VARIANCE COMPONENTS MODELS IN STATISTICAL GENETICS: EXTENSIONS AND APPLICATIONS

by

Feng Dai

B.E., Environmental Engineering,

Dalian University of Technology, 1999

M.S., Biostatistics, University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Feng Dai

It was defended on

June 4, 2007

and approved by

Daniel E. Weeks, Ph.D., Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Sati Mazumder, Ph.D., Professor, Departments of Biostatistics and Psycology, Graduate School of Public Health, University of Pittsburgh

Eleanor Feingold, Ph.D., Associate Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Candace Kammerer, Ph.D., Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Daniel E. Weeks, Ph.D., Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

VARIANCE COMPONENTS MODELS IN STATISTICAL GENETICS: EXTENSIONS AND APPLICATIONS

Feng Dai, PhD

University of Pittsburgh, 2007

Variance components linkage analysis is a powerful method to detect quantitative trait loci (QTLs) for complex diseases. It has the advantages of easy applicability to large extended pedigrees and provides a good flexible framework to accommodate more complicated models like gene-gene, gene-environmental interactions.

This dissertation consists of two major parts. In the first part, I propose two approaches for deriving relative-to-relative covariances that are indispensable for expanding the applications of standard variance components linkage approaches to more complicated genetic models such as those involving genomic imprinting. In the first approach, I extend 'Li and Sacks' ITO method to model ordered genotypes and derive some generalized linear functions of the extended transition matrices. I demonstrate the wide applicability of this extension by applying it to calculate the covariance in unilineal and bilineal relatives under genomic imprinting. In the second approach, I derive a general formula for calculating the genetic covariance using ordered genotypes for any type of relative pair, which does not have the limitation of the extended ITO method to biallelic loci and to unilineal and bilineal relatives. I also propose a recursive algorithm to calculate necessary coefficients in the formula, which opens up the possibility of calculating even inbred relative-to-relative covariance.

In the second part of my dissertation, I discuss linkage evidence for susceptibility loci for adiposity-related phenotypes in the Samoan population, an extensive summary of our multi-center study "Genome-scan for Obesity Susceptibility Loci in Samoans". Obesity, BMI ≥ 30 kg/m^2, in the U.S. has become a major and serious public health problem, affecting 33% of

adults in 2002. Obesity increases risks for serious diet-related diseases, such as cardiovascular disease, type-2 diabetes, and certain forms of cancers. Obesity is a typical multi-factorial disease with overwhelming evidence of genetic effects, yet their roles in obesity are largely unknown. Our current research findings will help further understand the genetics of obesity, which may have great influence on early prevention and later interventions for human obesity, making it a fundamentally important contribution to public health.

PREFACE

I am very grateful to my advisor, Dr. Daniel E. Weeks, for his great guidance, constant support and encouragement in my research work. During the past five years, he always impressed me by his foreseeing scientific insights and strict scientific attitudes. I would like to thank Dr. Eleanor Feingold, Dr. Sati Mazumdar and Dr. Candace Kammerer for serving as my committee members and their valuable discussions and suggestions regarding this dissertation. I want to thank Dr. Stephen T. McGarvey, Dr. Ranjan Deka, Ember D. Keighley, Dr. Karolina Åberg, and all other colleagues from Brown University, University of Cincinnati and University of Pittsburgh for their help in the Samoan project. I am also appreciative to many faculty, staff, and students of Department of Biostatistics and Human Genetics.

Special thanks to Dr. Daniel E. Weeks, Dr. Eleanor Feingold, and Dr. Stephen T. McGarvey for writing reference letters for me that greatly helped in my job search. I owe my success to them.

Finally, I dedicate this dissertation to my family, to whom I am forever indebted for the love and support throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 THE STRUCTURE OF THIS DISSERTATION

This dissertation is organized as follows. In chapter 1, I give a brief introduction to current genetic studies: backgrounds and statistical methods with a focus on variance component (VC) models. A motivation to extend VC models is discussed. In chapter 2, I propose to extend the ITO method to handle ordered genotypes and apply the extension to calculate the covariance in unilineal and bilineal relatives under genomic imprinting. I also derive a general formula for calculating the genetic covariance using ordered genotypes. In chapter 3, I derive a preliminary recursive algorithm for computing the detailed identity coefficients, which are necessary for the analytical calculation of general covariance discussed in chapter 2. In chapter 4, I report findings from our study "genome-scan for Obesity Susceptibility Loci in Samoans" and discuss the significance of those findings in the whole picture of genetics of obesity. In chapter 5, I introduce a simulation study investigating bias in multipoint linkage analysis from map misspecification. I also report findings in our sex-specific linkage analyses of Samoan data. Finally, I conclude the whole dissertation in the chapter 6 and give some discussion about future work.

Chapter 2 of this dissertation discussing the ordered genotype ITO method has been published and is dedicated to our late esteemed colleague, C. C. Li, one of the co-developers of the original ITO method. I would like to thank our reviewers for their helpful comments, and, in particular, to thank Kenneth Lange for suggesting the elegant approach for deriving the general covariance formula. This work was supported by the University of Pittsburgh and by NIH grants R01DK059642 and R01DC005630.

1.2 STATISTICAL GENETICS: CONCEPTS AND METHODOLOGIES

Since the successful completion of the Human Genome Project in 2003, genetic studies of complex human diseases such as cardiovascular diseases, obesity, and diabetes has increased dramatically [1]. In this chapter, necessary genetic models and two main genetic mapping methodologies, linkage analysis and association analysis, are briefly reviewed, followed by new emerging issues in the linkage analysis that motivate our proposed methods described in chapter 2.

1.2.1 Genetic terminology

Each individual has 23 paired *chromosomes*, one is from mother and the other is from father. A region of a chromosome that codes for a protein product is termed as *gene*, which can be passed from parent to child. At each chromosomal location or genetic *locus*, there may be several distinct variants for a gene, known as *alleles*. The composition of two alleles at a locus is called the *genotype*. The relative frequency of an allele in a population is defined as *allele frequency*. Two alleles at a locus are called *homozygous* if they are identical, and *heterozygous* if they are different [1]. An ordered arrangement of alleles found on a single chromosome is called *haplotype*. The probability of an allele or haplotype inherited from one generation to the next is called *transmission* probability.

When two alleles are sufficiently close together (linked) on the same chromosome, they tend to be transmitted to the same gamete (sperm or egg) in a process called *meiosis* [1]. If two alleles are far apart on the same chromosome or on the different chromosomes, they are transmitted to the same or different gametes with equal probability, *i.e.*, their recombination frequency is 50%. During the time of gamete formation, the chromosome strands pair up and each may swap a portion of its genetic material for the matching portion from its mate by a process known as *crossing over*, a form of *genetic recombination*, which is a source of *genetic diversity*.

The observable manifestation of an alleles or a specific genotype is denoted as the *phenotype* of an individual. Most commonly phenotype was used to describe the observable trait

(disease) manifestations. The term *penetrance* denotes the probability that an individual will express the disease phenotype given that they inherited the disease-causing genotype or allele(s). The phenotype may be discrete (*qualitative*) or continuous (*quantitative*) [1]. By a quantitative trait, we mean a measurable human trait that shows continuous variation. Typically we refer to the chromosomal loci that influence human quantitative traits in humans as quantitative trait loci (QTL). A typical genetic analysis concerns modeling the relationships between underlying genotypes and observed phenotype of an individual. Many statistical methods have been developed from the ad hoc scientific efforts of dissecting these relationships, and they can be divided into two broad categories: (i) genetic linkage analysis and (ii) association analysis. Both methods rely on the similar principles and assumptions [2], that is, "people who have similar trait values should have higher than expected levels of sharing of genetic materials near the genes that influence those traits" (page 1, [3]).

(i) Linkage analysis

Traditionally, the search for a disease gene begins with linkage analysis, which aims at finding the rough location of the gene relative to a DNA segment called a genetic *marker* with known position on a chromosome, an effort to narrow down the region of interest so that conventional molecular approaches can then be used to identify the specific defects underlying the disease. Linkage analysis asks the question of whether certain genetic material co-segregate or not with disease of interest in each family, and it consists of estimating the *genetic distance* (or *recombination fraction*) between a measured marker and a trait (disease) locus with an unknown genotype that is inferred from trait phenotype (*single-point analysis*). The purpose is to find a group of markers that give low recombination fractions with the trait locus, so as to identify the genomic region most likely to contain the trait locus. The power of single-point linkage analysis can be increased by computing IBD sharing probabilities at a locus using information from multiple linked markers (*i.e.*, *multipoint*) and then correlate those genotypic similarities between relatives to some measure of trait similarities. In linkage analysis, a LOD (logarithm of the odds to the base 10) score of three or more (the odds are a thousand to one in favor of genetic linkage) is generally taken to indicate that two gene loci are close to each other on the chromosome.

The linkage analysis approaches have been successful in mapping hundreds of monogenic Mendelian traits [4], and they generally fall into two main classes: model-based or parametric methods, which requires a prior specification of a precise genetic model, *e.g.*, penetrance, disease-allele frequency, phenocopy and the mode of disease inheritance; and model-free or nonparametric methods that do not require a prior specification of a genetic model. Parametric linkage analysis is often used when researchers have a good understanding of the genetic model for the trait. For complex traits, for which the underlying genetic models are too complicated and less known to researchers, model-free or nonparametric methods are often used. For most nonparametric linkage methods, "sharing of genetic material" is measured as *identity by descent* (IBD) in families [3]. Two related people share an allele IBD if that allele was inherited from a common ancestor. Two alleles can be *identical by state* (IBS) without being IBD. A pair of siblings can share 0, 1, and 2 alleles IBD, and most other types of relatives (unilineal relatives) can share 0 or 1 allele IBD. Because linked loci tend to cosegregate, if two related people share alleles IBD at one genetic locus, there is a high probability that they share alleles IBD at a second closely linked locus. Nonparametric methods test whether IBD sharing at a locus is greater than expected under the null hypothesis of no linkage.

There are two types of nonparametric or model-free linkage analysis, one for both qualitative traits and the other for quantitative traits. In this dissertation, we are mainly interested in the latter, especially those methods for detecting quantitative trait loci (QTLs) that influence human quantitative traits, with a focus on the statistical issues arising in variance components (VC) linkage analysis approach [5],[6] (see next section). A very detailed review of other human QTL gene mapping methods can be found in Feingold (2001) [3] and in Szatkiewicz (2004) [7].

One thing worthy to be noted is that gene regions identified by linkage analysis is often large (typically 10 cM) [8], and can contains hundreds of other biologically plausible candidate genes. Further fine-mapping methods are often needed to narrow down the linkage region. Linkage analysis is powerful to detect the effects of rare variants, but less powerful for common variants [8].

(ii) Association analysis

Genetic association analysis, in which allele or genotype frequencies at markers are determined in affected individuals and compared with those of controls, provides an effective and powerful method to detecting the effects of common variants with modest effects and often yields fine-scale location [9]. Generally there are two types of association tests: family-based analysis and population-based analysis. Family-based association analysis compares the transmission of sequence variants from parents to affected offspring. Instead, the population-based association analysis compares frequencies of sequence variants between unrelated cases and controls. The most prominent method in family-based association studies is transmission disequilibrium test (TDT) [10] and its various extensions; the population association studies can be classified into several types: candidate polymorphism method, candidate gene method, fine mapping method and genome-wide method [11].

Recent advances in molecular genetic techniques have made it possible to genotype many individuals for increasingly dense genetics markers (SNPs). With the availability of high-resolution genetic maps, human population-based genome-wide association (GWA) studies involving hundreds of thousands of SNPs in thousands of cases and controls are underway [12], albeit with little consensus on optimal research design and analysis strategies. As is true for general case-control study design, confounding is a problem. For population-based association analysis, population admixture and stratification can generate spurious genotype-phenotye associations. So far solutions like genomic control [13],[14], and structured association methods [15],[16],[17] have evolved, another alternative is to do family-based association test instead but with the cost of lower power per genotype. With millions of independent tests in typical GWA studies, multiple testing problem arise frequently as in other biomedical and genomic research. Bonferroni correction (too conservative), permutation procedure (subject to null hypothesis of no association of genotype with phenotype), and false discovery rate (FDR) approach are among the usual frequentist approaches to multiple testing problem [11]. The Bayesian approach might a be a "remedy" but is complex because of a need to choose a prior probability for every disease model [11].

1.2.2 Issues in variance components linkage analysis

One of the most popular approaches to quantitative trait linkage analysis, variance components (VC) linkage analysis [5],[6] has many attractive features that other quantitative trait linkage methods lack. It requires few assumptions and can be easily extended to accommodate multiple gene effects, environmental effects, and their interactions [18]. Here we only review two most often used VC linkage analysis approaches: (1) univariate linkage analysis and (2) bivariate linkage analysis. The reason is that we will apply these two methods into our whole genome scan for obesity susceptibility loci in adults from Samoan archipelago (chapter 4).

(1) Univariate VC Linkage Analysis

Quantitative trait model

The idea of VC methods is to specify the genetic covariances between relatives within a pedigree as a function of IBD sharing at a marker locus, which is assumed to be closely linked to a trait locus [5]. Let Y_i be the trait value of the ith individual in the pedigree, the genetic mean model is

$$Y_i = \mu + g_i + G_i + \sum_{k=1}^{K} \beta_k c_{ik} + e_i \tag{1.1}$$

where μ is the grand mean, g_i is the unobserved random major gene effect at the trait locus, G_i is the unobserved random polygenic effect, c_{ik} is the value of kth covariate measure for ith individual, β_k is the coefficient of regression of Y_i onto covariate c_{ik}, and e_i is the random residual deviation. We assume two allelic variants, A and B, with population frequencies p and $q\,(=1-p)$, respectively at a given trait locus. The major gene effect g_i is

$$g_i = \begin{cases} a, & \text{if individual } i \text{ has genotype AA} \\ d, & \text{if individual } i \text{ has genotype AB} \\ -a, & \text{if individual } i \text{ has genotype BB} \end{cases}$$

The usual assumption is that g_i, G_i and e_i are uncorrelated random variables with expectation 0. So the expected trait value is

$$E(Y_i) = \mu + \sum_{k=1}^{K} \beta_k c_{ik}$$

The covariance between individuals i and j is

$$Cov(Y_i, Y_j) = \begin{cases} \sigma_q^2 + \sigma_d^2 + \sigma_G^2 + \sigma_e^2 & \text{if } i = j \\ 2\phi_{ij}(\sigma_q^2 + \sigma_G^2) + \Delta_{7ij}\sigma_d^2 & \text{if } i \neq j \end{cases}$$

where $\sigma_q^2 = 2pq(a + (2p - 1))^2$ is the additive genetic variance due to the major gene, $\sigma_d^2 = 4p^2q^2d^2$ is the dominant genetic variance due to the major gene, σ_G^2 is the polygenic variance, $\phi_{ij} = \frac{\Delta_{7ij}}{2} + \frac{\Delta_{8ij}}{2}$, the *kinship coefficient* [19] between individuals i and j, is the probability that an allele selected randomly from individual i and an allele selected randomly from the same autosomal locus of individual j are IBD. The Δ_{kij}'s (k = 1, ..., 9) are the probabilities for the nine possible condensed IBD status between two individuals i and j as defined by Jacquard (1974) [20]. Also, $2\phi_{ij}$ is the expected coefficient of relationship between individuals i and j, which is the expected proportion of allele IBD for individuals (i, j) at this locus. Δ_{7ij}, the *fraternity coefficient*, is the probability that individuals i and j share 2 alleles IBD at an autosomal locus.

For a pedigree of n members, the covariance matrix has the form

$$\Sigma = 2\mathbf{\Phi}\sigma_q^2 + \boldsymbol{\Delta_7}\sigma_d^2 + \boldsymbol{I_n}\sigma_e^2 \tag{1.2}$$

where $\mathbf{\Phi}$ is the $n \times n$ kinship matrix for the pedigree, Δ_7 is the $n \times n$ matrix containing fraternity coefficient for the pedigree members, $\boldsymbol{I_n}$ is the $n \times n$ identity matrix. For linkage analysis, an additional QTL-specific component is introduced [5],[6]. If n QTLs and an unknown number of residual *polygenes* influence a trait, assuming that dominance variance is negligible (only individuals who share 2 alleles IBD contribute the dominant variance), the expected covariance is then

$$\Sigma = \hat{\mathbf{\Pi}}\sigma_q^2 + \mathbf{2\Phi}\sigma_G^2 + \boldsymbol{I_n}\sigma_e^2 \tag{1.3}$$

where $\hat{\mathbf{\Pi}}$ is a matrix with elements ($\hat{\boldsymbol{\pi}}_{\boldsymbol{ij}}$) providing the estimated allele-sharing proportions at a marker that is putatively linked to the QTL; σ_q^2 is the additive genetic variance due to the major locus; σ_G^2 now represents the residual additive genetic variance not explained by the QTL.

Parameter estimation

Assuming multivariate normality distribution for phenotypic trait Y within pedigrees, the log likelihood of a pedigree of n individuals is given by

$$lnL(\mu, \sigma_q^2, \sigma_G^2, \sigma_e^2, \beta|Y) = -\frac{n}{2}ln(2\pi) - -\frac{1}{2}ln|\Sigma| - -\frac{1}{2}(Y - E(Y))^T\Sigma^{-1}(Y - E(Y)) \quad (1.4)$$

A likelihood-ratio test for linkage is carried out by testing whether the additive genetic variance due to the QTL (σ_q^2) was significantly different from 0 by comparing the likelihood of the general model, in which genetic variance due to the QTL σ_q^2 is estimated, with that of the restricted model, in which σ_q^2 is constrained to 0. Twice the difference of the log-likelihoods of these two models yields a test statistic that is asymptotically distributed as a 1/2:1/2 mixture of a χ_1^2 and a unit point mass at the origin (χ_0^2) [21]. The classical LOD scores are obtained by converting the statistic into values of log to the base 10.

(2) Bivariate VC Linkage Analysis

Multiple traits that are correlated can add information to each other. Joint use of data from multiple traits can increase power to detect QTLs, and make it possible to test the genetic correlations between two traits [22]. Here we briefly review the genetic models used in bivariate quantitative trait linkage analyses [22],[23].

Quantitative trait model

Let X, Y be two pedigree quantitative trait vectors, for which the genetic trait model discussed in univariate linkage analysis applies. For simplicity, we assume both X and Y are measured for each pedigree member (total n), which can be relaxed by appropriate evaluations of the pedigree likelihood [23]. Assume the composite phenotype $Z = [X, Y]^T$, the analogous covariance matrix for Z is

$$\Sigma_z = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$$

where Σ_x and Σ_y are as in equation 1.3, and the matrix $\Sigma_{xy} = \Sigma_{yx}$ of cross-covariances is given by

$$\Sigma_{xy} = \hat{\mathbf{\Pi}}\sigma_{qxy}^2 + \mathbf{2\Phi}\sigma_{Gxy}^2 + \boldsymbol{I_n}\sigma_{exy}^2 \quad (1.5)$$

As shown in Williams et al. (1999) [23], the covariance matrix Z is

$$\Sigma_Z = \hat{\mathbf{\Pi}} \otimes \boldsymbol{Q} + \mathbf{2\Phi} \otimes \boldsymbol{G} + \boldsymbol{I_n} \otimes \boldsymbol{E} \tag{1.6}$$

where $\otimes$ is the Kronecker product operator (Searle 1971 [24]), $\boldsymbol{Q}$, $\boldsymbol{G}$, and $\boldsymbol{E}$ are the QTL, polygenic and environmental covariance matrix, respectively. In a bivariate analysis, $\boldsymbol{Q}$, $\boldsymbol{G}$, and $\boldsymbol{E}$ have the form

$$\begin{pmatrix} \sigma^2_{\theta x} & \sigma_{\theta x}\sigma_{\theta y}\rho_\theta \\ \sigma_{\theta x}\sigma_{\theta y}\rho_\theta & \sigma^2_{\theta y} \end{pmatrix}$$

where ρ_θ is the correlation between X and Y due to effect θ (q, G, and e).

Parameter estimation

Assume trait vector Z follows a $2n$-variate normal distribution with mean E(Z) and variance Σ_Z [23], the log-likelihood for the data is

$$lnL_Z = -\frac{n}{2}ln(2\pi) - -\frac{1}{2}ln|\Sigma_Z| - \frac{1}{2}(Z - E(Z))^T\Sigma_Z^{-1}(Z - E(Z)) \tag{1.7}$$

where lnL_Z is

$$lnL(\mu_x, \mu_y, \sigma^2_{qx}, \sigma^2_{qy}, \sigma^2_{Gx}, \sigma^2_{Gx}, \sigma^2_{ex}, \sigma^2_{ey}, \rho_q, \rho_G, \rho_e, \beta_x, \beta_y | X, Y)$$

Two trait-specific estimates of the mean, σ^2_q (major gene effects), σ^2_G (residual additive genetic effects), and σ^2_e (random environmental effects) as well as three associated correlations ρ_q (correlation caused by a major gene), ρ_G (correlation caused by residual additive genetic effects), ρ_e (correlation caused by random environmental effects) are estimated using maximum likelihood method.

As in univariate linkage analysis, the hypothesis of no linkage for either trait (*i.e.*, $\sigma^2_{q_x} = \sigma^2_{q_y} = 0$) is tested using likelihood-ratio tests, in which the log-likelihood of the restricted model was compared with that of the model in which σ^2_q was estimated for the two traits. However, the bivariate LOD score obtained this way is on a different scale than the usual univariate LOD score, as the bivariate LOD asymptotically has two degrees of freedom [25]. An asymptotic P-value for the bivariate LOD score can be calculated using a $1/4\chi^2_2{:}1/2\chi^2_1{:}1/4\chi^2_0$ mixture distribution [21],[22]. Using this P-value, it is then possible to derive a univariate equivalent LOD score, LOD_{eq}, which has the same P-value as the bivariate LOD score.

1.3 MOTIVATION AND CONTRIBUTION OF OUR PROPOSED METHODS

1.3.1 Ordered genotypes: an extended ITO method and a general formula for genetic covariance

Genomic imprinting is a phenomenon where the functional activity of the two copies of each gene is not equivalent, and depends on whether they have been inherited maternally or paternally. Genomic imprinting occurs when both maternal and paternal alleles are present, but while one is expressed, the other is inactive, *e.g.* maternal imprinting means the maternally inherited mutant allele is not expressed (not 100% off). A gene is imprinted if its allele-specific expression depends on its parental origin. Several dozen mammalian genes are affected by genomic imprinting, and those affected genes express from only one of the two parental chromosomes [26].

More accurate modeling of underlying biological processes should lead to more accurate and more powerful inference. Extending the standard quantitative trait locus (QTL) model to a model incorporating parent-of-origin effects or parental imprinting can result in a more powerful test for linkage[27],[28]. Hanson et al. (2001)[27] also reported that the incorporation of parent-of-origin effects within linkage analysis of quantitative traits would facilitate the genetic dissection of complex traits, and substantially increase the power to detect an imprinted locus as well. Methods for detection of linkage and imprinting have been developed in sibship data [27],[28],[29], and in extended pedigrees [30]. Shete and Amos (2002) [28] reported that imprinting model became more powerful when the imprinting was moderate to large compared with usual variance-components model [31]. They also recommended testing for imprinting only if significant evidence for linkage is observed. However, despite the increasingly availability of a few literature talking about parent-of-origin linkage analysis [27],[28],[29],[30],[32], there is not a general uniform relative-to-relative (including *inbred*) covariance equation that is necessary for generalizing the analysis to any pedigree.

We have extended the traditional "ITO" transition method (Li and Sacks 1954 [33]) to derive the covariance components equations both for sibs and parent-offspring (incorporating

imprinting) [34], and then including them in the variance components linkage analysis which is a new test of linkage and imprinting in nuclear families for quantitative traits, which has increased power than sibship methods (data not published). In chapter 2, I continue to derive some generalized linear functions of those transition matrices for both unilineal and bilineal relatives under genomic imprinting in extended pedigrees, which can be easily applied to calculate covariance between relatives in those pedigrees. Meanwhile, the current extension expands the application of ITO method to more contemporary genetic models such as those involving imprinting, *e.g.*, computing genetic recurrence risks to any relative of the proband. Although historically attractive, the ITO method we generalized is limited to biallelic loci and to unilineal and bilineal relatives. In order to tackle these limitations, in the Chapter 2 we introduce a more general way for calculating the covariance for any relative pair under genetic imprinting [35]. This is a generalization of the approach utilized by Gillois (1964) [36], which was more recently summarized by Lange (2002) [37].

1.3.2 A recursive algorithm for computing generalized kinship coefficients: an ordered genotype version

The derived general equation for the covariance needs the computation of the detailed identity state coefficients (defined in details in chapter 2). In chapter 3, we modify the definition of generalized kinship coefficients to whether the sampled genes are maternal or paternal and propose a recursive algorithm for computing them (we call it *parental generalized kinship coefficients*) in general. **Our ultimate goal** is to write a recursive software implementing the algorithm that opens up the opportunity of calculating covariance between even inbred relatives.

1.3.3 Genome-wide scan for adiposity-related phenotypes in adults from Samoan archipelago

Overweight and obesity have reached epidemic proportions on a global scale. Obesity, BMI $\geq$ 30 kg/m^2, in the U.S. has become a major and serious public health problem, affecting 33% of adults in 2002. Obesity increases risks for serious diet-related diseases, such as CVD,

type-2 diabetes, and certain forms of cancers. Obesity is a typical multi-factorial disease with overwhelming evidence of genetic component effects, yet their roles in obesity are largely unknown.

In chapter 4, I discuss linkage evidences of susceptibility loci for adiposity-related phenotypes in Samoan population, an extensive summary of our multicenter NIH funded study "Genome-scan for Obesity Susceptibility Loci in Samoans". Univariate and bivariate VC linkage analysis approaches are used to localize possible candidate genes that influence variation of those phenotypes independently or simultaneously (a phenomenon called *pleiotropy*). Our current research findings will help further understand the whole picture of genetics of obesity, which has great influences on early preventions and later interventions of human obesity, making it a fundamentally important contribution to public health.

1.3.4 Sex-specific linkage analysis: sex-averaged genetic maps vs. sex-specific genetic maps

Since multipoint linkage analysis usually assumes a known genetic map, a map misspecification might compromise the estimation and testing procedures in the linkage analysis [38], [39], [40]. In the chapter 5, we carry out a preliminary simulation study to investigate the bias of multipoint linkage analysis arising from map misspecification. We also perform linkage analyses of sex-specific subsets of Samoan data and report some sex-specific obesity susceptibility loci, which may be hard to detect in the regular genome scans that fail to model for sex-specific differences [41].

2.0 ORDERED GENOTYPES: AN EXTENDED ITO METHOD AND A GENERAL FORMULA FOR GENETIC COVARIANCE

This chapter has been published in American Journal of Human Genetics, volume 78, pages 10351045 (Dai and Weeks 2006 [35]). I have obtained the copyright permission from the Chicago Press. The content of the paper has been modified to fit the style of this dissertation. Parts of this chapter are extended from my master thesis (Dai 2004 [34]).

2.1 INTRODUCTION

The "ITO" paper (Li and Sacks 1954 [33]) provides an elegant algorithm for deriving joint genotype probabilities between pairs of relatives. With ITO method, given the genotype of an individual, it is possible to derive the conditional probability of the genotypes of any non-inbred relative of that individual, *i.e.*, $P(G_2|G_1)$, where G_i denotes the genotype of i^{th}person. The ITO method was extended to handle multiple alleles and generalized for inbred populations (Richardson 1964 [42]). The ITO method was generalized for multiple loci and was also extended to handle consanguinity (Campbell and Elston 1971 [43]).

Although the ITO method has been widely used to solve various problems in human genetics, it only uses unordered genotypes, *i.e.*, the genotypes are unordered in the sense that maternal and parental contributions are not distinguished. However, some new applications require the use of the ordered genotypes, *e.g.*, when one is modeling genomic imprinting, one must keep track of the parental (ordered) origin of alleles. Genomic imprinting occurs when the functional activity of a person's allele depends on whether it was inherited maternally or paternally. Strong genomic imprinting renders an imprinted locus effectively haploid and

thereby causes certain genetic diseases, including disorders affecting cell growth, development, and behavior (Reik and Walter 2001 [26]). From the point of view of quantitative genetics, the effect of genomic imprinting is to make the phenotypical values of reciprocal heterozygotes different, which means various basic values of genetic quantities, as well as correlations, are not the same as the standard values. This difference may be crucial, especially in human quantitative genetics.

While Campbell and Elston [43] did attempt to extend the ITO method to handle ordered genotypes, their extension is flawed due to an incorrect assumption that does not generalize (as we explain in detail below). Li (1998 [44]) revised the 4×4 Li-Sacks matrices [33] to 2×2 matrices by focusing on allele IBD instead of genotype IBD. However, Li [44] still did not consider ordered genotypes.

More accurate modeling of underlying biological processes should lead to more accurate and more powerful inferences. In this report, we extended the ITO method to handle ordered genotypes. We derived some generalized linear functions of the transition matrices for deriving the probabilities of an individual's genotype, conditional on a relative's genotype. In the application part of this paper, our extended method is applied to calculate the covariance between both unilineal and bilineal relatives under imprinting. While the ITO approach is pleasing in terms of its clarity and understandability, it is difficult to extend it to handle loci with multiple alleles, as well as to handle very complex inbred relative pairs. Therefore, we also derive a completely general formula for the genetic covariance using ordered genotypes for any type of relative pair; this uses the approach of Gillois (1964 [36]) as more recently elucidated by Lange (2002 [19]). The resulting covariance equations can be easily applied in a variance component-based linkage analysis that takes genomic imprinting into account (not shown here).

2.2 METHODS

Here we notationally distinguish ordered and unordered genotypes. Accordingly, we let A/a with a slanted slash represent an *unordered* genotype, and use a vertical bar to denote

ordered genotypes (*e.g.*, $a|A$ and $A|a$). Here, without loss of generality, the maternal allele is listed to the left of the vertical bar and the paternal allele is listed to the right.

2.2.1 The original ITO method

Two diploid outbred related individuals may share (i) both genes identical by descent (IBD), (ii) one gene IBD, or (iii) no genes IBD. If we denote transition matrices as matrices of conditional probabilities, then the three basic transition matrices corresponding to the number of identical genes shared in common by the two relatives are respectively: [33]

$$I = \begin{array}{c} \\ G_1\ A/A \\ A/A \\ A/a \\ a/a \end{array} \begin{array}{c} \begin{array}{ccc} G_2 & & \end{array} \\ \begin{array}{ccc} A/a & a/a \end{array} \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$T = \begin{array}{c} \\ G_1\ A/A \\ A/A \\ A/a \\ a/a \end{array} \begin{array}{c} \begin{array}{ccc} G_2 & & \end{array} \\ \begin{array}{ccc} A/a & a/a \end{array} \end{array} \begin{bmatrix} p & q & 0 \\ \frac{p}{2} & \frac{1}{2} & \frac{q}{2} \\ 0 & p & q \end{bmatrix}$$

$$O = \begin{array}{c} \\ G_1\ A/A \\ A/A \\ A/a \\ a/a \end{array} \begin{array}{c} \begin{array}{ccc} G_2 & & \end{array} \\ \begin{array}{ccc} A/a & a/a \end{array} \end{array} \begin{bmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{bmatrix}$$

Following convention, in these three matrices, p and q represent the allele frequencies of A and a in the population, respectively, with $p + q = 1$. The first matrix, I, $P(G_2|G_1, \text{share 2 IBD})$, gives the genotype transition probabilities for two relatives when they share two alleles IBD with person 1's genotype given. In such a case, their genotypes are necessarily identical. When one is given to be A/A, the other must be the A/A, etc. The second matrix, T, $P(G_2|G_1, \text{share 1 IBD})$, gives the transition probabilities from one relative to the other when they share one gene IBD. Suppose the given relative is of genotype A/a (second row of T), the other relative must share allele A or allele a in common with probability 0.5 plus a random allele from the population. The third matrix, O, $P(G_2|G_1, \text{share 0 IBD})$, gives the conditional probabilities when the two individuals do not share any alleles IBD in common. Hence, they are genetically unrelated individuals. Regardless of the genotype of one individual, the probabilities of the other individual having the genotypes (A/A, A/a, a/a) remain simply p^2, $2pq$, and q^2, respectively, under Hardy-Weinberg equilibrium.

With these three basic matrices, it is then straightforward to find the joint distribution and correlation between any pair of relatives (unordered genotypes) in a random-mating population [33]. For autosomal genes, the general expression of the transition matrix for a

specific pair of relatives is given as

$$R = c_I I + c_T T + c_o O \tag{2.1}$$

where c_I, c_T, and c_o are the probabilities that the two specified relatives share both, one, and no genes IBD, respectively, with $c_I + c_T + c_o = 1$.

2.2.2 The ordered genotype ITO method

Since the original ITO matrices were derived for unordered genotypes, they are not useful when one is modeling imprinting, since one must then keep track of parental origin. Thus, in order to take genomic imprinting into consideration, we must track where the IBD gene comes from by using ordered genotypes (where we list the maternal allele first). Since the heterozygote $A|a$ may have a different genotypic value from $a|A$, Campbell and Elston [43] introduced four basic transition matrices with dimension 4×4 (instead of three 3×3 matrices), which are listed below.

$$I = \begin{array}{c} \\ G_1 \\ A\,|\,A \\ A\,|\,a \\ a\,|\,A \\ a\,|\,a \end{array} \begin{array}{c} \begin{array}{cccc} G_2 & & & \\ A\,|\,A & A\,|\,a & a\,|\,A & a\,|\,a \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \qquad S_m = \begin{array}{c} \\ G_1 \\ A\,|\,A \\ A\,|\,a \\ a\,|\,A \\ a\,|\,a \end{array} \begin{array}{c} \begin{array}{cccc} G_2 & & & \\ A\,|\,A & A\,|\,a & a\,|\,A & a\,|\,a \end{array} \\ \begin{bmatrix} p & 0 & q & 0 \\ 0 & p & 0 & q \\ p & 0 & q & 0 \\ 0 & p & 0 & q \end{bmatrix} \end{array}$$

$$S_f = \begin{array}{c} \\ G_1 \\ A\,|\,A \\ A\,|\,a \\ a\,|\,A \\ a\,|\,a \end{array} \begin{array}{c} \begin{array}{cccc} G_2 & & & \\ A\,|\,A & A\,|\,a & a\,|\,A & a\,|\,a \end{array} \\ \begin{bmatrix} p & q & 0 & 0 \\ p & q & 0 & 0 \\ 0 & 0 & p & q \\ 0 & 0 & p & q \end{bmatrix} \end{array} \qquad O = \begin{array}{c} \\ G_1 \\ A\,|\,A \\ A\,|\,a \\ a\,|\,A \\ a\,|\,a \end{array} \begin{array}{c} \begin{array}{cccc} G_2 & & & \\ A\,|\,A & A\,|\,a & a\,|\,A & a\,|\,a \end{array} \\ \begin{bmatrix} p^2 & pq & pq & q^2 \\ p^2 & pq & pq & q^2 \\ p^2 & pq & pq & q^2 \\ p^2 & pq & pq & q^2 \end{bmatrix} \end{array}$$

Each matrix element represents the probability of sibling 2 having the specific ordered genotype conditional on the genotype of sibling 1. The subscript m and f represent male and female. We use the same notion I and O as defined in 3×3 matrices discussed above. The matrix S_m, $P(G_2|G_1, \text{share 1 allele IBD through father})$, specifies the probabilities of sibling 2's genotypes conditional on sibling 1 sharing 1 allele IBD through father. Similarly, the

matrix S_f, $P(G_2|G_1, \text{share 1 allele IBD through mother})$, specifies the probabilities of sibling 2's genotypes conditional on sibling 1 sharing 1 allele IBD through mother.

Campbell and Elston [43] proposed that the transition matrix R for any specified pair of relatives could be derived by the formula $R = c_I I + \frac{c_T}{2} S_m + \frac{c_T}{2} S_f + c_o O$. However, their formula for R is incorrect for some pairs of relatives, as we explain in detail below. Furthermore, a more complete derivation of the ITO method for ordered genotypes requires the specification of two additional matrices, T_f and T_m, which concern parent-offspring transitions (Dai 2004 [34]). These matrices are:

$$T_m = \begin{array}{c} \\ G_{father} \\ A\,|\,A \\ A\,|\,a \\ \\ a\,|\,A \\ a\,|\,a \end{array} \begin{array}{c} \begin{array}{cccc} G_{offspring} & & & \\ A\,|\,A & A\,|\,a & a\,|\,A & a\,|\,a \end{array} \\ \begin{bmatrix} p & 0 & q & 0 \\ \frac{p}{2} & \frac{p}{2} & \frac{q}{2} & \frac{q}{2} \\ \frac{p}{2} & \frac{p}{2} & \frac{q}{2} & \frac{q}{2} \\ 0 & p & 0 & q \end{bmatrix} \end{array} \qquad T_f = \begin{array}{c} \\ G_{mother} \\ A\,|\,A \\ A\,|\,a \\ \\ a\,|\,A \\ a\,|\,a \end{array} \begin{array}{c} \begin{array}{cccc} G_{offspring} & & & \\ A\,|\,A & A\,|\,a & a\,|\,A & a\,|\,a \end{array} \\ \begin{bmatrix} p & q & 0 & 0 \\ \frac{p}{2} & \frac{q}{2} & \frac{p}{2} & \frac{q}{2} \\ \frac{p}{2} & \frac{q}{2} & \frac{p}{2} & \frac{q}{2} \\ 0 & 0 & p & q \end{bmatrix} \end{array}$$

The derivations of the matrices T_m, $P(G_{offspring}|G_{father})$ and T_f, $P(G_{offspring}|G_{mother})$ are straightforward. For example, if the genotype of the mother is $A|a$, the conditional probabilities that the offspring's genotypes are $A|A$, $A|a$, $a|A$, and $a|a$ are $\frac{p}{2}$, $\frac{q}{2}$, $\frac{p}{2}$, and $\frac{q}{2}$ respectively. The reason is that the alleles A and a from the mother each have a 50% chance of being transmitted to her offspring.

Li [44] showed how the 4×4 S_m and S_f matrices [43] can be derived as 'external tensors' of the 2×2 matrices. However, the matrices T_f and T_m can't be derived via outer products. We have shown that $T_f(T_m)$ can be computed as a weighted sum of the $S_f(S_m)$ matrix and a permuted version of $S_f(S_m)$ where the middle two rows are switched (details omitted).

2.3 TRANSITION MATRICES FOR UNILINEAL RELATIVES IN EXTENDED PEDIGREES

Now with the above six ordered genotype 4×4 transition matrices, we can derive the conditional probabilities for two specified outbred relatives, in which we track the origin of both alleles at a locus. We first consider some simple unilineal relatives. For parent and offspring pair, since they share one gene identical by descent, the transition probabilities from parent to offspring for unordered genotypes are the elements of matrix T [33]. However, in ordered genotype method, we have to consider the transition of paternal and maternal alleles separately. The elements of matrix T_m are now the transition probabilities from father to offspring, and the elements of matrix T_f are the transition probabilities from mother to offspring. Note that with ordered genotypes $P(G_{father}|G_{offspring})$ is not equal to $P(G_{offspring}|G_{father})$, while these two are equal with unordered genotypes.

Consider the transition probabilities from a maternal grandmother to a grandchild GC through a mother M. If the genotype of the maternal grandmother is $A|A$, the resulting genotype of her daughter M will be $A|?$ (?-some unknown allele), so the total conditional probability that the grandchild GC is $A|a$ with allele A inherited from the mother M is $(\frac{1}{2} + \frac{1}{2}p)q$. The reason is that there is a 50% chance that the grandchild GC receives the grandmaternal A allele from her mother M and a 50% chance that she receives her maternal grandfather's allele, which is A with probability p. So the conditional probability of the grandchild's ordered genotype being $A|a$ is then $(\frac{1}{2} + \frac{1}{2}p)q$, where q is the probability of grandchild GC randomly inheriting the second allele (allele a) from his/her father.

It is much easier to understand if we do the above derivation using matrix manipulation. The probabilities of the mother are the elements of first row of matrix T_f, since we are conditioning on the grandmother's $A|A$ genotype. For each genotype of the mother M, the elements of second column of T_f are the probabilities of the genotype of the grandchild GC being $A|a$. The total conditional probability for the grandchild to have genotype $A|a$ given grandmother's $A|A$ genotype via his/her mother is the product of the first row and second

column of the transition matrix T_f as shown below:

$$\begin{pmatrix} p & q & 0 & 0 \end{pmatrix} \begin{pmatrix} q \\ \frac{q}{2} \\ \frac{q}{2} \\ 0 \end{pmatrix} = pq + \frac{q^2}{2} = (\frac{1}{2} + \frac{1}{2}p)q$$

By the same algorithm, the conditional probabilities for a grandchild given a specific genotype for the maternal grandmother are given by the elements of the product matrix $T_f \cdot T_f = T_f^2$. In the same manner, the conditional probabilities of a grandchild's genotypes given a specific genotype for the maternal grandfather (via his/her mother) are given by the elements of the product matrix $T_m \cdot T_f$.

For the above grandmother-grandchild pair relation, Campbell and Elston's [43] formula (page 229)

$$R = \frac{S_m}{4} + \frac{S_f}{4} + \frac{O}{2}$$

gives

$$(\frac{1}{4} + \frac{1}{2}p)q$$

as the probability of the genotype of the grandchild being $A|a$ given the genotype of the grandmother is $A|A$, which is clearly wrong. For another example, for an aunt-niece pair connected through the mother, $c_T = \frac{1}{2}$, but c_T doesn't split in half as Campbell and Elston suggested, but rather all its "weight" goes on the T_f, and the correct matrix is

$$R = \frac{T_f}{2} + \frac{O}{2}.$$

Li and Sacks [33] showed that

$$T^2 = \frac{T}{2} + \frac{O}{2},$$

which means a grandparent-grandchild pair shares 1 gene IBD and 0 gene IBD with an equal 50% chance. T^2 also gives the conditional probabilities for half sibs. They also showed that in general,

$$T^{n+1} = (\frac{1}{2})^n T + (1 - (\frac{1}{2}^n))O \tag{2.2}$$

where $n+1$ is the total number of generations between the two relatives. When ordered genotypes are used, similar equations hold. For example:

$$\begin{cases} T_m^2 = \frac{1}{2}T_m + \frac{1}{2}O, & T_f^2 = \frac{1}{2}T_f + \frac{1}{2}O \\ T_m T_f = \frac{1}{2}T_f + \frac{1}{2}O, & T_f T_m = \frac{1}{2}T_m + \frac{1}{2}O \\ T_m T_f T_f = \frac{1}{4}T_f + \frac{3}{4}O, & T_m T_f T_m = \frac{1}{4}T_m + \frac{3}{4}O \\ T_f T_m T_m = \frac{1}{4}T_m + \frac{3}{4}O, & T_f T_m T_f = \frac{1}{4}T_f + \frac{3}{4}O \end{cases}$$

where T_m^2 gives the conditional probabilities for half sibs who have same father but different mothers, and T_f^2 gives the conditional probabilities for half sibs who have same mother but different fathers. And in general

$$\begin{cases} T_{m(f)}^{n+1} = (\frac{1}{2})^n T_{m(f)} + (1-(\frac{1}{2})^n)O \\ (T_{i_1}T_{i_2}\cdots T_{i_n})T_{m(f)} = (\frac{1}{2})^n T_{m(f)} + (1-(\frac{1}{2})^n)O,\ i_j \in \{m,f\}, j = 1,...,n \\ T_{m(f)}^{n+1} \to O,\ as\, n \to \infty \end{cases}$$

where $n+1$ is the number of generations between two relatives. When n is infinitely large, the conditional probabilities for two relatives are given by the elements of the matrix O, *i.e.*, the two relatives could be treated as two random samples from the general human population who are all unrelated to each other.

2.4 TRANSITION MATRICES FOR BILINEAL RELATIVES IN EXTENDED PEDIGREES

Now we model bilineal relatives. Let us first consider the simple but most important type: full sibs. Since full sibs have 25% chance of sharing two genes IBD, 25% chance of sharing 0 IBD, they therefore have 50% chance of sharing 1 gene IBD. However, we are dealing with

ordered genotypes, so the two sibs have 25% chance of sharing 1 maternal allele IBD and no paternal allele IBD and vice versa. Thus the transition matrix for full sibs is as follows:

$$\boldsymbol{\mathcal{S}} = \frac{1}{4}I + \frac{1}{4}S_m + \frac{1}{4}S_f + \frac{1}{4}O \tag{2.3}$$

Another pair of relatives that can share 2 genes IBD is double first cousins whose parents are members of two sibships. There are six types of sibships in the general population due to six different mating types [33]. Following the same algorithm mentioned in Fig.1 in Li and Sacks's paper [33] but labeling maternal and paternal alleles, we derive the conditional matrix for double first cousins, which is,

$$D = \boldsymbol{\mathcal{S}}^2 = \frac{1}{16}I + \frac{3}{16}S_m + \frac{3}{16}S_f + \frac{9}{16}O \tag{2.4}$$

Next we try to model the relationship for some unlineal relatives in which the conditional matrix $\boldsymbol{\mathcal{S}}$ for full sibs, and the matrices T_m and T_f are involved. We use the broad sense "avuncular" term, which includes uncle-nephew, uncle-niece, aunt-nephew and aunt-niece relations. The conditional probabilities of a nephew's genotypes are then given by the product of $\boldsymbol{\mathcal{S}}T_{m(f)}$ when the uncle's genotypes are conditioned on. Conversely, the conditional probabilities of the uncle's genotypes are given by elements of the product of $T_{m(f)}\boldsymbol{\mathcal{S}}$ conditional on the genotypes of the nephew. Whether the transition matrix T_m or T_f is involved depends on whether the father or mother of the nephew is the "connecting" relative. We verified that $\boldsymbol{\mathcal{S}}T_{m(f)} = T_{m(f)}\boldsymbol{\mathcal{S}}$, which indicates that uncle-nephew matrix is same as the nephew-uncle transition matrix [33]. Through further multiplication of matrices, we can also prove "the most remarkable property" that is

$$\boldsymbol{\mathcal{S}}T_{m(f)} = T_{m(f)}\boldsymbol{\mathcal{S}} = T^2_{m(f)} \tag{2.5}$$

which indicates that uncle-nephew relations are same as those for grandparent-grandchild or half sibs[33], whether the transition matrix T_m or T_f is involved depends on whether the uncle is the nephew's paternal uncle or the nephew's maternal uncle. Extension of the above equations results in other important matrices given by

$$T_{m(f)}\boldsymbol{\mathcal{S}}T_{m(f)} = T^3_{m(f)} = \frac{1}{4}T_{m(f)} + \frac{3}{4}O \tag{2.6}$$

whose elements give the conditional probabilities for first cousins, and also the probabilities for the great-grandchild conditional on one given great-grandparent.

2.5 APPLICATION FOR DERIVING COVARIANCES BETWEEN RELATIVES UNDER GENOMIC IMPRINTING

Next, as an illustration of the utility of our extended ordered genotype ITO method, we now derive equations for the genetic covariance between sibs, and for the covariance between parent-child taking genomic imprinting into account (Dai 2004 [34]). We begin with the standard genetic model and extend it to consider the case when a locus is subject to imprinting. To derive the covariance formulae, it is necessary first to define the quantitative trait locus (QTL) model and its variance components. Here, we briefly review results of Spencer [32]. Assume that an unobserved major gene has two alleles, allele A and allele a, with $P(A) = p$ And $P(a) = q$. In the standard genetic model, the genotypic value is a if the genotype is A/A, that of the a/a homozygote is $-a$, and the genotypic value for heterozygote A/a is denoted as d. However, under imprinting, different genotypic values are possible for the two possible heterozygotes: d_1 for $A|a$ and d_2 for $a|A$ (Figure 2.1). It is usually assumed that $a \geq d_1$ and $d_2 \geq -a$. We have $d_1 = a$ ($d_2 = -a$) when there is complete inactivation of the maternally (paternally) derived allele. A measure of imprinting is denoted as $I = (d_1 - d_2)/2$ [28].

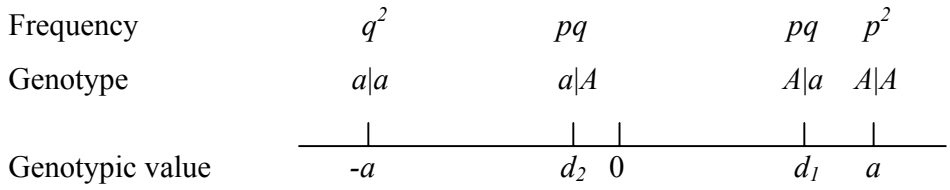


Figure 2.1: Genotypic values of the four possible genotypes

Spencer [32] derived many useful genetic components of variance under imprinting, which are summarized in Table 2.1. When $d_1 = d_2 = d$, *i.e.*, there is no imprinting ($I = 0$), the above various genetic values "revert" to their standard values. We further show that $\sigma^2_{Am} + \sigma^2_{Af} = 2\sigma^2_a$ and $\sigma^2_D + \sigma_{ADm} + \sigma_{ADf} = \sigma^2_d$ (no constraints on I), where σ^2_a and σ^2_d are the additive genetic variance and dominance genetic variance respectively under the standard genetic model with no imprinting (details not presented here).

Table 2.1: Values of genetic components of variance under genomic imprinting (after Spencer [32]).

NAME	EXPRESSION	DEFINITION
μ	$a(p-q) + (d_1 + d_2)pq$	The mean phenotype of a population in HWE
α_m	$a + d_2 q - d_1 p$	The average effect of a gene substituion for males
α_f	$a + d_1 q - d_2 p$	The average effect of a gene substituion for females
$\sigma^2_{A_m}$	$2pq\alpha^2_m$	The additive genetic variance for males
$\sigma^2_{A_f}$	$2pq\alpha^2_f$	The additive genetic variance for females
σ^2_D	$pq(pq(d_1+d_2)^2 + (d_1-d_2)^2)$	The dominance genetic variance
$\sigma^2_{D_m}$	$pq(pq(d_1+d_2)^2 + (d_1-d_2)^2)$	The dominance genetic variance for males
$\sigma^2_{D_f}$	$pq(pq(d_1+d_2)^2 + (d_1-d_2)^2)$	The dominance genetic variance for females
σ^2_G	$pq(2\alpha_m\alpha_f + pq(d_1 + d_2)^2 + (d_1 - d_2)^2)$	The "overall" genetic variance
σ_{AD_m}	$pq\alpha_m(d_1 - d_2)$	The covariance between dominance deviation and breeding value for males
σ_{AD_f}	$pq\alpha_f(d_2 - d_1)$	The covariance between dominance deviation and breeding value for females

2.5.1 Covariance between sibs under genomic imprinting

With the above definitions of different genetic variance components (Table 2.1), we now begin to derive the covariance between sibs and between parent-offspring under genomic imprinting using our transition matrices. Spencer [32] has derived the covariance between parent and offspring under genomic imprinting. However, he did not derive the covariance between a pair of full sibs. As an illustration of the utility of our ordered ITO method, we first derive the covariance between sib pairs and also include a short part on deriving covariance between parent and offspring to verify that our results match Spencer's results.

Let k_0 denote the probability of sib 1 and sib 2 sharing 0 allele IBD, k_{1m} denote the probability of sib 1 and sib 2 sharing 1 paternal allele IBD, k_{1f} denote the probability of sib 1 sharing 1 maternal allele IBD with sib 2, and k_2 denote the probability of sib 1 and sib 2 sharing 2 alleles IBD. The probability of sib 2 having a particular genotype (G_2) given the genotype of sib 1 (G_1) can be calculated as follows:

$$\begin{aligned} P(G_2|G_1) &= \sum_{i=0}^{2} P(G_2|G_1, \text{share } i \text{ alleles IBD}) P(\text{share } i \text{ alleles IBD}) \\ &= k_2 I + k_{1m} S_m + k_{1f} S_f + k_0 O \end{aligned} \tag{2.7}$$

which, for full sibs, is the same as matrix $\boldsymbol{\mathcal{S}}$ (in equation (2.3) above).

Given the above conditional matrix, the joint probability of sib 2 having a certain genotype s_2 and sib 1 having another certain genotype of s_1, $P(s_1, s_2)$ can be derived by multiplying the specific element of the above matrix $P(G_2|G_1)$ by the probability of having the certain genotype for sib1. The genetic covariance between a pair of sibs can be derived in the following equation (2.8).

$$\begin{aligned} Cov(s_1, s_2) &= E(s_1 s_2) - E(s_1)E(s_2) = \sum_{s_1}\sum_{s_2} s_1 s_2 P(s_1, s_2) - \mu \cdot \mu \\ &= \frac{(k_{1m} + k_2)}{2}\sigma^2_{Am} + \frac{(k_{1f} + k_2)}{2}\sigma^2_{Af} + k_2(\sigma^2_D + \sigma_{ADm} + \sigma_{ADf}) \end{aligned} \tag{2.8}$$

For full sibs, $k_0 = 1/4$, $k_{1m} = 1/4$, $k_{1f} = 1/4$, and $k_2 = 1/4$, we get the following model from equation (2.8):

$$Cov(s_1, s_2) = \frac{1}{4}\sigma^2_{Am} + \frac{1}{4}\sigma^2_{Af} + \frac{1}{4}(\sigma^2_D + \sigma_{ADm} + \sigma_{ADf}) \tag{2.9}$$

As stated earlier, $\sigma^2_{Am} + \sigma^2_{Af} = 2\sigma^2_a$ and $\sigma^2_D + \sigma_{ADm} + \sigma_{ADf} = \sigma^2_d$ (σ^2_a and σ^2_d are additive genetic variance and dominance genetic variance respectively under the standard genetic model). Thus, equation (2.9) simplifies to be exactly the standard genetic model defined as

$$Cov(s_1, s_2) = \frac{1}{2}\sigma^2_a + \frac{1}{4}\sigma^2_d \tag{2.10}$$

For half sibs, we distinguish between half sibs who share a mother ($k_0 = 1/2$, $k_{1f} = 1/2$, and $k_2 = 0$) and half sibs who share a father ($k_0 = 1/2$, $k_{1m} = 1/2$, and $k_2 = 0$). Equation (2.8) gives, respectively:

$$\begin{aligned} Cov(s_1, s_2) &= \frac{1}{4}\sigma^2_{Af} \\ Cov(s_1, s_2) &= \frac{1}{4}\sigma^2_{Am} \end{aligned} \tag{2.11}$$

2.5.2 Covariance between parent-offspring under genomic imprinting

In the next part, we derive equations for the genetic covariance for a parent-offspring pair, as follows. Let o and p_f denote the genotypic values of offspring and mother, respectively. The joint probability for the offspring having a certain genotype and the mother having another genotype, $P(o, p_f)$ can be derived by multiplying the specific element of the matrix T_f ($P(G_{offspring}|G_{mother})$) by the probability of having that certain genotype for the mother. Then, the covariance σ_{op_f} is calculated as:

$$\begin{aligned} Cov(o, p_f) &= E(op_f) - E(o)E(p_f) = \sum_o \sum_{p_f} op_f P(o, p_f) - \mu \cdot \mu \\ &= \frac{1}{2}\sigma^2_{Af} + \frac{1}{2}\sigma_{ADf} \end{aligned} \tag{2.12}$$

Similarly, we can derive the covariance equation for father and offspring:

$$Cov(o, p_m) = \frac{1}{2}\sigma^2_{Am} + \frac{1}{2}\sigma_{ADm} \tag{2.13}$$

Equations (2.12, 2.13) were also derived by Spence (2002) [32].

2.6 A GENERAL FORMULA FOR GENETIC COVARIANCE

Although historically attractive, the "ITO" method we generalized here is limited to biallelic loci and to unilineal and bilineal relatives. In order to tackle these limitations, we now introduce a more general way for calculating the covariance for any relative pair under imprinting. This is a generalization of the approach utilized by Gillois [36], which was more recently summarized by Lange [19], whose notation we use here.

First, suppose there are two or more alleles, let the k^{th} allele have population frequency p_k and the ordered genotype $k|l$ have the trait value $w_{k|l}$, then we can write

$$w_{k|l} \quad = \quad \alpha_k + \beta_l + \delta_{k|l} \tag{2.14}$$

where α_k is the additive impact of the maternal allele, β_l is the additive impact of the paternal allele, and $\delta_{k|l}$ is the residual departure from additivity. Under imprinting, the identity $w_{k|l} = w_{l|k}$ does not necessarily hold. No generality is lost if we adjust the trait mean to be zero, so that $\sum_k \sum_l w_{k|l} p_k p_l = 0$. The allelic contributions α_k, β_l are chosen to minimize the deviations $\delta_{k|l} = w_{k|l} - \alpha_k - \beta_l$. One way of doing this is to minimize the sum of squares

$$\sum_k \sum_l \delta^2_{k|l} p_k p_l \quad = \quad \sum_k \sum_l (w_{k|l} - \alpha_k - \beta_l)^2 p_k p_l \tag{2.15}$$

which is achieved by taking $\alpha_k = \sum_l w_{k|l} p_l$, and $\beta_l = \sum_k w_{k|l} p_k$ (See Appendix).

Next suppose individuals i and j are relatives. The covariance Cov(X_i,X_j) between the trait values X_i and X_j of i and j can be computed in the following steps. Here we use the fifteen detailed identity states of Gillois [36], Harris [45], and Jacquard [20]. Figure 2.2 (after Sobel et al. [46], Jacquard [47]) shows the 15 detailed identity states possible when maternal and paternal alleles are distinguished. The states vary from sharing no alleles IBD, S_{15}, to sharing all four alleles IBD, S_1.

Conditioning on these detailed identity states of the two relatives, and using the identities $\sum_k \alpha_k p_k = 0$, $\sum_l \beta_l p_l = 0$, $\sum_k \delta_{k|l} p_k = 0$, and $\sum_l \delta_{k|l} p_l = 0$, we can deduce

$$\begin{aligned} E(X_i, X_j) &= (\delta_1 + \delta_2 + \delta_4 + \delta_9 + \delta_{10}) \sum_k \alpha_k^2 p_k \\ &\quad + (\delta_1 + \delta_3 + \delta_5 + \delta_9 + \delta_{11}) \sum_k \beta_k^2 p_k \\ &\quad + (2\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5 + 2\delta_{12} + \delta_{13} + \delta_{14}) \sum_k \alpha_k \beta_k p_k \\ &\quad + (2\delta_1 + \delta_2 + \delta_4) \sum_k \alpha_k \delta_{k|k} p_k \\ &\quad + (2\delta_1 + \delta_3 + \delta_5) \sum_k \beta_k \delta_{k|k} p_k \\ &\quad + \delta_1 \sum_k \delta_{k|k}^2 p_k \\ &\quad + \delta_6 \sum_k \sum_l \delta_{k|k} \delta_{l|l} p_k p_l \\ &\quad + \delta_9 \sum_k \sum_l \delta_{k|l}^2 p_k p_l \\ &\quad + \delta_{12} \sum_k \sum_l \delta_{k|l} \delta_{l|k} p_k p_l \end{aligned} \tag{2.16}$$

where δ_i is the probability of the i^{th} detailed identity state [20]. A detailed derivation of equation (2.16) is given in Appendix. When there is no imprinting, equation (2.16) reduces to the general covariance equation derived by Gillois [36].

Since trait means $\mathrm{E}(X_i) = \mathrm{E}(X_j) = 0$, the covariance $\mathrm{Cov}(X_i,X_j) = \mathrm{E}(X_i,X_j)$, which is given in equation (2.16). If we assume that neither i nor j is inbred, we have $\delta_1 = \delta_2 = \cdots = \delta_8 = 0$. The covariance $\mathrm{Cov}(X_i,X_j)$ then simplifies to

$$\begin{aligned} Cov(X_i, X_j) &= (\delta_9 + \delta_{10}) \sum_k \alpha_k^2 p_k + (\delta_9 + \delta_{11}) \sum_k \beta_k^2 p_k \\ &\quad + (2\delta_{12} + \delta_{13} + \delta_{14}) \sum_k \alpha_k \beta_k p_k \\ &\quad + \delta_9 \sum_k \sum_l \delta_{k|l}^2 p_k p_l + \delta_{12} \sum_k \sum_l \delta_{k|l} \delta_{l|k} p_k p_l \end{aligned} \tag{2.17}$$

When there are only two alleles, we can rewrite the summations in terms of our notation used above:

$$\begin{cases} \sum_k \alpha_k^2 p_k = \frac{1}{2}\sigma_{Af}^2 \\ \sum_k \beta_k^2 p_k = \frac{1}{2}\sigma_{Am}^2 \\ \sum_k \alpha_k \beta_k p_k = \frac{1}{2}\sigma_{Am}^2 + \sigma_{ADm} = \frac{1}{2}\sigma_{Af}^2 + \sigma_{ADf} \\ \sum_k \sum_l \delta_{k|l}^2 p_k p_l = \sum_k \sum_l \delta_{k|l}\delta_{l|k} p_k p_l = \sigma_d^2 \end{cases}$$

Thus, from our general equation (2.17), we obtain the same covariances as we derived above for full sibs (equation (2.9)), half sibs (equation (2.11)), mother-offspring pairs (equation (2.12)), and father-offspring pairs (equation (2.13)). Furthermore, our equation (2.16) can be used to generalize the variance-component model developed by Shete et al. (2003) [30] to handle all possible types of inbred relative pairs.

Once the detailed identity coefficients [20] are computed, any relative-to-relative covariance is expressible in terms of the theoretical variances and covariances defined above. An algorithm for computing these detailed identity coefficients (assuming the entire pedigree structure connecting the two individuals is known) was derived by Nadot and Vaysseix [48].

2.7 DISCUSSION

In summary, in this chapter we extended the "ITO" method ([33],[43]), to handle ordered genotypes in an attempt to generalize this simple but useful method. We also showed that Campbell and Elston's previous formula [43] for the transition matrix R is incorrect for some pairs of relatives. In practice, a more complete derivation of the ITO method for ordered genotypes requires the specification of two additional matrices T_m and T_f, which we derived in this paper. By tracking the paternal or maternal origin of each allele, we now have six basic transition matrices, with the help of which it is possible derive conditional probabilities between two specified outbred relatives when we need to distinguish the two forms of the heterozygotes. In addition to providing an algorithm for deriving conditional probabilities

using ordered genotypes, the ITO approach can be used to derive formulae for the genetic covariance between a pair of relatives. To complement the more limited ITO approach, we also derived a completely general formula for the genetic covariance using ordered genotypes; this formula is applicable to multi-allelic loci and to any type of inbred relative pair.

We illustrated the utility of the extended ITO approach and our general covariance formula by using them to derive the genetic covariance under imprinting between parent-offspring and sib pairs. The derived formulas for the covariance between parents and offspring's genotypic values are the same as those given in previous work [32]. The derived formulas for the covariance between sibs are also independently derived by Santure and Spencer (2006) [49] later after we published our paper. The consistency of our equations with previous/very recent work proves the applicability of our proposed method for the calculation of covariance between two relatives when we have to deal with ordered genotypes, *e.g.*, when we try to model genomic imprinting in human quantitative genetic analysis. Also it should be noted that, our work could help accurately test genetic hypotheses or predict risk for genetic counselling given a known genetic model [43]. Our extended ordered genotype "ITO" method and our general covariance formula, with their easy applicability, will be helpful in modeling the complex relationship between relatives under the important biological phenomena (genomic imprinting) that need further statistical attention.

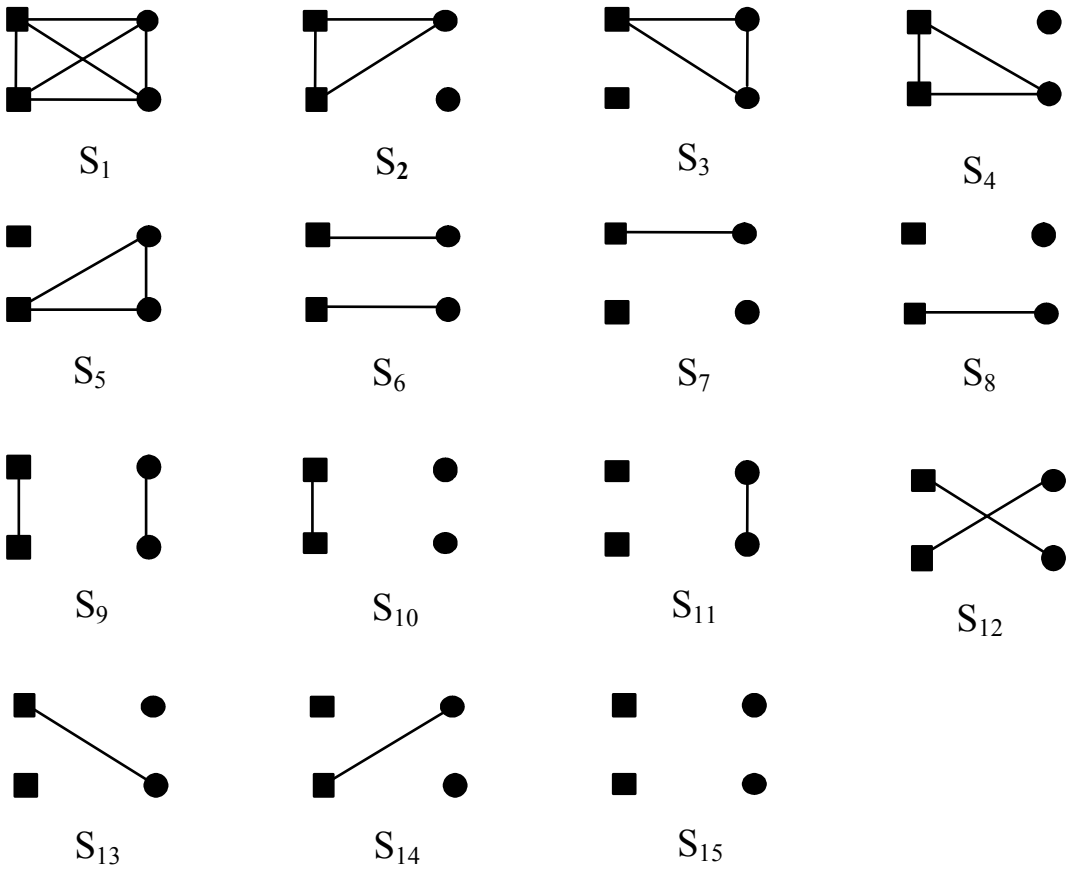


Figure 2.2: The 15 possible detailed identity states for individuals i and j with ordered genotypes. The squares are maternal alleles and the circles are paternal alleles. The above two symbols are i's alleles and the under two symbols are j's alleles. Lines connect alleles that are IBD (after Sobel et al. [46], Jacquard [47]).

3.0 A RECURSIVE ALGORITHM FOR COMPUTING GENERALIZED KINSHIP COEFFICIENTS: AN ORDERED GENOTYPE VERSION

In this chapter we first introduce a different way for computing the 15 detailed identity coefficients discussed in chapter 2 (Figure 2.1) by computing some modified generalized kinship coefficients [19], [50],[51]. Then we derive a recursive algorithm for computing those modified generalized kinship coefficient for two individuals with ordered genotypes in the same pedigree. As an example, we apply this algorithm to compute one specific modified generalized kinship coefficient in the pedigree discribed in Figure 3.1. Although the current algorithm works for our example, it is incomplete and we are still trying to improve it. Once the final algorithm is ready, recursive software implementing the algorithm will be prepared based on the X-linked version of Affected-Pedigree-Member (APM) program for linkage analysis [52].

3.1 INTRODUCTION

In chapter 2, we mentioned that an algorithm for computing the detailed identity coefficients (under the assumption that the entire pedigree struture connecting the two individuals is known) was already derived by Nadot and Vaysseix [48]. The algorithm was implemented in a program written in computer languages ALGOL and ASSEMBLER [48]. Although the program has been existed for a long time, it is not easy to understand by normal readers, and is not ready for general usage because an executable code is not available anymore.

Here we introduce a different way for computing those detailed identity coefficients, which involves modifying the definition of 'generalized kinship coefficients' [19] to include whether

the sampled genes are maternal or paternal (we call them ***parental generalized kinship coefficients***), and developing an algorithm for recursively computing those modified kinship coefficients.

3.2 METHODS

To describe the algorithm, we first define some necessary terms to be mentioned later (from Weeks et al. (1995) [52]). A 'founder' represents a person in a pedigree who has no parents; all founders are assumed to be noninbred and unrelated. For example, persons 1 and 2 in Figure 3.1 are founders. "A 'block' represents an equivalence class of the same alleles, where a gene (allele) picked at random from a block is identical by descent (IBD) to any other gene (allele) picked at random from the same block; however, it is not IBD to a gene (allele) picked at random from a different block" (page 27, [52]). Two genes (alleles) picked from two different blocks are not IBD, but may be identical by state (IBS). In general, as shown in chapter 5 of the book [19], with a list of four individuals i, j, k, l, if we randomly pick one gene from each individuals, there are 15 partitions of the four sampled genes G_i, G_j, G_k, G_l, which ranges from the block $[G_i, G_j, G_k, G_l]$, where all of the genes are IBD, to the four blocks $[(G_i), (G_j), (G_k), (G_l)]$, where no genes are IBD (for a list of two individuals i and j, if we randomly pick two alleles at one locus from one individual and pick two alleles at the same locus from the second individual, there are still 15 partitions of the four sampled alleles G_i^1, G_i^2, G_j^1, G_j^2). The probability of any such block, or "the probability of randomly picking a gene at random from each individual in a list and having those genes satisfying imposed equivalence conditions imposed by these blocks" (page 27, [52]), was defined as a *generalized kinship coefficient* Φ. Consider the generalized kinship coefficient $\Phi[(G_i, G_j), (G_k, G_l)]$. Here G_i and G_j are in block 1 and IBD while G_k and G_l are in block 2 and IBD, and the alleles in block 1 are not IBD to the alleles in block 2.

3.2.1 Generalized kinship coefficients

In the chapter 2, we discussed the 15 detailed identity states of Gillois [36], Harris [45], and Jacquard [20]. δ_i is defined as the probability of the ith detailed identity state. When the maternal and paternal origins of the two pairs of alleles are ignored, the 15 detailed identity states (Figure 2.1) collapse to 9 condensed identity states S_1, S_2,$\cdots$, S_9 (Figure 3.2) [19], [36]. Let Δ_k denote the probability of condensed states S_k (also called condensed identity coefficient [19], $\Delta_k = P(\textit{alleles of } i \textit{ and } j \textit{ are in a certain partition in state } S_k)$), we have,

$$\begin{cases} \Delta_1 = \delta_1, & \Delta_2 = \delta_6 \\ \Delta_3 = \delta_2 + \delta_3, & \Delta_4 = \delta_7 \\ \Delta_5 = \delta_4 + \delta_5, & \Delta_6 = \delta_8 \\ \Delta_7 = \delta_9 + \delta_{12}, & \Delta_8 = \delta_{10} + \delta_{11} + \delta_{13} + \delta_{14} \\ \Delta_9 = \delta_{15} & \end{cases}$$

Let Ψ_k denote the probability of a random condensed state S_k ($\Psi_k = P(\textit{randomly picked alleles of } i \textit{ and } j \textit{ end up in a certain partition in state } S_k)$, which is an integer multiple of a generalized kinship coefficient and can be expressed in terms of the Δ_k's. Therefore, the problem of the calculation of the Δ_k's can be solved by computing the coefficients Ψ_k's, or more specifically, by computing the generalized kinship coefficients Φ's [19].

Without loss of generality, we denote G^m as a maternal allele and G^p as a paternal allele. We also assume that G_i^m and G_i^p occupy the upper two allele positions and G_j^m and G_j^p occupy the lower two allele positions within each state in Figure 2.1. Suppose we randomly pick two alleles G_i^1, G_i^2 from person i and two alleles G_j^1, G_j^2 from person j (digits 1 and 2 here only represent the order of two alleles picked, so they can be either a maternal allele or a paternal allele or both). Denote the probability of a random detailed identity state $\mathbf{S}_k$ by $\boldsymbol{\psi_k}$, which then is an integer multiple of a ***parental generalized kinship coefficient***. As

mentioned before, there are 15 possible $\boldsymbol{\psi_k}$s, which are listed as below,

$$\begin{cases} \boldsymbol{\psi_1} = \Phi[(G_i^1, G_i^2, G_j^1, G_j^2)], & \boldsymbol{\psi_2} = \Phi[(G_i^1, G_i^2, G_j^1), (G_j^2)] \\ \boldsymbol{\psi_3} = \Phi[(G_i^1, G_i^2, G_j^2), (G_j^1)], & \boldsymbol{\psi_4} = \Phi[(G_j^1, G_j^2, G_i^1), (G_i^2)] \\ \boldsymbol{\psi_5} = \Phi[(G_j^1, G_j^2, G_i^2), (G_i^1)], & \boldsymbol{\psi_6} = \Phi[(G_i^1, G_i^2), (G_j^1, G_j^2)] \\ \boldsymbol{\psi_7} = \Phi[(G_i^1, G_i^2), (G_j^1), (G_j^2)], & \boldsymbol{\psi_8} = \Phi[(G_i^1), (G_i^2), (G_j^1, G_j^2)] \\ \boldsymbol{\psi_9} = \Phi[(G_i^1, G_j^1), (G_i^2, G_j^2)], & \boldsymbol{\psi_{10}} = \Phi[(G_i^1, G_j^1), (G_i^2), (G_j^2)] \\ \boldsymbol{\psi_{11}} = \Phi[(G_i^1), (G_j^1), (G_i^2, G_j^2)], & \boldsymbol{\psi_{12}} = \Phi[(G_i^1, G_j^2), (G_i^2, G_j^1)] \\ \boldsymbol{\psi_{13}} = \Phi[(G_i^1, G_j^2), (G_i^2), (G_j^1)], & \boldsymbol{\psi_{14}} = \Phi[(G_i^2, G_j^1), (G_i^1), (G_j^2)] \\ \boldsymbol{\psi_{15}} = \Phi[(G_i^1), (G_i^2), (G_j^1), (G_j^2)] \end{cases}$$

For each $\boldsymbol{\psi_k}$, in the block(s) each of four alleles can be maternal or paternal, which makes a combination of 16 possible 'labeled IBD configurations', each with a probability of $\frac{1}{16}$. For example, for $\boldsymbol{\psi_{15}}$, these IBD configurations can be $[(G_i^m), (G_i^m), (G_j^m), (G_j^m)]$, $[(G_i^m), (G_i^p), (G_j^m), (G_j^m)]$, $\cdots$, and $[(G_i^p), (G_i^p), (G_j^p), (G_j^p)]$. We define $P[()]$ as the probability for a certain IBD configuration.

These $\boldsymbol{\psi_k}$'s can be expressed in terms of the δ_i's by conditioning on which detailed identity state the four original alleles of person i and person j occupy, which is a one-to-one function mapping. For example,

$$\boldsymbol{\psi_1} = \delta_1 + \frac{1}{4}\delta_2 + \frac{1}{4}\delta_3 + \frac{1}{4}\delta_4 + \frac{1}{4}\delta_5 + \frac{1}{8}\delta_9 + \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \qquad (3.1)$$

Suppose the four alleles of person i and person j occur in detailed identity state $\mathbf{S}_1$, then the four alleles picked at random falls in $\mathbf{S}_1$ with probability 1, which accounts for the first term on the right of equation (3.1). The second term $\frac{1}{4}\delta_2$ arises because if the four alleles of person i and person j occur in detailed identity state $\mathbf{S}_2$, both G_j^1 and G_j^2 must be so picked from the lower left allele of $\mathbf{S}_2$ (each with a probability of $\frac{1}{2}$,with a total probability of $\frac{1}{4}$) to achieve state $\mathbf{S}_1$ for the randomly selected alleles. The terms $\frac{1}{4}\delta_3$, $\frac{1}{4}\delta_4$, $\frac{1}{4}\delta_5$ are accounted for similarly. The term $\frac{1}{8}\delta_9$ arises because if the four alleles of person i and person j occur in $\mathbf{S}_9$, the four randomly sampled alleles must be drawn either from the left side of $\mathbf{S}_9$ or the right side of $\mathbf{S}_9$. The term $\frac{1}{8}\delta_{12}$ is accounted for similarly, with the alleles on the two opposite diagonal ends in $\mathbf{S}_9$ are IBD. The term $\frac{1}{16}\delta_{10}$ arises because the four randomly sampled alleles can only be drawn from the left side of $\mathbf{S}_{10}$, where two alleles are IBD. The terms of $\frac{1}{16}\delta_{11}$, $\frac{1}{16}\delta_{13}$, $\frac{1}{16}\delta_{14}$ are accounted for similarly. Finally, there are no terms involving δ_6, δ_7, δ_8, and

δ_{15} because it is obvious that these states do not allow IBD between any allele of i and any allele of person j.

Based on the same reasoning, all $\boldsymbol{\psi_k}$'s in terms of the δ_i's are derived as follows,

$$
\begin{aligned}
\boldsymbol{\psi_1} &= \delta_1 + \frac{1}{4}\delta_2 + \frac{1}{4}\delta_3 + \frac{1}{4}\delta_4 + \frac{1}{4}\delta_5 + \frac{1}{8}\delta_9 + \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_2} &= \frac{1}{4}\delta_2 + \frac{1}{4}\delta_3 + \frac{1}{8}\delta_9 + \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_3} &= \frac{1}{4}\delta_2 + \frac{1}{4}\delta_3 + \frac{1}{8}\delta_9 + \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_4} &= \frac{1}{4}\delta_4 + \frac{1}{4}\delta_5 + \frac{1}{8}\delta_9 + \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_5} &= \frac{1}{4}\delta_4 + \frac{1}{4}\delta_5 + \frac{1}{8}\delta_9 + \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_6} &= \delta_6 + \frac{1}{4}\delta_2 + \frac{1}{4}\delta_3 + \frac{1}{4}\delta_4 + \frac{1}{4}\delta_5 + \frac{1}{2}\delta_7 + \frac{1}{2}\delta_8 + \frac{1}{8}\delta_9 + \frac{3}{16}\delta_{10} + \frac{3}{16}\delta_{11} + \frac{1}{8}\delta_{12} + \\
&\quad + \frac{3}{16}\delta_{13} + \frac{3}{16}\delta_{14} + \frac{1}{4}\delta_{15} \\
\boldsymbol{\psi_7} &= \frac{1}{2}\delta_7 + \frac{1}{8}\delta_{10} + \frac{1}{8}\delta_{11} + \frac{1}{8}\delta_{13} + \frac{1}{8}\delta_{14} + \frac{1}{4}\delta_{15} \\
\boldsymbol{\psi_8} &= \frac{1}{2}\delta_8 + \frac{1}{8}\delta_{10} + \frac{1}{8}\delta_{11} + \frac{1}{8}\delta_{13} + \frac{1}{8}\delta_{14} + \frac{1}{4}\delta_{15} \\
\boldsymbol{\psi_9} &= \frac{1}{8}\delta_9 + \frac{1}{8}\delta_{12} \\
\boldsymbol{\psi_{10}} &= \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_{11}} &= \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_{12}} &= \frac{1}{8}\delta_9 + \frac{1}{8}\delta_{12} \\
\boldsymbol{\psi_{13}} &= \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_{14}} &= \frac{1}{16}\delta_{10} + \frac{1}{16}\delta_{11} + \frac{1}{16}\delta_{13} + \frac{1}{16}\delta_{14} \\
\boldsymbol{\psi_{15}} &= \frac{1}{4}\delta_{15}
\end{aligned}
\tag{3.2}
$$

From the equation (3.2), one can compute all the detailed identity state coefficients δ_i's by computing the coefficients $\boldsymbol{\psi_i}$'s, which can be derived by computing 15 ***parental generalized kinship coefficients***. In the next section, we will develop the algorithm for calculating those ***parental generalized kinship coefficients***.

3.2.2 Recursive computation of parental generalized kinship coefficients

A straightforward recursive algorithm for computing the generalized kinship coefficients was described in chapter 5 of the book by Lange (2002) [19]. Weeks et al. (1995) [52], by modifying the two components of the recursive algorithm: boundary conditions and recurrence rules, derived the X-linked recursive algorithm for computing the generalized kinship coefficients between two individuals by tracking alleles picked at random from X chromosomes. Here we use some of the definitions in [19] and [52] and define a similar algorithm for computing ***parental generalized kinship coefficients***.

Without loss of generality, throughout the following description, we assume the members of a pedigree are numbered in an ascending order so that parents always precede their offspring. We define new boundary conditions and recurrence rules in terms of probability of 'labeled IBD configurations' where every allele has an 'm' (maternal) or 'p' (paternal) label on it. Note that there is no random sampling with replacement from each individual that is implicit in the definition of the traditional kinship coefficients [19],[52].

3.2.2.1 Boundary conditions Boundary conditions are usually used in the static phase of the recursive algorithm to evaluate the boundary kinship coefficients involving only randomly picked alleles from founders [19]. However, here our new boundary conditions are used to evaluate the probabilities of 'certain labeled IBD configurations' involving only labeled alleles from founders. Note that the boundary condition **1** we define below actually applies to any person including nonfounders.

Boundary condition **1**: if the same (maternal or paternal) allele of a founder (i, or j) occurs in two or more blocks, then the probability of this type of labeled IBD configurations is equal to 0, *e.g.*, $P[(G_i^m), (G_i^p),(G_i^m)] = 0$. The condition holds because a founder only has two alleles at a certain autosomal locus.

Boundary condition **2**: if there are two founders in the same block, then the probability of this type of labeled IBD configurations is also equal to 0, *e.g.*, $P[(G_i, G_j),\cdots,] = 0$. The condition holds because founders are by definition unrelated and therefore their alleles are not IBD.

Boundary condition **3**: maternal alleles and paternal alleles of a founder can not occur in the same block, *e.g.*, $P[(G_i^m, G_i^p), \cdots,] = 0$. The condition holds because we assume founders are noninbred, so maternal and paternal alleles of a founder are not IBD.

Boundary condition **4**: if each block contains only founders and none of the above three boundary conditions holds, then the probability of this type of IBD configurations is equal to 1, *e.g.*, $P[\textit{certain labeled IBD configurations}] = 1$.

3.2.2.2 Recurrence Rules Recurrence rules are by definition used in the recursive phase of the recursive algorithm, in which an appropriately selected individual (nonfounder) i in the current 'label IBD configuration' is replaced by his/her parent(s) (j is the mother, k is the father), generating new IBD configurations. This replacement process travels up a pedigree and a selected allele of i is substituted by a maternal and a paternal allele from j or k depending on the origin of the allele to be replaced [19],[52]. The process stops when it hits the boundary conditions defined above.

Recurrence rule **1**: suppose s ($s \geq 1$) same alleles G_i^1, G_i^2, $\cdots$, G_i^s of a nonfounder i occur in only one block, we have

$$\begin{cases} P[(\underbrace{G_i^m, \cdots, G_i^m}_{s}, \cdots), \cdots, ()] = \frac{1}{2}\{P[(G_j^m, \cdots), \cdots, ()] + \Phi[(G_j^p, \cdots), \cdots, ()]\} & \langle \mathbf{1} \rangle \\ P[(\underbrace{G_i^p, \cdots, G_i^p}_{s}, \cdots), \cdots, ()] = \frac{1}{2}\{P[(G_k^m, \cdots), \cdots, ()] + \Phi[(G_k^p, \cdots), \cdots, ()]\} & \langle \mathbf{2} \rangle \end{cases}$$

$\langle \mathbf{1} \rangle$: i's maternal allele (m) appears in the IBD configuration s times, G_j^m is i's mother j's maternal allele, and G_j^p is j's paternal allele;
$\langle \mathbf{2} \rangle$: i's paternal allele (p) appears in the IBD configuration s times, G_k^m is i's father k's maternal allele, and G_k^p is k's paternal allele.

Recurrence rule **2**: suppose that s ($s > 1$) alleles G_i^1, G_i^2, $\cdots$, G_i^s of a nonfounder i occur in only one block, if among those s alleles, there are n ($n \geq 1$) maternal alleles and t ($t \geq 1$) paternal alleles, then we have

$$P[(\overbrace{\underbrace{G_i^m, \cdots, G_i^m}_{n}, \underbrace{G_i^p, \cdots, G_i^p}_{t}}^{s}, \cdots), \cdots, ()] = \frac{1}{4} \sum_{l \in (m,p)} \sum_{r \in (m,p)} P[(G_j^l, G_k^r, \cdots), \cdots, ()]$$

Recurrence rule **3**: suppose that a nonfounder i occurs in two blocks, with n ($n \geq 1$) maternal alleles in one block and t ($t \geq 1$) paternal copies in the other block, then we have

$$P[(\underbrace{G_i^m, \cdots, G_i^m}_{n}, \cdots), (\underbrace{G_i^p, \cdots, G_i^p}_{t}, \cdots), \cdots, ()] = \frac{1}{4} \sum_{l \in (m,p)} \sum_{r \in (m,p)} P[(G_j^l, \cdots), (G_k^r, \cdots), \cdots, ()]$$

This rule holds because the paternal allele (p) and maternal allele (m) of i can not be present in the same block under the condition that some of them are in the other block at the same time.

Recurrence rule **4**: this might not be called a recurrence rule. When either n or t (not both) is equal to zero, recurrence rule **2** and **3** both reduce to rule **1**.

3.3 SAMPLE APPLICATION

As an example of implementing above rules, we now begin to compute the $15th$ ***parental generalized kinship coefficient*** $\boldsymbol{\psi_{15}}$, $\Phi[(G_5^1), (G_5^2), (G_6^1), (G_6^2)]$ for inbred sibling 5 and 6 in Figure 3.1 as follows.

As we mentioned before, $\boldsymbol{\psi_{15}}$ is a combination of 16 states with probabilities from $P[(G_5^m), (G_5^m), (G_6^m), (G_6^m)]$, $P[(G_5^m), (G_5^p), (G_6^m), (G_6^m)]$, $\cdots$, to $P[(G_5^p), (G_5^p), (G_6^p), (G_6^p)]$. Because of boundary condition **1**, 12 of 16 states for $\boldsymbol{\psi_{15}}$ have a probability zero because two or more same alleles occurring in different blocks (*e.g.* $P[(G_5^m), (G_5^p), (G_6^m), (G_6^m)]$). The only state that has its probability coefficient ($\frac{1}{4}$) great than zero is $P[(G_5^m), (G_5^p), (G_6^m), (G_6^p)]$. So the problem of computing $\boldsymbol{\psi_{15}}$, $\Phi[(G_5^1), (G_5^2), (G_6^1), (G_6^2)]$, ends up with calculating $P[(G_5^m), (G_5^p), (G_6^m), (G_6^p)]$, which we show as follows,

(a) By recurrence rule 3,

$$\begin{aligned} P[(G_5^m), (G_5^p), (G_6^m), (G_6^p)] &= \frac{1}{4}\{P[(G_4^m), (G_3^m), (G_5^m), (G_5^p)] + P[(G_4^m), (G_3^p), (G_5^m), (G_5^p)] \\ &\quad + P[(G_4^p), (G_3^m), (G_5^m), (G_5^p)] + P[(G_4^p), (G_3^p), (G_5^m), (G_5^p)]\} \end{aligned} \tag{3.3}$$

(b) Again by recurrence rule 3,

$$\begin{aligned} P[(G_4^m),(G_3^m),(G_5^m),(G_5^p)] &= \frac{1}{4}\{P[(G_4^m),(G_3^m),(G_4^m),(G_3^m)] + P[(G_4^m),(G_3^m),(G_4^m),(G_3^p)] \\ &\quad + P[(G_4^m),(G_3^m),(G_4^p),(G_3^m)] + P[(G_4^m),(G_3^m),(G_4^p),(G_3^p)]\} \end{aligned} \tag{3.4}$$

(c) The first three terms of right side of the equation (3.4), $P[(G_4^m), (G_3^m), (G_4^m), (G_3^m)]$, $P[(G_4^m), (G_3^m), (G_4^m), (G_3^p)]$, and $P[(G_4^m), (G_3^m), (G_4^p), (G_3^m)]$ are equal to 0. The reason is that for those three IBD configurations next recursive step would proceed by replacing alleles in them by maternal and paternal alleles from either founder 1 or founder 2, which results in founder 1 or founder 2 being involved in three or more blocks. By boundary condition 1, those probabilities are equal to 0. Then the equation (3.4) reduces to

$$\begin{aligned} P[(G_4^m),(G_3^m),(G_5^m),(G_5^p)] &= \frac{1}{4}P[(G_4^m),(G_3^m),(G_4^p),(G_3^p)] \\ &= \frac{1}{4}\times\frac{1}{4}\{P[(G_4^m),(G_2^m),(G_4^p),(G_1^m)] + P[(G_4^m),(G_2^m),(G_4^p),(G_1^p)] \\ &\quad + P[(G_4^m),(G_2^p),(G_4^p),(G_1^m)] + P[(G_4^m),(G_2^p),(G_4^p),(G_1^p)]\} \end{aligned} \tag{3.5}$$

(d) By recurrence rule 3, and then boundary condition 1 and 3,

$$\begin{cases} P[(G_4^m),(G_2^m),(G_4^p),(G_1^m)] = \frac{1}{4}P[(G_2^p),(G_2^m),(G_1^p),(G_1^m)] = \frac{1}{4}\times 1 = \frac{1}{4} \\ P[(G_4^m),(G_2^m),(G_4^p),(G_1^p)] = \frac{1}{4}P[(G_2^p),(G_2^m),(G_1^m),(G_1^p)] = \frac{1}{4}\times 1 = \frac{1}{4} \\ P[(G_4^m),(G_2^p),(G_4^p),(G_1^m)] = \frac{1}{4}P[(G_2^m),(G_2^p),(G_1^p),(G_1^m)] = \frac{1}{4}\times 1 = \frac{1}{4} \\ P[(G_4^m),(G_2^p),(G_4^p),(G_1^p)] = \frac{1}{4}P[(G_2^m),(G_2^p),(G_1^m),(G_1^p)] = \frac{1}{4}\times 1 = \frac{1}{4} \end{cases}$$

So the equation (3.5) reduces to

$$P[(G_4^m),(G_3^m),(G_5^m),(G_5^p)] = \frac{1}{4}\times\frac{1}{4}\times(4\times\frac{1}{4}) = \frac{1}{16}$$

(e) Repeat similar procedures within (c) to (d), we can derive that

$$\begin{aligned} P[(G_4^m),(G_3^p),(G_5^m),(G_5^p)] &= P[(G_4^p),(G_3^m),(G_5^m),(G_5^p)] \\ &= P[(G_4^p),(G_3^p),(G_5^m),(G_5^p)] \\ &= P[(G_4^m),(G_3^m),(G_5^m),(G_5^p)] \\ &= \frac{1}{16} \end{aligned}$$

So the equation (3.4) reduces to,

$$P[(G_5^m),(G_5^p),(G_6^m),(G_6^p)] \quad = \quad \frac{1}{4} \times \frac{1}{16} \times 4 = \frac{1}{16}$$

That is, $\boldsymbol{\psi_{15}} = \frac{1}{16} \times \frac{1}{4} = \frac{1}{64}$. From the equation (3.2), the detailed identity state coefficient δ_{15} is

$$\delta_{15} = 4 \times \boldsymbol{\psi_{15}} = 4 \times \frac{1}{64} = \frac{1}{16}$$

3.4 DISCUSSION

In this chapter, we generalized the methods described in Karigl (1981) [50], Karigl (1982) [51], and chapter 5 of the Lange's book [19] and introduced a new algorithm for computation of the detailed identity state coefficients, which opens up the possibility of calculating covariances between even inbred relatives. To illustrate the algorithm, we applied it to calculate the coefficient for one detailed identity state for one inbred sibling pair. However, further work is still needed to test the algorithm.

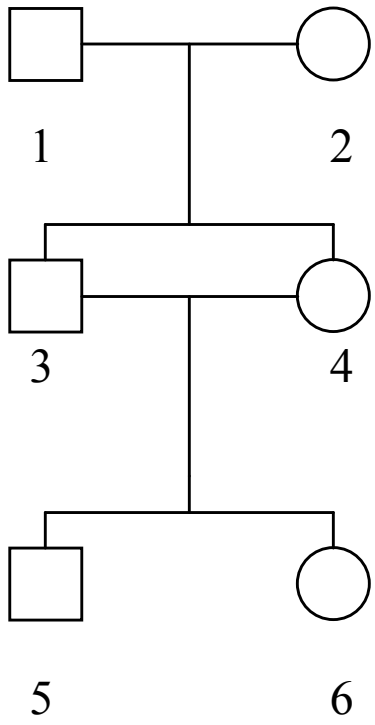


Figure 3.1: A Brother-Sister Mating (after Lange (page 72, [19])).

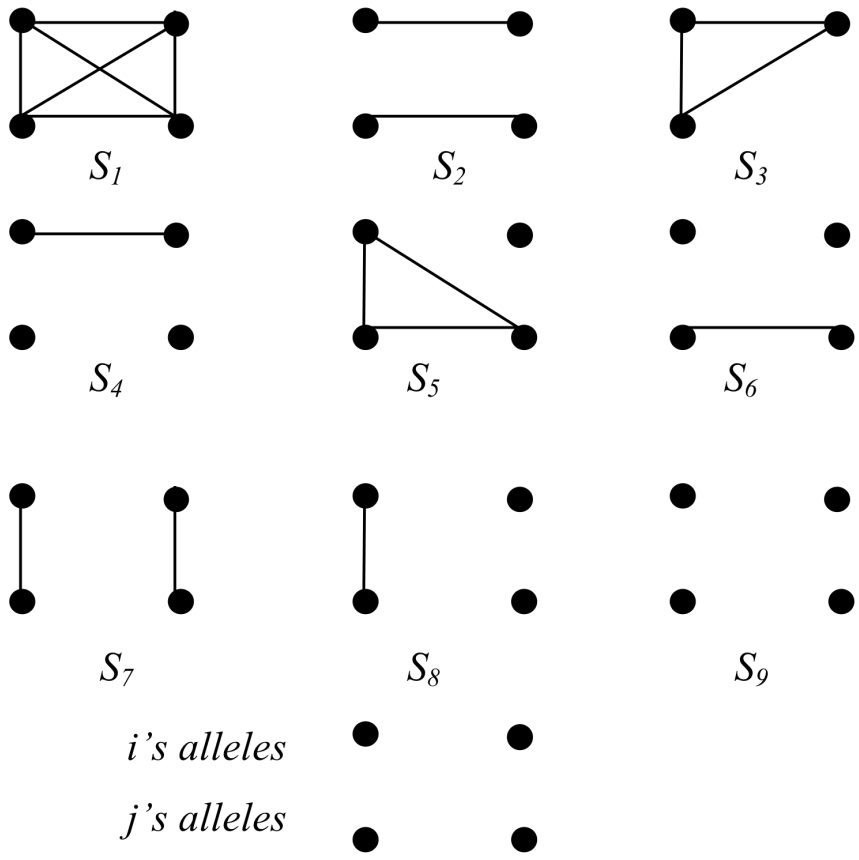


Figure 3.2: The nine condensed identity states (after Lange (page 74, [19])).

4.0 GENOME-WIDE SCAN FOR ADIPOSITY-RELATED PHENOTYPES IN ADULTS FROM SAMOAN ARCHIPELAGO

This chapter reports linkage results from a NIH funded project "Genome-Scan for Obesity Susceptibility Loci in Samoans" (R01-DK59642, PI: Dr. Stephen T. McGarvey from Brown University). Two manuscripts out of this chapter are being submitted for peer review [53], [54] with the one summarizing American Samoan results is accepted by Int J Obes [54]. Under the supervision of Dr. Daniel E. Weeks, I carried out the primary statistical analyses of the data and wrote much of the manuscripts. Dr. Stephen T. McGarvey is the corresponding author of the manuscripts. For issues related to the copyright of the manuscripts, please contact either Dr. Stephen T. McGarvey or Dr. Daniel E. Weeks for more information.

4.1 INTRODUCTION

Overweight and obesity have reached epidemic proportions on a global scale and are associated with economic modernization and the concomitant changes in physical activity and diet [55],[56],[57]. Obesity, body mass index (BMI) $\geq$ 30 kg/m^2, in the U.S. has become a major public health problem, affecting $\sim$33% of adults in 2002. The prevalence of obesity is much higher in U.S. minority ethnic groups, including Pacific islanders, and in those with low education [58]. Estimates based on National Health and Nutrition Examination Survey (NHANES) data show a marked increase in age-adjusted prevalence of obesity in US adults over years, from 13.4% in 1960-1962 to 30.5% in 1999-2000 [55],[58]. All these data also show significant disparities among racial groups with higher prevalence of both obesity and overweight among the minorities.

Compared to the US and most other populations, levels of overweight and obesity are remarkably high in Samoan adults and children residing in American Samoa and the independent nation of Samoa, and are strongly associated with measures of modernization [59], [60],[61],[62],[63]. In 2002 in American Samoa 89% and 92% of adult men and women, respectively, had BMI $>$ 26 kg/m^2 and in 2003 in Samoa 68% and 84% of adult men and women, respectively, had BMI $>$26 kg/m^2 [63]. (Body composition studies of Polynesians indicate that the criteria of BMI $>$ 26 kg/m^2 be used for overweight and BMI $>$ 32 kg/m^2 for obesity [64]).

Obesity is a complex phenotype resulting from interactions of multiple factors including genetic, environmental, cultural, nutritional, and life-style factors. Therefore, identification of underlying susceptibility genes and genetic variants associated with obesity has been difficult. Nonetheless, numerous studies using candidate gene approach and linkage analysis have identified genes and/or chromosomal regions that might harbor susceptibility genes, some of which have been replicated across studies and populations [57],[65]. With a reduced level of genetic variation, populations of recent ancestry that have remained isolated since their founding are likely to provide added advantage in understanding the genetic basis of complex phenotypes [66]. Samoans, with an evolutionary history of approximately 3000 years, relative isolation, large family sizes and their recent exposure to rapid modernization and the nutrition transition, offer an important opportunity for undertaking genetic studies in obesity related traits [59], [63], [67], [68], [69].

Due to its unique history of ancestry, allele frequencies at many loci in the Samoan population are significantly different from those in other population [70], therefore potential susceptibility loci to several obesity phenotypes are likely to be different than those in other populations. The primary purpose of our project was to identify the location of potential obesity susceptibility loci in two Samoan populations (American Samoa and (Western) Samoa). This research is of particular value in light of understanding the full picture of genetics of human obesity, that is, if the linkage results for obesity susceptibility loci are similar for both isolated groups (like Samoan population here) and cosmopolitan outbred human groups, then it offers hope for broad applicability of proteomic and drug discovery for future interventions of obesity; On the contrary, if there are considerable differences in

genetic influences of obesity across different populations, these differences must be taken into account in practice [69],[71].

We used maximum likelihood-based multipoint variance components linkage analyses to locate genes influencing complex adiposity-related phenotypes, which consists of BMI, abdominal circumference (ABDCIR), percent body fat (%BFAT, as measured by bioelectrical impedance), and fasting serum leptin and adiponectin. We also tested for linkage to some adjusted phenotypes, leptin and ABDCIR adjusted for BMI and %BFAT. Due to the complex role of genetic factors in common obesity, it is possible that these phenotypes might be under the influence of trait-specific genes, or shared single major genes (pleiotropic effects). In this study we resorted to the bivariate multipoint linkage analysis method to simultaneously examine linkage between genetic markers and two quantitative traits, a way to test for the pleiotropic effects of a single major gene. To investigate the sensitivity of linkage results to different pedigree structures, we also performed linkage analyses using nuclear pedigrees from breaking up the connections of higher generations in our original pedigree structures.

4.2 SUBJECTS AND METHODS

4.2.1 Study Population

As mentioned above, our study populations derive from both the U.S. territory of American Samoa and the less modernized independent country of (Western) Samoa, which is located just 100 km from American Samoa. In 2000, the population of American Samoa was 57,291, of which 88.2% are ethnic Samoans [62], [72] (57,794 by July 2006, estimated from CIA World Factbook). In 2001, the population of Samoa was 177,714, of which 92.6% are ethnic Samoans [73] (176,908 by July 2006, estimated from CIA World Factbook). American Samoa has higher levels of education, a higher proportion of adults in wage and salary occupations and higher economic and material lifestyle indicators than Samoa [59],[60],[62],[68]. For example, in 2000 52% of American Samoan adults were employed in the paid labor force,

while only 6.7% of adults participated primarily in subsistence activities. On the contrary, by 2001 only 36% of Samoan men and 26% of women were in wage or salary jobs [73].

All participants took part in the Samoan Family Study of Overweight and Diabetes in 2002-03. Recruitment in American Samoa in 2002 was based on random selection of probands seen in the 1990-94 cohort study in American Samoa [68],[71] and the presence of at least two adult siblings alive and residing in American Samoa. Participants in the 1990-94 cohort (which was originally designed as a longitudinal study of blood pressure change over time) had to (1) not have a medical diagnosis of hypertension or type 2 diabetes (based on doctor's report or current use of medications for either), and (2) self-report that all four grandparents were Samoan. In American Samoa we collected data from 1,311 individuals, including 755 adults, age $\geq$ 18 years, and 556 children in 212 households.

Recruitment in Samoa was from Feb-Sept 2003 and was first based on finding individuals in Samoa who were members of American Samoan pedigrees who had been recruited in 2002. We then selected villages throughout the nation to assess geographic and economic diversity, and chose families based on available number of adult siblings. In Samoa, we studied 1,465 individuals, 957 adults and 508 children from 96 households. This included 395 individuals who were members of 15 different families seen in American Samoan sample recruited in 2002.

Probands and families were unselected for obesity or related phenotypes. We did not recruit individuals who indicated at enrollment they wanted to participate because they had obese, diabetic or hypertensive members. Standard methods were used in the communities for collection of pedigree information [74]. Protocols for this study were approved by the Brown University Institutional Review Board, American Samoan Institutional Review Board, and the Government of Samoa, Ministry of Health, Health Research Committee. Written informed consent was obtained from all participants.

4.2.2 Pedigree information

After cleaning the pedigree structures constructed from our field work, there are 34 original pedigrees with a total of 1,968 individuals. Table 4.1 and Table 4.2 show for each original

pedigree the number of adults with phenotypic and genotypic data. The average family size is 57.88 with a range from 3 to 719 individuals (median 10) and an average of 3.71 generations (range 2 to 8). Some of our original pedigree structures are a bit larger than might be truly needed to connect only the American Samoans or Samoans, as some pedigree structures were constructed so as to connect members from both American Samoa and (Western) Samoa. Many of the upper generation connections are made through ungenotyped individuals, so relationship testing cannot be used to directly verify these untyped connections.

4.2.2.1 American Samoan pedigrees Intermediate American Samoan only pedigree structures were constructed using the PEDSYS database system (PEDSYS, Southwest Foundation for Biomedical Research): we first kept only phenotyped adults from American Samoa using the SUBSET command and then ran the INDEX command to construct pedigree structures (PEDTRIM command can do the same job). This procedure generated larger pedigree structures than the component nuclear pedigrees. In brief, 34 intermediate pedigrees (Table 4.1) were derived (190 nuclear pedigrees were also derived from breaking these intermediate pedigrees). The average intermediate family size is 45.94 with a range from 3 to 604 individuals (median 10) and an average of 3.57 generations (range 2 to 8). The average nuclear family size is 4.23 with a range from 3 to 8. In 34 intermediate pedigrees, there were $\geq 1,483$ relative pairs (trait leptin) that are useful and informative for linkage analysis, including ≥ 237 phenotyped sib pairs, ≥ 368 cousin pairs, and ≥ 302 avuncular pairs (Table 4.3).

4.2.2.2 Samoan pedigrees 46 intermediate Samoan pedigrees (Table 4.2) and 196 nuclear Samoan families were derived from breaking up our original pedigrees (using PEDSYS). The average intermediate family size is 35.09 with a range from 3 to 222 individuals and an average of 4.28 generations (range 2 to 8). In these 46 intermediate pedigrees, there were $\geq 1,633$ relative pairs (trait %BFAT) that are useful and informative for linkage analysis, including ≥ 251 phenotyped sib pairs, ≥ 503 cousin pairs, and ≥ 439 avuncular pairs (Table 4.4). The average nuclear family size is 4.61 with a range from 3 to 14.

In order to investigate the sensitivity of linkage results to different pedigree structures, we performed genome scans in both American Samoan and Samoan sample, using both larger intermediate pedigrees as well as smaller nuclear pedigrees. However, unless stated otherwise about the pedigree structures used in linkage analyses, the autosomal linkage results we discussed later in this chapter are derived using the samples of intermediate pedigrees, with comparisons to those linkage results derived using nuclear pedigrees.

4.2.3 Genotyping

Buffy coats were prepared from 10 ml of EDTA blood samples in the field, kept at -40^oC in American Samoa and then shipped on dry ice to the laboratory at Cincinnati. Genomic DNA was isolated using the Puregene Kit (Gentra Systems Inc., Minneapolis, MN) and quantitated, diluted to 20 ng/ml and arrayed in 96 well microtitre plates. The ABI PRISM Linkage Mapping Set MD10 (Applied Biosystem Inc., Forrest City, CA) consisting of 10 cM panels of microsatellite markers was used. We adopted a multiplex PCR (three to five markers) scheme developed in our laboratory. PCR was performed in 96-well plates and in each plate a sample of known genotype (CEPH DNA 1347-02) was used in two wells as a positive control and distilled water in one well as a negative control. PCR products from four 96-well plates were then assembled in one 384-well plate for analysis in the DNA sequencer. Thus, there were eight wells in the 384-well plate containing the CEPH sample with known genotype. In addition to serving as a positive control, this protocol ensured consistency of genotype data. Further, we used a Biomek FX liquid handling station to eliminate errors associated with manual handling of samples. Amplified products were separated on two automated DNA sequencers (Applied Biosystems)-3100 Genetic Analyzer that uses the POP4 polymer and internal size standard ROX400, and 3130XL Genetic Analyzer that uses POP7 polymer and internal size standard 500LIZTM. Raw genotype data generated in the two machines was transferred to a separate analysis computer for assigning individual genotypes using GeneMapper V4.0 software, which were manually checked by two persons ensuring correct calling of genotypes.

4.2.4 Phenotypes

Standard anthropometric techniques and measurements were used to measure stature,weight, skinfold thickness, and circumferences, and to calculate BMI [75]. Abdominal circumference was used as the primary measure of central fat distribution. Bioelectrical impedance measures of resistance were obtained with the RJL Systems, Inc. BIA-101Q device (Clinton, MI 48035), using standard procedures. These measures were then used to calculate fat-free mass and body fat percentage using equations established from body composition studies in Samoans [62],[64]. Fasting blood specimens were drawn after a 10-hour minimum overnight fast into Vacutainers, separated with a portable field centrifuge and stored in plastic storage tubes at 40 ^oC in local freezers. Serum was shipped on dry ice to Providence and the following assays were completed: serum leptin by radioimmunoassay (RIA) using a kit from ALPCO (Windham NH); serum insulin using standard RIA kits from Diagnostic Products, Inc; serum glucose using an automatic analyzer, Beckman CX4; serum adiponectin using RIA kits from Linco, Inc. (St. Charles, MI).

Interviews were used to collect information on years of education, occupation and physical activity from farm work. Farm work activity was coded as a dichotomous variable based on self-reported physical work at the farms (and in a few cases fishing) regularly each week or if subsistence work was the primary occupation [62]. Cigarette smoking data were collected by interview and coded as a dichotomous variable based on current smoking status.

These phenotypes were checked for data entry mistakes and for outliers. A power transformation was applied to the phenotypes that were not normally distributed [76],[77]. The power coefficients are phenotype-specific, but transformation for each phenotype was same in both American Samoan sample and Samoan sample.

4.2.5 Statistical Analyses: Error Checking and Data handling

Our analyses involved several steps, with a strong emphasis on validation of our data for consistency and integrity. Our genotyping data were subjected to several quality checks in order to help ensure accurate genotyping. First, we compared the heterozygosity rates observed with the rates expected on the basis of the estimated allele frequencies to assay

the quality of our genotyping panel data. We also performed checks for 'reasonable' allele sizes, numbers of alleles, and frequencies. Then, we used the relationship-inference program RELPAIR v2.0.1 [78],[79],[80] to check the accuracy of our self-reported pedigree relationships (after checking for internal consistency of ages). The relationship checking results were used to double-check our field-collected pedigrees, as well as to conservatively adjust the pedigree structure for maximal reduction of pedigree structure errors. Misclassification of half siblings as full siblings, unrelated persons as parent-offspring, and cousins as full siblings were identified and resolved. A total of 93 relationships were changed, with 13 individuals added as connecting parents. There was one coding error that was inconsistent with the field notes.

Another relation-estimation program PREST [81],[82] was also used concurrently to assess consistency of allele-sharing pattern with the specified relationship. No families were removed as a result of relationship inconsistency. Unresolved relationship errors led to the removal of the questionable subjects from the study. We also excluded from analysis two individuals who were each part of a monozygotic twin pair.

We used PEDCHECK [83] to check genotypes at each locus for Mendelian inconsistencies; for the analyses on the nuclear families this information was used to zero out families containing Mendelian inconsistencies at a specific marker. For the larger pedigree structures, LOKI [84] was used to remove a minimal set of genotypes so as to generate an internally consistent set of genotypes for all family members.

The PEDSYS database system (PEDSYS, Southwest Foundation for Biomedical Research) was used to prepare the pedigree structure file. Mega2 [85] and the statistical software R (The R Project for Statistical Computing) were used interactively to set up the other files for the analyses performed in this study.

4.2.6 Allele Frequency Estimation

Our planned analyses are sensitive to allele frequency estimates, and we have previously shown that Samoans have allele frequencies at microsatellite loci that differ from European-derived populations [67] on whom most allele frequency estimates in current databases are

based. Therefore, we estimated marker allele frequencies from our pedigree data, while taking the pedigree structure into account using LOKI [84] and simultaneously estimating the identity-by-descent (IBD) sharing matrices. Using LOKI in this manner has the advantage of properly modeling the variability in the estimates of the allele frequencies [86],[87]. For verification, we compared LOKI-derived estimates to those derived by the new BLUE algorithm of McPeek et al. (2004) [88]; this algorithm is particularly useful for large complex pedigrees like some of ours for which MLE calculation is computationally impractical. Tests of fit to Hardy-Weinberg expectations, including tests for excess homozygosity, were used to confirm the quality of the data.

4.2.7 Genetic Map

Our Kosambi genetic map was taken from the Rutgers Combined Linkage-Physical Map of The Human Genome [89]. Linear interpolation was used to predict the genetic position of polymorphic markers that were not present in the Rutgers map. However, please note that in the figures the genomic locations of all genetic markers were based on a Haldane map scale.

4.2.8 Multipoint Linkage Analysis

4.2.8.1 Autosomal Univariate Multipoint Linkage Analysis A multipoint variance components (VC) approach was used to test for linkage between marker loci and the adiposity-related phenotypes. In this approach, the phenotypic and genetic information from all the pedigree members is considered simultaneously, and the expected genetic covariances between relatives are specified as a function of the observed IBD at a given genetic position as estimated from the marker data [5].The multipoint IBD matrices were estimated by LOKI and were imported into SOLAR [6], in which the VC models were fit by simultaneously estimating the trait mean μ (as a function of the covariates), three variance components σ_q^2 (additive genetic variance due to the major locus), σ_g^2 (variance due to residual additive genetic effects), and σ_e^2 (variance caused by random environmental effects) using maximum-

likelihood techniques. This approach allows for locus-specific effects, residual effects, covariates, epistasis, multivariate phenotypes, and random environmental effects.

For a given phenotype, a likelihood-ratio test for linkage was carried out by testing whether the additive genetic variance due to the QTL σ_q^2 was significantly different from 0 by comparing the likelihood of the general model, in whichσ_q^2 is estimated, with that of the restricted model, in which σ_q^2 is constrained to 0. Twice the difference of the ln likelihoods of these two models yields a test statistic that is asymptotically distributed as a 1/2:1/2 mixture of a χ_1^2 and a point mass at zero [21]. The classical LOD scores are obtained by converting the statistic into values of log to the base 10. LOKI uses a stochastic approximation algorithm, which can result in minor variation in the LODs from run to run. Thus the reported LODs are the average of maximum LODs obtained from ten runs; ranges of those scores are also reported.

Violations of multivariate normality assumptions variance component approach can lead to non-robust results, *e.g.*, excessive Type I errors when the distribution of the trait is markedly leptokurtic [90]. To account for possible deviations of trait distributions from multivariate normality, we transformed our phenotypes to approximate normality prior to linkage analyses (See Phenotypes section). We also routinely used the multivariate t-distribution in SOLAR to guard against possible non-normality for each phenotype after transformation, which is more robust than using SOLAR's default multivariate normal distribution [6]. A LOD score $\geq$ 3.3 was taken as evidence of significant linkage, which is equivalent to a P value of .0001 or less. A LOD score $\geq$ 1.175 and LOD score $\geq$ 1.9 were considered to show potential linkage and evidence of suggestive linkage, respectively [91].

Three sets of multipoint linkage analyses were performed; these analyses differed in terms of which covariates were included in the mean model. In the first set of linkage analyses, only the age and sex effects were screened for statistical significance while modeling familial relationships using SOLAR, and only covariates with significant effects at the P-value ≤ 0.10 level were retained for the subsequent analyses. In the second set of linkage analyses, the individual environmental exposures farm work, smoking, and years of education, in addition to age and sex effects, were screened for inclusion in the statistical model. Since the power of the variance component linkage analysis approach is proportional to heritability, one might

increase power by reducing the environmental variance of the phenotype by adjusting for covariates [92]. Comparison of these two sets of analyses may help reveal genes that control the residual variation after the effects of behavioral or environmental covariates have been removed. The kurtosis of residuals of above phenotypes after adjusting for significant covariates was within normal range (< 0.8). Finally, in the third set of analyses, in order to search for genes influencing the propensity to deposit fat in the abdominal region, we analyzed ABDCIR after removing the effects of BMI or %BFAT, in addition to other covariates; we also tested for linkage to BMI (%BFAT)-adjusted leptin concentrations. The kurtosis of residuals of adjusted ABDCIR and leptin levels after adjusting for significant covariates were above normal range (> 1.5). These residuals were standardized.

4.2.8.2 Autosomal Bivariate Multipoint Linkage Analysis For linkage analyses performed with related traits, it is important to differentiate between pleiotropic effects and co-incident linkage when two or more traits show linkage to the same region, which can be done using bivariate linkage analysis approach [22]. In this approach, the trait-specific estimates of the mean, variance-components σ_q^2 (major gene effects), σ_g^2 (residual additive genetic effects), and σ_e^2 (random environmental effects) as well as three associated correlations ρ_q (correlation caused by a major gene), ρ_g (correlation caused by residual additive genetic effects), ρ_e (correlation caused by random environmental effects) are estimated simultaneously using maximum likelihood techniques. The hypothesis of no linkage for either trait (*i.e.*, $\sigma_{q_1}^2 = \sigma_{q_2}^2 = 0$) was tested using likelihood-ratio tests, in which the log-likelihood of the restricted model was compared with that of the model in which trait-specific σ_q^2 was estimated. The likelihood ratio statistic is asymptotically distributed as a $1/4\chi_2^2{:}1/2\chi_1^2{:}1/4\chi_0^2$ mixture distribution [22],[21]. For ease of interpretation, the bivariate LOD score can be adjusted to a univariate-equivalent LOD score, LODeq, which has an equal asymptotic P-value as the bivariate LOD.

In bivariate linkage analysis, likelihood-ratio tests for pleiotropy or coincident linkage were made at the chromosomal location where the highest linkage peaks reside. To test pleiotropy (*i.e.*, the same major gene affects the two phenotypes) or coincident linkage (*i.e.*, a set of clustered genes, each influencing a particular trait), the likelihood for the model

in which ρ_q was estimated was compared with the likelihood of the model in which ρ_q was constrained to 0 (coincident linkage) or ρ_q constrained to 1 or -1 (complete pleiotropy). For the test of complete pleiotropy, because ρ_q being constrained to 1 or -1 involves a boundary condition, twice the differences in likelihoods is distributed as a 1/2:1/2 mixture of a χ^2_1 and a point mass at zero [21]. When testing for coincident linkage, twice the difference in likelihoods is distributed as a χ^2 distribution with one degracccee of freedom. A chi-square P-value of 0.05 was suggested as a sufficiently conservative cut-off for the rejection of either pleiotropy or coincident linkage [22], however, one must remain aware that linkage disequilibrium within a region can reduce the power to reject pleiotropy [22], and that power of bivariate analyses to detect linkage is highest when the trait locus and the environment induce phenotypic covariation in opposite directions [97],[98].

At this time, covariate screening in bivariate models is not supported in SOLAR and researchers are expected to do covariate screening in univariate models first by including them in trait-specific analysis. This has been shown to improve both the power to detect susceptibility loci and precisely localize mapping of correlated traits to the same chromosomal region [22],[23]. For our bivariate analyses, the individual environmental exposures farm work, smoking, and years of education, in addition to age and sex effects, were screened for inclusion. Almasy et al. (1997) [22] also recommended that strong pleiotropy must be scrutinized for confounding linkage disequilibrium in the region, which should not be a problem for our 10 cM genome scan.

4.2.8.3 X-linked Multipoint Linkage Analysis The current version of SOLAR does not carry out correct X-linked variance components analysis, as it does not yet support fitting of the proper variance component models [93],[94]. Recently, Lange and Sobel (2006) [95] discussed a new model for mapping X-linked quantitative trait loci (QTLs), which is incorporated in latest version of the genetic analysis program Mendel [96], with which we performed X-linked QTL analysis. A plethora of model options in Mendel allowed us to try different genetic models in our X-linked analysis, in which different combinations of covariate sets, X-linked QTL, autosomal additive polygenic, X-linked additive polygenic, random environmental variance components were modeled assuming our traits follow multivariate t

distributions. However, since Mendel cannot handle our intermediate pedigrees, so we used our nuclear families in these analyses.

4.3 RESULTS

4.3.1 Results from American Samoan sample

We genotyped 377 autosomal microsatellite markers with an intermarker average spacing of 9.79 (8.36-12.16) cM (Haldane map), and 18 microsatellite markers on the X chromosome, in 580 American Samoan adults (247 males and 333 females).

Table 4.5 displays the phenotypic characteristics of the adult participants. The large BMI, %BFAT and ABDCIR values indicate the high levels of overweight and obesity among American Samoans and are characteristic of modern Polynesians. Our five 'primary phenotypes (BMI, %BFAT, ABDCIR, leptin, and adiponectin) are highly heritable with heritability ranges from 0.41 to 0.62, adjusted for different sets of significant covariates (Table 4.6). Four adjusted traits (BMI or %BFAT adjusted ABDCIR and leptin) also have medium heritability estimates in the range 0.31 to 0.45. These estimates indicate a major role of genes in the control of adiposity in the American Samoan population.

Univariate linkage analyses were performed on all the phenotypes with adjustment for the indicated covariates listed in Table 4.6. For BMI- or %BFAT-adjusted ABDCIR and leptin, we standardized the their residuals (with high kurtosis above normal (> 0.8)) before linkage analysis. The chromosomal regions with (average) maximum multipoint LOD scores ≥ 1.5 (based on 10 runs) are presented in Table 4.7. In addition, the ranges of the locations (cM) of the maximum LOD scores are also displayed. These show that as the LOD score gets higher, the maximum LOD score tends to occur in the same position from run to run, even though it may vary slightly in magnitude.

Figure 4.1 shows the univariate multipoint results of one single SOLAR/LOKI run for chromosomes 6, 13, and 16 where we observe maximum LOD scores greater than 2.30, or observe clustering of linkage peaks in the same regions for multiple traits. The results of the

multipoint genome scan for susceptibility loci of five 'primary' adiposity-related traits are plotted in Figure 4.2 and Figure 4.3.

The highest multipoint LOD score was 3.83 for leptin, close to marker D6S262 located in 6q23.2. The region harboring a putative QTL is broad, and the 1-LOD support interval surrounding the peak extends from 137.0 to 155.2 cM (Figure 4.1). The second strongest evidence for linkage was also for leptin with a LOD of 2.98 in 16q21 at location 91.0-91.9 cM near marker D16S503. Suggestive linkage, LOD $\geq$ 1.9, was detected for 13 other trait/region pairs including both %BFAT and leptin at D12S86 in 12q24.23. There were 7 other unique trait/region pairs with 1.5 $\leq$LOD$<$ 1.9, including both BMI and %BFAT at 14q12-q13.1. Adjusting for environmental covariates (*e.g.*, farm work, education, and cigarette smoking) led to larger LOD scores for leptin at 6q23.2 and 16q21, for %BFAT at 16q12-q21, for adiponectin at 13q33.1, and for %BFAT-adjusted ABDCIR at 1q31.1, 3q27.3-q28, and 12p12.3 (Table 4.7).

There is co-localization of linkage signals for phenotypes BMI, %BFAT, leptin, and ABDCIR to the same region involving markers D16S415-D16S515 (Table 4.7 & Figure 4.1). The bivariate analyses indicate that there are significant genetic and environmental correlations between pairs of four phenotypes (Table 4.8). Results of bivariate linkage analyses of the selected phenotype pairs for chromosome 16 are reported in Table 4.9 and the equivalent-univariate LOD_{eq} scores are plotted in Figure 4.4. The maximum bivariate LOD scores with 2-df vary from 2.13 for BMI-ABDCIR pair to 2.98 for %BFAT-leptin. Likewise, the LOD_{eq} scores range from 1.68 to 2.48. All linkage peaks in Table 5 overlap the region 16q21 at 90.9-91.9 cM, except for the ABDCIR-%BFAT pair (the second highest peak for the ABDCIR-%BFAT pair was detected on 16q21 at 90.9 cM with a LOD_{eq} of 2.23). Since the bivariate linkage analyses are computationally intensive, Table 4.9 and Figure 4.4 only display results from one single run. However, similar variations of LOD_{eq} scores as seen in univariate LOD scores (Table 4.7) would be expected should multiple analyses be performed.

As shown in Table 4.9, the $\rho_q s$, the correlations between pairs of adiposity phenotypes due to major gene effects are quite high, ranging from 0.97 to 1.00. Coincident linkage hypothesis was strongly rejected in favor of pleiotropy, and none of the tests for complete pleiotropy rejected the hypothesis that the locus-specific genetic correlation (ρ_q) was equal

to 1, indicating that variations in the given pair of phenotypes are mediated by common genetic factors (*e.g.*, a putative common QTL).

Results of our univariate VC linkage analyses using the 190 nuclear families are summarized in Table 4.10. Maximum LOD scores ≥ 1.5 (one run) were observed for only 8 putative QTLs, and a comparison of Table 4.10 to Table 4.7 shows that only two linkage peaks identified using the original pedigrees show suggestive linkage using the nuclear families (linkage of %BFAT to 12q24.23-q24.23 with LOD 2.17, and linkage of ABDCIR to 2p11.2 with LOD 2.09).

We performed X-linked variance components analyses incorporating different genetic models (Table 4.11) with no maximum LOD scores > 1.3 observed anywhere on the X chromosome for the primary phenotypes BMI, %BFAT, leptin and ABDCIR, except for the phenotype adiponectin. When an X-linked polygenic component was not modeled (models 1, 4 in Table 4.11), we detected elevated LOD scores to adiponectin across the entire X chromosome (Figure 4.2 , Figure 4.3). Instead, after an X-linked polygenic component was modeled (models 2,3,5,6, in Table 4.11), no linkage was detected for adiponectin with a mere maximum LOD score of 0.70 (data not shown). Modeling X-linked polygenic and autosomal polygenic backgrounds together or only an X-linked polygenic component in our X-linked analysis only gave marginal changes in the LOD scores (model 2 vs. model 3; model 5 vs. model 6, Table 4.11). Likewise, adjusting for environmental covariates or not gave marginal changes in the maximum LOD scores across the X chromosome (data not shown).

4.3.2 Results from Samoan sample

In total 378 autosomal microsatellite markers with an average intermarker spacing of 9.51 (8.06 -12.54) cM (Haldane), and 14 microsatellite markers on the X chromosome with an average intermarker spacing of 12.25 cM, were typed for 572 Samoan adults (278 males and 294 females) that were analyzed in this study.

Table 4.12 displays the phenotypic characteristics of the adult participants. The high mean values for BMI, %BFAT and ABDCIR are remarkable, as are the low serum adiponectin levels. Females tend to have higher leptin levels than males, which is similar to what we

observed in our previous study of American Samoans. All mean phenotype levels in males are lower than those in females, which might be partially due to the fact that over 80% of men in Samoa participate in farm work and participating in farm work is associated with lower adiposity [62]. The heritability estimates range from 0.26 to 0.43 for the traits included in this study (Table 4.13). Note that about 30% of the subjects have %BFAT record missing due to unexpected failure of the instrument used during fieldwork.

The univariate multipoint linkage results with (average) maximum multipoint LOD scores (≥ 1.5) are presented in Table 4.14. Figure 4.5 displays results (one run) for chromosomes 4, 7, 9, and 13 where we observed average maximum LOD scores that reached the suggestive linkage level (LOD ≥ 1.9). Multipoint LOD scores for primary traits are plotted in Figure 4.6 and Figure 4.7. As shown in Table 4.14, the highest (average) LOD score that we observed was 2.30 for leptin, close to marker D13S265 in 13q31.3. Linkages of BMI (LOD 2.09), %BFAT (LOD 1.62) and ABDCIR (LOD 1.66) were also mapped to 13q31.3.

Suggestive linkage (LOD ≥ 1.9) was detected for five other trait/region pairs including, %BFAT (LOD 2.09) near marker D4S414 in 4q22.1, %BFAT (LOD 2.19) near D7S484 in 7p14.3, ABDCIR (LOD 2.14) near D9S285 in 9p22.3-p22.2, adiponectin (LOD 1.96) near marker D2S160 on 2q13, and BMI-adjusted leptin in 19q12-q13.13 (LOD 2.03). Similar as found in American Samoan sample, adjusting for environmental covariates such as farm work, education, and cigarette smoking improved our ability to detect linkage to some phenotypes (*e.g.*, BMI, ABDCIR, and leptin in 13q31.3, to adiponectin at 2q13 and 18q22.3, and to ABDCIR at 9p22.3-p22.2, Table 4.14, Figure 4.5). However, the impact of adjusting for these covariates on linkage signals varies from trait to trait.

Similar as observed in American Samoan sample, significant genetic correlations exist between phenotypes BMI, %BFAT, ABDCIR and leptin (Table 4.15). For the combination adiponectin-ABDCIR we observed significant negative genetic correlation. Since genetic correlation is a measure of the extent to which same major genes affect two traits, the existence of significant genetic correlations suggest that there might be common genes influencing the variation between these correlated phenotypes.

A bivariate LOD score of 3.10 (P-value 2.9×10^{-4}) was obtained for %BFAT-ABDCIR in 9p22.2-p21.3. In this region BMI-%BFAT obtained a bivariate LOD score of 2.79 and

BMI-ABDCIR a score of 1.88 (Table 4.16). On chromosome 13q31.3 bivariate LOD scores ranged from 1.96 to 2.69 for the trait combinations of BMI, %BFAT, ABDCIR and leptin. All bivariate LOD scores were converted to the "equivalent-univariate" LOD scores (LODeq) so as to be comparable to univariate LOD scores. The LODeq scores are plotted for chromosome 9 and 13 in Figure 4.8. Notably, all linkage peaks on chromosome 13q31.3 are right at marker D13S365.

In chromosome 13q13.3 and 9p22.2-p21.3 no tests for complete pleiotropy, but all tests for coincident linkage were strongly rejected for phenotype pairs displayed in Table 4.16. In addition, the correlations between phenotype pairs due to QTL effects (ρ_q) range from 0.95 to 1.00, which is highly significant different from zero. These results are strongly in favor of pleiotropic effects between the respective phenotype pairs.

Table 4.17 presents the results (LOD $\geq$ 1.5) of our autosomal genome scan using Samoan nuclear pedigrees. We detected two QTLs with suggestive linkage for %BFAT on chromosome 11q13.2 (LOD 2.22) and 12q23.1 (LOD 2.18) as well as a QTL for ABDCIR (LOD 1.99) on 9p22.2-p21.3. No multipoint LOD scores >1.0 were detected anywhere across the X chromosome using nuclear families. Modeling either X-linked polygenic and autosomal polygenic backgrounds or only the latter in X-linked analysis as well as adjusting for environmental covariates or not gave only marginal changes in the LOD scores (Data not shown). X-linked linkage results (from using model 4 in Table 4.11) were plotted in Figure 4.6 and Figure 4.7.

4.4 DISCUSSION

In this chapter we discussed our genome scan study for obesity susceptibility loci in Samoans (American Samoans vs. Samoans). The American Samoan polity and the Samoan polity have a common population history but have recently been differently influenced by modernization which partly is reflected in differences in phenotypes (Table 4.7 vs. Table 4.14). In general, BMI, %BFAT, ABDCIR and serum leptin levels tend to be lower and serum adiponectin levels tend to be higher in the individuals from Samoa. The Samoan popula-

tion, which is less influenced by modernization, tends to perform more farm work and smoke less than American Samoans. These environmental factors might partly be involved in the physical differences observed.

4.4.1 Genome scan of American Samoans

In our genome-wide scan for adiposity-related phenotypes in adults from American Samoa consisting of 34 pedigrees containing 578 genotyped members, we detected linkage at several chromosomal regions: the principal regions with evidence of suggestive linkage (LOD $\geq$ 1.9) indicated by our univariate linkage analyses were 12q24.23-q24.32 and 16q12.2-q21 for %BFAT; 1q42.2, 5q11.2, 6q23.2, 12q24.23, 16q12.2-q21, and 19p13.3 for leptin; 14q12 for %BFAT-adjusted leptin; 2p13.1-p11.2 and 16q23.1 for ABDCIR; 1q31.1, 3q27.3-q28, and 12p12.3 for %BFAT adjusted ABDCIR; and 13q33.1 for adiponectin. Far fewer linkage signals were detected using nuclear pedigrees instead. To our knowledge, this genome scan and our accompanying scan in Samoans are the first genome-wide studies of adiposity-related traits in the population from the Samoan islands.

The strongest linkage signal, average LOD = 3.83, was found for leptin in 6q23.2 near marker D6S262. This is near the obesity candidate gene detected in French adults and children, ENPP1 (ectonucleotide pyrophosphate phosphodiesterase, OMIM 173335), an inhibitor of insulin receptor tyrosine kinase activity [99],[100]. Several other genome-scan studies have reported QTLs in the 6q region for type 2 diabetes, obesity, and metabolic syndrome in Mexican Americans [101],[102], Finns [103],[104], and French [99],[105],[106]. Given the high levels of adiposity in the adult American Samoans, it is noteworthy that one study reporting linkage of leptin to 6q24 took place among French nuclear families with severe adult obesity [99].

The second highest linkage signal, average LOD = 2.98, was for leptin in 16q21 near marker D16S503 (the same chromosomal region appears to harbor putative susceptibility loci for other phenotypes BMI, %BFAT and ABDCIR). This 16q21 region has been implicated by multiple other studies [107],[108],[109] and contains the candidate gene AGRP (Agouti-related protein) or ART (Agouti-related transcript) (OMIM 602311), which regulates body

weight via central melanocortin receptors. We also detected linkage between leptin and chromosome 5q11.2 near marker D5S407 (LOD = 2.08). The ISL1 gene (insulin-enhancer binding protein 1, OMIM 600366), a transcription factor involved in pancreas development and insulin production, has been mapped to this chromosomal region [110]. Evidence of linkage for adiponectin was observed with a LOD of 2.41 on chromosome 13q33.1 near marker D13S158, a region identified in other recent genome scan studies [111],[112].

Suggestive linkage with both %BFAT and leptin were detected in the region 12q24.23-q24.32 between the markers D12S86-D12S324, similar to what were found in a genome-scan study in European-American families [113]. Important candidate genes within the 1-LOD support region are the transcription factor TCF1 (OMIM 142410) and MODY3 (OMIM 600496), whose mutant variants are associated with maturity-onset diabetes [114], [115],[116],[117],[118]. The region may harbor a putative major gene that exhibits pleiotropic effects for correlated traits %BFAT and leptin.

We adjusted ABDCIR for %BFAT, in order to analyze a phenotype that isolates specific fat in the abdominal depots or locations that are assessed indirectly with circumference from the overall influence of total adiposity. This may allow for detection of novel QTLs, as suggested by Norris et al. [119]. We detected three chromosomal regions with average LOD scores > 2.0 for %BFAT-adjusted ABDCIR at 1q31.1, 3q27.3-q28, and12p12.3, which were different from the linkages detected to unadjusted ABDCIR. The 1-LOD support region for %BFAT-adjusted ABDCIR at 1q31.1 near marker D1S238 contains a possible candidate gene FCAMR (Fc fragment of IgA and IgM receptor, OMIM 605484), which plays important roles in host immunity, allergy, and autoimmunity, with elevated total IgM and IgA-levels have been observed in type-2 diabetes [120],[121]. However, our findings should be taken cautiously since the kurtosis of residuals of %BFAT-adjusted ABDCIR after adjusting for significant covariates were too high (>1.5), although procedures have been taken to alleviate possible false positives due to violations of normality assumption.

In a genome-wide scan of obesity in the Old Order Amish, Hsueh et al. (2001) [122] demonstrated the high heritability of BMI-adjusted leptin for the first time, and mapped the trait to chromosome 10p region (LOD 2.73). In our study, suggestive linkage to %BFAT-adjusted leptin was detected at 14q12 region, which was not identified for either leptin, or

%BFAT alone or BMI-adjusted leptin. Further work is needed to explore genetic and environmental influences on such adjusted leptin phenotypes as they may represent specific adiposity or obesity susceptibility phenotypes.

Significant genetic correlations between phenotypes BMI, %BFAT, ABDCIR, and leptin suggest that there might be common genes influencing their variations. We demonstrated by bivariate analyses that the region 16q21 near marker D16S503 harbor a putative susceptibility locus that may simultaneously affect these correlated phenotypes. Although all the LODeq scores were less significant than the peak univariate LOD score we observed for leptin alone (LOD 3.83), some of the LODeq scores were higher than the univariate LOD scores obtained for BMI, %BFAT, and ABDCIR alone. This is consistent with the finding of Almasy et al. (1997) [22] that, when analyzing a 'weaker' trait and a 'stronger' trait together in the presence of pleiotropy, the bivariate power is usually no more than the univariate power for the stronger trait. Because our phenotypes are genetically correlated, we made no attempt to correct for multiple testing.

Our genome scan on autosomal chromosomes was carried out by the VC approach implemented in SOLAR. Another genetic package LOKI was used to calculate the multipoint IBD matrix that was input into SOLAR to test linkage. As discussed earlier, using LOKI has the advantage of properly modeling the variability in the estimates of the allele frequencies. However, we found that the LOKI multipoint IBD matrix varies depends on the initial random seed, which results in variation of the LOD scores SOLAR later estimated. Therefore, this provides a range of LOD scores, which as we described becomes narrower as the LOD score increases.

Our univariate linkage analyses implicated a common chromosomal region at 16q12.2-q21, influencing the phenotypes BMI, %BFAT, ABDCIR, and fasting serum leptin. We also showed that there are significant genetic correlations between pairs of these four traits, suggesting that these traits might have common genes influencing their variability. Our subsequent bivariate linkage analyses implied that the region 16q21 near marker D16S503 harbors a putative susceptibility locus that simultaneously affects all four of these adiposity related phenotypes. Although all of the bivariate LOD_{eq} scores were less significant than the maximum LOD score we observed for the phenotype leptin alone, some of the bivariate

LOD_{eq} scores were higher than those univariated LOD scores obtained for BMI, %BFAT, and ABDCIR alone.

Significant linkage to Xq24 has been detected in a genome scan of Finnish population [123]. Later the positional candidate gene SLC6A14 was reported to show association with obesity in Finnish population [124],[57]. In contrast, our X-linked analyses detected no strong evidence of linkage anywhere on the X chromosome. One possible reason for the observed chromosome-wide elevated LOD scores for adiponectin under some models might be driven by the fact that male variance of it is higher than its female variance although the difference is very small. By taking the safeguards suggested in Lange and Sobel [95], the spurious linkage dropped. In contrast to autosomal results, we did not observe large changes in the linkage signals of putative QTLs when we adjusted for environmental covariates in the X-linked analyses. In X-linked analyses, we did not screen for significant covariates; however, this choice would not affect the resulting LOD scores that much because the likelihood of the model that includes extra nonsignificant covariates approximates that of the model without those extra covariates.

For a complex trait that is determined by multiple genetic and environmental factors such as diet, physical activity, and smoking, careful adjustment for significant environmental effects may increase signal-to-noise ratio in linkage analysis by decreasing the proportion of the residual phenotypic variation attributable to these adjusted factors [125]. In our genome scan, we compared the analyses with and without adjustment for significant environmental effects (*e.g.*, farm work activity, education, cigarette smoking), by which some putative genes may be identified regarding their direct importance for the adiposity trait per se or indirect importance via the influence on the covariates only. Notably, most of the genetic regions identified for significant or suggestive linkage were detected only when certain environmental effects had been adjusted for in our linkage analyses. For example, the highest linkage to leptin was observed on chromosome 6q23.2, LOD = 3.83, after adjusting for significant covariates of sex, farm work, education and cigarette smoking; When only smoking and sex effects were accounted for, the highest linkage to leptin still was observed on chromosome 6q23.3 with a LOD score of 2.52 (not shown in Table 4.7). However, when only a sex effect was accounted for, the LOD score dropped to 1.82. These results also imply a possible

genotype-by-smoking interaction for leptin levels, which has been reported in the San Antonio Family Heart Study [126], [127].

Finally, we emphasize that our findings of linkage derived from a study population characterized by excessive adiposity. Therefore, the regions reported here may or may not be identified in multiple independent studies elsewhere [65]. Our similar ongoing studies on adults from less economically developed Samoa may also yield different findings due to the lower adiposity levels and less altered nutritional environments.

4.4.2 Genome scan of Samoans

In our study of 572 genotyped adults (46 intermediate pedigrees) from Samoan polity, we have detected six chromosomal regions, 2q13, 4p22.1, 7p14.3, 9p22.3-p22.2, 13q31.3 and 19q12-q13.13, with suggestive LOD scores (LOD $\geq$ 1.90) for adiposity-related traits. When those pedigrees were divided into 196 nuclear families to repeat linkage analyses, only three chromosomal regions with suggestive LOD score were detected, which are 9p22.2-p21.3, 11q13.2 and 12q23.1.

The American Samoan polity and the Samoan polity have a common population history but have recently been differently influenced by modernization which partly is reflected in differences in phenotypes measured in our two samples. In general, BMI, %BFAT, ABDCIR and serum leptin levels tend to be lower and serum adiponectin levels tend to be higher in the individuals from Samoa. The Samoan population, which is less influenced by modernization, tends to perform more farm work and smoke less than American Samoans. These environmental factors might partly be involved in the physical differences observed.

The strongest univariate linkage signal, LOD = 2.30, was found for leptin near marker D13S265 in 13q31.3, after adjustment of environmental effects, as well as age and sex effects. When only age and sex effects were adjusted for, a lower signal (LOD =1.47) was detected (Figure 4.5, solid line), which is consistent with the fact that carefully adjusting for environmental effects can increase "signal-to-noise ratio" in genetic linkage analysis [125]. We also observed evidence of linkage for BMI and %BFAT in this chromosomal region. This region in 13q31.3 and its 1-LOD-drop support interval has been reported in other recent studies

[65], [128], [129]. For example, a genome-wide parent-of-origin linkage analysis by Dong et al. (2005) [129] found strong evidence (LOD of 3.72 for BMI) for an obesity susceptibility locus with paternal effect in 13q32 in an European American sample. Yet no specific gene variant has been found to be associated with obesity-related traits in the region 13q31.3.

One of the chromosomal regions that appears to be suggestively linked to %BFAT (LOD 2.19) in this study is located near marker D7S484 in 7p14.3. Additional evidence for linkage has previously been found between D7S484 and the serial changes of BMI from childhood to adulthood [112]. The neuropeptide Y gene (NPY) (7p15.1, MIM 162640) (which is implicated by supporting evidence to work synergistically with leptin in regulating body fat utilization and storage as well as hormone release [130]) located in this region has been reported to be linked [131] and associated[132] with both obesity and other obesity-related traits in Mexican Americans, yet its role in etiology of common forms of obesity in this population is unclear. In addition, in the flanking 7p15.3 region a suggestive linkage for fat-free mass (LOD 2.7) was detected at marker D7S1808 in the Quebec Family Study [133]. Another possible positional candidate gene, glucokinase (GCK) (MIM 138079), has been shown to be linked with maturity-onset diabetes of the young (MODY) [134], and mutations of the GCK gene have been detected in French MODY families [135].

The second chromosomal region exhibiting suggestive linkage (LOD 2.09) to %BFAT is near marker D4S414 in 4q22.1. In addition, there is a weaker linkage peak (LOD 1.85, Figure 4.5) between markers D4S412 and D4S2935 in 4p16.2-p16.1. Linkage of obesity phenotypes to this region or its 1-LOD-drop support region has been reported in recent studies [65],[136],[137]. It is worthwhile to note that two positional candidate genes, cholecystokinin A receptor (CCKAR) (MIM 118444) and peroxisome proliferative activated receptor (PPARGC1A) (MIM 604517) are located near this region; both genes have major roles in the development of obesity [137].

Suggestive linkage to ABDCIR (LOD 2.14) and potential linkage to %BFAT (LOD 1.76) were detected in 9p22.2-p21.3, where strong evidence of linkage (LOD = 3.4) to high density lipoprotein (HDL-C) levels has been found in Mexican Americans [137] and it is known that obesity is associated with lower HDL-C levels. The genomic region 2q13 may contain a QTL for the variation of adiponectin, with a LOD score of 1.96 near marker D2S160. At the

flanking 2q14 region, Deng et al. (2002) [125] obtained a maximum LOD score (MLS) of 4.44 for BMI in their genome-wide linkage scan for quantitative trait loci for obesity phenotypes.

There is evidence of linkage for adiponectin (LOD 1.87) and BMI-adjusted leptin (LOD 2.03) in the 19q12-q13.3 region, where Saar et al. (2003) [138] detected a peak LOD score for obesity in German children and adolescents. The nearby 19q13 region contains two prominent candidate genes for obesity: apolipoprotein E precursor (APOE) (MIM 107741) and transforming growth factor, beta 1 (TGFB1) (MIM 190180). APOE codes a glycoprotein that plays a central role in lipid metabolism and several studies have reported positive associations of APOE with obesity phenotypes [139],[140],[141],[142]; The TGFB1 peptide is a multifunctional cytokine with roles in cell differentiation, and immune modulation in many cell types including adipocyte precursor cells [143],[144]. TGFB1 has been reported to be closely associated with BMI in human adipose tissue during morbid obesity [145]. In addition, one polymorphism (T29C) in the TGFB1 gene was recently reported to be associated with abdominal obesity in Swedish men [146]. Recently Long et al. (2003) [144] reported positive associations between APOE and TGFB1 and obesity phenotype variation in a large sample of white Europeans, which adds new evidence of suggesting the possible effects of the two genes on obesity.

In the present study as well as in our similar study in American Samoa polity, we show that there are significant genetic correlations between all of the pair-wise combinations of BMI, %BFAT, leptin and ABDCIR (Table 4.15); this implies that there might be genes in common that influence the variation of these traits. Additional support for such pleiotropic effects is given by the bivariate results from both of our studies which suggest promising susceptibility loci for adiposity-related traits on chromosome 9p and 13q in Samoa and on 16q in American Samoa. However, also the environmental correlations between the traits are statistically significant, which supports the well-known fact that environmental factors are of great importance for the variation of obesity-related traits. Furthermore, despite the overall genetic homogeneity in the population from the Samoan islands, there is considerable variation in environmental exposures (*e.g.*, diet, exercise, etc) across the population ("westernized" American Samoa vs. rural Samoa). In the two genome-wide studies, we attempted to adjust for such environmental factors by including environmental covariates in the statisti-

cal models. In both studies, we detected varying linkage signals in the same genomic regions when environmental effects were adjusted for vs. unadjusted for. This further suggests that gene by environment interactions potentially could be detected using more elaborate sets of behavioral and environmental exposure variables not currently available for these datasets, such as dietary factors, general physical activity or even psychosocial stress factors. However, the covariates that turned out to be significant differed between the two sample sets, for the majority of phenotypes. It is therefore possible that adjustment for environmental variance of adiposity phenotypes was not equally successful in the two studies. Further more, it is possible that this environmental variation will fluctuate the genetic effect and result in varying disease penetrance, which in turn could be of major importance for the minimal overlap seen between the linkage results found in the Samoa and the American Samoa.

Another possible reason for the lack of reproducibility seen in Table 4.14 between the two sample sets from the Samoan islands could be due to genetic heterogeneity. However, since previous studies have shown that there is no evidence of population substructure [70] in the population from the Samoan islands, this is improbable. Finally, the relatively larger family size in the American Samoan study vs. the Samoan study might play a role in explaining the lack of reproducibility as an effect of variation in statistical power.

Pooling data across studies is one way to increase power in linkage analysis of complex disease [91],[107]. We are currently carrying out a genome-wide linkage scan for adiposity phenotypes based on a combined sample set of both the Samoan families and the American Samoan families. Some chromosomal regions of linkage evidences identified here have been confirmed by our preliminary analysis on the combined sample (Data not shown). However, because two genotyping platforms were used respectively for American families and Samoan families, several families have members from both polities, how to efficiently align alleles still remains challenging.

4.5 CONCLUSIONS

In summary, here we report the results from our genome-wide linkage scans to search for susceptibility loci for adiposity-related phenotypes in adults from both American Samoa and (Western) Samoa. Our linkage analyses reveal several chromosomal regions with suggestive linkage (significant linkage at 6q23.2) that may harbor genes for adiposity. Our studies show strong support for different chromosomal regions, respectively, that appear to harbor a gene that has significant pleiotropic effects on multiple adiposity traits. However, due to the uniqueness of the studied groups, its population history and potential natural selection, these susceptibility loci for adiposity found in the population from the Samoan islands may or may not be identified in multiple independent studies elsewhere [65]. Furthermore, the sample set from the Samoan islands with its homogenous population history but with its heterogeneous environmental settings offers an unique possibility to study gene by environmental interaction that should be taken advantage of. Further linkage and association studies of the susceptibility loci found in our studies may allow for identification of candidate genes and pathways that are of strong importance for variation in adiposity phenotypes.

Table 4.1: Description of American Samoan Families-Original Pedigrees and Intermediate Pedigrees.

Original Pedigrees				Intermediate Size Pedigrees			
ID	Size	Genotyped	Phenotyped*	ID	Size	Genotyped	Phenotyped*
1	553	208	211	1	531	208	211
2	20	4	4	2	12	3	3
3	719	204	206	3	604	203	205
4	55	4	4	4	11	4	4
5	36	4	4	5	26	4	4
6	47	4	4	6	35	8	8
7	78	8	8	7	19	4	4
8	127	14	14	8	34	14	14
9	96	26	26	9	57	26	26
10	9	3	3	10	22	8	8
11	22	8	8	11	7	4	4
12	7	4	4	12	8	4	4
13	8	4	4	13	9	5	5
14	9	5	5	14	13	5	5
15	13	5	5	15	4	2	2
16	4	2	2	16	14	11	10
17	14	11	10	17	8	2	2
18	8	2	2	18	10	4	3
19	10	4	3	19	19	7	7
20	19	7	7	20	8	2	2
21	8	2	2	21	13	5	5
22	13	5	5	22	4	2	2
23	4	2	2	23	7	2	2
24	7	2	2	24	5	3	3
25	7	4	4	25	7	4	4
26	10	4	4	26	10	4	4
27	5	2	2	27	5	2	2
28	6	2	2	28	6	2	2
29	4	2	2	29	4	2	2
30	26	6	6	30	26	6	6
31	13	9	9	31	13	9	9
32	5	4	4	32	5	4	4
33	3	2	2	33	3	2	2
34	3	3	3	34	3	3	3
Total		580	583			578	581

*: Phenotype is BMI.

Table 4.2: Description of the Samoan Families- 46 original Pedigrees and 47 intermediate Pedigrees.

Original Pedigrees				Intermediate Size Pedigrees			
ID	Size	Genotyped	Phenotyped*	ID	Size	Genotyped	Phenotyped*
1	553	9	12	1	12	4	5
2	20	4	5	2	28	7	8
3	719	42	49	3	40	27	30
4	28	7	8	4	32	13	13
5	55	27	30	5	4	3	3
6	36	13	13	6	52	16	18
7	4	3	3	7	58	21	26
8	52	16	18	8	42	14	17
9	58	21	26	9	58	21	26
10	47	14	17	10	15	4	4
11	78	21	26	11	12	3	3
12	15	4	4	12	57	16	20
13	12	3	3	13	43	21	24
14	57	16	20	14	36	16	22
15	43	21	24	15	16	2	4
16	36	16	22	16	18	6	9
17	16	2	4	17	40	10	12
18	18	6	9	18	22	9	11
19	40	10	12	19	10	3	6
20	22	9	11	20	9	3	3
21	10	3	6	21	19	7	10
22	9	3	3	22	44	11	13
23	19	7	10	23	53	21	26
24	44	11	13	24	47	17	24

Note.-The table is continued on the next page.

Table 4.2: Continued.

Original Pedigrees				Intermediate Size Pedigrees			
ID	Size	Genotyped	Phenotyped*	ID	Size	Genotyped	Phenotyped*
25	53	21	26	25	18	6	8
26	47	17	24	26	58	24	26
27	18	6	8	27	80	21	22
28	59	24	26	28	49	27	33
29	127	34	37	29	49	9	12
30	96	27	33	30	56	28	28
31	57	28	28	31	22	13	15
32	74	18	20	32	73	18	20
33	57	15	16	33	222	42	49
34	43	16	20	34	56	15	16
35	37	24	23	35	43	16	20
36	9	2	3	36	36	24	23
37	3	2	2	37	6	2	3
38	14	7	8	38	3	2	2
39	12	5	6	39	12	7	8
40	11	3	5	40	12	5	6
41	26	13	13	41	10	3	5
42	39	13	14	42	25	13	13
43	3	2	2	43	38	13	14
44	3	2	2	44	3	2	2
45	5	3	3	45	3	2	2
46	3	2	2	46	5	3	3
				47	3	2	2
Total	2787	572	669		1649	572	669

*: Phenotype is BMI.

Table 4.3: Pairwise relationships in 34 intermediate pedigrees, used in SOLAR/LOKI analyses of adults from American Samoa.

	Number of pairs counted				
Phenotyped relative pairs	BMI	%BFAT	Leptin	ABDCIR	Adiponectin
Parent-Offspring	306	304	299	309	285
Siblings	250	250	237	252	233
Half-Siblings	72	72	71	72	69
1st Cousins	391	391	368	391	377
Grandparent-Grandchild	33	33	32	33	33
Avuncular	309	309	302	315	298
Half-Avuncular	122	122	122	122	119
Great-Avuncular	55	52	52	52	51
Total	1538	1533	1483	1546	1465

Table 4.4: Pairwise relationships in 46 intermediate pedigrees, used in SOLAR/LOKI analyses of adults from Samoa

	Number of pairs counted				
Phenotyped relative pairs	BMI	%BFAT*	Leptin	ABDCIR	Adiponectin
Parent-Offspring	442	283	425	448	422
Siblings	365	251	349	365	338
Half-Siblings	42	16	37	42	37
1st Cousins	667	503	602	667	583
Grandparent-Grandchild	82	49	77	83	77
Avuncular	636	439	584	636	572
Half-Avuncular	69	31	68	70	66
Great-Avuncular	112	61	108	112	108
Total	2415	1633	2250	2423	2203

*: Large number of missing values for the phenotype (broken measurement kit).

Table 4.5: Characteristics of study participants from American Samoa.

	Males	Females
Characteristics	N = 249 (42.7%)	N = 334 (57.3%)
Age(years)	43.50 ± 16.79	43.07 ± 16.13
Body mass index (BMI) (kg/m^2)	33.45 ± 7.59	36.64 ± 8.50
Percent body fat (%BFAT)	33.62 ± 6.51	41.61 ± 6.41
Leptin(μ/ml)	11.36 ± 9.76	30.19 ± 16.10
Abdominal circumference (ABDCIR) (cm)	107.70 ± 16.22	111.35 ± 16.70
Adiponectin (mg/ml)	8.40 ± 6.45	11.10 ± 10.02
Education (years)	11.72 ± 2.34	11.93 ± 2.45
Smoking[a]	0.39 (0-1)	0.21 (0-1)
Farm work activity[b]	0.51 (0-1)	0.24 (0-1)

Note. - Data are mean ± s.d., unless otherwise indicated. Phenotypes were not transformed.

[a]Mean(range), 0 = no cigarette smoking, 1= cigarette smoking.

[b]Mean(range), 0 = no farm work, 1= farm work.

Table 4.6: Residual heritability of adiposity-related phenotypes in adults from American Samoa, adjusted for different covariates.

Phenotype	N[a]	h_r^2 (s.e.)	Variance explained by covariates (%)	Significant Covariates[b]
BMI	581	0.50 (0.10)	4.1	S
	558	0.41 (0.11)	7.7	S, F, E
%BFAT	580	0.60 (0.10)	31.6	A, S
	557	0.52 (0.11)	32.5	A, S, F, E
Leptin	571	0.62 (0.09)	40.4	S
	516	0.52 (0.10)	45.9	S, F, E, C
	569	0.31(0.01)	55.8	BMI, S
	536	0.36 (0.10))	57.0	BMI, S, F, C
	566	0.42 (0.01)	34.2	%BFAT, A, S
	533	0.45 (0.10))	32.4	%BFAT, A, S, C
ABDCIR	583	0.50 (0.09)	4.4	A, S
	560	0.42 (0.10)	6.6	A, S, F, E
	581	0.32 (0.09)	25.0	BMI, A, S
	581	0.33 (0.09)	22.9	BMI, A, S, F
	580	0.45 (0.10)	20.3	%BFAT, A, S
	547	0.44 (0.10)	20.9	%BFAT, A, S, C
Adiponectin	569	0.55 (0.09)	15.7	A, S
	547	0.54 (0.09)	23.3	A, S, F, E

Note. - All heritability estimates are significantly different from zero at P-value $< 10^{-4}$.

[a]Number of total phenotyped individuals in the heritability analysis.

[b]Significant covariates (P-value < 0.1) kept in polygenic model. A=age, S=sex, F=farm work, E=education, C=cigarette smoking.

Table 4.7: Summary of SOLAR/LOKI multipoint linkage analyses for adiposity-related phenotypes in adults from American Samoa (with max LOD score ≥ 1.5).

Trait	cytogenetic position	Flanking marker/s	Range(cM)[a]	LOD score (range)[b]	Covariates[c]
BMI	14q12-q13.1	D14S275, D14S70	17.8-19.7	1.47 (1.36-1.52)	S, F, E
	16q21	D16S503	91.0-105.9	1.57 (1.47-1.70)	S
%BFAT	8p22-p21.3	D8S549, D8S258	34.7-35.7	1.82(1.75-1.86)	A, S, F, E
	12q24.23-q24.32	D12S86, D12S324	162.4-163.4	**1.96 (1.94-1.98)**	A, S, F, E
	14q12-q13.1	D14S275, D14S70	20.7-23.6	1.63 (1.51-1.72)	A, S
			15.8-19.8	1.55 (1.50-1.61)	A, S, F, E
	16q12.2-q21	D16S415, D16S503	87.9-88.9	**2.24 (2.16-2.34)**	A, S
			88.9-89.9	1.56 (1.50-1.61)	A, S, F, E
Leptin	1q42.2	D1S2800, D1S2785	271.8-276.8	**1.97 (1.94-2.00)**	S, F, E, C
	5q11.2	D5S407	77.3	**2.08 (1.87-2.22)**	S, F, E, C
	6q23.2	D6S262	150.5	**3.83 (3.81-3.84)**	S, F, E, C
				2.06 (2.04-2.07)	S
	12q24.23	D12S86	154.4	**2.06 (2.03-2.08)**	S, F, E, C
	13q14.2-q22.1	D13S153, D13S156	71.7-73.8	1.77 (1.68-1.87)	S
	16q21	D16S503	90.9-91.9	**2.98 (2.92-3.03)**	S, F, E, C
	16q12.2-q21	D16S415, D16S503	85.9-86.9	**1.99 (1.94-2.03)**	S
	19p13.3	D19S209, D19S216	11.9	**2.05 (2.02-2.07)**	S, F, E, C
Leptin, adjusted by %BFAT	14q12	D14S275	16.8-17.8	**2.01 (1.92-2.16)**	A, S, F, E, C
				1.70 (1.63-1.82)	A, S

Note. - Table continued on the next page.

Table 4.7: Continued.

Trait	cytogenetic position	Flanking marker/s	Range(cM)[a]	LOD score (range)[b]	Covariates[c]
ABDCIR	2p13.1-p11.2	D2S286, D2S2333	116.4	**1.92 (1.88-1.96)**	A, S, F, E
	6q23.2	D6S262	150.5	1.55 (1.54-1.56)	A, S
	12p13.33	D12S352	0	1.82 (1.80-1.85)	A, S, F, E
	16q23.1	D16S515, D16S516	102.6	**1.95 (1.92-1.98)**	A, S, F, E
			103.6	**1.95 (1.87-2.03)**	A, S
ABDCIR, adjusted by %BFAT	1q31.1	D1S238	215.3-217.3	1.71 (1.63-1.85)	A, S
			215.3-219.3	**2.36 (2.29-2.50)**	A, S, C
	12p12.3	D12S310	41.4	1.68 (1.64-1.71)	A, S
				2.22 (2.20-2.24)	A, S, C
	3q27.3-q28	D3S1262, D3S1580	226.3-227.3	1.60 (1.55-1.68)	A, S
			227.3-228.3	**2.04 (2.03-2.08)**	A, S, C
Adiponectin	13q33.1	D13S158	110.8-111.8	**2.41 (2.36-2.51)**	A, S, F, E
				2.23 (2.18-2.31)	A, S

Note. - LOD scores (mean, range) are derived from 10 independent Solar/Loki runs. Regions showing significant (LOD $\geq$ 3.3) or suggestive (LOD $\geq$ 1.9) linkage are highlighted in bold.

[a] Range of the locations of the maximum LOD scores in centimorgans (Haldane) from p terminus.

[b] Mean of maximum LOD scores and their ranges.

[c] Significant covariates adjusted before linkage analysis. A=age, S=sex, F=farm work, E=education, C=cigarette smoking.

* Out of 10 separate analyses, highest LOD scores observed 2 times at 105.9 cM, 1 time at 91.0 cM, 7 times at 91.9 cM.

Table 4.8: Genetic (ρ_g) and environmental (ρ_e) correlations between selected adiposity-related phenotypes in adults from American Samoa.

Phenotype pairs	$\rho_g \pm s.e$	$\rho_e \pm s.e$
BMI-%BFAT	0.89 ± 0.03	0.90 ± 0.03
BMI-ABDCIR	0.94 ± 0.02	0.86 ± 0.03
BMI-Leptin	0.85 ± 0.06	0.63 ± 0.08
ABDCIR-%BFAT	0.85 ± 0.04	0.79 ± 0.05
ABDCIR-Leptin	0.78 ± 0.07	0.58 ± 0.08
%BFAT-Leptin	0.78 ± 0.06	0.64 ± 0.08

Note. - BMI was adjusted for sex; %BFAT was adjusted for age and sex; ABDCIR was adjusted for age, sex, farm work, and education; Leptin was adjusted for sex, farm work, education, and cigarette smoking. All correlations are significant at P-value $< 10^{-5}$.

Table 4.9: Multipoint bivariate linkage analyses of pairs of adiposity-related phenotypes in adults from American Samoa.

Phenotype pairs	cytogenetic position	cM[a]	Closest marker	Bivariate LOD score	P-value[b]	LOD^{c}_{eq} score	ρ^{d}_{q}	P-value[e]	P-value[f]
BMI-%BFAT	16q21	91.9	D16S503	2.41	0.00753	1.94	0.99	0.2721	$< 10^{-6}$
BMI-ABDCIR	16q21	91.0	D16S503	2.13	0.00717	1.68	1.00	0.5000	$< 10^{-6}$
BMI-leptin	16q21	89.9	D16S503	2.82	0.00080	2.32	1.00	0.5000	$< 10^{-6}$
ABDCIR-%BFAT*	16q23.1	101.7	D16S515	2.74	0.00130	2.25	0.97	0.2356	$< 10^{-6}$
ABDCIR-leptin	16q21	91.9	D16S503	2.94	0.00063	2.44	1.00	0.5000	$< 10^{-6}$
%BFAT-leptin	16q21	89.9	D16S503	2.98	0.00082	2.48	1.00	0.5000	$< 10^{-6}$

Note. - BMI was adjusted for sex; %BFAT was adjusted for age, sex; and ABDCIR were adjusted for age, sex, farm work, and education; Leptin was adjusted for sex, farm work, education, and cigarette smoking; Results are based on one particular run.

[a] Distance in centi Morgans (Haldane) from p-terminus.

[b] Asymptotic P-value, under the null hypothesis that the likelihood-ratio statistic (2ln10*bivariate LOD score) is distributed as $\frac{1}{4}\chi^2_2 : \frac{1}{2}\chi^2_1 : \frac{1}{4}\chi^2_0$ mixture.

[c] LODeq is the equivalent univariate LOD score (df =1) corresponding to the reported bivariate LOD score with 2 df.

[d] ρ_q is the correlation due to QTL effects.

[e] P-value of the test for complete pleiotropy, for which the likelihood for the linkage model in which ρ_q was estimated was compared with that of the model in which ρ_q was constrained to 1 or -1.

[f] P-value of the test for no coincident linkage, for which the likelihood for the linkage model in which ρ_q was estimated was compared with that of the model in which ρ_qs was constrained to 0.

* Second highest LOD_{eq} score was 1.97, occurring on 16q21 at 90.9 cM for trait pair ABDCIR-%BFAT.

Table 4.10: Summary of multipoint linkage results of adiposity-related phenotypes using nuclear pedigrees, American Samoa (with max LOD score ≥ 1.5).

Structure	Trait	cytogenetic position	Flanking marker/s	Genomic location (cM)[a]	Maximum LOD score	Covariates[b]
Nuclear	BMI	3q29	D3S1311	251.7	1.74	S
	%BFAT	8q12.2	D8S260	86.2	1.65	A, S, E
					1.64	A, S, E
		12q24.23-q24.32	D12S86, D12S324	155.4	**1.90**	A, S, E
				157.4	**2.17**	A, S
		16p13.13	D16S3075	32.6	1.52	A, S
	ABDCIR	2p11.2	D2S2333	120	**2.09**	A, S, E
					1.77	A, S
		3q29	D3S1311	251.7	1.83	A, S
	Adiponectin	1p36.22	D1S450	10.6	**2.32**	A, S, E
		9q34.3	D9S1826, D9S158	174.7	1.61	A, S

Note. - LOD scores were derived from one particular SOLAR/LOKI run. Regions showing suggestive (LOD ≥ 1.9) linkage are highlighted in bold.

[a] Genomic location of the maximum LOD score in centi Morgan (Haldane) from p terminus.

[b] Covariates included in the model: . A=age, S=sex, F=farm work, E=education, C=cigarette smoking.

Table 4.11: Genetics models in X-linked variance components linkage analysis.

Model	Predictor*	Standardize trait within sex groups	Variance Components			
			X-QTL	Autosomal additive	X-poly additive	Environmental
Model1	Yes		Yes	Yes		Yes
Model2	Yes		Yes		Yes	Yes
Model3	Yes		Yes	Yes	Yes	Yes
Model4	Yes	Yes	Yes	Yes		Yes
Model5	Yes	Yes	Yes		Yes	Yes
Model6	Yes	Yes	Yes	Yes	Yes	Yes

* Two sets of covariates were adjusted for in all models: one set includes sex and age; the other includes sex, age, education, farm work and cigarette smoking.

Table 4.12: Characteristics of phenotyped participants from Samoa.

	Males	Females
Characteristics	N = 331 (49.3%)	N = 341 (50.7%)
Age(years)	41.39 ± 16.27	45.56 ± 17.36
Body mass index (BMI) (kg/m^2)	28.91 ± 5.46	32.91 ± 7.53
Percent body fat (%BFAT)	28.33 ± 7.22	39.19 ± 6.70
Leptin(μg/ml)	6.43 ± 6.89	23.76 ± 13.70
Abdominal circumference (ABDCIR) (cm)	95.88 ± 15.14	106.73 ± 16.11
Adiponectin (mg/ml)	9.99 ± 7.85	12.83 ± 8.29
Education (y)	9.73 ± 3.40	10.04 ± 3.00
Smoking[a]	0.43 (0-1)	0.16 (0-1)
Farm work activity[b]	0.83 (0-1)	0.31 (0-1)

Note. - Data are mean $\pm$ s.d., unless otherwise indicated. Phenotypes are not transformed.

[a]Mean(range), 0 = no smoking, 1= smoking.

[b]Mean(range), 0 = no farm work, 1= farm work.

Table 4.13: Residual heritability of adiposity-related phenotypes in adults from Samoa, adjusted for different covariates

Phenotype	N[a]	h_r^2 (s.e.)	Variance explained by covariates (%)	Significant Covariates[b]
BMI	669	0.43 (0.08)	12.4	A, S
	662	0.43 (0.07)	13.5	A, S, E
%BFAT	414	0.38 (0.10)	46.3	A, S
Leptin	642	0.38 (0.08)	52.2	A, S
	534	0.27 (0.10)	54.5	A, S, E, C
	539	0.26 (0.10)	58.8	BMI, S, F, C
	638	0.38 (0.09)	55.9	BMI, S
ABDCIR	671	0.40 (0.08)	29.1	A, S
	662	0.39 (0.08)	30.4	A, S, E
	669	0.42 (0.07)	35.4	BMI, A, S
Adiponectin	639	0.35 (0.09)	10.1	A, S
	538	0.30 (0.09)	6.9	A, S, C

*Note.- All heritability estimates are significantly different from zero at P-value $< 10^{-5}$.

[a] Number of total phenotyped individuals in the heritability analysis.

[b] Significant covariates (P-value < 0.1) kept in polygenic model. A=age, S=sex, F=farm work, E=education, C=cigarette smoking.

Table 4.14: Summary of SOLAR/LOKI multipoint linkage analyses for adiposity-related phenotypes in adults from Samoa (with max LOD score ≥ 1.5).

Trait	cytogenetic position	Flanking marker/s	Range(cM)[a]	LOD score (range)[b]	Covariates[c]
BMI	13q31.3	D13S265	91.8	1.87 (1.81-1.95)	A, S
				2.09(2.03-2.17)	A, S, E
%BFAT	4q22.1	D4S414	115.3-116.3	**2.09 (2.03-2.19)**	A, S
	7p14.3	D7S484	60.0	**2.19 (1.92-2.24)**	A, S
	9p22.2-p21.3	D9S157, D9S171	46.1	1.76 (1.74-1.79)	A, S
	13q31.3	D13S265	91.8	1.62 (1.59-1.64)	A, S
Leptin	13q31.3	D13S265	91.8	**2.30 (2.23-2.37)**	A, S, E, C
ABDCIR	6q24.1-q25.2	D6S308, D6S441	175.5-177.5	1.58 (1.53-1.61)	A, S
			175.5-176.5	1.55 (1.51-1.58)	A, S, E
	9p22.3-p22.2	D9S285, D9S157	38.3	1.83 (1.80-1.85)	A, S
		D9S285	33.4-34.4	**2.14 (2.10-2.18)**	A, S, E
	12p13.31	D12S99	21.3-22.3	1.66 (1.65-1.68)	A, S
			15.2-16.3	1.55 (1.57-1.60)	A, S, E
	13q31.3	D13S265	92.8-93.8	1.54 (1.48-1.74)	A, S
			91.8-92.8	1.66 (1.61-1.87)	A, S, E
Adiponectin	2q13	D2S160	134.4-135.2	**1.96 (1.92-2.00)**	A, S, C
	8q12.2	D8S260	88.9	1.87 (1.85-1.89)	A, S
	18q22.3	D18S1161	121.0	1.81 (1.78-1.84)	A, S, C
	19q12	D19S414	59.3	1.87 (1.80-1.91)	A, S
				1.72 (1.64-1.77)	A, S

Note. - Table continued on the next page.

Table 4.14: Continued.

Trait	cytogenetic position	Flanking marker/s	Range(cM)[a]	LOD score (range)[b]	Covariates[c]
Leptin, adjusted by BMI	19q12-q13.13	D19S414, D19S220	63.4-64.4	**2.03 (2.00-2.05)**	A, S
ABDCIR, adjusted by BMI	3q21.1	D3S1267	146.3-147.3	1.50 (1.49-1.51)	A

Note. - LOD scores (mean, range) are derived from 10 independent Solar/Loki runs. Regions showing suggestive (LOD $\geq$ 1.9) linkage are highlighted in bold.

[a] Range of the locations of the maximum LOD scores in centimorgans (Haldane) from p terminus.

[b] Mean of maximum LOD scores and their ranges.

[c] Significant covariates included in the linkage model. A=age, S=sex, F=farm work, E=education, C=cigarette smoking.

Table 4.15: Genetic (ρ_g) and environmental (ρ_e) correlations between selected adiposity-related traits in adults from Samoa

Trait pairs	$\rho_g \pm s.e$	$\rho_e \pm s.e$
BMI-%BFAT	0.87 ± 0.04	0.92 ± 0.02
BMI-ABDCIR	0.91 ± 0.02	0.92 ± 0.01
BMI-Leptin	0.87 ± 0.05	0.76 ± 0.04
%BFAT-ABDCIR	0.85 ± 0.05	0.86 ± 0.03
%BFAT-Leptin	096 ± 0.08	0.71 ± 0.05
ABDCIR-Leptin	0.82 ± 0.07	0.72 ± 0.04
Adiponectin-ABDCIR	-0.48 ± 0.15	-0.27 ± 0.09
Adiponectin-Leptin	$-0.31 \pm 0.22^*$	-0.23 ± 0.10
Adiponectin-%BFAT	$-0.42 \pm 0.22^*$	-0.22 ± 0.12
Adiponectin-BMI	$-0.37 \pm 0.16^*$	-0.36 ± 0.09

Note. - BMI, %BFAT and adiponectin were adjusted for age and sex; ABDCIR was adjusted for age, sex, and education; Leptin is adjusted for age, sex, farm work, education, and cigarette smoking.

* Not significantly different from zero at P-values > 0.05.

Table 4.16: Summary of bivariate multipoint linkage analyses for chromosomes 9, and 13 and selected pairs of adiposity-related phenotypes in adults from Samoa.

Phenotype pairs	cytogenetic position	cM[a]	Closest marker	Bivariate LOD score	P-value[b]	LOD^{c}_{eq} score	ρ^{d}_{q}	P-value[e]	P-value[f]
BMI-%BFAT	9p22.2-p21.3	46.1	D9S171	2.79	5.7×10^{-4}	2.30	0.99	0.42	5.6×10^{-3}
BMI-ABDCIR	9p22.2-p21.3	41.2	D9S157	2.35	1.5×10^{-3}	1.88	1.00	0.50	2.4×10^{-3}
%BFAT-ABDCIR	9p22.2-p21.3	43.2	D9S157	3.10	2.9×10^{-4}	2.59	0.95	0.13	1.1×10^{-3}
BMI-%BFAT	13q31.3	91.8	D13S265	2.64	7.8×10^{-4}	2.15	0.97	0.16	$< 10^{-6}$
BMI-ABDCIR	13q31.3	91.8	D13S265	1.96	4.0×10^{-4}	1.52	1.00	0.50	2.6×10^{-3}
BMI-Leptin	13q31.3	91.8	D13S265	2.58	9.2×10^{-4}	2.10	0.98	0.31	8.8×10^{-4}
ABDCIR-%BFAT	13q31.3	91.8	D13S265	2.53	1.0×10^{-4}	2.05	0.99	0.40	1.3×10^{-3}
ABDCIR-Leptin	13q31.3	91.8	D13S265	2.33	1.6×10^{-4}	1.87	0.99	0.43	1.9×10^{-3}
%BFAT-Leptin	13q31.3	91.8	D13S265	2.69	7.4×10^{-4}	2.21	0.95	0.35	1.9×10^{-3}

Note. - BMI and %BFAT were adjusted for age and sex; ABDCIR was adjusted for age, sex, and education; Leptin is adjusted for age, sex, education, and cigarette smoking. Results are based on one particular run.Results are based on one SOLAR/LOKI run.

[a] Distance in centi Morgans (Haldane) from p-terminus.

[b] Asymptotic P-value, under the null hypothesis that the likelihood-ratio statistic (2ln10*bivariate LOD score) is distributed as a $\frac{1}{4}\chi^2_2 : \frac{1}{2}\chi^2_1 : \frac{1}{4}\chi^2_0$ mixture.

[c] LODeq is the equivalent univariate LOD score (df =1) corresponding to the reported bivariate LOD score with 2 df.

[d] ρ_q is the correlation due to QTL effects.

[e] P-value of the test for complete pleiotropy.

[f] P-value of the test for no coincident linkage.

Table 4.17: Summary of SOLAR/LOKI multipoint linkage results using nuclear pedigrees for adiposity-related phenotypes in adults from Samoa (with max LOD score ≥ 1.5).

Structure	Traits	cytogenetic position	Closest marker	cM[a]	Maximum multipoint LOD score	Covariates[b]
Nuclear	%BFAT	11q13.2	D11S987	87.4	**2.22**	A, S
		12q23.1	D12S346	126.3	**2.18**	A, S
	ABDCIR	9p22.2-p21.3	D9S157	41.2	**1.99**	A, S
				40.2	**1.92**	A, S, E
		13q31.3	D13S265	88.6	1.67	A, S
				90.8	1.71	A, S, E
	Adiponectin	19q12	D19S414	59.3	1.71	A, S
		20p13	D20S889	12.7	1.64	A, S

Note. - LOD scores were derived from one particular SOLAR/LOKI run. Regions showing suggestive (LOD ≥ 1.9) linkage are highlighted in bold.

[a] Genomic location of the maximum LOD score in centi Morgan (Haldane) from p terminus.

[b] Significant covariates included in the linkage model. A=age, S=sex, F=farm work, E=education, C=cigarette smoking.

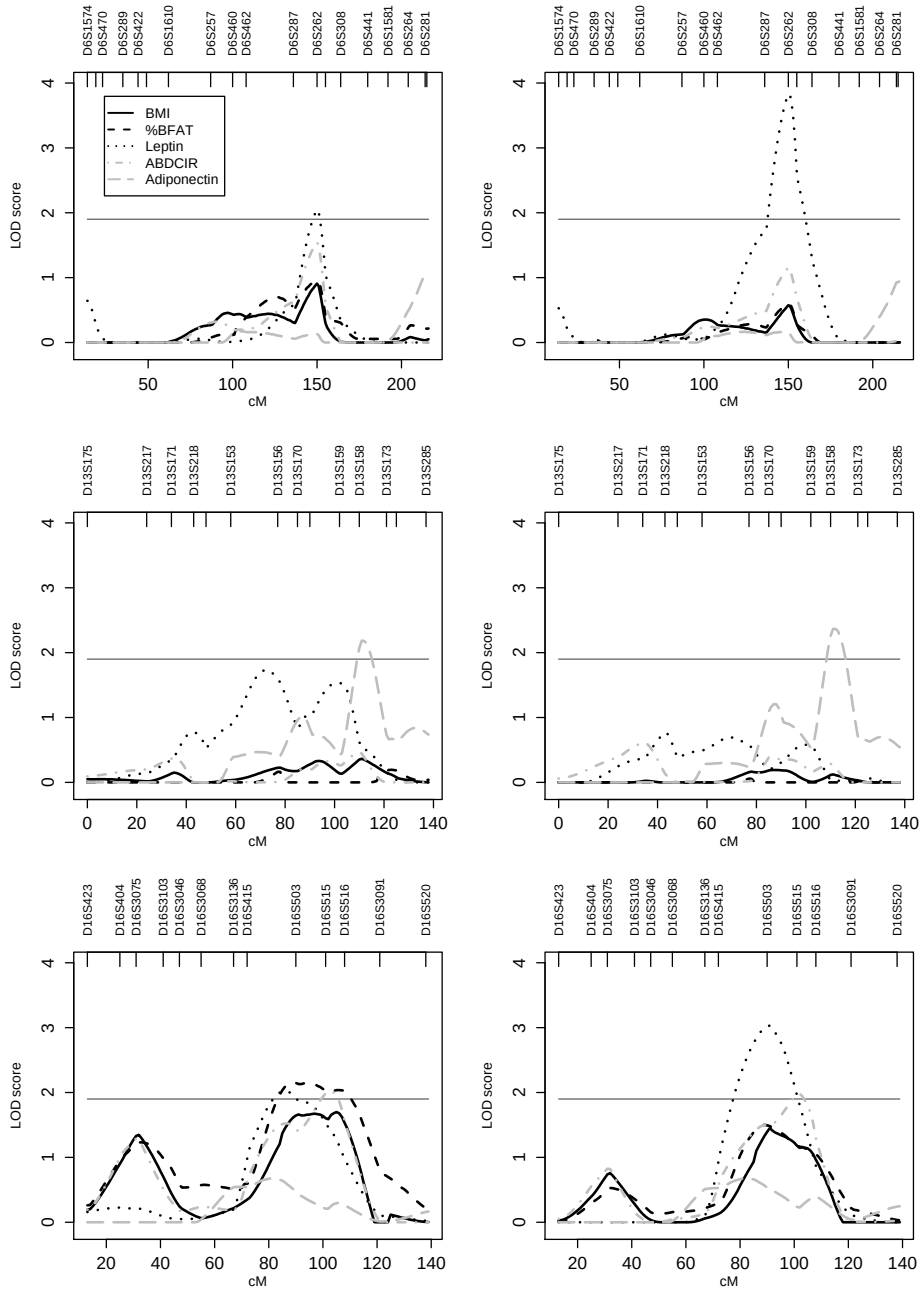
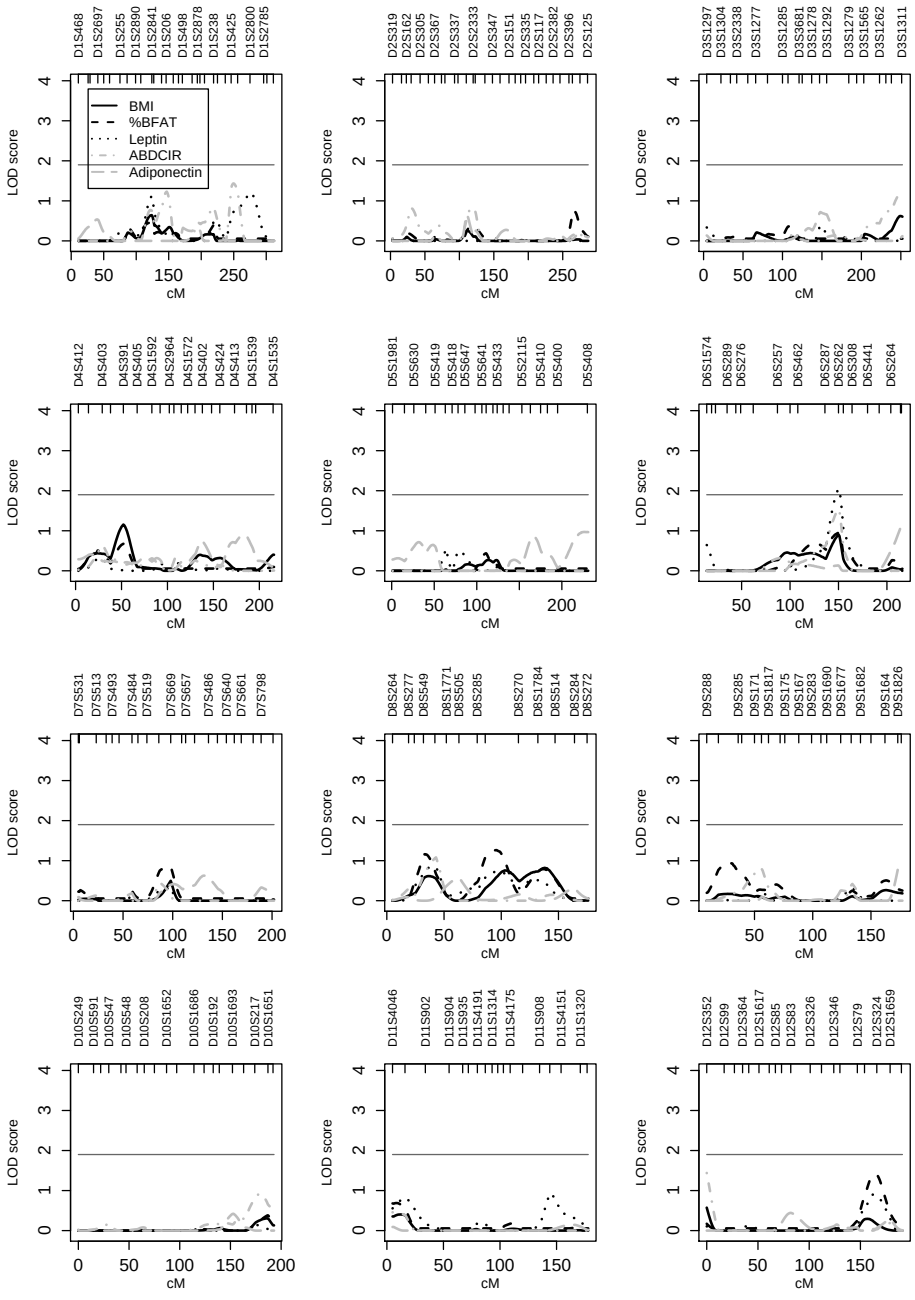


Figure 4.1: The multipoint LOD score results for chromosome 6 (top), chromosome 13 (middle) and chromosome 16 (bottom) for adiposity-related phenotypes (American Samoa). In graphs on the left, BMI and leptin were adjusted for sex; %BFAT, ABDCIR and adiponectin were adjusted for age and sex. In graphs on the right, BMI was adjusted for sex, farm work, and education; %BFAT, ABDCIR and adiponectin were adjusted for age, sex, farm work and education; leptin was adjusted for sex, farm work, education and smoking.


BMI
%BFAT
Leptin
ABDCIR
Adiponectin
LOD score
cM

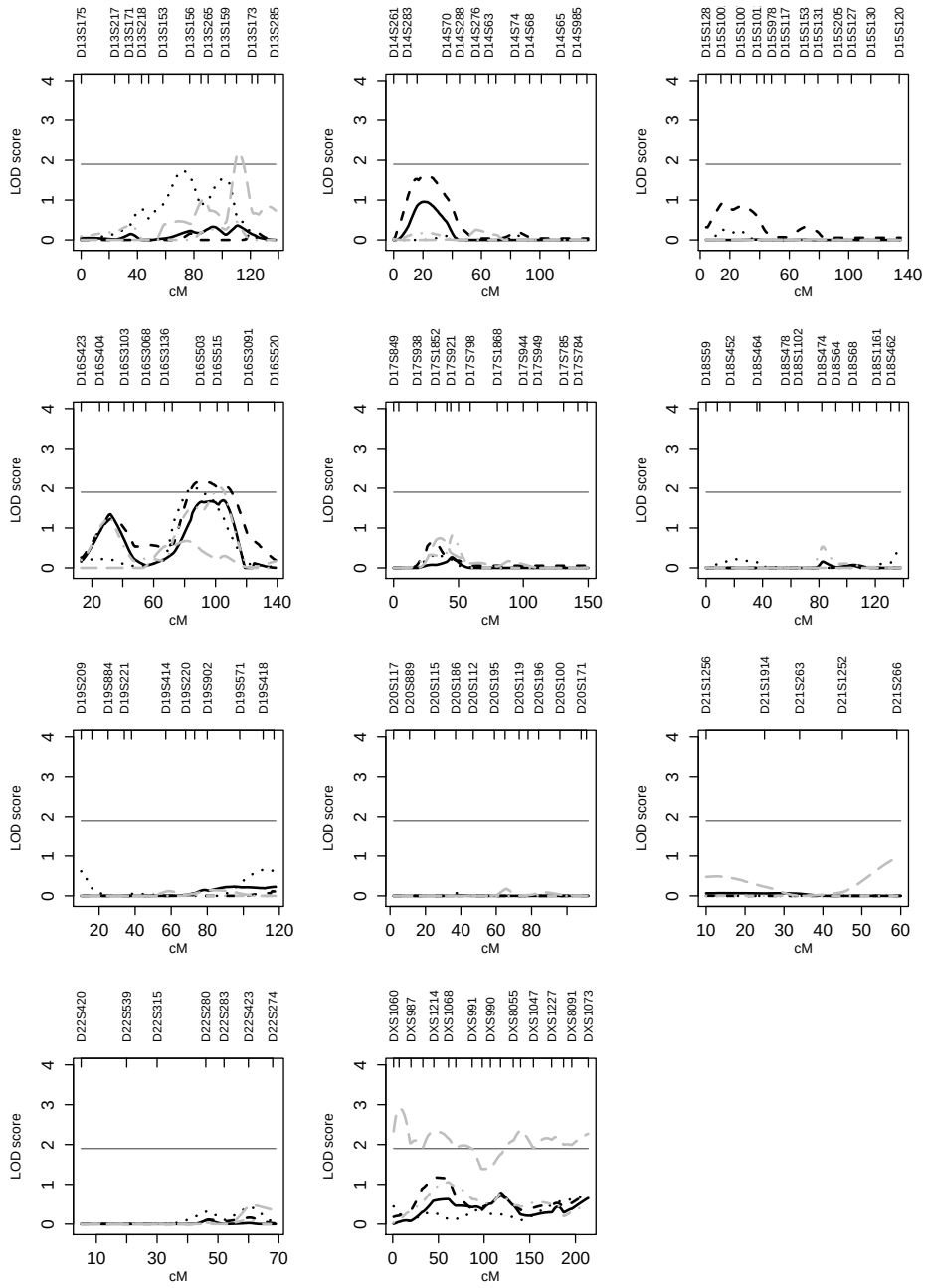
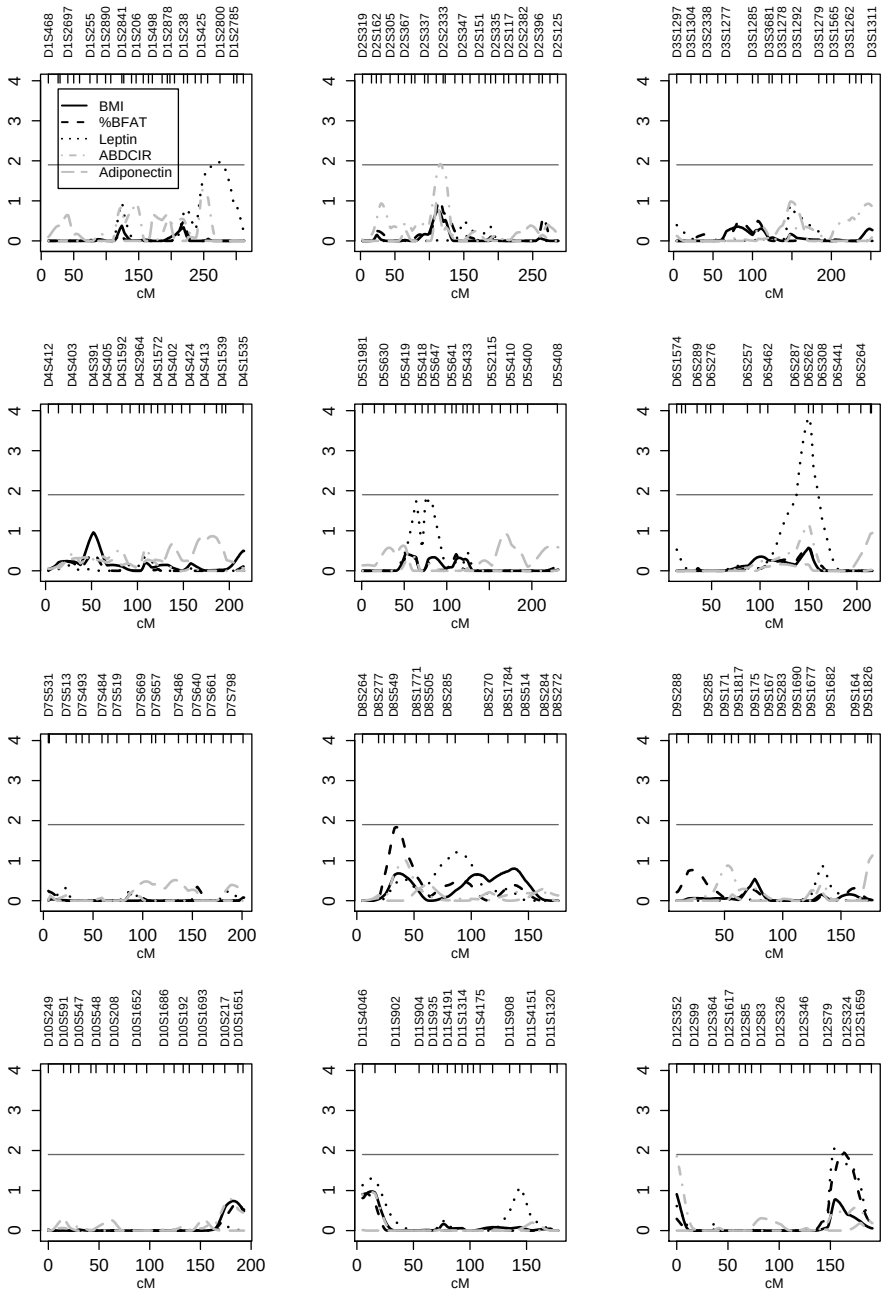


Figure 4.2: Results of genome scan for five adiposity-related phenotypes (American Samoa). For chromosomes 1-22, BMI and leptin were adjusted for sex; %BFAT, ABDCIR and adiponectin were adjusted for age and sex; for the X chromosome, results from Model 1 were plotted (see Table 4.11), and all phenotypes were adjusted for age and sex.


D1S468 D1S2697 D1S255 D1S2890 D1S2841 D1S206 D1S498 D1S2878 D1S238 D1S425 D1S2800 D1S2785
BMI
%BFAT
Leptin
ABDCIR
Adiponectin
D2S319 D2S162 D2S305 D2S367 D2S337 D2S2333 D2S347 D2S151 D2S335 D2S117 D2S2382 D2S396 D2S125
D3S1297 D3S1304 D3S2338 D3S1277 D3S1285 D3S3681 D3S1278 D3S1292 D3S1279 D3S1565 D3S1262 D3S1311
D4S412 D4S403 D4S391 D4S405 D4S1592 D4S2964 D4S1572 D4S402 D4S424 D4S413 D4S1539 D4S1535
D5S1981 D5S630 D5S419 D5S418 D5S647 D5S641 D5S433 D5S2115 D5S410 D5S400 D5S408
D6S1574 D6S289 D6S276 D6S257 D6S462 D6S287 D6S262 D6S308 D6S441 D6S264
D7S531 D7S513 D7S493 D7S484 D7S519 D7S669 D7S657 D7S486 D7S640 D7S661 D7S798
D8S264 D8S277 D8S549 D8S1771 D8S505 D8S285 D8S270 D8S1784 D8S514 D8S284 D8S272
D9S288 D9S285 D9S171 D9S1917 D9S175 D9S167 D9S283 D9S1690 D9S1677 D9S1682 D9S164 D9S1826
D10S249 D10S591 D10S547 D10S548 D10S208 D10S1652 D10S1686 D10S192 D10S1693 D10S217 D10S1651
D11S4046 D11S902 D11S904 D11S935 D11S4191 D11S1314 D11S4175 D11S908 D11S4151 D11S1320
D12S352 D12S99 D12S364 D12S1617 D12S85 D12S83 D12S326 D12S346 D12S79 D12S324 D12S1659
cM

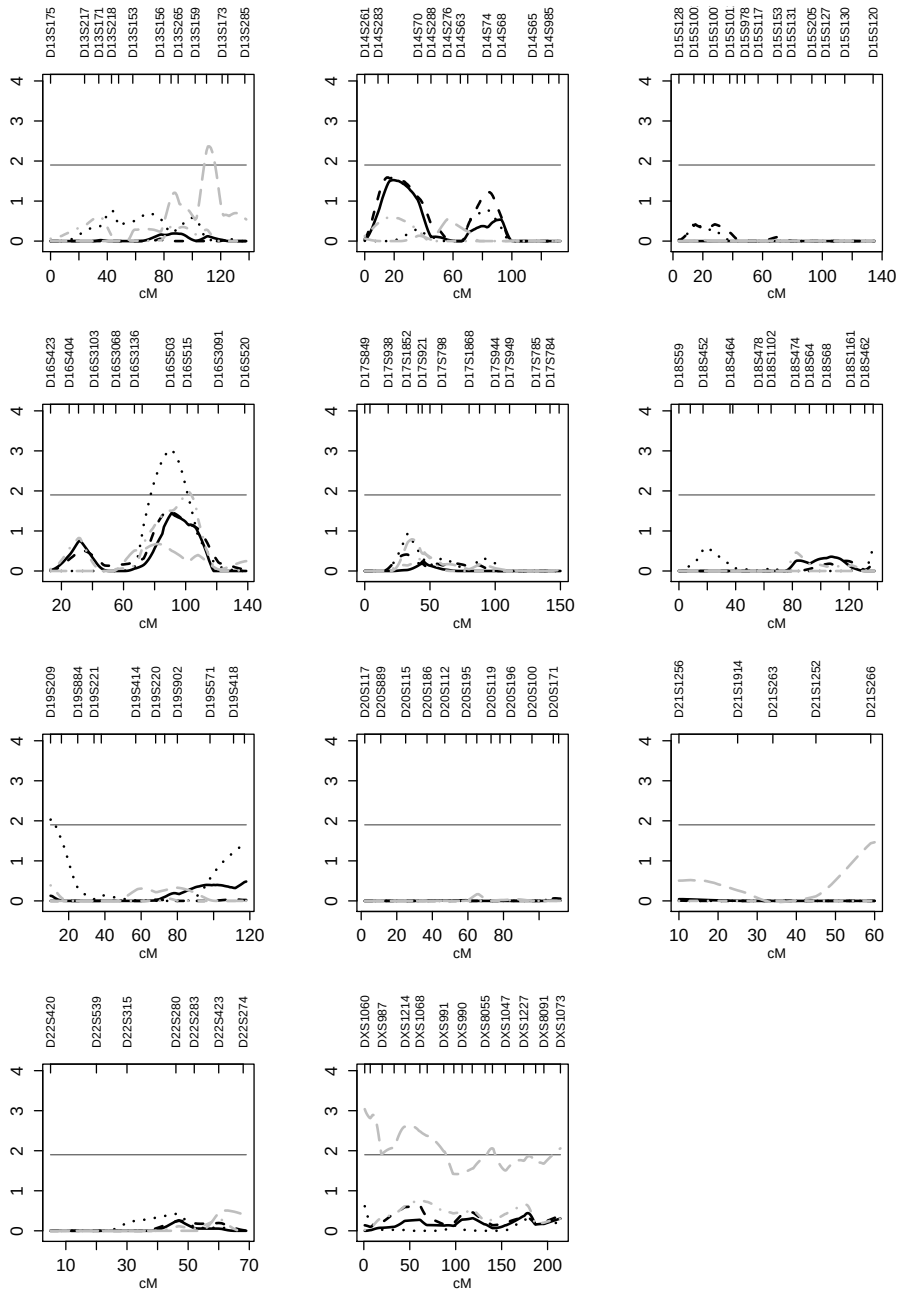


Figure 4.3: Results of genome scan for five adiposity-related phenotypes (American Samoa). For chromosomes 1-22, BMI was adjusted for sex, farm work, and education; %BFAT, ABDCIR and adiponectin were adjusted for age, sex, farm work and education; leptin was adjusted for sex, farm work, education and smoking; for the X chromosome, results from Model 1 were plotted (see Table 4.11), and all phenotypes were adjusted for age, sex, education, farm work and smoking.

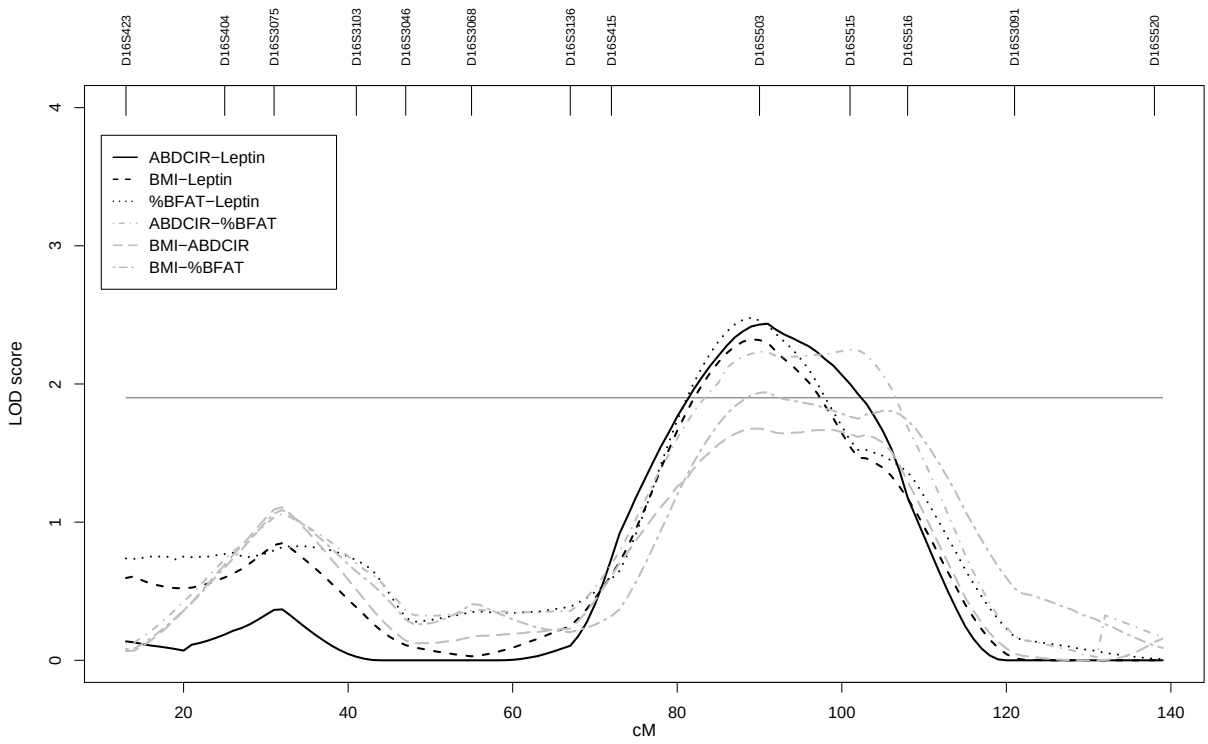


Figure 4.4: The bivariate LOD_{eq} score profiles of adiposity-related phenotype pairs for chromosome 16 (American Samoa). BMI was adjusted for sex; %BFAT was adjusted for age, sex; and ABDCIR were adjusted for age, sex, farm work, and education; Leptin was adjusted for sex, farm work, education, and cigarette smoking.

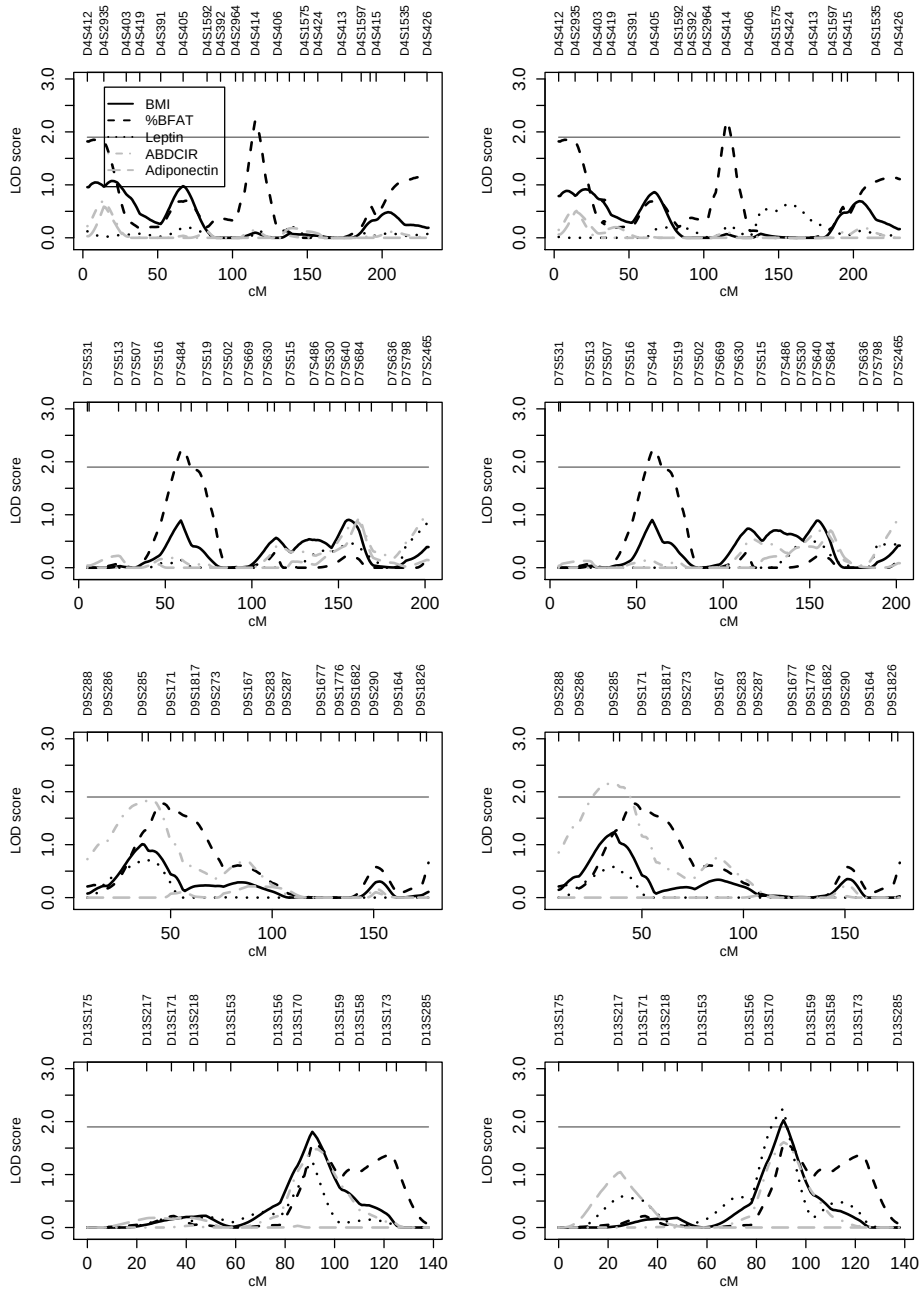
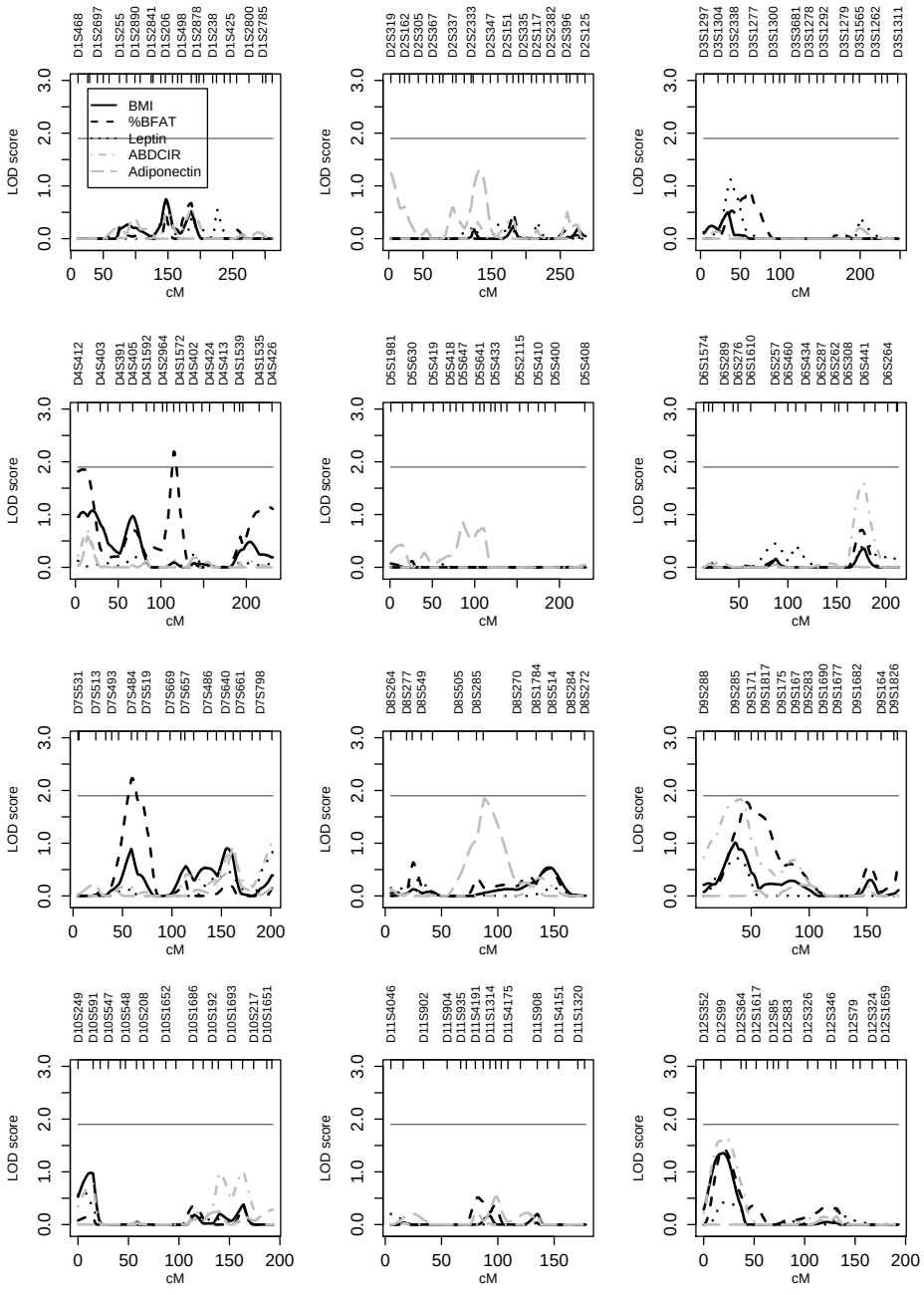


Figure 4.5: The multipoint linkage results from one single SOLAR/LOKI run for chromosome 4, 7, 9, and 13 for five 'primary' adiposity-related phenotypes (Samoa). In graphs on the left, all phenotypes were adjusted for age and sex. In graphs on the right, BMI and ABDCIR were adjusted for age, sex, and education; %BFAT was adjusted for age and sex; Adiponectin was adjusted for age, sex, and smoking; Leptin was adjusted for age, sex, education and smoking.


BMI
%BFAT
Leptin
ABDCIR
Adiponectin
LOD score
cM

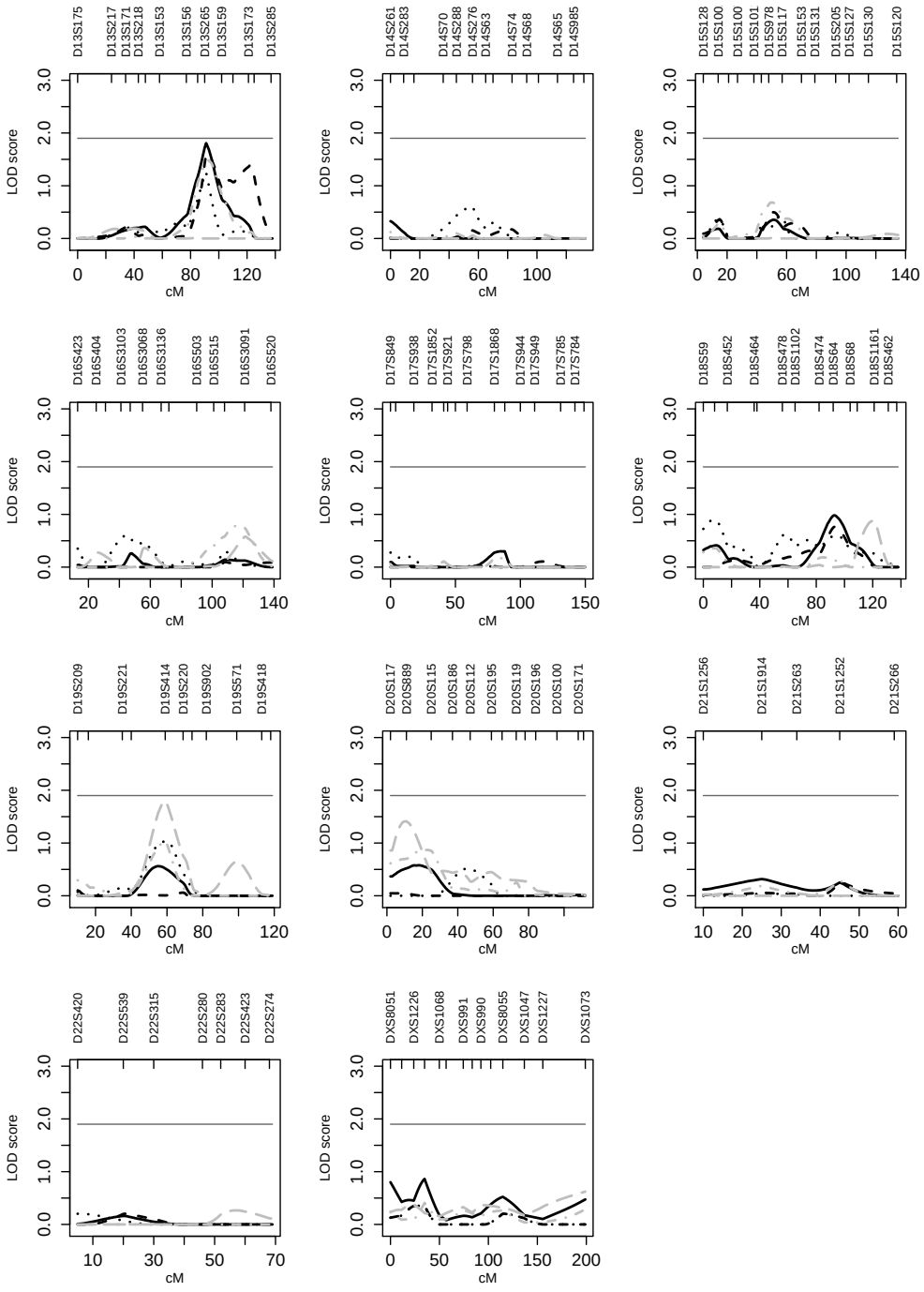
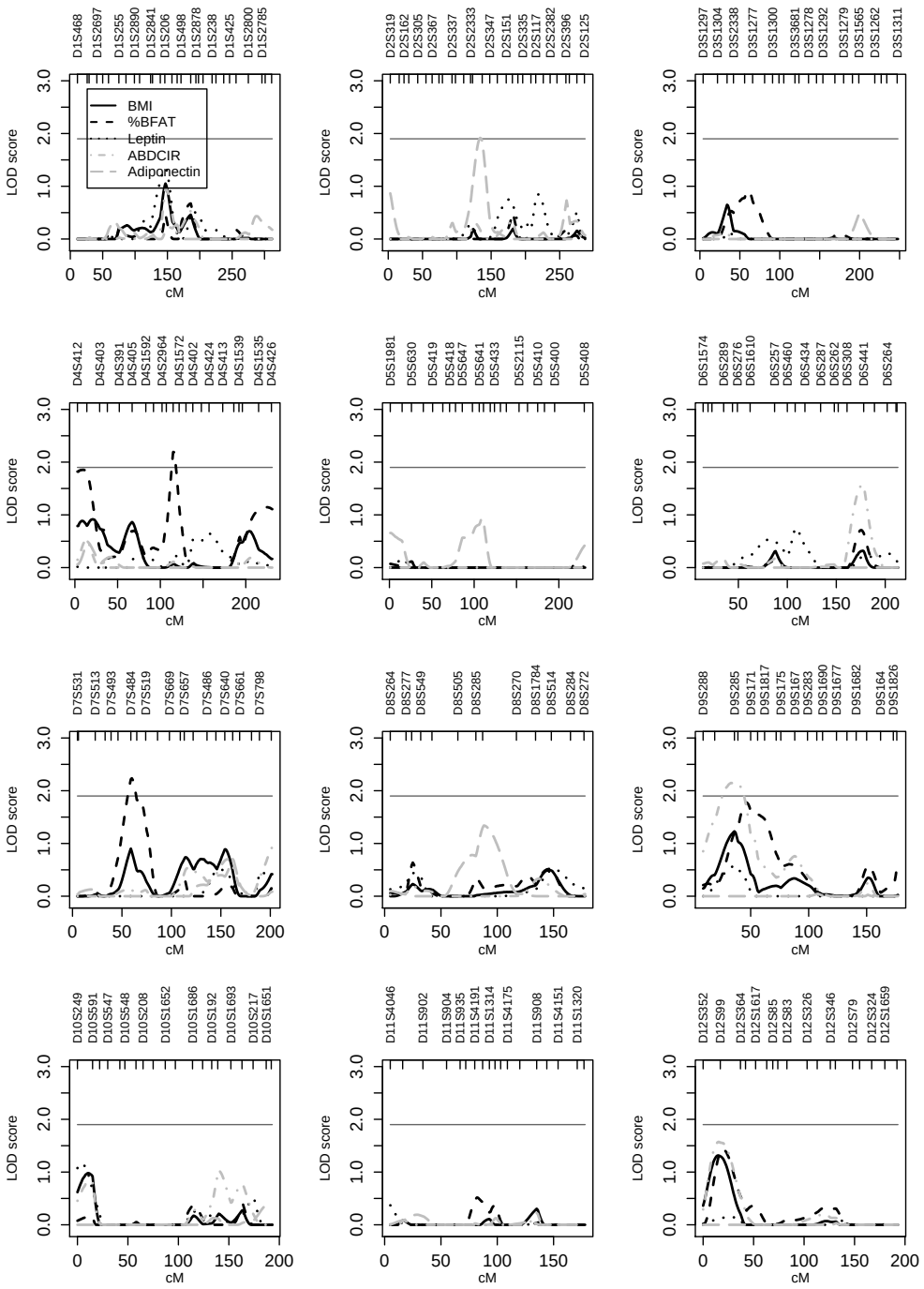


Figure 4.6: Genome scan results for five adiposity-related phenotypes from one single SOLAR/LOKI run (Samoa). For the X chromosome, results from Model 3 (see Table 4.11) were plotted. All phenotypes were adjusted for age and sex.


BMI
%BFAT
Leptin
ABDCIR
Adiponectin
LOD score
cM

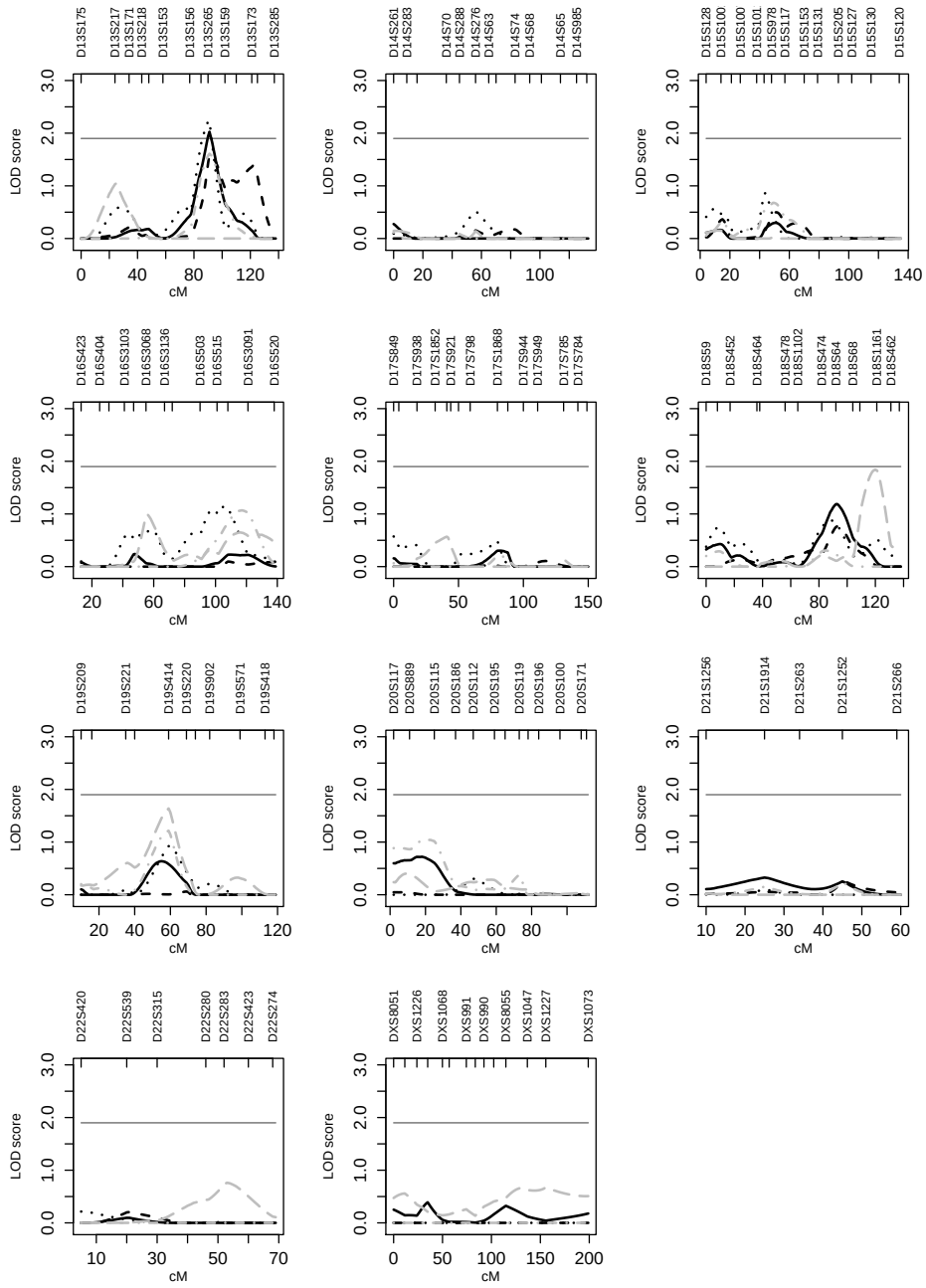


Figure 4.7: Genome scan results for five adiposity-related phenotypes from one single SOLAR/LOKI run (Samoa). For chromosome 1-22, BMI and ABDCIR was adjusted for age, sex, and education; %BFAT was adjusted for age and sex; Adiponectin were adjusted for age, sex, and smoking; Leptin was adjusted for age, sex, education and smoking; for the X chromosome, results from Model 4 were plotted (see Table 4.11), and all phenotypes were adjusted for age, sex, education, farm work and smoking.

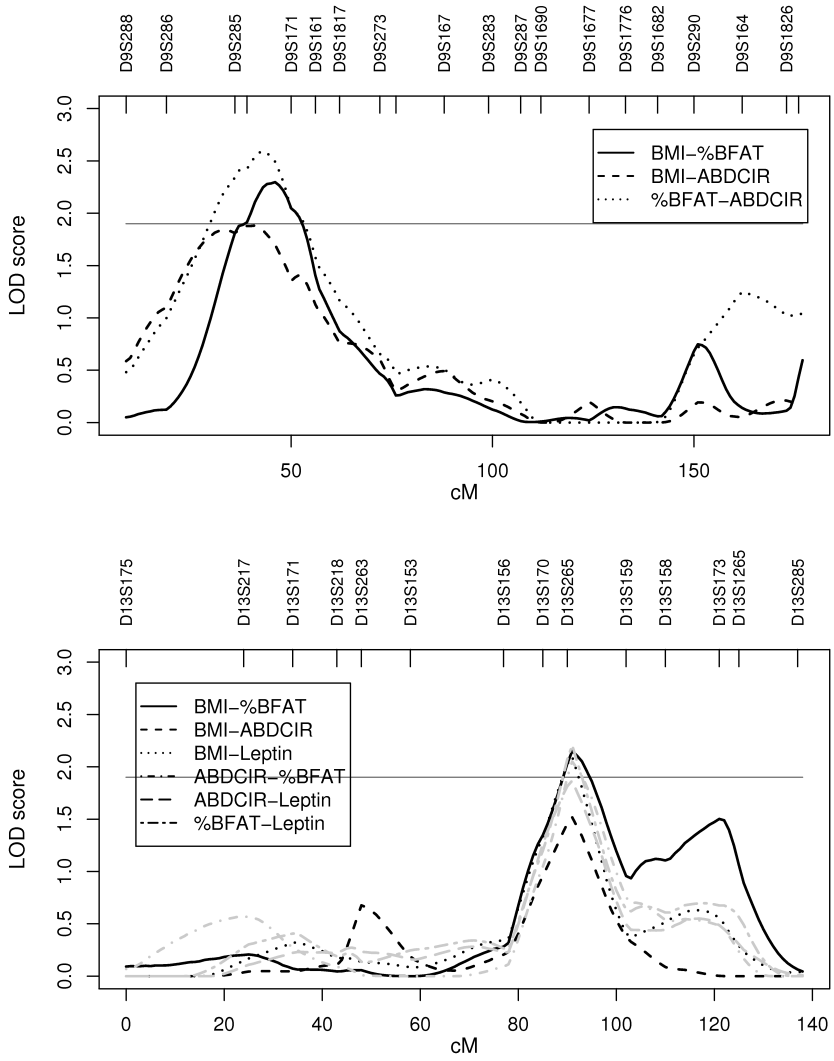


Figure 4.8: The LOD_{eq} scores of selected pairs of adiposity-related phenotypes from one single SOLAR/LOKI run for chromosome 9 and 13 (Samoa). BMI and %BFAT were adjusted for age and sex; ABDCIR was adjusted for age, sex, and education; Leptin is adjusted for age, sex, education, and cigarette smoking.

5.0 SEX-SPECIFIC LINKAGE ANALYSIS: SEX-AVERAGED GENETIC MAPS VS. SEX-SPECIFIC GENETIC MAPS

In this chapter, we carry out a preliminary simulation study to investigate the bias of multipoint linkage analysis arising from map misspecification. We also report sex-specific susceptibility loci for adiposity-related phenotypes from our linkage analyses of sex-specific subsets of our American Samoan and Samoan data.

5.1 INTRODUCTION

In the past decade, genomewide linkage analysis has been widely used in identifying the genetic variation underlying single-gene disorders, common human diseases, and other complex traits in which multiple genetic and environmental factors interact to influence disease risk [147]. Typically genome scans in the literature have been pursued in combined-sex samples assuming there are common susceptibility loci segregating in both sexes, however, few scans have been pursued in sex-specific subsets. One possible reason is that limiting the computation to same-sex pairs in linkage analysis would cause a large reduction in power [149]. However, failing to model the sex-specific architecture in genome-wide screens might hamper detection of susceptibility loci for complex traits [148]. Without modeling for sex-specific differences, autosomal genes that have sex-specific action or interaction may be hard to detect in the regular genome scans [41].

In the Framingham study, Atwood et al. (2006) [149] found that sex-specific effects of chromosomal regions on BMI are common. A few other research groups have also reported significant linkage signals in their linkage analyses of sex-specific subsets [41], [150]. Despite

different types of traits (quantitative or qualitative) being investigated, these reports share similar findings, that is, they all observed more significant linkage signals in the sex-specific subsets, and less or modest signals in the larger combined-sex data sets. For example, in their genome-wide linkage scan for 17 quantitative traits in the Hutterites, Weiss et al. (2006) [41] reported that only two of 12 total sex-specific significant linkage signals were detected in their combined analysis. All those findings are reasonable under the strong assumption that if there is a sex-specific effect, it will have to be quite strong to be detected in a relatively small subset [149].

Despite their important implications for mapping complex trait (sex-specific) genes, strong conclusions made by those studies might be weakened by their use of sex-averaged maps instead of sex-specific maps in their multipoint linkage analyses. It is known that recombination rates differ between male and female meiosis and on average the genetic map of females is about 90% longer than that of males, although in some regions of the genome recombination rates in male meiosis exceed those in female meiosis [151]. Since multipoint linkage analysis usually assumes a known genetic map, map misspecification might compromise the estimation and testing procedures in the linkage analysis [38], [39], [40].

In the following section, we begin to introduce a preliminary simulation study to investigate the bias of multipoint linkage analysis arising from map misspecification.

5.2 METHODS

5.2.1 Disease model and data

We first assume a biallelic dominant model at the disease locus, and the disease-causing allele D is very rare with frequency q (the frequency of d is p). The probability of the disease in an individual with i copies of the D allele is denoted by penetrance parameter f_i, for $i = 0, 1, 2$. The population risk (prevalence) is given by $K = f_0p^2 + 2f_1pq + f_2q^2$. We also assume differential sex-specific penetrances between males and females, that is, $f_{male} = Rf_{female}$, so

the population prevalence in males is proportional to the population prevalence in females with the same coefficient of R ($K_{male} = RK_{female}$).

We assume two markers (M_1, M_2) unlinked with a trait locus (d) with fully informative meioses. The recombination rate between M_1 and locus d is 0.5 in both females and males, with sex-specific recombination rates between M_1 and M_2 (θ_{male} vs. θ_{female}, Figure 5.1). Haldane's map function was used convert between recombination rates and map distances. Sex-average map distance was calculated by average two sex-specific map distances in males and females.

5.2.2 Simulation procedure

The purpose of our simulation is to compare the false-positive rates of multipoint qualitative linkage analysis using sex-specific genetic maps to those using sex-averaged maps. Our simulation work follows a step-wise procedure (10,000 iterations):

1. Set up replicate(s) with four types of nuclear family in ratios of $R^3 : R^0 : R^1 : R^2$. (Figure 5.2). The power coefficient is equal to the number of affected males in each family. One typical replicate in our simulation work consists of $64 : 1 : 4 : 16$ of those four families from left to right.

2. Use SLINK [152] to simulate two markers under sex-specific maps (Figure 5.1). Different recombination ratios between females and males ($\theta_{female} : \theta_{male}$) are assumed. A female-to-male distance ratio of 8.68 was used in the simulation.

3. Use Merlin [153] to analyze the whole data set, male-only subset, and female-only subset, respectively (S_{all}, S_{pair} statistics [154]), using both sex-averaged maps and sex-specific maps. In each category, we record the total number of LOD scores that are great than 1, 2, and 3, which will be used to calculate the probability $P(\text{LOD} > 1, 2, 3)$.

5.3 RESULTS

5.3.1 Simulations

Table 5.1 displays the false positive rates calculated by S_{all} statistic using sex-specific maps and sex-averaged genetic maps. Table 5.2 displays the false positive rates calculated by S_{pair} statistic using sex-specific maps and sex-averaged genetic maps. Two statistics S_{pair} and S_{all} give very similar results in different situations. The false positive rates (within each cell) calculated from using sex-specific genetic maps are pretty similar to those from using sex-averaged maps. When sample size is small (*e.g.*, one replicate of 85 families simulated, tables 5.1, 5.2), the false positive rates in female subset are much less than those in male subset. However, when sample size is increased (4 replicates simulated), most of them are pretty similar to each other, as well as what are seen in whole date set (tables 5.1, 5.2).

5.3.2 Sex-specific susceptibility loci for adiposity related phenotypes in adults from Samoan archipelago

We performed autosomal variance component linkage analyses on sex-specific subsets of American Samoan sample and Samoan sample, respectively. The approach to sex-specific genome scans is simple, the phenotype values for individuals of the opposite sex were set as missing data. Sex-average genetic maps were used in all analyses. We did not perform X-linked variance components analyses on sex-specific subsets as we did before in chapter 4.

Before linkage analyses, we screened for significant covariates (P-value < 0.1) to be contained the linkage model and estimated sex-specific heritability for five adiposity traits BMI, %BFAT, ABDCIR, leptin, and adiponectin, which are summarized in both table 5.3 and table 5.4. From table 5.3 we can see that in the American Samoan sample, female-specific heritability of leptin is about $\sim$1.5 times of that estimated in male subset; in the Samoan sample, male-specific heritability of BMI is about $\sim$2 times of that estimated in female subset (table 5.4). At present, it is too early to conclude whether or not those differences are significant. Further statistical testings are thus needed.

Table 5.5 displays all maximum lodscores greater than 1.5 from the sex-specific genome scans on American Samoan sample, in which we identified 13 sex-specific susceptibility loci of suggestive linkage (LOD $\geq$ 1.90). 5 sex-specific linkage signals (highlighted in red and underline, table 5.5) were detected in our combined analyses discussed in chapter 4. Table 5.6 displays all maximum lodscores greater than 1.5 from the sex-specific genome scans on Samoan sample, in which we observed only 4 sex-specific susceptibility loci of suggestive linkage (LOD $\geq$ 1.90), of which one region (4q22.1) was also detected in our combined analyses.

5.4 DISCUSSION

In this chapter we first discussed some simulations regarding bias in linkage analysis arising from genetic map misspecification. If there were substantial difference in the number of informative male and female meioses, previous studies have found that using sex-averaged map in place of the biologically more plausible sex-specific maps could lead to increased false positive rates [40]. However, our results did not indicate inflated false positive rates from using sex-averaged genetic maps instead of sex-specific maps. One possible reason is that the female:male map distance ratio was not substantially misspecified in our simulations, which make it unable to detect modestly increased false-positive rates [39]. We also should have simulated a large number of replicates as in previous studies [39],[40].

In the real data section, we reported a few chromosomal regions with suggestive linkage that may harbor sex-specific genes for adiposity related phenotypes in Samoans (American Samoans vs. Samoans). Some chromosomal regions (*e.g.*, 4q35.1 and 8q12.2 in table 5.5) might harbor a gene that has significant pleiotropic effects on multiple adiposity traits. Some sex-specific linkage regions (highlighted in red and underline, table 5.5 and table 5.6) were previously detected in our combined analyses discussed in chapter 4. For possible positional candidate genes harbored by these chromosomal regions, please refer to the discussion part of chapter 4.

5.5 FUTURE WORK

We have not had time to test whether the heritabilities of adiposity traits in one sex are significantly different from those in the other sex. We will work on this issue soon. We will also prepare sex-specific genetic maps for all the markers typed in our study, and then perform genome-wide screens for our phenotypes using those maps. In doing so, we would be able to assess how much the use of sex-averged genetic maps instead of sex-specific maps would change the linkage signals in our real data.

Table 5.1: Empirical false-positive rates using S_{pair} statistic

Statistic	Threshold[1]	All[2]	Male[3]	Female[4]
S_{pair}_R1	1	0.0341 (0.0325)	0.0347 (0.0331)	0.0009 (0.0021)
	2	0.0038 (0.0030)	0.0041 (0.0031)	0.0000 (0.0000)
	3	0.0000 (0.0000)	0.0001 (0.0001)	0.0000 (0.0000)
S_{pair}_R4	1	0.0331 (0.0304)	0.0319 (0.0304)	0.0402 (0.0379)
	2	0.0029 (0.0024)	0.0037 (0.0031)	0.0012 (0.0018)
	3	0.0004 (0.0006)	0.0003 (0.0004)	0.0000 (0.0000)
S_{pair}_R1_Half	1	0.0340 (0.0306)	0.0324 (0.0304)	0.0000 (0.0004)
	2	0.0029 (0.0031)	0.0027 (0.0030)	0.0000 (0.0000)
	3	0.0004 (0.0002)	0.0002 (0.0002)	0.0000 (0.0000)
S_{pair}_R4_Half	1	0.0293 (0.0275)	0.0290 (0.0264)	0.0341(0.0326)
	2	0.0039 (0.0037)	0.0031 (0.0028)	0.0002 (0.0006)
	3	0.0006 (0.0010)	0.0003 (0.0003)	0.0000 (0.0000)

Note.- Simulation was repeated 10000 times. In each cell, false positive rates in (out of) the parenthesis were calculated using sex-specific (sex-averaged) maps. R1: one replicate (85 families) was simulated. R4: four replicates (340 families) were simulated. Half: missing genotypes for one parent, father or mother depending on which sex-specific subset was being analyzed.

[1] Threshold LOD scores are 1, 2, and 3.

[2] Whole simulated data set.

[3] Male subset.

[4] Female subset.

Table 5.2: Empirical false-positive rates using S_{all} statistic

Statistic	Threshold[1]	All[2]	Male[3]	Female[4]
S_{all}_R1	1	0.0324 (0.0293)	0.0333 (0.0294)	0.0011 (0.0022)
	2	0.0032 (0.0023)	0.0032 (0.0024)	0.0000 (0.0000)
	3	0.0003 (0.0004)	0.0004 (0.0002)	0.0000 (0.0000)
S_{all}_R4	1	0.0301 (0.0273)	0.0296 (0.0263)	0.0359 (0.0380)
	2	0.0025 (0.0020)	0.0021 (0.0019)	0.0010 (0.0016)
	3	0.0002 (0.0003)	0.0004 (0.0005)	0.0000 (0.0000)
S_{all}_R1_Half	1	0.0291 (0.0257)	0.0289 (0.0250)	0.0001 (0.0004)
	2	0.0029 (0.0023)	0.0026 (0.0029)	0.0000 (0.0000)
	3	0.0001 (0.0001)	0.0002 (0.0004)	0.0000 (0.0000)
S_{all}_R4_Half	1	0.0280 (0.0258)	0.0266 (0.0231)	0.0336 (0.0337)
	2	0.0029 (0.0026)	0.0020 (0.0017)	0.0003 (0.0003)
	3	0.0003 (0.0002)	0.0002 (0.0001)	0.0000 (0.0000)

Note.- Simulation was repeated 10000 times. In each cell, false positive rates in (out of) the parenthesis were calculated using sex-specific (sex-averaged) maps. R1: one replicate (85 families) was simulated. R4: four replicates (340 families) were simulated. Half: missing genotypes for one parent, father or mother depending on which sex-specific subset was being analyzed.

[1] Threshold LOD scores are 1, 2, and 3.

[2] Whole simulated data set.

[3] Male subset.

[4] Female subset.

Table 5.3: Sex-specific heritability of adiposity-related phenotypes in adults from American Samoa, adjusted for different covariates.

	Male				Female			
Trait	N[1]	h^2 (s.e.)	Variance (%)[2]	Covariates[3]	N	h^2 (s.e.)	Variance (%)	Covariates
BMI	234	0.30* (0.27)	6.2	F, C	316	0.46 (0.18)	1.6	E
	248[4]	0.50 (0.25)	N/A	N/A	333	0.49 (N/A)	N/A	N/A
%BFAT	234	0.27* (0.22)	40.1	A, F, C	315	0.55 (0.17)	3.4	A, F, E
	248	0.57 (0.21)	3.5	A	332[4]	0.61 (0.15)	0.2	A
Leptin	227	0.35* (0.30)	11.8	A, F, C	313	0.77 (0.14)	5.5	F, E
	242	0.55 (0.25)	1.0	A	329	0.86 (0.13)	1.7	A
ABDCIR	249	0.56 (0.24)	8.1	A, F	315	0.49 (0.18)	6.9	A, C
	249	0.66 (0.23)	3.3	A	334	0.58 (0.17)	1.9	A
Adiponectin	239	0.46 (0.19)	5.0	A	311	0.39 (0.19)	29.8	A, E
	239	0.46 (0.19)	5.0	A	328	0.44 (0.19)	13.8	A

Note.- N/A: not applicable or none or no convergence.

*: Heritability estimates are NOT significantly different from zero, other heritability estimates are significantly different from zero at P-value < 0.05.

[1]Number of total phenotyped individuals in the heritability analysis.

[2]Variance explained by significant covariates.

[3]Significant covariates (P-value < 0.1) kept in polygenic model. A=age, F=farm work, E=education, C=cigarette smoking.

[4] Kurtosis is above normal (>0.8).

Table 5.4: Sex-specific heritability of adiposity-related phenotypes in adults from Samoa, adjusted for different covariates.

	Male				Female			
Trait	N[1]	h^2 (s.e.)	Variance (%)[2]	Covariates[3]	N	h^2 (s.e.)	Variance (%)	Covariates
BMI	330	0.64 (0.16)	7.9	A	339	0.32 (0.14)	N/A	A
	330	0.64 (0.16)	7.9	A	339	0.32 (0.14)	N/A	A
%BFAT	180	0.58 (0.24)	17.9	A, C	204	0.12[4] (0.20)	1.3	A
	210	0.60 (0.20)	19.8	A	204	0.12(0.20)	1.3	A
Leptin	279	0.64 (0.16)	21.9	A, E, C	326	0.20[4] (0.15)	0.5	A
	316	0.64 (0.13)	17.1	A	326	0.20 (0.15)	0.5	A
ABDCIR	328	0.61 (0.15)	29.9	A, E	340	0.24 (N/A)	15.5	A
	331	0.63 (0.15)	28.8	A	340	0.24 (N/A)	15.5	A
Adiponectin[4]	315	0.32 (0.19)	N/A	N/A	258	0.39 (0.23)	4.4	C
	315	0.32 (0.19)	N/A	N/A	324	0.42 (0.17)	5.0	A

Note.- N/A: not applicable or none or no convergence.

*: Heritability estimates are NOT significantly different from zero, other heritability estimates are significantly different from zero at P-value < 0.05.

[1]Number of total phenotyped individuals in the heritability analysis.

[2]Variance explained by significant covariates.

[3]Significant covariates (P-value < 0.1) kept in polygenic model. A=age, E=education, C=cigarette smoking.

[4] Kurtosis is above normal (> 0.8).

Table 5.5: Sex-specific susceptibility loci with suggestive linkage (LOD score ≥ 1.50) in adults from American Samoa.

			American Samoa			
		Cytogenetic	Male		Female	
Trait	Closest marker	position	LOD	Covariates[a]	LOD	Covariates[a]
BMI	D4S1535	4q35.1			1.68	N/A
	D6S460	6q14.1			1.75	N/A
	D7S513	7p21.3			1.60	N/A
	D7S640	7q32.3	1.94	N/A		
	D8S260	8q12.2			**2.01 (2.70)**	E (N/A)
	D9S175	9q21.13			1.87	E
	D16S515	16q23.1			**1.91 (1.91)**	E (N/A)
%BFAT	D4S1535	4q35.1			1.56 (1.86)	A, E, F (A)
	D8S260	8q12.2			1.88 (**2.84**)	A, E, F (A)
	D9S175	9q21.13			1.52	A, E, F
	D11S4046	11p15.5			**2.04**	A
	D16S3103	16p12.3			**2.12**	A
	D16S503	16q21	1.60	A		
Leptin	D6S1574	6p25.1			1.60	A
	D6S292	6q23.3			**1.95**	F, E
	D12S1659	12q24.32	1.70	A, F, C		
	D19S209	19p13.3			**2.65** (1.74)	F, E (A)
	D22S283	22q12.3	**1.90 (2.19)**	A, F, C (A)		

Note.- The table is continued on the next page.

Table 5.5: Continued

			American Samoa			
			Male		Female	
Trait	Closest marker	Cytogenetic position	LOD	Covariates[a]	LOD	Covariates[a]
ABDCIR	D1S206	1p21.2			1.51 (**1.91**)	A, C (A)
	D7S513	7p21.3			1.87 (**1.91**)	A, C (A)
	D12S352	12p13.33			1.54 (**1.90**)	A, C (A)
Adiponectin	D4S1572	4q24			1.78	A
	D4S406	4q25			1.66	A, E
	D5S416	5p15.1			1.67	A
	D13S158	13q33.1	**1.97**	A		
	D13S285	13q34			1.58	A

Note. - LOD scores were derived from one particular SOLAR/LOKI run. Regions showing suggestive (LOD $\geq$ 1.90) linkage are highlighted in bold. Region(s) found in whole sample including males and females are highlighted in and underline (see chapter 4).

[a] Covariates included in the model: A=age, F=farm work, E=education, C=cigarette smoking.

N/A: no significant covariates included in the model.

Table 5.6: Sex-specific susceptibility loci with suggestive linkage (LOD score $\geq$ 1.50) in adults from Samoa.

			Samoa			
		Cytogenetic	Male		Female	
Trait	Closest marker	position	LOD	Covariates[a]	LOD	Covariates[a]
BMI	D6S1581	6q25.3	**1.93**	A		
%BFAT	D4S414	4q22.1	**2.06**	A, C		
	D7S519	7p13	**2.04**	A		
	D10S1686	10q23.1	1.51	A, C		
	D12S364	12p13.1	**2.02**	A		
ABDCIR	D6S1581	6q25.3	1.74	A, E		
Adiponectin*	D5S407	5q11.2			1.56	A
	D10S217	10q26.2			1.65	C

Note. - LOD scores were derived from one particular SOLAR/LOKI run. Regions showing suggestive (LOD $\geq$ 1.90) linkage are highlighted in bold. Region(s) found in whole sample including males and females are highlighted in red and underline (see chapter 4).

[a] Covariates included in the model: A=age, E=education, C=cigarette smoking.

* Kurtosis is above normal ($>$ 0.80).

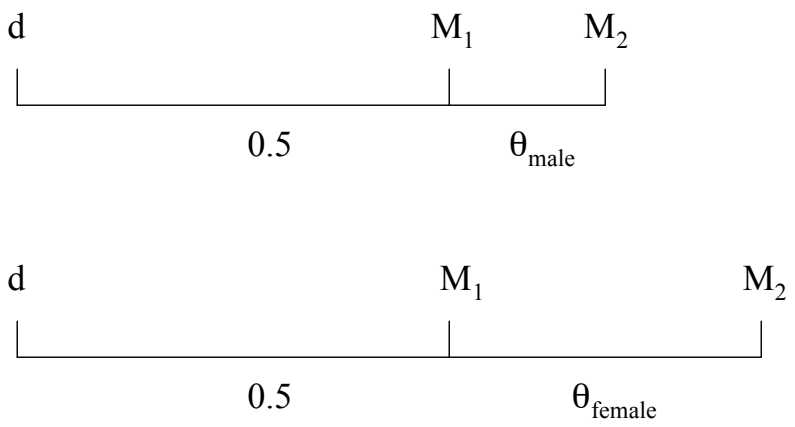


Figure 5.1: Unlinked case. The recombination rates between marker 1 (M_1) and disease locus (d) in males and females are 0.5, the recombination rate between marker 1 (M_1) and marker 2 (M_2) is sex-specific (θ_{male} vs. θ_{female}).

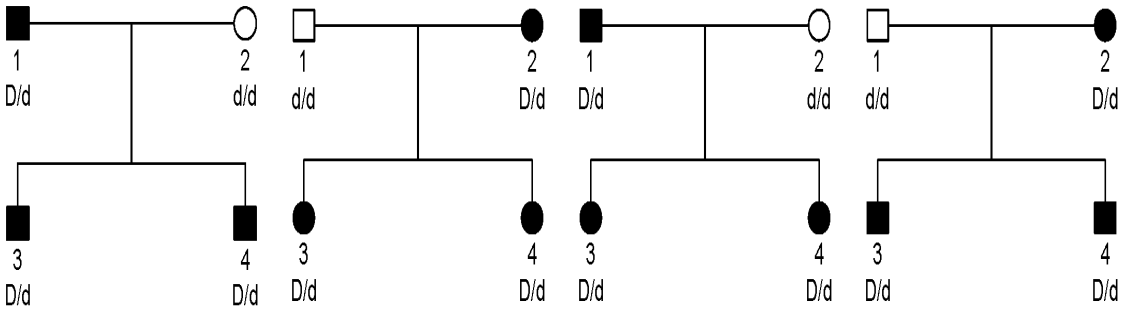


Figure 5.2: Four types of nuclear family in a simulated replicate. Capital D represents a rare disease-causing allele.

6.0 DISCUSSION AND FUTURE WORK

6.1 CONCLUSION

The proposed methods in the first part of this dissertation extend previous work in the following two aspects: **1**) we extended the Li and Sacks (1954) [33] ITO method to ordered genotypes and expanded the application of the method to more contemporary genetic models such as those involving genomic imprinting. **2**) we derived a general equation for calculating any relative-to-relative covariance. Our equation makes it possible to extend the variance components linkage analysis to model genomic imprinting in even inbred population.

Obesity is a typical multi-factorial disease with overwhelming evidence of genetic component effects, yet the role of genetics in obesity is complex, *e.g.*, multiple genes, gene x gene interventions. Given the fact that over 200 obesity QTLs have been reported to be linked with numerous obesity-related phenotypes [65], the genome scan results of adiposity-related traits described in the second part of my dissertation are not novel, however, it still provides novel insights about how different and rapidly changing environments would affect two culturally isolated Samoan populations with a homogenous genetic background. More importantly, our 10 cM linkage analysis in our Samoan pedigrees narrows down the promising chromosomal regions for fine mapping and paves the way for our future association study. A comparison between genetic risk factors for obesity in Samoan populations with those identified in cosmopolitan outbred human groups would help answer the question whether obesity susceptibility genes distribute at similar frequencies across all human groups, which in turn would provide scientists a more broad picture of the genetics of human obesity.

6.2 ISSUES RAISED FROM DATA ANALYSES

There are several theoretical issues raised by our analyses of Samoan data, which consist of:
1. How to efficiently select a list of covariates and adjust for their effects before doing linkage analysis?

In our study, adjusting for the effects of certain covariates intacting with potential susceptibility genetic loci have resulted in increased linkage signals. However, we only considered simple covariates like age, gender and a few other environmental measurements, further investigations of interactions between these effects might be needed. In the research of genetic component of complex disease like obesity here, more "accurate" environemental measurements are needed to help not only more accurately localize the casual genes but also clarify the complex network of gene-environmental interactions underlying the diseases. Since in large genetic epidemiological studies, there are often hundreds or thousands of phenotypic measurements, an efficient agorithm for selection of covariates for study of traits of interest would be highly beneficial.

2. Multivariate linkage analysis for dissection of complex disease.

In our study, several correlated traits have been collected and analyzed independently or at most analyzed in bivariate analyses, which may cause a potential loss in power [155]. Furthermore, any single trait might not explain all the phenotypic variation of the disease and how to best adjust for multiple testing of correlated traits remains a challenge because Bonferroni corrections are obviously overconservative when number of independent tests are large. Critical questions and futher complications also arise when to inteprete the univariate linkage analysis of different phenotypic measurements (*e.g.*, how to explain linkage to the same region with different but correlated traits?) [156].

As an alternative to univariate analysis, many research groups have performed linkage analyses using some composite phenotypes constructed by principle component analysis (PCA) or factor analysis of multiple phenotypic measurements. However, these composite scores were constructed without consideration of the genetic relationship of these measurements, and were fixed throughout the genome, which can not be justified in many situations.

Actually PCA approach is not optimal and has been found to be outperformed by multivariate analysis on a genomewide scale [156].

Extensions of the univariate variance component (VC) model to multivariate models have been described in the literature [22], [23] but have not been applied beyond the bivariate applications [156]. The full multivariate approach proposed by Marlow et al. (2003) [156] shows promise, as it may aid in identifying susceptibility genes in many complex traits on a genomewide scale. However, how can their approach be computationally implemented in pedigrees beyond nuclear families remains a question. Meanwhile, their approach shares the drawback of all multivariate VC linkage methods, which is that all of them are subject to the inherent assumption of multivariate normality [90], which can not be guaranteed in practice (nonnormality of some quantitative traits (like adiposity measurements in our study) are obvious). Very recently a robust regression based method for multivariate linakge analysis was proposed by Wang and Elston (2007) [157]. This model-free method is not subject to normality assumption and its statistic does not follow a complex mixture distribution as some VC-based statistics follow [157].

Although theoretically their method can be applied for any type and any number of traits and any type of pedigree data, Wang and Elston (2007) [157] only evaluated the empirical type I error of their method in a bivariate simulation study (in nuclear families) and then illustrated the use of their method by performing trivariate linkage analysis in nuclear families from the Beaver Dam Eye Study, so it raises the question how this method would perform in dealing with multiple traits in extended pedigrees. Furthermore, it is worth to note that in practice, as in any multivariate linkage analysis, the gains of power by this method will still vary depending on the specific temporal patterns of correlation between multiple traits and sources causing the correlation [97], [156], [157]. Thus further investigations are needed for fully exploring the phenotypic correlations between multiple traits as well as the source of these correlations.

3. Extensions to handle genomic imprinting in variance components linkage analyses. Is there a need for a whole genome-wide scan that takes imprinting in certain chromosomal regions into account?

Because so far only 1% of genes are currently known to be imprinted, Shete and Amos (2002) [28] recommended that one should usually perform genome scans with the usual variance-components method and test for imprinting only if significant evidence for linkage was observed. However, this may reduce the power to detect those genes that are strongly imprinted [28].

6.3 FUTURE WORK

6.3.1 Bayesian linkage analysis of Samoan data

Bayesian analysis has been proposed by many authors for human linkage and association analysis [84], [158]. Compared to traditional genetic analysis methods, Bayesian analysis provides a versatile framework to implementing and interpreting multilocus models, and a natural way to use prior information from other studies. Bayesian analysis also was highly recommended to be applied to avoid problems of model misspecification [159].

In future we will perform a Bayesian genome scan for susceptibility loci of adiposity-related traits in adults from American Samoans and Samoans, which is also our attempt to replicate our initial findings discussed in chapter 4. The computer program LOKI will be used to carry out similar linkage analyses summarized in a recent PNAS paper (Shmultewitz et al. 2006 [160]). Note that linkage results are now expressed as Bayes factors, or odds of posterior probability divided by prior probability [84], [161], [162]; no formal statistical significance like LOD score will be provided. To explain any future linkage signals from our Bayesian analysis of Samoan data using LOKI software, a L-score (*i.e.*, a Bayes factor calculated in 1-cM intervals) $\geq$ 20 will be considered as a strong evidence for linkage, with a L-score $\geq$10 considered as a suggestive evidence for linkage.

6.3.2 A recursive software for computing detailed identity coefficients

The recursive algorithm discussed in the chapter 3 so far does not work for all the detailed identity states. We are still testing new boundary conditions and recursive rules to have

a final functional algorithm. Once it is ready, an accompanying recursive software will be developed and be available by request.

APPENDIX

COVARIANCE BETWEEN INDIVIDUALS I AND J UNDER IMPRINTING

Let the k^{th} allele has population frequency p_k and the ordered genotype $k|l$ has trait value $w_{k|l}$, then we can write

$$w_{k|l} = \alpha_k + \beta_l + \delta_{k|l},$$

where α_k is the additive impact of the maternal allele, β_l is the additive impact of the paternal allele, and $\delta_{k|l}$ is the residual departure from additivity. Under imprinting, the identity $w_{k|l} = w_{l|k}$ does not necessarily hold. No generality is lost if we take the trait mean

$$\sum_k \sum_l w_{k|l} p_k p_l = 0$$

To minimize

$$\sum_k \sum_l \delta^2_{k|l} = \sum_k \sum_l (w_{k|l} - \alpha_k - \beta_l)^2 p_k p_l$$

the partial derivative with respect to α_k was taken, which gives

$$-2 \sum_k \sum_l (w_{k|l} - \alpha_k - \beta_l) p_k p_l = -2 \sum_k \sum_l \delta_{k|l} p_k p_l = 0$$

which is only true if

$$\begin{cases} \sum_k \delta_{k|l} p_k = 0, \sum_l \delta_{k|l} p_l = 0 \\ \sum_k \alpha_k p_k = 0, \sum_l \beta_l p_l = 0 \end{cases}$$

Using the above facts and because

$$\begin{aligned} 0 &= \sum_l \delta_{k|l} p_l \\ &= \sum_l w_{k|l} p_l - \sum_l \alpha_k p_l - \sum_l \alpha_l p_l \\ &= \sum_l w_{k|l} p_l - \alpha_k \end{aligned}$$

we can conclude that $\alpha_k = \sum_l w_{k|l} p_l$. Similarly, $\beta_l = \sum_k w_{k|l} p_k$ also holds.

If X_i and X_j are traits of individuals i and j, since $\mathrm{E}(X_i) = \mathrm{E}(X_j) = 0$, then the covariance is

$$\begin{aligned} Cov(X_i, X_j) &= E(X_i, X_j) \\ &= \delta_1 \sum_k (\alpha_k + \beta_k + \delta_{k|k})^2 p_k + \delta_2 \sum_k \sum_l (\alpha_k + \beta_k + \delta_{k|k})(\alpha_k + \beta_l + \delta_{k|l}) p_k p_l \\ &\quad + \delta_3 \sum_k \sum_l (\alpha_k + \beta_k + \delta_{k|k})(\alpha_l + \beta_k + \delta_{l|k}) p_k p_l \\ &\quad + \delta_4 \sum_k \sum_l (\alpha_k + \beta_l + \delta_{k|l})(\alpha_k + \beta_k + \delta_{k|k}) p_k p_l \\ &\quad + \delta_5 \sum_k \sum_l (\alpha_k + \beta_l + \delta_{k|l})(\alpha_l + \beta_l + \delta_{l|l}) p_k p_l \\ &\quad + \delta_6 \sum_k \sum_l (\alpha_k + \beta_k + \delta_{k|k})(\alpha_l + \beta_l + \delta_{l|l}) p_k p_l \\ &\quad + \delta_7 \sum_k \sum_l \sum_m (\alpha_k + \beta_k + \delta_{k|k})(\alpha_l + \beta_m + \delta_{l|m}) p_k p_l p_m \\ &\quad + \delta_8 \sum_k \sum_l \sum_m (\alpha_k + \beta_l + \delta_{k|l})(\alpha_m + \beta_m + \delta_{m|m}) p_k p_l p_m \\ &\quad + \delta_9 \sum_k \sum_l (\alpha_k + \beta_l + \delta_{k|l})^2 p_k p_l \\ &\quad + \delta_{10} \sum_k \sum_l \sum_m (\alpha_k + \beta_l + \delta_{k|l})(\alpha_k + \beta_m + \delta_{k|m}) p_k p_l p_m \\ &\quad + \delta_{11} \sum_k \sum_l \sum_m (\alpha_k + \beta_l + \delta_{k|l})(\alpha_m + \beta_l + \delta_{m|l}) p_k p_l p_m \\ &\quad + \delta_{12} \sum_k \sum_l (\alpha_k + \beta_l + \delta_{k|l})(\alpha_l + \beta_k + \delta_{l|k}) p_k p_l \end{aligned}$$

$$
\begin{aligned}
&+\delta_{13}\sum_k\sum_l\sum_m(\alpha_k+\beta_l+\delta_{k|l})(\alpha_m+\beta_k+\delta_{m|k})p_kp_lp_m\\
&+\delta_{14}\sum_k\sum_l\sum_m(\alpha_k+\beta_l+\delta_{k|l})(\alpha_l+\beta_m+\delta_{l|m})p_kp_lp_m\\
&+\delta_{15}\sum_k\sum_l\sum_m\sum_n(\alpha_k+\beta_l+\delta_{k|l})(\alpha_m+\beta_n+\delta_{m|n})p_kp_lp_mp_n\\
=\;&(\delta_1+\delta_2+\delta_4+\delta_9+\delta_{10})\sum_k\alpha_k^2p_k+(\delta_1+\delta_3+\delta_5+\delta_9+\delta_{11})\sum_k\beta_k^2p_k\\
&+(2\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+2\delta_{12}+\delta_{13}+\delta_{14})\sum_k\alpha_k\beta_kp_k\\
&+(2\delta_1+\delta_2+\delta_4)\sum_k\alpha_k\delta_{k|k}p_k+(2\delta_1+\delta_3+\delta_5)\sum_k\beta_k\delta_{k|k}p_k\\
&+\delta_1\sum_k\delta_{k|k}^2p_k+\delta_6\sum_k\sum_l\delta_{k|k}\delta_{l|l}p_kp_l\\
&+\delta_9\sum_k\sum_l\delta_{k|l}^2p_kp_l+\delta_{12}\sum_k\sum_l\delta_{k|l}\delta_{l|k}p_kp_l
\end{aligned}
$$

BIBLIOGRAPHY

[1] Olson JM, Witte JS, Elston RC (1999) Genetic mapping of complex traits. Statist Med 18:2961–2981

[2] Borecki IB, Suarez BK (2001) Linkage and association: basic concepts. Adv Genet 42:45–66

[3] Feingold E (2001) Methods for linkage analysis of quantitative trait loci in humans. Theor Popul Biol 60:167–180

[4] Peltonen L, McKusick VA (2001) Genomics and medicine: dissecting human disease in the postgenomic era. Science 291:1224–1229

[5] Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543

[6] Almasy L, Blangero J (1998) Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211

[7] Szatkiewicz JP (2004) Mapping genes for quantitative traits using selected samples of sibling pairs. PhD dissertation, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

[8] Teare MD, Barrett JH (2005) Genetic linkage studies. Lancet 366:1036–1044

[9] Newton-Cheh C, Hirschhorn JN (2005) Genetic association studies of complex traits: design and analysis issues. Mutat Res 573:54–69

[10] Spielman R, McGinnis R, Ewens J (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

[11] Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781-791

[12] Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662

[13] Devlin B, Roeder K (1999) Genomic Control for Association Studies. Biometrics 55:997–1004

[14] Devlin B, Roeder K, Wasserman L (2001) Genomic Control, a New Approach to Genetic-Based Association Studies. Theor Pop Biol 60:155–166

[15] Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

[16] Satten GA, Flanders WD, Yang QH (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466–477

[17] Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504

[18] Epstein MP, Lin X, Boehnke M (2003) A Tobit Variance-Component Method for Linkage Analysis of Censored Trait Data. Am J Hum Genet 72:611–620

[19] Lange K (2002) Mathematical and statistical methods for genetic analysis, 2nd Edition. Springer-Verlag, New York

[20] Jacquard A (1974) The genetic structure of populations. Springer-Verlag, New York

[21] Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc 82:605–610

[22] Almasy L, Dyer TD, Blangero J (1997) Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet Epidemiol 14:953–958

[23] Williams JT, Van Eerdewegh P, Almasy L, Blangero J (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation result. Am J Hum Genet 65:1134–1147

[24] Searl SR (1971) Linear models. John Wiley & Sons, New York

[25] Comuzzie AG, Mahaney MC, Almasy L, Dyer TD, Blangero J (1997) Exploiting pleiotropy to map genes for oligogenic phenotypes using extended pedigree data. Genet Epidemiol 14:975–980

[26] Reik W, Walter J (2001) Genomic imprinting: parental influence on the genome. Nat Rev Genet 2:21–32

[27] Hanson RL, Kobes S, Lindsay RS, Knowler WC (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. Am J Hum Genet 68:951–962

[28] Shete S, Amos CT (2002) Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. Am J Hum Genet 70:751–757

[29] Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait Models: application to mite sensitization. Am J Hum Genet 66:1945–1957

[30] Shete S, Zhou X, Amos CI (2003) Genomic imprinting and linkage test for quantitative-trait loci in extended pedigrees. Am J Hum Genet 73:933–938

[31] Williams JT, Blangero J (1999) Power of variance component linkage analysis to detect quantitative trait loci. Ann Hum Genet 6:545–563

[32] Spencer HG (2002) The correlation between relatives on the supposition of genomic imprinting. Genetics 161:411–417

[33] Li CC, Sacks L (1954) The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 10:347–360

[34] Dai F (2004) Imprinting in variance components-based linkage analysis. Master thesis, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

[35] Dai F, Weeks DE (2006) Ordered genotypes: an extended ITO method and a general formula for genetic covariance. Am J Hum Genet 78:1035–1045

[36] Gillois M (1964) La relation d'identité en génétique. Ann Inst Henri Poincaré B 2:1–94

[37] Lange K (2002) Mathematical and statistical methods for genetic analysis (2nd edition). Springer, New York

[38] Halpern J, Whittemore AS (1999) Multipoint Linkage Analysis. Hum Hered 49:194–196

[39] Daw EW, Thompson EA, Wijsman EM (2000) Bias in multipoint linkage analysis arising from map misspecification. Genet Epidemiol 19:366–380

[40] Fingerlin TE, Abecasis GR, Boehnke M (2006) Using sex-averaged genetic maps in multipoint linkage analysis when identity-by-descent status is incompletely known. Genet Epidemiol 30:384–396

[41] Weiss LA, Pan L, Abney M, Ober C (2006) The sex-specific genetic architecture of quantitative traits in humans. Nat Genet 38:218–222

[42] Richardson WH (1964) Frequencies of genotypes of relatives, as determined by stochastic matrices. Genetica 35:323–354

[43] Campbell MA, Elston RC (1971) Relatives of probands: models for preliminary genetic analysis. Ann Hum Genet 35:225–236

[44] Li W (1998) An exact calculation of the probability of identity-by-descent in two-locus models using an extension of the Li-Sacks' method. Am J Hum Genet 63:A297

[45] Harris DL (1964) Genotypic covariance between inbred relatives. Genetics 50:1319–1348

[46] Sobel E, Sengul H, Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. Hum Hered 52:121–131

[47] Jacquard A (1972) Genetic information given by a relative. Biometrics 28:1101–1114

[48] Nadot R, Vaysseix G (1973) Apparentement et identité. Algorithme du calcul des coefficients d'identité. Biometrics 29:347–359

[49] Santure AW, Spencer HG (2006) Influence of mom and dad: Quantitative genetic models for maternal effects and genomic imprinting. Genetics 173:in press

[50] Karigl G (1981) A recursive algorithm for the calculation of identity coefficients. Ann Hum Genet 5:299-305

[51] Karigl G (1982) A mathematical approach to multiple genetic relationships. Theor Popul Biol 21:379-393

[52] Weeks DE, Valappil TI, Schroedr M, Brown DL (1982) A mathematical approach to multiple genetic relationships. Hum Hered 45:25-33

[53] Dai F, Keighley ED, Sun G, Indugula SR, Roberts ST, Åberg K, Smelser D, Viali S, Tuitele J, Jin L, Deka R, Weeks DE, McGarvey ST (2007) A genome-wide scan for adiposity-related phenotypes in adults from Samoa. In preparation

[54] Dai F, Keighley ED, Sun G, Indugula SR, Roberts ST, Åberg K, Smelser D, Viali S, Tuitele J, Jin L, Deka R, Weeks DE, McGarvey ST (2007) Genome-wide scan for adiposity-related phenotypes in adults from American Samoa. Int J Obes (accepted)

[55] Hedley AA, Ogden CL, Johnson CL, Carroll MD, Curtin LR, Flegal KM (2004) Prevalence of overweight and obesity among US children, adolescents, and adults, 1999-2002. JAMA 291:2847–2850

[56] Popkin BM, Gordon-Larsen P (2004) The nutrition transition: worldwide obesity dynamics and their determinants. Int J Obes Relat Metab Disord 28:Suppl 3:S2-S9

[57] Bell CG, Walley AJ, Froguel P (2005) The genetics of human obesity. Nat Rev Genet 6:221–234

[58] Flegal KM, Carroll MD, Ogden CL, Johnson CL (2002) Prevalence and trends in obesity among US adults, 1999-2000. JAMA 288:1723–1727

[59] McGarvey ST (1991) Obesity in Samoans and a perspective on the etiology in Polynesians. Am J Clin Nutr 53:S1586–S1594

[60] McGarvey ST (2001) Cardiovascular disease (CVD) risk factors in Samoa and American Samoa, 1990-95. Pac Health Dialog 8:157–162

[61] Roberts ST, McGarvey ST, Quested C, Viali S (2004) Youth blood pressure levels in Samoa in 1979 and 1991-93. Am J Hum Bio 16:158–167

[62] Keighley ED, McGarvey ST, Turituri P, Viali S (2006a) Farming and Adiposity in Samoan Adults. Am J Hum Biol 18:1–11

[63] Keighley ED, McGarvey ST, Quested C, McCuddin C, Viali S, Maiava T (2006b) Nutrition and health in modernizing Samoans: temporal trends and adaptive perspectives. ED Keighley, In R Ohtsuka, SJ Ulijaszek (eds). Nutrition and Health Changes in the Asia-Pacific Region. Cambridge University Press (in press)

[64] Swinburn BA, Ley SJ, Carmichael HE, Plank LD (1999) Body size and composition in Polynesians. Int J Obes 23:1178–1183

[65] Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Prusse L, Bouchard C (2006) The Human Obesity Gene Map: The 2005 Update. Obesity 14:529–644

[66] Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. Nat Rev Genet 1:182–189

[67] Tsai HJ, Sun G, Smelser D, Viali S, Tufa J, Jin L, Weeks DE, McGarvey ST, Deka R (2004) Distribution of genome-wide linkage disequilibrium based on microsatellite loci in the Samoan population. Hum Genomics 1:327–334

[68] Galanis D, McGarvey ST, Quested C, Sio B, Afele-Faamuli S (1999) Dietary intake among modernizing Samoans: Implications for risk of cardiovascular disease. J Am Diet Assoc 99:184–190

[69] McGarvey ST (1994) The thrifty gene concept and adiposity studies in biological anthropology. J Polyn Soc 103:29–42

[70] Deka R, Shriver MD,Yu LM, Heidreich EM, Jin L, Zhong Y, McGarvey ST, Agarwal SS, Bunker CH, Miki T, Hundrieser J, Yin SJ, Raskin S, Barrantes R, Ferrell RE, Chakraborty R (1999) Genetic variation at twentythree microsatellite loci in sixteen human populations. J Genet 78:99–121

[71] McGarvey ST, Levinson PD, Rausserman L, Galanis DJ, Hornick CA (1993) Population change in adult obesity and blood lipids in American Samoa from 1976-1978 to 1990. Am J Hum Biol 5:17–30

[72] US Department of Commerce: Census of Population and Housing, American Samoa 2000. Washington DC, 2004

[73] Samoa Department of Statistics (2003) Census of population and housing 2001. Apia, Samoa: Government Printing House

[74] Williams-Blangero S, Blangero J (2006) Collection of pedigree data for genetic analysis in Isolated populations. Hum Biol 78:89–101

[75] Lohman TG, Roche AF, Martorell R (1988) Anthropometric Standardization Reference Manuals. Human Kinetics Press: Champaign, IL

[76] Box GEP, Cox DR (1982) An analysis of transformations revisited, rebutted. J Am Stat Assoc 77:209–210

[77] Cook RD, Wang PC (1983) Transformations and influential cases in regression. Technometrics 25:337–343

[78] Boehnke M, Cox NJ (1997) Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet 61:423–429

[79] Eptsein M, Duren WL, Boehnke M (2000) Improved inference of relationships for pairs of individuals. Am J Hum Genet 67:1219–1231

[80] Duren WL, Epstein M, Li M, and Boehnke M (2004) RELPAIR: A Program that Infers the Relationships of Pairs of Individuals Based on Marker Data. Version 2.0.1, June 2004

[81] McPeek MS, Sun L (2000) Statistical Tests for Detection of Misspecified Relationships by Use of Genome-Screen Data. Am J Hum Genet 66:1076–1094

[82] Sun L, Wilder K, McPeek MS (2002) Enhanced pedigree error detection. Hum Hered 54:99–110

[83] O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266

[84] Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760

[85] Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE (2005) Mega2: data-handling for facilitating genetic linkage and association analyses. Bioinformatics 21:2556–2557

[86] Sobel E, Weeks DE (1998) Multipoint IBD estimation at arbitrary locations using general pedigrees. Am J Hum Genet 63:A309

[87] Goring HH, Williams JT, Dyer TD, Blangero J (2003) On different approximations to multilocus identity-by-descent calculations and the resulting power of variance component-based linkage analysis. BMC Genet 4 Suppl 1:S72

[88] McPeek MS, Wu X, Ober C (2004) Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60:359–367

[89] Kong X, Murphy K, Raj T, He C, White PS, Matise TC (2004) A Combined Linkage-Physical Map of the Human Genome. Am J Hum Genet 75:1143–1148

[90] Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–544

[91] Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

[92] Zeegers M, Rijsdijk F, Sham P (2004) Adjusting for covariates in variance components QTL linkage analysis. Behav Genet 34:127–133

[93] Ekstrøm CT (2004) Multipoint linkage analysis of quantitative traits on sex-chromosomes. Genet Epidemiol 26:218–230

[94] Kent JW Jr, Dyer TD, Blangero J (2005) Estimating the additive genetic effect of the X chromosome. Genet Epidemiol 29:377–388

[95] Lange K, Sobel E (2006) Variance component models for X-linked QTLs. Genet Epidemiol 30:380–383

[96] Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E (2001) Mendel version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Am J Hum Genet 69(Suppl):A1886

[97] Evans DM (2002) The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. Am J Hum Genet 70:1599–1602

[98] Evans DM, Duffy DL (2004) A simulation study concerning the effect of varying the residual phenotypic correlation on the power of bivariate quantitative trait loci linkage analysis. Behav Genet 34:135–141

[99] Meyre D, Bouatia-Naji N, Tounian A, Samson C, Lecoeur C, Vatin V, Ghoussaini M, Wachter C, Hercberg S, Charpentier G, Patsch W, Pattou F, Charles MA, Tounian P, Clement K, Jouret B, Weill J, Maddux BA, Goldfine ID, Walley A, Boutin P, Dina C, Froguel P (2005) Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. Nat Genet 37:863–867

[100] Meyre D, Froguel P (2006) ENPP1, the first example of common genetic link between childhood and adult obesity and type 2 diabetes. Med Sci (Paris) 22:308–312

[101] Stern MP, Duggirala R, Mitchell BD, Reinhart LJ, Shivakumar S, Shipman PA, Uresandi OC, Benavides E, Blangero J, OConnell P (1996) Evidence for linkage of regions on chromosomes 6 and 11 to plasma glucose concentrations in Mexican Americans. Genome Res 6:724–734

[102] Duggirala R, Blangero J, Almasy L, Arya R, Dyer TD, Williams KL, Leach RJ, O'Connell P, Stern MP (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. Am J Hum Genet 68:1149–1164

[103] Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, et al. (2000) The Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics (FUSION) Study: I. An autosomal genome scan for genes that predispose to type 2 diabetes. Am J Hum Genet 67:1174–1185

[104] Watanabe RM, Ghosh S, Langefeld CD, Valle TT, Hauser ER, Magnuson VL, Mohlke KL et al. (2000) The Finland United States Investigation of Non Insulin-Dependent Diabetes Mellitus Genetics study. II. An autosomal genome scan for diabetes-related quantitative trait loci. Am J Hum Genet 67:1186–1200

[105] Hager J, Dina C, Francke S, Dubois S, Houari M, Vatin V, Vaillant E, Lorentz N, Basdevant A, Clement K, Guy-Grand B, Froguel P (1998) A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. Nat Genet 20:304–308

[106] Meyre D, Lecoeur C, Delplanque J, Francke S, Vatin V, Durand E, Weill J, Dina C, Froguel P (2004) A genome-wide scan for childhood obesity-associated traits in French families shows significant linkage on chromosome 6q22.31-q23.2. Diabetes 53:803–811

[107] Wu X, Cooper RS, Borecki I, Hanis C, Bray M, Lewis CE, Zhu X, Kan D, Luke A, Curb D (2002) A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program. Am J Hum Genet 70:1247–1256

[108] Geller F, Dempfle A, Gorg T (2003) Genome scan for body mass index and height in the Framingham Heart Study families. BMC Genet Suppl 1:S91

[109] Wu X, Luke A, Cooper RS, Zhu X, Kan D, Tayo BO, Adeyemo A (2004) A genome scan among Nigerians linking resting energy expenditure to chromosome 16. Obes Res 12:577–581

[110] Tanizawa, Y, Riggs AC, Dagogo-Jack S, Vaxillaire M, Froguel P, Liu L, Donis-Keller H, Permutt MA (1994) Isolation of the human LIM/homeodomain gene islet-1 and identification of a simple sequence repeat 1. Diabetes 43:935–941

[111] Feitosa MF, Borecki IB, Rich SS, Arnett DK, Sholinsky P, Myers RH, Leppert M, Province MA (2002) Quantitative-trait loci influencing body-mass index reside on chro-

mosomes 7 and 13: the National Heart, Lung, and Blood Institute Family Heart Study. Am J Hum Genet 70:72–82

[112] Chen W, Li S, Cook NR, Rosner BA, Srinivasan SR, Boerwinkle E, Berenson GS (2004) An autosomal genome scan for loci influencing longitudinal burden of body mass index from childhood to young adulthood in white sibships: The Bogalusa Heart Study. Int J Obes Relat Metab Disord 28:462–469

[113] Li WD, Dong C, Li D, Zhao H, Price RA (2004) An obesity-related locus in chromosome region 12q23-24. Diabetes 53:812–820

[114] Yamagata K, Oda N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, et al. (1996) Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3) Nature 384:455–457

[115] Vaxillaire M, Rouard M, Yamagata K, Oda N, Kaisaki PJ, Boriraj VV, Chevre JC, Boccio V, Cox RD, Lathrop GM, Dussoix P, Philippe J, Timsit J, Charpentier G, Velho G, Bell GI, Froguel P (1997) Identification of nine novel mutations in the hepatocyte nuclear factor 1 alpha gene associated with maturity-onset diabetes of the young (MODY3). Hum Mol Genet 6:583–586

[116] Urhammer SA, Rasmussen SK, Kaisaki PJ, Oda N, Yamagata K, Moller AM, Fridberg M, Hansen L, Hansen T, Bell GI, Pedersen O (1997) Genetic variation in the hepatocyte nuclear factor-1 alpha gene in Danish Caucasians with late-onset NIDDM. Diabetologia 40:473–475

[117] Frayling TM, Bulamn MP, Ellard S, Appleton M, Dronsfield MJ, Mackie AD, Baird JD, Kaisaki PJ, Yamagata K, Bell GI, Bain SC, Hattersley AT (1997) Mutations in the hepatocyte nuclear factor-1alpha gene are a common cause of maturity-onset diabetes of the young in the U.K. Diabetes 46:720–725

[118] Ellard S (2000) Hepatocyte nuclear factor 1 alpha (HNF-1 alpha) mutations in maturity-onset diabetes of the young. Hum Mutat 16:377–385

[119] Norris JM, Langefeld CD, Scherzinger AL, Rich SS, Bookman E, Beck SR, Saad MF, Haffner SM, Bergman RN, Bowden DW, Wagenknecht LE (2005) Quantitative trait loci for abdominal fat and BMI in Hispanic-Americans and African-Americans: the IRAS Family study. Int J Obes 29:67–77

[120] Toplak H, Wascher TC, Weber K, Lauermann T, Reisinger EC, Bahadori B, Tilz GP, Haller EM (1995) Increased prevalence of serum IgA Chlamydia antibodies in obesity. Acta Med Austriaca 22:23–24

[121] Toplak H, Haller EM, Lauermann T, Weber K, Bahadori B, Reisinger EC, Tilz GP, Wascher TC (1996) Increased prevalence of IgA-Chlamydia antibodies in NIDDM patients. Diabetes Res Clin Pract 32:97–101

[122] Hsueh WC, Mitchell BD, Schneider JL, St Jean PL, Pollin TI, Ehm MG, Wagner MJ, Burns DK, Sakul H, Bell CJ, Shuldiner AR (2001) Genome-wide scan of obesity in the old order Amish. J Clin Endocrinol Metab 86:1199–1205

[123] Ohman M, Oksanen L, Kaprio J, Koskenvuo M, Mustajoki P, Rissanen A, Salmi J, Kontula K, Peltonen L (2000) Genome-wide scan of obesity in Finnish sibpairs reveals linkage to chromosome Xq24. J Clin Endocrinol Metab 85:3183–3190

[124] Suviolahti E, Oksanen LJ, Ohman M, Cantor RM, Ridderstrale M, Tuomi T, Kaprio J, Rissanen A, Mustajoki P, Jousilahti P, Vartiainen E, Silander K, Kilpikari R, Salomaa V, Groop L, Kontula K, Peltonen L, Pajukanta P (2003) The SLC6A14 gene shows evidence of association with obesity. J Clin Invest 112:1762–1772

[125] Deng HW, Deng H, Liu YJ, Liu YZ, Xu FH, Shen H, Conway T, Li JL, Huang QY, Davies KM, Recker RR (2002) A genomewide linkage scan for quantitative-trait loci for obesity phenotypes. Am J Hum Genet 70:1138–1151

[126] Martin LJ, Cole SA, Hixson JE, Mahaney MC, Czerwinski SA, Almasy L, Blangero J, Comuzzie AG (2002) Genotype by smoking interaction for leptin levels in the San Antonio Family Heart Study. Genet Epidemiol 22:105–115

[127] Martin LJ, Kissebah AH, Sonnenberg GE, Blangero J, Comuzzie AG (2003) Genotype-by-smoking interaction for leptin levels in the Metabolic Risk Complications of Obesity Genes project. Int J Obes Relat Metab Disord 27:334–340

[128] Kraja AT, Rao DC, Weder AB, Cooper R, Curb JD, Hanis CL, Turner ST, de Andrade M, Hsiung CA, Quertermous T, Zhu X, Province MA (2000) Two major QTLs and several others relate to factors of metabolic syndrome in the family blood pressure program. Hypertension 46:751–757

[129] Dong C, Li WD, Geller F, Lei L, Li D, Gorlova OY, Hebebrand J, Amos CI, Nicholls RD, Price RA (2005) Possible genomic imprinting of three human obesity-related genetic loci. Am J Hum Genet 76:427–437

[130] Erickson J, Hollopeter G, Palmiter R (1996) Attenuation of the obesity syndrome of ob/ob mice by the loss of neuropeptide Y. Science 274:1704–1707

[131] Bray MS, Boerwinkle E, Hanis CL (1999) Linkage analysis of candidate obesity genes among the Mexican-American population of Starr County, Texas. Genet Epidemiol 16:397–411

[132] Bray MS, Boerwinkle E, Hanis CL (1999) Sequence variation within the neuropeptide Y gene and obesity in Mexican Americans. Obes Res 8:219–226

[133] Chagnon YC, Borecki IB, Perusse L, Roy S, Lacaille M, Chagnon M, Ho-Kim MA, Rice T, Province MA, Rao DC, Bouchard C (2000) Genome-wide search for genes related to the fat-free body mass in the Quebec family study. Metabolism 49:203–207

[134] Froguel P, Velho G (1993) Non-sense mutation of glucokinase gene. Lancet 341:385

[135] Froguel P, Zouali H, Vionnet N, Velho G, Vaxillaire M, Sun F, Lesage S, Stoffel M, Takeda J, Passa P, Permitt MA, Beckmann JS, Bell GI, Cohen D (1993) Familial hyperglycemia due to mutations in glucokinase: definition of a subtype of diabetes mellitus. New Eng J Med 328:697–702

[136] Stone S, Abkevich V, Hunt SC, Gutin A, Russell DL, Neff CD, Riley R, Frech GC, Hensel CH, Jammulapati S, Potter J, Sexton D, Tran T, Gibbs D, Iliev D, Gress R, Bloomquist B, Amatruda J, Rae PM, Adams TD, Skolnick MH, Shattuck D (2002) A Major Predisposition Locus for Severe Obesity, at 4p15-p14. Am J Hum Genet 70:1459–1468

[137] Arya R, Duggirala R, Jenkinson CP, Almasy L, Blangero J, O'Connell P, Stern MP (2004) Evidence of a novel quantitative-trait locus for obesity on chromosome 4p in Mexican Americans. Am J Hum Genet 74:272–282

[138] Saar K, Geller F, Ruschendorf F, Reis A, Friedel S, Schauble N, Nurnberg P, Siegfried W, Goldschmidt HP, Schafer H, Ziegler A, Remschmidt H, Hinney A, Hebebrand J (2003) Genome scan for childhood and adolescent obesity in German families. Pediatrics 111:321–327

[139] Parlier G, Thomas G, Bereziat G, Fontaine JL, Girardet JP (1997) Relation of apolipoprotein E polymorphism to lipid metabolism in obese children. Pediatr Res 41:682–685

[140] Oh JY, Barrett-Connor E; Rancho Bernardo Study Group (2001) Apolipoprotein E polymorphism and lipid levels differ by gender and family history of diabetes: the Rancho Bernardo Study. Clin Genet 60:132–137

[141] Yanagisawa Y, Hasegawa K, Dever GJ, Otto CT, Sakuma M, Shibata S, Miyagi S, Kaneko Y, Kagawa Y (2001) Uncoupling protein 3 and peroxisome proliferator-activated receptor gamma2 contribute to obesity and diabetes in palauans. Biochem Biophys Res Commun 281:772–778

[142] Nicklas BJ, Ferrell RE, Bunyard LB, Berman DM, Dennis KE, Goldberg AP (2002) Effects of apolipoprotein E genotype on dietary-induced changes in high-density lipoprotein cholesterol in obese postmenopausal women. Metabolism 51:853–858

[143] Petruschke T, Rohrig K, Hauner H (1994) Transforming growth factor beta (TGF-beta) inhibits the differentiation of human adipocyte precursor cells in primary culture. Int J Obes Relat Metab Disord 18:532–536

[144] Long JR, Liu PY, Liu YJ, Lu Y, Xiong DH, Elze L, Recker RR, Deng HW (2003) APOE and TGF-beta1 genes are associated with obesity phenotypes. J Med Genet 40:918–924

[145] Alessi MC, Bastelica D, Morange P, Berthet B, Leduc I, Verdier M, Geel O, Juhan-Vague I (2000) Plasminogen activator inhibitor 1, transforming growth factor-beta1, and BMI are closely associated in human adipose tissue during morbid obesity. Diabetes 49:1374–1380

[146] Rosmond R, Chagnon M, Bouchard C, Bjorntorp P (2003) Increased abdominal obesity, insulin and glucose levels in nondiabetic subjects with a T29C polymorphism of the transforming growth factor-beta1 gene. Horm Res 59:191–194

[147] Hirschhorn JN, Lindgren CM, Daly MJ, Kirby A, Schaffner SF, Burtt NP, Altshuler D, Parker A, Rioux JD, Platko J, Gaudet D, Hudson TJ, Groop LC, Lander ES (2001) Genomewide Linkage Analysis of Stature in Multiple Populations Reveals Several Regions with Evidence of Linkage to Adult Height. Am J Hum Genet 69:106–116

[148] Towne B, Siervogel RM, Blangero J (1997) Effects of genotype-by-sex interaction on quantitative trait linkage analysis. Genet Epidemiol 14:1053–1058

[149] Atwood L, Heard-Costa N, Fox C, Jaquish C, Cupples LA (2006) Sex and age specific effects of chromosomal regions linked to body mass index in the Framingham Study. BMC Genetics 7:7–19

[150] Suresh R, Ambrose N, Roe C, Pluzhnikov A, Wittke-Thompson JK, Ng MCY, Wu X, Cook EH, Lundstrom C, Garsten M, Ezrati R, Yairi E, Cox NJ (2006) New Complexities in the Genetics of Stuttering: Significant Sex-Specific Linkage Signals. Am J Hum Genet 78:554–563

[151] Lander ES, Lincoln SE (1988) The appropriate threshold for declaring linkage when allowing sex-specific recombination rates. Am J Hum Genet 43:396–400

[152] Weeks DE, Ott J, Lathrop GM (1990) SLINK: A general simulation. program for linkage analysis. Am J Hum Genet 47:A204

[153] Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

[154] Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127

[155] Boomsma DI, Dolan CV (1998) A Comparison of Power to Detect a QTL in Sib-Pair Data Using Multivariate Phenotypes, Mean Phenotypes, and Factor Scores. Behav Genet 28:329–340

[156] Marlow AJ, Fisher SE, Francks C, MacPhie IL, Cherny SS, Richardson AJ, Talcott JB, Stein JF, Monaco AP, Cardon LR (2003) Use of Multivariate Linkage Analysis for Dissection of a Complex Cognitive Trait. Am J Hum Genet 72:561–570

[157] Wang T, Elston RC (2007) Regression-based Multivariate Linkage Analysis with an Application to Blood Pressure and Body Mass Index. Ann Hum Genet 71:96–106

[158] Kilpikari R, Sillanp MJ (2003) Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet Epidemiol 25:122–135

[159] Sillanp MJ, Auranen K (2004) Replication in genetic studies of complex traits. Ann Hum Genet 68:646–657

[160] Shmulewitz D, Heath SC, Blundell ML, Han Z, Sharma R, Salit J, Auerbach SB, Signorini S, Breslow JL, Stoffel M, Friedman J (2007) Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and dyslipidemia on the island of Kosrae, Federated States of Micronesia. Proc Natl Acad Sci U S A 103:3502–3509

[161] Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM (1997) MCMC segregation and linkage analysis. Genet Epidemiol 14:1011–1016

[162] Leal SM, Heath SC (1999) Searching for alcoholism susceptibility genes using Markov chain Monte Carlo methods. Genet Epidemiol 17:S217–S222