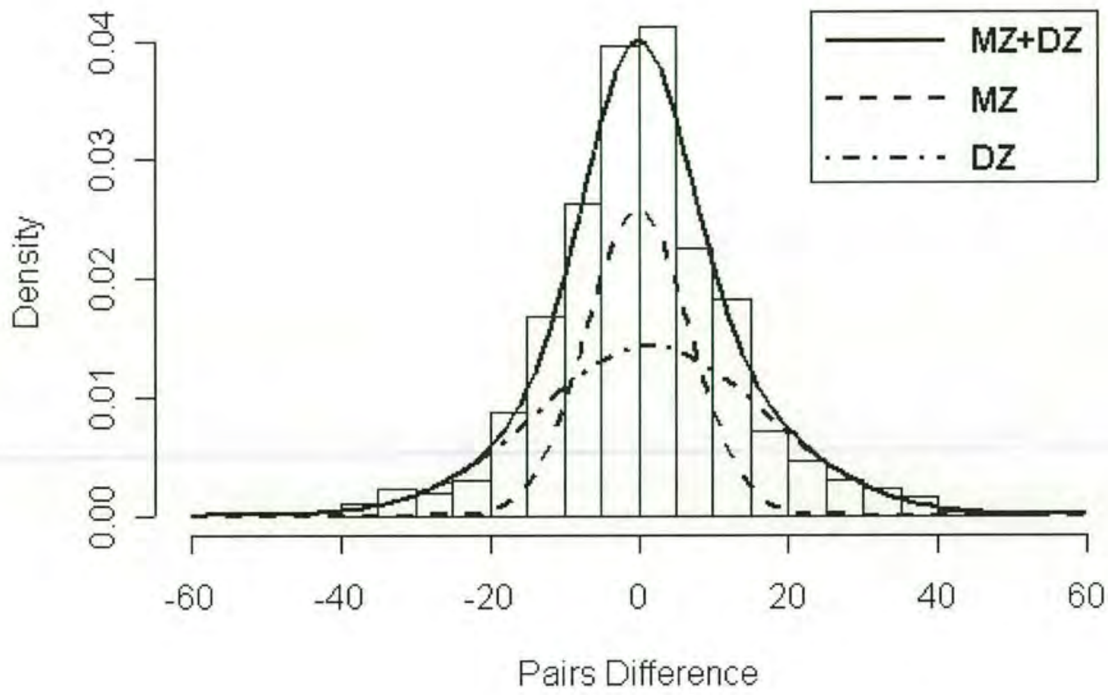

Figure 2.1: Probability density function of same sex (SS) pair difference for the combined Moray House Test score, after adjusting for sex and cohort effects, and the fitted curves for two underlying distributions, assumed to correspond to MZ and DZ pairs. The histogram represents the observed distribution of the SS pairs difference and the solid curve the sum of the fitted distributions from the mixture model. The estimated mean, variance, skewness and kurtosis from the single distribution of SS pair difference are 0.78 ± 0.49, 152.1 ± 8.6, 0.28 ± 0.10 and 1.40 ± 0.20, respectively. The estimated means (variances) of the inferred MZ and DZ distribution of pair difference are -0.12 ± 0.42 (48.4 ± 4.1), 1.52 ± 0.83 (236.1 ± 18.1), respectively. The estimated proportion of MZ among SS pairs of 0.45 was used to weight the analysis.

SMS1932, the proportion of monozygotic (MZ) among same-sex (SS) twins was 0.50 ± 0.04. The estimate was smaller in SMS1947 (0.38 ± 0.06), but the two are not significantly different ($P - value > 0.05$). When the twin data were combined, the proportion of MZ in SS was 0.45 ± 0.03, close to the estimated proportion of MZ in Caucasian populations (Imaizumi, 2003). A summary of the estimated proportion of MZ twins in the sample of SS twins is given in Table 2.2. The estimated proportions of MZ among SS twins were used in the mixture analysis.

Intraclass correlations for SS and opposite-sex (OS) twins for each trait are presented in Table 2.3. For all traits, the intraclass correlations of the SS twins were consistently higher than that of OS twins. This suggests that genetic variation contributed to the individual phenotypic differences.

2.5.1 Univariate Analyses

From univariate analyses, the additive genetic, shared environmental and specific environmental variance components and the corresponding 95% confidence interval (CI) of all traits are presented in Table 2.4. There were no significant differences between variance components for MHT scores estimated from the 1932 and 1947 populations ($P - value > 0.05$). Therefore the combined MHT score data after adjusting for the cohort effect provides more precise estimates of variance components for MHT score. A large proportion of phenotypic variance in MHT score, approximately 70% (95% CI 58% to 83%), was attributed to additive genetic effects. In addition, an estimated proportion of 21% (95% CI 10% to 32%) of the phenotypic variance was due to common environmental effects shared by twin pairs. These estimates imply a large repeatability ($\sim$0.90) of MHT score in the Scottish population at age 11 in the 1930s and 1940s. There

Table 2.1: Descriptive statistics of the Scottish Mental Surveys' twin data.

Traits	**Sex**	**N**	**Mean**	**SD**	**CV (%)**
Height (m)(SMS1947)	M	488	1.37	0.07	4.98
	F	521	1.37	0.07	5.12
	Total	1009	1.37	0.07	5.11
Weight (kg)(SMS1947)	M	480	31.14**	4.33	13.89
	F	519	30.13**	4.41	14.65
	Total	999	30.61	4.40	14.37
BMI (kg/m^2)(SMS1947)	M	480	16.68**	1.68	10.09
	F	519	16.11**	1.59	9.88
	Total	999	16.38	1.66	10.13
MHT Score (SMS1932)	M	508	28.45	15.31	53.81
	F	571	28.46	14.89	52.31
	Total	1080	28.46	15.08	52.99
MHT Score (SMS1947)	M	451	30.88*	16.77	54.31
	F	498	33.41*	15.81	47.32
	Total	949	32.21	16.31	50.64
Combined MHT Score	M	959	29.59	16.05	54.24
	F	1070	30.76	15.51	50.42
	Total	2029	30.21	15.78	52.23

Note: Column N is the number of individuals, and column SD and CV are the standard deviations and coefficient of variation, respectively. * and ** denote significant differences between males and females at 5% and 0.1%, respectively.

Table 2.2: Estimated proportion of monozygotic (MZ) twin pairs

Cohort	SS pairs	OS pairs	Total	pMZ (SE)	pMZ|SS (SE)
SMS1932	382	190	572	0.34 (0.04)	0.50 (0.04)
SMS1947	320	197	517	0.24 (0.04)	0.38 (0.06)
Combined	702	387	1089	0.29 (0.02)	0.45 (0.03)

Note: pMZ and pMZ|SS are the estimated proportion of MZ twins in the population of twin pairs and the proportion of MZ twins in SS twin pairs, respectively. pMZ was estimated using the formula $1 - 2\times$(proportion of OS twin pairs). pMZ|SS was estimated as pMZ/(proportion of SS twin pairs).

was no significant difference between variance components in males and females for all traits ($P - value > 0.05$, likelihood-ratio test). The genetic correlation between males and females was not significantly different from unity for all traits ($P - value > 0.05$). This suggests that the same set of genes influenced the phenotypes in males and females.

Eighty percent (95% CI 65% to 95%) of the phenotypic variation in height was due to genetic effects, under the assumed ACE model. Only 14% (95% CI 0% to 28%) of the height variation between individuals was due to common environmental effects shared by twin pairs. A large proportion of variance due to additive genetic effects was also observed for weight (73%, 95% CI 58% to 90%) and BMI (84%, 95% CI 70%-97%).

Table 2.3: Intraclass correlations for same-sex (SS) and opposite-sex (OS) twins from the Scottish Mental Surveys of 1932 and 1947 (SMS1932, SMS1947).

Traits	**Twin**	σ_b^2	σ_w^2	**t**	**SE(t)**
Height (m)(SMS1947)	SS	0.003	0.002	0.69	0.03
	OS	0.002	0.002	0.54	0.05
Weight (kg)(SMS1947)	SS	14.9	5.64	0.73	0.03
	OS	8.16	8.43	0.49	0.05
BMI (kg/m^2)(SMS1947)	SS	2.16	0.85	0.72	0.03
	OS	1.0	1.14	0.47	0.06
MHT Score (SMS1932)	SS	153.0	74.2	0.67	0.03
	OS	122.0	89.8	0.58	0.05
MHT Score (SMS1947)	SS	198.2	78.8	0.72	0.03
	OS	153.0	93.1	0.62	0.04
Combined MHT Score	SS	169.9	76.2	0.69	0.02
	OS	135.4	91.0	0.60	0.03

Note: σ_b^2, σ_w^2 are between and within twin pair variances obtained from ANOVA after adjusting for sex and cohort effects (for combined MHT Score); t and SE(t) are the estimated intraclass correlation and its corresponding approximate standard error.

Table 2.4: Variance components of MHT score and additional traits obtained from univariate analyses using the mixture distribution model.

Traits	A	C	E	T	a^2	c^2	e^2
	(95% CI)	**(95% CI)**	**(95% CI)**	**(95% CI)**	**(95% CI)**	**(95% CI)**	**(95% CI)**
Height	0.0040	0.0007	0.0003	0.0050	0.80	0.14	0.06
(SMS1947)	(0.0030-0.0050)	(0-0.0010)	(0.0002-0.0004)	(0.0040-0.0050)	(0.65-0.95)	(0-0.28)	(0.04-0.09)
Weight	13.90	3.62	1.42	18.94	0.73	0.19	0.08
(SMS1947)	(11.06-16.89)	(0.64-6.58)	(0.85-2.24)	(17.08-21.12)	(0.58-0.90)	(0.04-0.33)	(0.04-0.12)
BMI	2.22	0.29	0.13	2.63	0.84	0.11	0.05
(SMS1947)	(1.86-2.63)	(0-0.69)	(0.08-0.20)	(2.38-2.93)	(0.70-0.97)	(0-0.25)	(0.03-0.08)
MHT Score	164.8	36.3	25.1	226.3	0.73	0.16	0.11
(SMS1932)	(121.4-207.9)	(0-76.4)	(16.8-37.8)	(205.2-250.8)	(0.53-0.91)	(0-0.32)	(0.07-0.17)
MHT Score	176.6	68.1	17.3	262.0	0.67	0.26	0.07
(SMS1947)	(136.8-217.3)	(26.7-110.1)	(9.5-29.5)	(235.7-292.8)	(0.52-0.84)	(0.11-0.40)	(0.04-0.11)
Combined	170.9	51.0	21.6	243.4	0.70	0.21	0.09
MHT Score	(142.2-200.1)	(22.4-79.6)	(15.6-29.5)	(226.4-262.4)	(0.58-0.83)	(0.10-0.32)	(0.06-0.12)

Note: A, C, E, and T are the additive genetic, common and shared environmental and total variance for each trait, respectively; a^2, c^2 and e^2 are the standardized variance for A, C and E, respectively.

2.5.2 Multivariate Analysis

Multivariate analyses of MHT score with height, weight, and BMI were performed on data from the SMS1947 population. A full 4-trait analysis could not be carried out because of the dependence between BMI, height and weight. Therefore, 2 trivariate analyses were performed i.e. height, BMI and MHT score and weight, BMI and MHT score. From the multivariate analyses, 59% (95% CI 43% to 78%) of the phenotypic variance of MHT score of SMS1947 was attributed to genetic variance compared to 67% (95% CI 52% to 84%) from the univariate analysis. There was a slightly higher proportion of variance due to common environmental effects in the multivariate analyses (31% compared to 26%). Estimates of correlation coefficients [genetic (r_g), common (r_c) and specific (r_e) environmental and phenotypic (r_p)] were derived from the estimated (co)variance components. Estimated phenotypic correlations between the measured traits were consistent with estimates of Pearson's correlations that ignored the twin structure of the data. The estimated genetic correlation coefficients between MHT score and the additional traits (height, weight, BMI) were not significantly different from zero ($P - value > 0.05$). There was a significant phenotypic correlation coefficient between height and MHT score [0.28 (95% CI 0.21 to 0.35); $P - value < 0.001$], and the common environmental correlation coefficient between height and MHT score was high [0.74 (95% CI 0.29 to 1.0)]. Table 2.5 shows the estimates of all variance components from multivariate analyses, and the corresponding estimates of correlation coefficients are presented in Table 2.6.

2.6 Discussion

This study is the first to estimate genetic and environmental variance components for verbal reasoning ability (MHT score) in entire school-attending populations.

Table 2.5: Variance components obtained from multivariate analysis of SMS1947.

Traits	A	C	E	T	a^2	c^2	e^2
	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)
Height	0.0040	0.0009	0.0004	0.0050	0.74	0.18	0.08
(SMS1947)[1]	(0.0030-0.0040)	(0.0002-0.0020)	(0.0003-0.0006)	(0.0040-0.0050)	(0.59-0.88)	(0.05-0.32)	(0.06-0.13)
Weight	13.30	3.96	1.72	18.97	0.70	0.21	0.09
(SMS1947)[2]	(10.39-16.32)	(1.03-6.89)	(1.19-2.58)	(17.14-21.12)	(0.55-0.86)	(0.06-0.35)	(0.06-0.14)
BMI	2.21	0.29	0.14	2.64	0.84	0.11	0.05
(SMS1947)[1]	(1.86-2.59)	(0-0.70)	(0.09-0.21)	(2.39-2.94)	(0.70-0.96)	(0-0.25)	(0.04-0.08)
MHT Score	157.1	81.6	26.0	264.7	0.59	0.31	0.10
(SMS1947)[1]	(114.7-199.4)	(39.7-124.7)	(16.8-39.7)	(238.0-296.1)	(0.43-0.78)	(0.15-0.44)	(0.06-0.15)

Note: A, C, E and T are the additive genetic, common and shared environmental and total variance for each trait, respectively; a^2, c^2 and e^2 are the standardized variance for A, C, E, respectively. A full 4-trait analysis could not be carried out because of the dependence between BMI, height and weight. The results for height, BMI and MHT score are from a trivariate analysis of these traits. The results for weight are from a trivariate analysis of weight, BMI and MHT score.

Table 2.6: Phenotypic (r_p), additive genetic (r_g) common (r_c) and specific environmental (r_e) correlations estimated from multivariate analysis of SMS1947, and their 95% confidence interval (CI)

Traits	r_p	r_g	r_c	r_e
	(95% CI)	(95% CI)	(95% CI)	(95% CI)
Height and	0.28	0.15	0.74	0.06
MHT Score[1]	(0.21-0.35)	(−0.01-0.32)	(0.29-1.0)	(−0.21-0.31)
Weight and	0.19	0.09	0.47	0.13
MHT Score[2]	(0.11-0.26)	(−0.08-0.25)	(0.04-0.99)	(−0.13-0.37)
BMI and	−0.01	−0.003	−0.13	0.17
MHT Score[1]	(−0.09-0.06)	(−0.15-0.14)	(−1.0-1.0)	(−0.09-0.42)
Height and	0.71	0.65	0.88	0.85
Weight[3]	(0.67-0.74)	(0.56-0.72)	(0.45-1.0)	(0.75-0.91)
Height and	−0.02	−0.13	0.43	0.33
BMI[1]	(−0.09-0.06)	(−0.27-0.01)	(−1.0-1.0)	(0.05-0.56)
Weight and	0.69	0.68	0.79	0.73
BMI[2]	(0.65-0.73)	(0.61-0.74)	(−1.0-1.0)	(0.59-0.84)

Note: A full 4-trait analysis could not be carried out because of the dependence between BMI, height and weight. [1] is from a trivariate analysis of height, BMI and MHT score. [2] is from a trivariate analysis of weight, BMI and MHT score, and [3] is from a trivariate analysis of weight, height and MHT score.

Hence, it may be assumed that there was no bias in parameter estimates due to ascertainment. The estimates for the additive genetic contribution to differences in MHT scores were 70% in the combined analysis, with a 95% CI of 58 to 83%, and highly consistent across the two whole populations. Estimates of heritability in later-born cohorts of children of similar ages from twin studies with known zygosity vary from 40% to 70% (Bartels *et al.*, 2002; Bishop *et al.*, 2003; Knopik and DeFries, 1998; Plomin *et al.*, 2001).

The heritability estimate of height was consistent with the published estimates using twin designs with known zygosity (Schousboe *et al.*, 2004; Silventoinen, 2003b; Silventoinen, 2003a). The heritability estimates of weight and BMI in SMS1947 are similar with the estimates from previous twin studies (reviewed by Maes *et al.*, 1997). Analyses of these traits have been useful to provide a check on the reliability of the mixture distribution methods. In addition, the similarity of variance components estimated from multivariate analyses compared to univariate analyses indicated that the estimates are precise.

From multivariate analyses, although a significant phenotypic correlation between MHT score and height was estimated, there was no significant genetic component of this correlation. The significant phenotypic correlation between MHT score and height was attributed to common environmental correlation. It can probably be explained that the same common environmental effects (social economic status and better nutrition) influence both height and intelligence.

The mixture distribution maximum likelihood method that was used has not been used before to estimate variance components in twin studies without known zygosity. From simulation studies, the estimate of the heritability using a mixture distribution appeared slightly biased upwards (Neale, 2003). However, when new simulations were run with approximately the same population parameters as estimated in this study, the upward bias in the estimate of the heritability was only $\sim$2% (M. Neale, personal communication). The information to separate the distribution of same sex pair differences (for analysis within pairs) and pair sums (for analysis between pairs) into two underlying distributions comes from the contrast between the variance and kurtosis of the distribution (see Appendix 2A). Therefore, if the distribution of a trait is strongly kurtotic then the mixture distribution may result in biased results. For the MHT scores considered in this study, kurtosis was estimated as -0.71 (SE 0.15). Although this is significantly different from the expectation under normality, the estimate indicates that the distribution is platykurtic, which is the opposite of what is expected and observed when the differences between SS pair observations are analysed. The departure from normality is discussed further in Chapter 3.

Heath *et al.* (2003) showed that a latent class analysis can be used for zygosity diagnosis. The authors applied their analysis to discrete data from standard questions for zygosity diagnosis and fitted a 2-class latent class model, where the two classes are assumed to correspond to MZ and DZ groupings. In principle this method can be used for any discrete data on twins, and can be viewed as a discrete-trait version of the mixture model for quantitative traits that has been used.

An alternative approach would be to use separate ANOVA for OS and SS data to estimate intraclass correlations and to estimate heritability from these assuming that the mixture proportion in the SS pairs is known (see Appendix 2B for details). However, although these derivations help to understand and quantify the relationship between the population parameters and estimates in a least squares framework, this method does not use all information efficiently and could lead to severe bias if the male-female genetic correlation deviates from unity.

The proportion of MZ among same-sex twin pairs was assumed to be known in the analysis. A sensitivity analysis showed that varying this parameter from 0.35 to 0.65 has little effect on the estimate of heritability. For the combined analysis of MHT score, the estimate of the heritability for MZ proportions among same-sex twin pairs of 0.35 and 0.65 was 0.71 (95% CI 0.59 to 0.82) and 0.69 (95% CI 0.54 to 0.83), respectively.

In this study the commonly used ACE model was fitted. Other three-component parameterisation is also possible, for example a model with additive genetic (A), dominance (D), and residual environmental (E) effects. However, this ADE model predicts that MZ correlations are larger than twice the DZ correlations, which is not consistent with the reported OS and SS intraclass correlations in Table 2.3. For all traits, the estimate of the C component was greater than zero, further suggesting that an ACE model is more appropriate than an ADE model.

This study has wide implications for research into genetic variation of disease and non-disease related traits in human populations. Extremely large random samples, comprising 100,000s of individuals, have been collected or are being collected in a number of countries to answer fundamental questions in the fields of biomedical, educational and economics research. It has been shown that, with a minimum of required information (the most important of which are surname, date of birth, sex and a localized identifier such as school or household), a large number of twin pairs can be identified (see also Webbink *et al.*, 2006) and that appropriate statistical methods are available to estimate genetic parameters without knowing zygosity. Hence, a genetic element can be added to such studies, thereby greatly enhancing their value.

Appendix 2A. Variance and Kurtosis for a Mixture of Two Normal Distributions

For a trait x the i^{th} moment about the mean is defined as $m_i = E[x - E(x)]^i$. The variance and kurtosis are m_2 and $K = [\frac{m_4}{(m_2)^2} - 3]$, respectively. For a single distribution of a normally distributed trait, $m_2 = \sigma^2$, $m_4 = 3\sigma^4$ and $K = 0$. For a mixture of two distributions with mixture proportion p,

$$m_2 = p\sigma_1^2 + (1-p)\sigma_2^2 \tag{2.1}$$

and

$$m_4 = 3p\sigma_1^4 + 3(1-p)\sigma_2^4 \tag{2.2}$$

For example, consider the difference in observations between pairs of twins in a mixture of DZ and MZ twin pairs with mixture proportion 0.5 and an ACE model with heritability of 0.6 and the proportion of variance due to common environmental effects of 0.2. The phenotypic variance is unity. Then, $m_2 = 0.350, m_4 = 0.435$ and $K = 0.55$. The principle of the mixture distribution approach is that the two unknown variances are estimated from the observed variance and kurtosis.

Appendix 2B. Estimation of Parameters from ANOVA on OS and SS Pairs

Consider the ACE model and scaled phenotypes so that the phenotypic variance is unity. Let p be the proportion of MZ twins among same-sex (SS) twin pairs. The between (B) and within (W) mean square component in an ANOVA will then be a mixture from two distributions (MZ and DZ). The expected values are,

$$E(B_{SS}) = e^2 + 2c^2 + \frac{1}{2}h^2(3+p) = 1 + c^2 + \frac{1}{2}h^2(1+p) \tag{2.3}$$

$$E(W_{SS}) = e^2 + \frac{1}{2}(1-p)h^2 = 1 - c^2 - \frac{1}{2}h^2(1+p) \tag{2.4}$$

with h^2, c^2 and e^2 the proportion of phenotypic variance due to additive genetic (A), common environmental (C) and residual environmental (E) effects. The intra-class correlation (t) from the between and within pair analysis is,

$$t = \frac{\frac{[E(B)-E(W)]}{2}}{\frac{E[(B)-E(W)]}{2} + E(W)} = \frac{[E(B)-E(W)]}{[E(B)+E(W)]} \tag{2.5}$$

For the SS pairs,

$$t_{SS} = c^2 + \frac{1}{2}(1+p)h^2 \tag{2.6}$$

For the opposite-sex (OS) pairs, assuming an ACE model but allowing for a genetic correlation of less than unity between the sexes and different heritabilities for males (m) and females (f),

$$t_{OS} = c^2 + \frac{1}{2} r_g h_m h_f \tag{2.7}$$

If the heritability for males and females is the same then,

$$t_{OS} = c^2 + \frac{1}{2} r_g h^2 \tag{2.8}$$

Hence, under the assumption of equal heritabilities of males and females, there are two summary statistics (correlations), i.e. t_{SS} and t_{OS}, but three unknowns (c^2, h^2, r_g). It follows that,

$$2(t_{SS} - t_{OS}) = h^2[(1+p) - r_g] \tag{2.9}$$

and

$$\frac{[t_{OS}(1+p) - t_{SS} r_g]}{[(1+p) - r_g]} = c^2 \tag{2.10}$$

If one further assumes that $r_g = 1$, then the estimates of h^2 and c^2 satisfy

$$h^2 = \frac{2(t_{SS} - t_{OS})}{p} \tag{2.11}$$

and

$$c^2 = \frac{[t_{OS}(1+p) - t_{SS}]}{p} \tag{2.12}$$

Relative to the standard twin design with MZ and DZ pairs, the sampling variance of the estimate of the heritability from SS and OS pairs is increased by a factor of p^{-2}, for example by a factor of four if $p = \frac{1}{2}$.

3 Precision and Bias of a Mixture Distribution Model to Estimate Variance Components from Twins of Unknown Zygosity: Simulations and Application to IQ Measures from The U.K. Twins' Early Developments Study

3.1 Abstract

The classification of twin pairs based on zygosity into monozygotic (MZ) or dizygotic (DZ) twins is the basis of most twin analyses. When zygosity information is unavailable, a normal finite mixture distribution (mixture distribution) model can be used to estimate components of variation for continuous traits. The main assumption of this model is that the observed phenotypes on a twin pair are bivariately normally distributed. Any deviation from normality, in particular kurtosis, could produce biased estimates. Using computer simulations and analyses of a wide range of phenotypes from the U.K. Twins' Early Developments Study (TEDS), where zygosity is known, properties of the mixture distribution model were assessed. Simulation results showed that, if normality assumptions were satisfied and the sample size was large (e.g. 2,000 pairs), then the variance component estimates from the mixture distribution model were unbiased and the standard deviation of the difference between heritability estimates from known and unknown zygosity in the range of 0.02 to 0.20. Unexpectedly, the estimates of heritability of 10 variables from TEDS using the mixture distribution model were consistently larger than those from the conventional (known zygosity) model. This discrepancy was due to violation of the bivariate normality assumption. A leptokurtic distribution of pair difference was observed for all traits (except non verbal ability scores of MZ twins), even when the univariate distribution of the trait was close to normality. From an independent sample of Australian twins, the heritability estimates

for IQ variables were also larger for the mixture distribution model in 6 out of 8 traits, consistent with the observed kurtosis of pair difference. While the known zygosity model is quite robust to the violation of the bivariate normality assumption, this novel finding of widespread kurtosis of the pair difference may suggest that this assumption for analysis of quantitative trait in twin studies may be incorrect and needs revisiting. A possible explanation of widespread kurtosis within zygosity groups is heterogeneity of variance, which could be caused by genetic or environmental factors. For the mixture distribution model, violation of the bivariate normality assumption will produce biased estimates.

3.2 Introduction

The classical twin design is very useful in partitioning the observed phenotypic variance of complex traits in humans into genetic and environmental components (reviewed by Boomsma *et al.*, 2002). By comparing the resemblance of monozygotic (MZ) twin pairs to that of dizygotic (DZ) twin pairs, twin studies allow the causes of individual differences in complex traits to be quantified. Under the assumption that both types of twins share the same degree of common environmental experiences (the common environment assumption), a larger similarity of MZ pairs compared to DZ pairs indicates that genetic factors influence phenotypic variation (e.g. Evans *et al.*, 2002; Rijsdijk and Sham, 2002). The classification of twins based on zygosity is crucial in twin studies. A standard zygosity questionnaire (e.g. Peeters *et al.*, 1998) answered by twins or their parents is usually used to diagnose zygosity. With the advance of molecular genetic markers, such as microsatellites, DNA-based zygosity testing is now widely used and gives a greater accuracy (e.g. Forget-Dubois *et al.*, 2003).

Although zygosity information can now be easily and economically obtained, such information is not always available. Two examples are twin data that were collected before zygosity classification was routine [(e.g. the Scottish Mental Surveys 1932 (Scottish Council for Research in Education, 1933) and 1947 (Deary *et al.*, 2004; Scottish Council for Research in Education, 1949)] and data collected from large national studies in the fields of social sciences, economics or education where genetic study was not the main interest (e.g. Scarr-Salapatek, 1971). Twins from these studies can be identified by matching a pair with, for example, the same surname, birth date and location such as home address or school, if these identifiers are available. Assuming that an identifier of sex is also available, twin pairs from such studies can only be classified as same sex (SS) or opposite sex (OS) pairs. SS pairs are a mixture of MZ and DZ pairs whereas OS pairs are always DZ. For such studies, the conventional methods which rely on zygosity information cannot be used. Different methods have been proposed to analyse twin data where zygosity information is unavailable. Scarr-Salapatek estimated the correlations of MZ and DZ pairs by partitioning the z-transformed correlation coefficient of SS twins [An analogous method based upon ANOVA is described in Chapter 2 (Appendix 2B) and Benyamin *et al.* (2005)]. The method, however, assumed that the sample size and correlation of DZ SS twins were the same as those of the observed OS pairs, and is limited to univariate heritability (Neale, 2003). The OS correlation can substantially differ from the correlation of DZ SS, for example if the genetic or common environmental covariance is lower in OS pairs.

Neale proposed a method based upon a normal finite mixture distribution (mixture distribution) to estimate MZ and DZ correlations from SS twins. This method partitions the SS twin distribution into underlying MZ and DZ distributions by maximum likelihood. The estimated proportion of MZ among

SS twins (pMZ) is used to weight the likelihood. This method has been applied to analyse individual differences in cognitive ability (the Moray House Test No. 12) from twin data with unknown zygosity of the Scottish Mental Surveys 1932 and 1947 (Chapter 2; Benyamin *et al.*, 2005). In addition, Heath *et al.* (2003) proposed a latent class analysis to diagnose zygosity. This method can be used to analyse discrete data on twins by fitting a 2-class latent class model, which is assumed to correspond to MZ and DZ pairs (Chapter 2; Benyamin *et al.*, 2005).

The mixture distribution model of Neale assumes that the observed phenotypes on a pair follow a bivariate normal distribution in the population. Any deviation from normality, in particular kurtosis, could produce biased estimates because the partitioning of the observed within-pair and between-pair variation is based upon the contrast of the variance and kurtosis (Chapter 2; Benyamin *et al.*, 2005).

The purpose of the present study is to quantify the precision and bias of the mixture distribution model in estimating genetic parameters from twin data when zygosity is unknown. Simulation was used to quantify the precision of estimation of the mixture distribution model when the distributional assumptions were met, and to quantify bias when normality assumptions were violated. Finally, the known zygosity and mixture distribution models were applied to a range of IQ phenotypes from the U.K. Twins' Early Development Study (TEDS), a longitudinal study of a representative sample of all twins born in England and Wales between 1994 and 1996. Zygosity information is available on TEDS data. Therefore, the application of the mixture distribution model to these data afforded a check on variance components estimates from the previous application of the mixture distribution model on twins of unknown zygosity of cognitive ability from the Scottish Mental Surveys 1932 and 1947 (Chapter 2).

Table 3.1: Scenarios of simulated variance component proportions.

Variance Components	I	II	III	IV	V	VI	VII	VIII	IX
a^2	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
c^2	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05
e^2	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05

3.3 Methods

3.3.1 Simulation Study

MZ and DZ twin data were simulated using a standard ACE model of family resemblance, by sampling additive genetic (A), common environmental (C) and specific environmental (E) effects. No sex effects or other fixed effects were simulated, and for subsequent analyses it was assumed that there were only SS pairs. All simulations were replicated 1,000 times. To assess the precision of estimation of the mixture distribution model, twin data were first simulated under the assumed bivariate normal distribution. Nine different standardized variance component parameters (Table 3.1) were simulated for different sample sizes (500, 2,000 and 5,000 twin pairs, with equal proportions of MZ and DZ). Each simulated dataset was then analysed with the conventional (known zygosity) and mixture distribution (Neale, 2003) models, using the statistical package Mx (Neale *et al.*, 2002). An overall mean was the only fixed effect fitted in the model.

Although pMZ could in principle be estimated from the data when fitting a mixture distribution model, the estimate is very imprecise (result not shown). Therefore, in the mixture distribution model, an a priori estimate of pMZ is used to weight the likelihood. In a complete population survey, pMZ can be estimated using Weinberg's differential rule as 1 − 2×(proportion of OS twin pairs) (Weinberg, 1902). This formula assumes that the number of DZ SS twins is the same as DZ OS twins due to the distribution of sexes (Scarr-Salapatek, 1971). This proportion may not be accurately estimated in all studies. Therefore, in order to assess whether specifying a wrong proportion in the mixture distribution model has an effect on variance components estimation, different proportions (0.1-0.9) were used in the analyses when the true proportion was 0.5. For this simulation, the standardized A, C and E variance components were $a^2 = 0.50$, $c^2 = 0.25$, $e^2 = 0.25$ and simulations were based on 2,000 twin pairs with an equal proportion of MZ and DZ.

In order to assess the effect of kurtosis on the parameters' estimation of the mixture distribution model, normally distributed twin data were transformed into a distribution with a desired kurtosis value using the Cornish-Fisher expansion (Cornish and Fisher, 1937). For each value of an individual (x) drawn for a normal distribution, the transformation is:

$$y = x + \frac{c}{24}(x^3 - 3x) \tag{3.1}$$

where y is the transformed x with desired kurtosis given by the coefficient c. For positive c smaller than 1, the simulated data has a distribution with the kurtosis value similar to c. For larger positive c, the kurtosis value for the transformed distribution was larger than c. On the other hand, for negative c, the kurtosis value for the transformed distribution was slightly smaller than c. For examples, the corresponding average kurtosis values for c of -2, -1, -0.75, -0.50, -0.25, 0.25, 0.50, 0.75, 1 and 2 were -1.01, -0.70, -0.57, -0.42, -0.23, 0.27, 0.60, 0.99, 1.50 and 4.15, respectively. The exact relationship between the value of c and the kurtosis value of the transformed data is shown in Appendix 1. Data sets for different c (11 different values of c ranging from -2 to 2) were simulated. The standardized variance components were $a^2 = 0.50$, $c^2 = 0.25$, $e^2 = 0.25$ and simulations were based on 2,000 twin pairs with an equal proportion of MZ and DZ pairs. Data were then analysed using the conventional and mixture distributions models as before.

Parameter estimates obtained from all simulations were further analysed using the statistical package R (R Development Core Team, 2006).

3.3.2 Data Application

Variables and Zygosity Diagnosis:

TEDS is a large scale longitudinal study on language and cognitive developments involving a representative sample of all twins born in England and Wales in 1994 - 1996 (e.g. Trouton *et al.*, 2002). In the present study, eight variables related to language and cognitive developments of 7-year-old twins were available for analysis. These traits included scores on conceptual grouping, picture completion, similarities, vocabulary and test of word recognition (TOWRE). The composites of a number of the variables, i.e. language (a composite of similarities and vocabulary), non-verbal IQ (a composite of scores for conceptual grouping and picture completion) and general cognitive ability (g), which is the composite of language and non-verbal IQ, were also analysed. The cognitive abilities were measured on each child individually and separately using a telephone interview (Petrill *et al.*, 2002). The complete description and definitions of the IQ variables were presented previously (e.g. Harlaar *et al.*, 2005; Kovas *et al.*, 2005; Price *et al.*, 2000a; Spinath *et al.*, 2004). In addition to these variables, height and weight were included in the analyses and these variables were supplied by a parent/guardian of the twins, usually the mother. Parental ratings were used to ascertain the zygosity of SS twin pairs (Kovas *et al.*, 2005). This method has an error rate less than 5%, as validated by DNA typing using a multiplexed set of highly polymorphic markers (Harlaar *et al.*, 2005; Kovas *et al.*, 2005; Price *et al.*, 2000b).

Samples and Exclusions:

Individuals and their co-twin were excluded from the analysis if: (i) their data base entry had missing identifiers (for sex and zygosity), (ii) there were specific medical and genetic conditions recorded (as described Kovas *et al.*, 2005), (iii) they were of non-white ethnicity, (iv) English is not the language at home (v) either twin had an extreme phenotype (more than 3 standard deviations from the mean for any variable) and (vi) they were of the opposite sex. The reason for excluding the opposite sex twins from the analyses was to avoid possible (large) biases due to sex-limitation effects. If opposite sex pairs were included in the analysis then the parameter estimates for the DZ SS twins will be centered on the DZ OS twins because the OS intraclass correlation is estimated with more precision than the variance components from the mixture distribution (see Appendix 2B in Chapter 2 for more explanation).

The final data set comprised 3,582 SS twins, 1,904 MZ and 1,678 DZ pairs. The proportion of MZ pairs among all twin pairs in the selected dataset is larger than that of the excluded dataset (0.372 ± 0.007 compared to 0.313 ± 0.009). This difference could be due to a larger participation rate of MZ twins in the cognitive study. However, the proportion of MZ pairs among all pairs in the TEDS twin data is not significantly different from the whole twin population born in England and Wales between 1994 to 1996 (Imaizumi, 2003) [0.351 ± 0.005 (TEDS) vs 0.343 ± 0.003 (population)].

Analysis:

Descriptive statistics of the standardized residuals (after a general linear model correction for sex and age effects on all observations) of the IQ variables, height and weight were obtained using SPSS 12.0.2 for Windows (SPSS Inc., 1989 - 2003). The standardized residuals were then split into MZ and DZ groups. Pearson correlations for MZ and DZ boys and girls were computed after adjustment for age effects. To test for normality, a Kolmogorov-Smirnov normality test (implemented in SPSS) was performed for all phenotypes, after adjustment for sex and age effects. All phenotypes were analysed using the known zygosity and mixture distribution models. In the mixture distribution model, the observed proportion of MZ among SS twins (pMZ = 0.53) was used to weight the analyses. For all analyses, sex and age were fitted as fixed effects.

3.4 Results

3.4.1 Simulation Study

Mixture Distribution Model Under Normality:

For normally distributed twin data, heritability (a^2) estimates from the mixture distribution model were compared with that from the known zygosity model. Figure 3.1 shows the relationship between the estimates from the two models, for the range of heritabilities of 0.1 to 0.9 (other parameters as in Table 3.1), for samples of 2,000 twin pairs. For all sets of parameters, the mean estimate of the heritability was very similar for both models, i.e. there was no evidence of a bias in the estimate of heritability, unless the heritability was small ($a^2 \leq 0.4$). The results showed that the higher the heritability simulated, the more similar the estimates between the two models. For heritabilities ≤ 0.4, although the mean estimates between the two models were similar, the standard deviation of estimates from the mixture distribution was about three times that of the known zygosity model. A similar pattern was also observed for the standardized common environmental variance (c^2), i.e. the larger the heritability simulated, the more similar the c^2 estimates between the two models (Figure 3.2).

When the estimate of the heritability is unbiased, a useful criterion for precision of estimation of the mixture approach is the standard deviation of the difference in the estimate of the heritability between the two models. Figure 3.3 shows this standard deviation for a range of sample sizes from 500 to 5,000 pairs, for the range of population parameters as given in Table 3.1. As expected, the larger the sample size, the smaller the standard deviations of the difference between the two estimates. For a sample size as large as 5,000 pairs, the maximum standard deviation of the difference was 0.15 (for a low heritability) and for a large heritability little information is lost by not knowing zygosity. However, when the sample size is quite small (e.g. 500 pairs), the standard deviations of the difference between the two estimates of heritability were quite large even for heritability as high as 0.6.

The effect on bias in the estimates of variance components when specifying a wrong pMZ in the mixture distribution model is presented in Figure 3.4, when $a^2 = 0.50$, $c^2 = 0.25$ and $e^2 = 0.25$. The magnitude and direction of the effects on each parameter estimate were different. The effect on heritability and error variance estimates were small. For example, when the actual pMZ is 0.5 and it is specified as 0.6 or 0.4, the mean bias of heritability estimates from the mixture distribution was less than 2%. The effects were slightly larger on the common environmental variance estimates, especially when pMZ was underestimated.

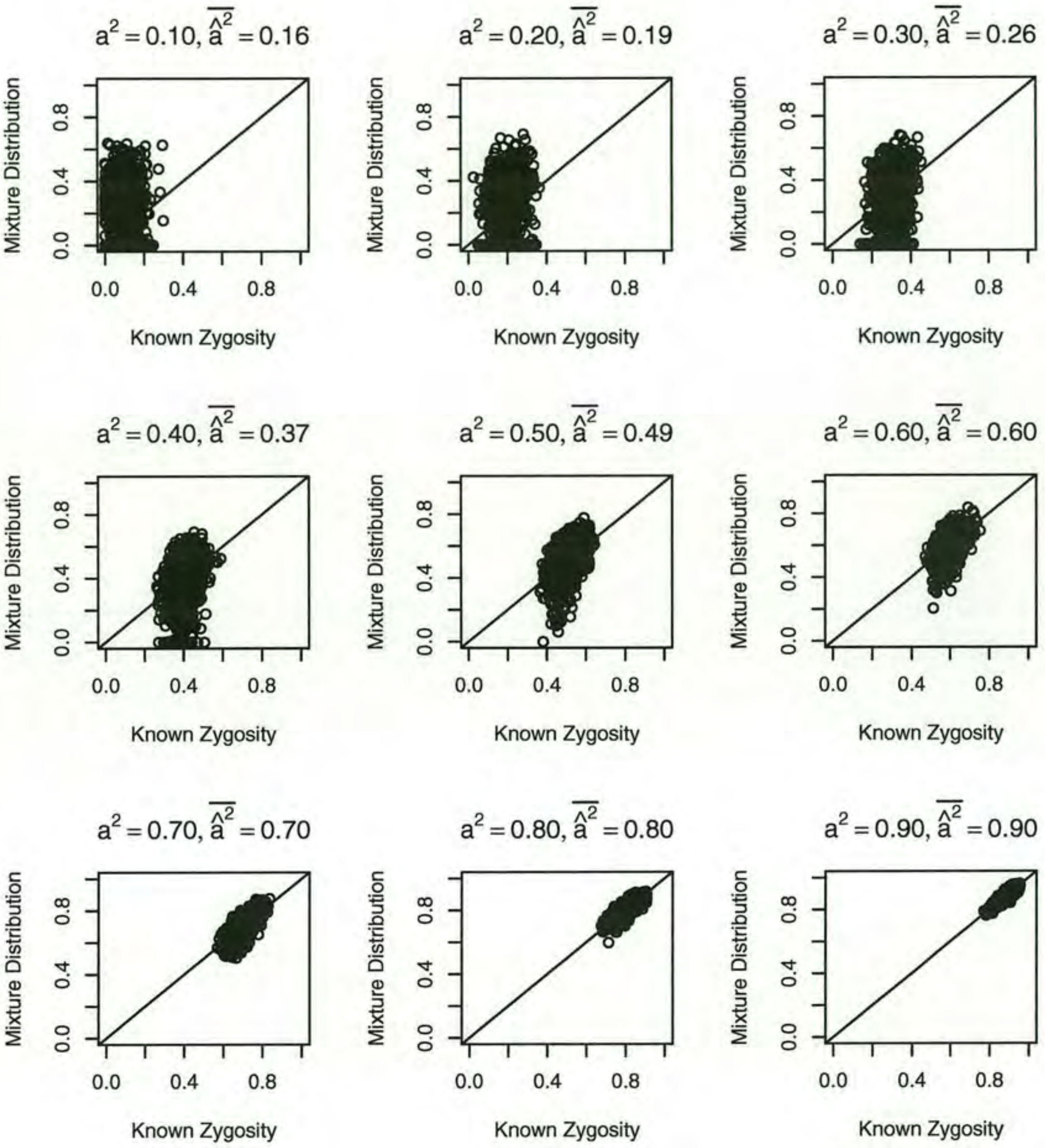

Figure 3.1: Heritability estimates from the mixture distribution and known zygosity models for different values of simulated heritability (Table 3.1) under an ACE model. The results were based on 2,000 pairs (1,000 MZs and 1,000 DZs) and 1,000 replicates. $\overline{\hat{a}^2}$ is the mean heritability estimate from the mixture distribution model.

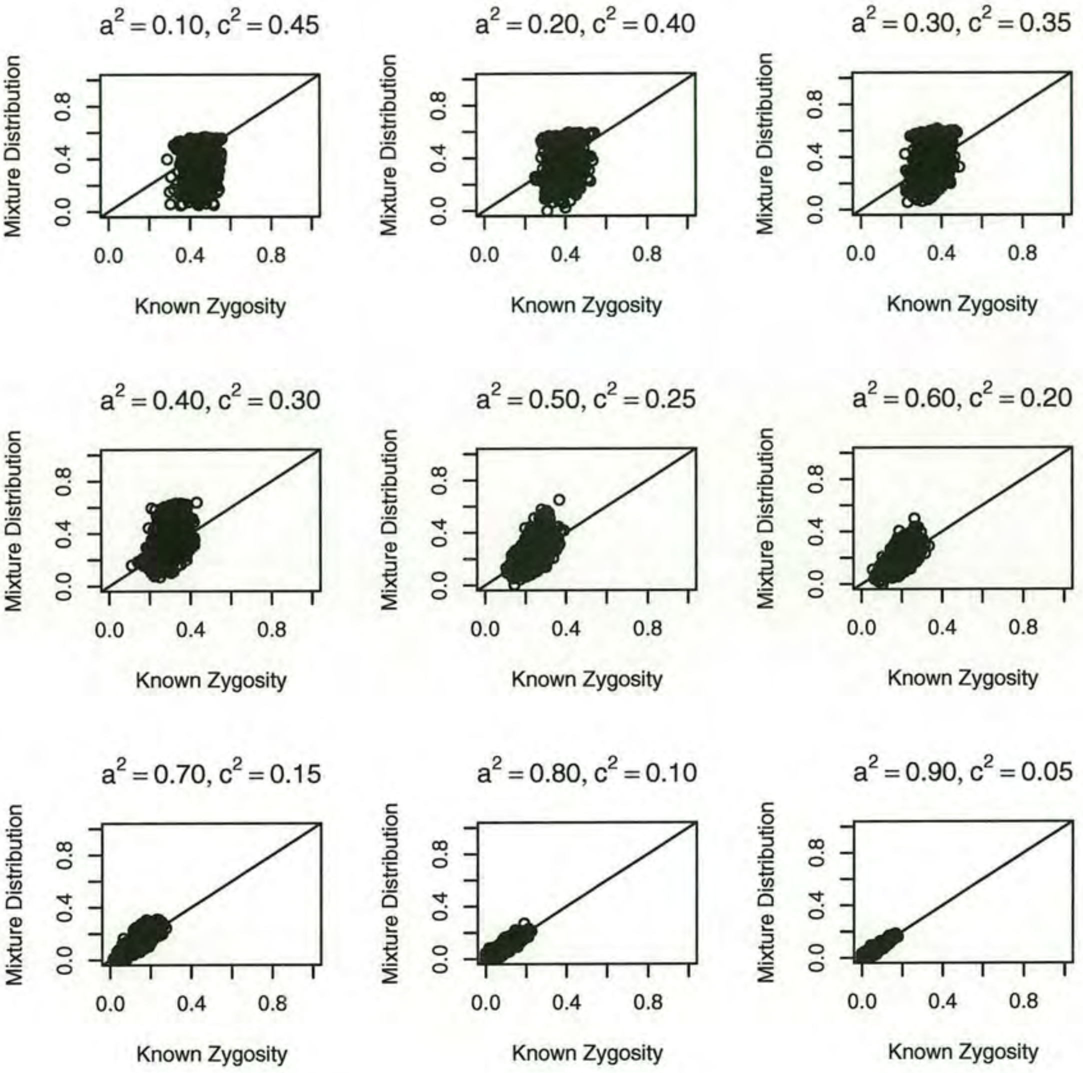

Figure 3.2: Standardized common environmental variance (c^2) estimates from the mixture distribution and known zygosity models for different values of standardized variance components (Table 3.1) under an ACE model. The results were based on 2,000 twin pairs (1,000 MZs and 1,000 DZs) and 1,000 replicates.

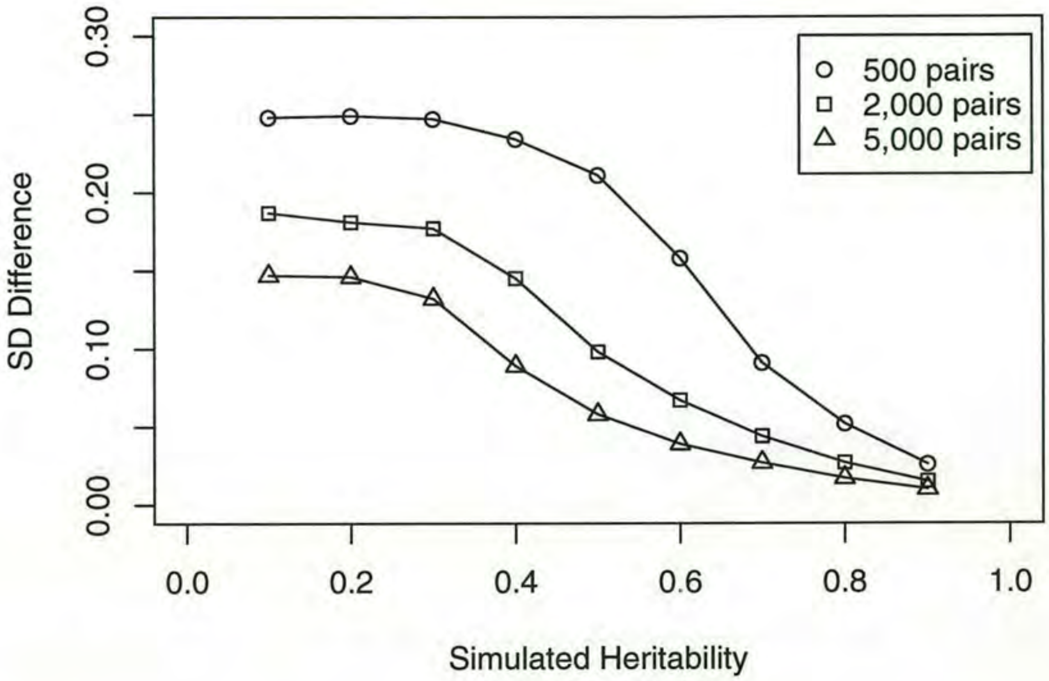

Figure 3.3: Standard deviation of the difference in heritability estimates between the mixture distribution and known zygosity models for different sample sizes.

Mixture Distribution Model When Twin Data is Kurtotic:

For twin data with a kurtotic distribution of the phenotype, estimates from the known zygosity model were unbiased (results not shown). Figure 3.5 presents the mean difference of parameter estimates between the mixture distribution and known zygosity models for platykurtic (negative kurtosis) and leptokurtic (positive kurtosis) distributions. As indicated from Figure 3.5, the mixture distribution model resulted in larger heritability and smaller c^2 and e^2 estimates on simulated data with a kurtotic distribution compared to the known zygosity model. However, for smaller kurtosis values ($-0.5 < k < 1$), the mean parameter estimates from the mixture distribution model did not differ substantially from that of the conventional analysis.

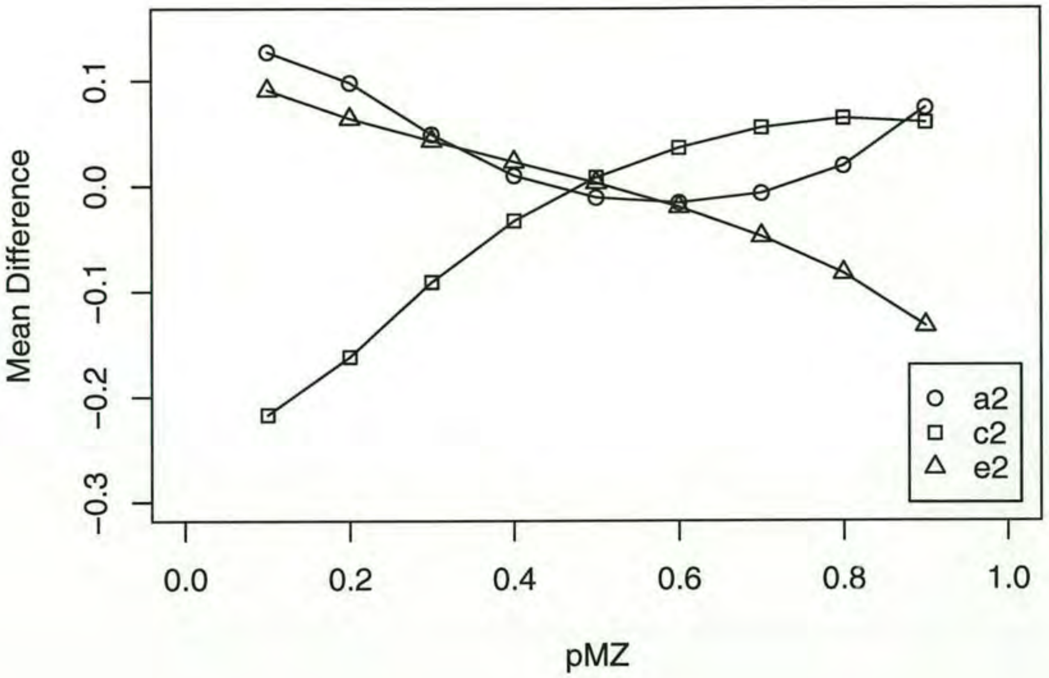

Figure 3.4: Mean difference of standardized variance component estimates between the mixture distribution and known zygosity models when the incorrect mixture proportion was assumed (true proportion = 0.5). The simulated parameters are: $a^2 = 0.50$, $c^2 = 0.25$, $e^2 = 0.25$ and simulations are based on 2,000 twin pairs (1,000 MZs and 1,000 DZs) and 1,000 replicates.

3.4.2 Analyses of TEDS Data

Descriptive Statistics and Phenotypic Distribution:

Descriptive statistics of the data after exclusions are presented in Table 3.2. Between 2,279 and 2,545 pairs for which both twins had a phenotype on any variable from a total of 3,582 pairs were available for analysis. The main reason for a considerable missing data is that not all twins were tested/interviewed at age 7. The age of the twins when the parents' booklet was returned, which was

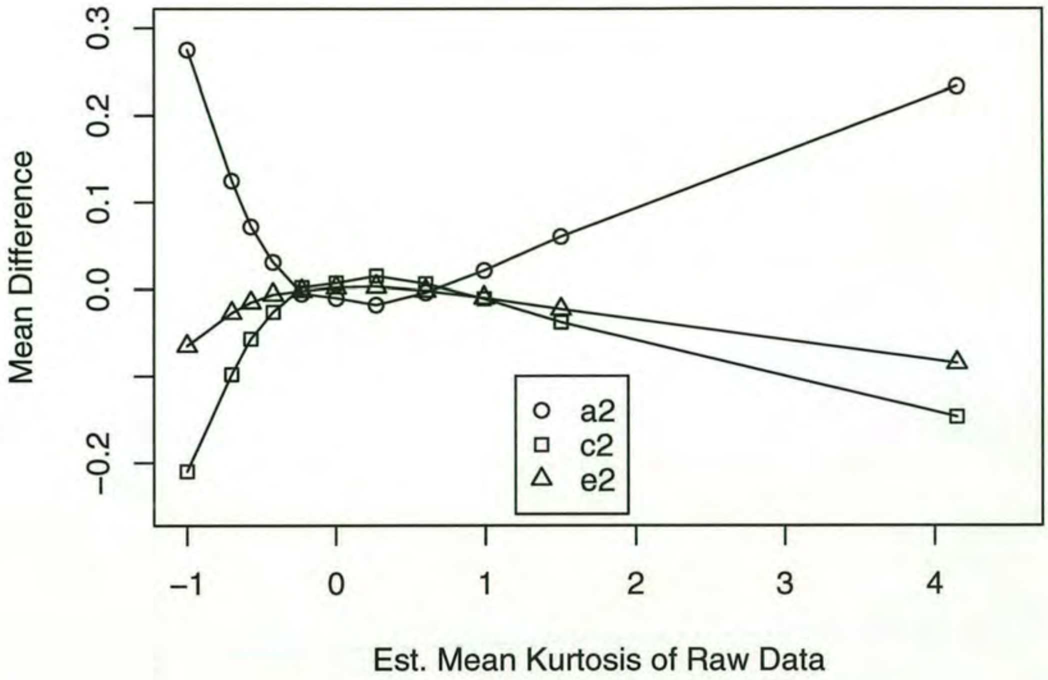

Figure 3.5: Mean differences of standardized variance component estimates between the mixture distribution and known zygosity models for a given estimated mean kurtosis. The simulated parameters are: $a^2 = 0.50$, $c^2 = 0.25$, $e^2 = 0.25$ and simulations are based on 2,000 twin pairs (1,000 MZs and 1,000 DZs) and 1,000 replicates.

used as a covariate in the genetic analysis, had a mean and SD of 7.05 and 0.25, respectively. Although the distribution of the phenotypes appeared normal, the Kolmogorov-Smirnov normality test showed that the trait distribution was significantly different from normality for all traits [except language (MZ, DZ) and g (DZ)]. The skewness and kurtosis values ranged from -0.38 to 0.48 and -0.75 to 1.12, respectively. Note that, with these kurtosis values, the simulations showed that the bias in the estimate of heritability of the mixture distribution model was less than 0.1 for a heritability of 0.5.

Table 3.2: Descriptive statistics of standardized residuals of TEDS variables after adjustment for sex and age effects.

Variables	Zygosity	N[a]	Mean (SD)	Skewness (SE)	Kurtosis (SE)	Skewness of pair difference (SE)	Kurtosis of pair difference (SE)
Age	MZ	1823	7.05 (0.25)				
	DZ	1580	7.06 (0.25)				
Weight (kg)	MZ	1360	−0.04 (0.99)	0.45 (0.05)	0.26 (0.09)	0.21 (0.07)	3.45 (0.13)
	DZ	1185	0.05 (1.01)	0.48 (0.05)	0.35 (0.10)	0.03 (0.07)	0.74 (0.14)
Height (cm)	MZ	1337	−0.02 (0.96)	0.00 (0.05)	0.59 (0.09)	−0.04 (0.07)	6.79 (0.13)
	DZ	1162	0.03 (1.04)	−0.24 (0.05)	1.12 (0.10)	0.17 (0.07)	1.54 (0.14)
TOWRE	MZ	1255	−0.02 (1.02)	0.10 (0.05)	−0.75 (0.10)	0.23 (0.07)	1.20 (0.14)
	DZ	1133	0.03 (0.97)	0.05 (0.05)	−0.65 (0.10)	0.01 (0.07)	0.26 (0.15)
Conceptual	MZ	1290	−0.01 (1.00)	−0.27 (0.05)	−0.67 (0.10)	0.06 (0.07)	0.16 (0.14)
grouping	DZ	1155	0.01 (1.00)	−0.34 (0.05)	−0.61 (0.10)	0.02 (0.07)	0.20 (0.15)
Similarities	MZ	1281	−0.03 (1.01)	0.02 (0.05)	0.19 (0.10)	−0.12 (0.07)	0.65 (0.14)
	DZ	1145	0.04 (0.98)	0.01 (0.05)	0.37 (0.10)	−0.19 (0.07)	1.12 (0.15)

Note: The Kolmogorov-Smirnov normality test showed that for most variables, both the trait distribution (except language (MZ, DZ) and g (DZ)) and the distribution of the pair difference (except language (MZ), and g (DZ), non verbal (DZ), TOWRE (DZ)) showed significant deviation from normality. [a] Number of pairs for which both twins had a phenotype.

Table 3.2: Continued.

Variables	Zygosity	N^a	Mean (SD)	Skewness (SE)	Kurtosis (SE)	Skewness of pair difference (SE)	Kurtosis of pair difference (SE)
Vocabulary	MZ	1284	−0.02 (1.00)	−0.01 (0.05)	−0.15 (0.10)	0.10 (0.07)	0.61 (0.14)
	DZ	1149	0.02 (1.00)	0.02 (0.05)	−0.09 (0.10)	−0.11 (0.07)	0.47 (0.15)
Picture	MZ	1291	−0.04 (0.98)	−0.29 (0.05)	−0.07 (0.10)	0.00 (0.07)	0.44 (0.14)
completion	DZ	1153	0.05 (1.02)	−0.38 (0.05)	0.06 (0.10)	0.11 (0.07)	0.71 (0.15)
g	MZ	1270	−0.04 (1.00)	−0.10 (0.05)	−0.19 (0.10)	−0.01 (0.07)	0.30 (0.14)
	DZ	1137	0.04 (1.00)	−0.13 (0.05)	−0.16 (0.10)	−0.04 (0.07)	0.70 (0.15)
Language	MZ	1274	−0.03 (1.01)	0.01 (0.05)	−0.20 (0.10)	0.07 (0.07)	0.59 (0.14)
	DZ	1141	0.03 (0.99)	0.06 (0.05)	−0.09 (0.10)	−0.11 (0.07)	0.58 (0.15)
Non verbal	MZ	1128	−0.03 (0.99)	−0.20 (0.05)	−0.31 (0.10)	−0.01 (0.07)	−0.10 (0.14)
	DZ	1151	0.04 (1.01)	−0.28 (0.05)	−0.24 (0.10)	0.06 (0.07)	0.33 (0.15)

Note: The Kolmogorov-Smirnov normality test showed that for most variables, both the trait distribution (except language (MZ, DZ) and g (DZ)) and the distribution of the pair difference (except language (MZ), and g (DZ), non verbal (DZ), TOWRE (DZ)) showed significant deviation from normality. ^{a} Number of pairs for which both twins had a phenotype.

Table 3.3: Twin correlations and their standard errors after adjustment for age.

Variables	MZ - Boy	MZ - Girl	DZ - Boy	DZ - Girl
Weight	0.84 (0.02)	0.85 (0.02)	0.47 (0.03)	0.52 (0.03)
Height	0.92 (0.02)	0.94 (0.02)	0.56 (0.03)	0.64 (0.03)
TOWRE	0.85 (0.02)	0.84 (0.02)	0.51 (0.03)	0.50 (0.04)
Conceptual Grouping	0.38 (0.04)	0.32 (0.04)	0.24 (0.04)	0.29 (0.04)
Similarities	0.51 (0.04)	0.43 (0.04)	0.37 (0.04)	0.35 (0.04)
Vocabulary	0.63 (0.03)	0.57 (0.04)	0.49 (0.03)	0.47 (0.04)
Picture Completion	0.47 (0.04)	0.48 (0.04)	0.39 (0.04)	0.40 (0.04)
g	0.68 (0.03)	0.61 (0.03)	0.49 (0.03)	0.48 (0.04)
Language	0.67 (0.03)	0.61 (0.03)	0.50 (0.03)	0.49 (0.04)
Non Verbal	0.45 (0.04)	0.45 (0.04)	0.39 (0.04)	0.38 (0.04)

Twin Correlations:

Twin correlations for all phenotypes, after adjustment for age, are presented in Table 3.3. The MZ and DZ twin correlations were similar across sexes. MZ correlations were consistently higher than DZ correlations. Results in Table 3.3 indicate strongly that genetic factors play a significant role in explaining phenotypic variance in most of the traits.

Variance Component Estimation:

Initially, the variance component estimation using the known zygosity model was performed with separate variance components for boys and girls. However, for most variables there was no significant difference between variance component estimates in boys and girls, except for weight, TOWRE, similarities and picture completion (results not shown). The pooled (boys and girls) estimates from the known zygosity and mixture distribution models are presented in Table 3.4.

Known Zygosity With the exception of TOWRE score for which a large heritability was estimated (about 0.6), the heritability estimates of other IQ phenotypes were small to moderate, ranging from 0.16 to 0.33. Shared environmental variance accounted for 19 to 37% of the phenotypic variance of IQ variables. Thus, most of the phenotypic variation in IQ related variables (except for TOWRE) was specific to individuals.

Genetic factors constituted a large proportion of the phenotypic variation in weight and height. About 70 % of the total phenotypic variance in weight and height were attributed to genetic factors. These findings are similar to previous studies on heritabilities of weight (reviewed by Pietilainen *et al.*, 2002) and height (reviewed by Silventoinen, 2003a).

Mixture Distribution The heritability estimates from the mixture distribution were consistently larger than those from the conventional model for all variables (Table 3.4). The confidence intervals of the estimates of heritability from the two models did not overlap for most traits. In addition, with the exception of height, the estimate of common environmental variance was zero (or close to zero) for all variables. For IQ phenotypes (except TOWRE), the mean difference of heritability estimates from the mixture distribution compared to the known zygosity model was 0.40. The difference in the average estimate of common environment variance was 0.29. However, the sum of the proportion of variance due to additive genetic and common environmental effects [giving the repeatability, the proportion of phenotypic variance of single measurements due to the effects of genetic and permanent environmental factors (Falconer and

Table 3.4: Standardized variance component estimates from the known zygosity and mixture distribution models (boys and girls are pooled).

Variables	**Model**	a^2 **(95%CI)**	c^2 **(95%CI)**	e^2 **(95%CI)**
Weight	Known	0.71 (0.63 - 0.79)	0.14 (0.06 - 0.22)	0.15 (0.14 - 0.17)
	Mixture	0.96 (0.93 - 0.97)	0.00 (0.00 - 0.03)	0.04 (0.03 - 0.05)
Height	Known	0.69 (0.63 - 0.76)	0.24 (0.17 - 0.31)	0.07 (0.06 - 0.07)
	Mixture	0.82 (0.75 - 0.91)	0.14 (0.05 - 0.22)	0.04 (0.03 - 0.05)
TOWRE	Known	0.63 (0.55 - 0.71)	0.21 (0.12 - 0.29)	0.16 (0.15 - 0.18)
	Mixture	0.88 (0.77 - 0.91)	0.01 (0.00 - 0.11)	0.11 (0.09 - 0.13)
Conceptual-	Known	0.16 (0.02 - 0.30)	0.19 (0.07 - 0.30)	0.65 (0.61 - 0.70)
Grouping	Mixture	0.41 (0.13 - 0.46)	0.00 (0.00 - 0.21)	0.59 (0.54 - 0.66)
Similarities	Known	0.19 (0.06 - 0.31)	0.27 (0.18 - 0.39)	0.54 (0.50 - 0.58)
	Mixture	0.59 (0.50 - 0.63)	0.00 (0.00 - 0.06)	0.41 (0.37 - 0.46)
Vocabulary	Known	0.26 (0.16 - 0.36)	0.35 (0.25 - 0.43)	0.40 (0.37 - 0.43)
	Mixture	0.72 (0.55 - 0.75)	0.00 (0.00 - 0.14)	0.28 (0.25 - 0.33)
Picture-	Known	0.20 (0.08 - 0.32)	0.29 (0.19 - 0.39)	0.51 (0.47 - 0.56)
Completion	Mixture	0.61 (0.54 - 0.65)	0.00 (0.00 - 0.04)	0.39 (0.35 - 0.44)
g	Known	0.33 (0.23 - 0.43)	0.32 (0.23 - 0.41)	0.35 (0.32 - 0.38)
	Mixture	0.74 (0.56 - 0.78)	0.01 (0.00 - 0.16)	0.26 (0.22 - 0.30)
Language	Known	0.26 (0.16 - 0.36)	0.37 (0.28 - 0.46)	0.37 (0.34 - 0.40)
	Mixture	0.71 (0.53 - 0.78)	0.04 (0.00 - 0.18)	0.26 (0.22 - 0.30)
Non Verbal	Known	0.16 (0.04 - 0.29)	0.30 (0.20 - 0.40)	0.54 (0.50 - 0.58)
	Mixture	0.55 (0.27 - 0.60)	0.00 (0.00 - 0.22)	0.45 (0.40 - 0.52)

Note: Known and mixture are the known zygosity and mixture distribution models, respectively.

Mackay, 1996)] was similar between the two models (Figure 3.6).

3.5 Discussion

Neale (2003) has shown that a mixture distribution model can be used to analyse twin data when zygosity information is incomplete or unavailable with little bias. His simulations were from (bivariate) normal distributions, an assumption of the mixture distribution model. In addition, Neale (2003) only simulated a single set of standardized variance components (i.e. $a^2 = 0.6$, $c^2 = 0.2$, $e^2 = 0.2$) for twins without zygosity information. The present study further explored the properties of this model for a wider range of parameters and different scenarios. It includes assessing the mixture distribution model when the data is not normally distributed by simulating platykurtic and leptokurtic distributions. Different (incorrect) proportions of MZ among twins were also used to weight the analysis in the mixture distribution for a given true proportion, to assess the bias introduced by misspecification of this parameter. The simulation results suggested that, if the normality assumption was satisfied and the sample size was large, then the variance component estimates from the mixture distribution are unbiased and accurate for analysing twin data where zygosity information is unavailable. However, if the heritability is small ($a^2 < 0.4$), then the estimates are imprecise.

If the distribution of the phenotypes is kurtotic then the mixture distribution produced biased estimates. However, this bias was small for kurtosis values in the range of -0.5 and 1. Specifying a wrong mixture proportion in the analysis had small impact, in terms of bias, on the estimates of variance components, unless the difference between the true and estimated proportion was very large

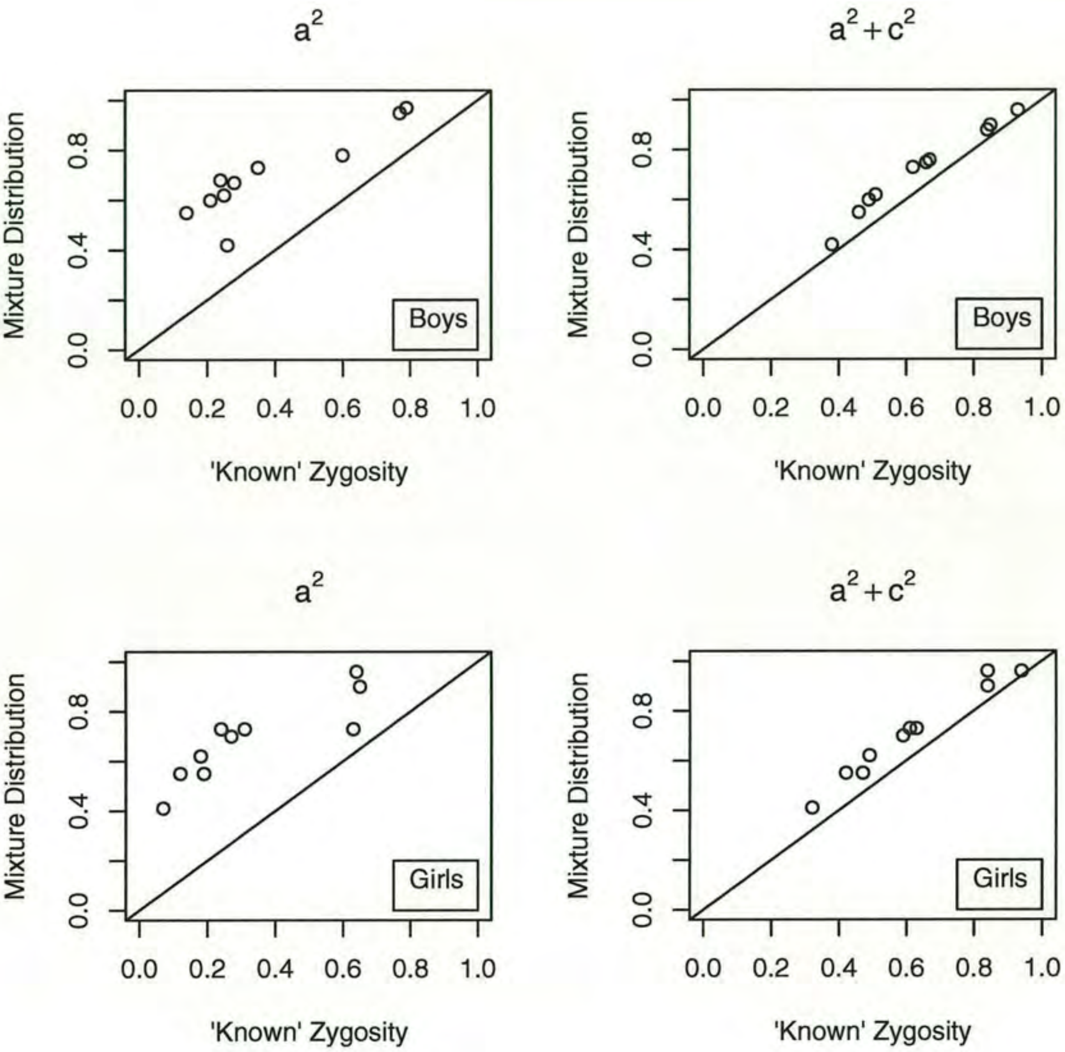

Figure 3.6: Heritability (a^2) and repeatability (defined as the sum of a^2 + c^2) estimates of nine IQ phenotypes, height and weight of TEDS data from the known zygosity and mixture distribution models for boys and girls. Sex and age were fitted as covariate.

(e.g. 0.5 and 0.2). For a population survey, the estimated proportion of MZ among SS twins can usually be estimated accurately. For the TEDS data, the estimated proportion of MZ twins among SS twins using Weinberg's differential rule (Weinberg, 1902) was very similar to the observed proportion, 0.523 and 0.519, respectively.

3.5.1 Post-hoc Analyses

The analyses of IQ phenotypes, height and weight from the TEDS data showed that the estimate of heritabilities from the mixture distribution were consistently larger than those from the conventional model. These results are inconsistent with those from all of the simulations and demand an explanation. For the observed kurtosis values of the TEDS variables (in the range of -0.75 to 1.12, see Table 3.2), the observed differences of variance component estimates between the two models were considerably larger than those from simulations. Thus, the observed differences could not be attributed to the kurtosis of the trait distributions. However, a further detailed dissection of the phenotypic distributions has shown that the distributions of pair difference were all leptokurtic [except non verbal ability scores of MZ twins (Table 3.2)], even for traits where the univariate (single twin) distribution was close to normality. This finding was unexpected and implies a violation of the usual assumption of bivariate normality of twins' phenotypes (e.g. Huggins *et al.*, 1998; Neale, 2003; Rijsdijk and Sham, 2002).

To verify that kurtosis of pair difference was the cause of the observed discrepancy, twin data that mimic the average parameter estimates of IQ phenotypes (except TOWRE) from the known zygosity model were simulated as an example (i.e. $a^2 = 0.22$, $c^2 = 0.30$, $e^2 = 0.48$ with a kurtosis of pair difference of 0.48). Phenotypes (y_1 and y_2) were simulated for a twin pair from a

Table 3.5: Mean estimates (SE) from 1,000 simulated twin data sets with parameters that mimic the average estimates of standardized variance components from the TEDS data (i.e. $a^2 = 0.22$, $c^2 = 0.30$, $e^2 = 0.48$).

Models	a^2 **(SE)**	c^2 **(SE)**	e^2 **(SE)**
Known (normal[a])	0.222 (0.002)	0.297 (0.002)	0.481 (0.001)
Mixture (normal[a])	0.211 (0.007)	0.306 (0.005)	0.483 (0.002)
Known (kurtosis[b])	0.257 (0.002)	0.236 (0.002)	0.507 (0.001)
Mixture (kurtosis[b])	0.581 (0.002)	0.005 (0.001)	0.413 (0.001)

Note: Known and mixture are the known zygosity and mixture distribution models, respectively. [a] Normally distributed twin data. [b] Transformed twin data with a kurtosis value of pair difference of 0.48.

normal distribution, their difference ($D = y_1 - y_2$) was transformed (to D^*) using the previously described Cornish-Fisher transformation, and finally individual observations were backtransformed to $y_i^* = y_i \times \frac{D^*}{D}$. This transformation was made to keep the means and variances of the individual observations approximately the same whilst creating kurtosis of the pair difference. The results (Table 3.5) clearly showed that the variance component estimates of the simulated data from the mixture distribution model resembled those of the IQ phenotypes of the TEDS study: the average estimates for the simulated data were $a^2 = 0.58$, $c^2 = 0.01$, $e^2 = 0.41$, whereas the estimates of the IQ phenotypes (except TOWRE) were $a^2 = 0.62$, $c^2 = 0.01$, $e^2 = 0.38$. The discrepancy on variance component estimates between the normal and transformed (kurtosis) data using the known zygosity model (Table 3.5) was a direct result of the transformation.

To assess further the effects of kurtosis of pair difference on the mixture distribution model, another simulation was carried out by simulating different values of kurtosis on pair difference. The results showed that kurtosis on pair difference had considerable effect on heritability estimations using the mixture distribution model (Figure 3.7). It can be seen clearly from the figure that the mixture distribution produced biased estimates even for small kurtosis values of the pair difference ($-0.5 < k < 0.5$). For a leptokurtic distribution of pair difference, the mixture distribution overestimated the heritability compared to known zygosity models. Even for a small positive kurtosis value ($k < 0.5$), the overestimation was not trivial (i.e. about 40%). On the other hand, the mixture distribution underestimated heritability if the distribution of pair difference was platykurtic. The bias produced by this type distribution was even larger than the bias from a leptokurtic distribution. These results are consistent with the observation in Chapter 2 that the information to separate the two mixtures from the mixture distribution model comes from the difference in the squared variance and kurtosis. Hence, if the pair difference within zygosity class is kurtotic, the mixture distribution will produce biased estimates because the model assumes that the only source of kurtosis is the mixture of two normal distributions. For the known zygosity model, the differences between the variance component estimates from the normal and transformed data are a direct result of the transformation of the data which changed the MZ and DZ correlations (results not shown), and merely show that the correlations depend on the scale of the observations.

Are these results particular to the TEDS data? To explore this possibility, a number of IQ variables from an independent smaller sample of 272 MZ and 191 SS DZ twins with known zygosity and an average age of 16 years old from the ongoing Brisbane Memory, Attention, and Problem-Solving (MAPS) twin study were analysed (Luciano *et al.*, 2003; Wright *et al.*, 2001). Eight IQ measures,

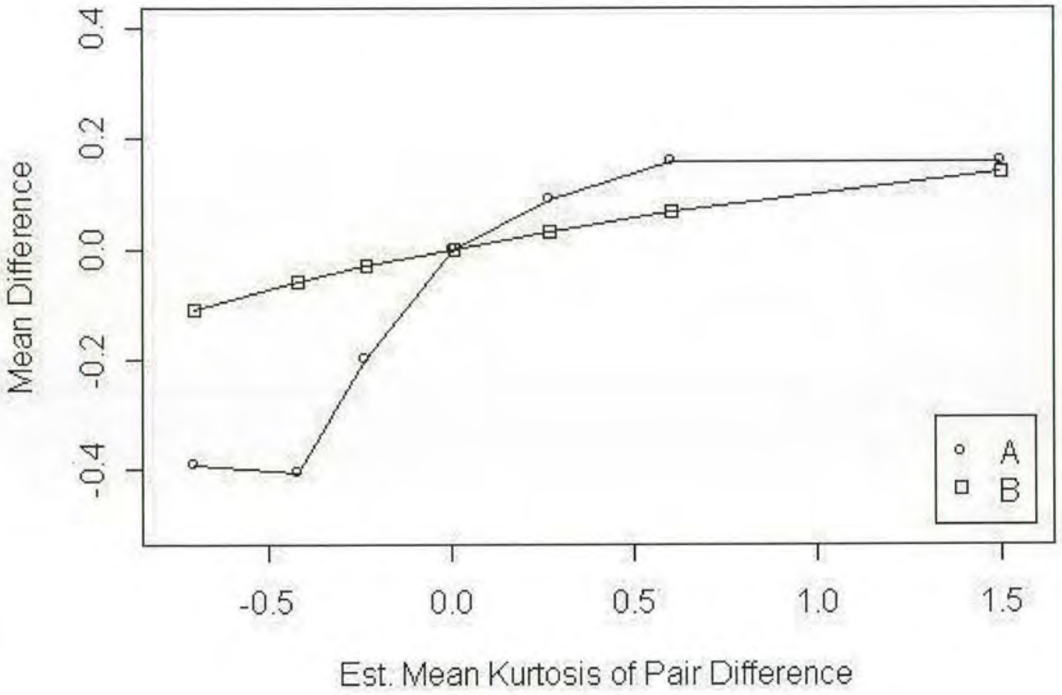

Figure 3.7: Mean differences of heritability estimates between the mixture distribution and known zygosity models for a given estimated mean kurtosis of the pair difference. The simulated parameters are: $a^2 = 0.50$, $c^2 = 0.25$, $e^2 = 0.25$ and simulations are based on 2,000 twin pairs (1,000 MZs and 1,000 DZs) and 1,000 replicates. **A** is the difference between the mixture distribution and known zygosity models for transformed data with specific kurtosis; **B** is the difference of the known zygosity model between transformed data with specific kurtosis and normally distributed data.

namely information, arithmetic, vocabulary, verbal IQ, spatial, object assembly, performance IQ and full scale IQ assessed with the Multidimensional Aptitude Battery II (MAB-II) (Jackson, 1998) were analysed using the known zygosity and mixture distribution model. Individuals with more than 3 standard deviations from the mean were excluded from the analysis, and sex and age were fitted as fixed effects. As with the TEDS data, an ACE model was fitted. The kurtosis of the pair difference of variables from the MAPS study were not different from zero for most traits but the SE was relatively large, ranging from 0.29 to 0.35. Heritability estimates ranged from 0.39 to 0.68 for the known zygosity analysis, and 0.01 to 0.85 for the mixture distribution model analysis. For six out of the eight traits, the estimate of the heritability from the mixture distribution model was larger than the estimate from the known zygosity model, consistent with an observed leptokurtic distribution of the pair difference, averaged over MZ and DZ pairs. For the other two traits, the lower estimate of the heritability from the mixture distribution model was consistent with the observed platykurtic distribution of the pair difference. For these traits, the average kurtosis of the pair difference from MZ and DZ pairs was -0.36 and -0.09, respectively. Although both the estimates of the heritability and their standard errors are larger in the MAPS study, making exact comparisons difficult, the results are qualitatively similar to those from the TEDS study, in that the difference in parameter estimates between the two models are consistent with the observed kurtosis of the pair difference.

What could be the cause of the observed kurtosis on the pair difference, and what are the consequences for twin studies in general? Kurtosis on the pair difference when there is no kurtosis in the population could be due to a 'known' zygosity group itself being a mixture with respect to within-family variances. This could be the case for example if MZ are 'contaminated' with DZ pairs, and vice versa.

This is not likely to be an explanation for the data analysed, because the zygosity protocol is well-established and ambiguities about zygosity were resolved by DNA typing. For $a^2 = 0.4$, $c^2 = 0.2$, $e^2 = 0.4$ and a bivariate normal distribution within zygosity group, a 5% error rate would create a kurtosis value of the pair difference of 0.034 and 0.016 within assigned MZ and DZ groups, respectively (Benyamin *et al.*, 2005). These predicted values are below what was observed from the TEDS data (Table 3.2). Heterogeneity of within-family variance could be due to many factors, including heterogeneity of environmental variance (both MZ and DZ) and heterogeneity of within-family genetic variance (DZ). One speculative biological cause of heterogeneity of variance for MZ pairs is that such pairs vary in the amount of genome-wide methylation or placentation effects that are shared.

The way in which data are collected or scored can also cause the observed kurtosis. For example, sum scores collected from questionnaires may not be multivariate normally distributed. For the TEDS data, the pair difference was extremely kurtotic for height and weight, traits that were reported by parents, and a histogram of the pair difference showed a huge peak at zero, both for MZ and DZ (results not shown). This suggests that the parents may report the average of their twins' height and weight correctly but not their difference. If this reporting bias is stronger in MZ than in DZ then parameter estimates will also be biased using the standard model with known zygosity. Although this may be an explanation for the height and weight data from the TEDS study, it is unlikely to be an explanation for the IQ phenotypes, which were measured on each child individually and separately using a telephone interview and material sent by post, presumably independently of parental input.

Although twin researchers may check normality assumptions of the data before embarking on a maximum likelihood analysis that assumes normality, it is unusual to check for the assumption of bivariate normality in zygosity groups. The results from this study suggest that while the known zygosity model is quite robust to the violation of bivariate normality assumption, a re-examination of bivariate normality for existing data may be prudent. For unknown zygosity data, consistency of the estimates of variance components with those from the known zygosity pairs should be checked. Finally, it is suggested that the possibility of extensive heterogeneity of within-family variance needs further attention.

Appendix 3. The Expected Kurtosis Value of Transformed Data for Given c

Let $x \sim N(0,1)$. The expected kurtosis value of the transformed variate y for given c is derived from the moments of y,

$$y = x + \frac{c}{24}(x^3 - 3x) \tag{3.2}$$

$$y^2 = \left(\frac{8-c}{8}\right)^2 x^2 + \left(\frac{8c-c^2}{96}\right) x^4 + \frac{c^2}{576} x^6 \tag{3.3}$$

$$\begin{aligned} y^4 &= \left(\frac{8-c}{8}\right)^4 x^4 + 2\left(\frac{8c-c}{8}\right)^2 \left(\frac{8c-c^2}{96}\right) x^6 \\ &+ \left[2\left(\frac{8-c}{8}\right)^2 \left(\frac{c^2}{576}\right) + \left(\frac{8c-c^2}{96}\right)^2\right] x^8 \\ &+ \left[2\left(\frac{8c-c^2}{96}\right)\left(\frac{c^2}{576}\right)\right] x^{10} + \left(\frac{c^2}{576}\right)^2 x^{12} \end{aligned} \tag{3.4}$$

Following Kendall and Stuart (1947), the expected value of $2r^{th}$ moment of the normal distribution is:

$E(x^{2r}) = \frac{(2r)!}{2^r r!}\sigma^{2r}$, $E(x^{2r+1}) = 0, r \geq 1$; $E(x^2) = \sigma^2$; $E(x^4) = 3(\sigma^2)^2$,
$E(x^6) = 15(\sigma^2)^3$, $E(x^8) = 105(\sigma^2)^4$, $E(x^{10}) = 945(\sigma^2)^5$; $E(x^{12}) = 10395(\sigma^2)^6$

The expected value of y, y^2 and y^4 are therefore:

$$E(y) = 0 \tag{3.5}$$

$$E(y^2) = \frac{\sigma^2}{576}\left[9(8-c)^2 + 18c(8-c)\sigma^2 + 15c^2\sigma^2\right] \tag{3.6}$$

$$\begin{aligned} E(y^4) = 3\left(\frac{8-c}{8}\right)^4(\sigma^2)^2 + 30\left(\frac{8-c}{8}\right)^2\left(\frac{8c-c}{96}\right)(\sigma^2)^3 \\ + \left[210\left(\frac{8-c}{8}\right)^2\left(\frac{c^2}{576}\right) + 105\left(\frac{8c-c^2}{96}\right)^2\right](\sigma^2)^4 \\ +1890\left(\frac{8c-c^2}{96}\right)\left(\frac{c^2}{576}\right)(\sigma^2)^5 + 10395\left(\frac{c^2}{576}\right)^2(\sigma^2)^6 \end{aligned} \tag{3.7}$$

Then, the expected kurtosis of y is:

$$k(y) = \frac{E\left(y - E(y)\right)^4}{\left(E(y^2)\right)^2} - 3 \tag{3.8}$$

$$= \frac{\left(\begin{array}{c} 3\left(\frac{8-c}{8}\right)^4(\sigma^2)^2 + 30\left(\frac{8-c}{8}\right)^2\left(\frac{8c-c^2}{96}\right)(\sigma^2)^3 \\ + \left[210\left(\frac{8-c}{8}\right)^2\left(\frac{c^2}{576}\right) + 105\left(\frac{8c-c^2}{96}\right)^2\right](\sigma^2)^4 \\ +1890\left(\frac{8c-c^2}{96}\right)\left(\frac{c^2}{576}\right)(\sigma^2)^5 + 10395\left(\frac{c^2}{576}\right)^2(\sigma^2)^6 \end{array}\right)}{\frac{\sigma^2}{576}\left[9(8-c)^2 + 18c(8-c)\sigma^2 + 15c^2\sigma^2\right]} - 3 \tag{3.9}$$

4 A Mixture Distribution Model to Estimate Variance Components from Twins of Unknown Zygosity: Multiple Traits

Abstract

A mixture distribution model was shown to provide reliable genetic and environmental variance component estimates from twin data of unknown zygosity provided that the data follow a bivariate normal distribution (Chapter 3). However, the standard error of the estimates are still larger than the estimates from a conventional method, where zygosity of the twins is known. One suggestion of a way to decrease the standard error of the estimates is to analyse multiple traits simultaneously in multivariate analysis. Additional phenotypes may provide additional zygosity classification as well as increase the effective sample size. Using computer simulations, this chapter assesses the precision of a multivariate mixture distribution model compared to a univariate model in analysing twins of unknown zygosity. The results show that a multivariate analysis reduces the standard error of variance component estimates. From the pattern of decreasing standard error of variance component estimates with the increase of number of traits analysed, it is suggested that if more than approximately 10 traits are analysed simultaneously, the mixture distribution model may provide variance component estimates that are comparable to conventional analysis of known zygosity. This study has opened the possibility of performing genetic analysis from large population based samples, where twin pairs can be identified but their zygosity is unknown.

4.1 Introduction

A large number of twins can be identified from population based surveys (Benyamin *et al.*, 2005; Scarr-Salapatek, 1971; Webbink *et al.*, 2006). Since genetic studies may not be the purpose of such surveys, zygosity of the twins are usually not available. Statistical methods to decompose phenotypic variance of a trait into genetic and environmental components that do not rely on zygosity are therefore needed. A mixture distribution model has been proposed to address this need (Neale, 2003; Benyamin *et al.*, 2005; Benyamin *et al.*, 2006).

In Chapters 2 and 3, the mixture distribution model was shown to be reliable in partitioning the phenotypic variance of twins of unknown zygosity into genetic and environmental components. Computer simulations (Chapter 3) have shown that for a reasonably large sample size, unless the (bivariate) normality assumption was violated, the mixture distributed provides unbiased variance components estimates. Those studies were concerned only with single traits (univariate analysis) and the standard error of the estimates from the mixture distribution was shown to be larger than that of the conventional (known zygosity) model, especially for small to moderate heritability. It has been suggested that by adding more traits in the analysis (multivariate analysis), the additional phenotypes may provide additional zygosity classification (Neale, 2003) and thereby lowering the standard error of the estimates. However, the amount gained by adding more traits in the analysis has not been evaluated. The aim of this chapter is to assess the precision of the multivariate mixture distribution model compared with that of a univariate model in analysing twin of unknown zygosity using computer simulations.

4.2 Simulation

The simulation protocol is an extension of that presented in Chapter 3. By assuming a standard ACE model (additive genetic (A), common environmental (C) and specific environmental (E) effects) of family resemblance, 2,000 twin pairs (equal proportion of MZ and DZ) were simulated from a multivariate normal distribution. MZ pairs were simulated as two individuals sharing the same additive genetic and common environmental effects, but different specific environmental effects. For DZ pairs, the additive genetic effect was simulated by drawing a random value for the first individuals. The additive genetic effect of the second DZ individuals was calculated as half of the additive genetic effect of the first DZ individual plus a Mendelian segregation effect. The Mendelian segregation effect had a variance of 3/4 of the additive genetic effect, so that the total additive genetic variance of the second DZ individual is equal to additive genetic variance. The shared 1/2 additive genetic effect gives an average correlation of 0.5 as defined by the DZ pair relationship. As with MZ pairs, the same common environmental effect was also shared between individuals in a DZ pairs.

A wide range of heritabilities (a^2) were simulated (i.e., 0.1, 0.3, 0.5, 0.7 and 0.9) and the standardised common environmental variance (c^2) was fitted as 0.2, except for the heritability of 0.9, where c^2 is 0.05. One, two and five traits were simulated by assuming no correlation between traits. In addition, to assess whether the degree of correlation between traits has an effect on variance component estimation, a range of different correlations between traits was also simulated assuming equal phenotypic (r_p), genetic (r_g), common environmental (r_c) and specific environmental (r_e) correlations, with values of 0.2, 0.5 and 0.9. No sex effects or other fixed effects were simulated. A rotation method (Barr and Slezak, 1972) was used to generate multivariate normal random vectors with a specified variance-covariance matrix.

As in Chapter 3, all simulated datasets were simultaneously analysed using the known zygosity (conventional) and mixture distribution models using the statistical package Mx (Neale *et al.*, 2002) and all simulations were replicated 1,000 times.

4.3 Results

4.3.1 Uncorrelated Traits

From the simulation, it is shown that by performing a multivariate analysis, the coefficient of variation of heritability estimates [the standard deviation of the estimates divided by the simulated heritability (CV)] from the mixture distribution model decreased as the number traits analysed increased (Figure 4.1). The decrease in CV was greater if the heritability was low to moderate and became almost non-existent if the heritability was greater than 0.5. For example, for heritability of 0.1, the CV decreased from 1.71 in a univariate analysis to 1.03 in a multivariate analysis with 5 traits. On the other hand, there was no difference in the CV of heritability estimates between univariate and multivariate analyses from the known zygosity model, except a small difference for very small heritability ($a^2 < 0.2$) (Figure 4.1). For both models, the CV decreased as the heritability increased.

The standard deviation of the difference between the heritability estimates from the mixture distribution and known zygosity models standardised by the heritability (CV difference) is shown in Figure 4.2. Similar to the pattern of the CV of heritability estimates from the mixture distribution model, the CV of the difference of heritability estimates also decreased as the number of traits increased. Also, the decrease is larger when the heritability is small to moderate.

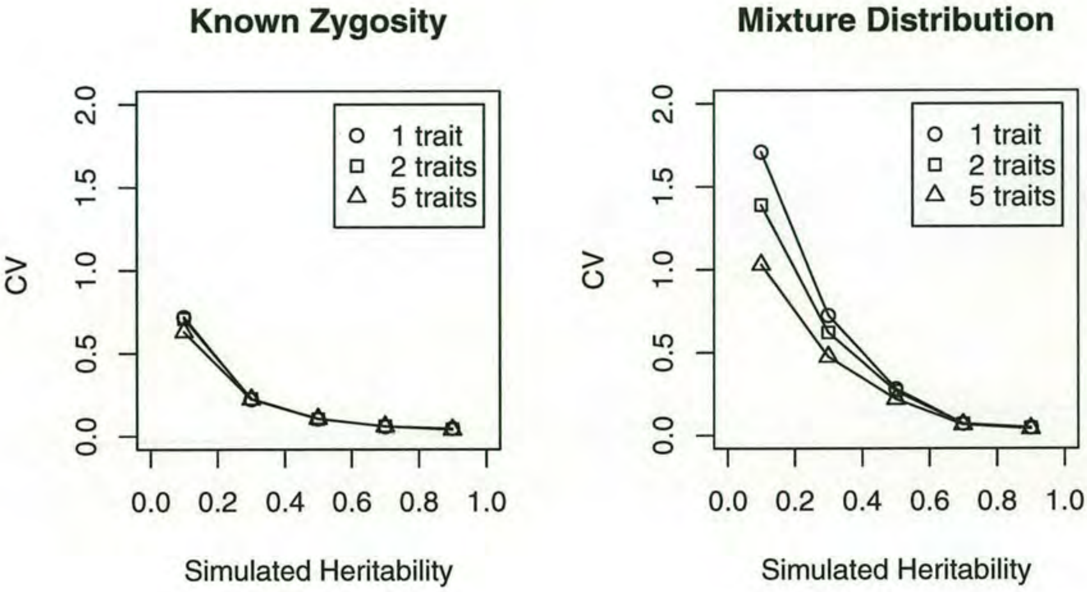

Figure 4.1: Coefficient of variation (CV) of heritability estimates from the known zygosity (left) and mixture distribution (right) models for given number of traits. The c^2 component is constant, i.e., 0.2, except for the heritability of 0.9, where $c^2 = 0.05$ and based on 2,000 twin pairs.

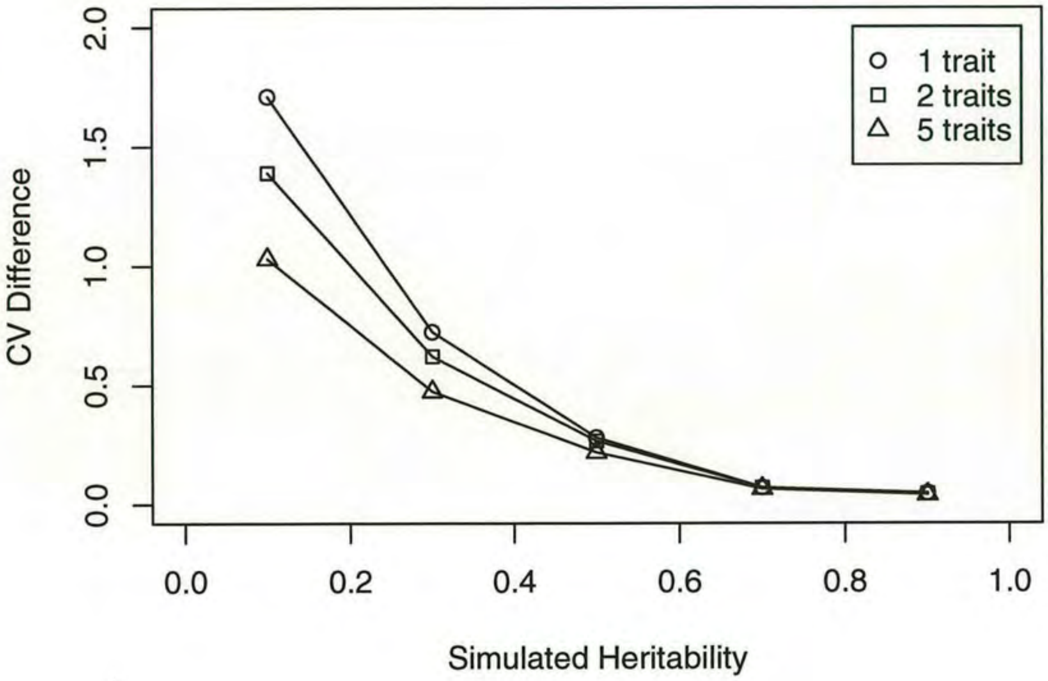

Figure 4.2: Standard deviation of the difference between the heritability estimates from the mixture distribution and known zygosity models standardised by the heritability (CV difference) for given number of traits. The c^2 component is constant (0.2), except for the heritability of 0.9, when it is 0.05. Results are from 1,000 replicate samples of 2,000 twin pairs.

4.3.2 Correlated Traits

Additional simulations for correlated traits showed that the correlation between traits did not have a large effect on variance estimation (Table 4.1). That is, the mean and variance of component estimates appeared to be independent of the correlation between traits.

For both correlated and uncorrelated traits, while multivariate analysis decreases CV, the mean estimate of variance components are slightly biased (Table 4.1). For multivariate analysis with 5 traits, the mean estimate of a^2 from the mixture distribution model was 0.48 compared to the true value of 0.50. The bias is also observed in the estimates of c^2 and e^2. For univariate analysis, a small bias was also reported by Neale (2003).

4.4 Discussion

It has been shown in this chapter that multivariate analysis decreased the variability (hence, the standard error) of the heritability estimates from the mixture distribution model. The more traits analysed simultaneously, the smaller the variability of the heritability estimates from the mixture distribution model. From the pattern observed in Figure 4.1, it is expected that if more than 10 traits were analysed simultaneously, then the mixture distribution would be as good as the known zygosity model with similar sample size.

Another interesting result from this study is that the multivariate mixture model performs well regardless of the magnitude of correlation between traits. However, it should be noted that the results were based on the same phenotypic, genetic, common and specific environmental correlations between traits. The behaviour

Table 4.1: Parameter estimates and SD in parenthesis from the known zygosity and mixture distribution models from multivariate analysis (5 traits) for given correlation coefficients. The simulated standardised variance components were $a^2 = 0.50$; $c^2 = 0.20$; $e^2 = 0.30$. Results were based on 1,000 replicates of 2,000 twin pairs.

Simulated parameters	$r = \mathbf{0}$		$r = \mathbf{0.2}$		$r = \mathbf{0.5}$		$r = \mathbf{0.9}$	
	Known	Mixture	Known	Mixture	Known	Mixture	Known	Mixture
$a^2 = 0.50$	0.502	0.481	0.500	0.480	0.499	0.478	0.505	0.485
	(0.053)	(0.109)	(0.053)	(0.110)	(0.053)	(0.106)	(0.053)	(0.108)
$c^2 = 0.20$	0.198	0.214	0.200	0.215	0.201	0.217	0.194	0.210
	(0.049)	(0.087)	(0.049)	(0.087)	(0.049)	(0.083)	(0.050)	(0.086)
$e^2 = 0.30$	0.300	0.305	0.300	0.305	0.300	0.305	0.301	0.305
	(0.015)	(0.028)	(0.015)	(0.029)	(0.015)	(0.028)	(0.015)	(0.028)

Note: r is the correlation coefficient between traits (identical values of phenotypic (r_p), genetic (r_g), common environmental (r_c) and specific environmental (r_e) correlations).

of the model for traits with different correlations requires further research.

The simulations assumed a multivariate normal distribution for both twin and trait values. For a single trait, the robustness of the mixture model against a violation of normality has been discussed in Chapter 2. It was shown in Chapter 2 that while the model is quite robust to the violation of normality assumption of the trait values, it is very sensitive to the bivariate normality assumption, i.e. the normality of twin pair difference. For multivariate analysis, although it was not evaluated in this chapter, similar results might be expected. This is because multivariate analysis is only an extension of univariate analysis in that if the distribution of the pair difference for one trait is kurtotic, then across traits pair differences may also be kurtotic. However, further research is required to investigate the effect of the violation of bivariate normality assumption on the estimates from multivariate analysis.

In a univariate mixture model (Chapter 2), it was shown that specifying the wrong proportion of MZ twins among same-sex twins in the model has a small effect on the estimates of heritability and error variance, but a moderate effect on the estimate of common environmental variance. Its effect on variance components in a multivariate analysis is expected to be smaller. This rationale is based on the notion that additional traits provide additional zygosity classification (Neale, 2003). However, the relationship between the increase in precision of zygosity assignment and the number of phenotypes analysed simultaneously, was not investigated in this chapter.

The availability of a mixture distribution model to analyse data on twins of unknown zygosity has opened the possibility of performing genetic analysis from large population based samples. Twins can be identified from large population databases with a minimum of required information (date of birth, sex and school or home address). For example, using a name identifier, date of birth, school, grade and year of survey, Webbink *et al.* (2006) successfully identified almost 3,000 twin pairs from about 300,000 pupils registered in a longitudinal survey in the Netherlands. With the availability of a mixture distribution method, these twins can be analysed to answer some of the questions about the genetic and environmental sources of variation of various different phenotypes.

5 A Multivariate Mixed Linear Model to Estimate Co(variance) Components from Twins with Known Zygosity: Understanding the Underlying Genetic and Environmental Aetiology of the Metabolic Syndrome

Abstract

The cluster of obesity, insulin resistance, dyslipidaemia and hypertension, called the metabolic syndrome, has been suggested as a risk factor for cardiovascular disease and type 2 diabetes. The aim of the present study is to quantify genetic and environmental (co)variation of endophenotypes of this cluster in a general population of twin pairs. Data on 13 endophenotypes associated with the metabolic syndrome (grouped into obesity, insulin, lipids and blood pressure related endophenotypes) from 756 adult twin pairs of the GEMINAKAR Study of the Danish Twin Registry were analysed by performing univariate and multivariate genetic analyses. All endophenotypes showed moderate to high heritability (0.34-0.73) and no significant common environmental variance, except for fasting glucose. In general, genetic, environmental and phenotypic correlations between the endophenotypes were strong only within the group, but weak to moderate between groups. However, moderate genetic and specific individual environmental correlations between fasting insulin and obesity related endophenotypes and a moderate specific individual environmental correlation between fasting insulin with lipids endophenotypes indicated that some common genetic or specific individual environmental background might be shared between those components. It is demonstrated that in a general population, the endophenotypes associated with the metabolic syndrome apparently do not share a substantial common genetic or familial environmental background.

5.1 Introduction

The metabolic syndrome or insulin resistance syndrome is characterised by clustering of a group of symptoms related to insulin resistance, impaired glucose tolerance/diabetes, obesity, hypertension and dyslipidaemia (raised triglyceride level and low high-density lipoprotein) (Eckel *et al.*, 2005; Roche *et al.*, 2005; Shaw *et al.*, 2005). The clustering of these symptoms has been suggested as a better predictor for type 2 diabetes (Laaksonen *et al.*, 2002) and cardiovascular disease (Girman *et al.*, 2005; Sundstrom *et al.*, 2006) than expected from its individual components. Currently, the prevalence of this syndrome is high, not only in developed countries but also in developing countries (Cameron *et al.*, 2004). In the United States, about 25% of the adult population have been identified as having the metabolic syndrome (Ford and Giles, 2003).

Growing evidence suggests that variation in the individual components of the metabolic syndrome is in part due to genetic effects. A wide range of heritabilities (mostly moderate to high) were estimated for obesity traits (mean: 0.55; range: 0.37 to 0.80) (Freeman *et al.*, 2002; Li *et al.*, 2006; Lin *et al.*, 2005; Poulsen *et al.*, 2001; Schousboe *et al.*, 2003a); insulin related traits (mean: 0.38; range: 0.08 to 0.75) (Freeman *et al.*, 2002; Henkin *et al.*, 2003; Li *et al.*, 2006; Lin *et al.*, 2005; Samaras *et al.*, 1999; Schousboe *et al.*, 2003a); blood pressure (mean: 0.41; range: 0.16 to 0.76) (Freeman *et al.*, 2002; Li *et al.*, 2006; Lin *et al.*, 2005; Martin *et al.*, 2003; Poulsen *et al.*, 2001) and lipid traits (mean: 0.46; range: 0.20 to 0.70) (Freeman *et al.*, 2002; Li *et al.*, 2006; Lin *et al.*, 2005; Martin *et al.*, 2003; Poulsen *et al.*, 2001). However, the underlying mechanism of the clustering of these characteristics in an individual remains unclear (Hong *et al.*, 1997).

Correlations between insulin related endophenotypes and obesity endophenotypes have been previously estimated to be moderate or high (Nelson *et al.*, 2000; Samaras *et al.*, 1999; Tregouet *et al.*, 1999), while correlations between other metabolic syndrome endophenotypes have been found to be weak or moderate (Martin *et al.*, 2003; Samaras *et al.*, 1999; Tregouet *et al.*, 1999; Rainwater *et al.*, 1997; Perusse *et al.*, 1997). These correlations have surprisingly enough often been interpreted as strong evidence for a common underlying factor underlying the metabolic syndrome.

In a joint statement from the American Diabetes Association and the European Association for the Study of Diabetes, Kahn *et al.* (2005) raised several provocative questions, including whether there is a metabolic syndrome. This question was raised due to the findings that correlations between the metabolic syndrome endophenotypes were weak to moderate. Also, factor analyses have suggested that more than one pathophysiological process underlies the syndrome, where more than one-third of the total variance in the clustering of the metabolic syndrome components was unexplained by latent factors identified from factor analyses (Kahn *et al.*, 2005).

Factor analysis, a data reduction method that explains a large set of observed variables by a smaller set of latent factors, has been widely used to understand the underlying factors of the clustering of the metabolic syndrome components, but its application and interpretation is often problematic due to the subjective nature of this method (Kahn *et al.*, 2005; Lawlor *et al.*, 2004). Two to four latent factors underlying the metabolic syndrome cluster have been reported, with these factors accounting for less than two-thirds of the total variance observed in the cluster (Austin *et al.*, 2004; Edwards *et al.*, 1994; Kahn *et al.*, 2005; Lin *et al.*, 2005; North *et al.*, 2003). Some of the latent factors identified are body mass/fat distribution, insulin/glucose, lipids and hypertension. In one study (Novak *et al.*, 2003), these factors were highly correlated, while in another, they were uncorrelated (Edwards *et al.*, 1997).

As an alternative method, multivariate genetic analysis may provide a complete and objective description of the underlying genetic and environmental architecture of the relationship between traits. In addition to partitioning the phenotypic variance into genetic and environmental components, the phenotypic covariance between the traits is partitioned in similar fashion.

Two approaches have been proposed for estimating the genetic and environmental variations of the metabolic syndrome and its components. First, the phenotypes related to the syndrome were treated as continuous traits (Freeman *et al.*, 2002; Hong *et al.*, 1997; Maison *et al.*, 2001; North *et al.*, 2003; Samaras *et al.*, 1999). Second, the phenotypes related to the metabolic syndrome were dichotomised into having and not having metabolic syndrome based on one of the formal definitions proposed. For example, Lin *et al.* (2005) estimated the heritability of the components of the metabolic syndrome as dichotomised traits based on the National Cholesterol Education Program Adult Treatment Panel III (NCEP/ATPIII) definition. While the latter approach might provide a direct insight into the genetic and environmental aetiology underlying the metabolic syndrome in the affected individuals, the standard errors of the estimates (e.g. heritability) were higher than those obtained from the former approach (Lin *et al.*, 2005). In addition, this approach is dependent on subjective clinical criteria.

An alternative to studying dichotomous traits is to study the underlying phenotypes (called endophenotypes or intermediate traits). Endophenotypes are continuous traits and thus overcome the problem of case definition. The aetiology may be genetically complex, but it will be less complex than that of a clinical endpoint (Dick *et al.*, 2006; Hasler *et al.*, 2006; Flordellis, 2005). This approach is preferable as it can provide a complete description of the underlying genetic and environmental aetiology of the investigated traits.

Phenotypic data on 13 endophenotypes associated with the metabolic syndrome from 756 adult twin pairs of the GEMINAKAR Study of the Danish Twin Registry were analysed in order to elucidate the underlying genetic and environmental relationships between these endophenotypes in general populations.

5.2 Subjects and Methods

5.2.1 Study Subjects

The data analysed in the present study were part of the GEMINAKAR Study. The study is a nation-wide Danish project investigating the genetic epidemiology of a wide variety of phenotypes among Danish twins, including endophenotypes of the metabolic syndrome (Schousboe *et al.*, 2003a; Schousboe *et al.*, 2004). The twins were recruited from two cohorts of the nation-wide, population-based Danish Twin Registry. Cohort I covers the birth cohorts 1931-1952, while cohort II covers the birth cohorts 1953-1982. Invitations to take part in a full day clinical investigation were sent to 2585 randomly chosen twin pairs who fulfilled the criteria that at least one twin should live within 100 km from one of the two clinical investigation sites (Odense and Copenhagen) and the pair should not take part in other studies at the same time. Cohort II was furthermore chosen based on a previous self report of being healthy. The invitation contained detailed information about the study and its exclusion criteria (i.e. known diabetes or cardiovascular disease, conditions making a progressive maximal bicycle test impossible, pregnancy, and breast feeding). A reply coupon was enclosed for the twins to give information about their present health status and to either consent or decline telephone contact. If one twin partner in a pair did not respond or was not willing to participate, the pair as such was excluded. In 1098 complete twin pairs (42%) both were willing and able to participate.

A stratified sample of 756 twin pairs underwent an extensive full day clinical examination of a variety of phenotypes. The main focus was on phenotypes related to insulin resistance, obesity, and cardiovascular risk factors. The population included 311 monozygotic (MZ) twin pairs, 445 dizygotic (DZ) twin pairs [314 same-sex (SS) twin pairs and 131 opposite-sex (OS) twin pairs] with a mean age of 38 years (range 18 to 67 years old). The examinations were running in parallel at The Danish Twin Registry in Odense and at The Institute of Preventive Medicine in Copenhagen from 1997 to 2000. The twins in a pair were examined on the same day. DNA-based microsatellite markers with the PE Applied Biosystems AmpFISTR Profiler Plus Kit were used to determine zygosity of the twins. The study was approved by all the Danish regional scientific-ethical committees, the Danish Data Protections Agency and conducted according to the principles of the Helsinki Declaration.

5.2.2 Phenotypic Studies

In this study, the endophenotypes were grouped into four different groups, i.e. obesity related endophenotypes, insulin related endophenotypes, blood pressure and lipid endophenotypes. Endophenotypes were grouped based on clear physiological relationships and supported by the earlier factor analyses (Novak *et al.*, 2003; Edwards *et al.*, 1997). Obesity was indicated by body mass index (BMI) and waist circumference (WAIST). Insulin related endophenotypes were represented by fasting, 30 and 120 minutes glucose (GLU0, GLU30, GLU120) and insulin (INS0, INS30, INS120) levels. Blood pressure included systolic (SBP) and diastolic blood pressure (DBP). Finally, the concentration of high (HDL), low (LDL) density lipoproteins and triglycerides (TG) were measured and are representative of lipid endophenotypes.

BMI was calculated as weight (kg) divided by square of height (m), where weight was measured using a standing beam scale to the nearest 0.1 kg and height was measured using a vertical scale with a horizontal moving headboard to the nearest centimetre. The measure (cm) at midway between the lowest rib and iliac crest was defined as waist circumference (Schousboe *et al.*, 2004). A WHO-standard oral glucose tolerance test was carried out after 10 - 12 hour overnight fast. Blood was taken before oral glucose ingestion, 30 and 120 minutes later, which was then analysed using the glucose dehydrogenase oxidation method. Then, a two site, two step, time-resolved immunofluorometric assay was used to measure serum insulin concentration (Schousboe *et al.*, 2003a). Cholesterol, HDL and TG were measured on fasting blood samples by a colorimetric method (VITROS, Johnson & Johnson). LDL was calculated (F-formula) by subtracting HDL and (0.45 × TG) from total cholesterol. Systolic and diastolic blood pressures were measured after 30 minutes rest using a conventional mercury sphygmomanometer. Measurements were taken three times and the mean was used for analysis.

5.2.3 Statistical Analyses

Descriptive statistics of the data were explored using SPSS version 12.0.2 for Windows (SPSS Inc., 1989 - 2003) to examine trait distributions (Table 5.1). For endophenotypes showing a non-normal distribution, including BMI, GLU0, GLU120, INS0, INS30, INS120, HDL, and TG, a 100 × natural logarithm transformation was carried out to make the trait distribution (near) normal, which is an assumption in a maximum likelihood method. To test for the effects of sex and zygosity on the endophenotypes related to the metabolic syndrome, a general linear model was fitted. Twin correlations were calculated after adjustment for age and sex effects.

By comparing the resemblance of MZ twin pairs to that of DZ twin pairs, a twin design allows the phenotypic variance of phenotypes to be partitioned into underlying additive genetic variance (A), common environmental variance shared by a twin pair (C) and environmental variance specific to individuals (E) (ACE model). The analysis assumes that both type of twins share the same degree of common environmental experiences (the so-called common environment assumption) and that the genetic effects on a trait are additive. An ACE model was selected since MZ twin correlations were generally less than twice that of DZ correlations (Table 5.2), which suggest that there is no appreciable dominance or epistatic genetic effect.

Univariate and multivariate genetic analyses were performed in this study. Univariate genetic analysis estimates the genetic and environmental variance components of each trait independently (ignoring the dependency between traits). On the other hand, multivariate genetic analysis accounts for dependency between traits by partitioning the phenotypic covariances between these in addition to partitioning their variances. Thus, multivariate genetic analysis estimates the proportion of phenotypic variance of individual traits due to genetic and environmental variances as well as the genetic and environmental correlations between traits. Variance component estimates from univariate analysis are still needed for comparison.

The partition of (co)variance components were estimated mainly by residual maximum likelihood (REML), fitting a mixed linear model (Visscher *et al.*, 2004) using the ASReml package (Gilmour *et al.*, 2002). The REML method takes into account that fixed effects or covariates are fitted when estimating the (co)variance components. If the analysis contains few such effects, the estimates are very similar to an analysis using maximum likelihood, as implemented in, for example, Mx (Neale *et al.*, 2002). This was verified for univariate analyses and multivariate analyses (results not shown). Variation due to age and sex was removed by including these as fixed effects in the models.

By including DZ-OS twins in the analyses, the possibility of sex-specific genes was tested for all endophenotypes in the univariate analyses. This was achieved by testing whether the coefficient of additive genetic covariance (CovA) between males and females for DZ-OS twin pairs was less than that of DZ-SS (i.e. 0.5), while the variance components in males and females are still allowed to be different. This test is equivalent to testing whether the genetic correlation between males and females for a particular trait deviates from unity. The log-likelihoods between the two models (OS CovA = 0.5 (H_0) vs OS CovA < 0.5 (H_1)) were compared. In addition, the heterogeneity of variance components across sexes was tested in univariate analyses for all endophenotypes by allowing the variance components to be different in males and females. Both of the tests for sex-specific genes and heterogeneity of variance components were performed simultaneously using the statistical package Mx (Neale *et al.*, 2002).

5.3 Results

5.3.1 Descriptive Statistics

Descriptive statistics of the endophenotypes related to the metabolic syndrome are presented in Table 5.1. There were no age differences (P-value < 0.05) across sexes and zygosity groups. No significance differences among zygosity groups were observed for all endophenotypes. However, across sexes, mean differences (P-value < 0.05) were observed for most endophenotypes, except for LDL, INS0 and INS30. Therefore, in the variance components analyses, the endophenotypes were adjusted for the effect of sex and heterogeneity of variance across sexes was modelled in the univariate analyses.

Table 5.2 shows that for all endophenotypes (except for GLU0), MZ correlations were consistently larger than DZ correlations, suggesting that genetic factors contributed to the phenotypic variation of the metabolic syndrome endophenotypes. Indeed, the results of statistical genetic modelling (described in the next sections) confirmed that a significant part of the phenotypic variation of most endophentypes was due to genetic factors. Furthermore, the pattern of MZ and DZ correlations, in which MZ correlations are generally less than twice DZ correlations, indicated that an ACE model is an appropriate model to explain the phenotypic variations of the metabolic syndrome endophenotypes.

Table 5.1: Mean (SD) and range of the metabolic syndrome endophenotypes by sex and zygosity.

Endophenotypes	**Sex**		**Zygosity**			**All**	**Range**
	Male	Female	MZ	DZ-SS	DZ-OS		
No. Subjects (range)	702-729	753-783	603-622	602-628	251-262	1465-1512	
Age (year)	38.0 (11.1)	37.5 (10.7)	37.7 (11.0)	37.9 (10.5)	37.7 (11.5)	37.8 (10.9)	18-67
Body Mass Index (BMI) (kg/m^2)	24.9 (3.1)	23.9 (3.8)	24.5 (3.4)	24.4 (3.6)	24.4 (3.4)	24.4 (3.5)	16.1-43.7
WAIST Circumference (cm)	89.3 (9.0)	78.6 (9.6)	83.8 (10.6)	83.7 (11.0)	84.0 (10.4)	83.8 (10.7)	58.0-122.0
Fasting Glucose (GLU0) (mmol/l)	4.9 (0.6)	4.7 (0.5)	4.8 (0.6)	4.8 (0.5)	4.8 (0.6)	4.8 (0.6)	2.7-13.0
Fasting Insulin (INS0) (pmol/l)	36.3 (19.6)	38.2 (19.4)	37.6 (18.5)	36.7 (19.7)	37.6 (21.4)	37.2 (19.5)	6.0-182.0
30 Minutes Glucose (GLU30) (mmol/l)	8.7 (1.6)	8.3 (1.5)	8.4 (1.6)	8.5 (1.5)	8.5 (1.7)	8.5 (1.6)	3.6-18.7
30 Minutes Insulin (INS30) (pmol/l)	301.2 (193.0)	320.8 (180.8)	308.6 (163.0)	311.8 (203.9)	316.0 (198.5)	311.2 (187.0)	36.0-1741.0
120 Minutes Glucose (GLUC120) (mmol/l)	5.9 (1.6)	6.4 (1.3)	6.3 (1.6)	6.1 (1.3)	6.1 (1.3)	6.2 (1.4)	2.5-23.6
120 Minutes Insulin (INS120) (pmol/l)	143.2 (143.4)	192.1 (125.9)	173.7 (143.1)	159.4 (125.3)	176.2 (146.8)	168.3 (136.9)	8.0-1992.0
Systolic Blood Pressure (SBP) (mmHg)	120.4 (14.0)	113.6 (13.6)	116.6 (14.7)	117.4 (14.4)	116.2 (12.6)	116.9 (14.2)	78.0-204.0
Diastolic Blood Pressure (DBP) (mmHg)	70.1 (10.5)	67.2 (10.1)	68.7 (10.8)	68.1 (10.3)	69.5 (9.6)	68.6 (10.4)	42.7-112.0
High Density Lipoprotein (HDL) (mmol/l)	1.4 (0.4)	1.6 (0.5)	1.5 (0.5)	1.5 (0.5)	1.6 (0.3)	1.5 (0.5)	0.4-5.4
Low Density Lipoprotein (LDL) (mmol/l)	3.3 (1.1)	3.2 (1.1)	3.2 (1.1)	3.3 (1.1)	3.4 (1.0)	3.3 (1.1)	-2.3-7.4
Triglyceride (mmol/l)	1.4 (0.8)	1.2 (0.7)	1.3 (0.7)	1.3 (0.6)	1.4 (1.0)	1.3 (0.8)	0.2-14.5

Note: MZ: monozygotic twins; DZ-SS: dizygotic same-sex twins; DZ-OS: dizygotic-opposite-sex twins.

Table 5.2: Twin correlations and their standard errors after adjustment for sex and age effects. MZ-M, MZ-F, DZ-M, DZ-F, and DZ-OS are MZ male, MZ female,DZ male, DZ female, and DZ opposite-sex twins, respectively.

Trait	**MZ-M (s.e.)**	**DZ-M (s.e.)**	**MZ-F (s.e.)**	**DZ-F (s.e.)**	**DZ-OS (s.e.)**
BMI	0.67 (0.06)	0.40 (0.07)	0.75 (0.06)	0.49 (0.07)	0.29 (0.08)
WAIST	0.55 (0.07)	0.34 (0.08)	0.65 (0.06)	0.46 (0.07)	0.15 (0.09)
GLU0	0.52 (0.07)	0.30 (0.08)	0.34 (0.08)	0.36 (0.07)	0.35 (0.08)
INS0	0.36 (0.08)	0.17 (0.08)	0.52 (0.07)	0.33 (0.08)	0.14 (0.09)
GLU30	0.55 (0.07)	0.28 (0.08)	0.44 (0.08)	0.27 (0.08)	0.25 (0.09)
INS30	0.51 (0.07)	0.29 (0.08)	0.50 (0.07)	0.32 (0.08)	0.40 (0.08)
GLU120	0.43 (0.07)	0.22 (0.08)	0.41 (0.08)	0.20 (0.08)	0.26 (0.09)
INS120	0.48 (0.07)	0.24 (0.08)	0.41 (0.08)	0.22 (0.08)	0.29 (0.08)
SBP	0.61 (0.06)	0.32 (0.08)	0.70 (0.06)	0.46 (0.07)	0.13 (0.09)
DBP	0.63 (0.06)	0.31 (0.08)	0.71 (0.06)	0.43 (0.07)	0.20 (0.09)
HDL	0.65 (0.06)	0.35 (0.08)	0.55 (0.07)	0.45 (0.07)	0.23 (0.09)
LDL	0.69 (0.06)	0.40 (0.07)	0.80 (0.05)	0.38 (0.07)	0.22 (0.09)
VLDL	0.34 (0.08)	0.32 (0.08)	0.54 (0.07)	0.38 (0.07)	0.07 (0.09)
TG	0.38 (0.08)	0.33 (0.08)	0.56 (0.07)	0.35 (0.07)	0.01 (0.09)

5.3.2 Univariate Genetic Analyses

The proportions of phenotypic variance of the endophenotypes due to additive genetic (a^2, heritability), common (c^2) and individual specific (e^2) environmental effects from univariate analyses are presented in Table 5.3. The variance component estimates from analysing all data and same-sex twins only were very similar for most endophenotypes, except for TG. With the exception of GLU0, moderate to high heritability were estimated for all endophenotypes (0.34 - 0.73). Most endophenotypes had very small common environmental variance (C) components, except for GLU0. Thirty percent of variation of fasting glucose was explained by common environmental variance shared by twin pairs, whereas the figures are less than 13% for the rest of the endophenotypes. For WAIST, INS0, LDL and SBP, the C components were close to zero.

The test for sex-specific genes showed that there was no evidence of sex-specific genes for all endophenotypes (results not shown). In addition, the difference in the variance component estimates and their ratios (e.g. heritability), when including or excluding DZ-OS twins, were mostly small (results not shown). Therefore, for subsequent analyses, DZ-OS twins were in included in the analyses with the DZ-SS twins (but adjusting for a sex effect on the mean).

The test for heterogeneity of variance components across sexes showed that the variance component profiles were significantly different in males and females for 8 out of 13 endophenotypes (P-value < 0.05), including for BMI, WAIST, GLU0, INS0, GLU120, INS120, SBP, and TG. However, this heterogeneity was ignored in the full multivariate analysis (13 traits) by pooling the data on males and females (after sex adjustment), because it would require too many mutually highly correlated parameters to take the heterogeneity properly into account.

Table 5.3: The proportions of phenotypic variance of endophenotypes related to the metabolic syndrome due to additive genetic effect (a^2), common environmental effects shared by twin pairs (c^2) and specific individual environmental effects (e^2) from univariate and multivariate analyses. The numbers in the brackets are the corresponding standard errors.

Endophenotypes	**Univariate Analysis**^{a}			**Univariate Analysis**^{b}			**Multivariate Analysis**^{a}		
	a^2 (s.e.)	c^2 (s.e.)	e^2 (s.e.)	a^2 (s.e.)	c^2 (s.e.)	e^2 (s.e.)	a^2 (s.e.)	c^2 (s.e.)	e^2 (s.e.)
BMI	0.68 (0.08)	0.05 (0.08)	0.27 (0.02)	0.61 (0.09)	0.14 (0.09)	0.26 (0.02)	0.67 (0.08)	0.07 (0.08)	0.26 (0.02)
WAIST	0.62 (0.10)	0.01 (0.09)	0.37 (0.03)	0.51 (0.11)	0.13 (0.09)	0.36 (0.03)	0.64 (0.10)	0.01 (0.08)	0.35 (0.03)
GLU0	0.10 (0.12)	0.30 (0.09)	0.60 (0.04)	0.04 (0.13)	0.37 (0.11)	0.59 (0.04)	0.08 (0.12)	0.33 (0.09)	0.59 (0.04)
INS0	0.46 (0.04)	0.00 (0.00)	0.54 (0.04)	0.47 (0.13)	0.00 (0.11)	0.53 (0.04)	0.49 (0.04)	0.00 (0.08)	0.51 (0.04)
GLU30	0.39 (0.12)	0.08 (0.10)	0.53 (0.04)	0.32 (0.13)	0.14 (0.11)	0.54 (0.04)	0.38 (0.12)	0.10 (0.08)	0.52 (0.04)
INS30	0.50 (0.11)	0.06 (0.09)	0.44 (0.04)	0.54 (0.12)	0.01 (0.11)	0.45 (0.04)	0.52 (0.10)	0.05 (0.10)	0.43 (0.04)
GLU120	0.34 (0.13)	0.06 (0.10)	0.60 (0.04)	0.36 (0.14)	0.04 (0.12)	0.60 (0.05)	0.37 (0.12)	0.05 (0.10)	0.58 (0.04)
INS120	0.45 (0.12)	0.02 (0.10)	0.53 (0.04)	0.45 (0.14)	0.01 (0.11)	0.54 (0.04)	0.46 (0.12)	0.02 (0.09)	0.52 (0.04)
SBP	0.62 (0.10)	0.01 (0.09)	0.37 (0.03)	0.53 (0.11)	0.12 (0.10)	0.35 (0.03)	0.63 (0.10)	0.01 (0.09)	0.36 (0.03)
DBP	0.61 (0.10)	0.04 (0.09)	0.35 (0.03)	0.55 (0.11)	0.11 (0.10)	0.34 (0.03)	0.62 (0.10)	0.03 (0.09)	0.35 (0.03)
HDL	0.50 (0.10)	0.13 (0.08)	0.37 (0.03)	0.53 (0.11)	0.12 (0.09)	0.35 (0.03)	0.49 (0.10)	0.14 (0.08)	0.37 (0.03)
LDL	0.73 (0.02)	0.00 (0.00)	0.27 (0.02)	0.71 (0.10)	0.03 (0.10)	0.26 (0.02)	0.72 (0.09)	0.00 (0.08)	0.28 (0.02)
TG	0.39 (0.12)	0.06 (0.10)	0.55 (0.04)	0.19 (0.13)	0.26 (0.10)	0.56 (0.05)	0.37 (0.12)	0.07 (0.10)	0.56 (0.05)

Note: ^{a}Analysis using all data (MZ, DZ-SS and DZ-OS). ^{b} Analysis using MZ and DZ-SS only.

The effect of ignoring this heterogeneity was investigated in detail for selected set of bivariate analyses.

5.3.3 Multivariate Genetic Analysis

Multivariate genetic analysis of all 13 endophenotypes was performed simultaneously in a single analysis using all data. The estimated proportion of variance components from the multivariate analysis are presented in Table 5.3. The estimates and their standard errors were similar to those obtained by univariate analyses.

To check the effects of ignoring the heterogeneity of variance components on the correlations between traits, a series of bivariate analyses were performed. Bivariate analysis between WAIST and INS0 is presented as an example (Table 5.4). In the univariate analyses, both traits showed significant differences between variance component estimates for males and females and were highly correlated. However, bivariate analysis showed that the genetic (r_g), phenotypic (r_p) and specific individual environmental (r_e) correlations were very similar in males and females. The pooled estimates were somewhere in the middle between the two estimates with narrower confidence intervals. Therefore, the possibility of a small unknown bias has been traded for the more accurate estimates.

Phenotypic, genetic and specific individual environmental correlations between the endophenotypes estimated from multivariate analysis are presented in Tables 5.5 and 5.6, and to aid the interpretation, summarised in Table 5.7. No common environmental correlations between pairs of endophenotypes were significantly different from zero (results not shown). The environmental correlations between endophenotypes were mostly due to correlations within individuals (r_e) rather

Table 5.4: An example of bivariate analysis by taking into account heterogeneity of variance components across sexes, while fixing the scaled additive genetic covariance in DZ-OS at 0.5 (the same as that of DZ-SS for both traits). WAIST and INS0 were chosen because both endophenotypes are highly correlated and show heterogeneity of variance components in males and females.

Sex	**Correlation between WAIST and INS0 (95% CI)**		
	r_g	r_e	r_p
Male	0.59 (0.22-1.00)	0.52 (0.39-0.62)	0.52 (0.46-0.57)
Female	0.41 (0.04-0.68)	0.32 (0.17-0.45)	0.41 (0.34-0.47)
Combined	0.51 (0.30-0.71)	0.41 (0.32-0.50)	0.46 (0.41-0.50)

Note: r_g, r_e, r_p are genetic, specific individual and phenotypic correlations, respectively.

than shared by twin pairs (r_c).

The main result from the multivariate analysis showed that the correlations between individual endophenotypes assigned to different groups were generally weak and not significantly different from zero. The strongest correlation between individual endophenotypes assigned to different groups was between obesity endophenotypes (WAIST) and insulin related endophenotypes (INS0) (r_g = 0.50 ± 0.11; r_e = 0.41 ± 0.05; r_p = 0.46 ± 0.02). The correlations between obesity endophenotypes and lipid endophenotypes were mostly weak, except for moderate specific individual environmental correlations between either BMI or WAIST with TG. Weak correlations were estimated between obesity and blood pressure related endophenotypes. For example, the phenotypic and genetic correlations between BMI and SBP were 0.28 ± 0.03 and 0.27 ± 0.09, respectively.

Table 5.5: Phenotypic correlations (r_p) between the metabolic syndrome endophenotypes. The numbers in the bracket are the corresponding standard errors.

	BMI	**WAIST**	**GLU0**	**INS0**	**GLU30**	**INS30**	**GLU120**	**INS120**	**SBP**	**DBP**	**HDL**	**LDL**
WAIST	0.86 (0.01)											
GLU0	0.11 (0.03)	0.11 (0.03)										
INS0	0.46(0.02)	0.46(0.02)	0.18(0.03)									
GLU30	0.01(0.03)	0.05(0.03)	0.38(0.02)	0.10(0.03)								
INS30	0.26(0.03)	0.28(0.03)	0.06(0.03)	0.53(0.02)	0.25(0.03)							
GLU120	0.02(0.03)	0.03(0.03)	0.27(0.03)	0.13(0.03)	0.30(0.03)	-0.07(0.03)						
INS120	0.21(0.03)	0.21(0.03)	0.11(0.03)	0.50(0.02)	0.19(0.03)	0.34(0.03)	0.69(0.01)					
SBP	0.28(0.03)	0.26(0.03)	0.12(0.03)	0.24(0.03)	0.13(0.03)	0.15(0.03)	0.15(0.03)	0.18(0.03)				
DBP	0.26(0.03)	0.23(0.03)	0.11(0.03)	0.23(0.03)	0.06(0.03)	0.14(0.03)	0.13(0.03)	0.19(0.03)	0.69(0.03)			
HDL	-0.17(0.03)	-0.19(0.03)	-0.05(0.03)	-0.15(0.03)	0.01(0.03)	-0.11(0.03)	-0.04(0.03)	-0.10(0.03)	-0.02(0.03)	0.03(0.03)		
LDL	0.11(0.03)	0.13(0.03)	0.02(0.03)	0.09(0.03)	0.07(0.03)	0.10(0.03)	-0.01(0.03)	0.06(0.03)	0.11(0.03)	0.10(0.03)	-0.19(0.03)	
TG	0.22(0.03)	0.27(0.03)	0.13(0.03)	0.35(0.02)	0.20(0.03)	0.27(0.03)	0.20(0.03)	0.33(0.03)	0.20(0.03)	0.20(0.03)	-0.22(0.03)	0.24(0.03)

Table 5.6: Genetic (r_g) and individual environmental (r_e) correlations between the metabolic syndrome endophenotypes in the upper and lower diagonal, respectively. The numbers in the brackets are the corresponding standard errors.

	BMI	**WAIST**	**GLU0**	**INS0**	**GLU30**	**INS30**	**GLU120**	**INS120**	**SBP**	**DBP**	**HDL**	**LDL**	**TG**
BMI		0.81(0.03)	0.12(0.28)	0.45(0.10)	-0.18(0.15)	0.31(0.10)	-0.30(0.16)	0.24(0.12)	0.27(0.09)	0.27(0.09)	-0.10(0.11)	0.11(0.09)	0.16(0.14)
WAIST	0.88(0.01)		0.14(0.31)	0.50(0.11)	-0.09(0.16)	0.31(0.11)	-0.31(0.18)	0.25(0.13)	0.26(0.11)	0.21(0.11)	-0.07(0.13)	0.10(0.10)	0.18(0.15)
GLU0	0.17(0.05)	0.18(0.05)		-0.11(0.44)	-0.14(0.53)	0.60(0.54)	-0.45(0.67)	0.27(0.43)	-0.30(0.41)	-0.20(0.37)	-0.43(0.45)	0.46(0.41)	-0.58(0.68)
INS0	0.48(0.04)	0.41(0.05)	0.26(0.05)		-0.03(0.20)	0.64(0.11)	-0.15(0.21)	0.60(0.13)	0.19(0.14)	0.04(0.14)	-0.08(0.16)	0.23(0.13)	-0.01(0.20)
GLU30	0.18(0.05)	0.20(0.05)	0.32(0.05)	0.15(0.05)		0.13(0.18)	0.36(0.21)	0.42(0.19)	0.04(0.16)	0.05(0.16)	0.29(0.18)	0.15(0.14)	0.26(0.22)
INS30	0.33(0.05)	0.29(0.05)	-0.04(0.06)	0.46(0.02)	0.36(0.05)		-0.01(0.19)	0.36(0.14)	0.16(0.13)	0.10(0.13)	-0.02(0.14)	0.30(0.12)	0.32(0.17)
GLU120	0.25(0.05)	0.24(0.05)	0.27(0.05)	0.19(0.05)	0.25(0.05)	-0.08(0.06)		0.75(0.11)	-0.08(0.17)	-0.11(0.17)	0.07 (0.19)	-0.04(0.15)	0.09(0.24)
INS120	0.33(0.05)	0.31(0.05)	0.09(0.05)	0.44(0.05)	0.17(0.05)	0.36(0.05)	0.64(0.03)		-0.03(0.15)	-0.03(0.15)	-0.22(0.16)	0.10(0.13)	0.34(0.18)
SBP	0.28(0.05)	0.20(0.05)	0.18(0.05)	0.17(0.06)	0.12(0.05)	0.08(0.06)	0.20(0.05)	0.21(0.05)		0.72(0.06)	-0.02(0.13)	0.11(0.10)	0.28(0.16)
DBP	0.27(0.05)	0.21(0.05)	0.12(0.05)	0.21(0.05)	0.04(0.06)	0.09(0.06)	0.19(0.05)	0.16(0.05)	0.72(0.03)		-0.01(0.13)	0.09(0.10)	0.15(0.16)
HDL	-0.20(0.06)	-0.23(0.05)	-0.01(0.06)	-0.15(0.06)	-0.05(0.06)	-0.08(0.06)	-0.07(0.06)	-0.10(0.06)	0.09(0.06)	0.11(0.06)		-0.03(0.11)	-0.32(0.17)
LDL	0.18(0.06)	0.17(0.06)	0.01(0.06)	0.02(0.06)	-0.01(0.06)	-0.05(0.06)	0.06(0.06)	0.05(0.06)	0.07(0.06)	0.06(0.06)	-0.30(0.05)		0.38(0.14)
TG	0.31(0.05)	0.32(0.05)	0.17(0.05)	0.44(0.05)	0.10(0.05)	0.21(0.05)	0.21(0.05)	0.34(0.05)	0.13(0.06)	0.16(0.05)	-0.24(0.05)	0.19(0.06)	

Table 5.7: Summary of genetic (r_g, upper diagonal) and specific individual environmental (r_e, lower diagonal) correlations between endophenotypes related to the metabolic syndrome assigned to different groups.

Groups of Endophenotypes	**Obesity**	**Insulin**	**Blood Pressure**	**Lipids**
Obesity (BMI, WAIST)	BMI and WAIST are strongly correlated ($r_p = 0.86 \pm 0.01$; $r_g = 0.81 \pm 0.03$; re = 0.88)	Weak or no significant correlations, except moderate correlations between INS0 with either BMI or WAIST (0.45 ± 0.10 and 0.50 ± 0.11, respectively)	Weak or no significant correlations	Weak or no significant correlations
Insulin (GLU0, GLU30, GLU120, INS0, INS30, INS120)	Generally weak or no significant correlations, except moderate correlations between fasting insulin with either BMI or WAIST (0.48 ± 0.04 and 0.41 ± 0.05, respectively)	Strong phenotypic and genetic correlations between insulin levels. Moderate phenotypic correlations between glucose levels, but their genetic correlations are not significant. No significant to moderate correlations between glucose and insulin levels, except strong correlation between INS120 and GLU120 ($r_p = 0.69 \pm 0.01$; $r_g = 0.75 \pm 0.11$; $r_e = 0.64 \pm 0.03$)	No significant correlations	Weak or no significant correlations, except moderate correlation between INS30 and LDL (0.30 ± 0.12)

Table 5.7: Continued.

Groups of Endophenotypes	**Obesity**	**Insulin**	**Blood Pressure**	**Lipids**
Blood pressure (SBP, DBP)	Weak correlations	Weak correlations	SBP is highly correlated with DBP ($r_p = 0.69 \pm 0.03$; $r_g = 0.72 \pm 0.06$; $r_e = 0.72 \pm 0.03$)	Weak or no significant correlations
Lipids (HDL, LDL, TG)	Generally weak correlations, except moderate correlations between either BMI or WAIST with TG (0.31 ± 0.05 or 0.32 ± 0.05)	Generally weak correlations, except moderate correlations between TG with fasting insulin (0.44 ± 0.05)	Weak or no significant correlations	Generally weak or no significant correlations between lipid traits

Note: The genetic, specific individual environmental and phenotypic correlations assigned to the same groups of endophenotypes are on the diagonal. Since no formal testing of correlations of groups was performed, the correlations between endophenotypes assigned to the same or different groups are provided descriptively.

Correlations between lipid endophenotypes and blood pressure were low. The correlations between insulin related endophenotypes and lipid endophenotypes were generally weak, except for moderate specific individual environmental correlations between TG and INS0 (0.44 ± 0.05). Estimated genetic, specific individual environmental and phenotypic correlations between blood pressure and insulin related endophenotypes were very weak.

In contrast to this, genetic and phenotypic correlations between endophenotypes assigned to the same groups were generally higher. Correlations between BMI and WAIST (both are indicators of some form of obesity) were high ($r_g = 0.81 \pm 0.03$; $r_e = 0.88 \pm 0.01$; $r_e = 0.86 \pm 0.01$). Strong correlations were estimated between INS0 with INS30 or INS120, but moderate between INS30 and INS120. The phenotypic correlations between various glucose levels were moderate, but their genetic correlations were not significantly different from zero. The correlations between various insulin and glucose levels were not significantly different from zero or moderate, except for a strong correlation between INS120 and GLU120 ($r_g = 0.75 \pm 0.11$; $r_e = 0.64 \pm 0.03$; $r_e = 0.69 \pm 0.01$). SBP and DBP were strongly related ($r_g = 0.72 \pm 0.06$; $r_e = 0.72 \pm 0.03$; $r_e = 0.69 \pm 0.03$). The correlations between various lipid measurements were mostly weak.

5.4 Discussion

In this chapter, the observed phenotypic (co)variation of endophenotypes associated with the metabolic syndrome has been partitioned into genetic and environmental components. The results showed that genetic factors largely contributed to the individual differences in most of the endophenotypes. At the same time, the results showed that there does not appear to be major common genetic or familial environmental background shared by the endophenotypes related to the metabolic syndrome.

Although the heterogeneity of variance across sexes in univariate analyses and the likely effects of ignoring this to the estimates of correlations between endophenotypes in selected traits using bivariate analysis were tested, the effects on whole correlation structures (13 endophenotypes) could not be easily inferred. Therefore, the present results should be viewed in the presence of this limitation.

Heterogeneity of variance components across sexes in the absence of sex-specific genetic effects for some traits could be caused by sex-specific environmental effects. One possible explanation for the sex differences in variance components estimates is that premenopausal women are largely protected against the metabolic syndrome or similar risk profiles. Although the investigation and understanding of this phenomenon will be very important, this is not addressed in the present study.

The endophenotypes related to the metabolic syndrome investigated in the present study were based on all individuals in the distribution. The underlying genetic and environmental relationships in the pertinent tail of the distribution representing the metabolic syndrome may be different to that in the normal range. The DeFries-Fulker extreme analysis (DeFries and Fulker, 1985), which is based on a multiple regression method, could be used to estimate (co)variance components of extreme scores from a tail of the distribution (Purcell and Sham, 2003; Viding *et al.*, 2003), to test whether the extremes are part of a continuum or form a discrete group. The application of this method to the endophenotypes studied could be explored, but given the large number of endophenotypes involved and the present sample size, power to distinguish between underlying discrete groups is likely to be low.

As with any other twin studies, this study should be viewed in the presence of the potential limitations, including whether twins under study are representative of the general population of singleton births. In our study, there is no reason to believe that twins are different from the general population as we have found the same prevalence of a number of diseases, e.g. Type 1 and Type 2 diabetes, thyroid diseases, skin diseases and mortality, in Danish twins as in the population (reference list available from the last author). In addition to that, twin studies assume that MZ and DZ twins share the same common environmental experiences. This assumption can be tested by comparing phenotypic similarity in twins of perceived versus true zygosity (Kendler *et al.*, 1993; Scarr, 1968). The assumption is violated if the phenotypic similarity of the twins is the result of perceived zygosity rather than the true zygosity. While this assumption has been tested empirically for some traits (Kendler *et al.*, 1993), this is not commonly practiced as part of most twin studies.

Both univariate and multivariate genetic analyses performed have shown that, except for GLU0, the heritability estimates for all 13 endophenotypes associated with the metabolic syndrome were moderate to high. These results indicate that phenotypic variation in individual endophenotypes is mostly due to genetic effects. Environmental factors that contribute to the variation of the endophenotypes between individuals appear to be mostly experienced by individuals, and not shared between family members.

Previous twin studies have also shown that phenotypic variations in most of the metabolic syndrome related endophenotypes were largely genetic in origin (Edwards *et al.*, 1997; Poulsen *et al.*, 2001; Samaras *et al.*, 1999). Those studies also reported that common environmental variance shared by family members did not contribute to the variations of the endophenotypes. Some family studies reported smaller genetic components influencing phenotypic variation in the phenotypes related to the metabolic syndrome (Freeman *et al.*, 2002; Henkin *et al.*, 2003; Martin *et al.*, 2003; Mitchell *et al.*, 1996), but the components were still significantly different from zero.

By partitioning the phenotypic covariance between endophenotypes into underlying genetic and environmental components, the multivariate genetic analyses have been useful for understanding the relationship among the endophenotypes. The results showed that the genetic and environmental correlations were strong only between endophenotypes assigned to the same group, but weak to moderate between endophenotypes assigned to different groups. The strong genetic correlations between endophenotypes in the same group indicate that these endophenotypes have genes in common. The findings may also be due to a direct causal physiologic relationship between endophenotypes assigned to the same group, e.g. insulin-glucose regulation and its consequences. Multivariate genetic analysis has also shown that environmental factors common to a pair of endophenotypes were specific to individuals rather than shared by family members.

While previous studies have suggested common underlying factors influencing the endophenotypes related to the metabolic syndrome (Hong *et al.*, 1997; Li *et al.*, 2006), genetic, environmental and phenotypic correlations between groups of endophenotypes estimated from the present study did not entirely support this finding. The correlations between the groups of endophenotypes were mostly weak, except moderate correlations between insulin related endophenotypes with either obesity or lipid endophenotypes. However, since insulin related endophenotypes and obesity have been suggested as the main factors underlying the metabolic syndrome, moderate correlations between these components may still indicate some common underlying genetic and environmental effects between the main components of the metabolic syndrome.

The genetic and environmental correlations estimated from the present study are mostly in line with previous studies (Hong *et al.*, 1997; Martin *et al.*, 2003; Nelson *et al.*, 2000; Perusse *et al.*, 1997; Rainwater *et al.*, 1997; Samaras *et al.*, 1999; Tregouet *et al.*, 1999). Moderate correlations between insulin related endophenotypes and obesity traits have previously been reported (Nelson *et al.*, 2000; Samaras *et al.*, 1999; Tregouet *et al.*, 1999). For example, using 110 female twin pairs, Samaras *et al.* (1999) have reported a genetic correlation of 0.41 between insulin resistance and central fat, while Nelson *et al.* (1997) reported a higher genetic correlation of 0.64 between waist to hip ratio (WHR) and fasting insulin. The correlations between blood pressure and other metabolic syndrome endophenotypes have been reported as weak (Hong *et al.*, 1997; Tregouet *et al.*, 1999), while Mitchell *et al.* (1996) have reported a near zero genetic correlation between fasting insulin with either systolic or diastolic blood pressure. The correlations between the other metabolic syndrome endophenotypes have also been reported as weak or moderate (Martin *et al.*, 2003; Samaras *et al.*, 1999; Tregouet *et al.*, 1999; Rainwater *et al.*, 1997; Perusse *et al.*, 1997). Although the correlations between endophenotypes reported in those studies were similar to the present study, these correlations have mostly been interpreted as strong evidence for a common underlying factor influencing the metabolic syndrome.

Knowing that a large proportion of phenotypic variance of individual endophenotypes of the metabolic syndrome traits is explained by genetic factors, finding genes responsible for these is the next challenge. Some studies have reported quantitative trait loci (QTL) responsible for the cluster of the metabolic syndrome or its components (An *et al.*, 2005; Kissebah *et al.*, 2000; Rich *et al.*, 2005; Shearman *et al.*, 2000). The BEACON gene located on chromosome 19p has been reported to be associated with the metabolic syndrome related endophenotypes (Jowett *et al.*, 2004). Goldin *et al.* (2003) summarised 12 studies reporting QTL related to the metabolic syndrome components and the QTL were located mostly on chromosomes 1, 3, 5, 6, 7, 10, 14 and 17. Some studies have tried to locate genes responsible for the metabolic syndrome as a composite variable (Loos *et al.*, 2003; McQueen *et al.*, 2003; Ng *et al.*, 2004; Tang *et al.*, 2003). Given that genetic correlations between the endophenotypes of metabolic syndrome were weak to moderate as suggested from this study, finding genes for the metabolic syndrome as a composite variable may not be a good strategy. This strategy may be appropriate only for endophenotypes within groups, where the genetic correlations were mostly high. Results from this study suggest that finding genes responsible for each component of the endophenotypes of the metabolic syndrome separately may be a better option.

In conclusion, while the individual endophenotypes of the metabolic syndrome were highly heritable, weak to moderate genetic correlations and no significant common environmental correlations between endophenotypes assigned to different groups suggest that the metabolic syndrome comprises a composite set of endophenotypes that apparently do not share a substantial common genetic and familial environmental background in the general population. However, moderate genetic and specific individual environmental correlations between fasting insulin and obesity endophenotypes and a moderate specific individual environmental correlation between fasting insulin and lipids endophenotypes indicated that some common underlying genetic or environmental variations may be shared between fasting insulin with either obesity endophenotypes or lipids endophenotypes. For the metabolic syndrome to be useful in clinical practices, much more study is still needed. These include understanding whether the metabolic syndrome is a unified syndrome with known underlying pathophysiology, the endophenotypes included or excluded, the value of diagnosing patients with the syndrome and the treatment itself (Kahn *et al.*, 2005).

6 A Genome-wide Linkage Analysis: Mapping Chromosomal Regions Influencing The Variation of Body Height

Abstract

Body height is an important anthropometric measure and an excellent model for studying the genetic architecture of complex traits. Genetic factors have been reported to account for most of the variation in body height, with a heritability estimated around 0.8. Previous linkage studies have identified several quantitative trait loci (QTL) that influence the variation of body height, mostly with moderate statistical support. The main objective of the present study is to identify chromosomal regions that influence the variation of body height in the Australian population. This study also provides a means to replicate previously reported QTL for body height. The subjects of the study were drawn from two Australian twin cohorts (Adolescent and Adult) consisting of 871 and 7,007 individuals, respectively. The combined data provided a total of 5,419 quasi-independent sib-pairs derived from 7,876 individuals in 2,628 families. A variance component linkage analysis revealed several chromosomal regions suggestive for linkage with body height, including 3q22.1 (LOD = 2.1) and 5q32 (LOD = 2.1). Sex-specific linkage analyses suggested that 1q32 (LOD = 1.9) and 15q23 (LOD = 1.9) were suggestive for linkage with body height in males, while 7p21.1 (LOD = 1.9) was only suggestive in females. Despite the relatively large sample size, the moderate statistical support for most of the identified chromosomal regions suggests that body height is influenced by several or many genes, each having a modest effect.

6.1 Introduction

Body height is one of the most important anthropometric measures. It has been suggested as an indicator of childhood living conditions (Peck and Lundberg, 1995) and has been associated with the risk of coronary heart diseases (McCarron *et al.*, 2002; Silventoinen *et al.*, 2006). Body height has also been associated with intelligence (Abbott *et al.*, 1998; Sundet *et al.*, 2005), educational attainment (Magnusson *et al.*, 2006; Silventoinen *et al.*, 2004) and longevity (reviewed by Samaras *et al.*, 2003).

Body height has also been considered as an excellent model trait for studying the genetic architecture of complex quantitative traits (Hirschhorn *et al.*, 2001; Visscher *et al.*, 2006). It is a normally distributed quantitative trait and highly heritable. The early report by Pearson and Lee (1903), which suggested that genetic factors were important sources of variation for body height, has been consistently supported by many twin and family studies. In these studies, the heritability of height was estimated to be about 0.8 (reviewed by Silventoinen, 2003a).

In order to identify the causal genes/polymorphisms underlying variation in body height, different strategies have been employed. From the results of segregation analyses, it has been suggested that several major genes (Ginsburg *et al.*, 1998), including major recessive genes (Xu *et al.*, 2002) influence the variation of body height. As suggested from association and candidate gene studies, several genes/polymorphisms have also been associated with body height. These include the short-stature homeobox-containing gene (SHOX) (reviewed by Rappold *et al.*, 2005), vitamin D receptor gene (D'Alesio *et al.*, 2005; Lorentzon *et al.*, 2000; Remes *et al.*, 2005; Xiong *et al.*, 2005), collagen type 1 alpha

1 gene (COL1A2) (Lei *et al.*, 2005), estrogen receptor alpha gene (ESR1) (Lei *et al.*, 2005; Schuit *et al.*, 2004), cytochrome p450 19 gene (CYP19) (Ellis *et al.*, 2001; Yang *et al.*, 2006), adiponectin receptor 1 gene (ADIPOR1) (Siitonen *et al.*, 2006), low density lipoprotein receptor-related protein 5 gene (LRP5) (Ferrari *et al.*, 2004), dopamine D2 receptor gene (DRD2) (Miyake *et al.*, 1999), peroxisome proliferator-activated receptor-γ gene (PPARγ) (Meirhaeghe *et al.*, 2003), BCHE locus of butyrylcholinesterase (Souza *et al.*, 2005) and the parathyroid hormone type 1 receptor gene (PTHR1) (Scillitani *et al.*, 2006).

Another popular approach for gene identification, linkage analysis, has been very successful in identifying single/major gene disorders (e.g. Burton *et al.*, 2005). However, its success for identifying loci affecting complex quantitative traits has been limited due to several factors, including low power because of the very large sample size needed to detect quantitative trait loci (QTL) with modest effect, inaccurate phenotyping, and low heritability (Palmert and Hirschhorn, 2003; Risch and Merikangas, 1996; Teare and Barrett, 2005). By avoiding the limitations of other complex traits, body height has been suggested as being amenable to linkage analysis for several reasons: 1) height has a very high heritability (about 0.8 in Caucasian populations); 2) it can be measured easily and accurately; 3) a large sample size with information on height can usually be obtained by combining comparable data across studies (Hirschhorn *et al.*, 2001; Palmert and Hirschhorn, 2003). In fact, results from previous studies suggested that chromosomal regions influencing body height can be mapped using a linkage approach (Beck *et al.*, 2003; Dempfle *et al.*, 2006; Geller *et al.*, 2003; Hirschhorn *et al.*, 2001; Liu *et al.*, 2006; Mukhopadhyay and Weeks, 2003; Perola *et al.*, 2001; Sale *et al.*, 2005; Sammalisto *et al.*, 2005; Shmulewitz *et al.*, 2006; Willemsen *et al.*, 2004; Wiltshire *et al.*, 2002; Wu *et al.*, 2003; Xu *et al.*, 2002). To date, numerous QTL for body height have been mapped to almost all chromosomes (Table 6.1), but with the exception of a few regions from a small number of studies, most of them were mapped with only moderate statistical support (i.e. LOD score < 3). This suggests that most of the genes influencing height have small to moderate effects.

For linkage analysis to be able to detect QTL with small to moderate effects, a large sample size is required (Risch and Merikangas, 1996). Therefore, using a large sample size of 7,876 individuals from 2,628 families, the present study aimed to detect chromosomal regions influencing body height, including those with smaller effect sizes that may be missed by studies using smaller samples. In addition, this study also provides an opportunity to replicate previously reported chromosomal regions affecting body height.

6.2 Subjects and Methods

6.2.1 Subjects

The subjects of the present study were drawn from two different cohorts, an adolescent and an adult cohort. The younger cohort constituted adolescent twins who participated in various studies conducted by the Genetic Epidemiology Unit at the Queensland Institute of Medical Research. These included a study of melanoma risk factors on 12 and 14 year old twins (Zhu *et al.*, 2004) and a study of cognition in 16 year old twins and their siblings (Wright and Martin, 2004). For these adolescents, height measured by a nurse using a stadiometer was available from 1,575 individuals. The adult cohort constituted twins and their families who had been drawn from the Australian Twin Registry for various studies, including asthma and allergy (Duffy *et al.*, 1998; Ferreira *et al.*, 2005),

Table 6.1: Summary of QTL affecting body height reported in previous studies where multipoint LOD score ≥ 1.9 [suggestive linkage (Lander and Kruglyak, 1995)].

Chromosome position	Peak markers	cM (interval)	LOD	Population	No. Individuals/ No. Families	Reference
1p21	D1S1631	125.75	4.25 (males only)	Botnia and Helsinki	659/277	Sammalisto et al. 2005
1p21.1	D1S1631	136 (115-143)	2.25	African American, Genoa	611/na	Wu et al. 2003
1p31	D1S1728	106.5 (100 - 120)	55.3 - 79.4 (L-score)	Island of Kosrae, Federated States of Micronesia	2188/na	Shmulewitz et al. 2006
1q12	D1S3723	140	2.02	Germany	184/92	Dempfle et al. 2006
2q11.2	D2S113	104	2.23	Botnia	379/58	Hirschhorn et al. 2001
3p14.2	D3S1766	72	2.31	Finland	702/183	Hirschhorn et al. 2001
3p26.1	D3S1297-D3S1304	8.9	3.17	UK	1377/573	Wiltshire et al. 2002
3q23	D3S1764	146 (134-158)	2.03	Combination of populations	6752/2508	Wu et al. 2003
3q26.1	D3S1763	175 (160-181)	2.06	European American, GENOA	749/na	Wu et al. 2003
4q25	D4S1564	108	2.28	Botnia	379/58	Hirschhorn et al. 2001
4q35	D4S426	202.69	2.18	Botnia and Helsinki	1417/277	Sammalisto et al. 2005
4q35.2	D4S3051-D4S426	201	1.89	Botnia	379/58	Hirschhorn et al. 2001
5p14.3-p13.3	D5S2845-D5S1470	36.25-45.34	2.04	Netherlands	513 sib-pairs/174	Willemsen et al. 2004
5q31	D5S2115	144	2.14	Caucasian (US)	671/53	Deng et al. 2002
5q31.1	D5S816	137 (127-178)	2.26	HyperGen European American	1100/na	Wu et al. 2003
5q34	D5S1471	160-qter	33.9 (L-score)	Island of Kosrae, Federated States of Micronesia	2188/na	Shmulewitz et al. 2006
6q12	D6S1053	78 (62-94)	2.66	GENOA European American	749/na	Wu et al. 2003
6q12-q14.1	D6S1053-D6S1031	80.45-88.63	2.32	Netherlands	513 sib-pairs/174	Willemsen et al. 2004
6q25	D6S2436	154.64	3.06	Netherlands	1184/200	Xu et al. 2002
6q25.3	D6S1007	159	3.85	Botnia	379/58	Hirschhorn et al. 2001

Table 6.1: Continued.

Chromosome position	Peak markers	cM (interval)	LOD	Population	No. Individuals/ No. Families	Reference
6q27	D6S503	200-201	2.45	Framingham Heart Study	2656/346	Geller et al. 2003
7q11-21	D7S669-D7S630	103.1	2.26	United Kingdom	1377/573	Wiltshire et al. 2002
7q35	D7S2195	150	3.40	Sweden	683/179	Hirschhorn et al. 2001
7pter	D7S2439-D7S1523	163.74-165.18	2.91	Finland	580/247	Perola et al. 2001
7qtel	D7S3058	174 (176-182)	2.46	Combination of populations	6758/2508	Wu et al. 2003
8	na	135 (126 - 149)	1.92	African American	580/221	Sale et al. 2005
8q24.2-q24.3	D8S1100-D8S373	159	2.52	Finland	702/183	Hirschhorn et al. 2001
9p21.1	D9S1868	42	2.01	Botnia	379/58	Hirschhorn et al. 2001
9p24	D9S2169	12.55	2.57 (males only)	Botnia and Helsinki	659/277	Sammalisto et al. 2005
9q21.1	D9S301	66.32	2.09	Netherlands	1184/200	Xu et al. 2002
9q22	GATA81C04M - ATA18A07M	97 (89 -104 cM)	4.34	Caucasian (US)	3726/434	Liu et al. 2006
9q34.2-q34.3	D9S1818-D9S1826	150.92-159.61	2.61	Finland	580/247	Perola et al. 2001
10q21	GATA121A08	88.5 (80-95)	188.8 (L-score)	Island of Kosrae, Federated States of Micronesia	2188/na	Shmulewitz et al. 2006
10q23.1	D10S1686	105.04	1.93	United Kingdom	1377/573	Wiltshire et al. 2002
10na-q26.3	D10S1248-D10S212	165.27-170.94	1.94	Netherlands	513 sibs/174	Willemsen et al. 2004
11p15.4	D11S2362-D11S1999	11	2.57	Sweden	683/179	Hirschhorn et al. 2001
12p13.3	D12S341	0.62	2.07	Finland	702/183	Hirschhorn et al. 2001
12q11	D12S1301	55-58	2.31	Germany	184/92	Dempfle et al. 2006
12q13.1	D12S1090-D12S398	56	3.35	Finland	702/183	Hirschhorn et al. 2001
12q15	D12S375	80.52	1.86	Netherlands	962/200	Xu et al. 2002
13q12	D13S221	16.26	2.66 (females only)	Botnia and Helsinki	1417/277	Sammalisto et al. 2005
13q33.1	D13S779-D13S797	82.93-na	3.56	Finland	702/183	Hirschhorn et al. 2001

Table 6.1: Continued.

Chromosome position	Peak markers	cM (interval)	LOD	Population	No. Individuals/ No. Families	Reference
14q11.2	GATA74E02A	0-22	2.38	Framingharm Heart Study, Massachusetts	2885/330	Beck et al. 2003
14q23.1	D14S592	67 (58-92)	3.67	GENOA European American	749/na	Wu et al. 2003
14q32.2	D14S1426	13	2.01	Framingharm Heart Study, Massachusetts	4693/330	Mukhopadhyay and Weeks 2003
15	D15S642	110-qter	43.7 (L-score)	Island of Kosrae, Federated States of Micronesia	2188/na	Shmulewitz et al. 2006
15q12	D15S1002	15.6	1.90	UK	1377/573	Wiltshire et al. 2002
15	na	35 (23 - 44)	2.61	African American	580/221	Sale et al. 2005
17q21.3	D17S958	66	2.69	Botnia	379/58	Hirschhorn et al. 2001
18q21	D18S60	86.76	2.39 (males only)	Botnia and Helsinki	1417/277	Sammalisto et al. 2005
18q21.3 - 18q22.1	D18S1270 - D18S1364	2 and 8 cM	1.99 - 2.26	Framingharm Heart Study, Massachusetts	4692/330	Mukhopadhyay and Weeks 2003
19p11	D19S250	46.5	28.5 (L-score)	Island of Kosrae, Federated States of Micronesia	2188/na	Shmulewitz et al. 2006
20q13.1	D20S96	58.48	2.51	Botnia	379/58	Hirschhorn et al. 2001
22q13	D22S282	50.81	2.85	Botnia and Helsinki	1417/277	Sammalisto et al. 2005
22centr	D22S420	0	1.95	Sweden	683/179	Hirschhorn et al. 2001
Xp22.2	DXS1060	15.12	1.95	Nebraska	671/53	Deng et al. 2002
Xp24	DXS1001	75.79	1.91	Nebraska	671/53	Deng et al. 2002
Xp22	AGAT144	11.3 cM pter	5.36 (two point LOD)	Caucasian (US)	3726/434	Liu et al. 2006
Xq24	GATA165B12P	133 cM pter	5.63 (two point LOD)	Caucasian (US)	3726/434	Liu et al. 2006

anxiety and depression (Kirk *et al.*, 2000), alcoholism (Heath *et al.*, 1997) and the factors for cardiovascular disease (Beekman *et al.*, 2003). A study of body mass index from this cohort has been extensively described by Cornes *et al.* (2005). From the adult cohort, clinical measurement and/or self-reported height were available from 36,427 unique individuals.

For the adult cohort, more than one measurement was available for some individuals due to their participation in multiple studies. Height discrepancies over time and extreme measures were checked carefully against original records and those which differed more than 3 *cm* from other height measures across studies were not included. Following Cornes *et al.* (2005), rules were implemented to select the most accurate measurement for further analysis. Briefly, if a clinical measurement was available for an individual, it was used for the analysis. Out of 36,427 individuals, clinical measurements were available for 5,129 individuals (14.1%). If there was more than one measurement, the most recent measurement was chosen (see Figure 6.1).

6.2.2 Genotyping

In both adolescent and adult cohorts, microsatellite marker genotypes were available from several genome-wide and fine mapping studies. For the adolescent cohort, a detailed description of genotyping, cleaning and merging of each of the smaller genome scans has been provided by Zhu *et al.* (2004). For the adult cohort, these details were described by Cornes *et al.* (2005). In addition, further genotypic data from a subset of individuals previously described by Cornes *et al.* (2005) and 4,575 new individuals from the adult cohort have been incorporated into the present study (Luciano *et al.*, 2006). The cleaning and integration procedures described by Cornes *et al.* (2005) were also applied when adding new

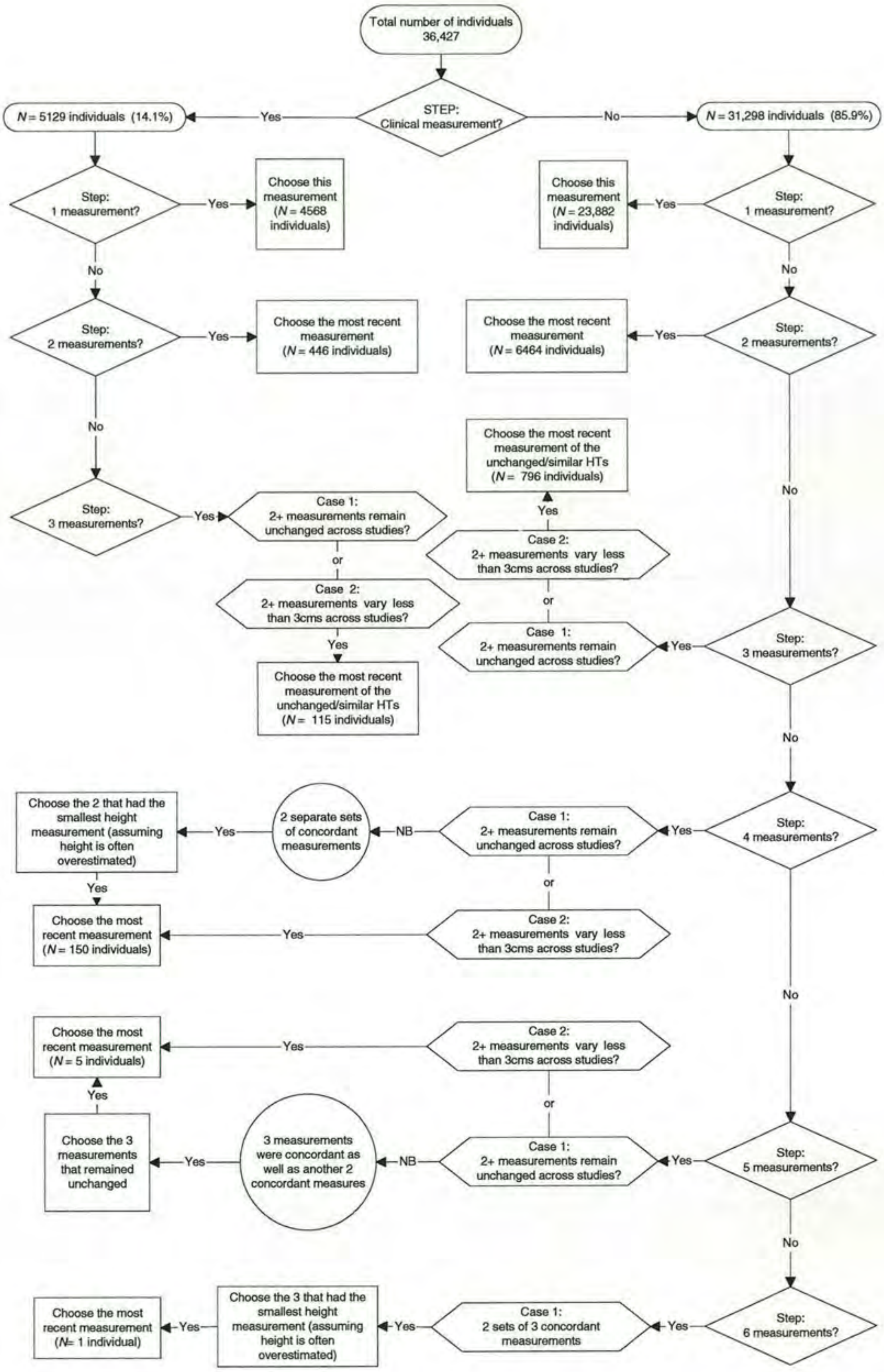
Total number of individuals 36,427
STEP: Clinical measurement?
Yes
No
N = 5129 individuals (14.1%)
N = 31,298 individuals (85.9%)
Step: 1 measurement?
Yes
Choose this measurement (N = 4568 individuals)
No
Step: 2 measurements?
Yes
Choose the most recent measurement (N = 446 individuals)
No
Step: 3 measurements?
Yes
Case 1: 2+ measurements remain unchanged across studies?
or
Case 2: 2+ measurements vary less than 3cms across studies?
Yes
Choose the most recent measurement of the unchanged/similar HTs (N = 115 individuals)
Step: 1 measurement?
Yes
Choose this measurement (N = 23,882 individuals)
No
Step: 2 measurements?
Yes
Choose the most recent measurement (N = 6464 individuals)
No
Step: 3 measurements?
Yes
Case 1: 2+ measurements remain unchanged across studies?
or
Case 2: 2+ measurements vary less than 3cms across studies?
Yes
Choose the most recent measurement of the unchanged/similar HTs (N = 796 individuals)
No
Step: 4 measurements?
Yes
Case 1: 2+ measurements remain unchanged across studies?
NB
2 separate sets of concordant measurements
Yes
Choose the 2 that had the smallest height measurement (assuming height is often overestimated)
Yes
Choose the most recent measurement (N = 150 individuals)
or
Case 2: 2+ measurements vary less than 3cms across studies?
Yes
No
Step: 5 measurements?
Yes
Case 1: 2+ measurements remain unchanged across studies?
NB
3 measurements were concordant as well as another 2 concordant measures
Yes
Choose the 3 measurements that remained unchanged
Yes
Choose the most recent measurement (N = 5 individuals)
or
Case 2: 2+ measurements vary less than 3cms across studies?
Yes
No
Step: 6 measurements?
Yes
Case 1: 2 sets of 3 concordant measurements
Yes
Choose the 3 that had the smallest height measurement (assuming height is often overestimated)
Yes
Choose the most recent measurement (N= 1 individual)

Figure 6.1: Schematic diagram of the cleaning procedure to obtain the most accurate measurement of height from the adult cohort.

genotypic data to the adult cohort.

Briefly, the procedure of merging the smaller genome scans into a combined genome scan can be described as follows. Firstly, raw data from each genome scan were integrated into one combined genome scan. Some individuals were genotyped using the same markers in different genome scans. Collapsing these duplicate markers into one unique marker was not an option since the same markers from different genome scans were genotyped using different primers, allele calling algorithms and measurement technologies. Thus, these markers were given unique identifiers and their map positions were separated by 0.001 cM to avoid zero-spacing. Secondly, for any marker, if Mendelian inheritance errors were detected at a given marker, all genotypes for that marker for a given family were removed. This procedure was performed by implementing the algorithm developed by O'Connell and Weeks (1999) for detecting and removing genotypes with Mendelian inconsistency. Finally, unlikely chromosomal recombination patterns were detected and removed using the −−error option in the Merlin package (Abecasis *et al.*, 2002).

6.2.3 Subjects with Phenotypic and Genotypic Data

For linkage analysis, the descriptive statistics for subjects with both phenotypic and genotypic information are summarised in Table 6.2. Subjects were selected if they were: 1) $\geq$ 16 years of age; 2) genotyped with more than 210 autosomal markers (i.e. the maximum number of autosomal markers from the smallest genome scan); 3) less than 4 standard deviations above or below the mean, after sex and age adjustments. In addition, bivariate (within family) outliers were also detected and removed. These outliers were defined as any sibling pair in a family for whom the height squared difference (after sex and age adjustments) was more than 4 standard deviations above the mean (see Appendix 3). If bivariate outliers were detected in families with more than 2 siblings, not all individuals in a pair were removed, but only the individuals causing the bivariate outliers (59 individuals). Subjects younger than 16 years old were mostly in their adolescent growth stage, so to minimize heterogeneity of the phenotypic definition, only subjects $\geq$ 16 years of age were included in the analyses.

The adolescent cohort constituted predominantly DZ twin pairs since these have been targeted for genotyping. From 871 individuals in 383 families, there were 519 possible pairings of sibs (quasi-independent sib-pairs, QISPs). On the other hand, additional family relationships were available for the adult cohort, including half-sibs, cousins, avuncular and grandparent-grandchild pairs. From 7,007 individuals in 2,245 families, there were 4,905 QISPs, 57 half-sibs, 8 cousins, 138 avuncular and 10 grandparent-grandchildren relationships. Hence, nearly all information on linkage comes from sibling pairs. In addition to DZ twin pairs and additional siblings, there were 58 and 165 MZ twin pairs in the adolescent and adult cohorts, respectively. In the calculation of the number of informative relative pairs, only one individual from an MZ pair was included. In total, the linkage analysis performed in the present study utilised 7,876 individuals with genotypic and phenotypic information who derived from 2,628 families. A description of pedigree structures is presented in Table 6.3.

6.2.4 Power Calculations

To assess the power of the present study to detect a QTL influencing the normal variation in body height, a series of genetic power calculations based on Sham *et al.* (2000), was performed using a genetic power calculator developed by Purcell *et al.* (2003). The power calculations were performed under the

Table 6.2: Descriptive statistics of subjects with phenotypic and genotypic information after removing univariate and bivariate outliers.

Variables	Cohort	Sex	N (Subjects)	Mean (SD)	Range
Age	Adolescent	Total	871	16.41 (0.72)	16 - 22
	Adult	Total	7007	41.46 (15.26)	16 - 90
	Combined	M	3165	38.42 (16.85)	16 - 88
		F	4711	38.86 (16.12)	16 - 90
		Total	7876	38.68 (16.41)	16 - 90
Number of autosomal markers /individual	Adolescent	Total	871	562	226 - 755
	Adult	Total	7007	544	211 - 1640
	Combined	M	3165	542	218 - 1617
		F	4711	549	211 - 1640
		Total	7876	546	211 - 1640
Height (cm)	Adolescent	Total	871	170.14 (8.38)	148- 195
	Adult	Total	7007	169.27 (9.78)	135 - 203
	Combined	M	3165	177.85 (6.86)	152 - 203
		F	4711	163.67 (6.55)	135 - 188
		Total	7876	169.37 (9.64)	135 - 203
Standardised residual of height (adjusted for sex and age)	Adolescent	Total	871	0.01 (0.97)	-2.66 - 2.88
	Adult	Total	7007	-0.01 (0.95)	-3.87 - 3.81
	Combined	M	3165	0.01 (0.99)	-3.76 - 3.81
		F	4711	-0.01 (0.94)	-3.90 - 3.60
		Total	7876	-0.01 (0.96)	-3.90 - 3.81

Table 6.3: Pedigree structure of individuals with phenotypic and genotypic information and after removing univariate and bivariate outliers.

Cohort	Adolescent	Adult	Combined-Male	Combined-Female	Combined-All
Families	383	2245	1897	2329	2628
Individuals[a]	813	6842	3102	4551	7653
QISPs[a]	519	4905	906	1946	5419
Half-sibs[a]	-	57	10	20	57
Cousins[a]	-	8	2	1	8
Grandparent-grandchild[a]	-	10	1	3	10
Avuncular[a]	-	138	28	45	138

Note: [a]Only one individual from each MZ pair was included in this calculation. QISPs is quasi independent sib pairs.

following assumptions: 1) an additive QTL is influencing the variation of body height; 2) recombination rate between the marker and the QTL is 0; 3) a heritability of 0.9 [estimated by Macgregor *et al.* (2006)]; 4) type 1 error rate (a) = 0.0001 (equivalent to LOD score of 3). The expected LOD score for the combined cohort was also calculated using the above assumptions as E(LOD) = (1 + NCP)/4.605, with NCP the QTL non-centrality parameter (Sham *et al.*, 2000).

6.2.5 Linkage Analysis

To identify and map chromosomal regions (quantitative trait loci, QTL) influencing variation in body height, a multipoint variance component linkage analysis was performed as implemented in Merlin 1.0.1 (for autosomes) and

MINX (for the X chromosome) programs (Abecasis *et al.*, 2002) for the adolescent, adult and combined cohorts. In addition, to examine the presence of sex-specific QTL influencing body height, separate analyses for males and females were performed in the combined cohort. This was done by including only phenotypes from one sex while retaining the genotypes from both sexes. As the framework for mapping the QTL, i.e. establishing the position and order of the markers in the chromosomes, a locally weighted linear regression map (http://www.qimr.edu.au/davidD) based on NCBI Build 35.1 physical map positions, deCODE and Marshfield maps was used (Duffy, 2006). In the linkage analyses, sex and age were fitted as covariates, except for the sex-specific analyses, where age was the only covariate.

Briefly, linkage analysis correlates the phenotypic similarity of relative pairs with their genotypic similarity (represented by the proportion of alleles shared identical by descent at a specific position in the genome) (e.g. Almasy and Blangero, 1998). The presence of a QTL at a specific chromosomal location was tested by comparing the likelihood of no QTL (H_0) with the likelihood that there is a QTL influencing the variation in body height (H_1). The LOD (log of the odds) score is defined as twice the difference between the $\log_{10}$ likelihood (H_0) and (H_1) (e.g. Almasy and Blangero, 1998). As suggested by Lander and Kruglyak (1995), the present study considered LOD score of 3.3 and 1.9 as significant and suggestive evidences of linkage between QTL and markers at the test position, respectively.

6.3 Results

6.3.1 Descriptive Statistics

Only subjects $\geq$ 16 years of age were selected in this study, so that most of them will have reached their final (body) height (Figure 6.2). The age distribution of the subjects from the adolescent cohort was different from that of the adult cohort. The subjects in the adolescent cohort were mainly 16 years of age. Some individuals in this cohort were older, but not more than 22 years. On the other hand, a wide range of age was seen in the adult cohort with a mean age of 41.5 years and a range of 16 - 83 years.

The mean height of males was larger than that of females for all cohorts, and estimated differences were 10.76 cm (SE 0.44), 14.66 cm (0.16) and 14.16 cm (0.15) for the adolescent, adult and combined cohorts, respectively. The estimated regression of age on height was 0.14 (SE 0.30), -0.07 (0.01), -0.05 (0.01) cm/year, for the adolescent, adult and combined cohorts, respectively. The negative slope for the adult cohort is consistent with Macgregor *et al.* (2006).

For all cohorts, skewness and kurtosis values of the standardised residuals of height after a general linear model adjustment for age and sex effects were very close to zero; skewness for the adolescent, adult and combined cohorts was 0.05 (SE 0.08), 0.04 (0.03) and 0.03 (0.03), respectively and kurtosis was -0.25 (SE 0.17), 0.10 (0.06) and 0.03 (0.06) were observed for the adolescent, adult and combined cohorts, respectively. This is a good indication that the data close to a normal distribution, which is an assumption of variance component linkage analysis by maximum likelihood.

The average numbers of markers genotyped per individual were 562 (range: 226

- 755) and 544 (range: 211 - 1640) microsatellite markers for the adolescent and adult cohort, respectively. For the combined cohort, the number was 542 (range: 211 - 1640). Overall, these numbers provided an average spacing of about 5 cM across the genome.

The estimated sib correlations of height adjusted for the effects of age and sex were 0.49 (SE 0.04) and 0.42 (0.01) for the adolescent and adult cohorts, respectively and for the combined cohort was 0.43 (0.01). The corresponding heritability estimates as calculated by Merlin were 0.92, 0.82 and 0.85. The same heritability of 0.89 was estimated for males and females in the combined cohort.

6.3.2 Power Calculations

In the adolescent cohort, the power to detect a QTL explaining 20% of height variation was only 6.9%, whereas the power of the adult cohort for detecting the same sized QTL was 99.9%. However, when the QTL to be detected is responsible for 10% of the variation in height, the power of the adult cohort reduced to only 38%. The power of the combined cohort to detect a QTL responsible for 10 and 5% of the phenotypic variation is 45 and 3%, respectively. For a power of 80%, the required sample size to detect a QTL explaining 10% of the phenotypic variation is about 8,768 sib-pairs. In the combined cohort, the expected LOD scores for a QTL explaining 5, 10 and 20% of the phenotypic variance were 0.9, 3.0 and 11.5, respectively. In these power calculations, the QTL and the marker are assumed to be completely linked (recombination rate = 0). For average spacing of 2.5 or 5 cM, which is commonly found in practice, the powers will be less than the above figures.

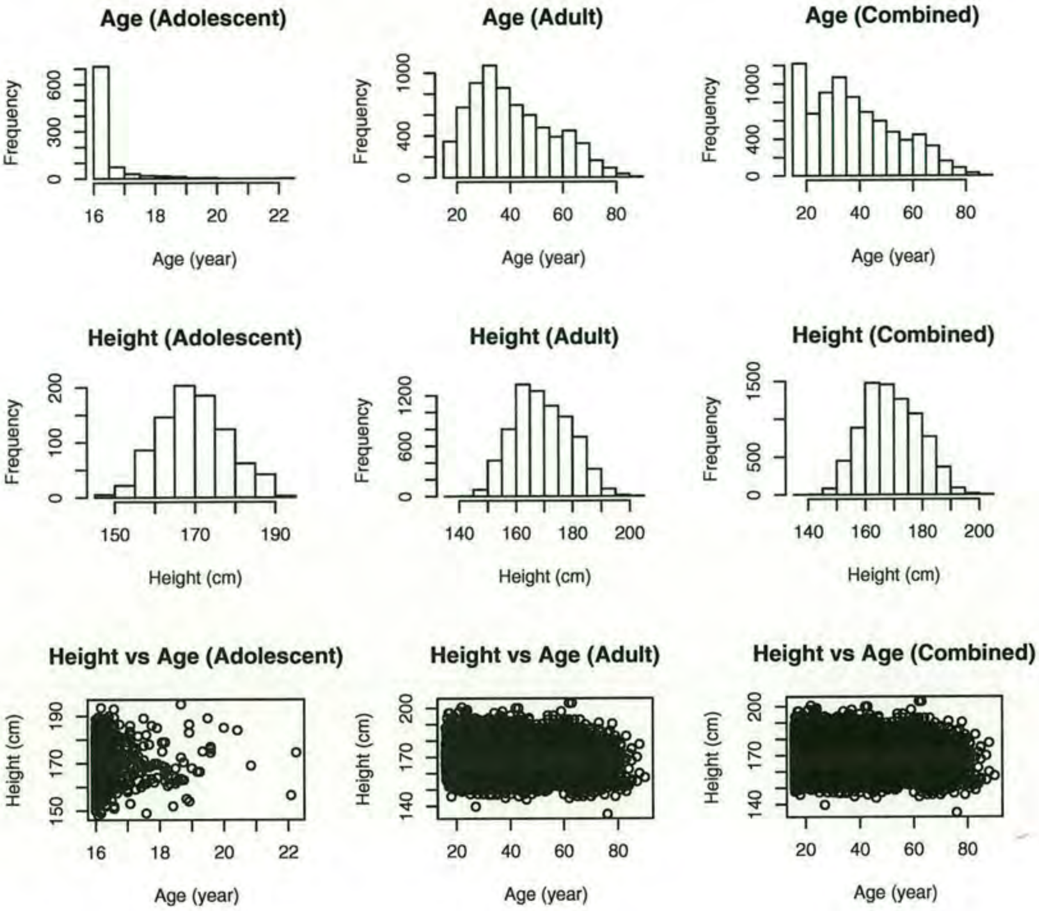

Figure 6.2: The age (top) and height (middle) distributions of the subjects from different cohorts. The plots of height for given age are presented in the bottom part of the figure.

6.3.3 Linkage Analysis

Adolescent and Adult Cohorts:

The multipoint LOD score profiles for all chromosomes are presented in Figure 6.3 and the chromosomal regions that reached a LOD score of 1.5 or greater are presented in Table 6.4. In the adolescent cohort, no chromosomal regions showed a suggestive linkage for body height. However, five chromosomal regions that reached LOD scores greater than 1 were identified on chromosomes 1, 2, 7, 8 and 9. From the adult cohort, 2 chromosomal regions [1q23.1 (LOD = 2.5); 5q32

(LOD = 1.9)] were suggestive for linkage with body height. In addition to that, 7 other chromosomal regions reached a LOD score of 1.5 or larger (Table 6.4).

There was little overlap between the results of the adolescent and adult cohorts (Figure 4A). When the two cohorts were combined, its LOD score profile was very similar to that of the adult cohort (Figure 4B), not surprising given that about 90% of the combined samples came from the adult cohort. In the combined cohort, only two chromosomal regions [3q22.1 (LOD = 2.1) and 5q32 (LOD = 2.1)] were suggestive for linkage.

The result of sex-specific linkage analysis in the combined cohort is presented in Figure 6.5. In males, two suggestive QTL were identified, including on 1q32.1 (LOD = 1.9) and 15q23 (LOD = 1.9) (Table 6.4). In females, only one such region was identified, namely 7p21.3 (LOD = 1.9).

Bivariate (Within Family) Outliers:

To guard against undue influence of a few extreme sib-pairs, all linkage analyses were performed by excluding individuals who were classified as bivariate (within family) outliers. Based on simple chi-square theory (Appendix 6A), in a given family, these outliers were defined as sibling-pairs of individuals for whom the absolute difference was 19 cm or greater. The difference of linkage results between including and excluding bivariate outliers in the combined cohort is presented in Figure 6.6. The results showed that by excluding the outliers, the LOD scores increased for most regions, except for the region on chromosome 15.

Table 6.4: Chromosomal regions showing multipoint LOD score ≥ 1.5.

Cohort	Chromosome	Adjacent Marker	Location and and CI^a (cM)	Multipoint LOD score
Adult	1q23.1	D1S1653	167 (155-174)	2.5
	3p22.1	D3S1768	59 (47-77)	1.6
	3q22.1	D3S1292	137 (128-153)	1.8
	5q32	GATA139B09	154 (141-164)	1.9
	7p21	D7S513	16 (2-24)	1.5
	8p23.1	D8S264	4 (0-18)	1.7
	12q13	D12S398	72 (41-114)	1.6
	16q21	D16S503	84 (74-92)	1.6
	20p11.23	D20S112	44 (37-76)	1.5
Combined-M	1q32.1	GGAA23C07	212 (203-231)	1.9
	2p21	D2S2259	69 (53-83)	1.5
	15q23	D15S131	70 (53-93)	1.9
	17q11.2	D17S798	59 (28-73)	1.7
Combined-F	7p21.3	GATA119B03	18 (9-26)	1.9
Combined	1q23.1	D1S1653	167 (152-191)	1.7
	3q22.1	D3S1292	137 (128-148)	2.1
	5q32	GATA139B09	150 (141-160)	2.1

Note: ^{a}CI was calculated using the one LOD drop-off method.

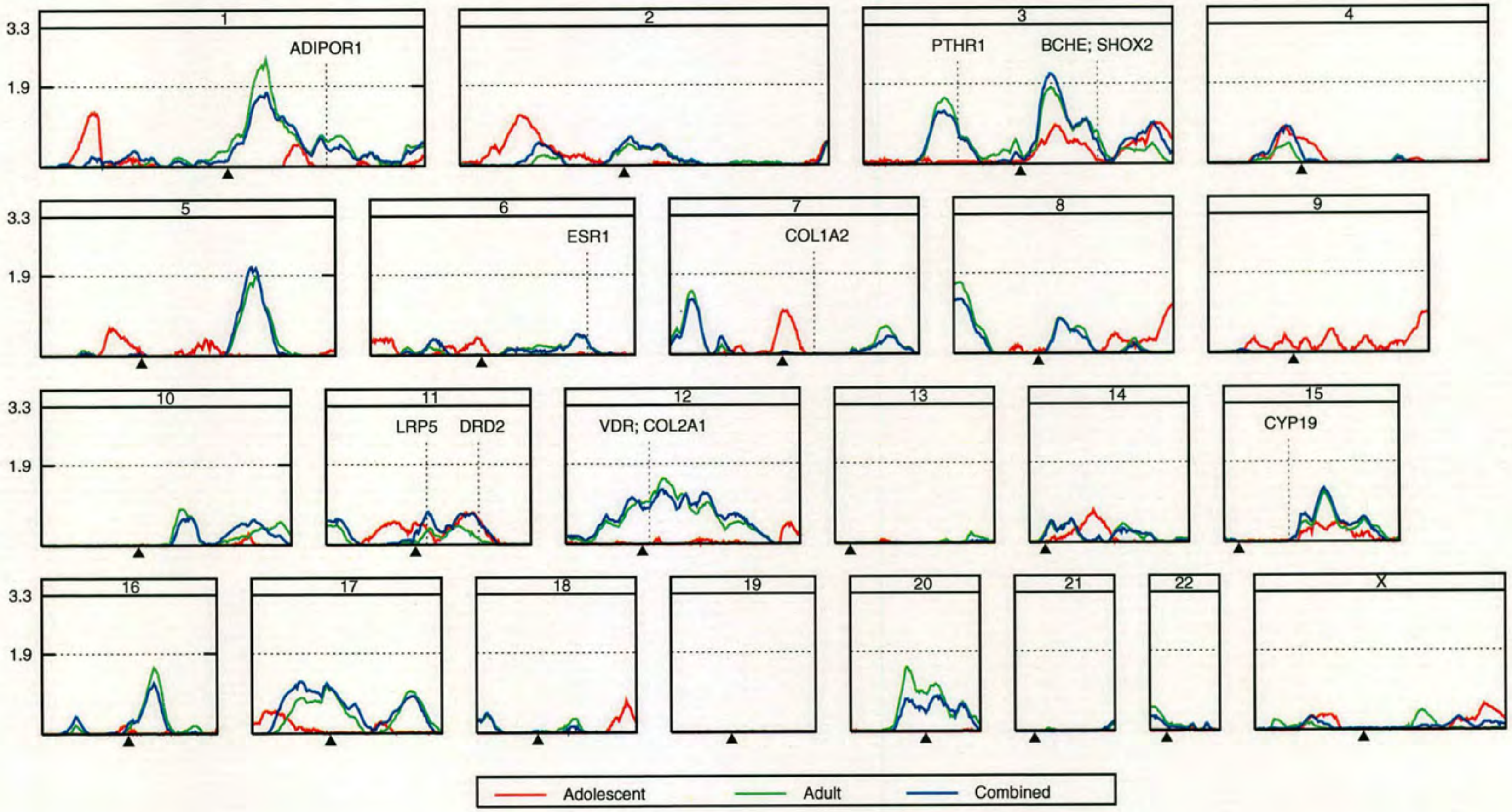

Figure 6.3: LOD scores from a genome-wide linkage analysis of body height for the adolescent, adult and combined cohorts.

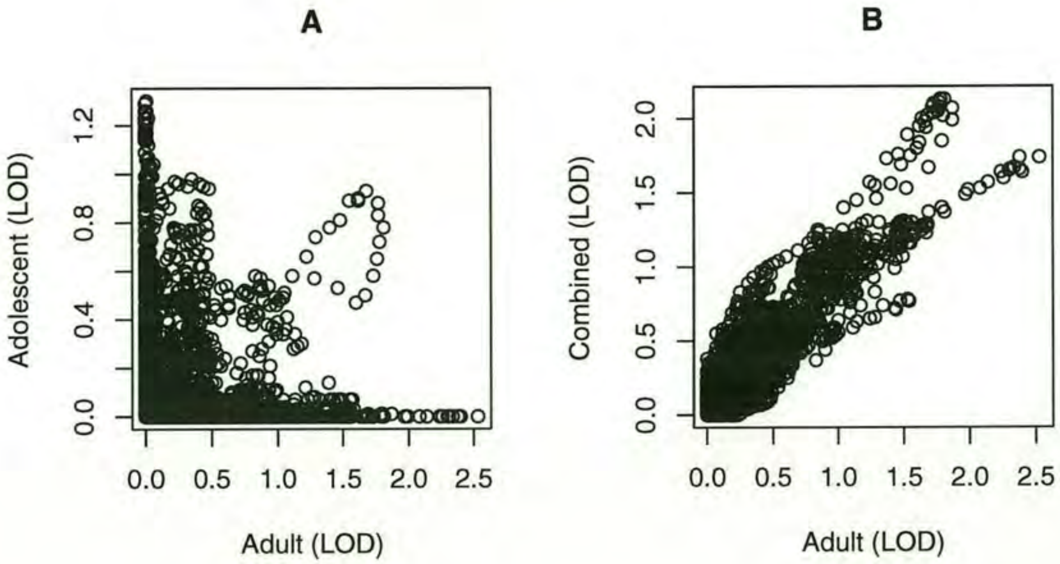

Figure 6.4: LOD scores at 1 cM grid between the adult and adolescent cohorts (**A**) and between the adult and combined cohorts (**B**).

6.4 Discussion

Genetic studies of normal variation in body height are medically important and useful for understanding the genetic architecture of complex quantitative traits. As was shown in the present study, body height is a normally distributed quantitative trait. In the view that the heritability of height is very high, this suggests that many genes influence its variation (e.g. Hirschhorn *et al.*, 2001). This genome-wide linkage analysis using a sample of 5,419 quasi-independent sib-pairs provides further evidence for the polygenic nature of body height. In the combined cohort, several chromosomal regions showed moderate linkage and two of them reached the asymptotic suggestive level of 1.9. However, despite using a large sample, none of them reached the asymptotic threshold of 3.3 indicating significant linkage. With the current sample size, power calculations indicated that if there was a QTL explaining 20% or more of the variation in body height segregating in the Australian twin sample and the QTL was

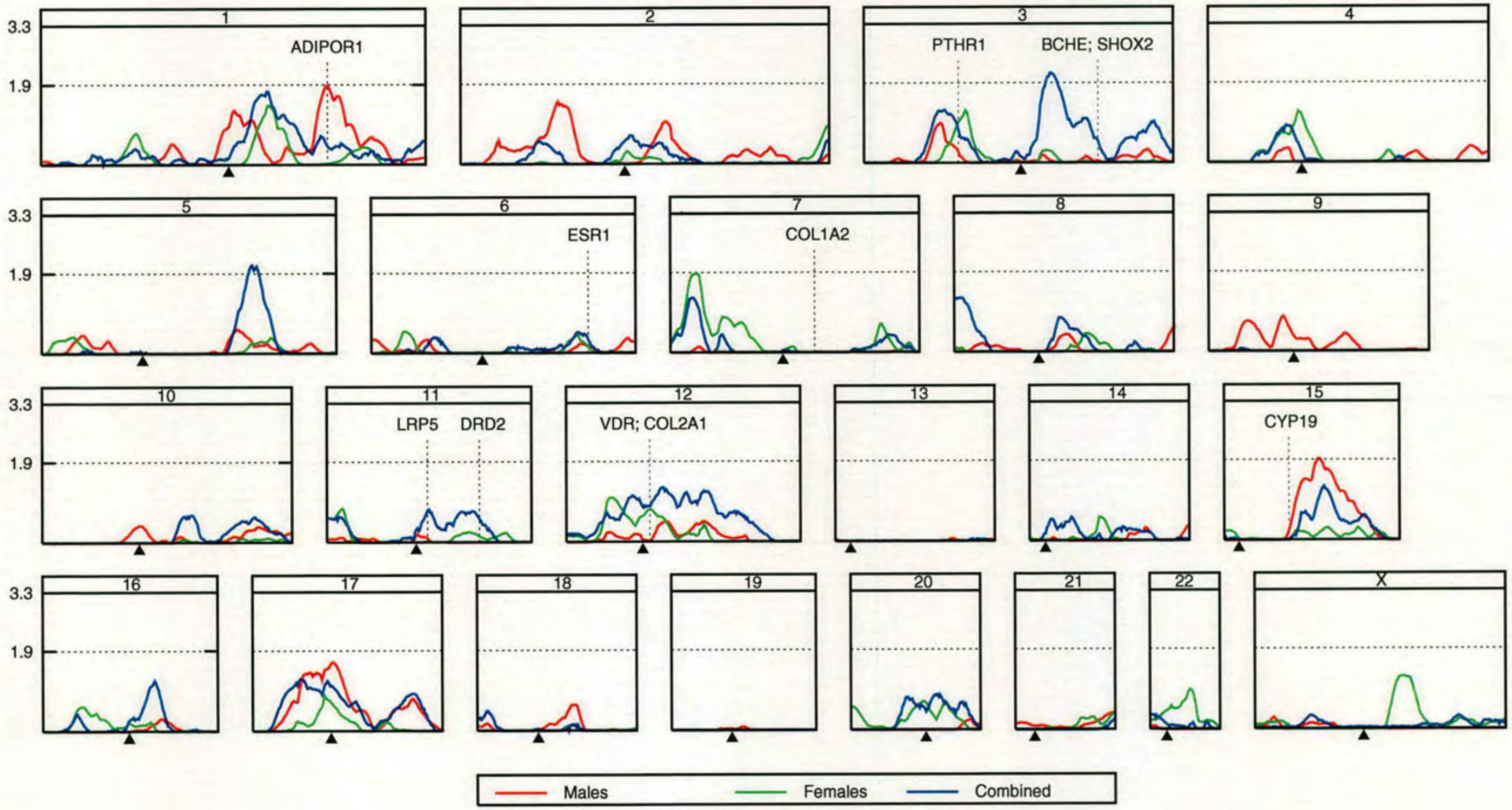

Figure 6.5: LOD scores from sex-specific genome-wide linkage analysis in the combined cohort.

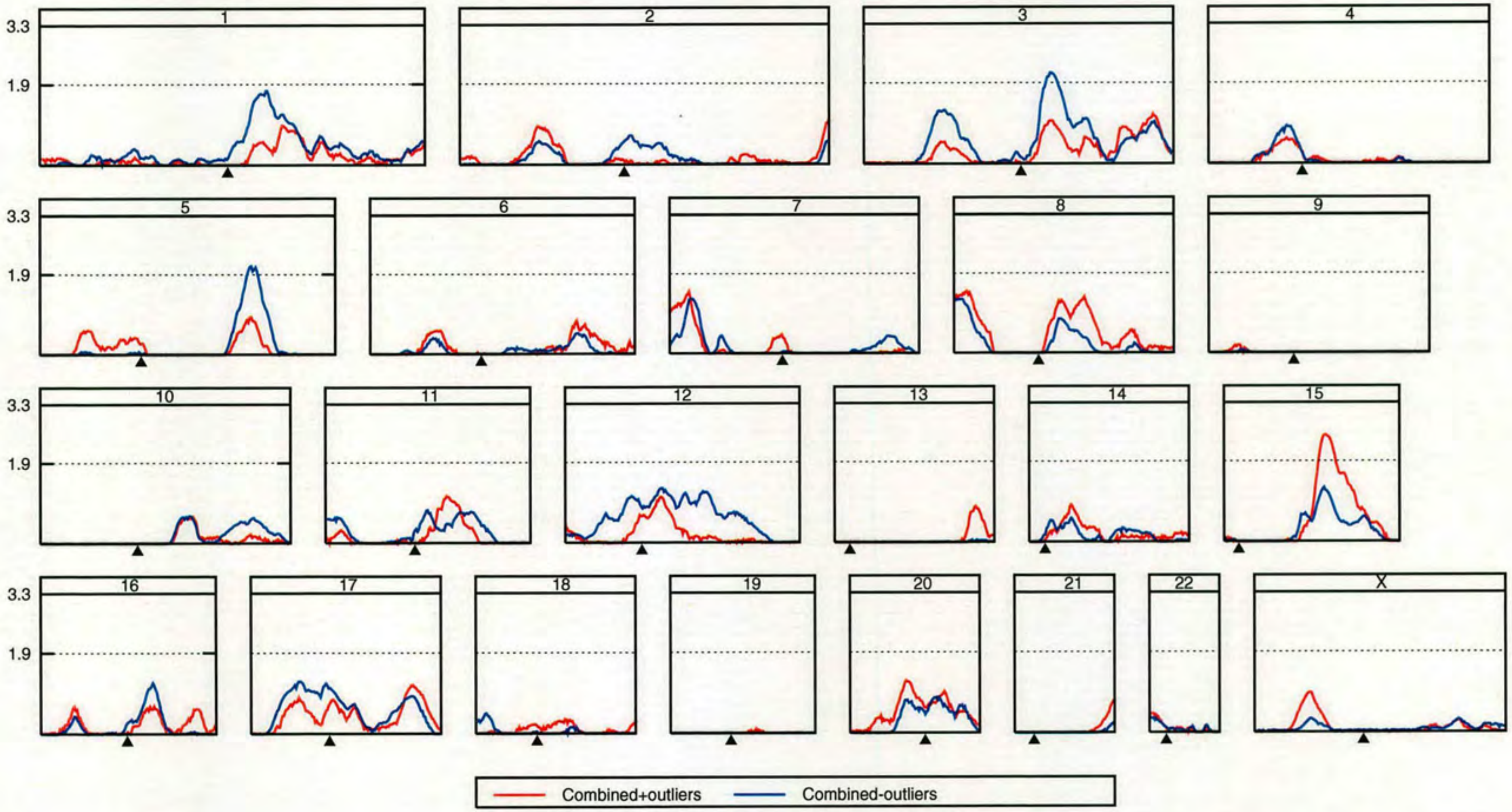

Figure 6.6: Comparison between LOD scores including and excluding bivariate outliers in the combined cohort.

completely linked to the marker, it was unlikely to be missed since the present study had 100% power to detect it. These results therefore suggest that normal variation in body height is influenced by several or many genes of small to modest effect.

The results of sex-specific linkage analyses are particularly interesting. Beside there being sex-specific QTL [1q32.1, 15q23 (males) and 7p21 (females)], the peaks on 1q23.1, 12q13 and 17q11.2 were almost equally contributed by each sex. In addition to those, two other chromosomal regions, 3q22.1 and 5q32 deserve more attention. The LOD scores from the combined samples for these regions were much greater than the sum of the LOD score from males and females. These can be explained by the fact that almost half of the samples were of the opposite-sex pairings, which were excluded from the sex-specific analyses. In other words, the LOD scores of these regions were mainly contributed by the opposite-sex pairings.

Previous studies have identified several chromosomal regions showing significant linkage (i.e. LOD score ≥ 3.3) with body height (Table 6.1). These included the regions on chromosome 1p21 (for males only) (Sammalisto *et al.*, 2005), 6q25.3 (Hirschhorn *et al.*, 2001), 7q35 (Hirschhorn *et al.*, 2001), 9q22 (Liu *et al.*, 2006), 12q13.1 (Hirschhorn *et al.*, 2001), 13q33.1 (Hirschhorn *et al.*, 2001), 14q23.1 (Wu *et al.*, 2003), and the X chromosome (Xp22 and Xq24) (Liu *et al.*, 2006). Many other studies have also supported these regions as being associated with body height, including the regions on chromosome 1 (Shmulewitz *et al.*, 2006; Wu *et al.*, 2003); chromosome 6 (Geller *et al.*, 2003; Xu *et al.*, 2002); chromosome 7 (Perola *et al.*, 2001); chromosome 9 (Xu *et al.*, 2002); chromosome 12 (Dempfle *et al.*, 2006; Sammalisto *et al.*, 2005; Xu *et al.*, 2002); and chromosome 14 (Mukhopadhyay and Weeks, 2003). Among these regions, 12q13.1 was replicated in this study. The LOD scores for this region were 1.6 and 1.3 in the adult and combined cohorts, respectively. The chromosomal regions of 3q22.1 and 5q32, which showed suggestive linkage in the combined cohort, have been previously mapped into a similar location. Using a combined population of 6752 individuals, Wu *et al.* (2003) reported that 3q23 and 5q31.1 were suggestive for linkage with body height (LOD = 2.0 for both). (Deng *et al.*, 2002) (2002) have also suggested that the region on chromosome 5 was associated with body height. Although these regions are not exactly the same, the one LOD drop-off confidence intervals overlap.

Among many genes that have been associated with body height, the vitamin D receptor (VDR) has recently received more attention. As one of the intracellular hormone receptors, the main role of VDR is to bind the active form of vitamin D. This gene was mapped to 12q12-q14 (Online Mendelian Inheritance in Man, 2006). Association studies have suggested that variation in this gene is associated with body height (e.g. D'Alesio *et al.*, 2005; Dempfle *et al.*, 2006; Lorentzon *et al.*, 2000; Remes *et al.*, 2005; Xiong *et al.*, 2005). Furthermore, a meta-analysis of published linkage studies on chromosome 12 has shown that there is significant evidence for the presence of a QTL on this chromosome (Dempfle *et al.*, 2006). In the present study, the LOD score for this region did not reach a suggestive linkage, but LOD scores of 1.6 and 1.3 were observed in the adult and the combined cohorts. For the combined cohort, power calculations suggest that the expected LOD scores for a QTL explaining 5 to 10% of the phenotypic variance are 0.9 and 3, respectively, so if the VDR has a moderate effect on height, the observed LOD scores in this region are not inconsistent with VDR being a candidate gene for height.

Other interesting genes that are located under or close to the linkage peaks were

ADIPOR1 (1q32.1), PTHR1 (3p21.3), BCHE (3q26.1) and CYP19 (15q21.2) (Figure 5). Variants in these genes have been associated with body height (Ellis *et al.*, 2001; Scillitani *et al.*, 2006; Siitonen *et al.*, 2006; Souza *et al.*, 2005; Yang *et al.*, 2006). In particular, Ellis *et al.* (2001) have reported that the association between the CYP19 gene was more evident in males than in females. As can be seen from Figure 5, the peak on chromosome 15 was male-specific. The concordance between this result and that of Ellis *et al.* provide evidence that the association between CYP19 and body height is male-specific.

Linkage analyses were performed by excluding individuals who were categorised as bivariate outliers. Within family, these individuals were discordant for body height, when the between sib absolute difference is about 19 cm. This strategy would seems to be counter-productive, because most information for linkage comes from extreme discordant sib pairs (Risch and Zhang, 1995). However, by removing these outliers, disproportionate contribution to linkage from extreme pairs, which could be due to measurement errors, is avoided.

In conclusion, a genome-wide linkage analysis has revealed several chromosomal regions suggestive for linkage with body height in a large sample of the Australian twin families. Among these regions, 3q22.1 (LOD = 2.1) and 5q32 (LOD = 2.1) reached the suggestive level for linkage. Separate linkage analyses for males and females suggested that three chromosomal regions were sex-specific. While the regions of 1q32 (LOD = 1.9) and 15q23 (LOD = 1.9) were suggestive for linkage with body height in males, 7p21.1 was suggestive in females only. Despite a large sample size, the moderate statistical support for most of the identified chromosomal regions suggests that body height is influenced by several or many genes, each having a modest effect.

Appendix 6A. A Simple Calculation for Detecting Bivariate (Within Family) Outliers

In a family consisting of sibling pairs, let y_1 and y_2 be the phenotypic (residual) values of sib$_1$ and sib$_2$, respectively. Their phenotypic difference is $D = y_1 - y_2$. By assuming that the phenotypic values are bivariate normally distributed, the expected mean and variance of the squared of D, (D^2) can be calculated using Chi-Square theory,

$$D^2 \sim 2(1-r)\sigma^2\chi^2_{(1)} \tag{6.1}$$

Where r is the sib correlation and σ is the standard deviation of the phenotypic values for y_1 and y_2. So, the mean of D^2 is $2(1-r)\sigma^2$ and the variance of D^2 is $8(1-r)^2\sigma^4$

Hence, the expected standard deviation (SD) of D^2 is $2\sqrt{2}(1-r)\sigma^2$

7 General Discussion

7.1 Summary

Understanding the genetic basis of human phenotypes is the principal goal in human genetics. One of the end results of this endeavor is to understand the biological function of genes. For complex traits, typical steps needing to be taken to achieve this goal include heritability estimation, genetic linkage, association and functional studies. For a phenotype of interest, the quest begins with understanding how the variation observed between individuals is related to its genetic variation. This step involves partitioning the phenotypic variance of a trait into underlying genetic and environmental components. After knowing that genetic factors are important sources of the observed variation between individuals, the next step is to locate the chromosomal regions influencing the trait. This step can be achieved by means of linkage mapping. The chromosomal regions identified by linkage mapping usually contain hundreds of genes. To finely map the regions or in some cases to find the actual genes, association studies can be conducted. Populations of twins are very useful to serve every step of this endeavour (see MacGregor *et al.*, 2000). The focus of this thesis is only on the first two steps, i.e. understanding the genetic variation of human quantitative traits and mapping the chromosomal regions influencing the variation of quantitative traits using samples of twins and (to some extent) their families.

It is shown in this thesis that populations of twins, with or without zygosity information, are useful for understanding the genetic and environmental sources of quantitative trait variation between individuals. If genotypic data are also available, twins can be used to identify genes/chromosomal regions affecting the variation of quantitative trait by means of linkage analysis.

In Chapters 2-4, novel applications of a finite mixture distribution model (Neale, 2003) to partition the phenotypic variance of a trait into genetic and environmental components using data from twins of unknown zygosity are presented. Variance component estimates of IQ from two whole-population surveys in Scotland, SMS1932 and SMS1947 are presented in Chapter 2. Consistent estimates of heritability of ~0.70 and shared environment (c^2) of ~0.21 were found in both surveys. The estimates decreased slightly when additional quantitative traits (height and weight) were added in a multivariate analysis. This study is the first to estimate genetic and environmental components of cognitive ability in entire school-attending populations and implies that large (national) data collections can provide sufficient information on twin pairs to estimate genetic parameters, even without known zygosity.

The precision and bias of the finite mixture distribution model were assessed using computer simulations and application to IQ measures from a large sample of twins with known zygosity (twins from the UK Twins' Early Development Studies) (Chapter 3). Simulation results showed that, if normality assumptions were satisfied and the sample size was large (e.g. 2,000 pairs), then the variance component estimates from the mixture distribution model were unbiased and the standard deviation of the difference between heritability estimates from known and unknown zygosity was in the range of 0.02 to 0.20. Unexpectedly, the estimates of heritability of 10 variables from TEDS using the mixture distribution model were consistently larger than those from the conventional (known zygosity) model. This discrepancy was due to violation of the bivariate normality assumption. A leptokurtic distribution of pair difference was observed for all traits (except non verbal ability scores of MZ twins), even when the univariate distribution of the trait was close to normality. From an independent sample of Australian twins, the heritability estimates for IQ variables were also larger for the mixture distribution model in 6 out of 8 traits, consistent with the observed kurtosis of pair difference. While the known zygosity model is quite robust to the violation of the bivariate normality assumption, the mixture distribution model produces biased estimates when the bivariate normality assumption is violated. The novel finding of widespread kurtosis of the pair difference may suggest that this assumption for analysis of quantitative trait in twin studies may be incorrect and needs revisiting. A possible explanation for widespread kurtosis within zygosity groups is heterogeneity of variance, which could be caused by genetic or environmental factors. In the mixture distribution model, the leptokurtosis-derived biases could perhaps be overcome by transforming the distribution of the pair difference to make it (near) normal, but further research is necessary to investigate whether it is feasible.

Although the finite mixture distribution model was shown to provide reliable genetic and environmental variance component estimates from twin data of unknown zygosity, the standard error of the estimates are still larger than the estimates from a conventional (known zygosity) method. One suggested way to decrease the standard error of the estimates is to analyse multiple traits simultaneously in a multivariate analysis. Additional phenotypes may provide additional zygosity classification as well as increase the effective sample size. It is subsequently shown in Chapter 4 that a multivariate analysis indeed reduces the standard error of variance component estimates. From the pattern of decreasing standard error of variance component estimates with the increase of number of traits analysed, it is hypothesized that if more than approximately 10 traits are analysed simultaneously, then the mixture distribution model provides variance component estimates with precision that are comparable to conventional analysis of known zygosity. However, further research is required to examine the behaviour of the mixture model for traits with different correlations or when the pair differences are not normally distributed.

As has been shown by Neale (2003), the mixture distribution model will be most useful for estimating variance components from twin data when the zygosity is known with misclassification, but the rate of misclassification is known. The model will also be useful when the posterior probability of zygosity of each pair of twins is calculated using e.g. a latent-class approach (Heath *et al.*, 2003) prior to analysis using the mixture distribution model (Neale, 2003). While most of the current and new twin data collected will have zygosity information, the mixture distribution will still have a value for analysing twins from large population-based surveys, where zygosity information is not available as shown in Benyamin *et al.* (2005).

In Chapter 5, another statistical model, a mixed linear model (Visscher *et al.*, 2004) is used to partition the phenotypic (co)variances of traits into genetic and environmental factors from twins of known zygosity (twins from the Danish Twin Registry). This model is applied to understand the underlying genetic and environmental aetiology of endophenotypes associated with the metabolic syndrome (the cluster of obesity, insulin resistance, dyslipidaemia and hypertension). All endophenotypes showed moderate to high heritability (0.34-0.73) and no significant common environmental variance, except for fasting glucose. In a general population, it is demonstrated that the endophenotypes associated with the metabolic syndrome apparently do not share a substantial common genetic or familial environmental background. It is suggested that much more studies are needed before categorising people as having the metabolic syndrome has a clinical utility. These include understanding the metabolic syndrome as a unified syndrome with known underlying pathophysiology, the endophenotypes included or excluded, the value of diagnosing patients with the syndrome and the treatment itself (Kahn *et al.*, 2005).

Following heritability estimations using twins unknown or known zygosity, the next focus is on the identification of QTL/chromosomal regions associated with body height (Chapter 6). Using a large sample of Australian twins, a variance component linkage analysis revealed several chromosomal regions suggestive for linkage with body height, including 3q22.1 (LOD = 2.1) and 5q32 (LOD = 2.1). Sex-specific linkage analyses indicated that 1q32 (LOD = 1.9) and 15q23 (LOD = 1.9) were suggestive for linkage with body height in males, while 7p21.1 (LOD = 1.9) was only suggestive in females. The results are mostly consistent with previous studies, in that statistical support for most of the identified QTL was low to moderate. It can be concluded from the present study that despite the relatively large sample size, the moderate statistical support for most of the identified chromosomal regions indicates that body height is influenced by several or many genes, each having a modest effect.

Data used in each chapter come from different studies in different countries. Although the procedures for data collection have been designed to minimise possible bias and error, it is still important to discuss possible limitations of each data. In Chapter 2, the twin data come from two population surveys in Scotland, i.e. SMS1932 and SMS1947. The twins from SMS1947 were explicitly ascertained, but not for the twins from SMS1932. They were instead identified by matching pairs of subjects for: surname, date of birth and school identifier. Although it is unlikely that any two individuals identified as twins are non biological twins, the possibility cannot be ruled out, in particular for large schools and common surnames.

In Chapter 3, the twins were collected from a representative sample of twins born in England and Wales who participated in the U.K. Twins' Early Development Study (TEDS). The main limitation of this data is that the measures/phenotypes were obtained from a telephone interview. It is possible that for some twins the answers were influenced by his/her co-twins or their parents. Although each twin was interviewed individually and presumably independent of his/her co-twin and parental input, such possibility cannot be ruled out. It is also possible that the TEDS families are not a random sample from the population with respect to the phenotypes investigated.

In Chapter 5, the twins were recruited from two cohorts of the nation-wide, population-based Danish Twin Registry. The younger cohort was selected as self reported healthy based on questionnaire data obtained three years before the study, while these data were not available for the older cohort. However, both cohorts were screened for known diabetes and cardiovascular disease before being invited to this clinical investigation. The non-participation among the selected and invited twins may have introduced biases in the distributions of the individual endophenotypes of the metabolic syndrome, but, although unverifiable, it is unlikely that the intrapair correlations and the mutual relationship between the endophenotypes are biased by this attrition of the study sample compared to the original study population.

In Chapter 6, a possible limitation of the height data from the Australian twins is that most of the data come from a self reported measure. It has been shown by Macgregor *et al.* (2006) that self reported height is more variable than clinically measured height, which leads to a slightly lower estimate of heritability. While this may also affect the result of the linkage analysis presented in Chapter 6, this was not quantified in this thesis.

7.2 Future Prospects

In this era of high throughput genetics, what is the future for twins in human genetic studies and in which direction is the field of gene identification going? These are two important questions that are going to be addressed in this last section of the thesis.

7.2.1 The Future of Twins in Human Genetic Studies

As has been shown in this thesis, twins are not only useful for traditional heritability estimation, but also provide additional advantages in the search for individual genes underlying variation in complex quantitative traits. Boomsma *et al.* (2003), Martin *et al.* (1997) and MacGregor *et al.* (2000) have argued for the importance of twins as a population of choice for genetic studies of complex traits. Populations of MZ and DZ twins can be useful and advantageous for genetic linkage and association studies. MZ twins by themselves are not informative for linkage since they are genetically identical. However, if sibling(s) are also available in an MZ family, they become informative for linkage. DZ twins are genetically the same as ordinary siblings, but they are matched for age and shared family environments. These are two important properties of twins in relation to linkage and association studies for the following reasons. It is recognised that age is an important factor for disease expression of most complex diseases. The matching of family environments also makes DZ twins have a greater similarity for other environmental variables, which might be important for disease expression (MacGregor *et al.*, 2000).

MZ twins alone can also be a valuable resource for gene expression (e.g. Sarkijarvi *et al.*, 2006) and epigenetic studies (Fraga *et al.*, 2005; Petronis, 2006). With the advance of gene array technologies, the expression of thousands of genes from MZ twins that are discordant for a disease can be compared and the association between the level of gene expression and a disease can be inferred. For example, greater than 2-fold increase in the level of expression in six genes was found in MZ twins discordant for multiple sclerosis (Sarkijarvi *et al.*, 2006). This type of research may lead to the identification of a number genes that can be further studied for their roles in influencing of the variation of complex traits using other methods, such as candidate gene association studies.

Epigenetics, heritable variations in gene function that is not caused by changes in DNA sequences, has been increasingly recognised as an important source of phenotypic difference between discordant MZ twins (reviewed by Kato *et al.*, 2005). In the presence of epigenetic modifications, individuals with the same genotypes can manifest different phenotypic expression. MZ twins are particularly suitable for studying these phenomena since MZ twins are genetically identical (Petronis, 2006). So, in the absence of environmental differences within pairs of MZ twins, any phenotypic difference could be attributed to epigenetic modifications.

In the field of variance components estimation, heritability estimates have been commonly reported for many human phenotypes (reviewed by Boomsma *et al.*, 2002). At first glance, it seems that there is no need for more heritability estimations. However, it should be noted that heritability is a property of a population at a given time (Falconer and Mackay, 1996). Since heritability is a function of allele frequency (which may change with e.g. the presence of selection) and the environmental effects on a trait in a population may also change, the heritability of a trait still needs to be estimated from population to population. Twin studies will also play a role for decomposing the sources of family resemblance of new phenotypes. One example of a group of phenotypes gaining widespread interest is gene expression. These phenotypes quantify the expression level of genes, which are represented by the amount of mRNA transcript in a particular tissue. Using 10 pairs of MZ and 5 pairs of DZ twins, York *et al.* partitioned the variation in expression of more than 6,500 genes into genetic and environmental components. Average intraclass correlation of 0.3 (SD 0.38) and -0.08 (0.54) were estimated for MZ and DZ pairs, respectively. Therefore, it is reasonable to argue that twins will still be important populations for decomposing the phenotypic variance into genetic and environmental factors.

It is recognised that heritability estimation in the classical twin design was based on several important assumptions. One of the most important being the equality of common environmental assumption between MZ and DZ pairs. If MZ twins are treated more similarly than DZ twins, then the estimate of heritability will be biased. This assumption can be tested by comparing phenotypic similarity in twins of perceived versus true zygosity (Kendler *et al.*, 1993; Scarr, 1968). The assumption is violated if the phenotypic similarity of the twins is the result of perceived zygosity rather than the true zygosity. While this assumption has been tested empirically for some traits (e.g. Kendler *et al.*, 1993), this is not commonly practiced as part of most twin studies.

Motivated by the limitations of the classical twin design and the availability of genome-wide genetic marker data, Visscher *et al.* (2006) proposed a new method for estimating heritability using the observed proportion of the genome that is shared by relatives. This method is different from the conventional approach, in that the proportion of variance due to genetic factors was estimated from the actual (e.g. the proportion of alleles shared identical by descent) rather than the expected coefficient of relationship. As a result, this method does not imply any assumptions about the sources of twin resemblance, e.g. no assumption is made about the equality of common environmental experiences shared by MZ and DZ twins. With sufficient data, dominance and epistasis genetic effects can also be modelled using this method. In Australian samples, they obtained an estimate of heritability for height of 0.8, which is consistent with the estimates from the classical twin design commonly reported in the literature (see Silventoinen, 2003a).

Another important point to be made about the place of twins in human genetic studies is the presence of twin registries worldwide. These registries maintain contacts and databases of twins, where large collections of phenotypes are recorded. Most of the established registries are in the Scandinavian countries, such as Denmark, Finland, Norway and Sweden. However, a number of twin registries have also been established outside Scandinavia, including in Italy, the Netherlands, UK and Australia. In Asia, the recognition of the importance of twins in large scale epidemiological studies has prompted countries such as Japan, South Korea and Sri Lanka to establish twin registries (Boomsma *et al.*, 2002; Busjahn and Hur, 2006). These large resources are very valuable for human genetic studies, since it is well known that a large sample size is a key to the success for gene identification.

7.2.2 The Future of Gene Identification

The identification of genes underlying complex traits has been a very slow and difficult endeavor. While traditional linkage analysis has been very successful in discovering genes responsible for most Mendelian traits, its success for complex traits has been limited. Many studies have failed to replicate previously reported linkage between markers and quantitative traits. Among the reasons for this failure is that complex traits may be influenced by many genes of small to moderate effects and most studies are under-powered to detect these genes (Altmuller *et al.*, 2001). Genetic association and linkage disequilibrium mapping have been suggested to be more suitable for detecting genes of small effects (Risch and Merikangas, 1996; Risch, 2000). Therefore, there is now a growing interest in using both methods to identify genes underlying complex traits. Genetic association studies have traditionally looked for an association between a single or several markers and a trait. However, with the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), the availability of SNP database in the public domain (The International SNP Map Working Group, 2001) and the completion of the first phase of the HapMap Project (The International HapMap Consortium, 2005), the prospect of a whole genome association using hundreds of thousands single nucleotide polymorphisms (SNPs) is very promising. Indeed, several genome-wide association studies have reported associations between SNPs and complex diseases (Duerr *et al.*, 2006; Klein *et al.*, 2005; Ozaki *et al.*, 2002).

The challenges faced by such large scale studies are both financial and technical. The funding required for a genome-wide association with 1,000 cases and controls are very expensive [in the order of several million USD (e.g. Palmer and Cardon, 2005)]. Although the costs of genotyping will continue to decrease, the costs of phenotyping are unlikely to decrease. Therefore, collaboration between research institutions is needed. The large numbers of SNPs to be tested also pose technical/analytical challenges, particularly the problem of multiple testing. Hundreds of thousands tests are performed in a genome-wide association test, which increase false positive rates if no correction is made. Therefore, an appropriate significance threshold based on the traditional Bonferroni correction or other methods (Thomas, 2004) as well as replication studies are required to ensure that positive results are real.

Another new development in genetics that has attracted much attention is the genetic studies of gene expression. The field, which is known as *genetical genomics* (Jansen and Nap, 2001), emerged from the realization that measures of gene expression can be treated as any other quantitative trait. This means that gene expression can be described using quantitative genetic methods as well as subjected to genetic linkage and association analyses (Rockman and Kruglyak, 2006). Compared to traditional phenotypes, there are several advantages attributed to the analysis of gene expression. These include the possibility of large numbers (thousands) of phenotypes being assayed simultaneously and the fact that variation in gene expression is directly related to variation in DNA sequence (Rockman and Kruglyak, 2006).

The studies of Morley *et al.* (2004) and Cheung *et al.* (2005) demonstrate how genome-wide linkage and association studies can be applied to gene expression phenotypes. Initially, a genome-wide linkage analysis was conducted on 3,554 expression phenotypes using sib-pairs of 14 large Centre d'Etude Polymorphism (CEPH) families (Morley *et al.*, 2004). Out of 3,554 expression phenotypes, 374 showed evidence for linkage. Subsequent regional and genome-wide association studies using 57 founder individuals from the same CEPH pedigrees confirmed 15 expression phenotypes which showed significant associations in the same regions indicated by linkage analysis (Cheung *et al.*, 2005). Genetic studies of gene expression offer a new insight into genetic architecture of quantitative traits as well as provide a more direct relationship between DNA sequence variation and phenotypic variation (Rockman and Kruglyak, 2006). At the same time, these studies also raise issues related to the analysis and interpretation of such experiments due to the large number of expression phenotypes analysed simultaneously.

The studies of epigenetics, genome-wide association analysis and gene expression are among the most promising areas in the search for molecular basis underlying human complex traits. Twins are also expected to play significant roles in these new areas of research. Inevitably, these new areas of genetic research pose challenges in term of analysis and result interpretation. Nonetheless, new and exciting areas of human genetic research are waiting to be explored.

The advances in genetics and genomics described above are expected to improve human health globally. However, genetic and genomic research described above and in this thesis mostly focuses on the needs of developed countries (Global Forum for Health Research, 2004). In other words, the phenotypes/diseases studied are those that are common in developed countries. This is perhaps because the key driver for allocation of funding is the burden of diseases in developed countries. On the other hand, genetic and genomic research on diseases that affect most people in developing countries, including genomics research of infectious agents and host-parasite relationships, have been very limited (The Advisory Committee on Health Research, WHO, 2002). This discrepancy is captured in the expression "the 10/90 gap", which states that only 10% of research funding is invested into the health problems that account for 90% of the global disease burden (Global Forum for Health Research, 2004). To correct the 10/90 gap, the commitment of researchers, research institutions, funding bodies, private companies, governments, media and NGOs in both developed and developing countries is required (Global Forum for Health Research, 2004).

This commitment would ensure that people from all countries and backgrounds benefit from the advances in genetics and genomics.

Bibliography

Abbott, R. D., White, L. R., Ross, G. W., Petrovitch, H., Masaki, K. H., Snowdon, D. A. and Curb, J. D. (1998). Height as a marker of childhood development and late-life cognitive function: the Honolulu-Asia Aging Study. *Pediatrics* **102**: 602–609.

Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**: 97–101.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**: 1198–1211.

Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. and Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics* **69**: 936–950.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics* **54**: 535–543.

An, P., Freedman, B. I., Hanis, C. L., Chen, Y. I., Weder, A. B., Schork, N. J., Boerwinkle, E., Province, M. A., Hsiung, C. A., Wu, X., Quertermous, T. and Rao, D. C. (2005). Genome-wide linkage scans for fasting glucose, insulin, and insulin resistance in the National Heart, Lung, and Blood Institute Family Blood Pressure Program. *Diabetes* **54**: 909–914.

Antonarakis, S. E. and Beckmann, J. S. (2006). Mendelian disorders deserve more attention. *Nature Reviews Genetics* **7**: 277–282.

Austin, M. A., Edwards, K. L., McNeely, M. J., Chandler, W. L., Leonetti, D. L., Talmud, P. J., Humphries, S. E. and Fujimoto, W. Y. (2004). Heritability of multivariate factors of the metabolic syndrome in nondiabetic Japanese Americans. *Diabetes* **53**: 1166–1169.

Barr, D. R. and Slezak, N. L. (1972). A comparison of multivariate normal generators. *Communications of the ACM* **15**: 1048–1049.

Bartels, M., Rietveld, M. J., van Baal, J. C. and Boomsma, D. I. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. *Twin Research* **5**: 544–553.

Beck, S. R., Brown, W. M., Williams, A. H., Pierce, J., Rich, S. S. and Langefeld, C. D. (2003). Age-stratified QTL genome scan analysis for anthropometric measures. *BMC Genetics* **4**: S31.

Becker, A., Busjahn, A., Faulhaber, H. D., Bahring, S., Robertson, J., Schuster, H. and Luft, F. C. (1997). Twin zygosity: automated determination with microsatellites. *Journal of Reproductive Medicine* **42**: 260–266.

Beekman, M., Heijmans, B. T., Martin, N. G., Whitfield, J. B., Pedersen, N. L., DeFaire, U., Snieder, H., Lakenberg, N., Suchiman, H. E., de Knijff, P., Frants, R. R., van Ommen, G. J., Kluft, C., Vogler, G. P., Boomsma, D. I. and Slagboom, P. E. (2003). Evidence for a QTL on chromosome 19 influencing LDL cholesterol levels in the general population. *European Journal of Human Genetics* **11**: 845–850.

Benyamin, B., Deary, I. J. and Visscher, P. M. (2006). Precision and bias of a normal finite mixture distribution model to analyze twin data when zygosity is unknown: simulations and application to IQ phenotypes on a large sample of twin pairs. *Behavior Genetics* **36**: 935–946.

Benyamin, B., Wilson, V., Whalley, L. J., Visscher, P. M. and Deary, I. J. (2005). Large, consistent estimates of the heritability of cognitive ability in two entire populations of 11-year-olds twins from Scottish Mental Surveys of 1932 and 1947. *Behavior Genetics* **35**: 525–534.

Berg, K. (1988). Variability gene effect on cholesterol at the kidd blood locus. *Clinical Genetics* **33**: 102–107.

Bishop, E. G., Cherny, S. S., Corley, R., Plomin, R., DeFries, J. C. and Hewitt, J. K. (2003). Development genetic analysis of general cognitive ability from 1 and 12 years in a sample of adoptees, biological siblings, and twins. *Intelligence* **31**: 31–49.

Boomsma, D., Busjahn, A. and Peltonen, L. (2002). Classical twin studies and beyond. *Nature* **3**: 872–882.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approach for complex disease. *Nature Genetics* **33**: 228–237.

Bouchard, T. J., Lykken, D. T., McGue, M., Segal, N. L. and Tellegen, A. (1990). Sources of human psychological differences: the Minnesota Study of Twins Reared Apart. *Science* **250**: 223–228.

Bouchard, T. J. and McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology* **54**: 4–45.

Burton, P. R., Tobin, M. D. and Hopper, J. L. (2005). Key concepts in genetic epidemiology. *Lancet* **366**: 941–951.

Busjahn, A. and Hur, Y.-M. (2006). Twin registries: an ongoing success story. *Twin Research and Human Genetics* **9**: 705.

Busjahn, A., Knoblauch, H., Faulhaber, H.-D., Aydin, A., Uhlmann, R., Tuomilehto, J., Kaprio, J., Jedrusik, P., Januszewicz, A., Strelau, J., Schuster, H., Luft, F. C. and Muller-Myhsok, B. (2000). A region on chromosome 3 is linked to dizygotic twinning. *Nature Genetics* **26**: 398–399.

Cameron, A. J., Shaw, J. E. and Zimmet, P. Z. (2004). The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and Metabolism Clinics of North America* **33**: 351–375.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor analytic studies.* Cambridge University Press, Cambridge.

Cavalli-Sforza, L. L. and Bodmer, W. F. (1999). *The genetics of human populations.* Dover Publication, Inc., Mineola, N. Y. 11501.

Cederlof, R., Friberg, L., Jonsson, E. and Kaij, L. (1961). Studies on similarity diagnosis in twins with the aid of mailed questionnaires. *Acta Genetica et Statistica Medica* **11**: 338–362.

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M. and Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**: 1365–1369.

Cornes, B. K., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Duffy, D. L., Heijmans, B. T., Montgomery, G. W. and Martin, N. G. (2005). Sex-limited genome-wide linkage scan for body mass index in unselected sample of 933 Australian twin families. *Twin Research and Human Genetics* **8**: 612–632.

D'Alesio, A., Garabedian, M., Sabatier, J. P., Guaydier-Souquieres, G., Marcelli, C., Lemacon, A., Walrant-Debray, O. and Jehan, F. (2005). Two single-nucleotide polymorphisms in the human vitamin D receptor promotor change protein-DNA complex formation and are associated with height and vitamin D status in adolescent girls. *Human Molecular Genetics* **14**: 3539–3548.

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R. and Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: follow-up of the 1932 Scottish Mental Survey. *Intelligence* **28**: 49–55.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J. and Fox, H. C. (2004). The impact of childhood intelligence on later life: following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology* **86**: 130–147.

DeFries, J. C. and Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behavior Genetics* **15**.

Dempfle, A., Wudy, S. A., Saar, K., Hagemann, S., Friedel, S., Scherag, A., Berthold, L. D., Alzen, G., Gortner, L., Blum, W. F., Hinney, A., Nurnberg, P., Schafer, H. and Hebebrand, J. (2006). Evidence for involvement of the vitamin D receptor gene in idiopathic short stature via a genome wide linkage study and subsequent association studies. *Human Molecular Genetics* Published Online 11 August 2006.

Deng, H.-W., Xu, F.-H., Liu, Y.-Z., Shen, H., Deng, H., Huang, Q.-Y., Liu, Y.-J., Conway, T., Li, J.-L., Davies, K. M. and Recker, R. R. (2002). A whole-genome linkage scan suggests several genomic regions potentially containing QTLs underlying the variation of stature. *American Journal of Medical Genetics* **113**: 29–39.

Dick, D. M., Jones, K., Saccone, N., Hinrichs, A., Wang, J. C., Goate, A., Bierut, L., Almasy, L., Schuckit, M., Hesselbrock, V., Tischfield, J., Foroud, T., Edenberg, H., Porjesz, B. and Begleiter, H. (2006). Endophenotypes succesfully lead to gene identifiaction: results from Collaborative Study on the Genetics of Alcoholism. *Behavior Genetics* Published online 10 December 2005.

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee, A., Gregersen, P. K., Barmada, M. M., Rotter, J. I., Nicolae, D. L. and Cho, J. H. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* Published Online October 26, 2006.

Duffy, D., Montgomery, G., Treloar, S., Birley, A., Kirk, K., Boomsma, D., Beem, L., de Geus, E., Slagboom, E., Knighton, J., Reed, P. and Martin, N. (2001a). Ibd sharing around the pparg locus is not increased in dizygotic twins or their mother. *Nature Genetics* **28**: 315.

Duffy, D. L. (2006). An integrated genetic map for linkage analysis. *Behavior Genetics* **36**: 4–6.

Duffy, D. L., Mitchell, C. A. and Martin, N. G. (1998). Genetic and environmental risk factors for asthma. *American Journal of Respiratory and Critical Care Medicine* **157**: 840–845.

Duffy, D. L., Montgomery, G. W., Hall, J., Mayne, C., Healey, S. C., Brown, J., Boomsma, D. I. and Martin, N. G. (2001b). Human twinning is not linked to the region of chromosome 4 syntenic with the sheep twinning gene FecB. *American Journal of Medical Genetics* **100**: 182–186.

Eckel, R. H., Grundy, S. M. and Zimmet, P. Z. (2005). The metabolic syndrome. *Lancet* **365**: 1415–1428.

Edwards, K. L., Austin, M. A., Newman, B., Mayer, E., Kraus, R. M. and Selby, J. V. (1994). Multivariate analysis of the insulin resistance syndrome in women. *Arteriosclerosis and Thrombosis* **14**: 1940–1945.

Edwards, K. L., Newman, B., Mayer, E., Selby, J. V., Krauss, R. M. and Austin, M. A. (1997). Heritability of factors of the insulin resistance syndrome in women twins. *Genetic Epidemiology* **14**: 241–253.

Ellis, J. A., Stebbing, M. and Harrap, S. B. (2001). Significant population variation in adult male height associated with the Y chromosome and the aromatase gene. *Journal of Clinical Endocrinology and Metabolism* **86**: 4147–4150.

Evans, D. M., Gillespie, N. A. and Martin, N. G. (2002). Biometrical genetics. *Biological Psychology* **61**: 33–51.

Evans, D. M. and Martin, N. G. (2000). The validity of twin studies. *GeneScreen* **1**: 77–79.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics.* Longman Group Ltd, Essex CM20 2JE England, 4th edition.

Ferrari, S. L., Deutsch, S., Choudhury, U., Chevalley, T., Bonjour, J.-P., Dermitzakis, E. T., Rizzoli, R. and Antonarakis, S. E. (2004). Polymorphisms in the low-density lipoprotein receptor-related protein 5 (LRP5) gene are associated with variation in vertebral bone mass, vertebral bone size, and stature in Whites. *American Journal of Human Genetics* **74**: 866–875.

Ferreira, M. A. R., O'Gorman, L., Le Souef, P., Burton, P. R., Toelle, B. G., Robertson, C. F., Visscher, P. M., Martin, N. G. and Duffy, D. L. (2005). Robust estimation of experimentwise p values applied to a genome scan of multiple asthma traits identifies a new region of significant linkage on chromosome 20q13. *American Journal of Human Genetics* **77**: 1075–1085.

Flordellis, C. S. (2005). The emergence of a new paradigm of pharmacogenomics. *Pharmacogenomics* **6**: 515–526.

Ford, E. S. and Giles, W. H. (2003). A comparison of the prevalence of the metabolic syndrome using two proposed definitions. *Diabetes Care* **26**: 575–581.

Forget-Dubois, N., Perusse, D., Turecki, G., Girard, A., Billete, J., Rouleau, G., Boivin, M., Malo, J. and Tremblay, R. E. (2003). Diagnosing zygosity in infant twins: physical similarity, genotyping, and chorionicity. *Twin Research* **6**: 479–485.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suner, D., Cigudosa, J. C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T. D., Wu, Y.-Z., Plass, C. and Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 10604–10609.

Freeman, M. S., Mansfield, M. W., Barrett, J. H. and Grant, P. J. (2002). Heritability of features of insulin resistance syndrome in a community-based study of healthy families. *Diabetic Medicine* **19**: 994–999.

Fulker, D. W., Cherny, S. S., Sham, P. C. and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics* **64**: 259–267.

Galton, F. (1875). The history of twins, as a criterion of the relative powers of nature and nurture. *Fraser's Magazine* pages 566–576.

Geller, F., Dempfle, A. and Gorg, T. (2003). Genome scan for body mass index and height in the Framingham Heart Study. *BMC Genetics* **4(Suppl 1)**: s91.

Gilmour, A. R., Cullis, B. R., Welham, S. J. and Thompson, R. (2002). *ASReml computer package*. Herpenden, England.

Ginsburg, E., Livshits, G., Yakovenko, K. and Kobyliansky, E. (1998). Major gene control of human body height, weight and BMI in five ethnically different populations. *Annals of Human Genetics* **62**: 307–322.

Girman, C. J., Dekker, J. M., Rhodes, T., Nijpels, G., Stehouwer, C. D. A., Bouter, L. M. and Heine, R. J. (2005). An exploratory analysis of criteria for the metabolic syndrome and its prediction of long-term cardiovascular outcomes: the Hoorn Study. *American Journal of Epidemiology* **162**: 438–447.

Global Forum for Health Research (2004). 10/90 report on health research 2003-2004. Technical report, Global Forum for Health Research, Geneva.

Goldin, L. R., Camp, N. J., Keen, K. J., Martin, L. J., Moslehi, R., Ghosh, S., North, K. E., Wyszynski, D. F. and Blacker, D. (2003). Analysis of metabolic syndrome phenotypes in Framingham Heart Study families from genetic analysis workshop 13. *Genetic Epidemiology* **25**: S78–S89.

Goldsmith, H. H. (1991). A zygosity questionnaire for young twins: a research note. *Behavior Genetics* **21**: 257–269.

Gottfredson, L. and Deary, I. J. (2004). Intelligence predicts health and longevity: but why? *Current Directions in Psychological Science* **13**: 1–4.

Gray, J. R. and Thompson, P. M. (2004). Neurobiology of intelligence: science and ethics. *Nature Reviews Neurosciences* **5**: 471–482.

Gringas, P. and Chen, W. (2001). Mechanisms for differences in monozygous twins. *Early Human Development* **64**: 105–117.

Hall, J. G. (2003). Twinning. *Lancet* **362**: 735–743.

Hankins, G. V. D. and Saade, G. R. (2005). Factors influencing twins and zygosity. *Paediatric and Perinatal Epidemiology* **19**: 8–9.

Harlaar, N., Spinath, F. M., Dale, P. S. and Plomin, R. (2005). Genetic influences on early word recognition abilities and disabilities: a study of 7-year-old twins. *Journal of Child Psychology and Psychiatry* **46**: 373–384.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait loci and a marker locus. *Behavior Genetics* **2**: 3–19.

Hasler, G., Drevets, W. C., Gould, T. D., Gottesman, I. I. and Manji, H. K. (2006). Toward constructing an endophenotype strategy for bipolar disorders. *Biological Psychiatry* Published online 9 January 2006.

Heath, A. C., Bucholz, K. K., Madden, P. A. F., Dinwiddie, S. H., Slutske, W. S., Bierut, L. J., Statham, D. J., Dunne, M. P., Whitfield, J. B. and Martin, N. G. (1997). Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. *Psychological Medicine* **27**: 1381–1396.

Heath, A. C., Nyholt, D. R., Neuman, R., Madden, A. F., Bucholz, K. K., Todd, R. D., Nelson, E. C., Montgomery, G. W. and Martin, N. G. (2003). Zygosity diagnosis in the absence of genotypic data: an approach using latent class analysis. *Twin Research* **6**: 22–26.

Henkin, L., Bergman, R. N., Bowden, D. W., Ellsworth, D. L., Haffner, S. M., Langefeld, C. D., Mitchell, B. D., Norris, J. M., Rewers, M., Saad, M. F., Stamm, E., Wagenknecht, L. E. and Rich, S. S. (2003). Genetic epidemiology of insulin resistance and visceral adiposity: the IRAS Family Study design and methods. *Annals of Epidemiology* **13**: 211–217.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**: 95–108.

Hirschhorn, J. N., Lindgren, C. M., Daly, M. J., Kirby, A., Schaffner, S. F., Burtt, N. P., Altshuler, D., Parker, A., Rioux, J. D., Platko, J., Gaudet, D., Hudson, T. J., Groop, L. C. and Lander, E. S. (2001). Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *American Journal of Human Genetics* **69**: 106–116.

Hong, Y., Pedersen, N. L., Brismar, K. and de Faire, U. (1997). Genetic and environmental architecture of the features of the insulin-resistance syndrome. *American Journal of Human Genetics* **60**: 143–152.

Huggins, R. M., Loesch, D. Z. and Hoang, N. H. (1998). A comparison of methods of fitting models to twin data. *Australian & New Zealand Journal of Statistics* **40**: 129–140.

Humphreys, L. G., Davey, T. C. and Park, R. K. (1985). Longitudinal correlation analysis of standing height and intelligence. *Child Development* **56**: 1465–1478.

Imaizumi, Y. (2003). A comparative study of zygotic twinning and triplet rate in eight countries, 1972-1999. *Journal of Biosocial Sciences* **35**: 287–302.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Jackson, R. W., Snieder, H., Davis, H. and Treiber, F. A. (2001). Determination of twin zygosity: a comparison of DNA with various questionnaire indices. *Twin Research* **4**: 12–18.

Jansen, R. C. and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics* **17**: 388–391.

Johnson, F. W. (1991). Biological factors and psychometric intelligence: a review. *Genetic, Social, and General Psychology Monographs* **117**: 313–357.

Joseph, J. (2002). Twin studies in psychiatry and psychology: science or pseudoscience. *Psychiatric Quarterly* **73**: 71–82.

Joseph, J. (2003). *The gene illusion: genetic research in psychiatry and psychology under the microscope.* PCCS, Ross-on-Wye.

Jowett, J. B., Elliot, K. S., Curran, J. E., Hunt, N., Walder, K. R., Collier, G. R., Zimmet, P. Z. and Blangero, J. (2004). Genetic variation in BEACON influences quantitative variation in metabolic syndrome-related phenotypes. *Diabetes* **53**: 2467–2472.

Kahn, R., Buse, J., Ferrannini, E. and Stern, M. (2005). The metabolic syndrome: time for a critical appraisal. Joint statement from the American Association and the European Association for the Study of Diabetes. *Diabetologia* **48**: 1684–1699.

Kato, T., Iwamoto, K., Kakiuchi, C., Kuratomi, G. and Okazaki, Y. (2005). Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders. *Molecular Psychiatry* **10**: 622–630.

Kendall, M. G. and Stuart, A. (1947). *The advanced theory of statistics*, volume 1. Charles Griffin & Company Limited, London, 3rd edition.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. and Eaves, L. J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness. *Behavior Genetics* **23**: 21–27.

Kirk, K. M., Birley, A. J., Statham, D. J., Haddon, B., Lake, R. I. E., Andrews, J. G. and Martin, N. G. (2000). Anxiety and depression in twin and sib pairs extremely discordant and concordant for neuroticism: prodromus to a linkage study. *Twin Research* **3**: 299–309.

Kissebah, A. H., Sonnenberg, G. E., Myklebust, J., Goldstein, M., Broman, K., James, R. G., Marks, J. A., Krakower, G. R., Jacob, H. J., Weber, J., Martin, L., Blangero, J. and Comuzzie, A. G. (2000). Quantitative trait loci on chromosome 3 and 17 influence phenotypes of the metabolic syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 14478–14483.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Lerris, F. L., Ott, J., Barnstable, C. and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385–389.

Knopik, V. S. and DeFries, J. C. (1998). A twin study of gender-influenced individual differences in general cognitive abilities. *Intelligence* **26**: 81–89.

Kovas, Y., Harlaar, N., Petrill, S. A. and Plomin, R. (2005). 'Generalist genes' and mathematics in 7-year-old twins. *Intelligence* **33**: 473–489.

Laaksonen, D. E., Lakka, H. M., Niskanen, L. K., Kaplan, G. A., Salonen, J. T. and Lakka, T. A. (2002). Metabolic syndrome and development of diabetes mellitus: application and validation of recently suggested definitions of the metabolic syndrome in a prospective cohort study. *American Journal of Epidemiology* **156**: 1070–1077.

Laird, N. M. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**: 385–394.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**: 241–247.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**: 2037–2048.

Lawlor, D. A., Ebrahim, S., May, M. and Smith, G. D. (2004). (Mis)use of factor analysis in the study of insulin resistance syndrome. *American Journal of Epidemiology* **159**: 1013–1018.

Lei, S.-F., Deng, F.-Y., Xiao, S.-M., Chen, X.-D. and Deng, H.-W. (2005). Association and haplotype analyses of the COL1A2 and ER-alpha gene polymorphisms with bone size and height in Chinese. *Bone* **36**: 533–541.

Li, J. K. Y., Ng, M. C. Y., So, W. Y., Chiu, C. K. P., Ozaki, R., Tong, P. C. Y., Cockram, C. S. and Chan, J. C. N. (2006). Phenotypic and genetic clustering of the metabolic syndrome in Chinese families with type 2 diabetes mellitus. *Diabetes/Metabolism Research and Reviews* **22**: 46–52.

Liew, S. H. M., Elsner, H., Spector, T. D. and Hammond, C. J. (2005). The first classical twin study? analysis of refractive error using monozygotic and dizygotic twins published in 1922. *Twin Research and Human Genetics* **8**: 198–200.

Lin, H. F., Boden-Albala, B., Juo, S. H., Park, N., Rundek, T. and Sacco, R. L. (2005). Heritabilities of the metabolic syndrome and its components in the Nothern Manhattan Family Study. *Diabetologia* Published online: 4 August 2005.

Liu, Y.-Z., Xiao, P., Guo, Y.-F., Xiong, D.-H., Zhao, L.-J., Shen, H., Liu, Y.-Z., Dvornyk, V., Long, J.-R., Deng, H.-Y., Li, J.-L., Recker, R. R. and Deng, H.-W. (2006). Genetic linkage of human height is confirmed to 9q22 and Xq24. *Human Genetics* **119**: 295–304.

Loos, R. J. F., Katzmarzyk, P. T., Rao, D. C., Rice, T., Leon, A. S., Skinner, J. S., Wilmore, J. H., Rankinen, T. and Bouchard, C. (2003). Genome-wide linkage scan for the metabolic syndrome in the HERITAGE Family Study. *Journal of Clinical Endocrinology and Metabolism* **88**: 5935–5943.

Lorentzon, M., Lorentzon, R. and Nordstrom, P. (2000). Vitamin D receptor gene polymorphism is associated with birth height, growth to adolescence, and adult stature in healthy Caucasian men: a cross-sectional and longitudinal study. *Journal of Clinical Endocrinology and Metabolism* **85**: 1666–1671.

Luciano, M., Wright, M. J., Geffen, G. M., Geffen, L. B., Smith, G. A., Evans, D. M. and Martin, N. G. (2003). A genetic two-factor model of the covariation among a subset of multidimensional aptitude battery and wechsler adult intelligence scale - revised subtests. *Intelligence* **31**: 589–605.

Luciano, M., Zhu, G., Kirk, K. M., Gordon, S. D., Heath, A. C., Montgomery, G. W. and Martin, N. G. (2006). "No thanks, it keeps me awake": the genetics of coffee-attributed sleep disturbance. *Submitted* .

Lynch, M. and Walsh, B. (1998). *Genetic and Analysis of Quantitative Traits.* Sinnauer Associates, USA.

Lyons, M. J. and Bar, J. L. (2001). Is there a role for twin studies in the molecular genetics era? *Harvard Review Psychiatry* **9**: 318–323.

MacGregor, A. J., Snieder, H., Schork, N. J. and Spector, T. D. (2000). Twins: novel uses to study complex traits and genetics. *Trends in Genetics* **16**: 131–134.

Macgregor, S., Cornes, B. K., Martin, N. G. and Visscher, P. M. (2006). Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Human Genetics* **120**: 571–580.

Maes, H. H. M., Neale, M. C. and Eaves, L. J. (1997). Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics* **27**: 325–351.

Magnusson, P. K. E., Rasmussen, F. and Gyllensten, U. B. (2006). Height at age 18 years is a strong predictor of attained education later in life: cohort study of over 950,000 Swedish men. *International Journal of Epidemiology* **35**: 658–663.

Maison, P., Byrne, C. D., Hales, C. N., Day, N. E. and Wareham, N. J. (2001). Do different dimensions of the metabolic syndrome change together over time? *Diabetes Care* **24**: 1758–1763.

Martin, L. J., North, K. E., Dyer, T., Blangero, J., Comuzzie, A. G. and Williams, J. (2003). Phenotypic, genetic, and genome-wide structure in the metabolic syndrome. *BMC Genetics* **4(Suppl.1)**: S95.

Martin, N., Boomsma, D. and Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics* **17**: 387–392.

McCarron, P., Okasha, M., McEwen, J. and Smith, G. D. (2002). Height in young adulthood and risk of death from cardiorespiratory disease: a prospective study of male former students of Glasgow University, Scotland. *American Journal of Epidemiology* **155**: 683–687.

McQueen, M. B., Bertram, L., Rimm, E. B., Blacker, D. and Santangelo, S. L. (2003). A QTL genome scan of the metabolic syndrome and its component traits. *BMC Genetics* **4(Suppl.1)**: s96.

Mehrota, S. N. and Maxwell, J. (1949). The intelligence of twins: a comparative study of eleven year-old twins. *Population Studies* **3**: 295–302.

Meirhaeghe, A., Fajas, L., Gouilleux, F., Cottel, D., Helbecque, N., Auwerx, J. and Amouyel, P. (2003). A functional polymorphism in a STAT5B site of the human PPAR gamma 3 gene promotor affects height and lipid metabolism in a French population. *Arteriosclerosis, Thrombosis and Vascular Biology* **23**: 289–294.

Merikangas, K. R. (1982). Assortative mating for psychiatric disorders and psychological traits. *Archives of General Psychiatry* **39**: 1173–1180.

Mitchell, B. D., Kammerer, C. M., Mahaney, M. C., Blangero, J., Comuzzie, A. G., Atwood, L. D., Haffner, S. M., Stern, M. P. and MacCluer, J. W. (1996). Pleiotropic effects of genes influencing insulin levels on lipoprotein and obesity measures. *Arteriosclerosis, Thrombosis, and Vascular Biology* **16**: 281–288.

Miyake, H., Nagashima, K., Onigata, K., Nagashima, T., Takano, Y. and Morikawa, A. (1999). Allelic variations of the D2 dopamine receptor gene in children with idiopathic short stature. *Journal of Human Genetics* **44**: 26–29.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.

Mukhopadhyay, N. and Weeks, D. E. (2003). Linkage analysis of adult height with parent-of-origin effects in the Framingham Heart Study. *BMC Genetics* **4 (Suppl 1)**: S76.

Neale, M. C. (2003). A finite mixture distribution model for data collected from twins. *Twin Research* **6**: 235–239.

Neale, M. C., Boker, S. M., Xie, G. and Maes, H. H. (2002). *Mx: Statistical Modelling.* Department of Psychiatry, Richmond, VA 23298, 6th edition.

Neale, M. C. and Maes, H. H. M. (2004). *Methodology for genetic studies of twins and families.* Kluwer Academic Publishers B.V., Dordrecht, The Netherlands.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J. and Urbina, S. (1996). Intelligence: known and unknowns. *American Psychologist* **51**: 77–101.

Nelson, T. L., Vogler, G. P., Pedersen, N. L., Y., H. and Miles, T. P. (2000). Genetic and environmental influences on body fat distribution, fasting insulin levels and CVD: are the influences shared? *Twin Research* **3**: 43–50.

Ng, M. C. Y., So, W. Y., Lam, V. K. L., Cockram, C. S., Bell, G. I., Cox, N. J. and Chan, J. C. N. (2004). Genome-wide scan for metabolic syndrome and related quantitative traits in Hong Kong Chinese and confirmation of a susceptibility locus in chromosome 1q21-q25. *Diabetes* **53**: 2676–2683.

North, K. E., Willliams, K., Williams, J. T., Best, L. G., Lee, E. T., Fabsitz, R. R., Howard, B. V., Gray, R. S. and MacCluer, J. W. (2003). Evidence for genetic factors underlying the insulin resistance syndrome in American Indians. *Obesity Research* **11**: 1444–1448.

Novak, S., Stapleton, L. M., Litaker, J. R. and Lawson, K. A. (2003). A confirmatory factor analysis evaluation of the coronary heart disease risk factors of metabolic syndrome with emphasis on the insulin resistance factor. *Diabetes, Obesity and Metabolism* **5**: 388–396.

O'Connell, J. R. and Weeks, D. E. (1999). An optimal algorithm for automatic genotype elimination. *American Journal of Human Genetics* **65**: 1733–1740.

O'Connor, T. P. and Crystal, R. G. (2006). Genetic medicines: treatment strategies for hereditary disorder. *Nature Reviews Genetics* **7**: 261–276.

Online Mendelian Inheritance in Man (2006). Mckusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bathesda, MD). World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

Ozaki, K., Ohnishi, Y., Iida, A., Sekina, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. and Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32**: 650–654.

Palmer, L. J. and Cardon, L. R. (2005). Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* **366**: 1223–1234.

Palmert, M. R. and Hirschhorn, J. N. (2003). Genetic approaches to stature, pubertal timing, and other complex traits. *Molecular Genetics and Metabolism* **80**: 1–10.

Pearson, K. and Lee, A. (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika* **2**: 357–462.

Peck, M. N. and Lundberg, O. (1995). Short stature as an effect of economic and social conditions in childhood. *Social Science and Medicine* **41**: 733–738.

Peeters, H., Van Gestel, S., Vlietinck, R., Derom, C. and Derom, R. (1998). Validation of a telephone zygosity questionnaire in twins of known zygosity. *Behavior Genetics* **28**: 159–163.

Perola, M., Ohman, M., Hiekkalinna, T., Leppavuori, J., Pajukunta, P., Wessman, M., Koskenvuo, M., Palotie, A., Lange, K., Kaprio, J. and Peltonen, L. (2001). Quantitative-trait-locus analysis of body-mass index and of stature, by combined analysis of genome scans of five Finnish study groups. *American Journal of Human Genetics* **69**: 117–123.

Perusse, L., Rice, T., Despres, J. P., Rao, D. C. and Bouchard, C. (1997). Cross-trait familial resemblance for body fat and blood lipids: familial correlations in the Quebec Family Study. *Arteriosclerosis, Thrombosis, and Vascular Biology* **17**: 3270–3277.

Petrill, S. A., Rempell, J., Oliver, B. and Plomin, B. (2002). Testing cognitive abilities by telephone in a sample of 6- to 8-year-olds. *Intelligence* **30**: 353–360.

Petronis, A. (2006). Epigenetics and twins: three variations on the theme. *Trends in Genetics* **22**: 347–350.

Phillips, D. I. W. (1993). Twin studies in medical research: can they tell us whether diseases are genetically determined? *Lancet* **341**: 1008–1009.

Pietilainen, K. H., Kaprio, J., Rasanen, M., Rissanen, A. and Rose, R. J. (2002). Genetic and environmental influences on the tracking of body size from birth to early adulthood. *Obesity Research* **10**: 875–884.

Plomin, R., DeFries, J. C., McClearn, G. E. and McGuffin, P. (2001). *Behavioural genetics*. Worth, New York, 4th edition.

Plomin, R. and Spinath, F. (2004). Intelligence: genetics, genes, and genomics. *Journal of Personality and Social Psychology* **86**: 112–129.

Posthuma, D., de Geus, E. J. C., Boomsma, D. I. and Neale, M. C. (2004). Combined linkage and association test in Mx. *Behavior Genetics* **34**: 179–196.

Poulsen, P., Vaag, A., Kyvik, K. and Beck-Nielsen, H. (2001). Genetic versus environmental aetiology of the metabolic syndrome among male and female twins. *Diabetologia* **44**: 537–543.

Price, T. S., Eley, T. C., Dale, P. S., Stevenson, J., Saudino, K. and Plomin, R. (2000a). Genetic and environmental covariation between verbal and nonverbal cognitive development in infancy. *Child Development* **71**: 948–959.

Price, T. S., Freeman, B., Craig, I., Petrill, S. A., Ebersole, L. and Plomin, R. (2000b). Infant zygosity can be assigned by parental report questionnaire data. *Twin Research* **3**: 129–133.

Purcell, S., Cherny, S. S. and Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**: 149–150.

Purcell, S. and Sham, P. C. (2003). A model-fitting implementation of the DeFries-Fulker model for selected twin data. *Behavior Genetics* **33**: 271–278.

R Development Core Team (2006). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Http://www.R-project.org.

Rainwater, D. L., Mitchell, B. D., Mahaney, M. C. and Haffner, S. M. (1997). Genetic relationship between measures of HDL phenotypes and insulin concentrations. *Arteriosclerosis, Thrombosis, and Vascular Biology* **17**: 3414–3419.

Rappold, G. A., Shanske, A. and Saenger, P. (2005). All shook up by SHOX deficiency. *Journals of Pediatrics* **147**: 422–424.

Remes, T., Vaisanen, S. B., Mahonen, A., Huuskonen, J., Kroger, H., Jurvelin, J. S. and Rauramaa, R. (2005). Bone mineral density, body height, and vitamin D receptor gene polymorphism in middle-aged men. *Annals of Medicine* **37**: 383–392.

Rende, R. D., Plomin, R. and Vandenberg, G. (1990). Who discovered the twin method. *Behavior Genetics* **20**: 277–285.

Rich, S. S., Bowden, D. W., Haffner, S. M., Norris, J. M., Saad, M. F., Mitchell, B. D., Rotter, J. I., Langefeld, C. D., Hedrick, C. C., Wagenknecht, L. E. and Bergman, R. N. (2005). A genome scan for fasting insulin and fasting glucose identifies a quantitative trait locus on chromosome 17p: the Insulin Resistance Atherosclerosis Study (IRAS) Family Study. *Diabetes* **54**: 290–295.

Rijsdijk, F. V. and Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in Bioinformatics* **3**: 119–133.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.

Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**: 1584–1589.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.

Roche, H. M., Phillips, C. and Gibney, M. J. (2005). The metabolic syndrome: the crossroads of diet and genetics. *Proceedings of the Nutrition Society* **64**: 371–377.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics* **7**: 862–872.

Sale, M. M., Freedman, B. I., Hicks, P. J., Williams, A. H., Langefeld, C. D., Gallagher, C. J., Bowden, D. W. and Rich, S. S. (2005). Loci contributing to adult height and body mass index in African American families ascertained for type 2 diabetes. *Annals of Human Genetics* **69**: 517–527.

Samaras, K., Nguyen, T. V., Jenkins, A. B., Eisman, J. A., Howard, G. M., Kelly, P. J. and Campbell, L. V. (1999). Clustering of insulin resistance, total and central abdominal fat: same genes or sama environment? *Twin Research* **2**: 218–225.

Samaras, T. T., Elrick, H. and Storms, L. H. (2003). Is height related to longevity? *Life Sciences* **72**: 1781–1802.

Sammalisto, S., Hiekkalinna, T., Suviolahti, E., Sood, K., Metzidis, A., Pajukanta, P., Lilja, H. E., Soro-Paavonen, A., Taskinen, M.-R., Tuomi, T., Almgren, P., Orho-Melander, M., Groop, L., Peltonen, L. and Perola, M. (2005). A male-specific quantitative trait locus on 1p21 controlling human stature. *Journal of Medical Genetics* **42**: 932–939.

Sarkijarvi, S., Kuusisto, H., Paalavuo, R., Levula, M., Airla, N., Lehtimaki, T., Kaprio, J., Koskenvuo, M. and Elovaara, I. (2006). Gene expression profiles in Finnish twins with multiple sclerosis. *BMC Medical Genetics* **7**.

Sarna, S., Kaprio, J., Sistone, P. and Koskenvuo, M. (1978). Diagnosis of twin zygosity by mailed questionnaire. *Human Heredity* **28**: 241–254.

Scarr, S. (1968). Environmental bias in twin studies. *Eugenics Quarterly* **15**: 34–40.

Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science* **174**: 1285–1295.

Schousboe, K., Visscher, P. M., Erbas, B., Kyvik, K. O., Hopper, J. L., Henriksen, J. E., Heitmann, B. L. and Sørensen, T. I. A. (2004). Twin study of genetic and environmental influences on adult body size, shape and composition. *International Journal of Obesity* **28**: 39–48.

Schousboe, K., Visscher, P. M., Henriksen, J. E., Hopper, J. L., Sørensen, T. I. and Kyvik, K. O. (2003a). Twin study of genetic and environmental influences on glucose tolerance and indices of insulin sensitivity and secretion. *Diabetologia* **46**: 1276–1283.

Schuit, S. C. E., Van Meurs, J. B. J., Bergink, A. P., Van Der Klift, M., Fang, Y., Leusink, G., Hofman, A., Van Leeuwen, J. P. T. M., Uitterlinden, A. G. and Pols, H. A. P. (2004). Height in pre- and postmenopausal women is influenced by estrogen receptor alpha gene polymorphisms. *Journal of Clinical Endocrinology and Metabolism* **89**: 303–309.

Scillitani, A., Jang, C., Wong, B. Y.-L., Hendy, G. N. and Cole, D. E. C. (2006). A functional polymorphism in the PTHR1 promoter region is associated with adult height and BMD measured at the femoral neck in a large cohort of young Caucasian women. *Human Genetics* **119**: 416–421.

Scottish Council for Research in Education (1933). *The intelligence of Scottish children.* University of London Press, London.

Scottish Council for Research in Education (1949). *The trend of Scottish intelligence.* University of London Press, London.

Scottish Council for Research in Education (1953). *Social implications of the 1947 Scottish Mental Survey.* University of London Press, London.

Sham, P. C., Cherny, S. S., Purcell, S. and Hewitt, J. K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics* **66**: 1616–1630.

Sham, P. C. and Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *American Journal of Human Genetics* **68**: 1527–1532.

Shaw, D. I., Hall, W. L. and Williams, C. M. (2005). Metabolic syndrome: what is it and what are the implications? *Proceedings of the Nutrition Society* **64**: 349–357.

Shearman, A. M., Ordovas, J. M., Cupples, L. A., Schaefer, E. J., Harmon, M. D., Shao, Y., Keen, J. D., DeStefano, A. L., Joost, O., Wilson, P. W. F., Housman, D. E. and Myers, R. H. (2000). Evidence for a gene influencing the TG/HDL-C ratio on chromosome 7q32.3-qter: a genome-wide scan in the Framingham Study. *Human Molecular Genetics* **9**: 1315–1320.

Shmulewitz, D., Heath, S. C., Blundell, M. L., Han, Z., Sharma, R., Salit, J., Auerbach, S. B., Signorini, S., Breslow, J. L., Stoffel, M. and Friedman, J. M. (2006). Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and dyslipidemia on the island of Kosrae, Federated States of Micronesia. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 3503–3509.

Siitonen, N., Pulkkinen, L., Mager, U., Lindstrom, J., Eriksson, J. G., Valle, T. T., Hamalainen, H., Ilanne-Parikka, P., Keinanen-Kiukaanniemi, S., Tuomilehto, J., Laakso, M. and Uusitupa, M. (2006). Association of sequence variations in the gene encoding adiponectin receptor 1 (ADIPOR1) with body size and insulin levels. The Finnish Diabetes Prevention Study. *Diabetologia* **49**: 1795–1805.

Silventoinen, K. (2003a). Determinants of variation in adult body height. *Journal of Biosocial Sciences* **35**: 263–285.

Silventoinen, K. (2003b). Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research* **6**: 399–408.

Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J. and Rose, R. J. (2003). Assortative mating by body height and bmi: Finnish twins and their spouses. *American Journal of Human Biology* **15**: 620–627.

Silventoinen, K., Krueger, R. F., Bouchard Jr, T. J., Kaprio, J. and McGue, M. (2004). Heritability of body height and educational attainment in an international context: comparison of adult twins in Minnesota and Finland. *American Journal of Human Biology* **16**: 544–555.

Silventoinen, K., Zdravkovic, S., Skytthe, A., McCarron, P., Herskind, A. M., Koskenvuo, M., de Faire, U., Pedersen, N., Christensen, K. and Kaprio, J. (2006). Association between height and coronary heart disease mortality: a prospective study of 35,000 twin pairs. *American Journal of Epidemiology* **163**: 615–621.

Souza, R. L. R., Fadel-Picheth, C., Allebrandt, K. V., Furtado, L. and Chautard-Freire-Maia, E. A. (2005). Possible influence of BCHE locus of butyrylcholinesterase on stature and body mass index. *American Journal of Physical Anthropology* **126**: 329–334.

Spinath, F. M., Price, T. S., Dale, P. S. and Plomin, R. (2004). The genetic and environmental origins of language disability and ability. *Child Development* **75**: 445–454.

Sundet, J. M., Tambs, K., Harris, J. R., Magnus, P. and Torjussen, T. M. (2005). Resolving the genetic and environmental sources of the correlation between height and intelligence: a study of nearly 2600 Norwegian male twin pairs. *Twin Research and Human Genetics* **8**: 307–311.

Sundstrom, J., Riserus, U., Byberg, L., Zethelius, B., Lithell, H. and Lind, L. (2006). Clinical value of the metabolic syndrome for long term prediction of total and cardiovascular mortality: prospective, population based cohort study. *BMJ* **332**: 878–882.

Tang, W., Miller, M. B., Rich, S. S., North, K. E., Pankow, J. S., Borecki, I. B., Myers, R. H., Hopkins, P. N., Leppert, M. and Arnett, D. K. (2003). Linkage analysis of a composite factor for the multiple metabolic syndrome. *Diabetes* **52**: 2840–2847.

Teare, M. D. and Barrett, J. H. (2005). Genetic linkage studies. *Lancet* **366**: 1036–1044.

Teasdale, T. W., Sorensen, T. I. A. and Owen, D. R. (1989). Fall in association of height with intelligence and education level. *BMJ* **298**: 1292–1293.

The Advisory Committee on Health Research, WHO (2002). Genomics and world health. Report, World Health Organisation, Geneva.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**: 1299–1320.

The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.

Thomas, D. C. (2004). *Statistical methods in genetic epidemiology.* Oxford University Press, New York.

Thorndike, E. L. (1905). *Measurements of twins.* Number 1 in Archives of Philosophy, Psychology and Scientific Methods, The Science Press, New York.

Tong, S., Caddy, D. and Short, R. V. (1997). Use of dizygotic to monozygotic twinning ratio as a measure of fertility. *Lancet* **349**: 843–845.

Tregouet, D. A., Herbeth, B., Juhan-Vague, I., Siest, G., Ducimetiere, P. and Tiret, L. (1999). Bivariate familial correlation analysis of quantitative traits by use of estimating equation: application to a familial analysis of the insulin resistance syndrome. *Genetic Epidemiology* **16**: 69–83.

Trouton, A., Spinath, F. M. and Plomin, R. (2002). Twins Early Development Study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood. *Twin Research* **5**: 444–448.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. and et al. (2001). The sequence of the human genome. *Science* **291**: 1304–1351.

Viding, E., Price, T. S., Spinath, F. M., Bishop, D. V. M., Dale, P. S. and Plomin, R. (2003). Genetic and environmental mediation of the relationship between language and nonverbal impairment in 4-year-old twins. *Journal of Speech, Language, and Hearing Research* **46**: 1271–1282.

Visscher, P. M., Benyamin, B. and White, I. (2004). The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin Research* **7**: 670–674.

Visscher, P. M. and Hopper, J. L. (2001). Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Annal of Human Genetics* **65**: 583–601.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2**: 316–325.

Webbink, D., Roeleveld, J. and Visscher, P. M. (2006). Identification of twin pairs from large population-based samples. *Twin Research and Human Genetics* **9**: 496–500.

Weinberg, W. (1902). Beitrge zur physiologie und pathologie der mehrlingsgeburten beim menschen. *Archiv Gesamte Physiol Menschen Tiere* **88**: 346–430.

Willemsen, G., Boomsma, D. I., Beem, A. L., Vink, J. M., Slagboom, P. E. and Posthuma, D. (2004). QTLs for height: results of a full genome scan in Dutch sibling pairs. *European Journal of Human Genetics* **12**: 820–828.

Wiltshire, S., Frayling, T. M., Hattersley, A. T., Hitman, G. A., Walker, M., Levy, J. C., O'Rahilly, S., Groves, C. J., Menzel, S., Cardon, L. R. and McCarthy, M. I. (2002). Evidence for linkage of stature to chromosome 3p26 in a large U.K. family data set ascertained for type 2 diabetes. *American Journal of Human Genetics* **70**: 543–546.

Wright, M., De Geus, E., Ando, J., Luciano, M., Posthuma, D., Ono, Y., Hansell, N., Van Baal, C., Hiraishi, K., Hasegawa, T., Smith, G., Geffen, G., Geffen, L., Kanba, S., Miyake, A., Martin, N. and Boomsma, D. (2001). Genetics of cognition: outline of a collaborative twin study. *Twin Research* **4**: 48–56.

Wright, M. J. and Martin, N. G. (2004). Brisbane Adolescent Twin Study: outline of study methods and research projects. *Australian Journal of Psychology* **56**: 65–78.

Wu, X., Cooper, R. S., Boerwinkle, E., Turner, S. T., Hunt, S., Myers, R., Olshen, R. A., Curb, D., Zhu, X., Kan, D. and Luke, A. (2003). Combined analysis of genomewide scans for adult height: results from the NHLBI Family Blood Pressure Program. *European Journal of Human Genetics* **11**: 271–274.

Xiong, D.-H., Xu, F.-H., Shen, H., Long, J.-R., Elze, L., Recker, R. R. and Deng, H.-W. (2005). Vitamin D receptor gene polymorphisms are linked to and associated with adult height. *Journal of Medical Genetics* **42**: 228–234.

Xu, J., Bleecker, E. R., Jongepier, H., Howard, T. D., Koppelman, G. H., Postma, D. S. and Meyers, D. A. (2002). Major recessive gene(s) with considerable residual polygenic effect regulating adult height: confirmation of genomewide scan results for chromosomes 6, 9, and 12. *American Journal of Human Genetics* **71**: 646–650.

Yang, T.-L., Xiong, D.-H., Guo, Y., Recker, R. R. and Deng, H.-W. (2006). Association analyses of CYP19 gene polymorphisms with height variation in a large sample of Caucasian nuclear families. *Human Genetics* **120**: 119–125.

York, T. P., Miles, M. F., Kendler, K. S., Jackson-Cook, C., Bowman, M. L. and Eaves, L. J. (2005). Epistatic and environmental control of genome-wide gene expression. *Twin Research and Human Genetics* **8**: 5–15.

Zhu, G., Evans, D. M., Duffy, D. L., Montgomery, G. W., Medland, S. E., Gillespie, N. A., Ewen, K. R., Jewell, M., Liew, Y. W., Hayward, N. K., Sturm, R. A., Trent, J. M. and Martin, N. G. (2004). A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q. *Twin Research* **7**: 197–210.