CO-MORBIDITIES AS QUANTITATIVE TRAITS

by

PAOLA RASKA

Submitted in partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Department of Epidemiology and Biostatistics

CASE WESTERN RESERVE UNIVERSITY

August 2010

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of Paola Raska

candidate for the Ph.D. degree *.

(signed) J. Sunil Rao (chair of the committee)

Joseph Nadeau

Xiaofeng Zhu

Catherine Stein

(date) 5 / 20 / 2010

*We also certify that written approval has been obtained for any proprietary material contained therein.

Dedication Page

I dedicate this work to my mother Doris. Her unconditional and unwavering love and belief in me since I first began what was to become a long and tempestuous relationship with science was what always kept me going. She has truly been through everything with me, every step of the way.

I also dedicate this work to my family, my husband Martin, our son Marko and, our soon to join us daughter, Niki. They are my purpose, my north. Everything I do I do for them.

TABLE OF CONTENTS

Table of contents ……………………………………………………………………... 1
List of tables and boxes ……………………………………………………….... 4
List of figures …………………………………………………………………….... 5
Acknowledgements ………………………………………………………………... 6
Abstract …………………………………………………………………………...... 8
Introduction ………………………………………………………………….…...... 10
Organizational comment …………………………………………………….…... 14
Section 1. PHYSIOLOGY …………………………………………………….….. 15
1.1 Obesity and associated phenotypes ……………………………….... 15
1.2 Heterogeneity in obesity and its relationship to other phenotypes …. 19
1.3 Importance of heterogeneity in disease prognosis ………………….. 22
1.4 Plausible mechanistic explanation underlying heterogeneity ………. 24
1.5 Challenging unidirectional relationships …………………………..... 26
1.5.1 Hypertension causing metabolic disease ……………………….. 26
1.5.2 Insulin resistance causing obesity …………………………….…... 27
Section 2. GENETICS ……………………………………………………….…..... 29
2.1 Mapping each and every trait ……………………………………….. 30
2.1.1 Univariate studies ……………………………………………… 30
2.1.2 Longitudinal studies …………………………………………… 36
2.1.3 Multivariate studies ……………………………………………... 40
2.1.3.1 Power …………………………………………………….... 41
2.1.3.2 Knowing what is being mapped ………………………….. 47
2.2 Trait Connections ………………………………………………………… 50
2.2.1 Connections through mosaic pleiotropy ………………………... 50
2.2.2 Concept of genetic correlation ………………………………….. 52
2.2.3 Localized genetic correlation ………………………………….... 53
2.2.4 Confounding with relational pleiotropy ……………………….... 57
Section 3. DESCRIPTION OF SPECIFIC AIMS ……………………………….... 61
3.1 AIM 1: The context ……………………………………………….…. 61
3.1.1 Uncaptured biological variation ………………………………... 61
3.1.2 The within and across individual distinction …………………... 63
3.1.3 Methods that have been described …………………………….... 64
3.2 AIM 2: The methods ……………………………………………………. 66
3.2.1 Assumptions of longitudinal study ……………………………... 67
3.2.2 The perturbation framework ……………………………………. 68
3.2.3 Correlations vs. slopes vs. covariances …………………………. 72
3.3 AIM 3: The data ……………………………………………………… 73

Section 4. THE CONTEXT ………………………………………………………..... 77
The within and across individual distinction in the genetics of correlated traits
Abstract ………………………………………………………………………… 77
4.1 Introduction ……………………………………………………………. 79
4.1.1 Unraveling causality …………………………………………….. 80
4.1.1.1 Causal relationships, a within individual process …………. 81
4.1.1.2 Across individual correlations can reflect within processes 82
4.1.1.3 Implications for methods based on causal model …………. 85
4.1.1.4 Evidence for need of within and across individual distinction ……………………………………………… 89
4.1.2 Unexploited Biological Variation ………………………………. 90
4.2 Data …………………………………………………………………….. 96
4.2.1 Increased false positive and false negative rates ………………… 96
4.2.2 The IL-6 gene: associative pleiotropy …………………………… 96
4.3 Methods ………………………………………………………………... 97
4.3.1 Increased false positive and false negative rates: data analysis …. 97
4.3.2 Increased false positive and false negative rates: simulations …... 98
4.2.3 The IL-6 gene: associative pleiotropy ……………………………100
4.4 Results …………………………………………………………………..102
4.4.1 Increased false positive and false negative rates: data analysis ….102
4.4.2 Increased false positive and false negative rates: simulations …...102
4.4.3 The IL-6 gene: associative pleiotropy ……………………………105
4.5 Discussion ……………………………………………………………106
4.5.1 Unraveling Causality ……………………………………………. 106
4.5.2 Unexploited biological variation …………………………………109
4.6 Conclusion ……………………………………………………………... 112
Section 5. THE METHODS ……………………………………………………… 113
Methods for testing genetic effects on within individual trait correlations
5.1 Introduction ……………………………………………………………..113
5.1.1 Relationships within vs. across individuals ……………………... 114
5.1.2 Adding the within and across individual distinction to the genotype interaction model …………………………………......117
5.2 Models ……………………………………………………………… 119
5.2.1 Interaction Model ………………………………………………...119
5.2.2 Correlation model ……………………………………………….. 120
5.3 Methods ………………………………………………………………... 121
5.3.1 Simulations ……………………………………………………… 121
5.3.2 Data analysis …………………………………………………….. 123
5.4 Results …………………………………………………………………..123
5.4.1 Type I error ……………………………………………………… 123
5.4.2 Power ……………………………………………………………. 124
5.4.3 Analysis of real data …………………………………………….. 125
5.5 Discussion …………………………………………………………… 127
5.5.1 Type I error and power comparisons ……………………………. 127
5.5.1.1 Simulations ………………………………………………... 127
5.5.1.2 Data ………………………………………………………………… 128

5.5.2 Comparison of methods aside from type I error and power considerations 129
5.6 Conclusion 132
Section 6. THE DATA 133

Association to disease and genetic architecture of metabolic trait correlations

6.1 Questions addressed 133
6.2 Data 134
6.3 Methods 136
6.3.1 Association to disease 136
6.3.1.1 Univariate 136
6.3.1.2 Multivariate 137
6.3.2 Genetic architecture 139
6.3.2.1 Univariate 140
6.3.2.2 Multivariate 141
6.4 Results 143
6.4.1 Association to disease 143
6.4.1.1 Univariate 143
6.4.1.2 Multivariate 145
6.4.2 Genetic Architecture 145
6.4.2.1 Univariate 145
6.4.2.2 Multivariate 150
6.5 Discussion 152
6.5.1 Association of correlations to disease 152
6.5.2 Mouse data vs. human data 155
6.5.3 Genetic variation underlying correlations 156
6.5.4 Advantages of multivariate analyses 157
6.5.5 Other considerations allowed by CSS panel 160
6.6 Conclusion 161
Section 7. FUTURE WORK 162
7.1 Interaction and correlation methods 162
7.2 The multivariate shape phenotype 164
APPENDIX A: Metabolic Syndrome Criteria 168
Bibliography 169

List of Tables and Boxes

Table 2.1 Results using SEM model …………………………………………………... 55

Table 4.1 Parameter estimates for the within, across and total observed correlations of traits with BMI ……………………………………………102

Table 4.2 Associations of BMI and blood glucose within correlations to diabetes status ……………………………………………….............. 106

Table 6.1 Pair-wise correlations that retained significant association to coronary heart disease and diabetes ……………………………………144

Table 6.2 Association of multivariate correlation phenotypes to coronary heart disease and diabetes …………………………………………… 145

Table 6.3 CSSs with at least one QTL for one of the 10 pair-wise correlations analyzed ………………………………………………………….. 147

Table 6.4 The average significant effect size for each pair-wise correlation total effect size …………………………………………………… 149

Table 7.1 Interaction test with insulin as the predictor and with BMI as the predictor ……………………………………………………….. 163

Box 4.1 Trait correlations are independent of trait values ……………………… 92

List of Figures

Fig. 1.1 Relationships between metabolic traits …………………………………. 18

Fig. 1.2 Age standardized prevalence of cardiometabolic abnormalities by body size and sex ……………………………………………………..... 20

Fig. 1.3 Prevalence of insulin resistance vs. number of comorbid conditions …… 21

Fig. 1.4 Cumulative incidence of type 2 diabetes and cardio vascular disease stratified by BMI ……………………………………….............. 23

Fig. 2.1 Multivariate analysis, a graphical representation ……………………….. 41

Fig. 2.2 Linear combinations ………………………………………………….. 45

Fig. 2.3 Correlation caused by mosaic pleiotropy ……………………………….. 51

Fig. 2.4 Example 1 of structural equation model ………………………………… 55

Fig. 2.5 Example 2 of structural equation model ………………………………… 57

Fig. 4.1 Within and across correlations are free to vary independently …………. 84

Fig. 4.2 Different models relating a SNP to two traits …………………………… 86

Fig. 4.3 A spurious association ………………………………………………... 89

Fig. 4.4 Gene by gene interaction vs. effect of associative pleiotropy gene ……... 94

Fig. 4.5 Power when testing causality between a trait and BMI through mendelian randomization ………………………………………………...103

Fig. 4.6 False positive rate when testing for associations between SNP and trait while controlling for BMI ……………………………………….. 104

Fig. 4.7 Density plots for BMI – blood glucose within individual correlations separated by genotypic group for the IL-6 polymorphism ……................ 105

Fig. 5.1 Comparison of type I error and power for correlation and interaction methods ……………………………………………………………….. 125

Fig. 5.2 BMI- blood glucose within individual correlations for the Framingham heart study data ……………………….................... 126

Fig. 5.3 QQ-plots of p-values obtained for the effect of chromosome 7 SNPs on within individual BMI – blood glucose correlations ………………… 126

Fig. 6.1 Multivariate analysis of pair-wise correlations …………………………. 138

Fig. 6.2 Results of variable transformations …………………………………... 140

Fig. 6.3 Histogram of effect sizes presented by 21 CSSs standardized by A/J effect size ……………………………………………………………….. 148

Fig. 6.4 Plots of the 10 pair-wise correlations against their respective composite traits ……………………………………………………….. 150

Fig. 6.5 Venn diagrams that compare approaches to finding significant CSSs for a system of correlations ……………………………………………151

Fig. 7.1 Visualization of shape change across two vectors of pair-wise correlations ……………………………………………………………167

Acknowledgements

I am very grateful to my advisor Dr. Joseph Nadeau in the department of Genetics for his continuous support throughout the past three years and for his leap of faith when he decided to mentor me towards obtaining a degree in a department different to his own. His observation of how trait correlations vary across the chromosome substitution strain panel which he studies was the seed that turned into the present work. I am indebted to the Nadeau lab for providing me with very helpful feedback throughout these years, especially conducive to the improvement in the presentation of my work to non-statistical audiences.

I want to give a very special thanks to Dr. David Sinasac, a former member of the Nadeau lab for sharing his mouse data with me. It is an essential part of the dissertation and it could not have been completed without it.

Dr. J. Sunil Rao my advisor in the department of Epidemiology and Biostatistics always had words of encouragement for me throughout the entire process of developing this dissertation and none of it would have been possible without him. He always managed to find the time within his busy schedule to meet with me and give me the academic and moral guidance that I needed to move forward. I also want to thank him for his help in obtaining the IRB approval needed for access to the Framingham heart study data.

I would like to thank Dr. Catherine Stein and Dr. Xiaofeng Zhu, the other members of my committee who took the time to meet with me and to look through the papers and dissertation manuscript and to make comments that substantially improved the work as a whole. Dr. Stein was particularly thorough in this process and without her feedback this dissertation would still be suffering from pervasive and detracting errors without my even being aware of them.

Dr. Zhu allowed me to take part of his lab meetings throughout the years. This kept me in the loop with statistical genetics in general and provided a source of endless learning and motivation. Thank you Dr. Zhu.

I'd like to give a special thanks to Dr. Nathan Morris for meeting with me to discuss the merit of the idea behind this dissertation from its inception. His critique and insight helped me in the development of my arguments and my ideas. Likewise I would like to thank Dr. Robert Igo who took the time to look at the first paper with me and convinced me that it was worthwhile pursuing. Any errors that persist in this paper are mine alone.

I cannot overstate the importance of the moral support of my friends, Ashok Panneerselvam, Vinay Cheruvu, AnnalisaVanderWyden-Piccorelli, Allison Baker and Joseph Magnotta. Thank you so much for being there.

Finally I would like to thank the Department of Genetics and their training grant through which the work behind this dissertation was funded.

Co-morbidities as Quantitative Traits

Abstract

by

PAOLA RASKA

Complex disease sometimes has system-wide level impact by affecting a constellation of physiologically interrelated phenotypes rather than a single phenotype, resulting in a set of co-morbidities. The physiological connection between the phenotypes, and the consequent co-occurrence of morbidities, varies from individual to individual in a way that can affect diseases prognosis. As an example, when obesity and its associated morbidities, dyslipidemia, hypertension and insulin resistance, do co-occur, this co-occurrence increases risk of diabetes and coronary heart disease in a way not explained by the presence of each individual morbidity alone. In this work, the physiological connections for each individual are characterized by the correlation values that the phenotypes present in their repeated measurements throughout the individual's life. The variation in the within-individual phenotypic relationships, from individual to individual, can then be studied as a new quantitative trait.

First, this study shows that traditional genetic approaches which target variation in the phenotypic values do not capture the variation in within individual phenotypic correlations. Secondly, two approaches designed to specifically model the new quantitative trait are statistically compared. Finally, the biological relevance of the phenotypic correlations underlying obesity and its associated morbidities is investigated

using the Framingham heart study data (human data) and the C57BL/6J and A/J chromosome substitution strain panel (mouse data). It is found that these phenotypic correlations are associated to diabetes and cardiovascular disease in a way not explained by the phenotypic values alone. It is also shown that there is genetic variation underlying these phenotypic correlations and that it is distinct and independent from that underlying the phenotypic values.

This work concludes that approaches that exclusively model phenotypic values when studying the genetics of co-morbidities in complex disease may be missing out on the biologically novel and relevant information contained in their correlations. Pursuing the genetics of phenotypic correlations as a new quantitative trait is therefore a worthwhile endeavor.

INTRODUCTION

Can genetics help us understand what ties the system of traits that underlie complex disease together?

Modern genetics provides us with a tool for getting at unknown biology underlying human variation. As long as this human variation has a heritable (hopefully meaning genetic) component to it we can use today's advanced computational and genotyping technology to localize loci that are associated or linked to the variation. Once these genes are localized or "mapped", functional studies can then determine the role they play in the variation being studied, opening a window into the biological mechanisms behind this variation. In this way, genetics can provide a very powerful tool for biological inquiry. For instance, when applied to human disease, it can pave the way for new possibilities in interventions and for the ability to predict individual risk. Hence the holy grail of genetics: the promise of personalized medicine.

Complex disease poses a special challenge for this investigative framework and it is the reason it is qualified as complex. The complexity lies in the multifactorial nature of this type of disease. It is multifactorial on two counts. First, unlike "simple" mendelian disease where one or two major loci account for the majority of the genetic variation, complex disease is typically characterized by having multiple associated loci with small effects. This makes finding them much more difficult and it is one of the reasons the ability to replicate findings has been rather limited, creating much discouragement.

Methods that take into account this aspect of complex disease are imperative. There have been some promising developments along this line such as multifactor dimensionality reduction (MDR) (Moore et al. 2006) which takes into account epistatic interactions between loci, or gene set enrichment analysis (GSEA) (Subramanian et al. 2005) types of methods, which incorporate biological pathway knowledge into the mapping effort and in so doing take into account the joint action of loci. Although this is an important problem facing the study of complex disease, it is not the one this body of work will be focusing on.

This work will focus instead on the second aspect that makes complex disease "complex", and that is its phenotype. Some complex diseases are characterized by having a system level impact on the individual. This means that when present it has an effect on numerous phenotypes, not just one. Sometimes all the phenotypes are present together and sometimes only a subset of them are. What is invariably true is that the phenotypes that do appear in the individual influence each other through, and serve as indicators for, their physiological interconnections. Sometimes their interactions may have an important impact on the individual's prognosis and disease development. This makes the definition of complex disease fuzzy and difficult. Some examples of this are mental illnesses such as schizophrenia, the constellation of autoimmune disorders related to arthritis, and the group of metabolic related morbidities that are associated to obesity and which are sometimes called metabolic syndrome.

When faced with such complexity, what can we hope to gain from this genetics tool we described above? How should we approach the problem of finding the important genes underlying these fuzzy constellations of phenotypes? Two approaches come to mind. First, we may try to simply ignore the complexity by analyzing each phenotype individually. There is value in this in that it will be sure to paint at least a partial picture of the biology underlying the disease as a whole. But it will certainly not be the complete picture. ***In order to get at the complete picture we have to somehow use the tool of genetics to get at an understanding of the physiological interconnections behind the phenotypes.*** How and why are these phenotypes coming together? Could this provide clues as to how they are compounding risk for the individual, and how in turn, the individual's risk could be lowered and his prognosis improved?

Herein lies the objective of the present work. I will provide a framework with which genetics can be used to help us understand the physiological connections between complex disease co-morbidities.

In order to do this I will use obesity and its associated morbidities as an example of complex disease with which to test the ideas presented in this work. I will begin by showing through results from the literature that the first prerequisite for being able to apply genetics as a study tool is fulfilled in this disease, this being that our object of study *varies*. Again, the object of study here is not the individual phenotypes in themselves, but rather *the physiological connections* between them. It is this connection which must vary from individual to individual. I will continue by going through how genetics has been used for the most part in tackling this disease's phenotypes, how they have been mapped individually, but also how the problem of connection between them has been addressed until now.

As my first aim I will present the genetic epidemiological context of studying physiological connections between phenotypes. We will see how current genetic methods do not generally target this variation and how this is just one consequence of the general lack of distinction made between within individual and across individual processes. This lack of distinction will be shown to be a source of wider methodological issues in some of these approaches.

In my second aim I will propose two methods that do target the variation of interest by making the within and across individual distinction. I will compare their behavior through simulations and contrast their assumptions to the assumptions made by other methods.

In my third aim I will again address the existence of variation in the physiological connections between traits and the relevance of this variation to disease, but in a more direct way by inquiring real data. I will also explore whether this variation fulfills the second prerequisite for being able to use genetics as a research tool and that is, of course, whether it has a genetic component to it. I will additionally consider the question of redundancy of this new type of variation with the variation in phenotypes that has been targeted before with other methods, and finally, I will explore the benefits of conducting an integrated analyses of the system of phenotypic connections as a whole.

Organizational comment

I will set up these three aims more fully in the section following the literature review (section 3). After this, the three papers corresponding to the three aims will be attached. Some redundancy between the rest of the dissertation and the first paper will be observed because it is a paper that has already been written to stand on its own and to be ready to be submitted. The second and third papers on the other hand still rely on the rest of the dissertation to serve as their introduction.

Section 1. PHYSIOLOGY

The purpose of this section is to:

1. Show that the physiological connections between obesity and associated phenotypes vary from individual to individual
2. Show how this variation may be relevant in the individual's prognosis
3. Show an example of a plausible, yet not necessarily proven, mechanistic explanation underlying this type of variation
4. Challenge the concept of causal direction between the phenotypes

1.1 Obesity and associated phenotypes

As we mentioned in the introduction, complex disease generally consists of a constellation of phenotypes. This led the field of phsychiatric genetics to develop the concept of "endophenotype" as a way of facilitating the etiological dissection of complex disease. The idea is to divide the disease into more stable, measurable phenotypes with a clear genetic connection. In their review, Gottesman and Gould list the characteristics that an endophenotype should have: it should be associated to the disease, it should be heritable and co-segregate with the disease in families and it should be present in healthy and diseased individuals (Gottesman and Gould 2003). Other terms that have been used to refer to endophenotypes are biological markers and subclinical traits.

One example of a complex disease that is best studied as a collection of multiple endophenotypes is obesity and its associated morbidities. Obesity is defined as having a body mass index (BMI) of 30 or more. Globally, there are more than 300 million obese individuals and obesity rates have risen three-fold or more since 1980 in some areas of North America with 35.1% of adults now being classified as obese in the United States (Catenacci, Hill and Wyatt 2009). Its increasing burden on public health and health costs makes it an extremely relevant disease. In fact, the Plain Dealer recently reported that “if current trends continue, more than 50% of Ohio’s adults will be obese by 2018 and the cost to the state’s health system could be as much as 21.7 billion annually” (Sarah Jane Tribble, The Plain Dealer, November 17, 2009).

This impact on health is only true because obesity generally presents itself with three other morbid conditions:

1. Dyslipidemia: abnormal levels of lipoproteins in the blood
2. Insulin Resistance: normal levels of insulin do not produce normal glucose uptake by muscle, liver, fat and other tissues, leading to high glucose levels in the blood
3. Hypertension: high blood pressure

When these conditions co-occur in the same individual they increase the risk for cardio-vascular disease events two-fold and they increase the risk for diabetes five-fold (Eckel, Grundy and Zimmet 2005).

Because of their co-occurrence and their joint effect on disease, the collection of morbidities are sometimes referred to as a “Metabolic Syndrome” or a “Metabolic Disease”. There are many differing criteria for defining this disease and these definitions and their clinical use are beyond the scope of this work. Despite this, I will be reviewing how these morbidities present themselves separately or together in individuals and in order to do this I will be using studies that refer to individuals that present some of the morbidities together as “metabolically diseased”. For these instances I will make clear the critieria that the particular study used when applying this label.

The endophenotypes as defined by Gottesman and Gould, or the biological markers for all of these morbidities are, cholesterol (CHOL), triglyceride (TG) and high density lipoprotein (HDL) levels in the blood for dyslipidemia, fasting blood sugar (BLSUG) and insulin levels for insulin resistance (INS), systolic (SBP) and diastolic blood pressure (DBP) for hypertension and body mass index (BMI) for obesity. From now on I will be referring to these markers as traits, or “obesity related traits”.

These traits influence each other within the individual through their physiological connections. Before getting into how these trait relationships vary from individual to individual it is desirable to paint a picture of what these relationships are generally thought to be. So the following may be, as we will see later, an oversimplified description of how these traits are physiologically connected as described in the Lusis et al. review (Lusis, Attie and Reue 2008).

.

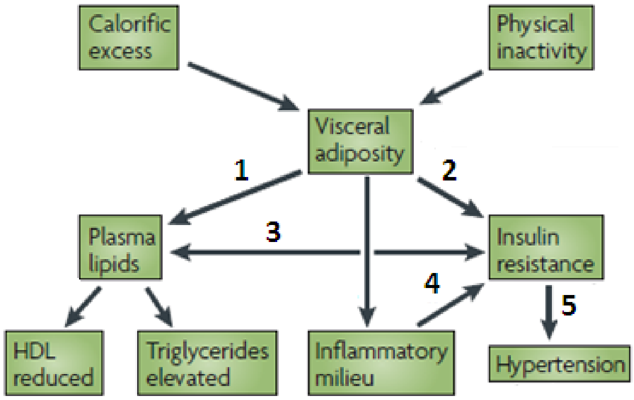


Fig. 1.1 Relationships between metabolic traits. Adapted from Box 1, in Lusis, Attie and Reue 2008, Adapted by permission from Macmillan Publishers Ltd: Nature Genetics Reviews, copyright 2008.

Legend for Figure 1.1

1. Increase in fat influences lipoprotein levels, for instance, increased flux of free fatty acids to the liver stimulates production of triglycerides
2. The increase in fatty acids and cytokines causes increased insulin resistance.
3. Hepatic insulin resistance causes increased production of triglycerides, while elevated fatty acids cause increased insulin resistance.
4. The proinflammatory state cause by excessive fat contributes to insulin resistance.
5. Activation of sympathetic nervous system by obesity and insulin resistance causes hypertension.

From this synopsis it seems that all of these conditions follow from increased fat due to excessive caloric intake and reduced physical activity. There appears to be a clear unidirectional causal connection, where it is all downstream from obesity. But it is in reality much more complex than this for two reasons:

1. Individual variation: for example there are healthy obese and morbid normal weight individuals.
2. Chicken and egg problem: determining which condition must be present first for the other to develop, what can be an indicator of causality, is not straightforward.

1.2 Heterogeneity in obesity and its relationship to other phenotypes

Obesity can be present in the absence of dyslipidemia, hypertension and insulin resistance and these three conditions may also be present in the absence of obesity. In fact, 31.7% of obese individuals are metabolically healthy and 23.5% of normal weight individuals (BMI < 25) are metabolically abnormal according to the analysis of the National Health and Nutrition Surveys 1999 – 2004 (Wildman et al. 2008). Metabolic abnormality in this study was defined as having 2 or more of the following: hypertension, elevated levels of triglycerides, glucose or insulin resistance, low HDL and high systemic inflammation gauged by level of high-sensitivity C-reactive protein. Figure 1.2 shows the Wildman et al. data broken down by gender.

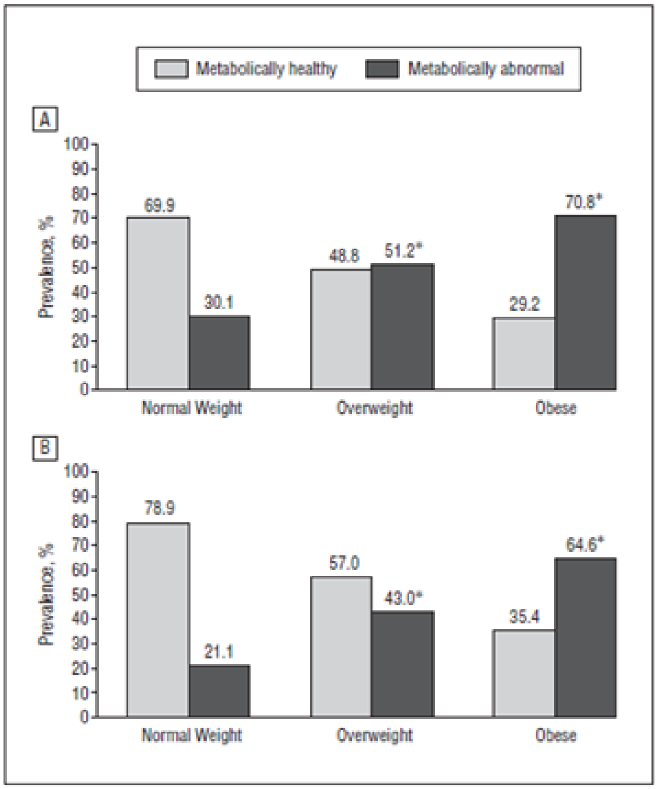


Fig. 1.2 "Age standardized prevalence of cardiometabolic abnormalities by body size and sex, A = women, B = men, *P<0.001 for proportion metabolically abnormal vs. normal weight." Taken from figure 2 in Wildman et al 2008. Reprinted with permission from the American Medical Association, copyright 2008.

The review by Sims also serves to show the heterogeneity in the co-occurrences of obesity and related conditions (Sims, 2001) (see figure 1.3). Although it is evident that obesity and insulin resistance tend to present themselves along with other conditions and when not present, individuals tend to be metabolically healthy, there is also substantial variation around this trend. All of the studies in this section provide evidence that one of the morbidities, obesity, can be separated from the rest. This is especially unexpected of obesity in particular because of its apparent causal role behind all the other conditions. But in addition, these studies show that like obesity, dyslipidemia, insulin resistance and hypertension can all present themselves independently of the rest within the individual.

The entire gamut of possible combinations of these four morbidities can be found in the population despite the fact that there is a general tendency for all of them to co-occur. *What factors make them present themselves together or independently from individual to individual are unknown.*

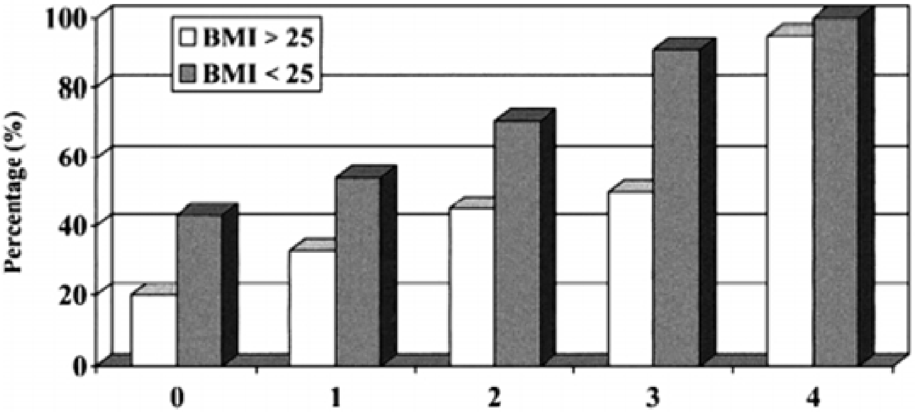


Fig. 1.3 Prevalence of insulin resistance on the y axis and number of comorbid conditions on the x axis. Comorbid conditions included are impaired glucose tolerance, T2D, dyslipidemia, hyperuricemia and hypertension. Taken from figure 1 in Sims, 2001. Reprinted with permission of Elsevier, copyright 2001.

In this section I am focusing on the link between obesity and the other morbidities because of the huge role that it is generally given in the whole of metabolic disease. Despite this, the arguments apply to the relationships between all the other phenotypes as well. The variation underlying all of these relationships has the potential of having a genetic component and of being biologically informative.

1.3 Importance of heterogeneity in disease prognosis

How exactly can this type of variation be informative? Let’s take for instance the occurrence of the healthy obese individuals. It brings into question the directional physiological relationships described above. Why does obesity causally contribute to the occurrence of the other morbidities in some individuals and not in others? What is it in their physiology that disconnects these phenotypes? Why is it that they can gain weight and maintain an otherwise healthy metabolism? Understanding how these individuals are genetically different from others may help give mechanistic/functional explanations to these questions. This in turn could potentially lead to the design of interventions for morbid obese individuals.

Additionally, simply knowing the individual’s value for a trait can sometimes not be enough in terms of determining his or her prognosis and the proper course of care. For example, Sims explains that “treatment” of a healthy obese individual can have negative effects that are similar to those that starvation would have in a normal weight individual (Sims 2001). Knowing how they are genetically different can therefore serve to predict what type of treatment may be more in line with the individual’s physiology regardless of obesity thereby providing an avenue for more personalized medicine.

Meigs et al. uses the Framingham study offspring generation data to look at how the presence of these morbidities influences the risk of type 2 diabetes (T2D) and cardio vascular disease events (CVD). They particularly look at whether obesity confers a higher risk on its own or whether it needs the other morbidities. Like the previous studies, Meigs et al. show that there is a prevalence of metabolically diseased normal weight individuals. Moreover, they show how these individuals *are at more risk for CVD than the obese with metabolic disease* (see figure 1.4). (They use the ATPIII criteria for determining metabolic disease, see Appendix A)(Meigs et al. 2006). Again, this is a case in which proper prognosis and treatment cannot be gauged by the levels of the traits on their own but rather by how the phenotypes combine and relate within the individual.

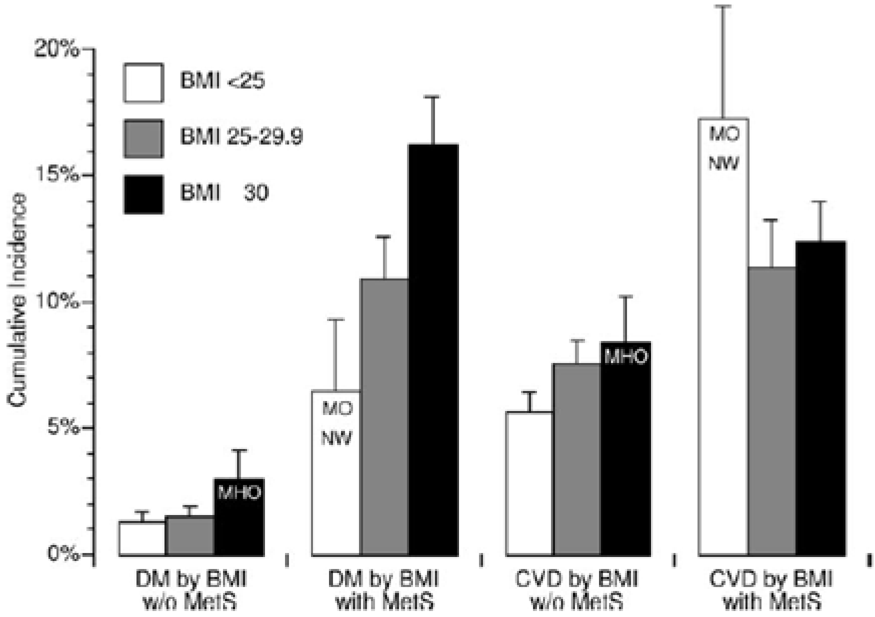


Fig. 1.4 Seven year age-sex adjusted cumulative incidence of type 2 diabetes T2D and 11 year adjusted cardio vascular disease (CVD) stratified by BMI and the absence of metabolic syndrome as defined by the ATP III criteria. MHO are obese subjects with metabolic syndrome and MONW are normal weight subjects with metabolic syndrome. Taken from figure 1 in Meigs et al., 2006. Reprinted with permission from The Endocrine Society, copyright 2006.

1.4 Plausible mechanistic explanation underlying heterogeneity

Chandalia and Abate find one genetic factor underlying this heterogeneity by studying a subpopulation which is predisposed to presenting metabolic morbidity in the absence of obesity, the Asian ethnic group. They determined that a point mutation in the gene encoding for ENPP1, a type II transmembrane glycoprotein which when overexpressed in cells impairs insulin receptor signal transduction, is much more common in migrant South Asians. They suggest this mutation may explain the development of insulin resistance and adipose tissue dysfunction in this group. They also show how this mutation predicts T2D in a different ethnic group as well as in the same ethnic group but in a different environment, providing support for its functional role. They conclude in their study that adipose tissue ***dysfunction***, not ***quantity*** is what may underlie metabolic morbidity (Chandalia and Abate 2007).

Following up on the idea of a predisposition to either adipose tissue dysfunction or robustness, one explanation for the Meigs et al. finding is that for an individual to present metabolic disease without obesity they must have some predisposition to adipose tissue dysfunction that all obese individuals do not necessarily have. Having this dysfunction may be what truly increases the risk for cardiovascular disease event.

While Chandalia and Abate provide an example of a genetic factor that may underlie the existence of individuals that are metabolically diseased but have normal weight, a potential *mechanistic* explanation behind the metabolically healthy obese and the concept of adipose tissue dysfunction, is given by Le Lay et al. (Le Lay, Ferre and Dugail 2003). In their paper the authors explain how there is a strong correlation between adipocyte cholesterol content and *fat cell size* despite the fact that the ratio of cholesterol content to triacylglycerol in the adipocyte is independent of BMI. Target genes of the SREBP transcription factors which are those responsible for regulating the transcription of genes needed for the uptake and synthesis of cholesterol, show an increased expression in the adipose tissue of obese individuals and rodents where adipocytes present a high cholesterol content and are enlarged. What these authors interestingly found is that these genes present a normal expression level in a transgenic obese mouse in which fat cell size was normal. An increase in *fat cell number* as opposed to *fat cell size* is one potential explanation for how some individuals can gain weight without the metabolic effects associated with obesity.

This is an illustrative example of a potential mechanism that may separate obesity from the other morbidities. It may or may not be the real or only explanation behind the healthy obese and the morbid normal weight individuals, but regardless, what it does provide is a way to showcase the relevance that the type of biological variation pursued in this work may have. It also provides a clue as to why finding associated genes pose a special challenge. In this example, an individual’s adipose tissue may have a tendency towards dysfunction or robustness. Some individuals like the southern Asians in Chandalia’s study may be more predisposed to adipocyte dysfunction independent of BMI while others may be predisposed to augmenting number of cells rather than size of cells and therefore maintaining a healthy adipose tissue independent of BMI. This is an example of a factor that would not be found through searching for associations to BMI directly or any of the other morbidities.

1.5 Challenging unidirectional relationships

An additional challenge in determining how these phenotypes relate becomes apparent when the causal direction between them within the individual is studied more closely.

1.5.1 *Hypertension causing metabolic disease*

Of all the morbidities, hypertension is probably the one that can be least thought of as being potentially causal of the rest, and yet even this directionality can be put into question. Julius and Jamerson do just this in their hypertension review paper. It has been shown that slight elevations in blood pressure in 40 + year olds were already apparent since their seventh year of age (Julius et al. 1990). The fact that hypertension is such a slow and gradual process combined with the fact that it has quick and systemic wide effects in the organism led them to ask the chicken and egg question: what comes first, insulin resistance and/or dyslipidemia which then cause hypertension, or does the enhanced sympathetic drive that causes hypertension then contribute to insulin resistance and dyslipidemia? Julius and Jamerson propose the latter to be the case (Julius and Jamerson 1994) and they provide three alternative pathophysiological mechanisms that have experimental support. Granted, all of these mechanisms can apply as the *initial* cause of morbidity only in those individuals that present an early sympathetic overactivity, which does not include the majority of hypertensive individuals (does not explain it in ~70% of individuals). But this demonstrates how looking for simple cause and effect relationships between interrelated physiological traits may be as futile as the chicken and egg question. Regardless of which comes first, as this may vary from individual to individual, what remains true is that once morbidity in one trait is established it can contribute to morbidity in the others and vice-versa, and this back and forth relationship can continue throughout the individual's life.

1.5.2 ***Insulin resistance causing obesity***

As already stated, one of the most unquestioned directional relationships in the picture of obesity and its related morbidities is that of obesity being causal of all the rest. Even though we have shown with the existence of the healthy obese and the metabolically morbid normal weight individuals, that obesity is not necessary or sufficient for the development of insulin resistance, hypertension and dyslipidemia, it may still be argued that when all four morbidities are present, obesity can be contributing to the other three morbidities but not the other way around. The Lazarus et al. study brings this assumption into question. They studied the temporal relations between obesity and insulin resistance using longitudinal data in order to determine a causal relationship between the two (Lazarus, Sparrow and Weiss 1998). Although weight was found to be a significant predictor of insulin, *insulin* was also found to be a significant *predictor* of weight. As the authors state, this suggests that the relationship between these two traits may constitute a complex bidirectional feedback rather than unidirectional causality. The carbohydrate craving produced by insulin resistance has been proposed as a potential mechanism by which insulin resistance may cause obesity and it has some experimental support in humans. Interestingly, the authors point out: “Physiologic counter-regulatory mechanisms must also be operating, because otherwise there would be a vicious cycle of increasing insulin levels and increasing obesity.”

These counter-regulatory mechanisms must also be present in the sympathetic drive – insulin resistance/dyslipidemia back and forth described above. In fact, they should be there for all the traits. They are what keep an individual’s metabolic homeostasis. Some individuals seem to have systems that are more robust to environmental perturbations (for example, instances of higher caloric intake) while others are more predisposed to having these positive feedback relationships take over more readily, pushing the individual’s system into a morbid metabolic state. It can be hypothesized that in the first case scenario, the traits will tend to present less variation and there will be less association between the traits. In the latter individuals, the traits will present higher variation and higher association levels, or correlations. This reveals the first clue as to how we may go about capturing the type of biological variation that has been showcased in this section.

Section 2. GENETICS

Now that we have a better idea of the complexities that underlie a real system of morbidities with physiologically interrelated phenotypes, we can better gauge the effectiveness of applying particular methodologies to the study of the genetics behind the system of traits.

In the introduction we mentioned that two questions may be asked:

1. What are the genetics underlying each and every single trait?
2. What are the genetics underlying the physiological connections between the traits?

This work is concerned with addressing the second question. We will begin by describing studies that have been conducted with the first question in mind. We will start out with univariate studies which look at each trait individually, continue on with longitudinal studies, which look at what each trait is doing through time, and we will end with multivariate studies which analyze all the traits together but still only look to answer that first question: what genes underlie each and all the traits.

Then we will focus on the problem of connection between the traits and how genetics has been used to address this question.

2.1 Mapping each and every trait

2.1.1 *Univariate studies*

The purpose of this section is not to give a thorough review of all the genetic studies carried out on obesity and related phenotypes. Rather I will show some univariate results for each of the phenotypes and illustrate some of the limitations involved in not taking into account the relationships between the traits when studying their genetics. Because it is the data set that this work will be analyzing, I will focus on, yet not completely restrict myself to, studies conducted on the Framingham heart study data set.

The Framingham heart study data set has spurred just under 2000 articles to date. The study began in 1948 led by the National Heart, Lung and Blood Institute. It is still carrying on today in collaboration with Boston University (Cupples et al. 2009).The objective of the study is to understand the factors underlying cardio vascular disease. So far three generations from Framingham, Massachussettes have been recruited and are currently being referred to as the cohort, the offspring and the Gen 3 generations (n=5209, n=5124, n=4095). The cohort generation has been given a physical examination and laboratory test every 2 years since recruitment, while the offspring generation has had examinations every 4 years. The third generation has only had one examination. These individuals span over 900 pedigrees. For the offspring generation spouses were recruited while for Gen3 they were not. Dense genotyping, ~550k SNPs, was also performed in approximately 10775 samples across the three generations (GeneChip® Human Mapping 500K Array Set and the 50K Human Gene Focused Panel). Included in the physical examination and laboratory testing data are the endophenotypes of interest in this study as well as covariates of interest such as use of medications, smoking, alcohol use, physical activity, age and sex (Cupples et al. 2009) (http://www.gaworkshop.org/README_Prob2_FHS_031908.pdf).

Standard analyses for getting at the genetics underlying obesity and related phenotypes are univariate in nature. Their objective is to map genes that associate to one of the phenotypes irrespective of the others, and as mentioned in the introduction, this is a worthwhile objective in that it tells part of the story. I will go through examples for each one of the morbidities, starting with the dyslipidemia traits, following with blood pressure and ending with insulin. I have reserved the study on BMI for the longitudinal section that follows because it serves to illustrate a point here.

For instance, Kathiresan et al. looked for associations to the blood lipid phenotypes of triglycerides (log transformed), HDL and LDL individually (Kathiresan et al. 2007). They used the mean values of these phenotypes for individuals having 4 or more measurements in the Framingham data set for the offspring generation. Two models and a 3 stage replication strategy were used. In the first model they only adjusted for age and age^2 and used the sex specific residuals to select a subset of most significant SNPs. They carried these along to the second stage in which they then used unrelated Framingham individuals for replication. For the third stage they selected yet another subset from stage 1 and 2 combined, and then use the GOLDN and MDC-CC data sets for replication. Here they switched to the second model where in addition to age they adjusted for smoking, alcohol, menopausal status and hormone replacement therapy and BMI. After their three stages of replication there was no convincing statistical evidence for association.

A pattern that we will see in these studies is their difficulty in being replicated. As stated in the introduction, this is in large part due to the small, single gene, *marginal* effect sizes associated with complex disease as well as genetic and environmental heterogeneity. Unfortunately addressing this issue is beyond the scope of this work and we will be subject to the same limitations when analyzing the data in a SNP by SNP fashion. Even though, we may still be able to find associations that at the very least only apply to this data, and we will be able to make the necessary comparisons to previous results using other methods. We will also see how these studies consistently “control” for BMI. BMI then is treated as an environmental variable that can lead to heterogeneity rather than a physiologically interrelated trait that has a genetic component of its own. We will explore this aspect more as we go along.

Levy et al. conducted a linkage analysis on the Framingham cohort and offspring generations (Levy et al. 2000). They used mean blood pressure measurements for individuals with at least 4 separate measurements and with at least a 10 year span between their first and last measurement. They adjusted this phenotype for age, and BMI and analyzed the sex and generation specific residuals. They obtained a LOD score of 4.7 for an interval on chromosome 17 and list as corroborating evidence for this interval results from previous human studies of hypertension and the fact that it contains rat and mouse homologues and good candidate genes. Levy et al. repeated this analysis in the Framingham study 100 K Project, a project developed to provide a web based resource to genome-wide (100 thousand SNPs) analysis results on the 987 phenotypes collected in Framingham throughout its 56 years (Levy et al. 2007). Thet could only replicate Levy et al.'s original result on chromosome 17 when modifying their long-term systolic blood pressure definition to include earlier cohort generation values and exclude later offspring generation values. Levy et al.'s earlier result was therefore contingent on the subset of the Framingham data that was used for the analysis. None of the associations in this new study reached genome-wide significance. Again in 2009 the search for a genetic association by Levy et al. was conducted, this time using the CHARGE consortium (n = 29,136), which includes the Framingham data, and then combining this in a meta-analysis with the Global BP gen consortium (n=34,433) (Levy et al. 2009). They identified 13 SNPs for SBP, 20 for DBP and 10 for hypertension in CHARGE and these were reduced to 4,6 and 1 SNP respectively after the meta-analysis. This progression of studies shows the brute force approach of conducting simple univariate analyses in bigger and bigger data sets all in an attempt at replication. In all studies the phenotype itself did not vary and in all the phenotype was controlled for BMI.

Kathiresan et al. also conducted a posterior GWAS and they were able to determine after using extensive sample sizes replication (n = 19,840 for first round, n = 20,623 for replication) 30 distinct loci involved with polygenic dyslipidemia. They did not adjust for BMI this time although they do not explain why, instead only controlling for age, age^2, population stratification using principal components and stratifying by sex. (Kathiresan et al. 2009)

Panhuysen et al. conducted a genome-wide scan to look for loci linked to insulin traits in non-diabetic Framingham offspring generation individuals (Panhuysen et al. 2003). They found suggestive evidence of linkage on chromosomes 9, 11, 17 and 19. They conducted their analyses on four models, one including BMI as a covariate and another model excluding it, both with and without full adjustment for other covariates. This study is the only study out of those researched for this dissertation where the author gives a thoughtful account of what the consequences of controlling for BMI may be. The LOD scores on chromosomes 9 and 11 decreased dramatically when adjusting for BMI and the authors hypothesized that the linkage to these regions is mediated in part by obesity. Adjusting for BMI in these cases may or may not be desirable depending on whether or not you want to control for these mediating effects so that only loci with marginal effects on insulin are detected. The authors pointed out though that there may be genes that have independent effects on both insulin and BMI and that in this case adjustment would obscure linkage signal. A third possibility arises in the evidence they presented for linkage to insulin on chromosome 11 which also drops upon adjustment with BMI. The authors stated that although independent studies have linked this region to BMI, it is not clear in their case "whether obesity leads to insulin resistance, or whether hyperinsulinaemia leads to obesity". The iddm4 locus for T1D is close to this signal they argued, suggesting that a gene in the region can be causing both T1 and T2 diabetes instead of obesity.

BMI can therefore be a mediator in effect, it can be affected independently or it can instead be caused by insulin. We will see these three possibilities more closely in the methodology section. The other univariate studies do not take these possibilities into account. It is of special note how the first Kathiresan study attempted to replicate SNPs that had been obtained without BMI adjustment in a separate data set using BMI adjustment, without realizing that they may have been mapping different genes in both cases, greatly reducing their chances of replication.

Herbert et al. shows yet another level of complication that can come about when considering the genetics of correlated phenotypes (Herbert et al. 2006). They looked at the interleukin (IL) – 6 gene polymorphism and its association to insulin resistance in the offspring generation. They found that the effect of BMI on insulin resistance in men depended on the IL-6 genotype, with higher insulin resistance in individuals with the CC instead of the GG genotype for high BMI (>27 kg/m^2), while at low BMI the CC genotype had lower insulin resistance than the GG genotype. They likewise found that genotype at this locus modulates association of BMI with diabetes prevalence, with the GG and GC genotypes being less affected by high BMI. This study then showed a fourth possibility with BMI, that of an interaction between BMI and the associated locus. We will revisit this possibility in our first aim.

What can be concluded from this section through our focus on issues with BMI, is that in general, it is important to take into account the relationships between physiologically connected traits even when conducting genetic studies in which the only interest lies in mapping genes associated to one of the phenotypes. The only way to do this is via the multivariate approaches which we will see in the integrative methodology section. We will outline several reasons for taking these relationships into account.

2.1.2 ***Longitudinal studies***

As previously mentioned, the Framingham cohort and offspring generations have data from periodic examinations. This means that not only are there measures of blood pressure, blood lipids, fasting insulin and glucose and BMI for each individual, but there is also information on how these measures change through time as the individual ages. One analysis that exploits this type of information is longitudinal analysis. Because it is a methodology that has been applied to the data and the phenotypes of interest, and because it presents apparent similarities with the methodology that will be proposed, I will use this section to describe previous longitudinal study results on obesity and related traits in the Framingham data. By doing this I hope to give a sense of what types of inquiry the methodology is generally used for. A more in depth contrast to the methodology being proposed in this work will be offered in the first aim. Like the univariate studies described above, the following studies analyze each trait independently of the other.

For example one longitudinal study conducted on the Framingham data is by Franklin et al. (Franklin et al. 1997). They looked at patterns of change of SBP and DBP throughout life using the cohort generation. They found a linear rise in SBP all throughout life (30 - 84 yrs) as well as an increase in DBP up until ~60 yrs of age after which there is a decline. They found that this pattern of having a late decline in DBP is magnified in individuals with overall high SBP. This led them to hypothesize that this decline may be caused by large artery stiffness that results from high SBP.

Here we see a longitudinal study that looks at two traits individually and how they relate through time. At a first glance then, it seems to fit well with our stated objective of wanting to capture variation in the relationships between traits. All that is missing is a search for a genetic component to the observed variation.

Another example of a longitudinal study but that in addition incorporates genetics is Strug et al. (Strug, Sun and Corey 2002). They conducted a linkage analysis on both mean BMI, which collapses the longitudinal information into one single measure (which is what many of the univariate studies mentioned above did), and on slope of BMI and mean gain of BMI, which summarizes the individual's weight gain throughout life (they excluded the period towards the end of life where some weight loss is encountered for their slope calculation.) They found a strong signal (LOD = 3.52) for their mean gain of BMI measure on chromosome 4 but failed to find anything with a LOD score greater than 3 for mean BMI. This study seems to suggest that what a trait does through age is a distinct trait from what it is on average for any particular individual and that it is a trait that may be worth looking at genetically.

Let's look at this concept more carefully. In contrast to the longitudinal approach, analyses that use the average value of the trait across all time-points, such as those summarized in the univariate analysis section, exploit the repeated measures aspect of the data in a different manner. Instead of considering each time point as a separate piece of information on its own, where the measure for your trait at time-point A is for a distinct quantity to that being measured at time-point B, these studies consider the values of the trait at all the time-points in the individual's life to be measures of the same quantity, only differing due to measurement error.

Which approach to take with repeated measures data may simply depend on the question of interest. But the data may also be inquired in this regard. For instance, we can compare the information at different time-points and evaluate how redundant or distinct they are before deciding whether to treat them as separate estimable quantities or not. The following studies offer such comparisons.

Havill and Mahaney looked at blood pressure (SBP), cholesterol and weight to see how much genetic variance was *shared* between two different age groups (30-39 yrs and 50-59 yrs) (Havill and Mahaney 2002). They found that the age groups shared 96% of genetic variance for weight, 57% for cholesterol and 20% for blood pressure. Only the heritability of SBP changed significantly with greater heritability in the older age group. In summary, there appeared to be a difference in which genes affect each age group and the level of expression of these genes across the two age groups for blood pressure.

Kraft et al. focused exclusively on blood pressure measurements across time-points in the Framingham data. They also found evidence that different genes affect blood pressure at different ages but unlike Havill and Mahaney's 20% shared genetic variance, they find 82% shared genetic variance between age groups 35-50 and 50-65 (Kraft et al. 2002). They also found that considering the two age groups in the linkage analysis (bivariate analysis) does not increase their power for identifying loci compared to using only one age group measure; *both* approaches peak at a LOD~ 2.2-2.4 and at the same location. They repeat Levy's linkage result on chromosome 17.

Atwood et al. looked for linkage to six separate measurements of BMI in the offspring generation. They found substantial evidence for linkage on chromosome 6 and chromosome 11 for all six measurements (Atwood et al. 2002). They concluded that linkage studies of BMI are robust to measurement error although there was some variation in LOD scores for the six measurements.

Some of the studies that used long term averages of the traits compared the heritabilities for single measurements with heritability for the average (Levy et al. 2000) (Kathiresan et al. 2007). The average SBP and DBP phenotypes had heritabilities of 0.57 and 0.56 while single examination heritabilities were 0.42 and 0.39 respectively. Average LDL, HDL and triglycerides were 0.66, 0.69 and 0.58, while they were 0.59,0.52 and 0.48 respectively for a single time point.

In summary, heritability is improved when looking at the mean of the traits rather than the individual time points, and different time-points tend to provide linkage to the same regions. These results may serve as support for looking at the repeated measures of the

Framingham data as measurements of one single heritable quantity with added error variance across time-points.

Furthermore, revisiting the Strug et al. study described above, although they emphasized the stronger LOD score for their mean gain of BMI measure, looking at their paper more thoroughly reveals that the mean BMI presents a suggestive LOD score for the *same* region. Additionally, their BMI slope, which is the only measure they use that actually incorporates time in some manner, has LOD scores that are lower than mean BMI. In conclusion, the Strugs et al. results may also be taken as support for the error variance model over the distinct time-point model, albeit with possibly greater error at the latter period of the individual's life.

2.1.3 ***Multivariate studies***

We mentioned in the univariate section how even when mapping individual traits it may be wise to take phenotypic relationships into account. Here we will see two reasons why not doing so may be problematic:

1. Power: Taking the multivariate structure of multiple phenotypes into account may enhance the power to detect genetic associations to any and all of the individual traits.
2. Knowing what is being mapped: associated loci may in reality be primarily associated to one of the correlated phenotypes not being taken into account.

We will look at multivariate approaches designed to address the issue of power and then we will look at how the relationships between the traits may affect studies that are only concerned with mapping the individual traits. This will provide a good sedgeway into methods that directly model these relationships when we switch our discussion to the study of the genetics underlying the connections between traits.

2.1.3.1 Power

To understand why a multivariate approach may enhance power we should look at a graphical example:

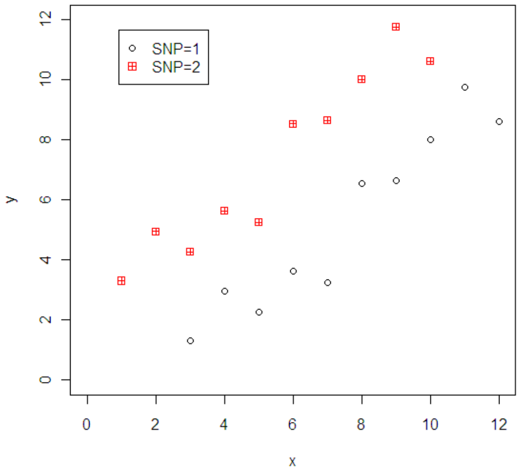


Fig. 2.1 Multivariate analysis, a graphical representation

Suppose x and y represent two correlated random variables and we are looking for evidence of their association to a SNP. (For the sake of simplicity we assume that the SNP is for a dominant gene with only two phenotypes, so for example, SNP=1 would refer to genotype CC while SNP=2 would refer to genotypes CG and GG). This hypothetical example is plotted in figure 2.1. If you look only at the values of x you will see how they completely overlap for the red and black groups representing alternative alleles for the SNP. This also happens for the values of y. This shows how in a univariate analysis of traits x and y the effect of this SNP would be undetectable. Yet from viewing the plot it is obvious that there are two distinct groups for alternative SNP alleles. In fact they do not even overlap, so there is clearly a SNP effect.

A bivariate analysis of this data on the other hand would reflect this separation between groups and would present high significance. An intuitive way of understanding this is to think of a line that can separate the two groups. In the univariate case you can only use a line along the x axis or along the y axis for x and y respectively. In both cases, the two groups as we said are not easily separable. In a bivariate analysis you are free to draw the line in any direction within the x-y plane. A line along the x=y axis can separate the two groups perfectly.

This is the reason multivariate analyses have increased power compared to univariate analyses when the phenotypes being studied present a correlation. For instance, in the case of obesity and associated phenotypes, their physiological connections produce a multivariate structure that can be exploited for power.

One example which showcases this increase in power is Arya et al. They conducted univariate and bivariate analysis of HDL levels and BMI in data from 1702 subjects distributed of both the cohort and offspring generations, using data from one examination for each. Their univariate analyses implicated a ~8cM region on chromosome 6q that influences both ln BMI and ln HDL (measures were transformed to minimize problem of non-normality). They found that the bivariate analysis improved power to co-localize the two phenotypes more precisely within this region.

One Important distinction that has to be made is that of using multivariate data to increase mapping power and using it to test for pleiotropy (Allison et al. 1998) which is defined as one gene having an effect on more than one phenotype. The Arya et al. study did both but here we are strictly referring to the power advantage offered by their multivariate approach. More about their testing for pleiotropy will be discussed below in the "genetics: trait connections" section.

Methods have been developed that also exploit the multivariate structure but without the need of a full multivariate analysis. They are all variations on the same idea which consists of coming up with a composite univariate trait or several composite traits that can then be analyzed univariately. To continue with our graphical example, a line in the x-y plane is in effect a "linear combination" of the two traits. We can analyze how the data is distributed along any line, or linear combination of the two traits, i.e. "using a composite trait", just as well as we generally do so by using the x and y axes, i.e. "the original traits". By allowing the freedom to choose any line on the x-y plane, this method although univariate, is able to exploit the multivariate structure in the data.

Variations on this idea differ on the criterion used to choose the line (or lines). For instance, one approach called principal component analysis (PCA) uses the direction of maximum variance in the data to choose the line – the first composite trait, or “principal component”. Subsequent lines, or principal components are chosen with criteria of maximum variance and perpendicularity to previous principal components.

Two examples of PCA analysis for the study of the genetics underlying obesity and its related traits that were conducted on different populations are Arya et al., on non-diabetic Mexican Americans (Arya et al. 2002), and Cox et al. on a Norfolk isolate (Cox et al. 2009). The first study included eight phenotypes: fasting glucose and insulin, BMI, systolic and diastolic blood pressure, HDL and triglycerides and leptin. The first factor or composite trait they analyzed, composed mostly by BMI, fasting insulin and leptin had significant linkage on chromosome 6 while their third factor, composed by triglycerides and HDL showed significant linkage in chromosome 7. The second factor was mainly composed by the blood pressures. The second study found a first factor for body size, a second one for cholesterol and triglycerides, a third for blood pressures and a fourth for cholesterol and LDL. They found suggestive linkage for the second factor only on chromosome 5.

Finally, and example of this approach using the Framingham data is the Liu et al. study. Their three factors grouped as BMI, systolic blood pressure and glucose for the first, HDL and triglycerides for the second and cholesterol and triglycerides for the third. They found significant linkage for the third composite trait on chromosome 2. When they ran the analysis on the original triglycerides and cholesterol traits separately the signal was lost, showing how taking into account the multivariate structure of these traits into the analysis increased power.

Another approach selects the composite trait, or the direction of the line, according to how it is maximally genetically associated. Going back to our graphical example in figure 2.2, the line that would best differentiate between the two SNP allele groups instead of the one that maximizes variance would be selected.

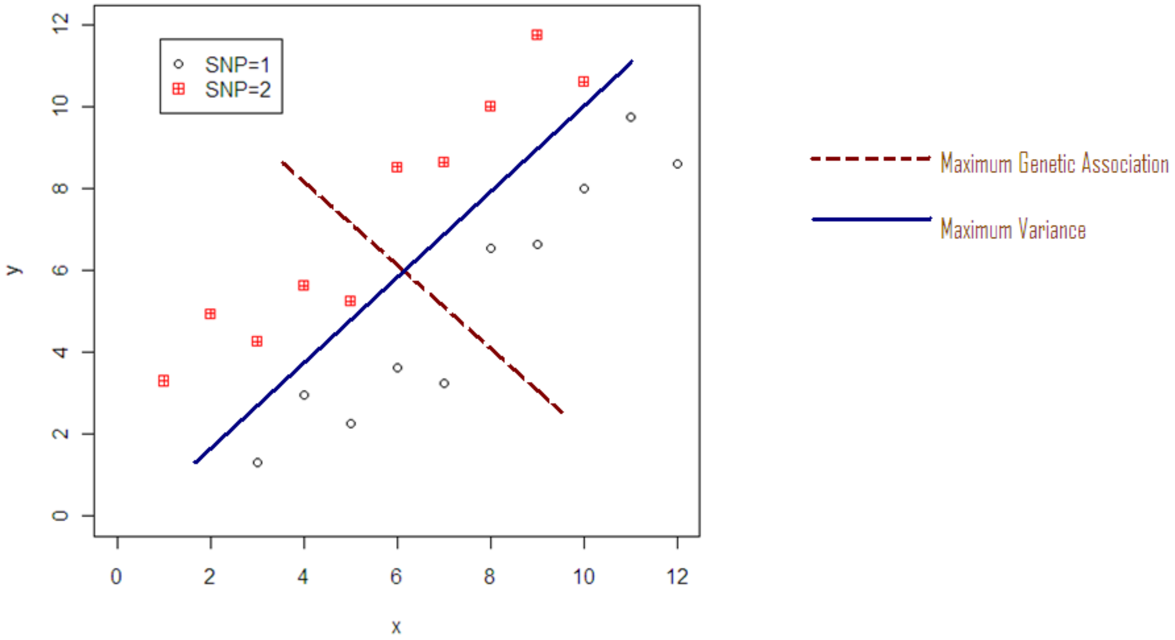


Fig. 2.2 Linear combinations

For example, Ott and Rabinowitz did this by optimizing according to overall heritability (Ott and Rabinowitz 1999). They explored this approach through simulations and determined that there is an increase in power when compared to a linear combination that simply maximizes variance as in principal component analysis. The limitation of this approach is that although the composite trait that it derives may be the most heritable choice, it may not be associated to any particular locus but rather to a large combination of loci.

Pedigree discriminant analysis looks to resolve this by looking for the linear combination that maximizes segregation (Miller et al. 1979, Elston et al. 1976).

This idea can be traced back to the proposed " genometric approach" where instead of looking for a gene that associates to a given phenotype, the search is instead for phenotypes that associate to a gene (Elston and Wilson 1990). The limitation with this approach is that "it is not possible for a single linear combination of traits to be powerful for all relevant loci" (Morris 2009) and to assume that there is a single genetic locus underlying complex disease is not believable.

In conclusion, the advantage of optimizing genetic association through linear combinations of the phenotypes is offset by the fact that it will either be artificially too localized or not localized enough in the absence of knowing the true genetic architecture that underlies the phenotypes.

Another general disadvantage to this group of methods is that reduction of the phenotypes into a composite trait compromises understanding how the traits are related. In other words, although they exploit the multivariate structure for power, they ignore it as a source of information of its own. Carrying out a full multivariate analysis is preferable if one wants to garner the additional advantage that may come from considering multiple phenotypes, such as, insight into the genetics underlying their relationships.

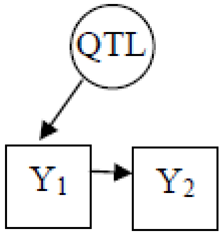
2.1.3.2 Knowing what is being mapped

We saw in the univariate section through examples in which BMI was controlled for when mapping one of its correlated traits, that many different things can occur if the relationships between the traits and the associated loci are not taken into account. For instance, the loci found to associate to the trait of interest may in reality be primarily associated to one of the correlated phenotypes not being taken into account. What's more, if the locus is primarily associated to BMI and BMI is controlled for, this will create a spurious association to the trait being studied! Moreover these relationships are also relevant when conducting studies on the individual traits because the power advantage that may be garnered by conducting a multivariate analysis depends on them.

Allison et al. usefully outlined all the possible case scenarios for the simplest of relationships: one gene and two traits. They also explained which models provide an improvement in power through multivariate analyses. Here we will concentrate on four different models (Allison et al. 1998):

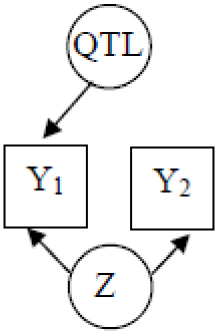
1. Relational Pleiotropy: The gene is associated to one of the traits directly and to the other trait indirectly through its relationship with the first. Joint analysis in this case will improve power over univariate analysis only in particular circumstances. One classic example of this type of pleiotropy is the FTO gene and its association

to BMI and Insulin Resistance. Although it was initially thought to directly influence diabetes prevalence, this association disappeared as soon as BMI was controlled for. The studies that we saw in the univariate section exclude loci that have this indirect influence on the trait of interest in their search when controlling for BMI.

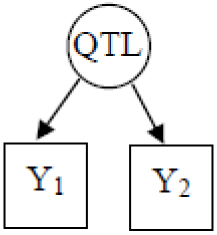


2. Latent Variable: This is actually just another model for relational pleiotropy and I am including it here to make the point that the relationship between the two traits themselves does not have to be direct. It can be caused by a latent, or unobserved variable that affects the two traits independently. This other variable could be environmental but it can also be another genetic (gene that is not being modeled). It is important to point this out because it brings into question any assumption about the directionality that we may want to make regarding the association between the two traits. For instance, as we saw in the univariate section, BMI is often taken as an independent variable that influences the rest of the traits and that has to be controlled for, potentially because of the downstream physiological causality that is often assumed. Although it may be difficult to imagine how insulin can directly be affecting BMI (despite the discussion about this in physiology section), it may be easier to see how this arrow may be pointing from insulin to BMI instead of the other way around through this latent variable model.

It only takes for the unobserved variable to be slightly more associated to insulin than BMI for the arrow to be pointing in this direction.

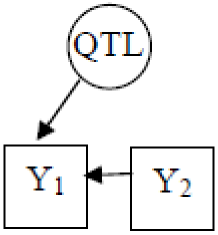


3. Mosaic Pleiotropy: The gene is associated to both traits directly. Joint analysis in this case will always improve power over univariate analyses. This is the only type of pleiotropy that is responsible for genetic correlations between the traits (relational pleiotropy cannot *cause* genetic correlations). If a locus is associated in this way to both BMI and another trait, controlling for BMI will as Panhuysen et al. points out "inappropriately obscure evidence for genes" that influence other obesity related traits (Panhuysen et al. 2003).



4. Exogenous Phenotype: The gene is only associated to one of the traits. The second trait (the exogenous one) also influences the first trait but is independent of the gene. This is not pleiotropy but including the two variables in a joint analysis will improve power. This is probably the model that explains the desire to control for BMI in many univariate studies. The reason taking into account the

exogenous variable improves power is that it controls for heterogeneity in the associated trait. But this model also showcases why much care has to be taken with this approach of controlling for BMI. For genes in which BMI is the associated variable, and the trait of interest is the exogenous one, controlling for BMI will cause a spurious association of the trait to the gene even though they are independent!



2.2 Trait Connections

We finally get to the question that is of interest in this work: what is the genetics underlying the connection between associated phenotypes? The methods that address this question are all by necessity multivariate in nature.

2.2.1 *Connections through mosaic pleiotropy*

One way in which these ties arise is, as mentioned above, through mosaic pleiotropy in which a gene influences two or more of the traits directly. For example, it may be the case that individuals with one genetic variant present all high mean values for the traits while individuals that have a different variant present all low mean values in the traits. If we look at trait values across individuals in a population, this will manifest itself as a correlation of these trait values. In the graphical example presented as figure 2.3 the red and black points represent individuals with alternative genetic variants. The individuals represented in red have mean low values for the two traits plus or minus some error while the individuals represented in black have mean high values also with some residual variance. The result of this grouping of trait values across individuals is a net correlation between the two traits.

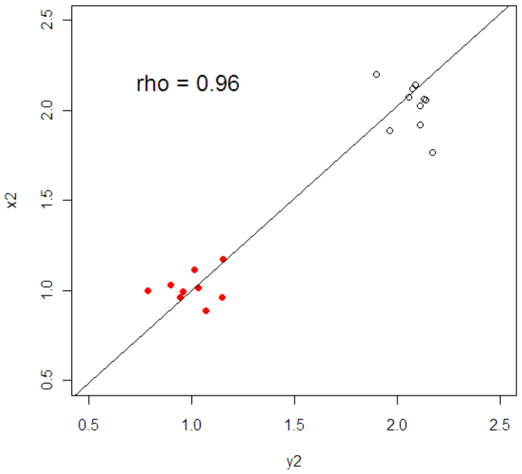


Fig. 2.3 Correlation caused by mosaic pleiotropy

A classic example of mosaic pleiotropy is the PKU (phenylketonuria) mutation, which produces a deficiency of the enzyme needed to convert phenylalanine to tyrosine. This mutation affects the synthesis of melanin, which results in a high percentage of individuals with blue eyes. Independently of this, the accumulation of phenylalanine causes mental retardation. Given a population with a high enough frequency of the mutation, the genetic connection between the two traits would cause them to correlate across individuals, i.e. present a genetic correlation.

2.2.2 *Concept of genetic correlation*

The following model was proposed by Fisher in 1918 in a paper that formed the basis of modern quantitative genetics and it shows how the phenotypic value of a trait (z) can be decomposed into the genetic (G) and environmental (E) value. The genotypic value is the average value for one genotype across the universe of environments (Balding et al. 2007, p534).

$$z = G + E \qquad \text{(eq. 2.1)}$$

The phenotypic variance-covariance likewise can be decomposed into genetic and environmental variance-covariance. For instance, in figure 2.3, if the genetic variants were not for just one gene but for all genes, for example strains of mice, then the genetic variance covariance would be given by the means of the groups while the environmental variance covariance would correspond to the residual deviations from the means within the groups.

For strains of mice where we have genetically homogeneous populations measuring genetic variance and covariance is straightforward. In humans this is a bit trickier. Instead of populations of genetically identical individuals we have groups of individuals that are genetically similar to varying degrees. We can still get an estimate of genetic variance-covariance by using known similarity measures. For instance sibs and parent-offspring share half of their genes and therefore twice their phenotypic variance-covariance provides an estimate of the underlying genetic variance-covariance (confounding factors such as shared environmental and maternal effects have to be taken into account). (Balding et al. 2007, p.535)

2.2.3 *Localized genetic correlation*

Genetic correlations are simply the standardized genetic covariances. Estimating genetic correlations in this way gives an idea of how the traits are genetically linked genomewide, that is, in a non-localized way. There are also ways of localizing pleiotropic genes through full multivariate analysis.

We already cited Arya et al. above as an example of a bivariate analysis that increased the power to map the two traits to a region more precisely than when analyzed individually. This paper also serves to show how particular loci can be tested for pleiotropy (Arya et al. 2003). In variance component linkage analysis the variance components for major gene effect, polygenic effect and environmental variance are all estimated simultaneously via maximum likelihood along with their respective correlations (Almasy et al). The following shows all the components modeled:

$$\Omega = \sum_{i=1}^{n} \Pi_i \sigma^2_{qi} + 2\Phi\sigma^2_g + I\sigma^2_e \qquad \text{(eq. 2.2)}$$

Let's see how it relates to our equation 2.1; phenotypic value = genotypic value + environmental noise. In variance component analysis, as the name implies, we model the variance-covariance matrices instead of the values themselves but the model still reflects the same components: Ω is the variance-covariance matrix of the phenotype, σ^2_q is the variance-covariance due to a major gene, σ^2_g is the variance covariance due to all other genetic effects and σ^2_e is the variance covariance that is left over, or that is due to environmental noise. The terms that precede the genetic variance covariance matrices are just the genetic similarity measures between the individuals that are needed for adjustment, since as we saw above, in humans we have varying degrees of relatedness instead of homogenous genetic groups. Although we present the model here in its univariate form, the extension to multivariate is straightforward. The genetic correlation values (ρ) can then be obtained by standardizing the covariance estimates.

This correlation value can be used to test whether there is a common major gene that affects both traits or whether it is one of two other possibilities: two linked genes may be affecting either trait, a situation called “co-incident linkage”, or the two traits may be tied through another unobserved factor (latent variable model above). As a first step a gene that is associated to both traits must be found. Then, in order to discard the other two possibilities mentioned, the hypotheses that $\rho = 0$ (no pleiotropy) and $\rho = -1$ or 1 (complete pleiotropy) are tested against the model in which this parameter is left free to vary (Almasy, Dyer and Blangero 1997). Arya et al. found that both ln BMI and ln HDL were associated to the same region on chromosome 6 and they rejected the null hypothesis of $\rho = 0$ for no pleiotropy but they also rejected $\rho=-1$. This is indicative of *incomplete* pleiotropy as defined by Almasy. More about what this incomplete pleiotropy may be indicating will be discussed in the first aim.

The two following studies search for common genetic factors underlying all the obesity related traits but using a structural equation modeling (SEM) framework:

1. Hong et al. examined a population of elderly twins from the Swedish Adoption/Twin Study of Aging (Hong et al. 1997). Because they had twin pairs reared together and reared apart they were able to add an additional component to the SEM model called “shared environment” (Es). The residual environmental effect (Ens) consisted of the individual deviations left over after the genetic effect (G) and the shared environmental effect (Es) were taken into account. Again, the common (c) and individual (i) influences on the phenotypes were modeled for each of these components.

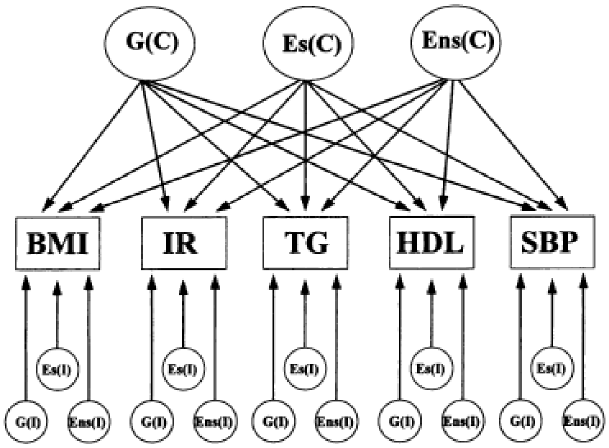
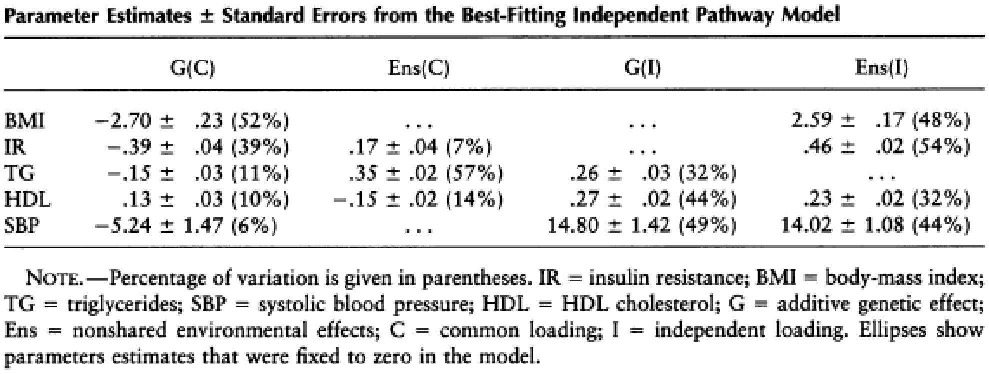


Fig 2.4 Example 1 of structural equation model (SEM). Taken from figure 1 in Hong et al. 1997. Reprinted with permission from Elsevier, copyright 1997.

Table 2.1. Results using SEM model. Taken from table 6 in Hong et al. 1997. Reprinted with permission from Elsevier, copyright 1997.



Parameter Estimates ± Standard Errors from the Best-Fitting Independent Pathway Model

	G(C)	Ens(C)	G(I)	Ens(I)
BMI	−2.70 ± .23 (52%)	. . .	. . .	2.59 ± .17 (48%)
IR	−.39 ± .04 (39%)	.17 ± .04 (7%)	. . .	.46 ± .02 (54%)
TG	−.15 ± .03 (11%)	.35 ± .02 (57%)	.26 ± .03 (32%)	. . .
HDL	.13 ± .03 (10%)	−.15 ± .02 (14%)	.27 ± .02 (44%)	.23 ± .02 (32%)
SBP	−5.24 ± 1.47 (6%)	. . .	14.80 ± 1.42 (49%)	14.02 ± 1.08 (44%)

NOTE.—Percentage of variation is given in parentheses. IR = insulin resistance; BMI = body-mass index; TG = triglycerides; SBP = systolic blood pressure; HDL = HDL cholesterol; G = additive genetic effect; Ens = nonshared environmental effects; C = common loading; I = independent loading. Ellipses show parameters estimates that were fixed to zero in the model.

There is some evidence for genetic correlations amongst obesity and related traits. The Hong et al. study showed 52% of variation in BMI to be due to a genetic influence that also directly accounts for considerable percentages of variance in the other traits (39% IR, 11% TG, 10% HDL, 6% SBP) (Hong et al. 1997). On the other hand, accounts of the importance of these genetic correlations differ across studies. Havill et al. reported an average of 10% of shared genetic effects for trait pairs for the traits of weight, triglycerides and SBP – this is for trait pairs, which means that only equal or less genetic variance could be shared by the three traits (Havill and Mahaney 2002). Benyamin et al. concluded no significant shared genetic or familial environmental effects on obesity and related traits in their twin study (Benyamin et al. 2007).

2. Stein et al. analyzed data from 5th exam of the offspring generation in the Framingham study. Gc stands for common polygenic effects on all the traits, QTL stands for common major gene effects and Ec stands for common environmental variation. Each phenotype's individual genetic and environmental components are also modeled (Stein et al. 2002). This particular SEM model allows for the search of ***direct*** major gene effects on the five phenotypes. Although no statistical significance was reached for any region, the study found tentative evidence for linkage on chromosomes 2, 3, 11, 13, and 15.

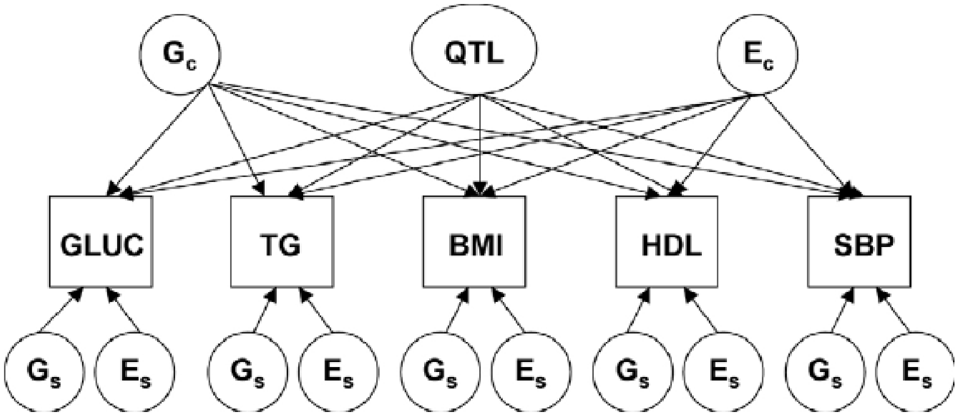


Fig. 2.5 Example 2 of structural equation model (SEM). Taken from figure 1 in Stein et al. 2002.

The difference in design between both models shows an adaptation to the available data: Hong et al.'s unique ability to factor environmental variance into shared and non-shared was afforded by his twin/adoption design and is not possible with most existent human data sets. On the other hand, their lack of genetic data precluded them from incorporating a common major gene effect as observed in Stein et al.'s study and as a consequence even if their results show an important genetic component underlying all the phenotypes, it would say little about the possibility of actually being able to map a common QTL.

2.2.4 *Confounding with relational pleiotropy*

Mosaic pleiotropy is therefore one way in which genetics can connect traits across individuals and, as we have seen, it can be detected and localized through its production of genetic correlations. When the traits are independent, such as in the PKU example provided above, this is a very useful tool. For instance, continuing this example, by localizing the gene responsible for the observed correlation between blue eyes and mental retardation, and following up with a functional study, the PKU mutation would in effect answer the question of “what is connecting these two traits across individuals?” But what happens when the traits are physiologically connected? Here enters the concept of relational pleiotropy. Conceivably, if a gene affects one of the traits, for instance BMI, this trait will be affecting the other traits it is physiologically tied to, creating an indirect link between the gene and the other traits. When this happens, the connection between the traits does not necessarily have anything to do with the associated gene, and yet this gene would also be picked up as causing a genetic correlation. Because the studies described above do not take trait relationships into account, they have no way of differentiating between relational pleiotropy, where the gene only affects one of the traits directly, and mosaic pleiotropy in which the gene is directly affecting both traits and thereby connecting them across individuals.

Structural equation modeling (SEM) is an extremely flexible modeling *framework,* rather than an actual model or a method. It allows for taking into account and modeling direct relationships between traits, such as those present in relational pleiotropy, and indirect relationships between the traits including but not exclusive to those that are due to a common gene, such as in mosaic pleiotropy. One challenge that comes with this flexibility is that “The selection of the actual model cannot always be entirely guided by the data…”(Todorov et al. 1998). To address this problem, what is generally attempted is that very general models are constructed. These models contain a wide range of more specific submodels that can then be compared against each other. Sometimes this means that certain complexities have to be overlooked. For instance, SEM studies described above partition phenotypic variance into all of its components in a way similar to the variance component linkage approach and they do not incorporate direct trait relationships.

Todorov et al. introduced an SEM framework in which to model both these trait to trait relationships and their genetic linkage. In his words: “One advantage of the present approach is that it enables us to test whether a certain gene influences a given phenotype directly, or only indirectly through other intervening phenotypes.” (Todorov et al. 1998) We can see how this is accomplished by comparing Todorov’s model with our general model, equation 2.1:

$$x_i = Bx_i + \Lambda g_i + e_i \qquad \text{(eq. 2.3)}$$

$$z = \quad ? \quad G + E$$

Here again x represents the vector of phenotypes, g the genetic effects and e the environmental residuals. What is different in this model is the term Bx_i. It represents the linear relationships between the phenotypes being modeled. In Todorov’s model they are assumed to be causal relationships and as such, unidirectional, making the matrix B a lower diagonal. This just means that if trait A influences trait B, it is assumed that trait B does not influence trait A. If for instance, we were looking at the phenotypes of insulin and BMI, this model represented as a system of equations (instead of representing them in matrix form) would look like:

$$\text{Insulin} = b{*}\text{BMI} + \lambda g + e \qquad \text{(eq. 2.4.1)}$$

$$\text{BMI} = \lambda g + e \qquad \text{(eq. 2.4.2)}$$

In this way, a gene will be considered to have an effect on insulin, only if this effect is present when controlling for BMI. Genes that truly influence both BMI and insulin independently can then be separated from those that only influence BMI directly.

Section 3. DESCRIPTION OF SPECIFIC AIMS

3.1 **AIM 1: The context**

We have reviewed what has been generally done in the study of the genetics underlying physiologically connected traits in general and we have also seen methods specifically designed to study the genetics behind the *connections* between these traits. But do the latter really capture the biological variation in the connections that we have described to be of interest, for instance, the variation in the connections between obesity and related traits that we have seen from the literature to be relevant to disease prognosis?

This is the subject of Aim 1 in this dissertation. Within this objective the study of the biological variation that has been described to be of interest is put into context by contrasting it with what is generally pursued in the field of genetic epidemiology. The paper corresponding to this objective is entitled "The within and across individual distinction in the genetics of correlated traits" and we will be referring to it in this section as "the context paper".

3.1.1 ***Uncaptured biological variation***

Let's begin by revisiting the mechanistic example proposed in the physiology section that is representative of the biological variation that we wish to study to then examine if we can indeed capture this variation with the methods that we have outlined.

Le Lay et al's study showed how normal sized fat cells of a transgenic obese mouse present normal expression levels of a transcription factor important in regulating cholesterol uptake (Le Lay et al. 2003). These normal sized fat cells may likewise present normal adipocyte function while their enlarged counterparts present adipocyte dysfunction along with all the co-morbidities that this entails. A gene that somehow predisposes to accumulation of normal sized fat cells, rather than an increase in size of existing fat cells, with increasing BMI, would *affect the relationship* between BMI and the other traits *within* the individual. Let's look at what this can imply a little closer.

An individual with fat tissue that is predisposed to increasing in total size through number of fat cells rather than through size of individual fat cells may go through life gaining and losing weight without this having a direct impact on the individual's obesity related traits. This individual will not develop high blood pressure, blood sugar or dyslipidemia *when* they gain weight. *Notice how this has no bearing on whether or not the individual is predisposed to obesity or to high blood sugar or any of the other morbidities*. This type of biological variation only has to do with whether or not BMI and the other traits are *connected*, and thereby associated throughout the individual's life irrespective of whether or not the *values* of these traits for the individual are high or not.

The methods that we have encountered focus on studying the genetics underlying the individual's *values* for the traits, without regard to the *relationship* that these traits present throughout the individual's life, i.e. the relationship between the traits within individuals. These physiological relationships or ties will manifest themselves as correlated measures for the traits throughout the individual's life. Even those methods that look at trait connections or correlations, do so by looking at the correlations between the *values* for the traits for each individual. As will be explained below, this results in a characterization of these relationships *across* individuals.

3.1.2 ***The within and across individual distinction***

In statistics this within and across distinction is made in order to take into account non-independence between observations. If there is some level of redundancy between observations, it is important to take into account when evaluating degree of evidence, i.e. significance, since redundancy means there is less independent evidence. Biologically speaking it is also important to make this distinction because across processes may be very different from within processes. The physiological connections between BMI and associated traits are *within* level processes. High BMI in one individual will only cause an effect on other traits within that same individual. Genotypic effects on the other hand are across level processes. Genotypes do not vary within the individual and they can therefore not explain variation within the individual. So although genotypic variation can cause associations between traits across individuals this says nothing about how the traits are relating within, i.e. physiologically.

Physiological causation is inferred in some instances in studies that only look at across individual patterns because these across relationships between traits are assumed to be a reflection of the within individual relationships. Rarely is the distinction between across and within relationships made, a lot of the times because having only one measurement of the trait per individual restricts us from doing so. Within individual variation can be captured only through repeated measurements on the same individual.

3.1.3 ***Methods that have been described***

All of the methods that we have described, whether univariate or multivariate, those that look for mosaic or relational pleiotropic genes, strictly look at genotypic variation that has an effect on the mean trait values for each individual. The type of variation described by the mechanistic example above would thereby go undetected by all of these methods. Let's see what additional limitations the methods involve more specifically.

The univariate methods to start with simplify away the multivariate relationships that may be present between traits altogether and in so doing are most proper for traits that are for the most part independent both across and within individuals (or those for which dependencies are unimportant or unobserved). When used to study physiologically connected traits univariate analyses do not capitalize on the statistical power that can be afforded by studying the traits together, and, as we saw in the methodology section, they run into the issue of uncertainty regarding what exactly is being mapped.

The multivariate methods that look for common genetic effects without taking into consideration direct relationships between traits (their physiological connections), can be thought of as assuming only within individual independence. The PKU mutation is a good example of a case for which these methods are most appropriate since traits like eye color and IQ do not covary within the individual, only across. Mosaic pleiotropic genes would capture most variation of interest here when looking for factors that connect the traits and that is what these methods find given that there aren't any within associations confounding them. When used to study physiologically connected traits multivariate methods that look for common genetic effects may end up finding genes that only really have an effect on one of the traits and then affect the other traits through their physiological connections with the first. In this case although it could be considered a common gene for both traits, it could not be considered a gene responsible for connecting the traits.

Todorov's approach, and in general multivariate methods that include direct trait relationships in what is being modeled, do not assume across or within individual independence of traits. What they do assume is that the within relationships are the same across individuals and that they are unidirectional. This means that there is no need to differentiate across from within associations. When this assumption is not met and there is variation in the within relationships between traits from individual to individual this approach can run into problems. By ubiquitously controlling for one trait, for example BMI, while looking for associations to another trait, for example insulin, there is the risk of picking up spurious associations to loci that are only associatied to BMI, on the one hand, and of interfering with the signal for loci independently associated to both BMI and INS on the other.

This assumption of non-varying within relationships across individuals is also required for studies that use the mendelian randomization paradigm. In this paradigm the causal effect of a modifiable exposure on disease is studied non-experimentally by making use of their known association to a gene and using mendelian segregation as a randomization process (Smith and Ebrahim 2003). Variation in the relationship between the exposure and the disease from individual to individual would compromise power in these studies.

The context paper discusses these issues with methodologies that assume non-varying within relationships, and that do not make the within and across distinction when studying physiologically connected traits, in more detail, in addition to addressing the benefit of capitalizing on the variation of the within relationships as novel biological variation with relevance to disease.

3.2 **AIM 2: The methods**

How exactly can we capitalize on this biological variation? How can we pursue the genetics underlying it and through the genes discovered potentially arrive at new clues into disease mechanism?

This is the subject of Aim 2 in this dissertation. Within this objective a new approach designed to study the biological variation we are interested in targeting is proposed. The paper corresponding to this objective is entitled “Methods for testing genetic effects on within individual correlations” and we will be referring to it in this section as “the methods paper”.

As we have mentioned before, within individual relationships can only be characterized through repeated measurements data. The methods paper therefore jumps off a random effects model which is a model designed to deal with repeated measurements data and which can be used to make the within from across individual distinction. The way the random effects model is usually used for repeated measurement data differs a bit from how we are interested in using it for our particular application, mainly because of differing assumptions designed for asking different biological questions. Let's look closer at these underlying assumptions.

3.2.1 ***Assumptions of longitudinal study***

It is useful to contrast the approach proposed in the methods paper as a way of capturing variation in within individual trait relationships to *longitudinal* type of approaches since longitudinal studies also partition within from across variation and require estimates of within correlations. In longitudinal studies homogeneity of trait correlations at one time-point across individuals is assumed. They are therefore estimated by calculating trait correlations across individuals at each time-point and this allows us to see how these trait correlations change through time. The variation that is observed across individuals is that of time-point correlations, i.e. patterns through time, not that of trait correlations. In our approach on the other hand, since the focus is on trait relationships independent of what these may be doing through time, homogeneity of correlations across time-points for each individual is assumed instead. In this way the trait correlations for each individual can be estimated by using measurements across time-points and the variation observed across individuals is in the correlations between traits.

Suppose that we have two individuals that are predisposed to keeping normal levels of insulin despite any gain in weight, and a third individual that, as is more common, develops insulin resistance with gain in weight. For the biological variation that we want to capture, we would hope that the methodology used would group the first two individuals together and separate them from the third. In longitudinal analyses this would entirely depend on the pattern of weight gain and loss throughout these individual's lifetimes. If the first individual gains weight in his teenage years and then loses the weight, while the second individual gains weight as an adult, this difference is going to be the focus of a longitudinal approach. The fact that they were both healthy obese would be missed unless specifically inquired. What's more, if the third individual also gains weight during the teenage years and then loses it, he would be looked at as being more similar to the first individual. Even though their insulin profiles would differ (the third individual develops insulin resistance during the weight gain while the first one doesn't), they would still be considered more similar because of their shared weight profile through time when compared to the second individual.

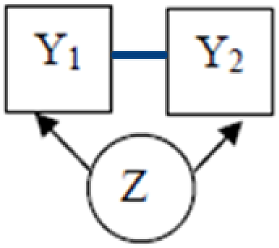
This focus on patterns of change through time is therefore not something we desire. Instead, we want the focus of our analysis to be the relationships between our traits *independent* of what they happen to do through time for any particular individual.

3.2.2 *The perturbation framework*

This does not necessarily mean that we want to factor out what the traits are doing through time because this could mean controlling away part of the information we are after. If the trait relationships also happen to be tied to time, i.e. they happen to all increase through time, controlling for time will also do away with the physiological tie between the traits. There is no way to separate the two and there is no need to because time can be looked at as a perturbing factor within the individual, a perturbation that allows us to see how the traits relate within the individual. Whether time is causing an effect on one of the traits or all of them, these effects will still be reverberating through the system of traits as a whole. So although in longitudinal studies each time-point for each trait is considered a distinct estimable quantity, for our application we can conceptualize each time-point as perturbations to the same quantity, as perturbations to the wiring, the physiological connections within the individual system.

Every individual’s system comprises a set of physiologically interconnected traits that respond in a coordinated manner to perturbations, whether these perturbations are contingent on time and development or if they are simply environmental perturbations. These coordinated responses will manifest themselves as correlations between the traits and these are what we can call *physiological* correlations.

With this new assumption we can conceive of the following model for within relationships. The blue line represents a non-directional tie between the traits within the individual and z represents all of the environmental and developmental/time factors that perturb the traits for this one individual. These perturbations only serve to show how it is that the traits are associated within the individual.



For instance, if an individual has periods in his or her life where they are more prone to overeating, this will cause a gain in weight during those times. We are not concerned with this change in trait value, but rather what happens with the second trait, for instance insulin, when this change occurs, that is, we focus only on the blue line. It is a non-directional association. So for example, if insulin resistance causes a carbohydrate craving that leads to a gain in BMI this phenomenon too is being modeled. The idea is that regardless of the factors that the individual encounters throughout life, whatever z may be, and however these factors may affect either of trait values, the traits are wired to respond to each other in a certain way within this individual's organism throughout his life.

It is true that these relationships may change through life, or that they may be dependent on the perturbing factors. This assumption of no change is what allows us to focus on the part of these relationships that does not change, the wiring so to speak that is present despite fluctuations. We make the same assumption when mapping a single trait like BMI which is a quantity that fluctuates all throughout life. With BMI we often take one measurement per individual (one random point in time) and proceed to make inferences on the genetics underlying it.

Some individuals will be more capable of a buffered response to perturbations, keeping their system under a homeostatic state more effectively. Other individuals may be more susceptible to positive feedback loops between the traits. The physiological correlations that the individual exhibits throughout his life should capture this system characteristic.

A very similar concept was used in Nadeau et al. to map the physiological correlations underlying a set of cardiovascular (CV) traits in mice (Nadeau et al. 2003). In place of environmental perturbations, they used the genetic perturbations present in a genetically randomized population of mice. In effect, what they did was use relational pleiotropy to get at the physiological relationships. The genetic perturbation on each individual trait, causes, through relational pleiotropy, an effect on a second, physiologically related trait. This in turn causes the two traits to cosegregate. Nadeau et al. were then able to map the system of physiological relationships on the basis of the observed patterns of co-segregation.

This study then compared the physiological relationships observed in normal mice to those present in mutant and pharmacologically treated (anesthesized) mice with compromised CV function. In Nadeau et al.'s words: “Trait relationships (correlations) may be maintained or lost depending on the way in which each component trait responds to the perturbation. With homeostatic responses, combinations of functionally related traits respond in correlated manners in an attempt to compensate for the effects of the perturbation. Alternatively, traits in mutant or treated individuals may change in manners opposite to those in the reference network because the nature of the perturbation compromises homeostasis.” (p. 2086)

Similar to Nadeau's approach, the approaches described in the methods paper allow us to compare observed trait relationships, or physiological correlations, across different groups of individuals, whether they be different genetic groups, which would allow us to find associations to genome-wide SNPs for instance or differentially affected groups such as those with and without diabetes.

The perturbation frame-work just described explains why for our application there is no interest in controlling for covariates at the within individual level. Instead the approaches described in the methods paper provide a way for controlling at the across individual level which is where confounders for tests of association to differences in physiological correlations from individual to individual would lie.

3.2.3 ***Correlations vs. slopes vs. covariances***

The following equation shows the relationship between correlations (ρ) and slopes (b):

$$\rho^2=b^2*\sigma^2(x)/\sigma^2(y) \qquad \text{(eq. 3.1)}$$

Slopes and correlations will differ between individuals when the ratio between the variances (σ^2) of the two traits change from individual to individual. Individual B can have the same correlation value as individual A, but if one of the traits varies more in individual B while the other trait's variance stays the same then the slopes between the two individuals will differ. Insulin for instance may be tracking BMI values just as closely in both individuals, but if individual B's insulin increases more per unit increase in BMI, then his slope will differ. Going back to our hypothetical biological model, the individual that increases BMI by increasing number of fat cells rather than their size will likely cause a disassociation between the two traits (a change in correlation) rather than a lower increase in insulin level per unit increase in BMI. This is why we focus on correlations and the methods paper describes a way of capturing variation in correlations. The methods described could be easily modified for studying variation in slopes as well if this were decided to be a better fit to the biology of interest.

Similarly, looking for associations to both correlations and covariances may be desired, depending on the hypothesis. If it is hypothesized that lower correlations are linked to lower variance because it is believed that individuals that have more stable traits (low correlations) may also have less variance in the traits, then weighting the correlations by variance would be detrimental since correlations from measurements with low variation would be deemed just as informative as correlations from highly variable measurements. If on the other hand it is hypothesized that variance is independent of correlation, for instance, if it is believed instead that those that have higher variance are more perturbed and therefore more informative in terms of their correlations, then weighting the correlations by their variance, i.e. using covariance instead, might be desirable.

3.3 **AIM 3: The data**

We have seen results from previous studies that provide indirect evidence for the existence of the biological variation that we are targeting in this dissertation and for its relevance to disease. The literature also provides some indirect evidence that there are genetic underpinnings to this biological variation (see context paper). In order to determine if there is more direct evidence for both we need to inquire real data.

This is the subject of Aim 3 in this dissertation. Although some data analysis was also conducted within the context and methods papers, it was not their main focus and it was therefore not as comprehensive as what is attempted in the third paper entitled "Association to disease and genetic architecture of metabolic trait correlations". We will refer to it in this section as "the data paper".

In addition to trying to find direct evidence of the existence and relevance to disease of the within individual correlations of obesity and its associated traits, and to exploring the genetic architecture of the system of correlations in a comprehensive manner the data paper allows us to ask a third and equally important question about the biological variation in these correlations. *How redundant is this variation with the variation in the trait values themselves?* As we see in the methods paper, correlations can be thought of as latent variables – variables that are not directly measurable the way more standard traits like BMI and insulin or blood glucose are. So although, as we see in the context paper, the variation in trait correlations is theoretically independent from variation in the trait values, we still need to establish empirically if this is the case in the existing variation. If in reality the trait values prove to predict disease in the same way that their correlations do and more effectively so, and if the trait correlations prove to only be associated to genes to which the trait values themselves are associated, the existence of genetic and disease relevant variation in trait correlations will be of little value. In the data paper we tackle this redundancy question by comparing our correlation results with that of the trait values.

The data paper also addresses the multivariate analysis of trait correlations, subject that this dissertation does not touch upon up to this point. One benefit that mapping the connection between traits brings with it is the potential of getting at the hubs that modulate the system of traits. This is a distinct advantage over mapping the trait values themselves. For instance, although mapping BMI brings with it a convenience of its own, which is its being considered a more easily measurable and reliable quantity, once loci underlying BMI are found, they can be anything ranging from a molecular phenotype to a behavioral trait. BMI is influenced by so many factors within the individual that the real question may be what *doesn't* influence BMI. Mapping BMI may therefore not bring us any closer to understanding the system of phenotypes associated to it. On the other hand, the number of factors that connect BMI and insulin can only consist of a much smaller, more system relevant set of factors, than the list of those that influence BMI and insulin independently. We would likewise expect the list of factors that influence the connection between BMI, insulin and dyslipidemia to be a subset of those that underlie the connection between BMI and insulin alone, and for the factors that connect the four morbidities to be a subset of this subset, each step further narrowing down the list to factors that have more of a system-wide relevance. A multivariate approach to the study of trait correlations would capitalize on this idea and possibly better target the hubs that modulate a system of physiologically connected traits or morbidities as a whole.

The data paper also describes additional benefits to a multivariate analysis of correlations. Biological insight can be garnered from the different multivariate phenotypes proposed. Which multivariate phenotype the disease or the genetics is associated to can give an idea as to *the type* of system-wide variation that is important in the prediction of the disease and that is influenced by the gene. There is also a strictly statistical motivation to carrying out multivariate analyses of correlations. Since the set of all pairwise correlations are necessarily not independent, this means that the multivariate analysis can capitalize on the additional power that taking into account their multivariate structure affords. We see the results of this increased power in the comparison of univariate and multivariate correlation analysis in the data paper.

Section 4. THE CONTEXT

The within and across individual distinction

in the genetics of correlated traits

Abstract

Objective: Making a distinction between within individual and across individual trait correlations may be useful on two counts. First, whether the two correlations differ may be informative regarding how much the across pattern of variation is a reflection of within individual processes. Only within individual processes can account for a direct and causal, physiological connection between two traits. This has implications for methods that rely on the relational pleiotropy model which requires that the across correlation reflect the within individual causal connections. It is also useful because it allows for the treatment of the within correlations as a quantitative trait. This opens up the possibility of targeting genes underlying variation that may be biologically relevant and that are generally not captured. This study examines both potential uses of making the within and across distinction using real data.

Methods: This within and across distinction is made for correlations between BMI and its associated traits, cholesterol, triglycerides, blood glucose, SBP, DBP and HDL using the Framingham heart study data. The implications of each of the trait pairs' within-across correlation profile on mendelian randomization tests and genetic association tests that systematically control for BMI, is examined through simulation. The association of the BMI – blood glucose *within* correlations ($\rho_{\text{bmi-gluc}}$) to an interleukin-6 polymorphism (rs1800795) and to diabetes status is tested using this data via permutation testing and logistic regression.

Results: There was variation of within-across correlation profiles for BMI and associated traits. Profiles in which the across correlation is lower than the within correlation were found to decrease power in mendelian randomization studies while those in which the across correlation is higher than the within correlation were found to increase the false positive rate in genetic association tests that control for BMI. The correlations of BMI with triglycerides and cholesterol fit the former profile while its correlations with blood glucose, SBP, DBP and HDL fit the latter. A suggestive p-value ($p<0.1$) was found for the effect of the IL-6 polymorphism on $\rho_{\text{bmi-gluc}}$ in women but not in men. $\rho_{\text{bmi-gluc}}$ was significantly associated to diabetes status even after controlling for BMI and blood glucose levels ($p<0.05$).

Conclusion: The within and across individual trait correlation profiles may inform on the risk of increased false positive and negative rates when conducting genetic studies that rely on the relational pleiotropy model and should be taken into account upon the availability of repeated measurement data. Variation in within individual trait correlations has the potential of being biologically relevant (predictive of disease) and non-redundant with the variation present in the trait values. Because of this, this variation should likewise be explored in genetic association studies.

4.1 **Introduction**

There are complex diseases with system-wide impact that affect a constellation of physiologically interrelated phenotypes rather than a single phenotype. Some examples of this are mental illnesses such as schizophrenia, the constellation of autoimmune disorders related to arthritis, and the group of metabolic related morbidities that are associated to obesity and which are sometimes called metabolic syndrome. The genetics underlying this type of complex disease can be uncovered by mapping biological markers underlying each of the phenotypes.

The correlations present between these biological markers, or traits, merit special consideration when searching for genes associated to the disease as a whole or to individual phenotypes. In particular, it is useful to distinguish between the patterns of these correlations within and across individuals. We will see two reasons for this:

1. Unraveling causality: where the correlation pattern lies may be indicative of the type of biological process that is responsible for it and whether or not it involves a causal relationship between the traits.
2. Unexploited biological variation: within correlations vary from individual to individual and this variation may be informative regarding disease.

4.1.1 *Unraveling causality*

When studying the genetics of a set of physiologically interconnected traits we often want to unravel the causal relationships underlying the traits and their correlations, and their connections to the genes.

For instance, we may only want to pursue through functional studies, those genes that have a *direct* influence on the trait of interest. In this case it is necessary to filter out genes that only affect the trait of interest via their effect on a second trait and its physiological ties to the first. An example of this indirect influence on a trait is the FTO gene and its effect on diabetes incidence (Xi and Mi 2009). This effect is entirely mediated through FTO's influence on BMI. Once BMI is controlled for, the association between FTO and diabetes incidence disappears. This process of filtering out indirect genetic relationships by controlling for one of the traits is a common practice in genetic association studies (Levy et al. 2009, Kathiresan et al. 2007, Duggirala et al. 2001).

Another example of when we may wish to understand causality is when the goal is to design an intervention that targets one of the traits. It may be that a second, correlated, trait is more easily modified. If this is the case, it would be of interest to know if the correlation between the traits represents a causal connection between the two that can be exploited in the intervention. Using BMI as an example again, if it were determined that the correlation between BMI and insulin resistance is causal in nature, then this would open up the non-pharmacological possibilities for regulating BMI, such as diet and exercise, as potential preventative and treatment measures for insulin resistance. Genetic associations to the traits provide a tool with which to gauge this causality through the process termed mendelian randomization.

4.1.1.1 Causal relationships, a within individual process

When two physiologically interconnected traits present a causal relationship between them, a change in the value of one of the traits causes a change in the value of the second within an individual organism. If an increase in BMI has the physiological consequence of an increase in blood glucose, and a decrease in BMI likewise produces a drop in blood glucose, then one expects these two trait values to track each other throughout an individual's life. This within pattern of correlation between the traits is therefore a necessary condition for inferring a causal relationship between them. Note though, that although this relationship is a *necessary* condition for causality, it is not sufficient. For instance, if a third trait is causal to both x and y, a relationship will be present between the two even though there is no causality between them.

The following equation shows how the within individual relationship between two traits may be depicted:

$$Y1 = \beta_1 + \beta_2 * Y2 + \varepsilon \qquad \text{(eq. 4.1)}$$

Although the real relationship may be non-linear, this simple linear equation can be applied by finding a proper transformation for the traits. Repeated measurements of Y1 and Y2 throughout the individual's life should present a "β_2" parameter that is different from zero for potential causality to be inferred. If the Y1 and Y2 measurements are standardized, then this β_2 parameter represents the within individual correlation between the two traits. Keep in mind that this relationship can hold irrespective of whether or not the traits present a pattern through time, i.e. whether or not they are associated to age.

4.1.1.2 Across individual correlations can reflect within processes

We seldom get to observe and measure these within individual relationships between traits because of the cohort study design it requires. Taking repeated measurements on the same individual for many individuals requires years of follow-up study that demand resources that are often not available or that simply do not outweigh the benefits in precision afforded by a cohort design (Feldman and McKinlay 1994). Because of this, many of the available data sets provide only one measurement per individual.

Despite not being able to directly observe the within individual correlations with only one measurement per individual, the observed correlation pattern can still reflect the within, potentially causal, relationship between traits. If the correlation across individuals, ρ(real across), which corresponds to the correlation of the individuals' trait expectations, is equivalent to the correlation within individuals, ρ (real within), which corresponds to the correlation of the repeated measurements of the traits within each individual, then the correlation observed using a single set of measurements per individual would likewise be the same (see equation 4.2). If on the other hand the across and within correlations differ then the single observed correlation is a weighted average of the across and within correlations (Snijders and Bosker 1999):

$$\rho \text{ (observed)} = \lambda * \rho \text{ (real across)} + (1 - \lambda) * \rho \text{ (real within)} \qquad \text{(eq. 4.2)}$$

λ is the intraclass correlation coefficient:

$$\lambda = \tau^2 / (\tau^2 + \sigma^2) \qquad \text{(eq. 4.3)}$$

where (τ^2) refers to the across individual variance for the trait and σ^2 represents the within individual variance (we assume the same τ^2 for both traits and the same σ^2 for both traits in order to simplify equation 4.2).

Within individual and across individual correlations between traits can differ because they can be product of entirely different processes (Snijders and Bosker 1999). Figure 4.1 presents simulated data used to show this independence between within and across correlations. A single individual (top left panel) consists of a set of repeated measurements (in black) and a mean for these measurements (in red). In this way the within correlations are depicted in black as the patterns created by the repeated measurements within individuals, and the across correlations as the patterns in red created by the means of these repeated measurements across individuals. Although the across pattern can entirely be a reflection of the within individual processes (bottom left panel), this is not necessarily the case. Across processes can disassociate two traits that have a within association or even associate the two traits in a manner opposite to that of how they are associated by within processes (top right panel). Likewise, across processes can create a correlation between traits even in the absence of any within and potentially causal relationship between them (bottom right panel).

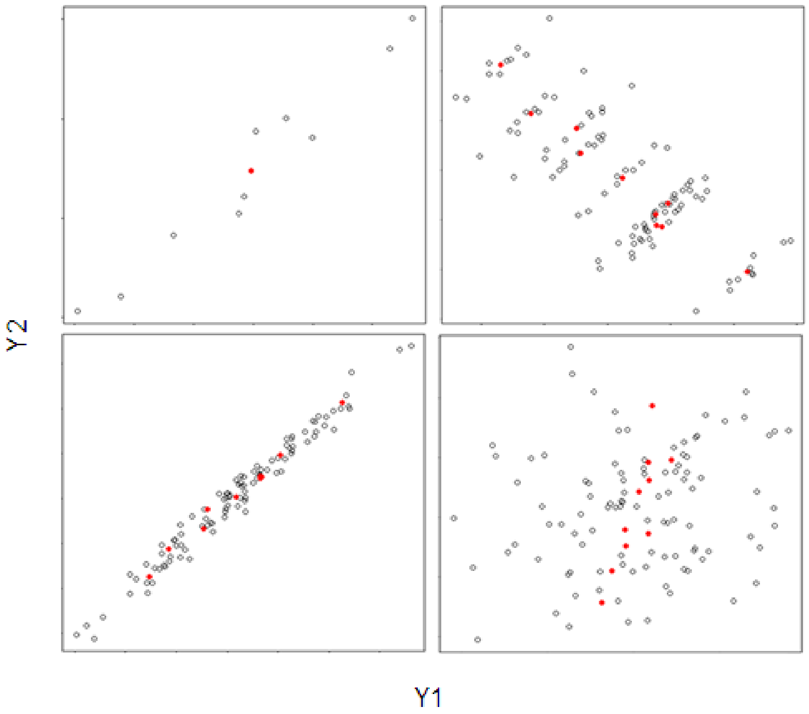


Fig. 4.1 Within and across correlations are free to vary independently. Red represents trait means for the individual while black represents repeated measurements for each individual. The top left panel shows a single individual and their "within" correlation between Y1 and Y2. The bottom left panels shows multiple individuals and how the correlation across, shown by red, can be strictly a reflection of within processes. The right panels show when across processes keep the across correlation from reflecting the within processes. In the top right the correlation across is negative even when all the within correlations are positive. The bottom right shows when across processes create a correlation despite the absence of any within relationship between the traits.

When we only have access to one measurement per individual we cannot tell whether the across correlation is reflecting the within potentially causal processes or not. But upon the availability of repeated measurements data, the across – within correlation profile can be obtained and used to infer this. The across correlation pattern will reflect the within processes when the across correlation and the within correlation are equivalent and the following two assumptions hold:

1. There is no variation in the within relationships from individual to individual, that is, that the β_1 and β_2 parameters of equation 4.1 remain the same from individual to individual.
2. There is no factor creating an association between the traits across individuals while not creating it within (for example a gene with a direct influence on both traits).

If only assumption 1 does not hold, then the across correlation is less than the within correlation and if only assumption 2 does not hold, then the across correlation is greater than the within correlation. If both assumptions do not hold, then the across correlation can be greater than, less than or coincidentally the same as the within correlation without being a reflection of the within processes. In the latter case the within-across correlation profile will be less informative.

4.1.1.3 Implications for methods based on causal model

This has implications for methods for which inference is based on the existence of a causal relationship between the traits and yet observations of the traits are necessarily only made across individuals. It becomes of interest to know how much of the correlation between traits across individuals is product of a within, and potentially causal, process.

Such is the case for methods that rely on the relational pleiotropy model of figure 4.2A. What is particular about this model is that it combines a within individual process with an across individual process. Here, the gene or SNP has an effect on the second trait only through its influence on the first, meaning that a causal relationship, which is a within level biological process, must exist between the two traits. Because genotypes do not vary within individuals, a gene can only have an effect on a trait value across individuals and this effect can only be gauged through its measurement across individuals. Genotypic effects are therefore strictly due to across level biological processes. Methods that are based on this model will therefore be prime examples of where the degree to which the across pattern of correlation between traits reflects the within pattern becomes important. It is only when this reflection exists that both the within and the across processes can be studied at the same across level and that the relational pleiotropy model can make sense. In particular, the two approaches mentioned above, the mendelian randomization method and the common practice of systematically controlling for one trait while conducting a genetic association analysis on a primary trait, rely on this model.

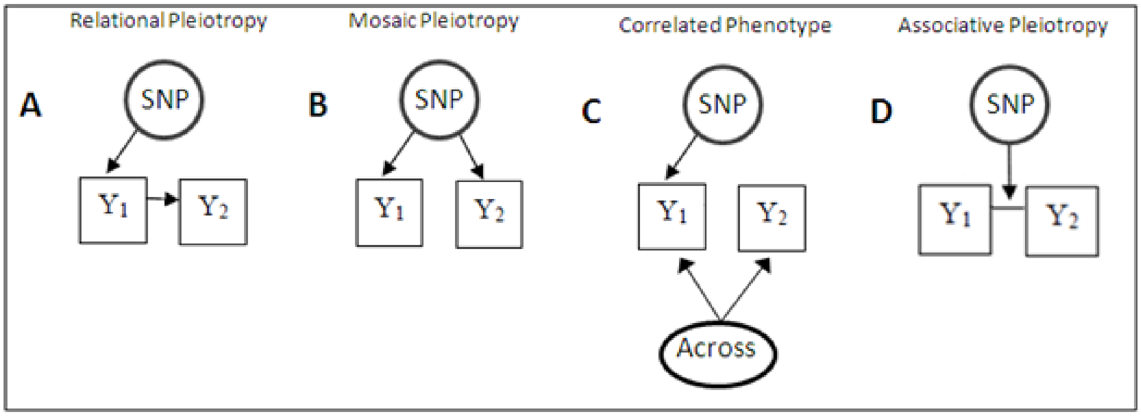


Fig. 4.2 Different models relating a SNP to two traits. Arrows represent a causal relationship.

In mendelian randomization, genes that are known to be associated to the trait that is amenable to intervention (Y1 in figure 4.2A) are used to test causality between this trait and the trait for which a treatment is being sought (Y2 in figure 4.2A) by using mendelian segregation of the gene as an in situ process of randomization (Lawlor et al. 2008). For example, in the case of BMI and insulin resistance, if the genes that are known to be associated to BMI also show an association to insulin resistance, a causal relationship from BMI to insulin resistance is inferred. The mendelian segregation method would therefore consist of testing for an association between the gene and insulin resistance.

Mendelian randomization studies would suffer from a loss of power in the case in which the β_1 parameter, the intercept, varies from individual to individual and the correlation across is less than the correlation within. The reduced across correlation would mask the real within individual correlations produced by the causal relationship between the traits. The following equations show why this reduction in power takes place. Y1 represents the trait that is known to be associated to the gene and the association between Y2 and the gene is the mendelian randomization test of causal relationship between Y1 and Y2.

$$Y\ 1(\text{means}) = \beta_0 + \beta_1 {*}\ \text{SNP} + \varepsilon_1 \qquad \text{(eq. 4.4.1)}$$

$$Y2 = \beta_{00} + \beta_{11} {*}\ Y\ 1 + \varepsilon_2 \qquad \text{(eq. 4.4.2)}$$

$$E[Y\ 2\ ] = \beta_{00} + \beta_{11} {*}\ E[Y\ 1] \qquad \text{(eq. 4.4.3)}$$

$$Y2(\text{means}) = \beta_{00} + \varepsilon_{\text{intercepts}} + \beta_{11} {*}\ (\beta_0 + \beta_1 {*}\ \text{SNP} + \varepsilon_1) \qquad \text{(eq. 4.4.4)}$$

In the above equations “means” refers to the expected values for the trait for each individual throughout their lifetime, for a group of individuals. Y1(means) are therefore a collection of E[Y1]’s and Y2(means) are a collection of E[Y2]’s. These expected values are those upon which the SNP may have an influence on as an across individual process as shown in equation 4.4.1. The β's with a single subscript describe relationships across and the β's with double subscripts describe relationships within individuals. Equation 4.4.2 then shows the relationship between repeated measurements of the traits for a single individual. In equation 4.4.3 we take the expectation at either side of equation 4.4.2 to show the relationship between the trait expectations for this one individual (the relationship between two single values) while equation 4.4.4 shows the relationship between these expected values across individuals. If there is variation in the intercepts β_{00} from individual to individual, this will come up as an additional error term that can swamp the signal of association between the SNP and Y2.

In genome-wide studies that systematically control for one of the traits while testing association to the other trait, an across correlation that is greater than the within correlation can increase the false positive rate. A higher across correlation occurs when a factor exists that associates the two traits across individuals, although it does not do so within the individual, thereby creating a non-causal tie between the traits. An example of such an "across" factor could be a mosaic pleiotropy gene (see figure 4.2B), that is, a gene which has a *direct* influence on both traits. Figure 4.2C also shows a non-causal relationship between the two traits: Y1 is associated to both the SNP and Y2 while Y2 is not associated to the SNP. This type of relationship has been previously termed the correlated phenotype model (Allison et al. 1998). In the correlated phenotype model, an association test between Y2 and the SNP that is conducted controlling for Y1 can result in a spurious significant association. The graphical example depicted in figure 4.3 serves to show how this occurs. Although Y2 is clearly not associated to the SNP in this figure, because Y1 is, the residuals of Y2 after regressing it on Y1 *are* associated to the SNP.

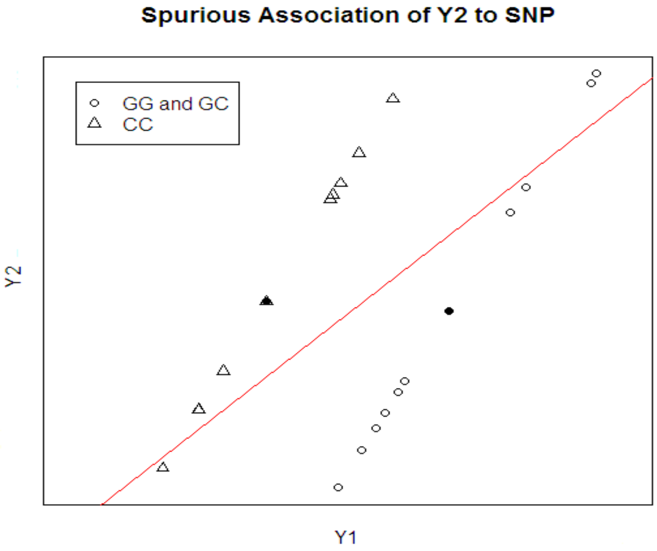


Fig. 4.3 A spurious association. The residuals of Y2 after regressing on Y1 are associated to the SNP even though Y2 is not associated to the SNP. The filled circle and triangle represent the means for the "GG and GC" and the "CC" genotypic groups respectively. They show an association of the SNP to Y1 and no association to Y2. The red line represents the regression line of Y2 on Y1 and the vertical distances from the points to the red line represent the residuals of Y2. Most distances from the triangles to the red line are positive while the distances from the circles to the red line a negative, resulting in an association of the residuals to the SNP.

4.1.1.4 Evidence for need of within and across individual distinction

Although it has been shown how across patterns of correlation in theory may not reflect within individual and potentially causal processes, and how this would have implications for methods that rely on the relational pleiotropy model, it is still possible that there is no need for this distinction in real data.

In order to test this, real repeated measurement data on obesity and physiologically related traits for which causal connections are biologically plausible was used. The within and across correlations for BMI with the traits of fasting blood glucose, triglycerides, cholesterol, high density lipoproteins (HDL), systolic blood pressure (SBP), and diastolic blood pressure (DBP) were estimated using the Framingham heart study data. Simulations were then conducted using real data parameter values in order to show the potential for a reduction in power for mendelian randomization tests used to gauge a causal relationship between BMI and associated traits, and for an increased false positive rate in genetic association tests that systematically control for BMI.

4.1.2 ***Unexploited Biological Variation***

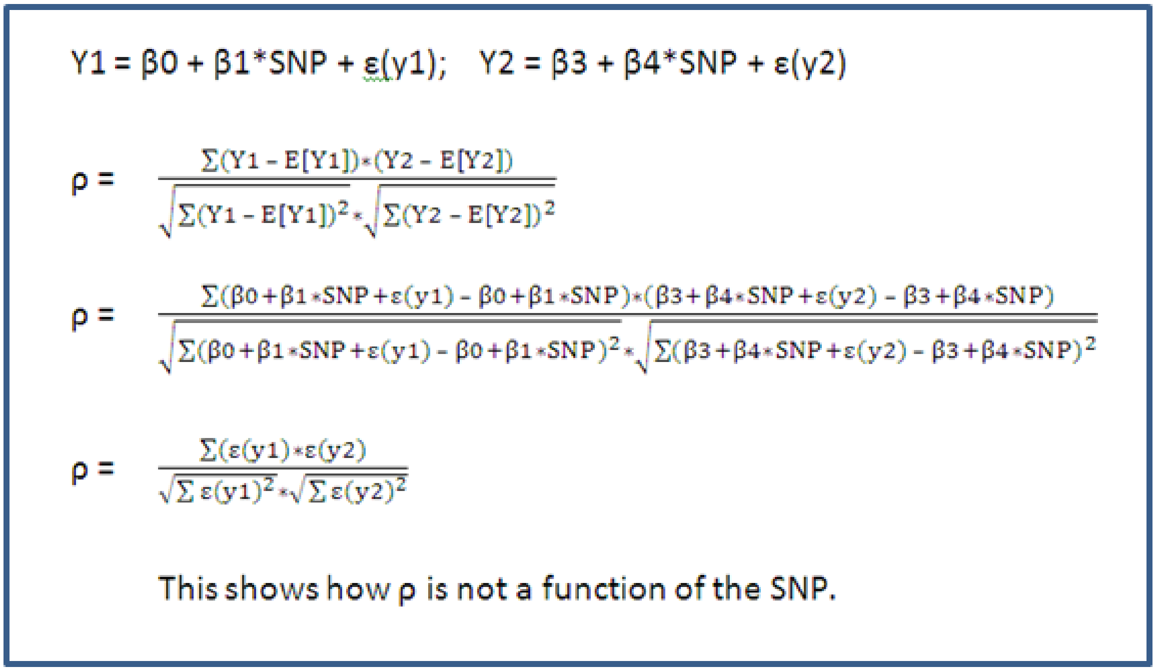
We have described how genes are exclusively across processes by virtue of the fact that they can only vary from individual to individual. For instance, they can be responsible for the correlation of trait expectations across individuals even in the absence of any connection between the traits within individuals, by means of a direct influence on both traits. Genes can also potentially explain the existence of variation in intercepts in the within trait relationships from individual to individual, despite an unchanging within correlation, or b parameter. These would correspond to genes that have an effect on one of the traits but not the other, regardless of any within causal relationship. Along with genes like FTO where there is an effect on diabetes via BMI, there are examples of genes that have an effect on BMI without having any effect on blood sugar or other related traits (Lusis et al. 2008).

A third type of across individual variation for which genes may be responsible is variation in within correlations or slopes, parameter b in equation 4.1, from individual to individual. There is evidence for the existence of this type of variation. For example, although most individuals have a high within correlation between BMI and metabolic traits like cholesterol, blood sugar and blood pressure, where an increase in BMI raises the levels in these traits, while loss of BMI decreases these levels, there is also a subset of individuals, in which fluctuations in BMI do not cause the same effect on the related metabolic traits. They have been termed “the healthy obese”, because of their tendency to remain healthy despite high BMI level (Sims 2001, Meigs et al. 2006). It has also been shown that this type of variation is an important marker for disease prognosis and for the proper course of treatment (Wildman et al. 2008). Genes that underlie this type of variation would therefore be valuable for prediction. Their functional study may also provide a novel source of biological insight into disease mechanism.

If these genes exist, they are not those generally targeted by traditional methods. To understand this let’s look closer at how they compare to genes for which traditional methods are designed. Figure 4.2D is a graphical representation of how the new type of gene relates to the two traits. Such a relationship may be termed associative pleiotropy in order to distinguish it from relational (Figure 4.2A) and mosaic pleiotropy (Figure 4.2B). These genes do not have an effect on the trait values as can be seen by the lack of arrows connecting the gene and the traits themselves. Instead, they have an effect on the trait relationships or correlations within individuals. Because these correlations are independent of trait values, genes that only have an effect on correlations can be easily missed by methodologies where only the genetic effects on trait values are modeled.

Even in the case in which the gene causes an across individual correlation between two traits through its influence on expected trait values across individuals (mosaic pleiotropy) the resulting across correlation is not a function of the gene. Box 4.1 shows how this is the case mathematically: a gene that influences two traits directly is modeled and then its independence to the trait correlations is demonstrated.

Box 4.1 Trait correlations are independent of trait values. Even though the SNP is associated to both traits, the correlation between the two traits is not a function of the SNP.



$$Y1 = \beta0 + \beta1{*}SNP + \varepsilon(y1); \quad Y2 = \beta3 + \beta4{*}SNP + \varepsilon(y2)$$

$$\rho = \frac{\Sigma(Y1 - E[Y1])*(Y2 - E[Y2])}{\sqrt{\Sigma(Y1 - E[Y1])^2}*\sqrt{\Sigma(Y2 - E[Y2])^2}}$$

$$\rho = \frac{\Sigma(\beta0+\beta1*SNP+\varepsilon(y1) - \beta0+\beta1*SNP)*(\beta3+\beta4*SNP+\varepsilon(y2) - \beta3+\beta4*SNP)}{\sqrt{\Sigma(\beta0+\beta1*SNP+\varepsilon(y1) - \beta0+\beta1*SNP)^2}*\sqrt{\Sigma(\beta3+\beta4*SNP+\varepsilon(y2) - \beta3+\beta4*SNP)^2}}$$

$$\rho = \frac{\Sigma(\varepsilon(y1)*\varepsilon(y2)}{\sqrt{\Sigma\,\varepsilon(y1)^2}*\sqrt{\Sigma\,\varepsilon(y2)^2}}$$

This shows how ρ is not a function of the SNP.

Although repeated measures data are sometimes used to search for genetic associations to slopes, the slopes used tend to model the relationship between individual traits and time rather than the relationship between the traits themselves (Strug et al. 2002). Trait correlations may be present regardless of whether or not the traits are associated to time so these studies do not capture the variation in trait associations that we are describing.

Making the within vs. across distinction therefore allows us to bring focus on what may be an underexploited and valuable source of biologically informative variation, i.e. variation in the within individual correlations. If genetics proves to be a factor behind this variation then the treatment of within individual correlations as quantitative traits would be a worthwhile pursuit. Indirect evidence that genes underlie variation in within individual correlations can be obtained from the literature.

Herbert et al. concluded in their study that BMI modifies the association between interleukin-6 (IL-6) genotype and insulin resistance (Herbert et al. 2006). Instead, this could be looked at as an example of associative pleiotropy, where IL-6 genotype modifies the association between BMI and insulin resistance. Because this study only uses one measurement (exam 5, offspring generation in Framingham data) per individual and does not partition within from across variance, it cannot be determined if what is being observed is in fact associative pleiotropy. It can also be a gene-gene interaction with IL-6 that only affects one of the traits. In other words, what is being observed may entirely be an across process where the genes affect the expectation of trait values in individuals rather than how the traits associate within an individual.

The left panel of figure 4.4 shows a hypothetical example with two separate dominant genes where each only produces two phenotypes: A and a for the first gene and B and b for the second gene. We can see how, independently from the within correlations, the interaction effect between these two genes on the trait means, can result in different across correlations for A vs. a. The gene for IL-6 can similarly be showing different *across correlations* for different genotypes. The right panel of figure 4.4 shows how a similar difference between A and a could be observed with a change of within correlations by a single gene. One way to determine if it is in fact an example of associative pleiotropy is to see if the same effect occurs when looking only at the within associations of BMI and insulin resistance.

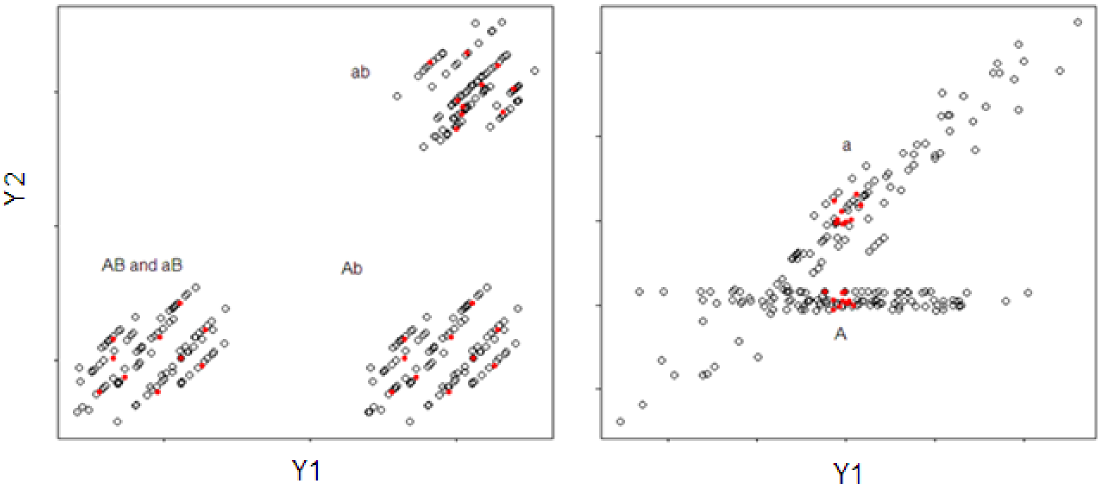


Fig. 4.4 Gene by gene interaction (left panel) vs. the effect of an associative pleiotropy gene (right panel). Red represents trait means for the individual while black represents repeated measurements for each individual. Big A and little a represent alternative genotypic groupings at one locus while big B and little b represent groupings at a different locus.

Another example from the literature of where this type of genetic variation may have been encountered is by Arya et. al. which described a case of “incomplete pleiotropy” of a SNP on the traits of BMI and HDL. Almasy et al. gave their interpretation of incomplete pleiotropy: “had there been epistatic or gene by environment interactions affecting the action of MG4 [the gene in their simulation] on one of Q4 or Q5 [the two traits in their simulation], but not both, the genetic correlation between Q4 and Q5 would have appeared incomplete…” (Almasy et al. 1997). The Arya et al. SNP may very well belong to a gene like gene B in the first plot which fits this Almasy et al. explanation. Alternatively, the SNP could belong to a relational pleiotropy gene, where the gene has an effect on one of the traits directly and the second trait only through its effect on the first. If this kind of gene interacts with an associative pleiotropy gene such as the one depicted in the second plot, then how closely it affects the second trait will depend on this interaction and this too will come through as incomplete pleiotropy.

In order to determine whether IL-6 is truly an associative pleiotropy gene, we revisited the Framingham heart study data and explicitly looked at the effects of this gene on the within individual correlations between blood sugar and BMI. We also evaluated the potential that this variation in the within correlations has for prediction of disease prognosis, with and without accounting for trait values.

4.2 Data

4.2.1 *Increased false positive and false negative rates*

The within individual and across individual correlations (ρ) for BMI and its associated traits were estimated using a subset of 777 unrelated individuals from the offspring generation of the Framingham study data. This subset was created by choosing the eldest individual from each pedigree with complete phenotype data. Measurements of fasting blood glucose, cholesterol, triglycerides, HDL, SBP, DBP, weight and height taken for each subject at Exams 1, 3, 5 and 7 were used.

The Framingham heart study was approved by the Boston University Institutional Review Board and every subject provided informed consent. The author's use of this data was also subject to approval by the Case Western Reserve University Institutional Review Board.

4.2.2 *The IL-6 gene: associative pleiotropy*

Of these 777 unrelated individuals, 723 had complete genotype data for the SNP rs1800795. This data was used for the within individual BMI – blood glucose correlation analysis of the IL-6 polymorphism. Participants were genotyped with the Affymetrix 500k SNP array. Measurements of fasting blood glucose, weight and height and their diabetes status in exams 1, 3, 5 and 7 were used for this analysis.

4.3 **Methods**

4.3.1 ***Increased false positive and false negative rates:*** **data analysis**

The intraclass correlation coefficients λ for BMI (λ1) and its associated traits (λ2) were estimated using equation 4.3. Each individual parameter estimate is given by:

$$\tau^2 = \tau^2\,(\text{observed}) - \sigma^2\,(\text{observed})/n \qquad \text{(eq. 4.5.1)}$$

$$\sigma^2 = \sigma^2\,(\text{observed}) \qquad \text{(eq. 4.5.2)}$$

where τ^2 (observed) and σ^2 (observed), are the variance of the individual means and the mean within individual variance respectively.

The within individual correlations (ρ (within real)) estimate is given by the observed within correlations (R (within observed)) which correspond to the Pearson correlation of the individual deviations for each associated trait with the individual deviations for BMI. The total observed correlation R (total) corresponds to the Pearson correlation of all the measurements across all individuals. Finally, the across individual correlations (ρ (across real)) were estimated using the observed total correlation, R (total) and R(within observed) through the following equation (Snijders and Bosker 1999):

$$\rho\ (\text{across real}) = \frac{\mathbf{R}\ (\text{total}) - \sqrt{(1-\lambda 2)*(1-\lambda 1)} * \mathbf{R}\ (\text{within observed})}{\sqrt{(\lambda 1)*(\lambda 2)}} \qquad \text{(eq. 4.6)}$$

4.3.2 *Increased false positive and false negative rates:* simulations

Using the parameter estimates of within and across correlations for BMI and associated traits, the following were gauged through simulation:

1. The effect of an across correlation that is lower than a within correlation on the power to detect a causal relationship between BMI and the associated trait through a mendelian randomization test.
2. The effect of an across correlation that is higher than a within correlation on the false positive rate when controlling for BMI in a genetic association test.

As a first step measurements of BMI were simulated with a fixed correlation to a SNP. Measurements of the second trait (cholesterol, triglycerides, blood sugar, SBP, DBP or HDL) were then generated with the known ***within*** correlation between BMI and the trait:

$$\text{BMI} = \text{SNP} + \varepsilon_1 \qquad \text{(eq. 4.7.1)}$$

$$\text{Second trait} = \text{BMI} + \varepsilon_2 \qquad \text{(eq. 4.7.2)}$$

The desired correlations are produced by manipulating the variance of the error terms ε_1 and ε_2. Given an explanatory variable with a variance of 1, the variance that is needed in the error term to produce a correlation of ρ is given by:

$$\text{Var}(\varepsilon) = (1\text{-}\rho)^2/\ \rho^2 \qquad \text{(eq. 4.7.3)}$$

This first step represents the relational pleiotropy model depicted in figure 4.2A. The SNP is associated to the second trait only through the second trait's *within* relationship with BMI. This first step was conducted for both simulations. This makes the across correlation equal to the within correlation between the two traits. Because the estimated across correlations differ from the within, a second step is required to account for the difference.

For the cases in which the across correlation is less than the within correlation, perturbations of the second trait that are not associated to BMI or the SNP are generated and added on to the initial data set. The variance of these perturbations is controlled in order to make the total across correlation equal to the estimated across correlation for the two traits. Traits with a lower across correlation with BMI relative to their within correlation will have a greater proportion of the perturbations, lowering the power to detect the influence of the SNP on the second trait through its effect on BMI.

For the cases in which the across correlation is greater than the within correlation, perturbations of BMI and the second trait that themselves have a correlation of 0.7 are added to the initial data set. The BMI perturbations have the same fixed correlation to the SNP while the second trait is not associated to the SNP. This corresponds to the correlated phenotype model in figure 4.2C. Again, the variance of these perturbations is controlled in order to make the total across correlation equal to the estimated across correlation for the two traits. In this way, traits that have a lower within correlation with BMI relative to their across correlation will need a greater proportion of the correlated phenotype model in order to attain their total estimated across correlation. The greater the proportion of the correlated phenotype model is relative to the relational pleiotropy model, the greater the increase in spurious associations of the second trait to the SNP.

These data sets were simulated 1000 times with n=800 and the following tests were conducted:

1. $Y2 = \beta_1 * SNP + \varepsilon$ (eq. 4.8.1)

   The proportion of the 1000 tests that give a significant p-value for β_1 at the 0.05 level is equivalent to power since there is a true association between the second trait (Y2) and the SNP through the causal influence of BMI (Y1) on the second trait (Y2).

2. $Y2 = \beta_1 * SNP + \beta_2 * Y1 + \varepsilon$ (eq. 4.8.2)

   The proportion of the 1000 tests that give a significant p-value for β_1 at the 0.05 level is equivalent to the type I error rate since there is no real association between the second trait (Y2) and the SNP that would remain after controlling for BMI (Y1).

For the traits that have a lower across than within correlation with BMI, and therefore a reduction in power to detect its causal relationship to BMI, a hypothetical trait was simulated with an across correlation equivalent to the within (the most power given the within, i.e. no perturbations) to serve as a point of comparison. For both cases, the simulations were repeated for different values of correlation between BMI and the SNP (range: 0.16-0.85).

4.2.3 *The IL-6 gene: associative pleiotropy*

Within individual Pearson correlations for BMI and fasting blood glucose were computed for each individual using their four repeated measurements. The difference between the mean of these correlations for individuals with the “CC” genotype for the IL-6 gene polymorphism and the mean of the correlations for individuals with the “GC” or “GG” genotype was computed. The significance of this statistic was evaluated via permutation testing. Specifically, the correlations were permuted across the 723 individuals while holding the order of the genotypes 1000 times and a new statistic was calculated each time. Where the observed statistic lies with respect to the empirical distribution obtained from the permutation statistics, gives a measure of its significance. This analysis was stratified by sex.

The association between diabetes status and the within individual correlations for BMI and fasting blood glucose ($\rho_{\text{bmi-gluc}}$) was evaluated using logistic regression. The generalized linear model (GLM) function was implemented using a binomial error distribution in R. First diabetes status was regressed on $\rho_{\text{bmi-gluc}}$ alone and then BMI and fasting blood glucose levels were controlled for. The two respective models were:

$\text{logit (diabetes status)} = \beta_0 + \beta_1 * \rho_{\text{bmi-gluc}}$ MODEL 2.1

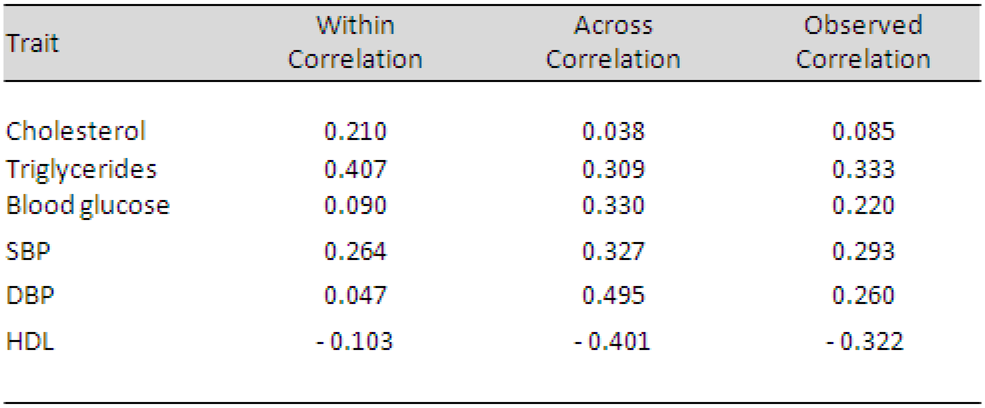
$\text{logit (diabetes status)} = \beta_0 + \beta_1 * \rho_{\text{bmi-gluc}} + \beta_2 \text{ *BMI} + \beta_3 \text{ *blood glucose}$ MODEL 2.2

4.4 Results

4.4.1 ***Increased false positive and false negative rates:*** **data analysis**

Triglycerides and cholesterol levels present a higher within than across correlation to BMI while, SBP, DBP, HDL and blood glucose all present higher across than within correlations (see table 4.1). This variation in within and across correlation profiles allowed for a comparison on their effects on power and type I error via simulation.

Table 4.1 Parameter estimates for the within, across and total observed correlations of traits with BMI.



Trait	Within Correlation	Across Correlation	Observed Correlation
Cholesterol	0.210	0.038	0.085
Triglycerides	0.407	0.309	0.333
Blood glucose	0.090	0.330	0.220
SBP	0.264	0.327	0.293
DBP	0.047	0.495	0.260
HDL	- 0.103	- 0.401	- 0.322

4.4.2 ***Increased false positive and false negative rates:*** **simulations**

Triglycerides and cholesterol were the only traits for which there was a reduction in power to detect their causal relationship to BMI (see figure 4.5). Triglycerides present a much more modest reduction relative to cholesterol. In fact, cholesterol, with the lowest across correlation with BMI ($\rho = 0.035$), has no power beyond the 0.05 alpha level regardless of increases in SNP signal and despite its potentially causal within relationship with BMI (ρ = 0.210). Its curve is no different to that of DBP which has the lowest estimated within correlation (ρ = 0.047). The power curves for SBP, DBP, HDL and blood glucose are a reflection of their total within correlations to BMI since all of the within correlation could be translated to across correlation and no perturbations had to be added. Among these, SBP and DBP have the highest and lowest within correlations with BMI respectively (ρ = 0.264 and ρ = 0.047) and therefore the highest and lowest potential for power in a mendelian randomization test among the four traits.

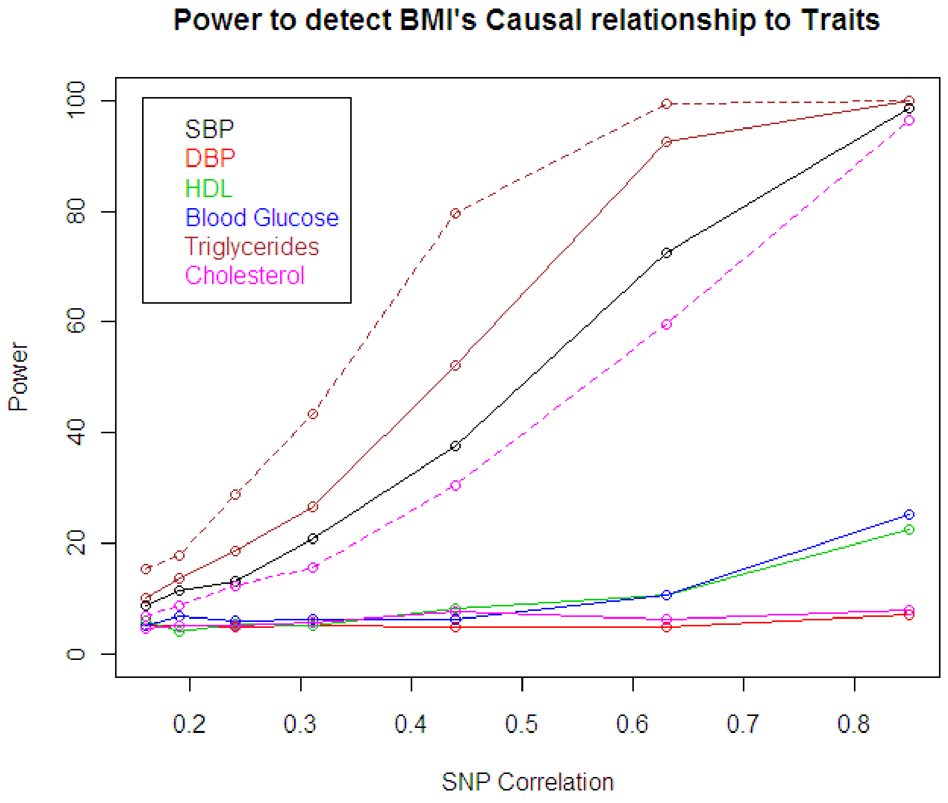


Fig. 4.5 Power when testing causality between a trait and BMI through the mendelian randomization paradigm. The x axis shows the SNP's correlation to BMI. The solid lines represent the power observed when simulating the estimated within and across correlation for the trait with BMI. The dotted lines represent what the power would be given an across correlation that is equal or greater to the estimated within correlation. These only appear for triglycerides and cholesterol, traits in which the estimated across was less than the within.

Triglycerides and cholesterol presented no increase in the false positive rate beyond the 0.05 alpha level when testing their association to the SNP while controlling for BMI since their across correlation can be entirely attributed to within and potentially causal processes (see figure 4.6). SBP, DBP, HDL and blood glucose on the other hand all necessarily have a mix of the relational pleiotropy model with the correlated phenotype model and therefore an increased false positive rate. DBP has the highest across correlation and the lowest within correlation with BMI, giving it the highest potential for spurious associations.

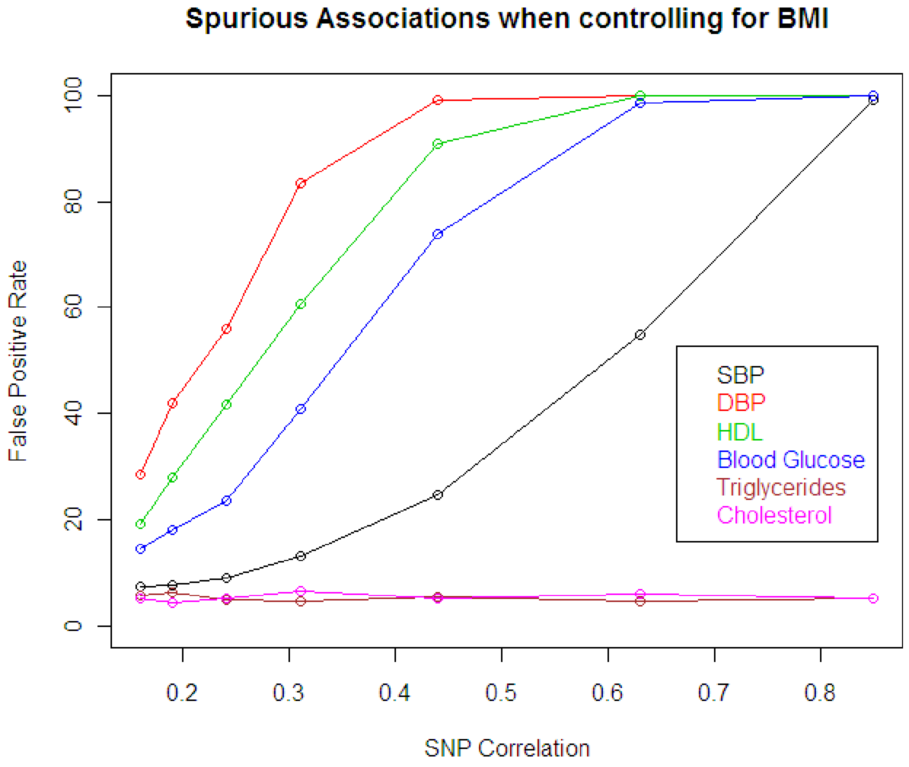


Fig. 4.6 False positive rate when testing for associations between the SNP and the trait while controlling for BMI. The SNP is only directly associated to BMI and its correlation to BMI is represented by the x axis.

4.4.3 *The IL-6 gene: associative pleiotropy*

The sample of unrelated individuals was comprised of 375 men and 348 women, with 49 men and 44 women having the "CC" genotype. There was a suggestive difference (p-value < 0.1) between the within individual correlations for females with the IL-6 genotype "CC" (mean = 0.165) and females with the "GC" or "GG" genotypes (mean = 0.011). Males presented no significant difference.

The density plots show a bimodal distribution of these correlations, with a subset of individuals having a positive within correlation and the other subset having a negative within correlation (see 4. 7). This is particularly evident in individuals with an IL-6 genotype of "CC". Males are evenly distributed among the two modes while females show a higher frequency of individuals with a positive within correlation.

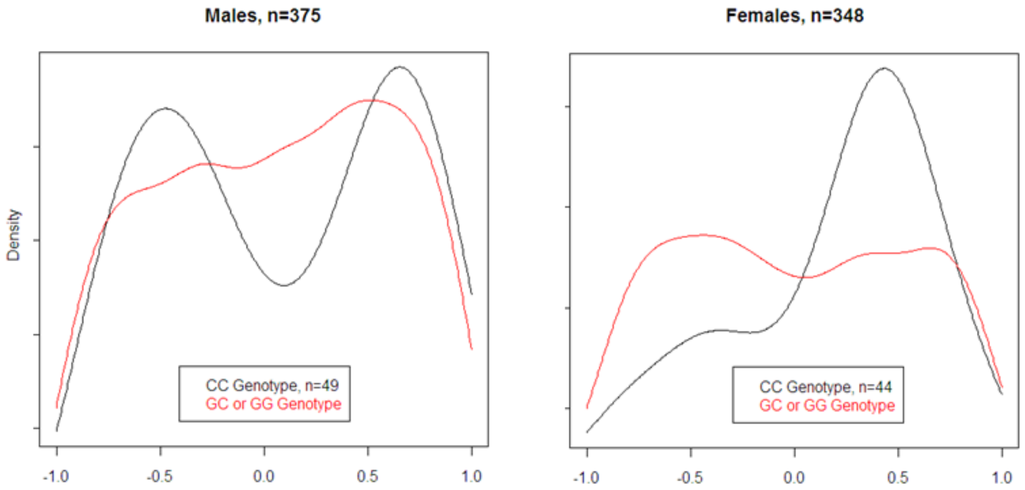
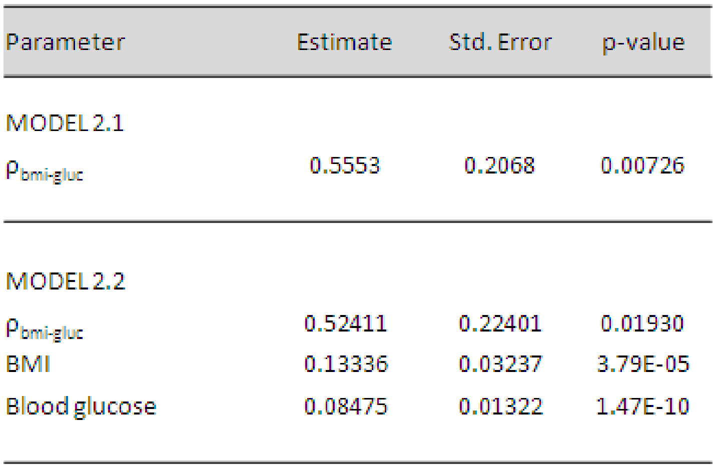


Fig. 4.7 Density plots of within individual correlations for BMI and blood glucose for males and females separated by genotypic group for the IL-6 polymorphism. The x axis shows the within correlation.

The association of $\rho_{bmi\text{-}gluc}$ to diabetes status was very significant ($p<0.01$) (see table 4.2) and retained significance ($p<0.05$) after controlling for both BMI and blood glucose.

Table 4.2 Associations of BMI and blood glucose within correlations to diabetes status, with and without controlling for the trait values.



Parameter	Estimate	Std. Error	p-value
MODEL 2.1			
$\rho_{bmi\text{-}gluc}$	0.5553	0.2068	0.00726
MODEL 2.2			
$\rho_{bmi\text{-}gluc}$	0.52411	0.22401	0.01930
BMI	0.13336	0.03237	3.79E-05
Blood glucose	0.08475	0.01322	1.47E-10

4.5 **Discussion**

4.5.1 ***Unraveling Causality***

The variation in within and across correlation profiles for BMI and its associated traits serves to illustrate the implications that particular profiles can have on methods that assume the relational pleiotropy model. In order to show this the simulations assume that the within correlations are entirely due to a causal influence of BMI on the trait and that if the across correlation is less than the within correlation, then it can entirely be accounted for by this within causal relationship. Both of these assumptions represent the best case scenario for making the relational pleiotropy assumption. If either of these assumptions do not hold, then the power to detect causality will be even less than that simulated and the false positive rates in association tests will be even greater. The results of the simulations therefore contrast the BMI associated traits by showing the highest power and the lowest false positive rates possible given their within and across correlation estimates with BMI.

For instance when the across correlation is the same as the within correlation there is the potential that the across pattern is entirely a reflection of within individual processes. This is the ideal situation for implementing the relational pleiotropy model methods. Triglycerides and SBP correlations with BMI have the closest within and across values. Consequently triglycerides show the lowest reduction in power in mendelian randomization tests and SBP shows the smallest increase in false positives when testing its genetic associations while controlling for BMI (see figures 4.5 and 4.6).

The cholesterol-BMI relationship exemplifies the danger that can underlie implementing the mendelian randomization model without considering the within and across correlation distinction. The within correlation of cholesterol with BMI, $\rho = 0.210$, can very well be due to a causal influence of BMI on cholesterol, and yet, because of the very low across correlation it would go undetected by mendelian randomization regardless of the strength of BMI's association to the SNP and/or sample size (see 4. 5). A high within correlation and low across correlation profile can easily occur if there are genes that influence one of the traits without influencing the other. Some genes may cause variation in cholesterol directly without influencing BMI, disrupting the tie of cholesterol with BMI across individuals. Other genes may cause variation in what constitutes the baseline, healthy

BMI from individual to individual (Sims, 2001) so that some individuals are genetically predisposed to having a high BMI without this having an influence on their cholesterol level (normal cholesterol).

BMI and DBP represent a good example of traits in which the across pattern of variation that is observed cannot be reflecting the within pattern of correlation because the across correlation is higher than the within. When this occurs, there must necessarily be across, non-causal processes accounting for the additional across correlation. We have described how this can cause false positives: if the gene is associated to BMI, then controlling for BMI while testing the association of the gene to DBP will result in a spurious association. But traits like these may also tend to have genes that directly affect them both since this would serve to explain their across correlation, i.e. genes that have a direct influence on both BMI and DBP would cause an increased across correlation pattern regardless of their within relationship. Such genes would result in false positives in mendelian randomization studies of BMI and DBP. They could also result in false negatives if BMI is controlled for in their test of association to DBP. This type of within and across correlation profile can therefore result in more pitfalls when implementing the relational pleiotropy model than the increased false positive rate depicted in figure 4.6.

The point estimates afforded by the limited sample size used (n=777) in the present study for the within and across correlations between BMI and associated traits could be greatly improved by incorporating not only more of the data offered by the Framingham heart study (using methods that take into account family structure), but also by taking advantage of how extensively this set of traits has been addressed in the literature and integrating additional data through meta-analysis. The present estimates still serve the purpose of this study in depicting the utility of making the within and across distinction in the genetics of correlated traits. Table 4.1 shows how not making this distinction and simply looking at the correlation between single measurements of the traits across individuals, i.e. the total observed correlation, completely misses the variation that exists in the within and across correlation profiles for this set of traits.

Although we are often constrained by the nature of the data to not making a within and across distinction simply because one measurement per individual is all that is available, it is also true that this distinction is generally overlooked even in the face of repeated measurements data. The Framingham heart study data provides examples of genome-wide association studies that systematically control for BMI, and of mendelian randomization studies, where this is the case (Levy et al. 2009, Kathiresan et al. 2007, Morris, Gray-McGuire and Stein 2009, (Freathy et al. 2008). A simple look at the within and across patterns of trait variation may be informative when interpreting the results of these studies and any other study that may rely on the relational pleiotropy model. A substantial difference between the across and within correlations, would warn against the implementation of this model.

4.5.2 ***Unexploited biological variation***

By looking at within individual correlations we were able to distinguish whether the IL-6 genotype has an effect on the within individual correlation between BMI and blood glucose, that is, whether it changes this within correlation for a subset of individuals, or if it instead has an effect on the *across* correlations of the traits via, for instance, a gene by gene interaction. Although these results are not quite significant at the alpha = 0.05 level, females do show a suggestive p-value ($p < 0.1$). If Herbert et al.'s result had been due to strictly effects on across correlations, it is unlikely we would have been able to pick up any level of genetic effect after factoring out all of the across variation in the trait values. These results therefore suggest that IL-6 may represent an example of an associative pleiotropy gene.

Females with the CC genotype tend to have a greater correlation between BMI and blood glucose throughout their lives (see figure 4.7). This may mean that they are metabolically less resilient to changes in BMI. This interpretation of higher within correlations would in turn explain why higher within correlations are associated to higher diabetes risk (see table 4.2). An individual whose blood glucose levels remain stable despite increases in BMI will be more resistant to developing diabetes than an individual whose blood glucose levels tend to spike with the slightest BMI increase. This kind of metabolic robustness or fragility is a system level characteristic that is not captured by BMI levels or blood glucose levels alone. Making the within and across distinction in the correlations between the two trait values is what allows us to target this system level characteristic in the form of within correlations. As a quantitative character that varies from individual to individual, the within correlations provide us a window into the genetics that may underlie system robustness or fragility.

Although $\rho_{\text{bmi-gluc}}$ represents a character that is distinct from BMI and blood glucose levels, it could still be the case that it varies in accordance to BMI and blood glucose levels (that the three are correlated) and that it therefore contributes no new information with respect to disease risk. Within correlations, become of particular interest if they prove to be relevant to disease prognosis, i.e. diabetes status, in a way that cannot be explained by the trait values alone. We showed this to be the case with the within correlations for BMI and blood glucose ($\rho_{\text{bmi-gluc}}$) (see table 4.2). This gives associative pleiotropy genes for these within correlations the potential of leading to understudied and possibly unknown disease biology. For example, even if connections between the IL-6 polymorphism and BMI and blood glucose levels are made, as have been made by some studies in the past (Wernstedt et al. 2004, Herbert et al. 2006), follow-up functional studies would be designed to pursue these connections to trait *values*. How this gene modulates the traits' within individual *correlations* would involve an entirely different biological inquiry that would not necessarily be included in these follow-up studies. What is more, in reality, the associations of the gene to BMI and blood glucose values in the literature have been equivocal at best (Huth et al. 2009), with many studies that are unable to replicate previously found associations or that instead find the opposite associations. There may likewise exist other associative pleiotropy genes that are not at all associated to the BMI and blood glucose values in themselves (see box 4.1) and that have therefore until now gone undetected. It would be of interest to see what other genes underlie this understudied type of variation, as well as variation in other trait-pair correlations relevant to metabolic disease as a whole.

4.6 **Conclusion**

Faced with the opportunity of analyzing repeated measurements data, it is of value to study the patterns of within and across variation of the traits of interest and to make a distinction between the across and within biological processes that can be generating these patterns. In particular, it should be kept in mind that genes underlying trait variation are in essence across processes. When trying to unravel causal relationships between the traits and associated genes, this underlines the need to check whether the observed across trait relationships are reflecting the within relationships. Making this distinction also brings to bear that genes may not only be underlying the variation in trait values across individuals, but also variation in the slopes (correlations) of the relationships between traits from individual to individual. This variation in within individual correlations may provide a way of targeting potentially biologically informative genes that are not generally captured by traditional approaches.

Section 5. THE METHODS

Methods for testing genetic effects on within individual trait correlations

5.1 **Introduction**

Faced with a disease that is characterized by the co-occurrence of a collection of traits rather than a single trait we generally conduct univariate or multivariate genetic association studies of the trait *values* to try to get at the genetic variation underlying the disease. These studies do not target the genetic variation that may be underlying the physiological ties between the traits within the individual. These ties can be thought of as within individual correlations. These within individual correlations have been shown to be free to vary independently from the trait values and to have the potential to predict disease in a way not explained by trait values. They have also been shown to have underlying genetic variation that is different from that underlying the trait values. How then can we capture this genetic variation and in so doing open a window into potentially new disease biology? This paper proposes two methods by which to study the genetic variation behind within individual correlations.

Genetic variation in trait correlations may be looked at as genotype by environment interaction problem. For instance, if the trait correlation of interest is between BMI and a metabolically associated trait Y, then BMI can be considered to be part of the internal, physiological "environment" (Panhuysen et al. 2003, Herbert et al. 2006). The corresponding model is:

$$Y= \beta_0+ \beta_1 BMI + \beta_2 G + \beta_3 G_* BMI + \varepsilon \qquad \text{(eq. 5.1)}$$

If variation in genotype changes the slope between BMI and Y, then the interaction term β_3 will be significant. Given variables standardized within each genotypic group, a change in slope is equivalent to a change in correlation between BMI and trait Y. Herbert found the interaction between BMI and genotype at the interleukin-6 locus to have a significant effect on insulin resistance (Herbert et al. 2006). This can be just as easily interpreted as the interleukin-6 locus having a significant effect on the slope between BMI and insulin resistance.

5.1.1 ***Relationships within vs. across individuals***

The model of equation 5.1 confounds within individual correlations with the pattern of correlation observed across individuals. The within individual correlations refer to the correlations between the traits observed throughout each individual's life while the across individual correlations refer to the correlations between the expected values of the traits for each individual. Since what we are interested in capturing is the genetic effects on the physiological relationship between the traits it is important to make this distinction (see paper 1, section 4). Only repeated measurement data can help us differentiate between the two by allowing us to model within and across individual effects independently.

The following equation represents a single individual's relationship between BMI and trait Y. We are able to characterize this relationship for this individual through the repeated measurements denoted by subscript i:

$$Y_i = \beta_0 + \beta_1 BMI_i + \varepsilon \qquad \text{(eq. 5.2)}$$

The β_0 and β_1 parameters are free to vary from individual to individual. What we are interested in biologically is how genetics may underlie variation in the β_1 parameter, the slope, from individual to individual. Once again, since our interest lies in the correlations, this parameter can serve to study these correlations if the trait values are standardized.

The random effects model as described by Snijders allows us see how these within individual relationships can be modeled while factoring out any across individual effects.

Modifying the notation from Snijders (Snijders and Bosker 1999), the following describes a model with random effects at both the repeated measurement level and individual level. The explanatory variables at the measurement level are denoted by $X_1,\ldots, X_p$. Examples of this would be BMI and age since they vary within the individual and could have an effect on the repeated measurements of other within individual time – varying traits such as fasting blood glucose. Those at the individual level are denoted by $Z_1,\ldots Z_q$. An example of such a variable would be genotype since it only varies across individuals. U_{0j} represents the random effects at the individual level, with j being the index for individuals, and R_{ij} represents the random effects at the repeated measurement level with i being the index for measurements. Both are assumed to be independent and normally distributed with a mean of zero. Finally, $\beta_{10},\ldots, \beta_{p0}$ and $\beta_{01},\ldots, \beta_{0q}$ are the parameter estimates for the p measurement level variables and the q individual level variables respectively.

$$Y_{ij} = \beta_{00} + \beta_{10} X_{1ij} + \ldots + \beta_{p0} X_{pij} + \beta_{01} Z_{1j} + \ldots + \beta_{0q} Z_{qj} + U_{0j} + R_{ij} \qquad \text{(eq. 5.3)}$$

Another possibility for an across level variable is the aggregate of a within level variable. An aggregate measure such as the mean for each individual can only vary across individuals and can be modeled to have its own independent effect on the outcome variable (Snijders and Bosker 1999). In order to allow the within individual slope parameter estimate to be different to the across individual slope we must include this aggregate measure of the variable, $\bar{x}_{.j}$, within the model. The following equations show why this is the case:

$$Y_{ij} = \beta_{00} + \beta_{10} x_{ij} + U_{0j} + R_{ij} \qquad \text{(eq. 5.4.1)}$$

$$Y_{ij} = \beta_{00} + \beta_{10} x_{ij} + \beta_{01} \bar{x}_{.j} + U_{0j} + R_{ij} \qquad \text{(eq. 5.4.2)}$$

$$\bar{Y}_{.j} = \beta_{00} + \beta_{10} \bar{x}_{.j} + \beta_{01} \bar{x}_{.j} + U_{0j} + \bar{R}_{.j} \qquad \text{(eq. 5.4.3)}$$

$$\bar{Y}_{.j} = \beta_{00} + (\beta_{10} + \beta_{01}) \bar{x}_{.j} + U_{0j} + \bar{R}_{.j} \qquad \text{(eq. 5.4.4)}$$

Equation 5.4.1 shows the model without the aggregate across level variable, while equation 5.4.2 shows a slope parameter that is modeled separately for all of the measurements x_{ij} and for the mean of x for each individual $\bar{x}_{.j}$, (parameters β_{10} and β_{01} respectively). β_{10} corresponds to the within individual slopes between x and Y, or said differently, the effects of measurements of x on measurements of Y. By taking the means for each individual on both sides of the equation and rearranging terms (equations 5.4.3 and 5.4.4) we can see how $\beta_{10} + \beta_{01}$ is the effect of the means of x on the means of Y. β_{01}

therefore corresponds to the difference between the within individual slopes and the slope across individual means of x and Y(equation 5.4.4). By not including this β_{01} parameter you are in effect making it zero and forcing the within and across individual slopes to be the same. If these are in reality different, forcing them to be the same subsumes the effects of across processes in the parameter estimate for the within individual slopes. Since our interest lies in properly estimating the within individual physiological relationship between the traits it is necessary to factor out the across processes when they exist.

5.1.2 *Adding the within and across individual distinction to the genotype interaction model*

Going back to our interaction model of equation 5.1, we can modify it in order to differentiate between within and across individual effects on slopes. First we must add the random effects component which allows for individual variation in the within individual relationships between BMI and trait Y.

$$Y_{ij} = \beta_{00} + \beta_{10}\, BMI_{ij} + \beta_{01}\, G_j + \beta_{11}\, G_j\, BMI_{ij} + U_{0j} + U_{1j}\, BMI_{ij} + R_{ij} \qquad \text{(eq. 5.5)}$$

In equation 5.5 we have added another independent and normally distributed random effects term, $U_{1j}\, BMI_{ij}$, to account for differences in slope between BMI and Y from individual to individual. The interaction term $\beta_{11}\, G_j\, BMI_{ij}$ gauges the effect of genotype (G) on this individual variation in slope. But once again, this model does not differentiate between across and within effects of BMI on Y thereby restricting them to be the same.

If we add the aggregate measure of mean BMI ($\overline{\mathrm{BMI}}_{.j}$) we get:

$$Y_{ij} = \beta_{00} + \beta_{10}\, BMI_{ij} + \beta_{01}\, G_j + \beta_{02}\, \overline{\mathrm{BMI}}_{.j} + \beta_{03}\, \overline{\mathrm{BMI}}_{.j} G_j + \beta_{11}\, G_j\, BMI_{ij} + U_{0j} + U_{1j}\, BMI_{ij} + R_{ij} \quad \text{(eq.5.6)}$$

Equation 5.6 differentiates between the effect of the genotype on the slope across individuals (β_{11}+β_{03}) and the effect of genotype on differences in within slopes from individual to individual β_{11}. Regrouping the terms in this equation we can see which have an effect on the individual intercepts for the relationship between BMI and Y (equation 5.7 first parenthesis) and which have an effect on the individual slopes (equation 5.7 second parenthesis):

$$Y_{ij} = (\beta_{00} + \beta_{01}\, G_j + \beta_{02}\, \overline{\mathrm{BMI}}_{.j} + \beta_{03}\, \overline{\mathrm{BMI}}_{.j} G_j + U_{0j}) + (\beta_{10} + \beta_{11}\, G_j + U_{1j})BMI_{ij} + R_{ij} \qquad \text{(eq. 5.7)}$$

This allows us to see that the interaction term between mean $\overline{\mathrm{BMI}}_{.j}$ and genotype only affects the intercepts of the within individual relationships without having any effect on the within individual slopes. By not allowing for the within and across individual distinction, it is entirely possible for the Herbert et al. result to have been product of this type of across level interaction. For instance, an interaction between the interleukin 6 gene (IL-6) and another gene could account for the difference in slopes across individual means without this pertaining to any change in the physiological relationship between BMI and insulin resistance from individual to individual.

5.2 Models

5.2.1 *Interaction Model*

Equation 5.7 also allows us to see how we can greatly simplify the model while still factoring out the across processes from the within individual slope estimation. By centering the traits BMI and Y for each individual, all the random and fixed effects on intercepts of the first parenthesis drop out without this having any effect on the terms for the within individual slopes in the second parenthesis. The within individual deviations can therefore replace the actual trait values giving the following equation:

$$(Y_{ij} - \overline{Y}_{.j}) = \beta_{10}\,(BMI_{ij} - \overline{BMI}_{.j}) + \beta_{11}\,G_j\,(BMI_{ij} - \overline{BMI}_{.j}) + U_{1j}(BMI_{ij} - \overline{BMI}_{.j}) + R_{ij} \qquad \text{(eq. 5.8)}$$

U_{1j} corresponds to the random effects in slopes that are not explained by the gene. Since we are not interested in estimating this parameter separately we can subsume U_{1j} into a general error term with R_{ij} to be left with:

$$(Y_{ij} - \overline{Y}_{.j}) = \beta_{10}\,(BMI_{ij} - \overline{BMI}_{.j}) + \beta_{11}\,G_j\,(BMI_{ij} - \overline{BMI}_{.j}) + \varepsilon \qquad \text{(eq. 5.9)}$$

It is very similar to the simple gene by environment interaction model of equation 5.1 but minus the effects of the across processes. Note that the marginal effects of the gene on trait Y are also lost since these correspond to an effect on the intercept of the BMI – Y relationship for the individual. Finally, as mentioned previously, in order to apply the model in a way that actually captures the within correlations between the traits through the estimation of slopes, the traits must be standardized for each individual.

Model inference: As with any simple regression, the interaction parameter β_{11} can be tested for significance through a marginal t-test where the parameter estimate is divided over its standard error in order to obtain the test statistic.

Controlling for covariates: We are not interested in controlling for covariates at within individual level. The reasoning behind this is that the more the traits are perturbed within an individual's life the better estimate we should be able to obtain on their physiological correlation or tie. We instead want to control for across level covariates that can be having an effect on the within individual relationships, covariates that can cause variation in slope from individual to individual and can thereby confound the genotypic effects on the slopes that we want to detect. We can control for these covariate effects by including additional interaction terms of the covariate with BMI. For instance, the effects of sex and age on slopes can be controlled for by adding an interaction term of BMI with sex and of BMI with the aggregate variable mean age.

5.2.2 ***Correlation model***

Another approach to getting at just the within relationships between the traits is a departure from the usual genotype by environment approach and from modeling of the trait values altogether. It consists of modeling the individual trait correlations directly.

Correlations range from -1 to 1 and can have multi-modal distributions making it impossible to model them directly as the response variable in a linear regression model where the assumption of a normally distributed error term must hold. A solution to this is provided by the Fisher-Z transformation which effectively makes correlations normally distributed (Fisher 1915, Ramasundarahettige, Donner and Zou 2009). The following is Fisher's Z transform (F(r)) where r is the correlation value and ln is the natural logarithm function:

$F(r) = ½ * \ln((1+r)/(1-r))$ (eq. 5.10)

The transformed within individual Pearson correlation values (ρBMI-Y_j) can then be modeled with the following simple linear regression, where j is the individual, and G is the individual's genotype.

ρBMI-$Y_j = \beta_1 G_j + \varepsilon_j$ (eq. 5.11)

Model Inference: marginal t-test of the β_1 parameter.

Controlling for covariates: Both across level and aggregate within level covariates can be straightforwardly added to the model as is done with any simple linear regression.

5.3 Methods

5.3.1 ***Simulations***

Y_{ij} and x_{ij} data were simulated with a distribution of N(0,Σ), where $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and i and j are the indexes for the repeated measurements and individuals respectively using the mvrnorm function in R. This resulted in Y and x data with a bivariate standard normal distribution and a *within* individual correlation of ρ between Y and x. For testing type I error, k repeated measurements and n individuals were simulated for correlation values, ρ

= 0.1, 0.3, 0.5 and 0.7 and were compared to another identically generated data set using the interaction and correlation methods described (effect size = 0). A dummy variable (g) was used to keep track of which generated set the data originated from. In order to asses power, k measurements and n individuals were simulated for correlation values of ρ = 0.1, 0.3, 0.5 and 0.7 and compared to another set with correlation values of ρ = 0.2, 0.4, 0.6 and 0.8 (effect size = 0.1). For 8 repeated measurements per individual these simulations were repeated for total sample sizes of n = 800, 1000, 1200 and 1400. For 4 repeated measurements per individual these sample sizes were doubled and for 16 measurements halved so that the total number of data points remained the same across repeated measurements.

The interaction method was applied by first centering and scaling the individuals generated and then using the lm function in R to test for significance of the interaction term at the 0.05 alpha level (the interaction of x with the dummy variable g, through a marginal t-test (see equation 5.9).

The correlation method was applied by computing the Pearson correlation for each individual generated, fisher z transforming them and then using the lm function in R to test for significance of the dummy variable g through its marginal t-test, also at the 0.05 alpha level (see equation 5.11).

5.3.2 *Data analysis*

Both methods were applied to real data using the Framingham heart study. A subset of 777 unrelated individuals were selected from the offspring generation by choosing the eldest individual from each pedigree with complete data. Repeated measurements on BMI and fasting blood glucose from Exams 1, 3, 5 and 7 were used. The individuals were genotyped using Affymetrix 500 SNP array. The effect of ~25 k SNPs on chromosome 7 on within individual BMI - blood glucose correlations were evaluated using the described methods while controlling for sex, age and age^2.

The Framingham heart study was approved by the Boston University Institutional Review Board and every subject provided informed consent. The author's use of this data was also subject to approval by the Case Western Reserve University Institutional Review Board.

5.4 Results

5.4.1 *Type I error*

The interaction method displayed a departure from the nominal type I error of 5% with a low number of repeated measurements per individual (k = 4) (see figure 5.1, left panel). This was not apparent at the higher repeated measurement sizes of 8 and 16 even though the total number of data points was kept constant across all correlations (see figure 5.1, center and right panels). This increased type I error became more pronounced for higher correlation values (9.3% at ρ = 0.1 to 37.1% at ρ = 0.7). The correlation method presented the appropriate type I error rate across all parameter variations.

5.4.2 ***Power***

Only k= 8 and 16 can be used to compare power between the methods since the deviation in type I error at the k = 4 level renders its power results meaningless. The correlation method appeared to have greater power than the interaction method for higher correlations and lower power than the interaction method for lower correlations with constant effect size (effect size = 0.1) (see figure 5.1 center and right panels). The difference in mean power across all n sample sizes between the correlation and the interaction methods for correlations ρ = 0.1, 0.3, 0.5, 0.7 with an effect size of 0.1, were -6.0, -2.3, 3.0 and 8.2 for 8 repeated measurements, -4.1, -1.0, 5.0 and 5.4 for 16 repeated measurements and the same total number of data points and -3.8, -0.22, 1.8, and 0.32 for 16 repeated measurements and double the number of data points. This shows that the influence of correlation level on the difference in power between methods was less pronounced for higher number of repeated measurements, and that given the same number of repeated measurements it was less pronounced for greater number of individuals.

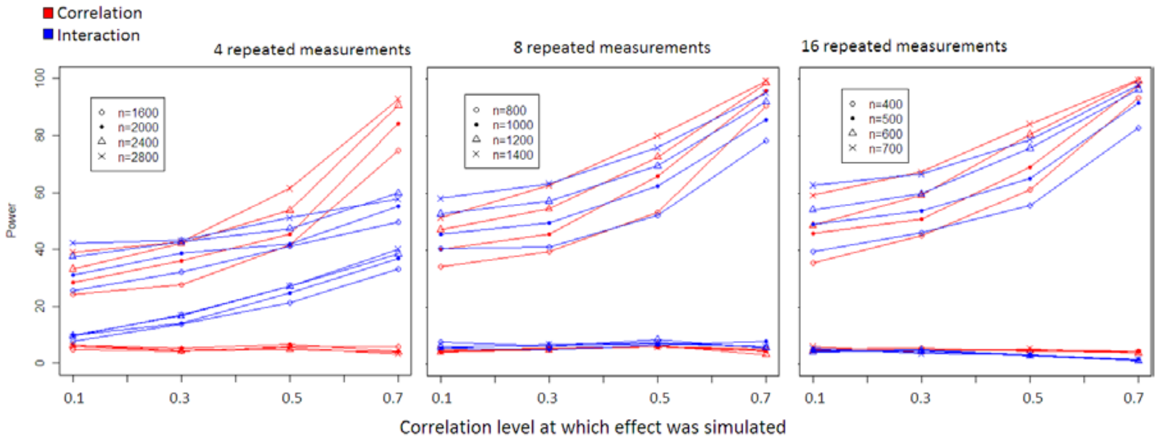


Fig. 5.1 Comparison of type I error and power for correlation and interaction methods. The y axis shows the percentage of tests that showed a significant difference in correlations at the 0.05 level between Y and x for the two data sets generated. The x-axis represents the correlations levels at which effects were simulated. For the bottom curves an effect size (difference in correlations between data sets) of 0 across all correlation values were simulated, making them the type I error curves, while the top curves show an effect size of 0.1 across all correlation values and correspond to the power curves. A total of 6400, 8000, 10600 and 11200 data-points were simulated across all plots distributed differently in terms of number of repeated measurements and number of individuals (n).

5.4.3 *Analysis of real data*

The mean within individual BMI – blood glucose correlation for the Framingham data was found to be 0.06 with a standard deviation of 0.57. Its showed close to a uniform distribution with a -1 to 1 range but was effectively transformed into a normal distribution via Fisher's Z transform (see figure 5.2).

QQplots were generated in order to explore the distribution of the p-values obtained for chromosome 7 SNPs when testing their effect on the within individual correlations of BMI and blood glucose (see figure 5.3). The interaction method clearly shows an overall departure from the uniform distribution, evidence of an increased false positive rate. The correlation method shows better adherence to the uniform distribution but no sign of a significant SNP (a p-value smaller than that expected for the uniform distribution).

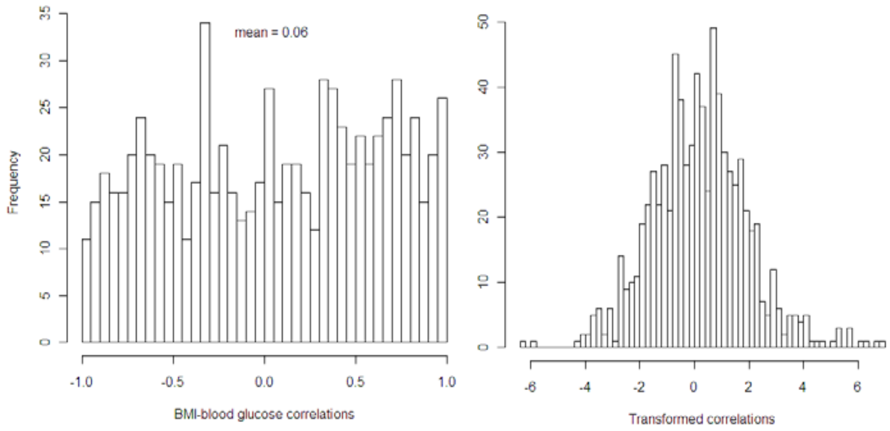


Fig. 5.2 BMI- blood glucose within individual correlations for the Framingham heart study data (n=777, k=4), before (left panel) and after (right panel) applying Fisher's Z transformation.

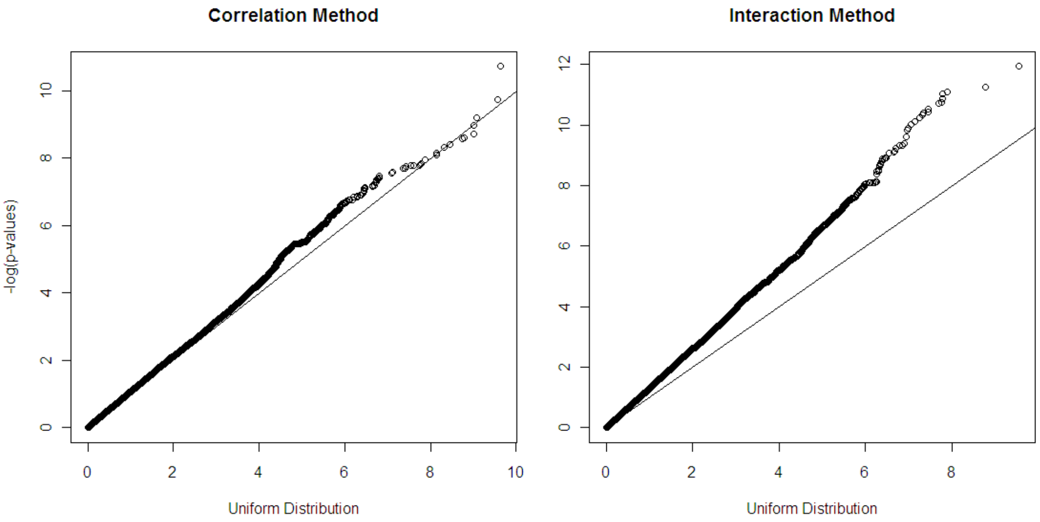


Fig. 5.3 QQ-plots of p-values obtained for the effect of chromosome 7 SNPs on within individual BMI – blood glucose correlations using the correlation and interaction methods.

5.5 Discussion

5.5.1 *Type I error and power comparisons*

5.5.1.1 Simulations

Although the interaction method presents an inflated type I error rate at a low number of repeated measurements (k=4), the power comparisons show it may still be preferable to the correlation method given sufficient repeated measurements when working with lower correlation values. This improvement in power of the interaction model goes away with increasing repeated measurements given a constant total number of data points and/or overall increasing number of data points. This means that there may be a small window of k and n (sample size) in which this increased power in the interaction method can be of benefit.

The type I error problem for the interaction method may at first glance be attributed to a lack of independence in the error terms that is usually associated to repeated measurement data. It has to be pointed out that this problem cannot in itself explain the type I error observed in the simulations where all the individuals within the same group were simulated as having the same means for x and y and the same slopes. In other words, all the measurements simulated were sampled independently from the same distribution with no individual random effects.

Furthermore, when type I error problems do occur due to non-independence of repeated measurement data, they do not subside upon increasing the number of repeated measurements per individual. This is what is observed in our simulations and it points to where the origin of the problem may lie. It may be that with very low number of repeated measurements a spurious non-independence is created as a by-product of the sampling procedure.

The correlation method presents adequate type I error rates across the board and this in itself may outweigh the increase in power evident in the interaction method for lower correlations. Even applying the full random effects model to the interaction model as a way of addressing the type I error problem would entail making assumptions about the covariance matrix for the error term that are difficult to justify. With the correlation method, having to make this type of assumption is by-passed. If working with high correlations ($\rho > 0.5$) the correlation method should be the method of preference.

5.5.1.2 Data

The Framingham data's within individual BMI – blood glucose correlations presented a low overall mean (0.06) which would indicate a potential benefit in using the interaction method. Because of the limited number of repeated measurements available (k=4), this method provides instead an increased false positive rate that is apparent in the qq-plot of the p-values obtained for chromosome 7 SNPs (see figure 5.3 left panel). The absence of any significant results for chromosome 7 when using the correlation method (see figure 5.3 right panel) may be the result of the low power afforded by the sample size of 777 individuals. For a k=4, this power does not exceed 40% (see figure 5.1, left panel).

Population stratification is a common source of inflated type I error rates in association studies and is caused by a systematic difference in allele frequencies between subsets of the population being studied due to non-random mating. Methods such as genomic control and structured association methods have been used to control for population stratification (Zhu et al. 2008, Devlin and Roeder 1999, Pritchard and Rosenberg 1999). Although population stratification was not controlled for in these analyses it is unlikely that it could justify the increased type I error rate observed in the interaction method. The Framingham study population is very homogeneous and the degree of stratification it does present could not cause the level of departure from the uniform distribution observed in the interaction qq-plot. Furthermore, the same type I error problem would have been observed in the application of the correlation method.

5.5.2 *Comparison of methods aside from type I error and power considerations*

Modeling the correlation between traits may go against the natural inclination towards modeling trait values directly, an approach that is still afforded by the interaction method and which may be seen as an advantage. Trait values such as BMI and blood glucose are real physical quantities that can be measured directly. The within individual correlations on the other hand can only be measured with the sampling error imposed by the limited number of repeated measurements on the traits that can be taken for each individual. Real within individual correlations are therefore not directly observable or measureable quantities, a characteristic that qualifies them as a latent variable. By not modeling the within individual correlations as the latent variable they really are but rather taking the *observed* correlation as the real parameter value for each individual, the correlation method confounds sampling variation with individual variation in the real parameter values, i.e. variation in the real underlying correlations.

In the simulations presented here this issue was not addressed. All individuals were simulated as having the same correlation value with no variation from individual to individual. The only variation present was therefore due to sampling. It would be interesting to study the effects of real parameter variation from individual to individual when making inferences with the correlation method via simulation in future work. The increased variation will likely have the effect of lowering power.

The interaction method, on the other hand, presents a disadvantage in that it models a random variable as an independent variable that is being held fixed or measured without error. In our application, where we are interested in modeling the relationship between physiologically tied traits within the individual, the traits in question are by definition all random variables. For instance, in our equations we have considered BMI to be the internal, physiological environment that is interacting with genotype and in so doing we are assuming that we are measuring BMI without error. For linear models this assumption leads to an underestimate of the effect of the predictor variable on the response variable, a situation known as *attenuation bias* (Chesher 1991). A more appropriate approach would model the real BMI value as a parameter separate from the observed, error containing BMI value as is done in measurement error models. This is very similar to the limitation described for the correlation method. In both methods the variance of what is being modeled, be it the latent correlation for the correlation method or the predictor variable for the interaction method, is being underestimated. The fact that the trait values are being modeled directly in the interaction method does not get around what in the end amounts to a variance underestimation issue.

Furthermore, this limitation may be considered especially problematic in the interaction method where it manifests itself in one of the traits being considered and not the other. Only the predictor variable is assumed to be measured without error. This can lead to issues with interpretation. For example, an interaction can appear to be significant when one of the traits is used as the predictor but not the other. It would be unclear what should be concluded in such a case since choice of predictor should not have a bearing on inference on the relationship between the two traits. The real difference in the within individual relationships, the parameter being estimated, does not change with choice of predictor.

The correlation method does not present the asymmetry just the described for the interaction method. Although both methods involve assumptions that disregard existing variance in the data, the correlation method cannot give two contradictory results as can occur with the interaction method. Simulations designed to study the attenuation bias effect on inferences on interactions and how to circumvent the problem with measurement error modeling will be the subject of future work.

In addition to all of the advantages already listed for the correlation method, unlike modeling the interaction, the approach of modeling the correlation value is easily extendable to the large diversity of genetic epidemiological methods used to study variation in trait values. As with the simple regression model shown here, all that is required is to model the within individual correlations as the new quantitative trait after applying the fisher's Z transformation to them to assure normality.

5.6 Conclusion

The correlation method provides a clear advantage over the interaction method for modeling effects on within individual trait correlations. First, it circumvents type I error problems and the assumptions on the error covariance matrix that would have to be made in the interaction method in order to avoid them. Additionally, although the correlation method assumes away the underlying unknown variance in the real within individual correlations from individual to individual, this does not incur the problems in interpretation that the interaction method does when it assumes that there is no underlying unknown variance in the trait used as a predictor. Finally, unlike the interaction method, the correlation method is extendable to the variety of existing genetic epidemiological methods in a straight forward fashion. All of these factors taken together suggest that even though modeling correlations may not have the appeal that many investigators find in modeling tangible physical quantities, it should be the approach of preference when targeting variation in the physiological connections between traits from individual to individual.

Section 6. THE DATA

Association to disease and genetic architecture

# of metabolic trait correlations

6.1 Questions addressed

1. Is variation in metabolic trait correlations (correlations between obesity and its associated traits) relevant to disease?
2. Are there genes underlying this variation? If so, what is the genetic architecture?
3. Is variation in trait correlations redundant to variation in the traits themselves in explaining disease and/or in their genetics?
4. Can the system of trait correlations be studied in an integrated manner and is there an advantage to doing so?

We need to analyze real data in order to answer all of these questions. For the first question it is ideal to use human data where there are clear definitions of what constitutes the disease of interest. The Framingham heart study data will be analyzed with the first question in mind. For the second question human data can have its limitations, the main one being how it can be significantly underpowered. The Framingham heart study data available for these analyses (777 unrelated individuals and 4 repeated measurements) in particular was shown to not exceed 40% power for low correlations in paper 2. A negative answer to question 2 using this data would leave the question unanswered since undetected genetic variation may still exist. Additionally, even if significant associations were found using human data in general, getting at the overall genetic architecture would be a challenge because of the complexity that is characteristic of human data. Chromosome substitution strains of model organisms have been designed with these limitations in mind (Nadeau et al. 2000, Singer et al. 2004). The A/J – C57BL/6J consomic panel will be used for answering the second question. Both sets of data will be used for answering questions 3 and 4.

6.2 Data

For the human data analysis, the Framingham heart study data was used. A subset of 777 unrelated individuals were selected from the offspring generation by choosing the eldest individual from each pedigree with complete data. Repeated measurements on BMI, fasting blood glucose, cholesterol, triglycerides, systolic blood pressure (SBP), diastolic blood pressure (DBP) and high density lipoproteins (HDL) from Exams 1, 3, 5 and 7 were used for obtaining the set of 21 pair-wise correlations for each individual. Information on the use of cholesterol and hypertension medication at each exam and the presence or absence of diabetes and hard coronary heart disease event was also obtained for each individual. The Framingham heart study was approved by the Boston University Institutional Review Board and every subject provided informed consent. The author's use of this data was also subject to approval by the Case Western Reserve University Institutional Review Board.

For the mouse data analysis, 21 chromosome substitution strains, also referred to as "consomics" and their host strain, C57BL/6J, and donor strain A/J were used. Chromosome substitution strains consist of a homozygous genome in which one chromosome is entirely derived from A/J while the rest of the genome is identical to C57BL/6J. How these chromosome subtitution strains (CSSs) were created is described elsewhere (Singer et al. 2004). An average of 37 mice per strain were weaned and started on a high fat and high sucrose diet (Surwit diet) at 5 weeks of age, with a total of 849 mice. After 16 weeks on the diet, at 21 weeks of age, measurements on BMI, liver weight, plasma glucose, insulin and cholesterol and liver triglycerides per gram of liver were taken on the mice. The homeostasis model assessment method quantifies insulin resistance (HOMA-IR) by estimating it with the product of blood glucose and blood insulin measurements divided by a constant. Matthews et al. developed the HOMA model on the basis of physiological studies that described glucose regulation (Matthews et al. 1985).

6.3 Methods

6.3.1 *Association to disease*

6.3.1.1 Univariate

Within individual pair-wise correlations were obtained from the human data by using the 4 repeated measurements for each individual to get at a distinct Pearson correlation value for each individual. The 7 phenotypes of BMI, blood glucose, cholesterol, triglycerides, SBP, DBP and HDL resulted in 21 pair-wise correlations for each of the 777 individuals. Each of these correlations was first transformed using Fisher's Z transformation (F(r)) in order to obtain normality (equation 6.1) and then tested for association to coronary heart disease and diabetes through logistic regression. The generalized linear model (glm) function in R was implemented with a binomial error distribution.

$$F(r) = \tfrac{1}{2} * \ln\left((1+r)/(1-r)\right) \qquad \text{(eq. 6.1)}$$

Association to coronary heart disease was also tested while controlling for medication use. A dummy variable was constructed to indicate which individuals reported using either cholesterol or hypertension medication during any of the exams.

In order to assess whether or not the association to disease observed for the correlations is accounted for by the associations of the trait values to the disease, multiple logistic regression was conducted on the correlations that showed a significant association to disease while controlling for its corresponding trait values to see if this significance was retained.

6.3.1.2 Multivariate

An integrated approach to analyzing the set of 21 correlations was taken by making each individual a vector in p = 21 dimensional space. The vectors could then be grouped and compared according to Euclidean distance in this space, to overall magnitude of correlations and to difference in angle between the vectors. The last two measures are components of the first as can be seen in figure 6.1. In this way we obtained three separate multivariate correlation phenotypes for each individual that we could then test for association to disease. The magnitude phenotype corresponds to overall magnitudes across all pair-wise correlations while the difference in angle corresponds to a "shape" phenotype, where what is being characterized is which correlations are stronger *relative* to other correlations. The Euclidean phenotype is simply a composite of magnitude and shape (see figure 6.1) and corresponds to the vectorized original correlation values. Magnitudes for each individual were obtained by summing the absolute values of each of the correlations. The shape phenotype for each individual was obtained by standardizing each of the correlations by the overall magnitude for that individual.

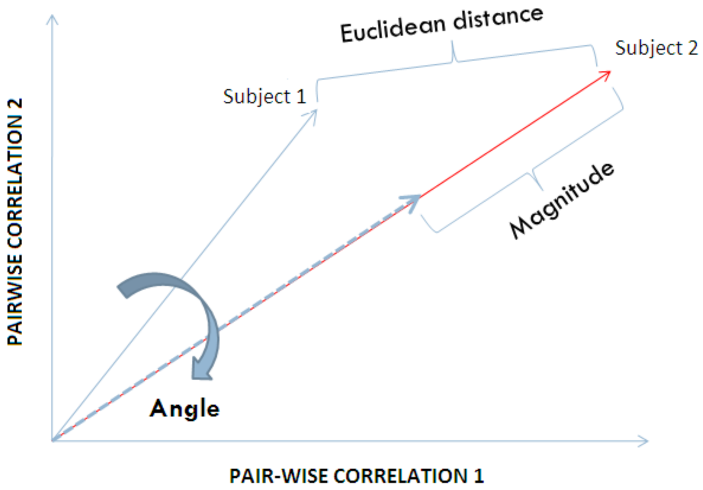


Fig. 6.1 Multivariate analysis of pair-wise correlations. The vectors are composites of pair-wise correlation 1 and 2. An integrated analysis of 2 or more pair-wise correlations is possible through the comparison of these vectors. 3 distinct ways of comparing them are shown. Euclidean distance can be decomposed into a difference in the magnitude and angle. Subjects 1 and 2 can represent different individuals or different strains of mice.

The magnitude phenotype is unidimensional and so was tractable to the same analysis applied to the univariate pair-wise correlations for testing association to coronary heart disease and diabetes: logistic regression. The shape and the Euclidean phenotypes are multidimensional and cannot be analyzed in the same way. Instead, hierarchical clustering was conducted on the 777 individuals using both phenotypes as a way of collapsing their dimensionality into more tractable categorical variables. In hierarchical clustering initially, each individual is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters (the two most distant individuals in the two clusters are compared, i.e. complete linkage method), continuing until there is just a single cluster. In this way the resulting dendrogram organizes all of the data according to Euclidean distance in multidimensional space. The top most clusters in the dendrogram are therefore indicative of the macro-structure of the data if this structure exists, and cluster assignment for each individual can be used as a categorical variable indicative of this macro-structure. The association to coronary heart disease and diabetes of both the top 3 and the top 4 clusters (k= 3 or 4) formed by each of these phenotypes was tested using a chi-square test or a Fisher's exact test where appropriate. The reasoning behind this being that if these phenotypes are of relevance to disease, then they should group individuals in a way that is non-random with respect to disease and the tests conducted should give significant p-values.

6.3.2 *Genetic architecture*

The five phenotypes of BMI, liver size (liver), HOMA, cholesterol (chol) and liver triglycerides per gram of liver (livertri) resulted in 10 pair-wise correlations for each of the 23 strains of mice. The non-linearity of these correlations was easily improved upon via transformations of three of the variables: -1/liver, log(HOMA) and log(livertri). These transformations also improved the distribution of the original traits by making their distribution closer to normal (see figure 6.2). Because of this these transformations were implemented prior to all analyses of the mouse data.

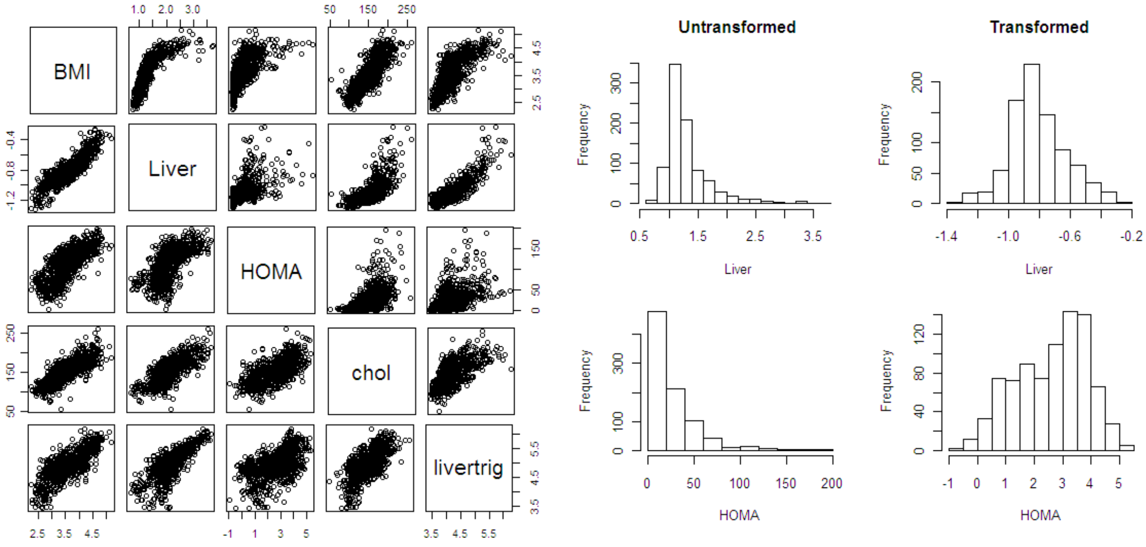


Fig. 6.2 Results of variable transformations. The matrix on the left panel shows variables plotted against each other, before the transformation above the diagonal and after the transformation below the diagonal. The right panel shows the histograms for the variables liver and HOMA before and after their transformations.

6.3.2.1 Univariate

Each of the 10 pair-wise correlations was tested for a significant difference to the host strain C56BL/6J for each of the 22 remaining strains (21 CSSs and 1 donor strain A/J) of mice. Each of the 5 original traits were tested in the same way for a total of 15 variables by 22 strains = 330 tests. A false discovery rate of 5% was implemented for gauging significance over the 330 tests. Significant tests are interpreted as evidence of a single or of multiple QTLs in the genomic segment in which the strain differs from C56BL/6. For each of the CSSs this corresponds to the single chromosome they have from the donor strain A/J for which they are named.

The traits were tested through a t-test conducted with the lm function in R. The correlations were first Fisher's Z transformed (see equation 6.1) and the known standard error of the new normally distributed values were used for their testing. The standard error for a Fisher Z transformed variable is $1/\sqrt{(n-3)}$, where n is the sample size.

The overall genetic architecture of these correlations was explored in a manner which has been implemented before for standard traits (Shao et al. 2008). The effect sizes across CSSs were standardized with the effect size of A/J for each individual pair-wise correlation in which A/J presented a significant effect. This allowed the aggregation of all the test results into a single histogram that serves to show the distribution of both significant and non-significant effects for all 10 pair-wise correlations across CSSs. The total effect sizes for each pair-wise correlation was computed by summing the effect sizes across the 21 CSSs.

Redundancy between genetic variance for trait values and trait correlations was explored by plotting the pair-wise correlation values for all the strains against the means of the trait values of the strains for the two component traits. For example, the values of the BMI-HOMA correlations for the strains were plotted against the mean values of BMI for the strains, and the mean values of HOMA. The corresponding correlations between the pair-wise correlations and each of its component traits were then tested for significance using the cor.test function in R which uses Fisher's Z transform to compute the confidence interval. This was done for all pair-wise correlations.

6.3.2.2 Multivariate

Euclidean, magnitude and shape phenotypes as described in the human data section were also tested for in all 22 strains for a significant difference to C57BL/6J. The significance of the difference between strains was assessed through permutation testing. The differences for each of the multivariate phenotypes between each consomic and C57BL/6J were computed. For the one-dimensional magnitude phenotype this entailed a simple subtraction, while the difference in multidimensional Euclidean and shape phenotypes consisted of Euclidean distance for the former and a correlation value for the latter. The significance of these differences was evaluated via permutation testing. First the data for each consomic was grouped with that of C57BL/6J. Then the rows of trait values were permuted 1000 times in this grouped data while holding the order of the strain labels. New pair-wise correlations for each strain and new multivariate differences were calculated each time. Where the true difference lay with respect to the derived empirical distribution gave a measure of its significance.

Venn diagrams were constructed to show which CSSs show evidence of at least one QTL for all the univariate pair-wise correlations combined, and for each of the multivariate phenotypes in order to compare all approaches as alternative ways of studying the system of correlations as a whole. A bonferroni correction for the 22 strains * 10 correlations = 220 tests was applied to the pair-wise correlations approach and a bonferroni correction of 22 tests was applied to each multivariate phenotype approach. The comparison was repeated using an FDR of 5% as the measure of significance for each approach.

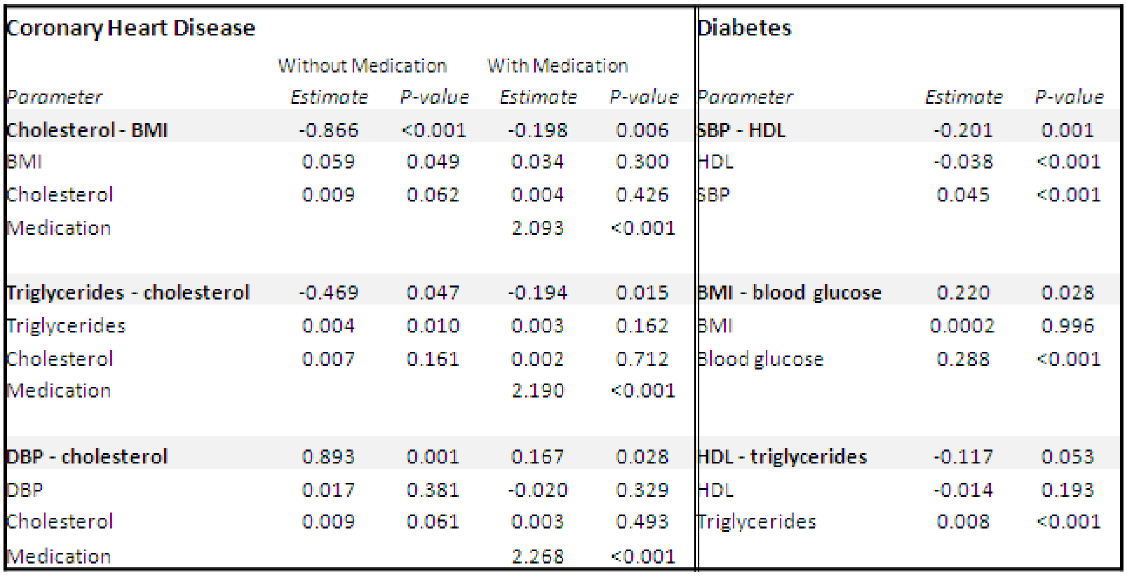
6.4 Results

6.4.1 *Association to disease*

6.4.1.1 Univariate

Although only one of the 21 possible pair-wise correlations for the human data should appear to be significant at the 0.05 alpha level out of chance alone, 9 of them showed this level of significance when testing their association to diabetes. These included the correlations of blood sugar with SBP, DBP, HDL, triglycerides and BMI and the correlations of HDL with SBP, cholesterol, triglycerides and BMI. With bonferroni correction only 3 of these correlations retained significance: SBP - HDL, HDL - triglycerides and BMI - blood glucose. These correlations all showed an increase in diabetes risk with an increase in level of association between the traits; for HDL correlations this means the more negative the correlation the greater the diabetes risk. Of these 3, SBP - HDL and BMI – Blood glucose retained significance at the 0.05 level after controlling for the component traits. HDL – triglycerides was only marginally significant (see table 6.1, right panel).

Table 6.1 Pair-wise correlations that retained significant association to coronary heart disease and diabetes after accounting for the effects of the respective trait values. For coronary heart disease, additional analyses that take into account the effect of hypertension and cholesterol medication were also conducted.



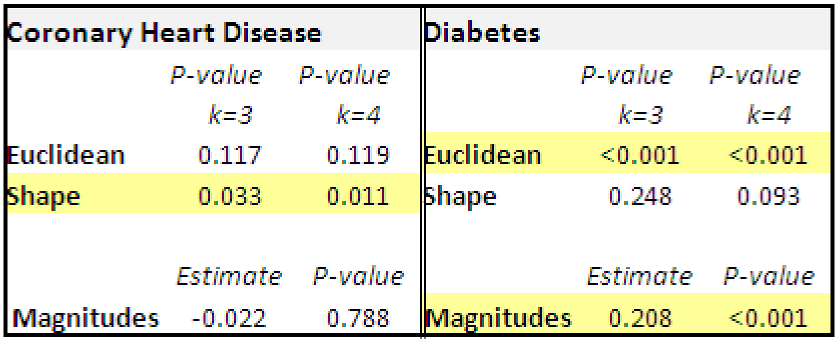
Coronary Heart Disease					**Diabetes**		
	Without Medication		With Medication				
Parameter	*Estimate*	*P-value*	*Estimate*	*P-value*	*Parameter*	*Estimate*	*P-value*
Cholesterol - BMI	-0.866	<0.001	-0.198	0.006	**SBP - HDL**	-0.201	0.001
BMI	0.059	0.049	0.034	0.300	HDL	-0.038	<0.001
Cholesterol	0.009	0.062	0.004	0.426	SBP	0.045	<0.001
Medication			2.093	<0.001			
Triglycerides - cholesterol	-0.469	0.047	-0.194	0.015	**BMI - blood glucose**	0.220	0.028
Triglycerides	0.004	0.010	0.003	0.162	BMI	0.0002	0.996
Cholesterol	0.007	0.161	0.002	0.712	Blood glucose	0.288	<0.001
Medication			2.190	<0.001			
DBP - cholesterol	0.893	0.001	0.167	0.028	**HDL - triglycerides**	-0.117	0.053
DBP	0.017	0.381	-0.020	0.329	HDL	-0.014	0.193
Cholesterol	0.009	0.061	0.003	0.493	Triglycerides	0.008	<0.001
Medication			2.268	<0.001			

When testing associations to coronary heart disease only 3 correlations presented significance at the 0.05 level, the correlations of cholesterol with BMI, triglycerides and SBP. Only 2, cholesterol with BMI and DBP retained significance after bonferroni correction. The same results were obtained when controlling for medications. All three correlations retained significance at the 0.05 level after controlling for their component traits (see table 6.1, left panel). Cholesterol – DBP showed an increased risk of coronary heart disease with an increase in association but cholesterol – BMI and cholesterol – triglycerides showed the opposite effect, an increase in risk with a lower level of association between the traits.

6.4.1.2 Multivariate

The shape phenotype showed a significant association to coronary heart disease while the Euclidean and magnitude phenotypes showed significance in relation to diabetes risk (see table 6.2). The tests for Euclidean and shape phenotype associations were not contingent on choice of number of clusters (choice of k = 3 or 4). The significant association of the magnitude phenotype with diabetes shows that the overall strength of association between traits across all pair-wise correlations results in an increase in diabetes risk.

Table 6.2 Association of multivariate correlation phenotypes to coronary heart disease and diabetes. The phenotypes that were significantly associated to disease are highlighted in yellow.



Coronary Heart Disease			Diabetes		
	P-value	*P-value*		*P-value*	*P-value*
	k=3	*k=4*		*k=3*	*k=4*
Euclidean	0.117	0.119	Euclidean	<0.001	<0.001
Shape	0.033	0.011	Shape	0.248	0.093
	Estimate	*P-value*		*Estimate*	*P-value*
Magnitudes	-0.022	0.788	Magnitudes	0.208	<0.001

6.4.2 ***Genetic Architecture***

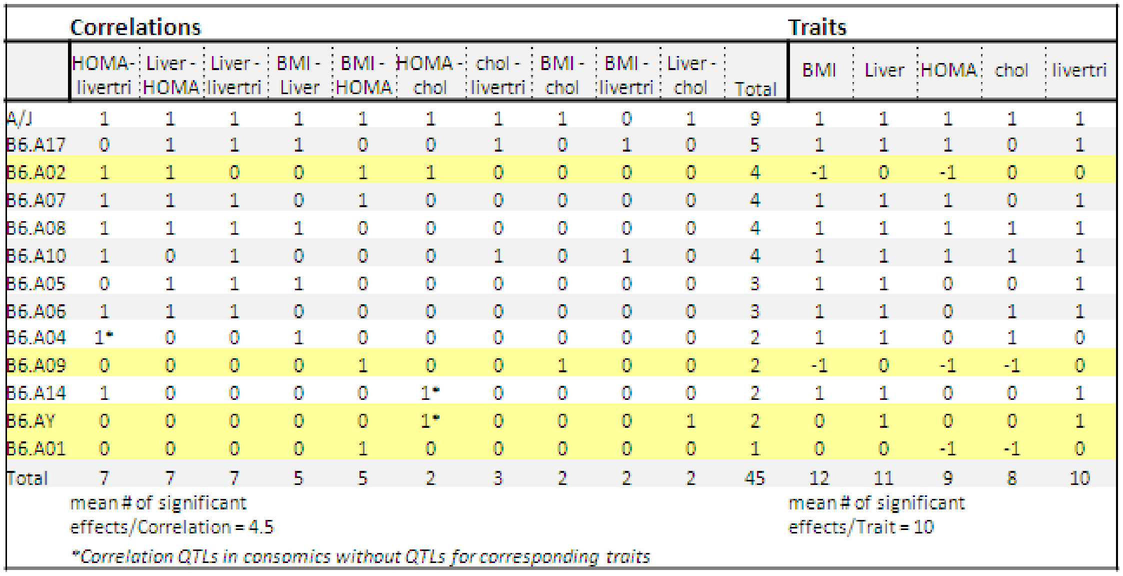
6.4.2.1 Univariate

Out of the 21 CSSs and A/J, 13 strains were found to contain QTLs for at least 1 pair-wise correlation (see table 6.3). All correlations presented at least 2 significant CSSs, with HOMA – livertri, liver – HOMA and liver - livertri at the top of the range with 7 significant CSSs each. An average of 4.5 significant CSSs were found for each correlation while the average number of significant CSSs for the component traits was 10 (p-value<0.001). The CSSs with QTLs for the greatest number of correlations were B6.A17 with QTLs for 5 correlations, and B6.A02, B6.A10, B6.A07 and B6.A08, each with QTLs for 4 correlations.

Although there was a suggestion of an association between genetic susceptibility to obesity and harboring QTLs for correlations, where 4 out of the 10 obese CSSs and 8 out of the 11 lean CSSs show a significant effect on at least one correlation, this association was not significant (p = 0.2).

There was no evidence for redundancy between significance of CSSs for correlations and significance of CSSs for their respective component traits. Although the traits presented the most CSSs with significant effects (average of 10 out of the 13 CSSs that showed significant effects for correlations), it was still possible to find instances in which the traits presented no QTLs in the CSS while the correlations for the traits did. Specifically, HOMA and liver triglycerides do not present QTLs in B6.A04 and yet this CSS presents a QTL for their correlation. The same occurs for B6.A14 and B6.AY and the traits HOMA and cholesterol. The opposite, where there are QTLs for the traits but not for the correlations, was also true and more common given the higher average of significant CSSs for the traits compared to the correlations.

Table 6.3 CSSs with at least one QTL for one of the 10 pairwise correlations analyzed. The corresponding trait value significant CSSs are also displayed. The positive 1s represent a CSS with an effect towards that of the donor strain AJ and negative 1s represent an effect towards the host strain C57BL/6J. Both the CSSs and the pair-wise correlations are sorted according to total number of significant effects. The obese strains are highlighted in yellow.



	Correlations											Traits				
	HOMA-livertri	Liver - HOMA	Liver - livertri	BMI - Liver	BMI - HOMA	HOMA - chol	chol - livertri	BMI - chol	BMI - livertri	Liver - chol	Total	BMI	Liver	HOMA	chol	livertri
A/J	1	1	1	1	1	1	1	1	0	1	9	1	1	1	1	1
B6.A17	0	1	1	1	0	0	1	0	1	0	5	1	1	1	0	1
B6.A02	1	1	0	0	1	1	0	0	0	0	4	-1	0	-1	0	0
B6.A07	1	1	1	0	1	0	0	0	0	0	4	1	1	1	0	1
B6.A08	1	1	1	1	0	0	0	0	0	0	4	1	1	1	1	1
B6.A10	1	0	1	0	0	0	1	0	1	0	4	1	1	1	1	1
B6.A05	0	1	1	1	0	0	0	0	0	0	3	1	1	0	0	1
B6.A06	1	1	1	0	0	0	0	0	0	0	3	1	1	0	1	1
B6.A04	1*	0	0	1	0	0	0	0	0	0	2	1	1	0	1	0
B6.A09	0	0	0	0	1	0	0	1	0	0	2	-1	0	-1	-1	0
B6.A14	1	0	0	0	0	1*	0	0	0	0	2	1	1	0	0	1
B6.AY	0	0	0	0	0	1*	0	0	0	1	2	0	1	0	0	1
B6.A01	0	0	0	0	1	0	0	0	0	0	1	0	0	-1	-1	0
Total	7	7	7	5	5	2	3	2	2	2	45	12	11	9	8	10

mean # of significant effects/Correlation = 4.5

mean # of significant effects/Trait = 10

Correlation QTLs in consomics without QTLs for corresponding traits

A/J presented QTLs for 9 out of 10 pair-wise correlations, the most out of all the strains compared. The only correlation that presented no significant difference between A/J and C57BL/6J was BMI – livertri. CSSs B6.A17 and B6.A10 were found to contain QTLs for this correlation.

All of the significant genetic effects for the correlations in which A/J did present QTLs were found to be in the direction of the donor strain A/J, with all effects in the C57BL/6J direction being non-significant. Most effects are bounded by the A/J and C57BL/6J phenotypes. The non-significant effects show a distribution that is shifted from a mean of zero, suggesting that there are more significant effects that are not being accounted for (see figure 6.3).

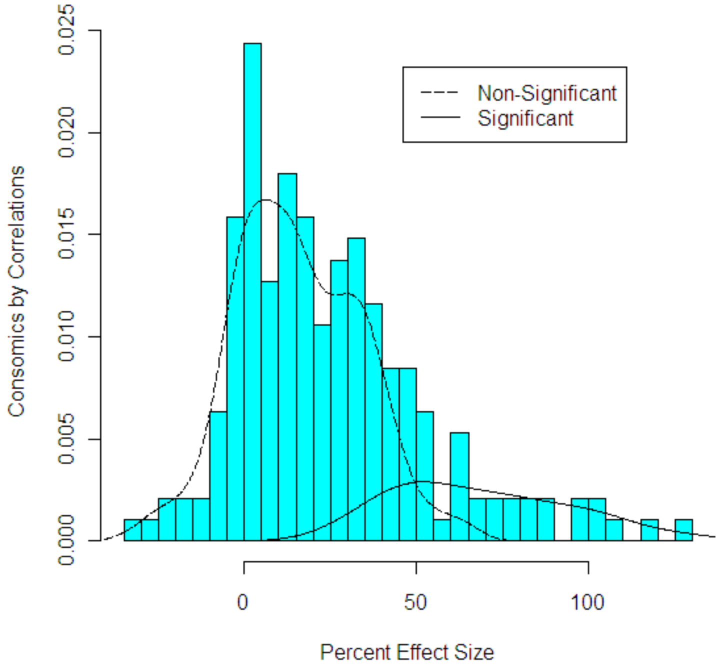
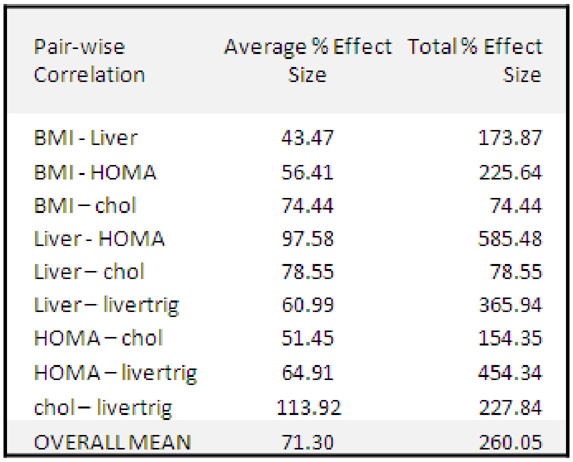


Fig. 6.3 Histogram of effect sizes presented by 21 CSSs standardized by effect size of AJ, for all 10 pair-wise correlations (CSSs by correlations = 210 effect sizes) that were significantly different in AJ compared to C57BL/6J. The smoothed density curves overlaid on histogram are for both significant and non-significant effect sizes.

The overall average effect size for all significant effects was 71.3% which resulted in total effect sizes of over 100% for all correlations except BMI – chol and liver – chol which presented only 1 significant CSS each. These significant CSSs had an effect size of ~ 75% which suggests that additional significant effects may have likewise resulted in a total effect size greater than 100% (see table 6.4).

Table 6.4 The average significant effect size for each pair-wise correlation that showed a significant difference between A/J and C57BL/6J and the total effect size summing across all significant CSSs.



Pair-wise Correlation	Average % Effect Size	Total % Effect Size
BMI - Liver	43.47	173.87
BMI - HOMA	56.41	225.64
BMI – chol	74.44	74.44
Liver - HOMA	97.58	585.48
Liver – chol	78.55	78.55
Liver – livertrig	60.99	365.94
HOMA – chol	51.45	154.35
HOMA – livertrig	64.91	454.34
chol – livertrig	113.92	227.84
OVERALL MEAN	71.30	260.05

There is also a lack of evidence for redundancy between correlations and traits when looking at both significant and non-significant genetic effects. The only significant correlations between the pair-wise correlations and their respective component traits were found for liver – liver triglycerides and for BMI – liver (see figure 6.4). The plots in figure 6.4 also show the distribution of effect sizes for each individual correlation, for instance, for liver – HOMA there are effect sizes that are close to A/J’s for both the traits and the correlations, while for BMI – HOMA there is a clear separation of A/J from the rest of the group. Here too we can see some differentiation between what occurs for correlations and for their respective component traits. For example for liver – HOMA there is a phenotypic separation between A/J and the rest of the strains when looking at the trait liver but not the correlation, while for HOMA – cholesterol the separation exists for the trait HOMA but not for the correlation.

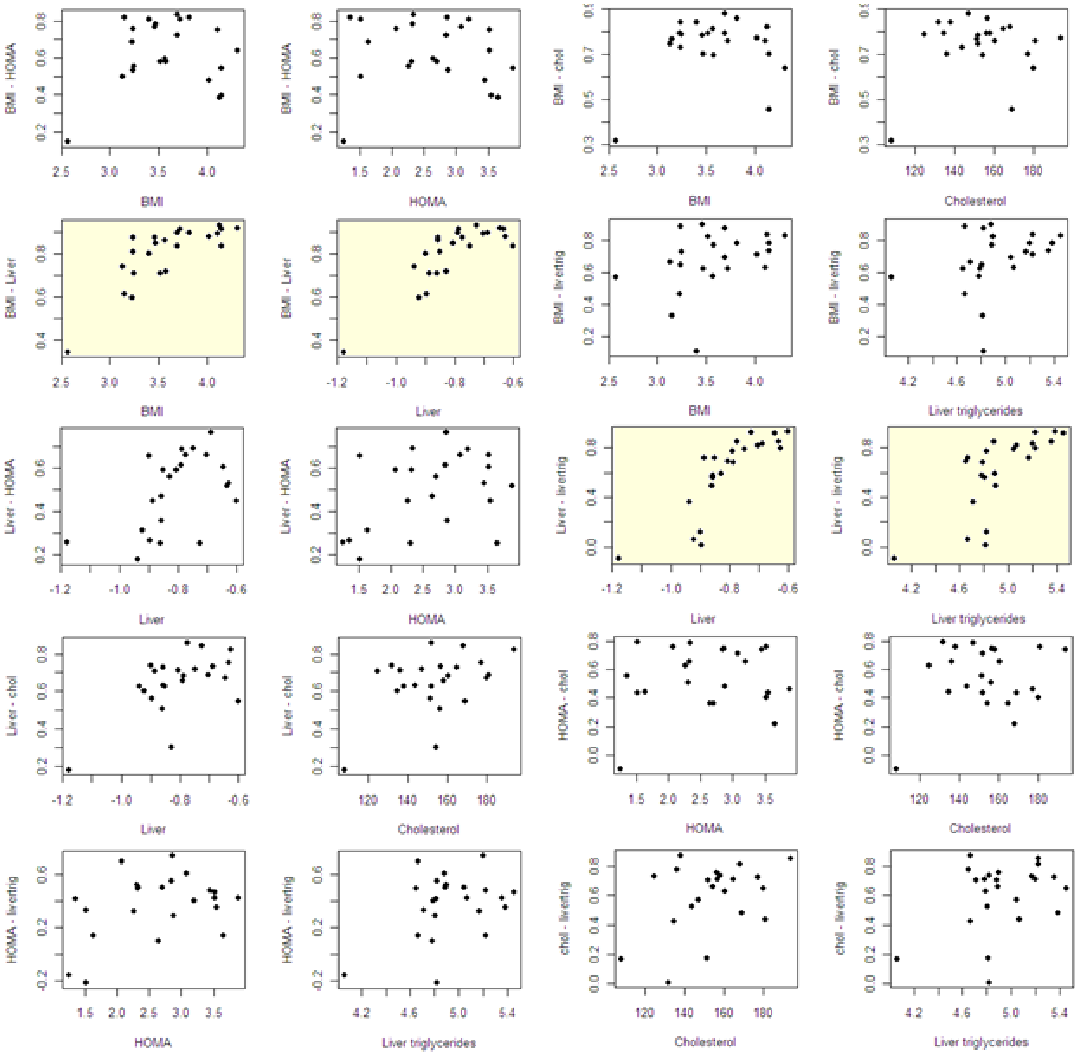


Fig. 6.4 Plots of the 10 pair-wise correlations against their respective composite traits using their mean values across all genetic groups (CSSs as well as A/J and C57BL/6J). The only significant correlations at the 0.05 level are highlighted in yellow.

6.4.2.2 Multivariate

Both the bonferroni corrected and the FDR controlled Venn diagrams (see figure 6.5) show that conducting an analysis of all pair-wise correlations can miss significant CSSs when studying the system of correlations as a whole and that this result is not due to being overly conservative or not conservative enough when dealing with the multiple testing issue. In the bonferroni case, the univariate approach captures 3 out of the 10 significant CSSs while in the FDR case it captures 10 out of the 14 detected. All of the multivariate phenotypes manage to capture CSSs not detected by the univariate method but the Euclidean phenotype method is the one that captures them all. Shape and magnitude phenotype methods are complementary in that they both manage to capture CSSs not detected by the other. B6.AY has an effect on overall magnitudes of the correlations while B6.A04 presents more of a difference from C57BL/6J in the relative strengths of the correlations.

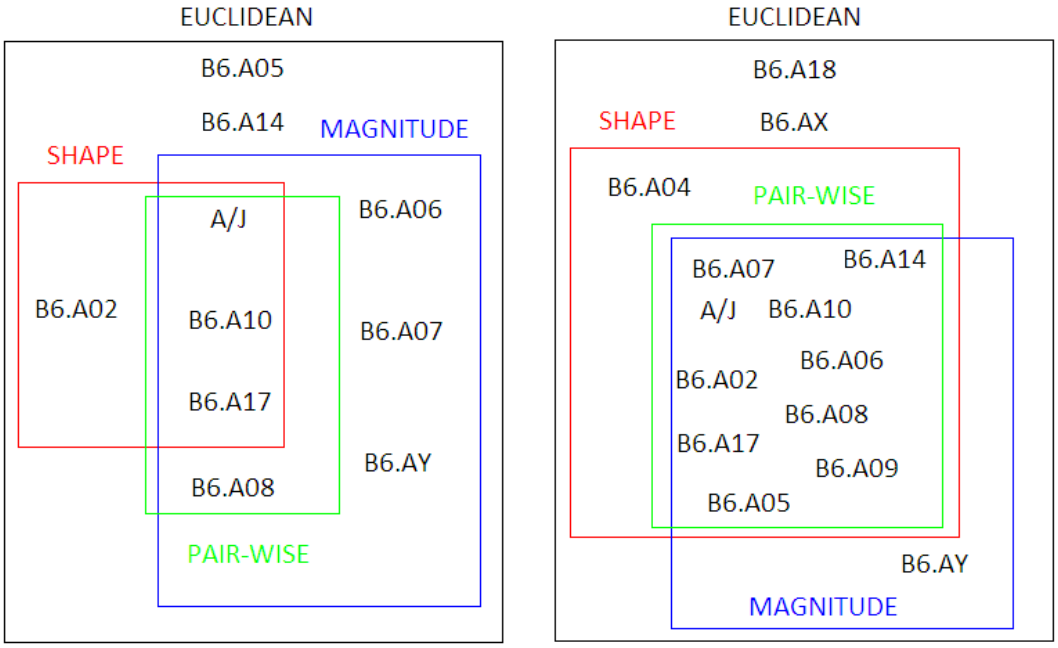


Fig. 6.5 Venn diagrams that compare approaches to finding significant CSSs for a system of correlations. The 4 approaches include testing for differences of the Euclidean phenotype, of the shape phenotype, of the magnitude phenotype and of all the possible pair-wise correlations. The panel on the left shows the CSSs found to have significant QTLs with each method after Bonferroni correction for the total number of tests conducted for each method and the panel on the right shows these while maintaining FDR at ~5%.

6.5 Discussion

6.5.1 *Association of correlations to disease*

The results of the association analyses of pair-wise correlations to disease provide evidence that there is disease relevant information contained in within individual correlation variation. The significance at the 0.05 level for 9 out of the 21 tests showed a considerable enrichment over the single test expected to be significant out of chance alone. This serves to underline the conservativeness of the bonferroni correction when applied to the diabetes analyses. Out of these 9 pair-wise correlations, the three correlations that did retain bonferroni significance also showed a non-redundancy with how the trait values associate to diabetes (see table 6.1, right panel). Even though HDL and SBP are highly associated to the disease, their correlation maintained a highly significant association after controlling for the traits themselves. Blood glucose explains such a high proportion of the risk for diabetes that BMI is no longer significantly associated to the disease when controlling for blood glucose. The BMI - blood glucose correlation on the other hand retains its significance after controlling for both blood glucose and BMI. This suggests that although the individual's mean value of BMI may consist of redundant information regarding diabetes risk once blood glucose level is taken into account, how these two traits track each other throughout the individual's life is not. Something similar happens with the HDL – triglycerides correlation which retains marginal significance even though HDL loses it after controlling for triglycerides, a trait highly associated to diabetes. All of this points to the relevance of looking at trait correlations in the individual rather than just the actual trait values when considering risk for diabetes.

The positive effect of these 3 trait correlations on risk for diabetes indicates that with greater tracking of the traits throughout the individual's life, the greater the risk for the disease. This is in agreement with the result for the multivariate tests (see table 6.2) where diabetes is associated to the magnitude phenotype, the overall magnitude of the correlations, rather than the shape phenotype. The magnitude phenotype is a measure of how strongly the 7 metabolic biological markers considered here as a whole, track each other. A high value in the magnitude phenotype may be indicative of a lack of resilience in the organism's system towards fluctuations in individual trait values. For instance, a spike in BMI may tend to throw the entire system of traits into a positive feedback loop that increases the values of all the traits together, more so in some individuals than in others, putting them at a higher risk for diabetes. The significant association of overall trait correlation magnitudes to diabetes may be indicating that such a susceptibility to positive feedback loops is characteristic of this disease's etiology.

Coronary heart disease did not present the same enrichment of significant associations to pair-wise correlations that diabetes did. Only 3 out of the 21 tests were significant at the 0.05 level, but 2 out of these 3 retained bonferroni significance. All 3 associations retained significance at the 0.05 level after controlling for the trait values and medication. Furthermore, 2 of these pair-wise correlations do a better job at predicting coronary heart disease than the trait values themselves after controlling for medication (see table 6.1, left panel). Both triglycerides and BMI lose their association to coronary heart disease while their correlations to cholesterol do not.

Cholesterol itself was not associated to coronary heart disease in any of the tests shown and yet all the significant pair-wise correlations for coronary heart disease involved the *connection* between cholesterol and another metabolic biological marker. Not only are trait correlations not redundant to the trait values themselves in explaining coronary heart disease, but in some cases they prove to be more informative than the trait values.

Unlike what happens with diabetes where the overall magnitude of the correlations increases disease risk, for coronary heart disease the direction of the effect on risk depends on the particular pair-wise correlation. For the DBP – cholesterol correlation, a higher correlation does imply a higher disease risk, but for cholesterol's connection to BMI and triglycerides, lower correlations increase disease risk. This is in agreement with the result of the multivariate analysis where there was only a significant association to the shape phenotype and not to magnitude phenotype. What is relevant to coronary heart disease is how the traits relate to each other relative to other traits rather than the overall strength of all associations (see table 6.2). One interpretation of this is that the dis-regulation in the physiological connections between particular biological markers plays more of a role in coronary heart disease etiology. This is in contrast to the result for diabetes where overall higher correlations increase risk of disease.

6.5.2 *Mouse data vs. human data*

Mouse data provides distinct advantages to the human data when used to explore the genetics underlying metabolic trait correlations. It optimizes power by providing the possibility of controlling for environment and other confounders, and it allows for the analysis of the correlation's genetic architecture on a genome-wide basis, by providing a way of partitioning the genome into individual chromosomes through engineered chromosome substitution strains.

The mouse data differs from the human data in another very important way that needs to be taken into account when interpreting the results: it does not consist of repeated measurements for a single individual as does the human data. Instead, many genetically identical mice, raised under a controlled environment, are measured at a single time-point. One objection that may be raised is whether the correlations observed across such mice can qualify as a characterization of the *physiological connections* between the traits. Paper 1 (section 5 in this dissertation) describes how these connections can only be inferred from the within individual trait correlations. But it also shows how the across individual correlations reflect the within individual processes when there is no variation in the within individual relationship between the traits from individual to individual. This is the assumption that needs to be made when analyzing the mouse data for physiological trait connections. The connections, or trait relationships, within each mouse should not differ from mouse to mouse when considering genetically identical mice reared in the same environment. Under this assumption, the correlations across mice within the same strain should reflect the within physiological relationships between the traits within individual mice for that strain. Each strain of mice can in this way represent what a single human being did in the human data analysis: a genetically non-varying entity for which multiple measurements on the traits exist.

Another difference to the human data follows from this and that is that the number of measurements used to gauge a single “entity’s” trait correlation. In the case of the human data, 4 repeated measurements were obtained for each individual, while for the mouse data and average of 37 measurements were obtained per strain. One possibility that this opens up for the mouse data and that was not present in the human data is the use of transformations to optimize linearity in the trait relationships. 37 points, unlike 4, can be considered an adequate number for establishing whether there is some non-linearity occurring in the within entity trait relationship. Transformations can then be applied in order to optimize linearity. It is important to point out that the transformations are not conducted in such a way as to optimize the significance of the correlations being studied. In fact, highly significant correlations can occur for completely non-linear relationships where the correlation values are meaningless. This is why it is important to inspect the data and conduct these transformations when studying trait correlations whenever possible, as was done for the mouse data analyses in this study.

6.5.3 ***Genetic variation underlying correlations***

The CSS panel shows trait correlations to have a substantial amount of potential genetic variance underlying them. It remains to be seen how much of the genetic variance represented in the consomic panel is also present in the human population. Nevertheless, the evidence of significant effects on correlations by the CSSs does provide evidence of genes whose biology can have an effect on trait correlations. The gene's biology is very likely to translate from the mouse model to humans even if its variation does not. All pair-wise correlations investigated showed at least two CSSs with significant effects. Three of the correlations, the correlations between the traits HOMA, liver and livertri, presented 7 significant effects each, the highest number of significant effects for any pair-wise correlation. These are very promising results that should spur optimism when considering using genetics as a tool for studying the biology underlying variation in trait correlations.

Although the average number of significant effects for correlations (4.5 per correlation) was significantly lower than the average for the component traits (10 per trait), these effects were not redundant between correlations and traits. Studying the genetic variation underlying the trait values will therefore not capture the same genes that underlie variation in the trait correlations. Considering that there is also evidence that the trait correlations may be of biological relevance to disease in a way not explained by the traits themselves, this suggests that specifically looking for the correlation genes and not just the trait genes may provide novel biological insight into disease etiology.

6.5.4 ***Advantages of multivariate analyses***

One advantage in conducting an integrated analysis of the system of metabolic trait correlations, and in particular, of decomposing the multivariate Euclidean phenotype into the phenotypes of shape and magnitude was already shown. These multivariate phenotypes proved to relate in different ways to coronary heart disease and diabetes in the human data, providing insight into what could be differences in disease etiology. An additional advantage to this decomposition can be pointed out in table 6.2. If only Euclidean phenotype had been tested for an association to disease, no significant association would have been found for coronary heart disease. The variation in overall magnitudes of correlations across individuals which does not seem to be associated to coronary heart disease, was enough to drown out the association to variation in shape when both phenotypes were clumped together as a single Euclidean phenotype. This phenotype's decomposition allowed for more power to detect the coronary heart disease association to the shape phenotype.

The method used to test for associations of the Euclidean and shape phenotypes to disease is unconventional but based on solid reasoning. Unlike the magnitude phenotype where individuals can only differ from one another in one dimension (either greater or lower overall magnitudes), individuals can differ in multiple dimensions for the Euclidean and shape phenotypes. Hierarchical clustering allows us to see the structure of the data in all dimensions by organizing the individuals into groups and hierarchies of these groups according to how close or distant they are. If individuals that have the disease of interest tend to be closer to each other than to individuals that do not have the disease in this multidimensional space, there will be a greater frequency of them is some of the clusters than that expected by chance alone. This is what the association test between cluster assignment and disease status tests for. Although the number of clusters used for the analysis is completely arbitrary, only k=3 and k=4 were tested and both gave a positive result. Even if k=3 or k=4 are not statistically adequate ways of grouping all of the individuals, that is, even if there is too much heterogeneity within these clusters to justify making them single groups, the fact remains that as groups they were found to be enriched for individuals with disease in a way not explained by chance. Disease relevant structure therefore was shown to exist in the data for the Euclidean and shape phenotypes at the macro level, the top 3 and 4 groups of the individuals' hierarchical organization according to the phenotype.

Multivariate analysis was also shown to provide statistical benefits over the analysis of each pair-wise correlation individually in the mouse data (see figure 6.5). The bonferroni and FDR comparison simply shows that this result is not contingent on the level of conservativeness of the tests. The Euclidean phenotype in particular allows for the detection of more QTLs than any other method alone. The shape and magnitude phenotypes also allow for greater power in addition to their offering clues as to the type of effect the CSS may have on the system of trait correlations as a whole, similarly to the way they did for disease.

It is important to note that for the CSS analyses the comparisons being made in multi-dimensional space for the Euclidean and shape phenotypes were necessarily in the direction in which each CSS and A/J differed from C57BL/6J. All of these comparisons could have been made in different directions. For instance, although A/J, B6.A10 and B6.A17 all presented significant effects for shape, the three may differ from C57BL/6J in completely different ways. The traits that are more highly correlated relative to other traits in A/J when compared to C57BL/6J, may not be the same for the B6.A10 and B6.A17 comparisons to C57BL/6J . It would be of great use to develop a way of characterizing these phenotypes in a way that would allow their qualitative comparison as well as their quantitative, so that we could see not only *when* there is a significant difference but also what changes in particular correlations these differences entail.

6.5.5 ***Other considerations allowed by CSS panel***

The A/J and C57BL/6J consomic panel allows for additional insights into the genetics of trait correlations. First it shows that despite the unconventionality of correlations as quantitative traits, the similarities of their genetic architecture when compared to that of more standard traits in chromosome substitution strains, is substantial. The CSSs showed an effect size distribution for the correlations that is largely bounded by the phenotypes of the host and donor strains and with most effects in the direction of the host strain. These results are shared with those of Shao et al.'s when conducting similar analyses on 41 standard traits in the same consomic panel (Shao et al. 2008). The average significant effect size for all correlations was found to be 71.3%, also not far from the 76% figure found by Shao et al. for the standard traits. In correlations, as in traits, there is evidence of pervasive epistasis throughout the entire genome with average cumulative effect sizes of 260% for the correlations.

The relationship between genetic predisposition to obesity and metabolic trait correlations is also something this particular CSS panel allows us to gauge since an important part of the genetic variance segregating has to do with susceptibility or resistance to obesity in the presence of a high fat diet. Because obesity is often thought of as having a causal effect on metabolic traits it would be reasonable to hypothesize that genetic susceptibility to a gain in BMI in the presence of a high fat diet would in turn result in increases in the associated metabolic traits, thereby causing greater correlations between all the traits. This would explain the significantly higher values for all the correlations (with the only exception being the BMI – livertri correlation) in the obesity susceptible strain C57BL/6J when compared to the correlations observed in A/J, an obesity resistant strain. The data did not support this hypothesis. No evidence for an association between obesity susceptible strains and higher correlations was found (p-value = 0.2). Additionally there was a lack of redundancy between the genetic variance for BMI across all the CSSs and A/J when compared to the genetic variance for its correlations, as presented in figure 6.4. Only the BMI-liver correlation presented a positive correlation to the expected value of BMI for each strain.

6.6 Conclusion

Although at first glance studying the genetics underlying variation in trait correlations may seem like a roundabout way of getting at the genetics underlying the traits, the two endeavors are clearly distinct and complementary. Trait correlations, and furthermore, the *system* of trait correlations are currently an underexploited source of biological variation. Considering that this disease relevant variation in correlations exists and that there are unexplored genetics underlying it, as we have shown in this paper, it would be misguided to not pursue correlations as sources of novel insight into disease mechanism.

Section 7. FUTURE WORK

7.1 Interaction and correlation methods

Simulations will be used to narrow down the window in which the interaction method provides greater power than the correlation method without an increased type I error rate.

Simulations will also be conducted to explore the effect of assuming away real variance in the data in both methods.

In the correlation method this will involve simulating variation in real correlations across individuals. In the interaction method different measurement errors will be simulated for both traits and then type I error and power will be compared when using either trait as predictor.

Evidence of the asymmetry that ensues from the assumption of no measurement error in the predictor variable for the interaction method was found in the study of the CSS data. Consomic B6.A11 shows that the significance of its interaction term for strain by trait when compared to C57BL/6J *depends on the choice of trait as predictor*.

For the mice data we do not use a random effects model, instead, within strain variation is analogous to our within individual variation for humans. The model simplifies to:

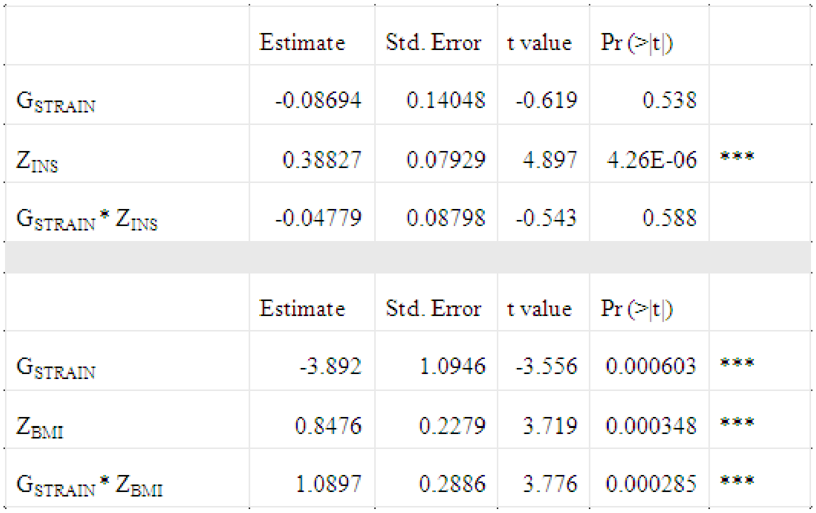
$$Z_{BMI} = G_{STRAIN} + Z_{INS} + G_{STRAIN} * Z_{INS} \qquad \text{(eq. 7.1)}$$

or

$$Z_{INS} = G_{STRAIN} + Z_{BMI} + G_{STRAIN} * Z_{BMI} \qquad \text{(eq. 7.2)}$$

If what is of interest is to test the significance of the difference in the slopes between the two strains, both models should provide the same result. Instead what we find is that the interaction term is highly significant only when using BMI as the predictor variable.

Table 7.1 Interaction test with insulin as the predictor (above) and with BMI as the predictor (below) using C57BL/6J and B6.A11 strains of mice. Significance of test is only observed using BMI as the predictor variable.



	Estimate	Std. Error	t value	Pr (>|t|)	
G_{STRAIN}	-0.08694	0.14048	-0.619	0.538	
Z_{INS}	0.38827	0.07929	4.897	4.26E-06	***
$G_{STRAIN} * Z_{INS}$	-0.04779	0.08798	-0.543	0.588	
	Estimate	Std. Error	t value	Pr (>|t|)	
G_{STRAIN}	-3.892	1.0946	-3.556	0.000603	***
Z_{BMI}	0.8476	0.2279	3.719	0.000348	***
$G_{STRAIN} * Z_{BMI}$	1.0897	0.2886	3.776	0.000285	***

It seems that the trait that is more closely associated to the genetic effect, even if not significantly so, is more likely to give a significant interaction term if used as predictor. The simulations will be geared towards shedding light on this aspect as well.

7.2 The multivariate shape phenotype

A better description of how the shape phentoype contrasts to the magnitude phenotype is required to explain the objective of the future work. Suppose (0.2, 0.5, 0.8) represents the vector of correlations for 3 traits (a total of 3 pair-wise correlations). One possible type of difference can be seen when compared to another vector (0.1, 0.25, 0.4). The difference lies in the extent of the correlations, where the second vector has correlations that are half the size of the first. The magnitude phenotypes for these vectors are accordingly different (1.5 and 0.75 respectively). On the other hand there is no difference in which traits group the most or the least, i.e. the third pair-wise correlation is the greatest in both, and the first is the lowest. In fact, the correlation between these two vectors is exactly 1, which shows that they are exactly the same. A second type of difference can be seen when compared to (0.8, 0.2, 0.5). The overall extent of the correlations is exactly the same, i.e. the magnitude phenotype of both vectors is 1.5. What differs is which traits have the higher correlations and which the lower which makes the correlation between the two vectors less than 1.

When the correlation between the 2 vectors is less than 1, some way of understanding which traits are closer and which are farther apart in one group compared to the other is required. A low correlation between the groups’ vectors of pair-wise correlations only serves to know that they are different, but it says nothing about where the differences lie in terms of the traits. The values for the individual correlations can also be inspected, but it would become too difficult to synthesize what the overall change is as the number of traits being analyzed goes up. For instance, for 3 traits, only 3 pair-wise correlations would have to be inspected, but for 8 traits, 28 values would have to be mentally compared.

A visual aid to understanding where the differences lie can be borrowed from the shape analysis field. Thin plate spline (TPS) software was designed by Rohlf at Stony Brook University to aid in the analysis of landmark data in morphometrics, the field of statistics that studies shape. With this software two or more shapes can be visually compared. When there are more than two shapes, it allows the visualization of their single axis of major shape variation (their first shape principal component). In order to get at this principal component, the software first uses the tangent space approximation to shape space, process which requires certain assumptions to be met by the data. We can use the TPS software to visually compare two vectors of correlations and we can do this by first converting our vectors into shapes.

Correlations are measures of association, but as such they can also be thought of as measures of closeness, a distance measure. Correlation matrices can therefore be conceptualized as summarizing the relative distances between traits. A two dimensional approximation to a vector of correlations could then entail a plot where each of the traits is represented by a point and the relative distances between the points represent their correlations. This two dimensional lay-out of points makes up a shape. A different shape would consist of the same number of points with a different lay-out: different points are closer relative to others. This is exactly the type of information that we want to compare and it is the reason that the shape application works. We are not concerned here with differences in absolute distances between vectors since that aspect is already being compared separately through the magnitude phentoype.

Exactly such a graph can be created by taking the first two principal components of variation across measurements of the traits and plotting the traits through their scores on these two components. The traits that tend to vary together across measurements will be plotted closer together in the two dimensional graph.

Below is an example of two vectors of correlations represented in this way. The green grids serve to show which directions contracted and which expanded between the two vectors. This helps to spot where the major changes occurred. In figure 7.1 a vector of pair-wise correlations is represented by the shape on the right and shows how blood glucose (GLUC) has become much more distant (lower pair-wise correlations) to all the other traits, when compared to the first vector, the shape on the left. The blood pressure traits (SBP and DBP) are tracking age much more in the second vector, as are the other grouped traits, cholesterol (CHOL), triglycerides (TRIG) and weight (WGT) (we have not used age as a trait before in this dissertation but it is an option especially when we want to understand where differences between systems of traits lies). It is much more difficult to discern these patterns by visual inspection of the 28 values making up the each correlation vector.

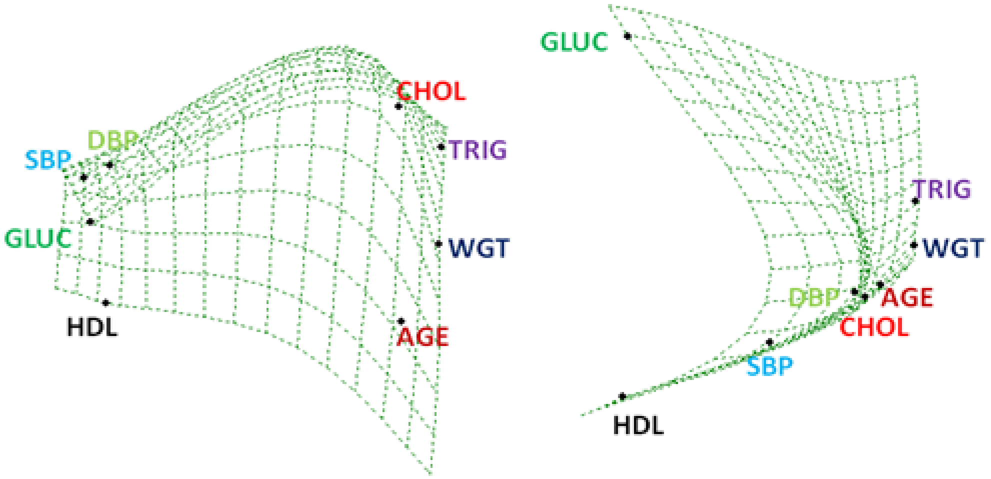


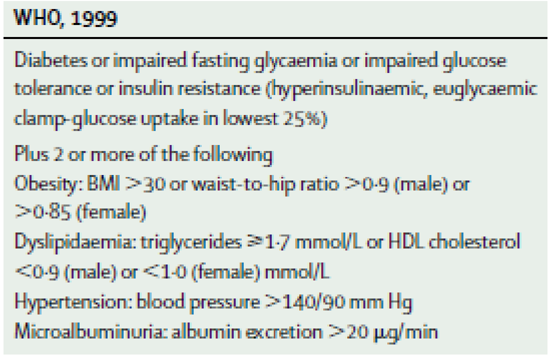
Fig. 7.1 Visualization of shape change across two vectors of pair-wise correlations.

These shapes can also be plotted in shape space which allows for one-dimensional comparisons across CSSs and establishing what type of change is occurring in a single dimension, an otherwise untractable problem for the multidimensional shape phenotype.

More simulations have to be pursued for this application. One issue that was recently discovered is that the method induces a variation in shape that is picked up by the TPS software and that is not real or meaningful and that is the different mirror images of the same shape.

APPENDIX A: Metabolic Syndrome Criteria

Taken from table 1 in Eckel et al 2005.



WHO, 1999

Diabetes or impaired fasting glycaemia or impaired glucose tolerance or insulin resistance (hyperinsulinaemic, euglycaemic clamp-glucose uptake in lowest 25%)

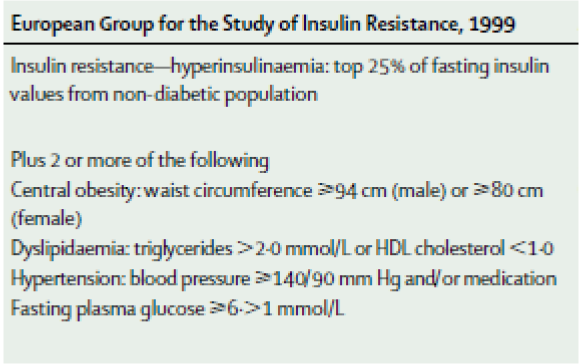
Plus 2 or more of the following

Obesity: BMI >30 or waist-to-hip ratio >0·9 (male) or >0·85 (female)

Dyslipidaemia: triglycerides ≥1·7 mmol/L or HDL cholesterol <0·9 (male) or <1·0 (female) mmol/L

Hypertension: blood pressure >140/90 mm Hg

Microalbuminuria: albumin excretion >20 μg/min



European Group for the Study of Insulin Resistance, 1999

Insulin resistance—hyperinsulinaemia: top 25% of fasting insulin values from non-diabetic population

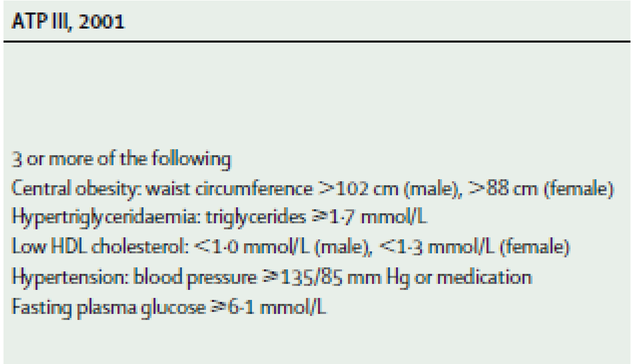
Plus 2 or more of the following

Central obesity: waist circumference ≥94 cm (male) or ≥80 cm (female)

Dyslipidaemia: triglycerides >2·0 mmol/L or HDL cholesterol <1·0

Hypertension: blood pressure ≥140/90 mm Hg and/or medication

Fasting plasma glucose ≥6·>1 mmol/L



ATP III, 2001

3 or more of the following

Central obesity: waist circumference >102 cm (male), >88 cm (female)

Hypertriglyceridaemia: triglycerides ≥1·7 mmol/L

Low HDL cholesterol: <1·0 mmol/L (male), <1·3 mmol/L (female)

Hypertension: blood pressure ≥135/85 mm Hg or medication

Fasting plasma glucose ≥6·1 mmol/L

BIBLIOGRAPHY

Allison, D. B., B. Thiel, P. St Jean, R. C. Elston, M. C. Infante & N. J. Schork (1998) Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. *American Journal of Human Genetics,* 63**,** 1190-1201.

Almasy, L., T. D. Dyer & J. Blangero. 1997. Bivariate quantitative trait linkage analysis: Pleiotropy versus co-incident linkages. 953-958. Wiley-Liss.

Arya, R., J. Blangero, K. Williams, L. Almasy, T. D. Dyer, R. J. Leach, P. O'Connell, M. P. Stern & R. Duggirala (2002) Factors of insulin resistance syndrome-related phenotypes are linked to genetic locations on chromosomes 6 and 7 in nondiabetic Mexican-Americans. *Diabetes,* 51**,** 841-847.

Arya, R., D. Lehman, K. J. Hunt, J. Schneider, L. Almasy, J. Blangero, M. P. Stern & R. Duggirala. 2003. Evidence for bivariate linkage of obesity and HDL-C levels in the Framingham Heart Study. Biomed Central Ltd.

Atwood, L. D., N. L. Heard-Costa, L. A. Cupples, C. E. Jaquish, P. W. F. Wilson & R. B. D'Agostino (2002) Genomewide linkage analysis of body mass index across 28 years of the Framingham Heart Study. *American Journal of Human Genetics,* 71**,** 1044-1050.

Balding ,D. J., M. Bisho, C. Cannings (eds) (2007) Handbook of statistical genetics – 3rd Edn.Volumes 1 and 2, Chichester, England, John Wiley & Sons Ltd.

Benyamin, B., T. I. A. Sorensen, K. Schousboe, M. Fenger, P. M. Visscher & K. O. Kyvik (2007) Are there common genetic and environmental factors behind the endophenotypes associated with the metabolic syndrome? *Diabetologia,* 50**,** 1880-1888.

Bonora, E., F. Saggiani, G. Targher, M. B. Zenere, M. Alberiche, T. Monauni, R. C. Bonadonna & M. Muggeo (2000) Homeostasis model assessment closely mirrors the glucose clamp technique in the assessment of insulin sensitivity - Studies in subjects with various degrees of glucose tolerance and insulin sensitivity. *Diabetes Care,* 23**,** 57-63.

Catenacci, V. A., J. O. Hill & H. R. Wyatt (2009) The Obesity Epidemic. *Clinics in Chest Medicine,* 30**,** 415-+.

Chandalia, M. & N. Abate (2007) Metabolic complications of obesity: inflated or inflamed? *Journal of Diabetes and Its Complications,* 21**,** 128-136.

Chesher, A. (1991) The effect of measurement error. *Biometrika,* 78**,** 451-462.

Cox, H. C., C. Bellis, R. A. Lea, S. Quinlan, R. Hughes, T. Dyer, J. Charlesworth, J. Blangero & L. R. Griffiths (2009) Principal Component and Linkage Analysis of Cardiovascular Risk Traits in the Norfolk Isolate. *Human Heredity,* 68**,** 55-64.

Cupples, L. A., N. Heard-Costa, M. Lee & L. D. Atwood (2009) Genetics Analysis Workshop 16 Problem 2: the Framingham Heart Study data. *BMC Proc,* 3 Suppl 7**,** S3.

Devlin, B. & K. Roeder (1999) Genomic control for association studies. *Biometrics,* 55**,** 997-1004.

Duggirala, R., J. Blangero, L. Almasy, R. Arya, T. D. Dyer, K. L. Williams, R. J. Leach, P. O'Connell & M. P. Stern (2001) A major locus for fasting insulin concentrations and insulin resistance on chromosome 6q with strong pleiotropic effects on obesity-related phenotypes in nondiabetic Mexican Americans. *American Journal of Human Genetics,* 68**,** 1149-1164.

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

Elston, R. C., J. B. Graham, C. H. Miller, H. M. Reisner & B. N. Bouma (1976) Probabilistic classification of Hemophilia-A carriers by discriminant analysis. *Thrombosis Research,* 8**,** 683-695.

Elston, R. C. & A. F. Wilson (1990) Genetic linkage and complex diseases - a comment. *Genetic Epidemiology,* 7**,** 17-19.

Feldman, H. A. & S. M. McKinlay (1994) Cohort versus cross-sectional design in large field trials - precision, sample-size, and a unifying model. *Statistics in Medicine,* 13**,** 61-78.

Fisher, R. A. (1915) Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika,* 10**,** 507-521.

Franklin, S. S., W. Gustin, N. D. Wong, M. G. Larson, M. A. Weber, W. B. Kannel & D. Levy (1997) Hemodynamic patterns of age-related changes in blood pressure - The framingham heart study. *Circulation,* 96**,** 308-315.

Freathy, R. M., N. J. Timpson, D. A. Lawlor, A. Pouta, Y. Ben-Shlomo, A. Ruokonen, S. Ebrahim, B. Shields, E. Zeggini, M. N. Weedon, C. M. Lindgren, H. Lango, D. Melzer, L. Ferrucci, G. Paolisso, M. J. Neville, F. Karpe, C. N. A. Palmer, A. D. Morris, P. Elliott, M. R. Jarvelin, G. D. Smith, M. I. McCarthy, A. T. Hattersley & T. M. Frayling (2008) Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes,* 57**,** 1419-1426.

Gottesman, II & T. D. Gould (2003) The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry,* 160**,** 636-645.

Havill, L. M. & M. C. Mahaney. 2002. Pleiotropic effects on cardiovascular risk factors within and between the fourth and sixth decades of life: Implications for genotype x age interactions. In *13th Genetic Analysis Workshop*. New Orleans, Louisiana.

Herbert, A., C. Y. Liu, S. Karamohamed, J. Liu, A. Manning, C. S. Fox, J. B. Meigs & L. A. Cupples (2006) BMI modifies associations of IL-6 genotypes with insulin resistance: The Framingham Study. *Obesity,* 14**,** 1454-1461.

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

Huth, C., T. Illig, C. Herder, C. Gieger, H. Grallert, C. Vollmert, W. Rathmann, Y. H. Hamid, O. Pedersen, T. Hansen, B. Thorand, C. Meisinger, A. Doring, N. Klopp, H. Gohlke, W. Lieb, C. Hengstenberg, V. Lyssenko, L. Groop, H. Ireland, J. W. Stephens, I. W. Asterholm, J. O. Jansson, H. Boeing, M. Mohlig, H. M. Stringham, M. Boehnke, J. Tuomilehto, J. M. Fernandez-Real, A. Lopez-Bermejo, L. Gallart, J. Vendrell, S. E. Humphries, F. Kronenberg, H. E. Wichmann & I. M. Heid (2009) Joint analysis of individual participants' data from 17 studies on the association of the IL6 variant-174GC with circulating glucose levels, interleukin-6 levels, and body mass index. *Annals of Medicine,* 41**,** 128-138.

Julius, S. & K. Jamerson (1994) Sympathetics, insulin-resistance and coronary risk in hypertension - the chicken-and-egg question. *Journal of Hypertension,* 12**,** 495-502.

Julius, S., K. Jamerson, A. Mejia, L. Krause, N. Schork & K. Jones (1990) THE ASSOCIATION OF BORDERLINE HYPERTENSION WITH TARGET ORGAN CHANGES AND HIGHER CORONARY RISK - TECUMSEH BLOOD-PRESSURE STUDY. *Jama-Journal of the American Medical Association,* 264**,** 354-358.

Kathiresan, S., A. K. Manning, S. Demissie, R. B. D'Agostino, A. Surti, C. Guiducci, L. Gianniny, N. P. Burtt, O. Melander, M. Orho-Melander, D. K. Arnett, G. M. Peloso, J. M. Ordovas & L. A. Cupples (2007) A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *Bmc Medical Genetics,* 8**,** 10.

Kathiresan, S., C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, B. F. Voight, L. L. Bonnycastle, A. U. Jackson, G. Crawford, A. Surti, C. Guiducci, N. P. Burtt, S. Parish, R. Clarke, D. Zelenika, K. A. Kubalanza, M. A. Morken, L. J. Scott, H. M. Stringham, P. Galan, A. J. Swift, J. Kuusisto, R. N. Bergman, J. Sundvall, M. Laakso, L. Ferrucci, P. Scheet, S. Sanna, M. Uda, Q. Yang, K. L. Lunetta, J. Dupuis, P. I. W. de Bakker, C. J. O'Donnell, J. C. Chambers, J. S. Kooner, S. Hercberg, P. Meneton, E. G. Lakatta, A. Scuteri, D. Schlessinger, J. Tuomilehto, F. S. Collins, L. Groop, D. Altshuler, R. Collins, G. M. Lathrop, O. Melander, V. Salomaa, L. Peltonen, M. Orho-Melander, J. M. Ordovas, M. Boehnke, G. R. Abecasis, K. L. Mohlke & L. A. Cupples (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics,* 41**,** 56-65.

Kraft, P., L. Bauman, J. Y. Yuan & S. Horvath. 2002. Multivariate variance-components analysis of longitudinal blood pressure measurements from the Framingham Heart Study. In *13th Genetic Analysis Workshop*. New Orleans, Louisiana: Biomed Central Ltd.

Lawlor, D. A., R. M. Harbord, J. A. C. Sterne, N. Timpson & G. D. Smith (2008) Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine,* 27**,** 1133-1163.

Lazarus, R., D. Sparrow & S. Weiss (1998) Temporal relations between obesity and insulin: Longitudinal data from the normative aging study. *American Journal of Epidemiology,* 147**,** 173-179.

Le Lay, S., P. Ferre & I. Dugail. 2003. Adipocyte cholesterol balance in obesity. In *44th International Conference on Bioscience of Lipids*, 103-106. Oxford, ENGLAND: Portland Press.

Levy, D., A. L. DeStefano, M. G. Larson, C. J. O'Donnell, R. P. Lifton, H. Gavras, L. A. Cupples & R. H. Myers (2000) Evidence for a gene influencing blood pressure on chromosome 17 - Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension,* 36**,** 477-483.

Levy, D., G. B. Ehret, K. Rice, G. C. Verwoert, L. J. Launer, A. Dehghan, N. L. Glazer, A. C. Morrison, A. D. Johnson, T. Aspelund, Y. Aulchenko, T. Lumley, A. Kottgen, R. S. Vasan, F. Rivadeneira, G. Eiriksdottir, X. Q. Guo, D. E. Arking, G. F. Mitchell, F. U. S. Mattace-Raso, A. V. Smith, K. Taylor, R. B. Scharpf, S. J. Hwang, E. J. G. Sijbrands, J. Bis, T. B. Harris, S. K. Ganesh, C. J. O'Donnell, A. Hofman, J. I. Rotter, J. Coresh, E. J. Benjamin, A. G. Uitterlinden, G. Heiss, C. S. Fox, J. C. M. Witteman, E. Boerwinkle, T. J. Wang, V. Gudnason, M. G. Larson, A. Chakravarti, B. M. Psaty & C. M. van Duijn (2009) Genome-wide association study of blood pressure and hypertension. *Nature Genetics,* 41**,** 677-687.

Levy, D., M. G. Larson, E. J. Benjamin, C. Newton-Cheh, T. J. Wang, S. J. Hwang, R. S. Vasan & G. F. Mitchell (2007) Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *Bmc Medical Genetics,* 8**,** 11.

Lusis, A. J., A. D. Attie & K. Reue (2008) Metabolic syndrome: from epidemiology to systems biology. *Nature Reviews Genetics,* 9**,** 819-830.

Matthews, D. R., J. P. Hosker, A. S. Rudenski, B. A. Naylor, D. F. Treacher & R. C. Turner (1985) HOMEOSTASIS MODEL ASSESSMENT - INSULIN RESISTANCE AND BETA-CELL FUNCTION FROM FASTING PLASMA-GLUCOSE AND INSULIN CONCENTRATIONS IN MAN. *Diabetologia,* 28**,** 412-419.

Meigs, J. B., P. W. F. Wilson, C. S. Fox, R. S. Vasan, D. M. Nathan, L. M. Sullivan & R. B. D'Agostino (2006) Body mass index, metabolic syndrome, and risk of type 2 diabetes or cardiovascular disease. *Journal of Clinical Endocrinology & Metabolism,* 91**,** 2906-2912.

Miller, C. H., J. B. Graham, L. R. Goldin & R. C. Elston (1979) Genetics of classic Von Willebrands disease.2. Optimal assignment of the heterozygous genotype (diagnosis) by discriminant-analysis. *Blood,* 54**,** 137-145.

Moore, J. H., J. C. Gilbert, C. T. Tsai, F. T. Chiang, T. Holden, N. Barney & B. C. White (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology,* 241**,** 252-261.

Morris, N. J. (2009) Multivariate and structural equation models for family data. Ph.D. dissertation. Department of Epidemiology and Biostatistics, Case Western Reserve University. Cleveland, OH, USA

Nadeau, J. H., L. C. Burrage, J. Restivo, Y. H. Pao, G. Churchill & B. D. Hoit (2003) Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. *Genome Research,* 13**,** 2082-2091.

Nadeau, J. H., J. B. Singer, A. Matin & E. S. Lander (2000) Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics,* 24**,** 221-225.

Ott, J. & D. Rabinowitz (1999) A principal-components approach based on heritability for combining phenotype information. *Human Heredity,* 49**,** 106-111.

Panhuysen, C. I. M., L. A. Cupples, P. W. F. Wilson, A. G. Herbert, R. H. Myers & J. B. Meigs (2003) A genome scan for loci linked to quantitative insulin traits in persons without diabetes: the Framingham Offspring Study. *Diabetologia,* 46**,** 579-587.

Pritchard, J. K. & N. A. Rosenberg (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics,* 65**,** 220-228.

Ramasundarahettige, C. F., A. Donner & G. Y. Zou (2009) Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Statistics in Medicine,* 28**,** 1041-1053.

Shao, H. F., L. C. Burrage, D. S. Sinasac, A. E. Hill, S. R. Ernest, W. O'Brien, H. W. Courtland, K. J. Jepsen, A. Kirby, E. J. Kulbokas, M. J. Daly, K. W. Broman, E. S. Lander & J. H. Nadeau (2008) Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 19910-19914.

Sims, E. A. H. (2001) Are there persons who are obese, but metabolically healthy? *Metabolism-Clinical and Experimental,* 50**,** 1499-1504.

Singer, J. B., A. E. Hill, L. Burrage, K. R. Olszens, J. H. Song, M. Justice, W. E. O'Brien, D. V. Conti, J. S. Witte, E. S. Lander & J. H. Nadeau (2004) Genetic dissection of complex traits with chromosome substitution strains of mice. *Science,* 304**,** 445-448.

Smith , G. Davey and S. Ebrahim (2003) Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology 32: 1-22

Snijders, T. & R. Bosker (1999) Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage Publications, Thousand Oaks, CA. 266 pp.

Stein, C. M., Y. Song, R. C. Elston, G. Jun, H. K. Tiwari & S. K. Iyengar. 2002. Structural equation model-based genome scan for the metabolic syndrome. In *13th Genetic Analysis Workshop*. New Orleans, Louisiana.

Strug, L., L. Sun & M. Corey. 2002. The genetics of cross-sectional and longitudinal body mass index. In *13th Genetic Analysis Workshop*. New Orleans, Louisiana: Biomed Central Ltd.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander & J. P. Mesirov (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 15545-15550.

Todorov, A. A., G. P. Vogler, C. Gu, M. A. Province, Z. Li, A. C. Heath & D. C. Rao (1998) Testing causal hypotheses in multivariate linkage analysis of quantitative traits: General formulation and application to sibpair data. *Genetic Epidemiology,* 15**,** 263-278.

Wernstedt, I., A. L. Eriksson, A. Berndtsson, J. Hoffstedt, S. Skrtic, T. Hedner, L. M. Hulten, O. Wiklund, C. Ohlsson & J. O. Jansson (2004) A common polymorphism in the interleukin-6 gene promoter is associated with overweight. *International Journal of Obesity,* 28**,** 1272-1279.

Wildman, R. P., P. Muntner, K. Reynolds, A. P. McGinn, S. Rajpathak, J. Wylie-Rosett & M. R. Sowers (2008) The obese without cardiometabolic risk factor clustering and the normal weight with cardiometabolic risk factor clustering - Prevalence and correlates of 2 phenotypes among the US population (NHANES 1999-2004). *Archives of Internal Medicine,* 168**,** 1617-1624.

Xi, B. & J. Mi (2009) FTO Polymorphisms Are Associated with Obesity But Not with Diabetes in East Asian Populations: A Meta-analysis. *Biomedical and Environmental Sciences,* 22**,** 449-457.

Zhu, X. F., S. C. Li, R. S. Cooper & R. C. Elston (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. *American Journal of Human Genetics,* 82**,** 352-365.