2002

Mapping Genes for Complex Traits: Obesity, Diabetes, Hypertension, and Dyslipidemia on the Pacific Island of Kosrae

Dvora Shmulewitz

CELEBRATING 100 YEARS

1901

THE ROCKEFELLER UNIVERSITY

PRO BONO HUMANI GENERIS

Mapping Genes for Complex Traits: Obesity, Diabetes, Hypertension, and Dyslipidemia on the Pacific Island of Kosrae

A thesis presented to the faculty of
The Rockefeller University
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

by

Dvora Shmulewitz

The Rockefeller University
New York
April, 2002

Dedication

בהכרת הטוב לד' יתברך. אם ד' לא יבנה בית שוא עמלו בוניו
בו, אם ד' לא ישמר עיר שוא שקד שומר. שוא לכם משכימי קום
מאחרי שבת אוכלי לחם העצבים, כן יתן לידידן שנא.
תהילים קכז; א-ב

I dedicate this thesis to Aba and Ema, who made this (and everything) possible. Thank you for your love, confidence, and constant support.

Thank you to friends and family who encouraged me to pursue my dreams: Yitzchak and Faige; Rafael and Chaim; Oma; Bobby and Zeidy; Ezra, Faigy, and Leah; Ben and Hindi; Rivka; and אחרונה אחרונה חביבה, the incomparable Mrs. Press.

Acknowledgements

Jeffrey Friedman was the ideal thesis advisor for me, as he allowed me the freedom to explore everything that interested me, even if it was far from his vision for this project. Jeff is amazingly generous with his time and advice (as well as Rangers tickets), and he taught me to think scientifically. He has an uncanny ability to see to the heart of a problem and to come up with creative and exciting solutions. If I could be half the scientist and educator he is, I would be happy.

I was unbelievably lucky to have advisors who were really teachers, collaborators, colleagues, and friends. There is no way a simple thank you can show all that I owe them, but that is really all I can say. Markus Stoffel has an infectious love for science, taught me all that I know about experimental molecular biology, and shared my obsession with the Jets. Jan Breslow taught me how to express myself clearly and precisely as a scientist, partially by going through my first manuscript line by line, and by showing me how to put together a presentation. When I first came to Simon Heath, I was totally ignorant about statistical genetics and computers, and he had the patience (and the courage) to take me on anyway, and teach me everything that I know in these areas.

The Friedman lab attracts some of the most wonderful people and scientists I have ever met. Thank you Elizabeth Stephenson, Barbara Saatkamp, Jeff DeFalco, Mac Ishii, Paul Cohen, Alex Soukas, Alena Pithart, Rat Sharma, and of course the "girls"- Agnes Viale, Maude Blundell, Esra Asilmaz, Shirly Pinto, and Silvia Novelli. Your support, caring, and friendship made all the difference. Susan Korres has helped me in innumerable ways over the years - thanks.

This project could not have been done without the diligent work of the following people: Arlene Auerbach, Steven Auerbach, Maude Blundell, Stephanie Boger, Rebecca Breslow, Brandie Galke, Derek Gordon, Zhihua Han, Maria Karayiorgou, Thomas Lehner, Mark Levenstien, Sandra Oplanich, Jurg Ott, Lynn Petukhova, Ratnendra Sharma, Stefano Signorini, Elizabeth Stephenson, Zoltan Takacs, Peter Verlander, Jeffrey Winick, Chavi Wolfish, and Merav Yifrach.

Thank you Dr. Mary Jeanne Kreek and Dr. John Blangero for being on my committee and for helpful comments and suggestions.

I am grateful to the Dean's office for support and a wonderful student environment, especially as provided by Marta Delgado and Kristen Cullin.

Table of Contents

Directory of Figures

Directory of Tables

Directory of Abbreviations

ADA	American Diabetes Association
APOA-I	apolipoprotein A-I
APOB	apolipoprotein B
APOE	apolipoprotein E
BMI	body mass index
BP	blood pressure
ch	chromosome
DB	diabetes
DBP	diastolic blood pressure
DYSL	dyslipidemia
FBS	fasting blood sugar
FSM	Federated States of Micronesia
GAW	Genetic Analysis Workshop
GCF	Genotyping Core Facility
HDL-C	high-density lipoprotein cholesterol fraction
HIP	hip circumference
HT	height
IBD	identity-by-descent
IGT	impaired glucose tolerance
INS	insulin concentration
LA	linkage analysis
LDL-C	low-density lipoprotein cholesterol fraction
LEP	leptin concentration
MAP	mean arterial pressure
MG1/2/3	major gene 1, 2, or 3
MCMC	Markov chain Monte Carlo
NIDDM	non-insulin dependent diabetes mellitus
OB	obesity
OGTT2	2-hour oral glucose tolerance test
PP	pulse pressure
Q1/2	quantitative trait 1 or 2
QTL	quantitative trait locus
SA	segregation analysis
SBP	systolic blood pressure
SI	segregation indicator
SOLAR	Sequential Oligogenic Linkage Analysis Routines
TC	total cholesterol concentration
TG	total triglyceride concentration
WHR	waist-to-hip ratio
WST	waist circumference
WT	weight

Publications

Han Z, Heath SC, Shmulewitz D, Li W, Auerbach SB, Blundell ML, Lehner T, Ott J, Stoffel M, Friedman JM, and JL Breslow. (2002) Candidate genes involved in cardiovascular risk factors by a family-based association study on the island of Kosrae, Federated States of Micronesia. *in press, Am J Med Gen*

Shmulewitz D and SC Heath (2001) Genome scans for Q1 and Q2 on general population replicates using Loki. Genetic Epidemiology 21(Suppl 1): S686-S69

Shmulewitz D, Auerbach SB, Lehner T, Blundell ML, Winick JD, Youngman LD, Skilling V, Heath SC, Ott J, Stoffel M, Breslow JL, and JM Friedman. (2001) Epidemiology and factor analysis of obesity, type II diabetes, hypertension, and dyslipidemia (Syndrome X) on the island of Kosrae, Federated States of Micronesia. Human Heredity, 51;8-19

Shmulewitz D and SC Heath (2002) Genome scans using Loki. *In preparation.*

Shmulewitz D, Heath SC, Han Z, Petukhova L, Auerbach SB, Verlander PC, Auerbach AD, Lehner T, Blundell ML, Winick JD, Stephenson EA, Yifrach M, Oplanich, S, Boger S, Youngman LD, Skilling V, Ott J, Stoffel M, Breslow JL, and JM Friedman. (2002) Mapping genes for obesity, type II diabetes, hypertension, and dyslipidemia (Syndrome X) on the island of Kosrae, Federated States of Micronesia. *In preparation.*

Abstracts

Han Z, Li W, Shmulewitz D, Heath SC, Auerbach SB, Blundell ML, Lehner T, Ott J, Stoffel M, Friedman JM, and JL Breslow. (2001) Family based association study (qTDT) on lipid abnormality candidate genes on an isolated, admixed population. Am J Hum Genet, 69(4):572

Han Z, Heath SC, Shmulewitz D, Auerbach SB, Blundell ML, Lehner T, Ott J, Stoffel M, Friedman JM, and JL Breslow. (2000) Family based association study of candidate genes regulating lipid and apolipoprotein levels on the island of Kosrae. Am J Hum Genet, 67(4):351

Shmulewitz D, Heath SC, Lehner T, Auerbach SB, Asilmaz E, Blundell ML, Petukhova L, Winick JD, Verlander PC, Han Z, Breslow JL, Ott J, Stoffel M, and JM Friedman. (2000) Gene mapping and admixture on the island of Kosrae. Am J Hum Genet, 67(4):236

Winick JD, Tuggle JT, Gilbreath AE, Xia J, Gwynne P, Gaskin M, Klimecki W, Peters T, Shmulewitz D, Heath SC, Bonner MR, Gallagher S,Friedman JM, and TJ Raich. (2000) Assay development of a 3-dimensional microarray system to detect mitochondrial SNPs. Am J Hum Genet, 67(4):23

Shmulewitz D, Heath SC, Auerbach SB, Blundell ML, Winick JD, Signorini S, Breslow JL, Ott J, Lehner T, Stoffel M, and JM Friedman. (1999) Genetic epidemiology of Syndrome X disorders on Kosrae. Am J Hum Genet, 65(4):A86

Abstract

One of the current challenges in human genetics is to map genes for common, complex diseases. For powerful mapping of such phenotypes, the suggestions are to analyze the underlying quantitative traits with covariate corrections in large extended pedigrees with multipoint analysis. This is virtually impossible to do with the current linkage programs, due to the computation difficulties of exactly calculating every possibility. Instead, sampling methods that sample the most likely data configuration from all the possibilities need to be used. This has been implemented in the reversible jump Markov chain Monte Carlo method Loki, which can carry out segregation and linkage analysis on quantitative traits in large pedigrees with multipoint analysis. Loki can model the trait with covariates, identify the number of quantitative trait loci, position linked loci, and estimate allele frequencies and gene effects. This method has a lot of promise but has not been vigorously tested for complete genome scans.

The first part of this study was to develop a strategy for carrying out genome scans using Loki and to evaluate the output. This was first done using the Genetic Analysis Workshop 12 simulated dataset with known answers. This resulted in a number of suggestions, such as initial single chromosome analysis, correction for polygenic effect, joint analysis of positive signals, and convergence analysis. Next these suggestions were applied to a real dataset from the population of Kosrae, the Federated States of Micronesia. This is a study of the population on the island of Kosrae, which has one large extended pedigree and high prevalence of the common complex disorders that are known as Syndrome X: obesity, type II diabetes, hypertension, and dyslipidemia.

This resulted in a number of additional suggestions, such as phenotypic and genotypic corrections, dealing with mixing issues, and inspection of L-graphs for signal reliability.

Once this strategy was developed, the second part of this study was to use Loki to identify quantitative trait loci for the continuous traits associated with Syndrome X and stature. This resulted in quantitative trait loci for body mass index, hip circumference, weight, fasting blood sugar, systolic blood pressure, arterial blood pressure, apolipoprotein B, total cholesterol, and height. This also identified interesting chromosomal regions with slight signals for correlated traits on chromosomes 1, 2, 7, 9, 13, and 16. This study shows that Loki is a program that can powerfully and reliably carry out linkage analysis on quantitative traits that was previously impossible to do and finds loci for many of the quantitative traits related to common metabolic disorders as well as height.

Chapter 1: Introduction

Complex disease gene mapping

A current challenge in human genetics is to map genes that increase susceptibility to common complex diseases. Complex diseases are caused by multiple factors, such as a number of genes, age, sex, lifestyle, other environmental covariates, and interactions between them. Examples include obesity, type II diabetes, hypertension, and dyslipidemia, which are major causes of morbidity and mortality in the United States and throughout the world. Despite countless studies on each of these diseases, little progress has been made in elucidating their molecular bases, and it is unknown why these diseases often aggregate in the same individuals—a syndrome known as Syndrome X (Reaven 1988).

There have been many suggestions, yet there is still no clear way to define Syndrome X, primarily because the underlying mechanisms are so poorly understood. Indeed, the precise molecular genetic defects responsible for each of these individual disorders, or of Syndrome X, have not been elucidated except in a small number of instances. The identification of genes that are involved in the causation of Syndrome X and/or its constituent disorders is expected to enhance the understanding of their pathogenesis and lead to improved management of susceptible individuals in the population with more rational therapies. In addition, the use of genetics to stratify these diseases into genetically defined subgroups may reveal subtle phenotypic differences between them and differences in individual response to therapy. Finally, the identification of causal genes will likely reveal new physiological pathways and thus allow us to better analyze gene-gene and gene-environment interactions.

Syndrome X diseases: prevalence, phenotypes, and genetic epidemiology

Obesity

Obesity (OB) is a major worldwide health problem and has been suggested as the most pressing problem in the United States. Obesity is mainly described in terms of ponderosity which is body weight (WT) relative to height (HT) measured by body mass index (BMI, in kg/m^2). In the United States, individuals are classified as overweight when BMI is 25-29, as obese with BMI 30-34, and as extremely obese with BMI $\geq$ 35. According to these criteria, 33% of the adult US population is overweight, 14% is obese, and 8% is extremely obese. From 1985 to 1995 there was a 33% increase in the number of obese individuals (Flegal et al. 1998). Threshold values for obesity traits are population specific, such as BMI $\geq$ 30 kg/m ^{2} as obese in Caucasian populations, but $\geq$ 27 in Asians (Neel et al. 1998). Other measurements include those of fat distribution, such as waist (WST) and hip (HIP) circumference and waist/hip ratio (WHR), which measure a more central fat distribution, and those calculated from skin fold measures, such as visceral adipose tissue and trunk to extremity ratio (Durnin and Womersley 1974; Van der Kooy and Seidell 1993). Intermediate phenotypes include total fat mass, measures of energy expenditure, and levels of leptin (LEP), an adipose tissue hormone (Comuzzie and Allison 1998). Leptin may be a very good physiologic marker, as it has been suggested that a significant portion of obese humans appear to be leptin resistant, with very high endogenous levels of the hormone.

Numerous studies have confirmed that genetic factors play an important role in the development of obesity. Familial aggregation and twin studies for obesity have shown increased risk for relatives of obese individuals and resemblance of adopted children to their biological parents, as well as heritability for BMI ranging from 0.4-0.8 (Biron et al. 1977; Stunkard et al. 1986; Sorensen et al. 1992; Allison et al. 1996; Bouchard 1997; Comuzzie and Allison 1998). Other measures of this phenotype are highly heritable, such as estimates of 0.39-0.55 for LEP, 0.72-0.82 for WST and 0.19-0.61 for WHR (Rotimi et al. 1997; Rose et al. 1998; Narkiewicz et al. 1999). Close to 40% of the variation in BMI has been shown to be due to a major gene, with another 40% due to polygenic effects (Moll et al. 1991; Comuzzie and Allison 1998). Other phenotypes, such as visceral fat and fat mass, show large percentage of the variance due to major gene as well as multifactorial effects (polygenes and environmental interactions) (Mitchell et al. 1996b; Lecomte et al. 1997). Though genes for monogenic forms of obesity and related syndromes have been found and linkage to general obesity and / or its related traits has been reported in some studies, there has been little replication and virtually no causal genes have been identified (Comuzzie and Allison 1998). Some of these studies will be discussed later.

Diabetes

Type II diabetes, or Non-Insulin Dependent Diabetes Mellitus (NIDDM), called here diabetes (DB), is a leading cause of death and morbidity, affecting more than 135 million people worldwide, and has a huge economic impact (McCarthy and Zimmet 1994; Harris 1995). Diabetes is usually defined based on fasting blood sugar (FBS or glucose) levels (≥ 126 mg/dl), as well as a two-hour oral glucose tolerance test (OGTT2, ≥ 200 mg/dl) to confirm the diagnosis, and it occurs in close to 8% of adult Americans (Harris et al. 1998). A less severe form, impaired glucose tolerance (IGT), based on intermediate glucose levels (FBS 100 – 125), is sometimes used to define an in-between phenotype, which may either be a disease state on its own or pre-diabetes. More specific measures of glucose, like post-prandial or post-challenge, are used to further define possible subsets. Diabetes is usually associated with insulin resistance and concomitant hyperinsulinemia, so various insulin related phenotypes can also be used. These include fasting insulin (INS), post challenge insulin, insulin secretion (pancreatic β-cell function), and insulin sensitivity. These various measures may prove to be important in defining Syndrome X, which may be due to insulin resistance.

Familial and population studies of diabetes point to a major genetic component by showing high concordance rates in twins, lifetime risk of greater than 40% for first degree relatives of patients and significant heritability of the underlying continuous variables (Hamann 1992). These include insulin secretion (0.4-0.7), abnormal glucose tolerance (0.61), insulin resistance (0.26), INS (0.53), and FBS (0-0.32) (Rice et al. 1990; Mayer et al. 1996; Elbein et al. 1999; Poulsen et al. 1999). There is also clear familial clustering of IGT and diabetes. There is evidence of a major gene component and polygenic effects for FBS, as well as for two-hour insulin and fasting insulin (Schumacher et al. 1992; Hanson et al. 1995; Mitchell et al. 1996b). To date, the only specific diabetes-susceptibility genes that have been identified are responsible for relatively rare, early-onset forms with monogenic or mitochondrial inheritance (Ballinger et al. 1992; Vionnet et al. 1992; Kadowaki et al. 1994; Yamagata et al. 1996a; Yamagata et al. 1996b; Horikawa et al. 1997; Stoffers et al. 1997). While several reports of linkage have been shown for various susceptibility loci, including some replicated regions such as on ch 1, the genes for the most common forms of late-onset type 2 diabetes have not yet been identified. These and other linkage studies will be discussed later.

Hypertension

Hypertension, defined as systolic blood pressure (SBP) ≥140mm or diastolic blood pressure (DBP) ≥90mm, is the strongest risk factor for heart disease and one of the most prevalent disorders in the US, occurring in 24% of adults (Anonymous 1993; Burt et al. 1995). Though systolic and diastolic blood pressure levels are often used for diagnosis, alternate measures that combine both, such as pulse pressure (PP, SBP-DBP) or mean arterial pressure (MAP, DBP+[SBP-DBP]/3), have also been used. Hypertension has many subsets, such as high/low renin and salt sensitivity or insensitivity, which may be used to reduce complexity. Other intermediate phenotypes can include urinary kallikrein secretion and various measures of erythrocyte ion exchange, such as $Na^+/Cl^-/K^+$ cotransport and Na^+/Li^+ countertransport (Lifton 1995).

Many studies confirm a genetic basis for blood pressure (BP) control. Heritability estimates range from 0.10 to 0.64 and 0.13 to 0.82 for DBP and SBP, respectively (Hong et al. 1994). Major gene effects have been shown for blood pressure, as well as for many of the intermediate phenotypes based on erythrocyte ion transport (Cusi et al. 1991; Williams et al. 1993; Williams et al. 1994; Cheng et al. 1995; Gu et al. 1998). To date, the identification of hypertension genes has also been limited to rare Mendelian forms. These include glucocorticoid-remediable aldosteronism, caused by the fusion of the aldosterone synthase and 11-βhydroxylase genes, and Liddle Syndrome, caused by various mutations in the β and γ amiloride-sensitive epithelial sodium channel genes

(Lifton et al. 1992; Shimkets et al. 1994; Hansson et al. 1995). Linkage analyses have suggested some possible regions with little consensus which will be discussed later.

Dyslipidemia

Risk for heart disease, the most frequent cause of death in the US, is also incurred with derangement in the levels of triglycerides (TG), total cholesterol (TC), high density and low density cholesterol fractions (HDL-C, LDL-C) and their constituent apolipoproteins, apolipoprotein A-I (APOA-I) and apolipoprotein B (APOB), which are all included in the term dyslipidemia (DYSL) (Castelli et al. 1986; Higgins et al. 1996). TC levels greater than 240 mg/dl occur in 20%, and TG levels greater than 200 mg/dl occur in 18% of adult men (40-59 years old) and 8% of adult women (40-59) (Anonymous 1980; 1989). Other indicators of dyslipidemia are cholesterol particle size and turnover rate, as well as enzyme activity in various synthetic or degradative pathways for these factors (Perusse et al. 1997). Each parameter may be used separately or combined in some way, thus complicating diagnosis. Furthermore, it is not clear whether derangement in these factors is a disease state in itself or is merely a risk factor for heart disease. Either way, these phenotypes may be useful in studying the genes that underlie the general physiology of these factors and their interactions.

Heart disease (particularly myocardial infarction) has been shown to cluster in families, and family history is one of the predominant risk factors, suggesting that genetic factors play an important role in the development of heart disease (Barrett-Conner and Khaw 1984; Hopkins et al. 1988; Colditz et al. 1991). Twin and family studies have shown genetic contribution for many of the other risk factors (dyslipidemia), such as TC (0.46-0.62), HDL-C (0.42-0.83), LDL-C (0.42-0.50), and TG (0.21-0.55) (Brenn 1994;

Mitchell et al. 1996a; Perusse et al. 1997). Strong single gene effects were shown for the factors involved in dyslipidemia, HDL-C, LDL-C, TC, APOA-I, APOB, and TG, a well as various measures of interactions between these factors (Amos et al. 1987; Williams et al. 1993). Only a small fraction of those with elevated LDL-C levels have mutations in the LDL receptor or other genes (Soutar 1998). A number of linkage studies for many dyslipidemia measures have been done, again with little consensus, and these will be discussed later.

Syndrome X: Clustering

Epidemiological and clinical studies have shown that obesity, diabetes, hypertension, and dyslipidemia often cluster in individuals and in families, and are often collectively known as Syndrome X (Reaven 1988; Defronzo and Ferrannini 1991; Ferrannini et al. 1991; Muller et al. 1993; Reaven 1993; Wannamethee et al. 1998). A wider definition actually includes elevated levels of the variables that do not necessarily reach the threshold assigned for disease state. Therefore, one way to define Syndrome X may be to choose six or seven characteristics, assign high but not extreme cutoffs, and individuals with levels above the cutoffs for four or five of the variables are defined as affected (Suzuki et al. 1996). This is problematic, as there may be many differences in diagnosis based on the variables selected or the thresholds used.

Another idea is to use factor analysis, through principal component analysis, on the continuous variables to cluster them into a few independent factors that describe the phenotype (Edwards et al. 1994; Meigs et al. 1997; Edwards et al. 1998; Gray et al. 1998; Leyva et al. 1998). This approach makes analysis much easier (few factors instead of many), includes all variables, and maintains the informativeness of quantitative traits by eliminating false dichotomization. These factors have been shown to be genetically determined, with heritability values for the body mass/ fat distribution factor ranging from 0.61-0.71, for the insulin/glucose factor 0.57-0.92, and for the lipid factor 0.25-0.32 (Edwards et al. 1997).

Genetic effects contribute significantly to the qualitative and quantitative traits that underlie these disorders and their clustering in Syndrome X, suggesting that the causal genes can be found. This syndrome is also referred to as insulin resistance syndrome (IRS) since insulin resistance has been suggested as the underlying cause for this clustering (Haffner et al. 1992). As this name suggests a cause that has not been conclusively proven, we use the more general term, Syndrome X. It is not clear whether Syndrome X is due to a single underlying cause, common independent causes, or partially overlapping causes that explain one or more of the features of Syndrome X without accounting for the whole syndrome.

Syndrome X: Environmental Effects

The available evidence suggests that all of the Syndrome X component diseases are oligogenic and heterogeneous. Moreover, environmental factors have been shown to influence the onset and severity of these disorders. Interestingly, it has been observed that in traditionally non-Western populations the incidence of these disorders increases when a high fat "Western" diet is introduced (Neel 1962). Similar findings were also noted when certain strains of mice (Akr, C57BL/6J) were switched to a high fat "Western" diet (West et al. 1992). One idea as to why modernity, particularly the shift to a "Western" diet, would cause Syndrome X disorders is the thrifty gene hypothesis (Neel

1962). The evidence suggests that descendants of populations who have endured periodic shortages of food and water exhibit a predisposition for these metabolic disorders when exposed to a high fat diet. This thrifty gene hypothesis suggests that the same genetic variants that confer a selective advantage in times of famine and drought, by superior ability to conserve nutrients and water, can lead to obesity, diabetes, hypertension, and dyslipidemia in times of surfeit.

Other lifestyle shifts, such as increases in high salt and fat diets, alcohol, smoking, less exercise, changes in socioeconomic status, and urbanization, could also affect the prevalence of these disorders. These and additional environmental factors may be involved in various ways. Others include physiologic "environment", such as the differences seen in disease risk for males or females or in different age groups. Identification of loci so affected could lead to a fuller understanding of the relationship between genes and environment in the pathogenesis of Syndrome X and might also shed light on the increasing incidences of obesity in this century.

Genetic epidemiology analyses show gene-gene and gene-environment interaction among the traits encompassed by Syndrome X. For example, studies have shown different genetic effects in different ethnic and racial backgrounds, with higher heritabilities for blood pressure measurements in African-Americans than for the Caucasians in the same study (Gu et al. 1998). There is a major gene effect found in Pima Indians that influences risk for diabetes by affecting age of onset and, in another study, genes found to affect age-related changes in DBP were found (Cheng et al. 1995; Hanson et al. 1995). Pleiotropic loci were shown for APOA-I and HDL-C, TG and BMI, and SBP and BMI (Blangero et al. 1993; Gauderman and Faucett 1997; Cheng et al. 1998).

Gene mapping in Kosrae

To identify the genes underlying Syndrome X and its components, a comprehensive epidemiological and genetic study was undertaken on the Pacific Island of Kosrae, one of four states in the Federated States of Micronesia (FSM). Kosrae has a high prevalence of obesity, diabetes, hypertension, and dyslipidemia, and is ideal for gene mapping studies.

Kosrae history

This section is from the following references: (Hezel 1983), (Segal 1989), (Cordy 1993), and (Irwin 1994). Kosrae, located 2500 miles northeast of Australia, was originally settled by a small number of founders (estimated to be 50) from Polynesia around 50 CE. Over the ensuing centuries, the islanders interacted with the inhabitants of several neighboring islands, including Nauru, Pingelap and Pohnpei. Kosrae was first sighted by Westerners in 1804 and first visited in 1824. During the 19th century, the combined effects of a typhoon (in 1835) and exposure of the native population to Western communicable diseases reduced the indigenous population from greater than 3000 to around 300 individuals by 1888. Historical and genealogical records indicate that these 300 survivors were the result of extensive admixture between native Kosraean females and male Caucasian whalers from New England and Europe who visited the island in the mid to late 19th century. (This admixture may prove useful for localizing the genes underlying the disorders that are being studied (Chakraborty and Weiss 1988; McKeigue 1997; 1998).)

Spain originally claimed Kosrae, but the island was purchased by Germany in 1899 following the Spanish-American War. In the beginning of World War I, the Japanese gained control of Kosrae, and there is evidence of some Japanese intermarriage with the

Kosraean population. After World War II, the FSM, including the islands of Kosrae, Truk, Yap, and Pohnpei, became a protectorate of the US government. At that time, the island population was estimated to be 1,550.

A number of visitors to the island in the 19th and early 20th century noted the natives to be thin, and available records do not show that obesity or other Syndrome X diseases were a major health problem on the island prior to World War II. As late as 1945, the Kosraean consumed a diet consisting mostly of fish and fruits and vegetables. The designation of Kosrae as a US protectorate led to a drastic life style change on the island. Kosraeans no longer had to fish for sustenance and assumed a more sedentary lifestyle. In addition, the diet on the island changed to a more Western diet with large quantities of high fat foods supplied through US aid (such as Spam, turkey tails, hamburgers, and ice cream). These changes resulted in a dramatically increased prevalence of obesity, an outcome similar to that seen in other indigenous populations, such as the Pima Indians of Arizona and the Nauruans (Zimmet et al. 1977; Zimmet et al. 1978; Ravussin et al. 1994). Correspondingly, other Syndrome X diseases became significant health problems. However, the Kosraean population is different from other indigenous populations in that there is greater phenotypic variability on Kosrae compared to the other populations. This may be a result of the extensive Caucasian and Micronesian admixture on Kosrae.

Methods of linkage analysis of complex traits

There are hundreds of candidate genes for the Syndrome X disorders. These include genes identified as causing the rare Mendelian forms, as well as many genes or syntenic regions from animal (particularly rodent) models. Examples are leptin and leptin receptor which are involved in genetically obese mice (*ob* and *db* respectively), the MODY loci for diabetes, ion-exchange genes and the SA locus (identified from rat models) in hypertension, and the familial combined hyperlipidemia loci for dyslipidemia (Zhang et al. 1994; Nabika et al. 1995; Chen et al. 1996; Dallinga-Thie et al. 1997). Yet there are many more, as virtually any gene related to the many aspects of metabolism and hemodynamics could play a role. Therefore complete genome scans for the quantitative traits underlying these four disorders: obesity (BMI, WT, WST, HIP, and LEP), diabetes (FBS and INS), hypertension (SBP, DBP, and MAP), and dyslipidemia (TC, TG, APOA-I, and APOB), as well as height (HT), were carried out.

Though many linkage studies in different populations have been done for these disorders, results have been inconsistent with replication in a limited number of instances (Guo 2000; Altmuller et al. 2001). This could be due to lack of consensus on exactly the best way to conduct many aspects of these studies, to both maximize power and reduce false positives. To that end, the Genetic Analysis Workshop (GAW) 10 was designed to assess relative power, accuracy of parameter estimates, and error rates for methods of complex disease gene mapping (Wijsman and Amos 1997). As each gene has a smaller, perhaps non-independent effect, there may not be a clear, direct connection between phenotype and genotype, so methods need to elicit the most information possible from the complete dataset.

The first suggestion was to analyze continuous (quantitative) traits related to the disease. Often, a lot of information is lost when quantitative traits are dichotomized to determine affection status, so it is preferable to use the continuous measure itself. For diseases that are not directly determined by an underlying distribution, there may be a simpler endophenotype (often quantitative) correlated to the disease. Further suggestions pertain to both qualitative and quantitative trait analysis. Correcting for covariates either before or during the analysis has a sizable effect, as it explains much of the non-genetic effects, making it easier to model the effects of the major genes. The use of large extended pedigrees increases power both because of the increased sample size and more phase known meioses, as well as the ability to trace genes through many generations. Additionally, there is a huge benefit to using many linked markers in multipoint analysis, as extra information is obtained from neighboring loci, leading to better localization and accuracy. To a lesser extent, modeling epistasis, pleiotropy, multivariate traits, and other interactions further increases power and accuracy and decreases error rates. The problems are the many challenges (mostly computational) to multipoint analysis of quantitative data while correcting for covariates in large extended pedigrees. Linkage analysis (LA) methods to do this are needed.

Linkage analysis: history and methods

Linkage analysis is a powerful method used to map disease genes, as it looks for correlated segregation of marker alleles and disease genes within families. The principal is that the linked markers (haplotype) surrounding the mutation will be transmitted along with the mutant gene within a pedigree; therefore, the identification of markers linked to a disease suggests a gene there as well. This method of finding genes based on their chromosomal location began with Morgan's 1911 discovery of linkage, and Sturtevant's 1913 use of it to map trait-causing genes to drosophila chromosomes (Morgan 1911; Sturtevant 1913). Linkage tests for major genes were developed starting in 1931, with the first one applied in 1935 (Penrose 1935; Ott 1999). Traditional parametric analysis (model-based) formally proposes a model, that a marker locus is linked (recombination fraction, θ, <0.5) to a disease locus, based on recombination events. These parametric tests require pedigree information, one-locus inheritance models, allele frequencies, penetrance functions, and allowance for heterogeneity. Linkage is measured in terms of the LOD score, the base 10 logarithm of the likelihood ratio between likelihood of linkage to the null of no linkage ($\theta = 0.5$), with a genome-wide p-value of 0.05 indicated by a score of 3.

These model-based methods have been very successful in localizing many monogenic Mendelian disorders, but less so for complex disorders or the associated quantitative traits. One reason is the inability to define the correct inheritance models for these traits, which has been shown to reduce mapping power (Elston 1998). Therefore, model-free (non- or quasi- parametric) methods that do not require complete parameter definition have been developed. These are mainly allele-sharing methods, which look for pairs of relatives that share disease or trait status and alleles at a locus identity-by-descent (IBD, same alleles are of same grandparental origin). Linkage is suggested by phenotypically similar pairs sharing alleles in excess to that expected by chance, inconsistent with random segregation, as opposed to by evidence of actual cosegregation between the alleles and trait. The allele-sharing methods best able to analyze quantitative traits with covariate corrections and multipoint markers in large pedigrees are based on variance components, such as those implemented in SOLAR (Sequential Oligogenic Linkage Analysis Routines) (Blangero and Almasy 1997; Almasy and Blangero 1998; Blangero et al. 2001). Variance components methods model the variance of a trait attributable to a locus by modeling covariance between relative pairs as a function of

IBD. These methods are excellent for complex disorders, as they can model multiple loci, epistasis, pleiotropy, and gene-environment interactions. Pedigree size is still a problem as it is necessary to be able to write out the multipoint pairwise IBD-sharing matrix.

Another main problem with parametric methods is the inability to perform multipoint analysis in extended pedigrees. (This section is mainly summarized from Heath 2002). This is mainly a missing data problem, as for each unknown every possibility has to be specified exactly. Solutions for each aspect (pedigree size and multipoint analysis) have been developed separately, both of which are actually special cases of Lauritzen and Spiegelhalter's method of calculating probabilities on graphical structures, utilizing local dependencies (Lauritzen and Spiegelhalter 1988).

For pedigree size, pedigrees can be drawn first as undirected acyclic graphs, then as undirected graphs, where the nodes are the ordered (phase-known) genotypes for each individual. Each node is connected to others only locally, with spouses to each other and children to parents (Figure 1.1). Elston and Stewart used this to enable fast summation over genotypes in a pedigree, called pedigree "peeling" (Elston and Stewart 1971). This is done by sequentially removing nodes, including their linkage information in their neighbors' probability functions, until the information for the whole pedigree is

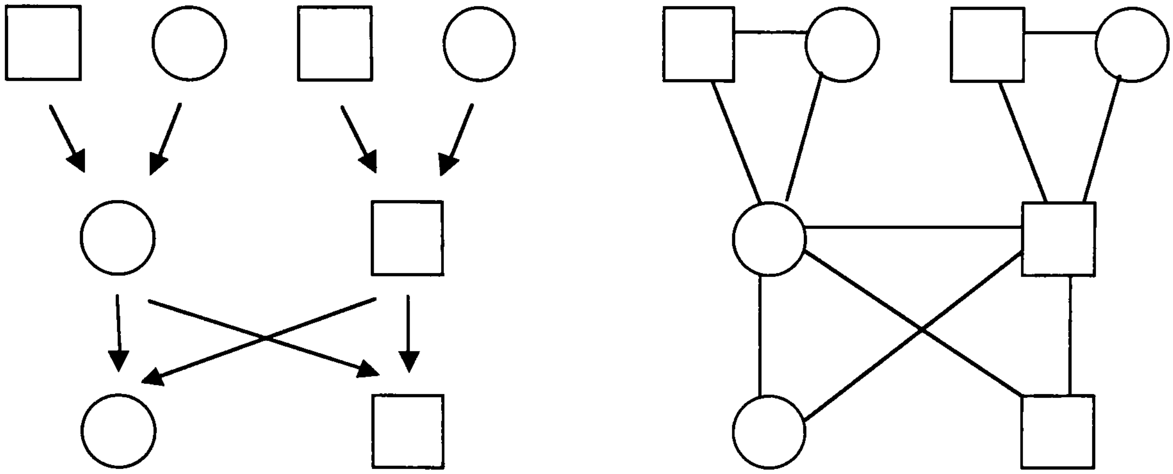

Figure 1.1: Pedigree drawn as a directed acyclic graph, showing the direction of allele transmission, and then with the edges connected to give a undirected cyclic graph.

combined into one individual. This method was extended for general pedigrees in 1978 (Cannings et al. 1978). Efficiency of peeling is based on complexity of the pedigree structure and order of combining nodes, but mainly on the number of possible combinations at each node. Pedigree complexity depends on size and the number of loops and the peeling should be done to remove and not add edges. The number of combinations increases exponentially with the number of loci, as it is based on the number of alleles at each locus. So though the Elston-Stewart algorithm works well for large pedigrees, as implemented in programs such as LINKAGE and FASTLINK, it cannot be used for multipoint peeling for more than 3-6 loci.

As a way to perform multipoint analysis, the chromosome is drawn as a linear graph, with each node (locus) sequentially connected only to those immediately adjacent. Lander and Green developed an algorithm to use this to sum over segregation indicators (SI) for each non-founder (Lander and Green 1987). Segregation indicators specify the grandparental origin of the alleles at a locus, as a binary digit, 0 for grandmaternal and 1 for grandpaternal (Figure 1.2).

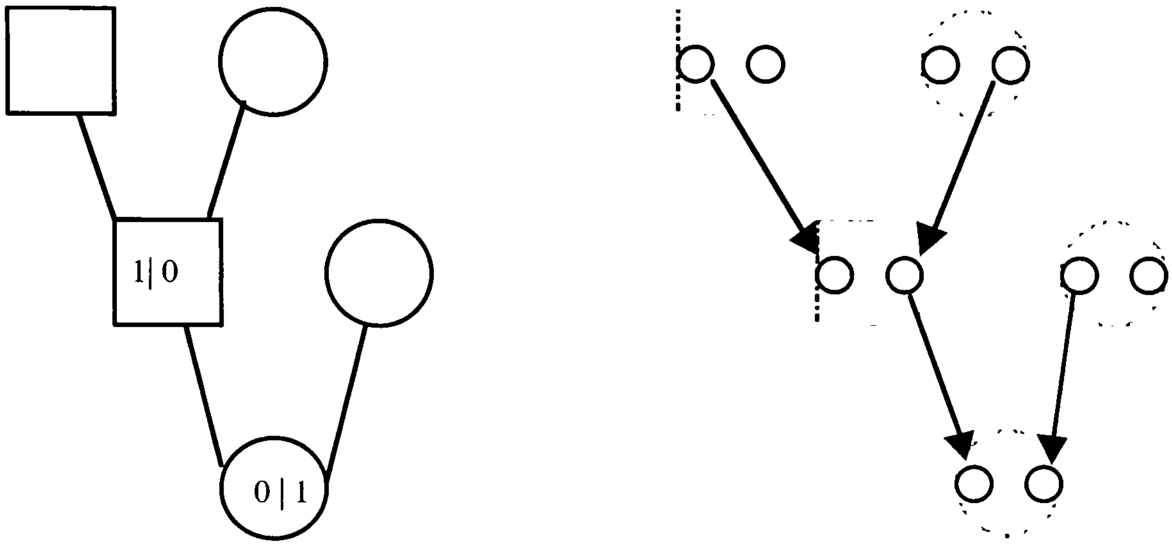

Figure 1.2: Pedigree drawn with SIs for non-founders, showing the transmission of alleles as well.

If the SIs for a meiosis at 2 loci are different, a recombination event has occurred. For each locus, the SIs can be written out as string of size 2n (n=number of non-founders). These strings can be compared for two loci, with the differences indicating recombination events. Here, the number of possible node states increases exponentially with the number of non-founders. So even though the Lander-Green algorithm (as implemented in Genehunter) is excellent for multipoint analysis, it only works for small pedigrees of less than16 non-founders.

Both Elston-Stewart peeling and Lander-Green SI recoding are actually special cases of Lauritzen-Spiegelhalter with different latent variables and directions of movement through the graph. For Elston-Stewart, the genotypes are the latent variables and the calculation moves through the pedigree, ignoring the loci dependencies on the chromosome. On the other hand, for Lander-Green, the latent variables are the SIs and the calculations move along the chromosome, ignoring the dependencies on relatives. The main limitation for both of these summation methods is the number of possible combinations at each node. If all the data are known, one does not need to sum over all possibilities, making these methods much more efficient. Even with missing data, the observed data can often reduce the number of possible configurations to make the summation more tractable (Lange and Goradia 1987; Lange and Weeks 1989). Another suggestion is to modify the dependency graphs by drawing the pedigree graph and then splitting each node into separate ones for alleles and SIs (Figure 1.3).

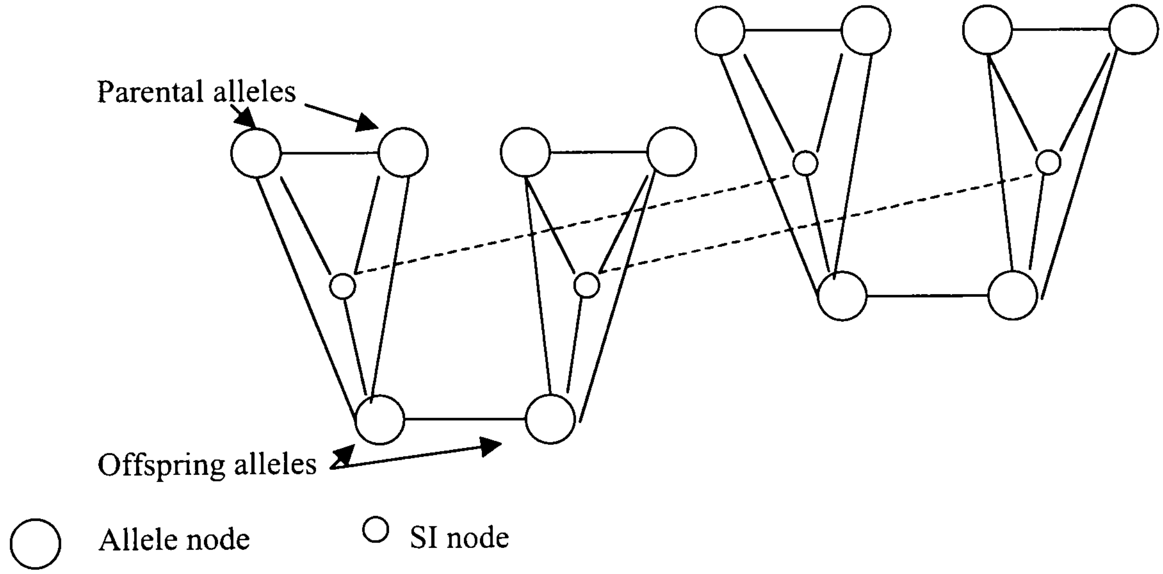

Figure 1.3: Trio drawn for two loci, separating allele and SI nodes. Allele nodes have pedigree dependencies (solid lines) while SI nodes have chromosome dependencies (dotted lines).

This adds total nodes but as more are known, the total of unknowns is reduced. Now the local dependencies can be separated out as well, as allele nodes are only dependent on pedigree members and SI nodes are only dependent on chromosomal location. By using both graphical structures separately but simultaneously, it may be possible to use large pedigrees and multipoint analysis. Yet, in real life situations with missing data, at some point there will be too many possible states to exactly sum over all combinations. Instead, sampling methods should be used.

Markov Chain Monte Carlo sampling

Exact methods calculate the probability of every possible configuration to determine the most probable data state. Sampling methods instead sample from all the possible data configurations, and the amount of times each configuration comes up is proportionate to its probability. Therefore the most likely configuration will be sampled most often, with the assumption that if the sampler is run infinitely long, the answer would be infinitely close to that extracted from the exact calculations (Gilks et al. 1996; Roberts 1996).

One such method is Markov chain Monte Carlo (MCMC), which provides a framework to analyze such complex problems without having to force simpler models (Gilks et al. 1996). This includes the ability to do both multipoint analysis and use large pedigrees, but also permits the use of complicated inheritance models, another limitation in traditional linkage analyses. So instead of doing exact integration over high dimension probability distributions to make inferences about model parameters and predictions, Monte Carlo integration is used to draw samples from the required joint distribution and the sample averages approximate the parameter estimates. In a genetic context, the probability distributions for the underlying parameters include phase-known genotypes for markers and quantitative trait loci (QTLs), allele frequencies, covariate effects, number of QTLs, position of linked QTLs, variance attributable to linked QTL, and residual variance. All these are sampled together to model a trait, and suggest the most likely regions to contain a QTL with particular parameters, such as allele frequencies and percent variance explained.

Many of the sampled configurations, though, may be impossible due to the restrictions on the sample space imposed by the data. Therefore, a Markov chain is used to move from one state (configuration) to the next, thereby directing the sampling and increasing efficiency. Using a Markov chain method, one state is dependent only the one before it (local dependency), and therefore it does not move from a possible state to an impossible one (Metropolis et al. 1953; Hastings 1970; Gilks et al. 1996; Roberts 1996).

Markov Chain updating

The Markov chain has to be constructed appropriately to have several important characteristics. First, it should converge, to settle down without major fluctuations, which suggests that the proportions of the underlying distribution has been reached (Gilks et al. 1996; Roberts 1996). Second, the chain has to mix, to be able to move over all of the possible sample space. This is necessary so that the sampler can find all the areas of high probability and not get stuck at a local maximum. This is even harder to ensure when there are correlations among the samples, so it is hard to move from that space. One improvement, which will be discussed further in the genetic context, is to perform block updates to change all the correlated samples together (Gilks et al. 1996). Then, the chain has to be run long enough for the sampled distribution to reach the underlying distribution (Gilks et al. 1996; Roberts 1996). Often the easiest way to diagnose these issues is by visual inspection. By looking at the iterations, convergence can be seen, as well as how often the sampler leaves and returns to the distribution to which it converged (mixing). Based on this it can be decided if the sampler was run long enough. This will be shown on real data later.

Construction of the chain means choosing how to move from one configuration to the next, by update steps. This consists of suggesting a change in state, and either accepting or rejecting the move to the new configuration, based on the acceptance ratio. There are a number of methods implemented in various samplers to do this, many of which are generalization of the Metropolis algorithm, and hybrids of different algorithms. The main samplers used in the genetic application discussed below, Loki, are the Metropolis-Hastings samplers and the Gibbs sampler (Hastings 1970; Kong 1991; Gilks et al. 1996; Roberts 1996; Heath 1997).

The Metropolis-Hastings acceptance ratio is based on the state and proposal probabilities, which of the states (present or proposed) is more likely, and which move, from the present to the proposed or vice versa, is more likely. Using this step, a change can be to a single element of the joint distribution or to groups of elements. For example, a change suggested could be in the position of a linked QTL, or in linkage status, percent variance attributable to a previously unlinked QTL, and allele frequencies all at once. A special case is the Gibbs sampler, which is mainly used to make a particular update, as it is always accepted. This typically changes one element at a time, with the new value sampled from the joint distribution, conditional on the rest. Additionally, in Loki, the number of QTL in the model can also be changed. This changes the dimensions of the sample space, and requires an extension to Metropolis-Hastings called reversible jump, with an acceptance ratio that takes the change of dimension into account (Green 1995; Richardson and Green 1997).

Applications to genetics

As MCMC utilizes local dependency to move from one state to the next, it makes sense to apply these methods to genetic data that is characterized by local dependencies, between pedigree members and loci as discussed above. The use of MCMC for genetic analysis was first suggested in 1989 (Sheehan et al. 1989) and has been discussed since in the literature, such as in the following: (Lange and Sobel 1991), (Thompson 1991), (Kong 1991), (Sheehan and Thomas 1993), (Geyer and Thompson 1995), (Lin 1995), (Thompson 1996), (Thomas and Gauderman 1996), (Heath 1997; Heath et al. 1997), and (Thompson and Heath 1999). The earliest human genetic application was a single locus Gibbs sampler, where one individual is updated conditional on the other pedigree members (Geman and Geman 1984; Sheehan et al. 1989). This does not mix very well, both due to the presence of states impossible to move through and because of the correlation between relatives. This can be improved by block sampling, by updating the whole pedigree at once using Elston-Stewart peeling (Ploughman and Boehnke 1989). With extension to multiple loci, mixing is again a problem due to the correlation between adjacent loci. Here, the block sampler is used with single-locus peeling to sample genotypes at a locus for the whole pedigree at once, conditional on the Lander-Green SIs of the other loci, or on the ordered genotypes at adjacent loci (Kong 1991). Another suggestion is to sample SIs for a meiosis instead of genotypes, conditional on the data and SIs for other meioses (Thompson 1994a; 1994b; Thompson and Heath 1999).

Two genotype samplers used in Loki are the locus and meiosis samplers (Heath 1997; Szydlowski and Heath 2000)). The locus sampler updates one locus at a time, conditional on the pedigree and adjacent loci. Efficiency of mixing is dependent on how close the markers are, as there is still correlation between the segregation patterns at adjacent loci. The meiosis sampler updates SIs at all loci for a particular meiosis, conditional on the data and SIs for other meioses. As this is dependent on the number of meioses, it mixes poorly with a lot of data. So depending on the type of data available, which sampler or combination of the two can be selected (Szydlowski and Heath 2000). These samplers are not yet optimized, and there is a need for good sampling schemes that condition on the minimum number of variables to allow efficient updating of the rest.

Loki

The LA method used in this study, called Loki, uses MCMC simulation to generate multilocus inheritance patterns, conditional on the marker data, to test for the best model, including the most likely locus for a trait (Heath 1997). Through the sampling of the underlying distributions, Loki generates estimates for the number of trait loci, allele frequencies, number and positions of linked QTLs, and the percent variance explained by each QTL. As this is a sampling method, Loki can realistically model quantitative traits by correcting for covariates and allowing for multiple QTLs and do multipoint analysis in extended pedigrees. (The limit on the pedigree structure is that the pedigree has to be single-locus peelable, which limits it to less then about eight interlocking loops.) Therefore, this method can fulfill the main suggestions from GAW10, to be able to powerfully and accurately map genes for complex diseases.

In the Loki analysis, the trait is modeled as $y = \mu+X\beta+\Sigma_{i=1 \text{ to } k}Q_i\alpha_i+e$, where μ is the grand mean, β are the covariate effects, X is the incidence matrix for the covariates, k is the number of QTL for the trait, Q is the linked QTL, α is the QTL effect (dominance or additive effect), and e is the residual variance. Loki samples from the joint posterior probability distribution for all unknown parameters and inferences for the individual parameters can be drawn from the marginal distributions. The joint probability is written as p $(k,G,M,\beta,\lambda,\delta,\eta,\alpha,\sigma^2e,\mu,Y)$, where G and M are the ordered (phase-known) genotypes at the QTL and markers respectively, λ is the position of a linked QTL, δ is linkage status for the QTL, η are the allele frequencies, σ^2e is the residual variance, and Y is the observed data, which includes the trait, covariate, and genotype data.

All these parameters go into the Metropolis-Hastings acceptance ratio, into the ratio of the state probabilities, as the probability of the observed data (Y) given the rest of the parameters with the new parameter state, divided by the probability of the data given the rest in the current state. For example, if that update step proposes a change to the QTL position, to λ^*, the numerator will be p $(Y \mid k,G,M,\beta,\lambda^*,\delta,\eta,\alpha,\sigma^2e,\mu)$, while the denominator will be p $(Y \mid k,G,M,\beta,\lambda,\delta,\eta,\alpha,\sigma^2e,\mu)$. How well the configuration models the trait will affect the acceptance. At some steps, it is more efficient to sample conditional on a portion of the joint distribution, and then there is a Gibbs step to update the elements that were not conditioned on. The trait model parameter approximates can be estimated from the distributions generated in the sampling scheme.

At the start of an analysis, there is the observed data, which includes phenotypes, genotypes at the markers, and covariate values, and a model (which covariates to use and a general genetic effect, such as trait = age + QTL). The sampler is initialized with marker allele frequencies as they are, covariate effects at 0, and residual variance as phenotypic variance minus the grand mean. Then genotypes are sampled at each locus alone, generating genotypes for those that were missing, and new allele frequencies are calculated. Next, working across the chromosome, the genotypes at each marker are updated (whole pedigree at once) conditional on neighboring loci and allele frequencies, using the locus sampler. Then a random position is picked, a random mode of inheritance (additive or dominance) that explains a proportion of the residual variance is selected, allele frequencies for a gene with this size effect are generated, and all that is put into the sampler.

Then the sampler (Markov chain) runs for tens to hundreds of thousands of iterations, with the update steps as follows. At each iteration, the ordered genotypes are updated at the marker loci. Then for each QTL the following parameters are updated: gene effect, position, linkage state (linked or not) and then ordered genotypes at the QTL are generated by reverse peeling. The first three QTL updates are done independent of the QTL genotypes, as they are integrated out of the acceptance ratio, as if the conditioning is on the sum of all genotypes in the pedigree. Then these are updated: allele frequencies, covariate effects and grand mean, and residual variance. The last step is adding or removing a QTL, which is a reversible jump step, as it changes the dimensions of the joint distribution. Many parameters are added or removed along with a QTL, such as percent of variance attributable to it, position, linkage status, and allele frequencies, and how well they explain the data will affect the acceptance ratio.

To identify chromosome regions likely to contain QTLs, the chromosomes are divided into equal sized bins of 1cM. At each sampling iteration, the prior probability of linkage of at least 1 QTL to a particular bin is calculated. For a single QTL, the probability of linkage to any bin is s/t, where s is the bin size and t is the total map length of the genome. If there are n QTLs in the model at a particular iteration, the prior probability of at least 1 QTL being located in the bin is $p = 1\text{-}(1\text{-}s/t)^{n}$. The posterior probability *q* is simply 1 or 0 depending on whether at least 1 QTL is located in the bin, and the Bayes factor (ratio of posterior probability to prior probability) for a bin is then estimated by averaging *q/p* for that bin over all iterations. This estimate is termed an L-score, and the plot of L-scores across a chromosome is called an L-graph. L-scores can be understood as how much excess evidence there is for real linkage to that region (bin) compared to being unlinked (random).

It is tempting to compare Log_{10} (L-score) with a conventional LOD score, but this is not valid, as an L-score of 1000 does not have the same meaning as a LOD score of 3. Firstly, test statistics only have real meaning for the tests for which they were developed, as the cut-offs are chosen to give a particular p-value based on the underlying distribution of the test statstic. So it is difficult to directly compare disparate test statistics from very different methods. Secondly, there is not a clear definition of LOD scores when there are multiple trait loci, while the L-score concept extends naturally to multiple (and even variable numbers of) QTLs. Thirdly, in contrast to LOD scores, the L-score takes into account the uncertainty in the estimation of the covariate effects, marker allele frequencies and the number of QTLs. As L-scores are genome-wide (because the prior probability of linkage is dependent on genome size), a score of 20 would have a p-value of 0.05, theoretically comparable to a LOD of 3. But the distribution of the L-score (to give empirical p-values) has not yet been evaluated. Therefore, exactly how L-scores and L-graphs are used to identify regions of interest will be discussed in detail later.

In general, what is really being asked is if a QTL explains some of the variance in the trait, and localization is based on summing over QTL genotypes. QTL alleles are symmetrical, meaning that in one iteration one allele is the "disease" allele and then it switches to the "normal" and vice versa. As the dominant and additive effects are based on the particular combination of disease / non-disease alleles at that iteration and they are reassigned each iteration, there are often four possible models. This is even more evident when allele frequencies are extreme, and therefore one allele combination (homozygote rare allele) is virtually non-existent. Therefore, these estimates are not useful on their own. Rather a combination of these effects is calculated to give the QTL effect size. This is the square root of the weighted average of the variance explained by the dominance and additive models, weighted by the respective allele frequencies. As this is the root of a variance, it has no sign, which means that no statement can be made as to the direction of the QTL effect, but it is in the same units as the trait being measured. Once there is evidence of a QTL, there are experiments that can be done to investigate the mode of inheritance, which is beyond the scope of this work.

As discussed before, one of the most important characteristics for a sampling scheme is to make sure the sampler is mixing. This is done by visually inspecting the graphs of iteration by centimorgans (by chromosome) to see if the same locus is found and that the sampler moves off that spot but comes back. This is even worse with markers closer than 5 cM, which cannot be used in the current implementation. Improvements for mixing include using the meiosis sampler to update genotypes, but this is under constant investigation. Additional improvements to Loki include allowing greater than two alleles at a QTL and modeling epistasis, gene-environment interactions, undetected errors, and heterogeneity.

Loki is a tremendously promising sampling LA method that can model quantitative traits with multiple QTL and correction for covariates and carry out multipoint analysis in large complex pedigrees. As it has not yet been widely used for genome scans on real data, it is still unclear exactly how to carry out such a study, and how to understand the output. This study aims to evaluate how Loki works and develop a strategy for a genome scan, first in simulated data (GAW12) with known answers, and then in real data using the Kosrae dataset (Almasy et al. 2001). Included in developing this strategy is to learn how to judge which linkage signals are real, and decide how to prioritize linkage regions that are reliable. Once this strategy is elucidated, we will use it to identify real signals for the metabolic traits under study in the Kosrae population.

To map genes in the Kosrae population, first all the necessary data had to be collected. This included data from the population screen for familial information, phenotypes, and covariates. This information was used to draw the pedigree, define the traits being studied, and calculate covariate effects for trait models. Then individuals were genotyped for the genome scan markers and the data (both pedigree and marker) was cleaned-up prior to analysis. Once the dataset was complete, it was analyzed using Loki, as described below.

Chapter 2: Materials and Methods

Epidemiology and phenotype definition

The data collection and phenotype definitions on the island of Kosrae is explained in detail in Shmulewitz et al (Shmulewitz et al. 2001). The screen was carried out by Steven Auerbach of the US Health Department, Maude Blundell of the Starr Center for Human Genetics at the Rockefeller University, and Vita Skilling of the Kosrae hospital, as well as the medical staff on the island of Kosrae. To summarize the salient features, 2188 Kosraeans (over 90% of adult population) were screened for Syndrome X traits. All participants signed the informed consent prior to filling out questionnaires about medical and family history, physical examinations and blood drawings. Each person was assigned a unique identification number and filled out a questionnaire which noted the participant's sex, family data including listing of biological parents, siblings, and children, smoking status, village of residence, age, and health status. The family data was used to derive the parity status of the women (0 to 5 and ≥6 children). Though there are five villages on Kosrae, because of similarities in life style four were considered as one group (Lelu, Malem, Tafunsek, and Utwe), with Welung as the other group. Three intervals were set up for age: I (20-34), II (35-49), and III (≥50).

HT, WT, WST, and HIP were measured. BMI was calculated as WT (kg) divided by HT squared (m^2). Although in the US recently the obese phenotype has been defined as a BMI ≥30, due to the distribution of BMI in Kosraeans obesity was defined as a BMI ≥ 35 kg/m^2 (Flegal et al. 1998) for the purposes of this study. WST and HIP were used to detect more central patterns of obesity. The average of three SBP and DBP measures, taken manually with a stethoscope and sphygmomanometer, was used in the analysis. MAP was calculated as DBP+[SBP-DBP]/3) and the hypertensive phenotype was defined as SBP ≥140 mmHg or DBP ≥90 mmHg (Anonymous 1993). Diabetes was defined according to the American Diabetes Association (ADA) recommendations as FBS ≥ 126 mg/dl, as measured by fingerstick with a glucometer or OGTT2 ≥ 200 mg/dl (Anonymous 1997).

A venipuncture was done for determination of serum LEP, INS, TC, TG, APOA-I, and APOB, as well as for the isolation of DNA. LEP levels were determined using a commercial radioimmunoassay (Linco, St. Louis, MO) by Jeffrey Winick in Jeffrey Friedman's lab at the Rockefeller University. INS concentrations were measured on 740 randomly selected samples using a sensitive enzyme-linked immunosorbent assay (ELISA), at Graeme Bell's lab at the University of Chicago (Hartling et al. 1985). TC and TG were measured with commercially available enzymatic kits (Boehringer Manheim, Indianapolis, IN, Sigma Diagnostics, St. Louis, MO) in Jan Breslow's lab at the Rockefeller University. The cutoff for hypercholesterolemia was defined as TC ≥ 240 mg/dl and for hypertriglyceridemia as TG ≥ 200 mg/dl, as discussed in the National Cholesterol Education Program (Anonymous 1994). LDL-C and HDL-C quantification was not possible due to the lack of an adequate centrifuge on the island and the inability to ship serum overnight at 4° C. As substitutes, the major apolipoproteins of LDL-C and HDL-C (APOB, and APOA-I, respectively) were measured using a standard double antibody immunoassay technique at Linda Youngman's lab at the Oxford University. Both APOB and APOA-I are highly correlated with their respective cholesterol fractions, making them excellent surrogates (Bachorik et al. 1997). High risk for cardiovascular disease was considered to be APOB ≥ 120 mg/dl, as suggested by the Framingham Offspring Study, and APOA-I ≤ 88 mg/dl, the lower tenth percentile of the Kosraean distribution, as Framingham suggestions for cutoffs for APOA-I were not appropriate due to the distribution in Kosraeans (Contois et al. 1996a; Contois et al. 1996b).

Statistical analysis was carried out using the JMP package (SAS Institute Inc., Cary, NC). Means for all quantitative traits were calculated for the total population, as well as for subgroups determined by the covariates: sex, parity (0-5 children vs. ≥6 children), smoking (male smokers vs. male non-smokers), village (other four or Welung), and age (20-34, 35-49, and ≥50). Disease frequency was determined by previously mentioned cutoffs in the quantitative traits, and the effect of covariates on disease risk was evaluated through univariate logistic regression analysis. Multivariate logistic regression analysis was also done to evaluate the effects of all 5 covariates together on disease risk. Univariate logistic regression analysis was used to calculate the risk of one disease phenotype given another. In all further analyses, the 4 quantitative variables that were not normally distributed, LEP, FBS, INS and TG, were log transformed. Quantitative variables were corrected for covariates before Pearson correlations were calculated.

Factor analysis

Multivariate factor analysis was performed to determine which quantitative traits changed together in this population (Stevens 1986). Factor analysis was carried out on 628 individuals, 255 males and 373 females. This sample set was obtained after excluding subjects missing information on INS levels or other variables. This subset had similar features to the entire dataset (data not shown). Diabetics were excluded from the factor analysis because their FBS levels did not conform to the normal distribution of the log transformed FBSs of the rest of the population. Factor analysis was performed using principal components analysis (PCA) followed by orthogonal rotation of the following continuously distributed variables: WT, WST, LEP, FBS, INS, DBP, SBP, TC, TG, APOB, and APOA-I. This analysis results in uncorrelated factors, which are interpreted based on factor loadings, the correlation between the factor and the original variable. Loadings of 0.4 (absolute value) or greater are considered definitive of the factor ($p<0.001$), as they share at least 15% of variance with the factor, while loadings of 0.2 (absolute value) or greater show variables that are significantly correlated to the factor ($p <0.01$) (Stevens 1986). Additional factor analyses were carried out, changing both samples used (excluding INS to allow for more) and including HIP and HT, with very similar results (data not shown). Lastly, factor analysis was carried out on the residuals, after correcting for sex, smoking, age, parity, and village. This mainly affected the fourth factor, which is reported.

Genetic Epidemiology

Prior to undertaking a complete genome scan, to ensure that there were genetic factors involved in the quantitative traits under study, heritability (h^2) was calculated and segregation analysis (SA) was carryied out. h^2 is a measure of the percentage of the variance in the trait that is attributable to the additive effects of genes, and is calculated in families, based on the relationship between the known genetic correlations of relative pairs and their phenotypic similarity. Heritabilities were estimated using the MORGAN computer package using all available pedigree information, with corrections for sex, age, parity, and smoking simultaneous to the estimation of heritability (Thompson and Shaw

1992). SA aims to show the amount of genes involved, their mode of inheritance, and magnitude of effect. SA involves fitting various inheritance models based on the distribution of a variable (phenotypic information) within a pedigree. SA was carried out for each trait, with corrections as suggested in the regression analysis described above (models for each trait are listed in Table 5.1) using the linkage package Loki (Heath 1997). Loki was used for SA in the same way as linkage analysis (discussed below) but without any genotype data, resulting in the number of QTL and the amount of variance explained by each.

Microsatellite marker genotyping

Genomic DNA from each of the Kosraean participants was extracted from a 20 ml peripheral blood sample using standard 'salting-out' procedures by Stefano Signorini, in Markus Stoffel's lab at the Rockefeller University. Approximately 300-500 micrograms of genomic DNA were extracted from each sample. DNA concentrations were determined by the modified fluorometric assay (Hopcroft, D.W. et al, 1985) using bisbenzimidazol (compound Hoechst 33258, Aldrich Chimica, Milano, Italy) as fluorochrome and human placental DNA (Sigma) as a standard. Each DNA sample was diluted to 10 ng/μl and aliquoted into a 96-well MicroAmp plate prior to amplification. 1564 of these samples (those part of the large pedigree) were put into five 384-well plates and genotyped in two stages. (An additional 120 individuals were typed for a number of the initial markers.) The first stage included 1102 individuals with about 450 microsatellite markers typed and the second stage included 462 individuals typed for about 390 markers. This was because the first set was a training set, used to develop a working set for all future projects in the lab.

A high-throughput center for genotyping using microsatellite markers has been established at the Rockefeller University under Markus Stoffel. The director of the Genotyping Core Facility (GCF) is Lynn Petukhova, with technical help from Sandra Oplanich and Merav Yifrach. The protocol for genotyping is available on their website (www.rockefeller.edu/genotype.html) and summarized below. For the genome scan, the markers started with were from the Marshfield set 9 (www.marshfieldclinic.org /research/genetics/). This set was modified as about 100 of the markers were replaced or removed for a number reasons discussed later. Panels (markers loaded together) were set up after allowing for four alleles to each side of the CEPH ranges, as in many cases novel alleles of those sizes were found in this population. Panels used in the second stage are available on the GCF website as well.

The protocol for genotyping is as follows. First, the 384 well DNA plates were replicated to a number of plates equal to the number of markers in the panel, using a Hydra 384 well micropipetter (Robbins Scientific, Sunnyvale, CA). Second, per marker, a PCR master mix (described below) was set up and added to a plate, using a Genosys 100 automated sample processor (Tecan, Research Triangle Park, NC). Then amplification was performed on Applied Biosystems 9700 thermocyclers (Foster City, CA). After PCR, the products were pooled, aliquoted, and combined with loading buffer using the Hydra 384. Next, products were heat denatured. Last, products were run first on Applied Biosystems 373 excel slab gel electrophoresis machines, and then on the Applied Biosystems 3700, which uses capillary gel electrophoresis to separate fluorescently labeled PCR products.

All PCR reactions were carried out using 20 ng of template DNA (10ng/μl) in a 10 μl total reaction volume, with a final MgCl2 concentration of 2.0 mM in the PCR buffer (10 mM Tris-HCl (pH 9.2) and 50 mM KCl). Each reaction contained 250 nM each of forward and reverse primer, 250 μM of each dNTP (dATP, dCTP, dGTP, dTTP), and 1 unit of Taq DNA polymerase. The standard PCR protocol was: three step PCR + 10 min extension at 72°C, 94°C for 5 min, followed by 3 cycles of 94°C for 30 sec, 60°C for 30 sec, and 72°C for 1 min, followed by an additional 29 cycles of 89°C for 30 sec, 55°C for 30 sec and 72°C for 1 min, plus a final extension step at 72°C for 10 min. The forward strand of each primer pair was labeled with one of three phosphoramidites: 6-FAM (blue), HEX (green), or NED (yellow). For the second set, VIC replaced HEX and PET (red) was added.

Markers were pooled (typically, total volume 10μl) adjusting individual marker volumes (range 1-5 μl) to give electropherogram peak intensities of 1000-2000 fluorescence units. 1.5 μl of the PCR pools was added to 2.5 μl of loading dye cocktail, comprised of formamide, blue dextran loading buffer, and DNA size-standard at a ratio of 5:1:1, respectively. The GS-400 size standard, with size range from 35-400 bp, was ROX-labeled (red) for the first set, and orange in the second set.

The resulting sample files from the electrophorsis were extracted using the Applied Biosystems Genescan program, and each panel was scored using templates defined with the Applied Biosystems Genotyper program. The template was set to define the individual marker's alleles, and the fluorescent dye used; alleles were called automatically on the basis of those parameters. Then the alleles were checked by two people independently for the first set. Other scorers for this included Jeffrey Winick, Elizabeth

Stephenson, Chavie Wolfish (Friedman Lab), Rebecca Breslow, Stephanie Boger (Stoffel lab), Zhihua Han (Breslow lab), Brandie Galke, Zoltan Takacs (Maria Karayiorgou lab at the Rockefeller University) and Merav Yifrach. After allele peaks were called, each genotype data set was imported into a Filemaker database, written by Peter C. Verlander of Arlene Auerbach's lab at the Rockefeller University, and the two sets of scored alleles were compared to resolve discrepancies. Then the consensus data was screened for nuclear family Mendelian inconsistencies, and errors were checked manually. Next, whole pedigree Mendelian errors were detected and removed by Loki, as described later. For the second set, the genotyping was very clean, therefore there was only one scorer for these data, all inheritance checks were done using Loki, and errors were checked and corrected manually.

Linkage analysis

Preliminary marker analysis was carried out using the SOLAR package, as described in the manual (www.sfbr.org/sfbr/public/software/solar). This analysis was carried out on a subset of the typed markers. As SOLAR cannot handle pedigrees as complex as the Kosrae one, first the pedigree was broken up into nuclear families (Dale Nyholt in the Ott Lab at the Rockefeller University) and then into 2-3 generation families (Mark Levenstien and Derek Gordon in the Ott Lab). The larger pedigrees were generated mainly by removing the individuals with no information, those which had been added in as connectors. This resulted in close to 630 individuals with information (out of the 1102 available). Next the genotype data was checked for Mendelian inconsistencies using the PedCheck program and errors were zeroed out manually (O'Connell and Weeks 1998). All traits were Box-Cox transformed (for normality) using the Unicorn program prior to analysis (Allison et al. 1995). The trait models were set up by using SOLAR to test which covariates explained a significant proportion of the trait variance. Complete genome scans with a subset of the scan markers (about 250) were carried out on the quantitative traits associated with Syndrome X by Mark Levenstien.

Complete linkage analysis for the genome scans on the same quantitative traits was carried out with the MCMC program Loki. This performs linkage analysis on quantitative traits using oligogenic trait models, where the number of QTLs as well as the QTL frequencies, positions, and effects, are parameters to be estimated. The method is described in detail in Heath (Heath 1997). Detail on exactly how the analyses were carried out for this project is described in Chapters 4 and 5.

Chapter 3: Population characteristics

Introduction

The first phase of this study was to analyze the clinical data from the population screen. First the covariates such as sex, parity, smoking, village of residence, and age, were looked at. Next the prevalence of each of the Syndrome X disorders (obesity, diabetes, hypertension, and dyslipidemia) and the quantitative traits associated with each were investigated. After that the covariate effects on disease risk and the quantitative traits were calculated. Clustering of the qualitative and quantitative traits using regression and factor analyses was also studied.

General covariates

The general characteristics of the population are shown in Table 3.1.

Table 3.1: Characteristics of the population sample

Covariates		
n (M/F)		2188 (908, 42%/1259, 58%)
Parity*	0-5 children	809 (64%)
	≥6 children	450 (36%)
Smoking**	yes	368 (42%)
	no	509 (58%)
Village	Lelu	768 (35%)
	Malem	497 (23%)
	Tafunsek	485 (23%)
	Utwe	334 (16%)
	Welung	70 (3%)
Age (years)		42 ± 14
Intervals:	I (20-34)	758 (35%)

II (35-49)	810 (37%)
III (≥50)	599 (28%)
Quantitative traits	
BMI (kg/m^2)	31.0±5.7
WT (lbs)	168.6±34.1
WST (in)	35.9±5.0
HIP (in)	40.7±4.9
LEP (ng/ml)	25.0±24.0
FBS (mg/dl)	94.7±36.2
OGTT2 (mg/dl, n = 943)	163.4±92.7
INS (mU/ml, n = 740)	17.3±19.9
SBP (mmHg)	119.9±17.2
DBP (mmHg)	77.6±10.3
MAP (mmHg)	91.7±11.2
TC (mg/dl)	174.9±36.7
TG (mg/dl)	96.8±53.2
APOB (mg/dl)	87.4±21.6
APOA-I (mg/dl)	117.2±24.6
HT (in)	61.9±3.3
Phenotypes	
Obesity (BMI ≥ 35)	518 (24%)
Diabetes (FBS ≥ 126 or OGTT2 ≥ 200)	260 (12%)
Hypertension (SBP ≥ 140 or DBP ≥ 90)	366 (17%)
Dyslipidemia (TC ≥ 240 or TG ≥ 200 or APOB ≥ 120 or APOA-I ≤ 88)	432 (20%)
Hypercholesterolemia (TC ≥ 240)	82 (4%)
Hypertriglyceridemia (TG ≥ 200)	98 (5%)
Hyperapobetalipoproteinemia (APOB ≥ 120)	153 (7%)
HypoapoA-I (APOA-I ≤ 88)	208 (10%)

Data for means are ± SD. *In females only ** In males only

More females (1259, 58%) than males (908, 42%) participated in the study, presumably because middle-aged men are more likely to leave the island for employment. The majority of women were married (1038/1259, 82%) and multiparous. In general, nulliparous women were younger and more likely to be unmarried or newly married. Two parity categories were assigned: 0-5 children (809, 64%) or 6 or more (450, 36%), as these groupings revealed significant difference in the frequency of the features of Syndrome X (see below).

Forty-two percent (368) of the men and 0.6% (7) of the women reported that they were smokers. On Kosrae, smoking is socially unacceptable for women, so it is possible that there was underreporting of smoking by women. Thus the effects of smoking were analyzed only in the male respondents.

The vast majority of the population considered themselves to be ethnically Kosraean (2094 individuals, 96%). There are five municipalities on the island. The most populous municipality was Lelu (768, 35%), followed by Malem (497, 23%) and Tafunsek (485, 23%), then Utwe (334, 16%), with Welung having the smallest percent of the population (70, 3%). Welung is only accessible by boat and the residents of this village generally eat a native (i.e. less Western) diet as compared to the other municipalities. As residents of Lelu, Malem, Tafunsek, and Utwe had similar lifestyles, they were considered as one group in the analyses.

The average age at ascertainment was 42 ± 14 years. Three age intervals that correspond to young adulthood, middle, and old age were used to examine age as a covariate: 20-34 (758 individuals, 35% of the population), 35-49 (810, 37%), and 50 or over (599, 27%).

Indices of Obesity

Several markers of adiposity were measured including BMI, WT, WST and HIP circumferences, and plasma LEP concentration. The average BMI was 31.0±5.7 kg/m^2 (Table 3.1) with 29% having a BMI 25-29, 35% BMI 30-34 and 24% BMI ≥ 35. For this population obesity was defined as a BMI ≥ 35.

A number of covariates strongly influenced the risk of obesity, including sex, parity, smoking, village of residence, and age (Table 3.2). Women had a higher average BMI (31.7±5.9) than men (30.1±5.2), corresponding to a 1.78 fold increased risk for obesity in females ($p<0.0001$). Females with ≥6 children had a higher average BMI (33.1±5.9) compared to those with 0-5 children (30.9±5.8) and thus had a 1.83 fold increased risk of obesity ($p<0.0001$). Among males, nonsmokers had a higher average BMI (31.0±4.7) than smokers (28.7±5.5), corresponding to 1.79 fold increased risk for obesity ($p=0.002$). Residents of the other 4 villages had a higher BMI (31.3±5.7) compared to Welung (28.0±5.4), corresponding to a 2.86 fold increased risk of obesity ($p=0.009$). This difference emphasizes that environmental factors contribute to the development of obesity. BMI was highest in middle age (32.4±4.9), and lower in both young adulthood (29.2±5.2) and old age (31.3±5.4). Thus in Kosraeans the highest risk for obesity is in middle age. The effects of covariates on the risk of obesity were also analyzed in multivariate regression analysis and a similar pattern was seen. In this case, the only significant difference was that parity was no longer significant (Table 3.3).

Table 3.2: Means by covariates

	Sex			Parity‡			Smoking‡‡		
	M	F	p	0-5	≥6	p	Yes	No	p
BMI	30.1±5.2	31.7±5.9	**	30.9±5.8	33.1±5.9	**	28.7±5.5	31.0±4.7	**
WT	178.8±34.4	161.2±31.8	*	158.7±31.4	165.7±32.2	**	172.8±37.5	182.8±31.8	**
WST	36.3±4.7	35.7±5.2	*	34.7±5.1	37.5±4.9	**	34.7±4.9	37.4±4.3	**
HIP	39.1±4.1	41.9±5.0	**	41.1±4.8	43.3±5.1	**	37.9±4.3	39.9±3.8	**
LEP	13.4±13.3	33.3±26.4	**	31.8±25.3	35.9±28.2		10.6±12..7	15.3±13.5	**
FBS	96.2±36.4	93.8±36.0		85.0±30.0	104.2±43.0	**	86.3±23.2	103.4±42.6	**
INS	18.3±28.1	16.6±10.4		16.8±10.8	16.3±9.7		15.0±28.3	20.5±28.3	
SBP	123.1±15.7	117.6±17.9	**	112.9±14.6	126.0±20.0	**	120.5±13.9	125.1±16.9	**
DBP	80.2±10.2	75.6±9.9	**	73.7±9.2	79.2±10.3	**	78.5±9.8	81.6±10.5	**
MAP	94.5±11.3	89.6±11.9	**	86.8±10.4	94.8±12.6	**	92.5±10.5	96.1±11.8	**
TC	168.4±40.5	173.5±33.8		169.3±31.8	180.8±35.8	**	171.1±38.6	181.1±41.8	**
TG	110.4±61.2	87.1±44.0	**	82.5±42.5	95.2±45.5	**	111.2±64.3	110.6±60.1	
APOB	90.2±21.6	85.3±21.3	**	82.6±19.9	90.1±22.8	**	87.0±21.6	92.5±21.5	**
APOA-I	108.4±22.0	123.3±24.5	**	123.7±24.7	122.5±24.1		108.5±23.4	109.3±21.1	
HT	64.7±2.5	59.9±2.2	**	60.2±2.2	59.4±2.2	**	65.1±2.3	64.4±2.6	*

Means ±SD. ‡ In females only. ‡‡ In males only * p=0.01 ** p≤ 0.001

Table 3.2: Means by covariates cont.

	Age Intervals				Village		
	I (20-34)	**II (35-49)**	**III (≥50)**	**Significance**	**Other**	**Welung**	p
BMI	29.2±5.2	32.4±5.9	31.3±5.4	a, b, c	31.1±5.7	28.0±5.4	**
WT	160.2±30.4	179.4±35.3	163.2±32.0	a, b	169.0±34.1	155.9±30.4	*
WST	33.6±4.5	37.0±4.8	37.2±4.7	a, c	36.0±5.0	34.3±4.8	*
HIP	39.4±4.8	41.4±5.0	41.3±4.7	a, c	40.8±4.8	38.9±4.8	*
LEP	22.6±20.2	25.0±25.0	28.1±26.1	b, c	25.4±24.3	12.4±9.8	**
FBS	80.8±15.6	95.8±36.6	110.8±46.1	a, b, c	94.8±36.6	92.6±22.4	
INS	15.9±21.3	17.6±11.9	18.4±25.8		17.3±19.9	NA	
SBP	110.6±10.8	118.8±13.6	133.1±19.7	a, b, c	120.1±17.3	114.6±14.9	*
DBP	72.9±8.6	79.2±9.9	81.4±10.6	a, b, c	77.7±10.3	73.3±8.6	**
MAP	85.4±8.8	92.3±10.6	98.5±12.7	a, b, c	91.8±11.9	87.1±9.9	**
TC	160.7±29.3	178.2±38.3	188.2±37.0	a, b, c	174.8±37.0	178.0±26.6	
TG	81.6±44.0	105.3±57.5	104.5±53.2	a, c	97.6±53.5	74.3±31.9	**
APOB	79.7±19.4	89.4±20.4	94.3±22.7	a, b, c	87.0±21.6	98.3±18.8	**
APOA-I	118.2±25.5	116.5±24.4	117.0±23.6		116.7±24.5	132.1±22.7	**
HT	62.2±3.2	62.5±3.3	60.7±3.4	b, c	61.9±3.3	62.7±3.4	

Means ±SD. * p=0.01 ** p≤ 0.001

Significant differences between age intervals, p<0.01- a – I vs. II, b – II vs. III, c – I vs. III

Table 3.3: Multivariate Logistic Regression Odds Ratios

	Obesity	**Diabetes**	**Hypertension**	**Dyslipidemia**
Sex [female-male]	0.65* [0.49-0.87]	1.66* [1.14-2.45]	1.74* [1.23-2.47]	2.95* [2.16-4.04]
Parity [0-5-≥6]	1.25 [0.93-1.66]	1.17 [0.78-1.74]	1.32 [0.92-1.91]	1.54* [1.08-2.20]
Smoking [Yes-No]	1.69* [1.16-2.50]	2.56* [1.54-4.55]	0.93 [0.64-1.39]	1.21 [0.89-1.65]
Village [Welung-Others]	2.86* [1.35-7.14]	1.37 [0.60-3.70]	3.57* [1.41-12.50]	2.78* [1.30-7.14]
Age [I-III]	3.31* [2.31-4.79]	42.14*‡ [17.4-124.9]	22.94*‡ [12.50-45.24]	2.01* [1.39-2.91]
Age [II-III]	0.36* [0.28-0.48]	2.56* [1.52-4.76]	1.37 [0.92-2.08]	1.27 [0.93-1.69]

Covariates [group 1- group 2], read as “group 2 as compared to group 1”

Odds ratio [95% confidence interval]

*$p<0.05$

‡ high value due to small number of individuals in the affected and age interval I group

The means for the other obesity measures were 168.6±34.1 lbs for WT, 35.9±5.0 in for WST, 40.7±4.9 in for HIP, and 25.0±24.0 ng/ml for LEP (Table 3.1). The covariates all had effects on these as well, as seen in Table 3.2. The mean LEP and HIP was significantly increased in females (33.3±26.4 and 41.9±5.0) compared to males (13.4±13.3 and 39.1±4.1), while males had slightly higher mean WT (178.8±34.4) and WST (36.3±4.7) than females (161.2±31.8 and 35.7±5.2, respectively). Parity had no significant effect on LEP levels, but multiparous (≥6) women had higher mean WT (165.7±32.2 vs. 158.7±31.4), WST (37.5±4.9 vs. 34.7±5.1), and HIP (43.3±5.1 vs. 41.1±4.8) than women with 0-5 children. Smokers had significantly lower mean WT (172.8±37.5 vs. 182.8±31.8), WST (34.7±4.9 vs. 37.4±4.3), HIP (37.9±4.3 vs. 39.9±3.8), and LEP levels (10.6±12.7 vs. 15.3±13.5) than nonsmokers. Residents of the 4 villages had higher mean LEP (25.4±24.3) as compared to residents of Welung (12.4±9.8), while means for WT (169.0±34.1 vs. 155.9±30.4), WST (36.0±5.0 vs. 34.3±4.8), and HIP (40.8±4.8 vs. 38.9±4.8) were slightly elevated as well. Mean WT was highest in the middle age group (179.4±35.3) with no difference between the youngest (160.2±30.4) and the oldest (163.2±32.0). WST was lowest in the youngest group (33.6±4.5) with no difference between middle (37.0±4.8) and oldest (37.2±4.7). HIP followed the same pattern, lower in the youngest (39.4±4.8) and the same in the middle (41.4±5.0) and oldest (41.3±4.7) age groups. LEP levels increased with age (22.6±20.2, 25.0±25.0) and were significantly higher in the oldest age group (28.1±26.1).

Indices of Diabetes

The mean FBS was 94.7±36.2 mg/dl (Table 3.1), but the distribution was skewed to the right with increased numbers of individuals with elevated FBS. According to the standard ADA criteria, 12% of the Kosraean population was diabetic, with a fasting FBS $\geq$ 126 or a FBS $\geq$ 200 after an OGTT2 (Table 3.1).

Risk for diabetes (based on FBS levels) was influenced by parity, smoking, and age, while sex and village of residence had no significant effect (Table 3.2). Among women, parity had a large effect, as the group with $\geq$6 children had a higher plasma FBS (104.2±43.0) than the group with 0-5 children (85.0±30.0). This corresponds to a 3.27 fold increase in risk of diabetes ($p<0.0001$). Nonsmokers had significantly higher FBS levels (103.4±42.6) than smokers (86.3±23.2), with a 4.55 fold increased risk for diabetes ($p<0.0001$). There was a significant step-wise increase in FBS levels and risk of diabetes with increasing age (80.8±15.6, 95.8±36.6, 110.8±46.1, in each group respectively). Individuals >50 years of age had a 35 fold increased risk of diabetes compared to those 20 to 34 years of age and a 3.1 fold increased risk compared to 35 to 49 year olds ($p<0.0001$). Multivariate regression analysis indicated similar effects of covariates on risk of diabetes, except that risk was significantly increased in males (OR 1.66, $p<0.05$) while, as was also seen with obesity, parity was no longer significant (Table 3.3).

Plasma INS was measured on 740 samples. The mean INS was 17.3±19.9 μU/ml (Table 3.1). INS was positively correlated with FBS ($r=0.23$, $p<0.001$). Covariates did not significantly affect INS levels (Table 3.2).

Indices of Hypertension

In Kosraeans the average SBP was 119.9±17.2 mmHg, DBP was 77.6±10.3 mmHg, and MAP was 91.7±11.2 mmHg (Table 3.1). As defined by SBP ≥ 140 or DBP ≥ 90, 17% of Kosraeans were hypertensive. All of the covariates significantly influenced blood pressure (Table 3.2). Men had higher mean SBP, DBP, and MAP than women (123.1±15.7 vs. 117.6±17.9, 80.2±10.2 vs. 75.6±9.9, and 94.5±11.3 vs. 89.6±11.9, respectively) corresponding to a 1.52 fold increased risk for hypertension in males (p=0.0003). Women with ≥6 children had higher mean SBP, DBP, and MAP than those with 0-5 children (126.0±20.0 vs. 112.9±14.6, 79.2±10.3 vs. 73.7±9.2, and 94.8±12.6 vs. 86.8±10.4, respectively) with a 3.86 fold increased risk of hypertension (p<0.0001). Nonsmokers had higher average SBP, DBP, and MAP than smokers (125.1±16.9 vs. 120.5±13.9, 81.6±10.5 vs. 78.5±9.8, and 96.1±11.8 vs. 92.5±10.5, respectively) with a 1.89 fold increased risk of hypertension (p=0.0004). Residents of the 4 villages had higher mean SBP, DBP, and MAP than residents of Welung (120.1±17.3 vs. 114.6±14.9, 77.7±10.3 vs. 73.3±8.6, and 91.8±11.9 vs. 87.1±9.9, respectively) leading to a 3.45 increase in risk of hypertension (p=0.01). SBP (110.6±10.8, 118.8±13.6, 133.1±19.7), DBP (72.9±8.6, 79.2±9.9, 81.4±10.6), MAP (85.4±8.8, 92.3±10.6, 98.5±12.7), and the risk of hypertension increased in a step-wise fashion with age in Kosraeans. In individuals over the age of 50 there was a 23 fold increased risk of hypertension compared to 20 to 34 years of age with a 3.63 fold increased risk compared to 35 to 49 year olds (p<0.0001). The effects of covariates on the risk of hypertension were confirmed for sex, village and age, but the effects of parity and smoking became insignificant (Table 3.3).

Indices of Dyslipidemia

In this population dyslipidemia was defined as TC ≥ 240, TG ≥ 200, APOB ≥ 120 or APOA-I ≤ 88. Each of these indices will be considered in turn.

Total Cholesterol

The mean TC was 174.9±36.7 mg/dl on Kosrae and hypercholesterolemia was evident in 4% of Kosraeans (Table 3.1). Several covariates including parity, smoking, and age affected cholesterol levels (Table 3.2). Women with ≥6 children had higher mean TC (180.8±35.8 vs. 169.3±31.8) with a 2.52 fold increase in risk of hypercholesterolemia ($p=0.002$). Non-smoking males had higher cholesterol levels than smokers (181.1±41.8 vs. 171.1±38.6) although there was no difference in the risk of hypercholesterolemia. There was a significant step-wise increase in TC levels (160.7±29.3, 178.2±38.3, 188.2±37.0) and the risk of hypercholesterolemia with advancing age. In individuals over the age of 50, there was a 9.4 fold increased risk of hypercholesterolemia compared to 20 to 34 year olds and a 2.5 fold increased risk compared to 35 to 49 year olds ($p<0.0001$). Multivariate regression analysis indicated that only age was a significant predictor of hypercholesterolemia (data not shown).

Triglycerides

The mean TG was 96.8±53.2 mg/dl and 5% of individuals on Kosrae were hypertriglyceridemic (Table 3.1). Sex, parity, village, and age all influenced TG levels (Table 3.2). Men had higher mean TG levels than women (110.4±61.2 vs. 87.1±44.0) with a 3.43 fold increased risk of hypertriglyceridemia (p<0.0001). Women with ≥6 children had higher mean TG levels than women with 0-5 (95.2±45.5 vs. 82.5±42.5), but with no significant increased risk of hypertriglyceridemia. Residents of the 4 villages had greater mean triglyceride levels than residents of Welung (97.6±53.5 vs. 74.3±31.9), but the odds ratio for risk of hypertriglyceridemia could not be calculated because no one in Welung had triglycerides greater than 200. TG levels increased with age (81.6±44.0, 105.3±57.5, 104.5±53.2), but did not differ between 35 to 49 and ≥50 year olds. The highest risk of hypertriglyceridemia was in middle age with an odds ratio of 2 compared to the oldest age group and a 3.2 fold over the youngest adult group. Multivariate regression analysis showed that only sex and age were significant predictors of the risk of hypertriglyceridemia (data not shown).

Apolipoprotein B

The mean level of APOB was 87.4±21.6 mg/dl and 7% of Kosraeans had hyperapobetalipoproteinemia (Table 3.1). Sex, parity, smoking, village and age all had effects on APOB levels (Table 3.2). APOB was higher in males than females (90.2±21.6 vs. 85.3±21.3) and males were at increased risk for hyperapobetalipoproteinemia (1.44, p=0.03). Women with ≥6 children had higher APOB levels than those with 0-5 children (90.1±22.8 vs. 82.6±19.9) and had a 3.54 fold increased risk of hyperapobetalipoproteinemia (p<0.0001). Nonsmokers had higher APOB levels than smokers (92.5±21.5 vs. 87.0±21.6) and residents of the 4 villages had lower levels than residents of Welung (87.0±21.6 vs. 98.3±18.8). Neither smoking nor village of residence were associated with a higher risk of hyperapobetalipoproteinemia. APOB levels and the risk of hyperapobetalipoproteinemia increased in a step-wise fashion with age in

Kosraeans. In individuals over the age of 50 (94.3±22.7), there was a 4.6 fold increased risk of hyperapobetalipoproteinemia compared to 20 to 34 year olds (79.7±19.4) and a 2 fold increased risk compared to 35 to 49 year olds (89.4±20.4). In multivariate regression analysis the covariates of sex, parity, and age still had significant effects on the incidence of hyperapobetalipoproteinemia (data not shown).

Apolipoprotein A-I

In Kosraeans, the mean level of APOA-I was 117.2±24.6 mg/dl and hypoAPOA-I was defined as the lower 10% of the population (Table 3.1). Sex and village of residence were significant covariates for APOA-I levels (Table 3.2). Males had lower APOA-I levels than females (108.4±22.0 vs. 123.3±24.5) with a 3.61 fold increased risk of hypoAPOA-I among females (p<0.0001). Residents of the 4 villages had lower APOA-I levels than residents of Welung (116.7±24.5 vs. 132.1±22.7). The risk ratio for hypoAPOA-I could not be calculated as no one in Welung had APOA-I levels below 88. After multivariate regression analysis, sex remained the only significant predictor of risk of hypoAPOA-I (data not shown).

Combined Dyslipidemia

The dyslipidemic phenotype, as defined above, was present in 20% of Kosraeans (Table 3.1). Significant covariate effects were found with sex, village, and age. Males had a higher prevalence of dyslipidemia than females (29% vs. 13%) with a 2.64 fold increased risk ratio (p<0.0001). Residents of the 4 villages had a higher risk of dyslipidemia than residents of Welung (20% vs. 9%) with a risk ratio of 2.70 (p=0.02). The risk of dyslipidemia did not differ significantly between the middle and old age groups, with relatives risk among the oldest and middle age groups of 2.00 fold and 1.75

fold, respectively (both p<0.0001) vs. the youngest group. The effects of sex, village and age remained significant after multivariate regression analysis with borderline significance seen for an effect of parity. However, the age effect was diminished as it was only significant for increased risk in the old age group compared to the young adult group (Table 3.3).

Height

The mean HT was 61.9±3.3 in or about 5 foot 2 inches (Table 3.1). Mean HT was higher in males (64.7±2.5) than females (59.9±2.2) and in the young and middle age groups (62.2±3.2 and 62.5±3.3) than the oldest (60.7±3.4). Mean HT was slightly higher in smokers (65.1±2.3) than non-smokers (64.4±2.6) and in women with 0-5 children (60.2±2.2) than those with ≥6 (59.4±2.2), probably because both those groups (smokers, 0-5 children) were younger (Table 3.2).

Clustering of Syndrome X component disorders

On Kosrae, 0.6% of the population had all 4 diseases associated with Syndrome X, 5.7% had 3, 22.7% had 2, 20.3% had one, and 50.7% had none. In subjects with 2 or more disorders a univariate logistic regression analysis was done to determine the odds ratio of a Kosraean having 2 of the disorders together rather than either one alone (Table 3.4). The odds ratio for having obesity and diabetes was 1.57 fold (p=0.002) and obesity and hypertension 2.62 fold (p<0.0001). However, obesity did not increase the odds ratio for dyslipidemia or any of the individual lipid related phenotypes. The odds ratio for having diabetes and hypertension was 3.16 fold (p<0.0001), diabetes and dyslipidemia

2.3 fold (p<0.0001) (hypercholesterolemia 3.71 fold, p<0.0001, hypertriglyceridemia 2.50 fold, p<0.0001, and hyperapobetalipoproteinemia 2.72 fold, p<0.0001), and hypertension and dyslipidemia 1.90 fold (p<0.0001) (hypercholesterolemia 3.07 fold, p<0.0001, hypertriglyceridemia 2.70 fold, p<0.0001, and hyperapobetalipoproteinemia 2.35 fold, p<0.0001).

Table 3.4:Univariate logistic regression

Compares the frequency of being affected with two disorders as compared to one but not the other.

	Diabetes	**Hypertension**	**Dyslipidemia**
Obesity	1.57*	2.62**	1.11
Diabetes		3.16**	2.30**
Hypertension			1.90**

*p<0.01 **p<0.001

Clustering of Syndrome X quantitative traits: Factor analysis

Factor analysis was used to delineate the clustering of the quantitative traits associated with Syndrome X. A total of 11 quantitative traits were analyzed. Each of the quantitative traits was significantly correlated with between four and nine of the others, showing that they were all interrelated. After principal component analysis and orthogonal rotation the factor patterns were extracted from a correlation matrix (Table 3.5). Four uncorrelated factors that explained 73.1% of the total variance in the dataset were found. Factor 1 (OB/DB) had strong contributions from WT (loading = 0.78), WST (0.81), LEP (0.66), and INS (0.76) with additional significant correlations with FBS (0.34) and TG (0.38). This factor explained 23.6% of the total variance. Factor 2 (DYSL) had strong contributions from TC (0.92), TG (0.49), and APOB (0.90) with additional significant correlation with INS (0.21) and APOA-I (0.28). This factor explained 18.9% of the total variance. Factor 3 (BP) had strong contributions from SBP (-0.93), DBP (-0.88), and FBS (-0.44) with additional significant correlation with WT (-0.22) and WST (-0.34). This factor accounted for 18.5% of the total variance. Factor 4 had strong contributions from APOA-I (0.82), plasma LEP (0.56), and TG (-0.47), as well as significant correlation with WT (–0.27). This factor explained 12.1% of the total variance. These data were not altered when factor analysis was performed for males and females separately (data not shown). Further clustering analyses were done, with the addition of HT, removal of INS (to allow for use of the whole dataset), and on the residuals corrected for the significant covariates. For the first three factors, the only change seen was a significant correlation between the OB/DB factor and HT (0.45). But the fourth factor was very different, suggesting the correlations seen were due to the covariates only. This factor (HT/DYSL) correlated strongly to HT (-0.45) as well (not shown) and was characterized by decreased TG (-0.65) and elevated APOA-1 (0.65). The HT/DYSL factor was used in further analyses, as Factor 4 itself showed no evidence for any major gene effect.

Table 3.5: Rotated Factor Pattern

	OB/DB	DYSL	BP	Factor 4	HT/DYSL
WT	**0.78****	0.11	-0.22*	-0.27*	0.04
WST	**0.81****	0.12	-0.34*	-0.11	-0.08
LEP‡	**0.66****	-0.07	-0.03	**0.56****	-0.12
FBS‡	0.34*	-0.01	**-0.44****	-0.04	-0.13
INS‡	**0.76****	0.21*	0.00	0.05	N/A
SBP	0.03	0.08	**-0.93****	-0.02	-0.05
DBP	0.16	0.16	**-0.88****	-0.08	-0.04
TC	0.10	**0.92****	-0.08	0.13	-0.12

TG‡	0.38*	**0.49****	-0.13	**-0.47****	**-0.65****
APOB	0.11	**0.90****	-0.10	0.09	-0.05
APOA-I	-0.10	0.28*	0.08	**0.82****	**0.65****
% variance	23.6	18.9	18.5	12.1	

Loadings ≥ |0.4| shown in bold

‡log transformed

*p<0.01, **p<0.001

Discussion

To gain insight into the molecular basis of these disorders, an epidemiologic analysis of obesity, diabetes, hypertension, and dyslipidemia on the Pacific Island of Kosrae was undertaken. Toward this end, a clinical database of the quantitative phenotypes associated with these disorders as well as important covariate has been established. These data were used to define disease phenotypes in this population, determine disease frequency, examine the effects of covariates, and assess the clustering of disease phenotypes as well as their underlying quantitative traits. Factor analysis strongly suggested that Syndrome X in this population is not due to a single underlying cause but is the result of interactions among several common ones.

The frequency of obesity and diabetes is higher among the Kosraeans than in the US population. The mean BMI in Kosrae was 31.0, a level that is considerably higher than the value among Caucasians of 24 (Deurenberg et al. 1998). In the US, 22% of the population is classified as obese with a BMI ≥ 30. On Kosrae, since 59% of the population has a BMI ≥ 30, a BMI ≥ 35 was used as the threshold for obesity, corresponding to 24% of Kosraeans. Population specific BMI cutoffs for obesity (BMI ≥ 40) have also been used in other populations with mean BMI higher than in the US, such as in the Pima Indians (mean BMI 34) and Nauruans (mean BMI 35) (Price et al. 1992; Hodge et al. 1993).

The prevalence of diabetes on Kosrae was 12% as compared to 8% in the US using the ADA criteria for diabetes (Harris et al. 1998). In contrast, the prevalence of hypertension and dyslipidemia is lower on Kosrae than in the US. 17% of Kosraeans were hypertensive (standard cutoffs for hypertension of SBP≥140 and/or DBP≥90 were used), a prevalence lower than that in the US (24%) (Burt et al. 1995). 4% of Kosraeans were hypercholesterolemic (cutoff of TC ≥ 240) compared to 20% of US and 5% of Kosraeans were hypertriglyceridemic (TG ≥ 200) compared to greater than 10% of US (Anonymous 1980; 1989). In summary, some, but not others, of the conditions that comprise Syndrome X have a higher prevalence among Kosraeans vs. the US population.

In spite of increased BMI and diabetes, the prevalence of hypertension and dyslipidemia are actually decreased. A similar phenomenon has been seen in Nauruans and Pima Indians (Nelson et al. 1990; Hodge et al. 1993). In the US population, BMI correlates with diabetes, hypertension, and the high triglyceride, low HDL phenotype. Thus Kosraeans appear to be more similar to the Nauruans and Pima Indians than to the US populations. This latter observation is not surprising in light of the geographic proximity and ethnic similarity of Kosrae and Nauru.

A number of covariates influence the prevalence of obesity, diabetes, hypertension, and dyslipidemia. The effect of these covariates on disease phenotypes was examined using univariate and multivariate regression analyses. Univariate regressions were used to generate models for the linkage analysis (see models in Table 5.1). Multivariate results are discussed below. Males were more likely to be diabetic, hypertensive, hypertriglyceridemic, hyperapobetalipoproteinemic, and hypoAPOA-I, while females were at greater risk for obesity. Multiparity ≥6 was associated with increased risk of hyperapobetalipoproteinemia only as the effects of parity on obesity, diabetes, hypertension, and hypercholesterolemia evident in univariate analysis became nonsignificant when corrected for age. Smoking was associated with a decreased risk of obesity and diabetes, consistent with previous studies showing that the effect of smoking on obesity is the basis of the decreased risk of diabetes and hypertension in smokers (Liu et al. 1999; Stroup-Benham et al. 1999). The village of residence was associated with decreased risk of obesity, hypertension, and dyslipidemia. The reduced risk of these disorders among the residents of the village of Welung is likely caused by the maintenance of a more traditional lifestyle there (due to the inaccessibility of this village). The frequency of Syndrome X components has been shown to increase in other populations when lifestyle becomes westernized and to decrease with the resumption of the native lifestyle (Shintani et al. 1991; Price et al. 1993; Ravussin et al. 1994; Shintani et al. 1994). In general there was increased disease risk with increasing age, as seen in previous studies (Wilson 1994; Muller et al. 1996). However, obesity was more frequent in middle age than in the elderly, and there was no effect of age on APOA-I levels. The effects of covariates on disease prevalence in Kosrae generally confirm what has been seen in other populations, suggesting that the epidemiological tools that were used were robust and that conclusions derived from Kosrae are more generally valid.

It was also notable that very few individuals on Kosrae had all four of the component disorders of Syndrome X. Thus only 0.6% of middle aged or older individuals (64% of the population) had all 4 disorders indicating that the presence of all of the features of Syndrome X are rarely evident in this high risk population. However, the finding that 28.4% of adult Kosraeans had 2 or 3 disorders compared to 20% with just one disorder confirmed that there is some clustering of the Syndrome X components on Kosrae. These data suggest that Syndrome X is not one disorder but is rather the coincidence of several common ones. Univariate logistic regression analysis was also done which showed that obesity was associated with increased risk of diabetes and hypertension, but not dyslipidemia, while diabetes was associated with increased risk of hypertension and dyslipidemia, and hypertension was associated with increased risk of dyslipidemia. In general these results agree with other population based epidemiological studies other than the fact that obesity usually clusters with dyslipidemia, particularly high TG and low APOA-I (Reaven 1988; Defronzo and Ferrannini 1991; Ferrannini et al. 1991). In general, despite their extreme obesity, Kosraeans tend to have much lower TG than Caucasians, suggesting that local environmental or genetic factors may protect Kosraeans from the usual effect of obesity on TG (Anonymous 1980).

To assess clustering of Syndrome X component disorders, relationships among quantitative traits related to Syndrome X diseases were studied using multivariate factor analysis. Four independent factors were identified that in aggregate explained 73% of the total variance. The OB/DB factor was characterized by elevated WT, WST, LEP, and INS with some contribution from FBS and TG. This factor is principally related to obesity with a component of diabetes and hypertriglyceridemia. The DYSL factor was characterized by elevated TC, TG, and APOB with some contribution from INS and APOA-I. This factor principally related to hyperlipidemia and to a lesser extent, diabetes. The BP factor was characterized by elevated blood pressure and FBS with some contribution from WT and WST circumference. This factor is primarily related to hypertension with a lesser relationship to diabetes and obesity. The HT/DYSL factor was characterized by elevated APOA-I and decreased TG with a correlation to tallness. This could be another form of dyslipidemia.

This analysis suggests that several components (rather than one) underlie Syndrome X, and that these components cannot simply be described as obesity, diabetes, hypertension and dyslipidemia. Rather the components of Syndrome X are apparently caused by the four factors referred to above. Clustering of obesity, diabetes, hypertension, and dyslipidemia is due to both the underlying factors sharing some of the same quantitative traits and the coincidence of more than one of these common factors.

Data from several other population-based studies have been analyzed using factor analysis with similar results. In the Kaiser Permanente Women Twins Study, four factors were identified: i) body mass/fat distribution, ii) FBS/INS, iii) BP, and iv) HDL-C/TG (Edwards et al. 1994). In the Framingham Offspring Study three factors were isolated: i) BMI/WHR/INS /HDL-C/TG, ii) FBS/INS, and iii) BMI/SBP/DBP (Meigs et al. 1997). In the Strong Heart Study of American Indians, three factors were identified: i) BMI/FBS/INS, ii) BP, and iii) INS/HDL-C/TG (Gray et al. 1998). Lastly, the Honolulu Heart Program of Japanese Americans identified four factors: i) WT/WST/INS, ii) FBS/INS, iii) DBP/SBP, and iv) HDL-C/TG (Edwards et al. 1998). Both the Kaiser Permanente Women Twins Study and the Honolulu Heart Program of Japanese Americans identified a HDL-C/TG factor similar to the APOA-I/TG factor identified here, and this could even be the same as the INS/HDL-C/TG factor from the Strong Heart Study of American Indians, except insulin was not included. However, the blood pressure factor on Kosrae also included components of diabetes and obesity, whereas the Kaiser, Strong Heart, and the Honolulu Studies revealed only a pure hypertension factor. In addition, the factor analysis reported here did not reveal a distinct diabetes (FBS and INS) related factor, whereas this was found in the Kaiser Permanente Women Twins Study, Framingham Offspring Study, Strong Heart Study, and the Honolulu Heart Program Study. This may be due to the use of slightly different quantitative traits in the Kosrae factor analysis compared to the others. Alternatively, these differences may be due to genetic and environmental differences in these disparate populations.

In summary, this population based study on the Island of Kosrae suggests that Syndrome X is a composite of 4 independent factors: obesity with diabetes and hypertriglyceridemia, combined hyperlipidemia with diabetes, hypertension with obesity and diabetes, and increased HDL-low triglycerides with thinness associated with stature. If this is merely a measure of phenotypic correlation or if there are genes underlying these factors will be seen after the genome scan. With an understanding of the phenotypes and covariate effects in this population, the next step is the genetic analysis.

Chapter 4: Initial genetic analyses

Introduction

Preliminary genetic analyses were carried out prior to the complete genome scan analyses. First two types of genetic epidemiology studies were preformed to ensure that there are genetic factors involved that affect the quantitative traits of interest in this population. These included measuring heritability, the amount of variance due to genetic effects, and carrying out SA, which suggests how many QTLs control each trait, and the percent of trait variance attributable to each QTL. Next a screening set of markers, for a genome scan with average 10 cM density and no gaps greater than 12 cM, was developed. The complete pedigree was genotyped for this minimal set of markers. Then the Kosrae data, both the pedigree and the genotypes, were cleaned up as a preparation for marker analysis. Last, initial genome scans were carried out using the SOLAR package.

Heritability

Significant heritability (percent variance of trait attributable to genetic effects) was seen for all the traits (Table 4.1) with corrections for the five covariates (sex, parity, smoking, village, and age) done simultaneous to the calculations. The highest h^2 was seen for HT (0.64), then WT (0.53), BMI (0.45), HIP (0.45), TC (0.41), and WST (0.40). The others were slightly lower, with APOA-I (0.36), APOB (0.31), DBP (0.29), SBP (0.25), and FBS (0.24), and the lowest for TG (0.21) and LEP (0.20). As h^2 measures all the genetic effects, when it appears to be low, it could be due to the effect of one major gene, and then it would actually be pretty high for linkage analysis, or it could be due to the effect of many small genes, then even a high h^2 would not mean there is a greater chance to find genes. How many genes seem to be coding for each trait, and how much variance is explained by these, is seen in the next analysis, SA.

Table 4.1: Heritabilities

BMI	0.45	**DBP**	0.29
WT	0.53	**TC**	0.41
WST	0.40	**TG***	0.21
HIP	0.45	**APOB**	0.31
LEP*	0.20	**APOA-I**	0.36
FBS*	0.24	**HT**	0.64
SBP	0.25		

*log transformed

Segregation Analysis

Segregation analyses for each trait model (correcting for significant covariates for each trait) were carried out using Loki, and results are shown in Table 4.2. This gives a very general idea of the major genes affecting the traits, and suggests which would be able to be mapped (explaining a significant portion of the total variance). The highest number of QTL was seen for HT (5.5), which explain 0.81 of the total variance, with the largest gene explaining 0.28. The other traits showed a significant percent of the variance due to major genes: WT (0.78), WST (0.74), BMI (0.63), FBS (0.54), HIP (0.52), APOA-I (0.51), MAP (0.49), INS (0.43), TG (0.40), APOB (0.38), TC (0.36), LEP (0.23), and SBP (0.16). These traits had between 2 to 3.5 QTLs, except for SBP (1), with about 0.20 to 0.30 of the total variance explained by the largest QTL (WT 0.32, BMI 0.27, WST 0.31, HIP 0.26, LEP 0.20, FBS 0.31, INS 0.29, DBP 0.20, MAP 0.24, TC 0.26, TG 0.24, APOB 0.18, and APOA-I 0.29), again with the exception of SBP (0.14). In general, similar results were seen for the polygenic effect, with the number of QTLs reduced by about 1 (not shown).

Table 4.2: Segregation Analyses

Trait	mean % variance attributable to environmental effects	mean % variance attributable to genetic effects	mean number of QTL	mean % variance attributable to largest QTL
HT	19	81	5.5	28
WT	22	78	3	32
BMI	37	63	3	27
WST	26	74	3.5	31
HIP	48	52	3.5	26
LEP*	77	23	1.5	20
FBS*	46	54	2.5	31
INS*	57	43	2	29
DBP	61	39	3	20
SBP*	84	16	1	14
MAP	51	49	3	24
TC	64	36	2	26
TG*	60	40	2	24
APOB	62	38	3	18
APOA-I	49	51	2.5	29

* log transformed

Genome scan marker set

The genome scan marker set was developed and optimized at the GCF at the Rockefeller University. The Marshfield 9 screening set was modified for use in the GCF lab for the Kosrae population. First, any marker with product greater than 400 bp was removed, as at the onset the 3700 capillary machine was not able to handle these markers.

A number of markers close to candidate genes for the traits discussed were genotyped, so Marshfield markers close to these were removed. There were a few chromosomal areas with larger gaps (greater than 12 cM), so markers in those regions (if available) were added. Then the markers were tested on CEPH DNA (Corriel Cell Repository, New Jersey) to make sure they worked under the standard conditions in the lab. For any marker that failed or had low intensity fluorescence (low product yield), the PCR was redone with additional Taq polymerase. For those that had unamplified second alleles the PCR was redone with glycerol-based reaction buffer. The last optimization done was for markers with an excess of non-specific amplification, which lead to artifact peaks, where the PCR was redone with betaine containing buffer. If these problems were not resolved, the markers were replaced. Additionally, some primers were redesigned to generate different size products to allow for more efficient multiplex gel or capillary loading. Once the markers were all working, they were organized into tentative panels, with the size ranges greatly expanded to allow for the possibility of non-CEPH alleles in the Kosrae population. Then, after amplification in a portion of the Kosrae samples, a number of markers either failed completely, were unscoreable, or had high error rates (described later). These were replaced as well. Finally, the forty panels were set to allow for the allele ranges observed in the Kosrae population. Panels used in this final stage are available on the GCF website as mentioned above. Scoring was described in Chapter 2, and the data cleanup is described below.

Kosrae Pedigree

The pedigree was constructed by Maude Blundell from the questionnaires, family visits, and genealogical records. This pedigree goes back five generations and includes a total of 2286 individuals with DNA available for 1564 of these. The average sibship size on Kosrae was four with some families having as many as twelve children. The inbreeding coefficient for individuals who have been genotyped was zero with the exception of one individual who had an inbreeding coefficient of 0.00781.

At the beginning of the project, the complete pedigree was not known, so 1102 individuals were genotyped (the first set). For some of the early markers, an additional 120 individuals were genotyped. 462 more individuals were subsequently placed on the pedigree and genotyped (the second set). About a quarter of these were siblings in nuclear families already on the pedigree, the remaining were new families that connected. A majority of those were women whose parents had not been known previously. Most of the analyses discussed were carried out on the first subset, and the final results were confirmed on the complete pedigree available (see below). Prior to analysis, the pedigree was prepared by Simon Heath, by removing relationships that appeared to be improbable and breaking the loops.

After checking the markers for Mendelian inconsistencies (see below), a set of nuclear families was identified that had errors for many markers, suggesting that these were pedigree problems and not genotyping errors. There was usually one person who was problematic, so that individual was removed and kept as a singleton, with a "dummy" inserted in its place in the family if it was needed (such as a parent). Though singletons give no linkage information, these were kept in the analysis for the population-wide estimates, such as allele frequencies and phenotype variances. On the first set, about 30 families were modified, and this pedigree was further modified (about 25 families) after addition of the second set, to maintain consistency for signal comparison. Ideally, the pedigree would be modified only once, using all available information.

Though Loki can handle large complex pedigrees, it needs to be single locus peelable and is therefore limited to a small number of interlocking pedigree loops. Loops also slow down the analysis significantly, and do not add that much information to justify inclusion. This pedigree contained greater than 110 loops, mostly marriage loops, which were all broken by duplicating selected individuals. When possible, these were individuals for whom no data was available.

Microsatellite Marker Genotypes

Next, the genotype data for each marker were cleaned up to remove all detected Mendelian inheritance problems. This was done automatically in Loki, which goes through the pedigree, removing the nuclear family for each identified individual in error. It then goes back to each family and puts back non-problematic individuals one at a time. Though there is an element of randomness, as it removes families based on the first error it sees and then adds individuals back randomly, it is an acceptable trade-off to checking thousands of genotypes manually. This information was first used to detect families that appeared to be problems with the pedigree as described above, then the errors were reidentified on the modified pedigree.

Markers with observed error rates of greater than 1.9% were rescored, as this suggests a systematic allele-calling error, rather than random errors. A majority of the time this was the case, the alleles were rescored correctly and a reduction in the observed error rate was seen. A few markers were regenotyped completely, as there was a systematic error that could not be corrected on the available data, but was able to be scored correctly using the new data. Finally, most markers with unresolved errors were not used in the analysis as there was probably a non-obvious systematic error that could not be corrected. The same steps were carried out with the second set, which resulted in the removal of three more markers from the analysis, that gave high error rates (close to 5%) for no obvious reason. Genotypes that were still detected as errors were zeroed out prior to analysis, so that individual was considered to be unknown for that marker. There may still be many undetected errors, those that are Mendelianly consistent.

After data cleanup, the marker statistics for the two sets are as follows: for the first set, the average number of genotypes per marker was 1043, with a range of 835 to 1179, average error rate 0.43%, range 0-3.9%, and average heterozygosity 0.67, range 0.12 to 0.9; with the addition of the second set, the average number of genotypes was 1453, range 951 to 1604 (12 markers were only typed on the first set), average error

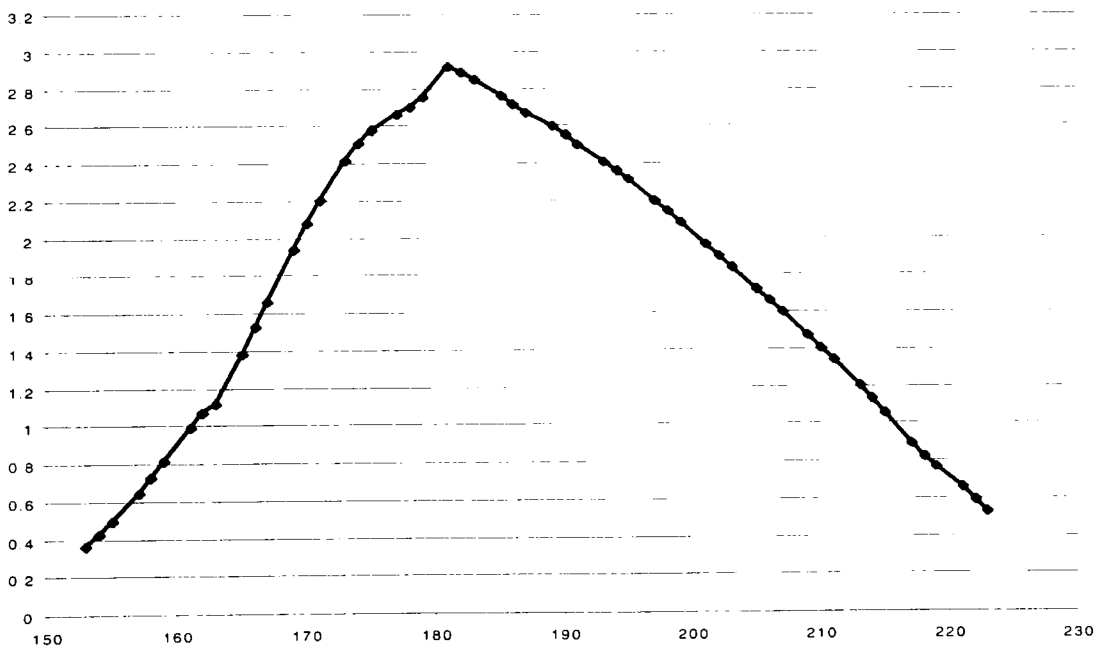

Figure 4.1: BMI on chromosome 1

cM on the X-axis, LOD score on the Y-axis

0.49%, range 0 to 2.2%, and heterozygosity was the same.

SOLAR analysis

The SOLAR analysis on the nuclear pedigrees yielded no significant results. The analyses on the larger pedigrees yielded one highly significant LOD score, ~2.9, for BMI on chromosome 1 (Figure 4.1). There were a few suggestive scores (LOD ~1 to ~2) as well: BMI and DBP on ch 17, TC on ch 7, and TG on ch 1.

Discussion

Before mapping genes, it is important to make sure that there are genetic factors influencing these traits in this pedigree. The quantitative traits associated in Syndrome X have significant heritability, ranging from 0.20-0.64 in Kosrae, suggesting that important alleles are likely to be segregating in this population. Significant evidence for multiple quantitative trait loci (QTLs) was found for all traits, with the variance attributable to the gene with the largest effect size at least 20% (except SBP) or more of the total variance, the theoretical limit for localization using LA. This indicates that there is a high probability that genes strongly associated with the traits being investigated are segregating in the population. As the actual mapping is done on the residual variance, after correcting for the covariates, these QTL actually explain a larger percentage of the "total" (residual) variance during the analysis, so there is strong evidence for are major genes for these traits that LA should be able to identify. Furthermore, different models with a larger number of QTLs or increased variance explained by the QTL were analyzed, as described later. These genetic epidemiology measures are similar to other studies, as discussed in the introduction. Power studies indicated that the Kosrae dataset has high power to detect linkage to genes with this magnitude of effect using Loki (data not shown). Thus there was confidence that it would be possible to localize using LA several genes that predispose individuals to the diseases associated with Syndrome X by affecting the underlying quantitative traits.

For the marker analysis, a genome scan set of markers, which included about 380 markers that fit into 40 panels and were easy to score, was developed. Up to 1684 of the individuals on the Kosrae pedigree were typed for markers from this set, and the pedigree and genotyping errors were removed, yielding a powerful dataset for further analysis. The errors detected were similar in type and magnitude to those found in other similar studies. The presence of errors has been shown to affect the power to map and accuracy to localize QTL though the effect of errors in Loki analysis is not known (Sobel et al. 2002). One suggestion to detect more errors is to identify at obligate recombinants, both single on a small scale, and double on a larger scale, and recheck those genotypes. This is currently under investigation, how to identify and report all types of recombinants using Loki. Another idea is to model errors into the analysis, but this is quite complicated to do in the current implementation of Loki and still get efficient mixing (Heath 2002).

The initial analysis with SOLAR with the nuclear families had no signal, and the larger pedigrees gave one highly significant signal, which suggests that there are genes that could be found, but power is reduced by breaking up the pedigree. Therefore, the genome scans on the entire dataset were carried out using Loki, as described in the next chapter.

Chapter 5: Strategy for Genome Scan Using Loki

Introduction

The purpose of this study was to learn how to carry out a complete genome scan and understand the results using Loki. The questions asked were how many markers to analyze at once, for how many iterations, what models to use (polygenic or not), effects of adding whole pedigree, and then how to be sure which positive signals are real, by multi-chromosome analysis and repeat runs. These were answered partially in a simulated dataset, and then that information was applied to the Kosrae dataset to see if the same strategy works well in real data. Further tests with the real data were done, such as more stringent model testing with many different corrections (polygenic effect, major gene, and quantitative traits), mixing, and a number of other analysis parameters. This resulted in a working strategy for a genome scan, as well as reliable signals for many of the traits associated in Syndrome X (discussed in the next chapter).

Genetic Analysis Workshop 12

The simulated dataset is from the GAW12, which was arranged by the Southwest Foundation for Biomedical Research to develop and test methods of genetic data analysis (Almasy et al. 2001). This set models a complex disease, with information on five quantitative risk factors, two environmental covariates, age, sex, and affection status. Each of fifty replicates includes 1497 people contained in 23 pedigrees, of whom 1000 have phenotype information, and 497 are "founders" (unsampled). This study is described in detail in Shmulewitz and Heath and discussed below (Shmulewitz and Heath 2001).

The initial analysis was performed on replicate 1 without access to the answers. To simulate a 'real' genome scan, markers were selected to give an inter-marker spacing of at least 10cM and marker data from unsampled individuals were deleted, leaving 1000 typed individuals. The MCMC LA package Loki was used to perform genome scans for two models: Quantitative trait (Q)1 = sex + age and Q2 = ln (environmental factor 1) + sex + age, these models being selected based on the regression analysis. Marker allele frequencies and covariate effects were estimated jointly with the QTL parameters. The linkage analyses were performed on one or two chromosomes, using all selected markers on each chromosome jointly. For each analysis, 101,000 sampling iterations were run, with the first 1000 iterations being discarded. An L-score ≥ 20 (genome-wide $p \leq 0.05$) was the natural choice for the cutoff for linkage, as the vast majority of the scores were around 0-5, with a few above 20.

These initial 10cM genome scans on replicate 1, which were done on only sampled individuals and without the polygenic effect, resulted in four scores above 20. The focus of the secondary analysis was to use the answers to assess the effects of using all the marker data, fitting a polygenic effect, and analyzing multiple chromosomes simultaneously, on the power and specificity of the analysis. A series of analyses with 200,000 iterations were performed on replicate 1 and on the 'best replicate', replicate 42. The initial complete genome scans for both models were performed using a polygenic effect and the complete pedigrees. This resulted in four signals for Q1 and four for Q2. The signals were essentially the same with the polygenic effect, but addition of the polygenic effect consistently shifted the distribution of the number of putative QTL down by one (data not shown). Genotype information for the complete pedigree increased the scores, though it did not find any different QTL. Then the effect of analyzing chromosomes jointly was looked at. For Q1, analyzing the chromosomes with real loci together raised the score for one but lowered it for the other, and the polygenic models had higher signals for both. When all the positive chromosomes were run together, with the polygenic effect, one false positive disappeared, another false positive was greatly reduced, one real locus was slightly lower, and the second real locus was increased. For Q2, analyzing the real chromosomes together, with the polygenic effect, resulted in higher scores for two of the loci, but no signal for the third locus was found. Analyzing the positive chromosomes together removed the two false positives, and raised the scores at the two real loci (discussed below).

In general, QTL allele frequencies and variance due to each locus were correctly estimated for Q1. Estimates were more similar to the generating models as more information was included in the analysis, such as the whole pedigree and polygenic effect (data not shown). This could not be estimated for Q2, as the correct (interactive) model could not be specified.

The simple single-chromosome scans which were initially performed, without knowledge of the answers, detected and correctly localized 2 of the 3 trait loci affecting Q1 and Q2. There are 3 major genes affecting Q1 and Q2, Major Gene (MG) 1 affects Q1, MG3 affects Q2 and MG2 affects both. All the analyses performed on replicate 1 detected MG1 and MG3, but failed to detect MG2 for either trait. Conversely, all analyses on replicate 42 found MG1 and MG2 (for both traits), but not MG3. It therefore appears that there are differences between replicates that affect which trait loci can be found. However, adding more information, such as marker data on the unsampled individuals, fitting a polygenic effect, or analyzing multiple chromosomes simultaneously tended to clean up the signals, increase L-scores, and remove false positives, but did not detect any signals missed using the initial analysis. The failure to detect new loci may be due to the model used in Loki, which does not allow for interactions between trait loci. In this context, the trait loci that appeared to be harder to find, MG2 and MG3, did interact. Further suggestive evidence that the interactions cause problems for the analysis is evident with MG2, where the L-score peak is clearly bi-modal. The 'dip' between the 2 peaks is on top of the marker in the scan that is closest to the true location of MG2, indicating that a deficiency in the model is preventing accurate mapping (data not shown).

Results indicate that, at least with this dataset, a strategy of performing single chromosome scans, under realistic conditions (10cM marker map, no marker data on unsampled individuals and no knowledge of the answers), followed by a joint analysis of all chromosomes showing indications of linkage, works well at detecting the true signals. Adding additional information, with genotypes for the unsampled founders and polygenic effect, tends to increase L-scores at previously identified real loci, and therefore may be helpful in prioritizing scores.

To test whether the differences in L-scores between runs were real, an effect of Monte Carlo sampling error, or because the sampler was not run for long enough, one analysis was run six times with different random number seeds as starting points. For this test Q1 on replicate 42 was analyzed with marker data on all individuals and the polygenic effect, fitting all four chromosomes that showed evidence of linkage from the single chromosome runs. Each run produced 500,000 iterations, as opposed to the 200,000 iterations used for the other joint analyses.

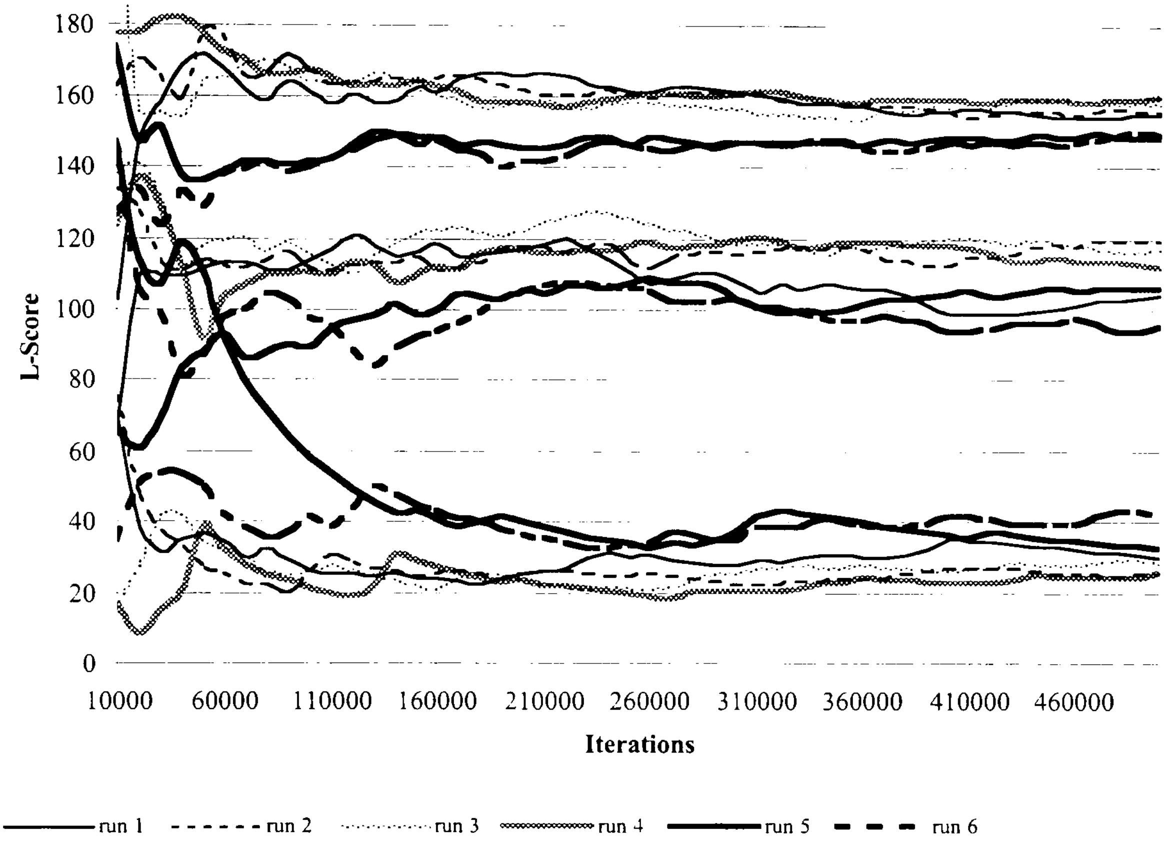

Figure 5.1: convergence of 6 runs for Q1 = sex + age. Signals for ch 2/MG2 are in the region of the graph between L-scores of 140 and 180, for ch 19/MG1 between 80 and 130, and for ch 4 between 20 and 60.

Results from the convergence analyses are shown in Figure 5.1. Most of the signals stabilized by 200,000 iterations, though some did so around 150,000, and others continued fluctuating. One of the false positives disappeared, and is not shown in the figure. The second false positive shows consistently lower scores then the two real loci. This convergence testing indicates that the differences in L-scores between the chromosomes are reproducible and therefore appear to be real. It appears that 200,000 iterations are sufficient to distinguish the four different chromosomes. However, close examination of Figure 5.1 shows that there are differences between runs that are still present after 500,000 iterations. More precise estimates of L-Scores could be obtained by running the sampler for longer, though the estimates obtained after 200,000 iterations are likely to be sufficient to determine which chromosomes merit further investigation. As this is a sampling method, and there are differences between analysis runs, repeating the same analysis a few times to ensure that the results are real is always necessary.

Kosrae analysis

This is the general strategy applied to the real data in the Kosrae study: to do quick single chromosome scans, then to refine the results with the polygenic effect, joint analyses of all positive chromosomes, and repeat runs. Therefore, the linkage analyses were done in a hierarchical manner, first with complete genome scans (with the original models, polygenic effects, and other model changes) and then with selected traits and multiple chromosome analyses, multiple times. Furthermore, the L-graphs and L-peak localizations were looked at carefully as another way to judge signal reliability. The purpose of these multiple tests is to see what is most likely to be real, to be able to prioritize the signals for further investigation.

The Kosrae dataset includes phenotype and covariate information on greater than 2200 individuals, around 1700 of whom are part of one extended pedigree. Up to1684 of these individuals were genotyped for markers to give at most a 10 cM map, as described in Chapters 2 and 4. Prior to linkage analysis, both the pedigree and genotypes were "cleaned-up" to be as error-free as possible. The phenotype models, including corrections for significant covariates, were obtained from the epidemiology study discussed in chapter 3. Real data is harder to analyze, as there can be a lot of missing information and undetected errors in all aspects of the study, such as genotypes, phenotypes, and familial relationships.

For the autosomal genome scan, a set of markers to give a 10 cM map based on the Marshfield sex-averaged map was chosen (Broman et al. 1998). As discussed above, due to the characteristics of the locus sampler implemented in Loki, markers that are too close together may have trouble mixing properly. In areas where more than one marker was available, the one with the highest heterozygosity and/or number of samples typed was selected. For the first set, 360 of the typed markers were used, with an average density of 10.3 and a range of 5 to 15 cM. For the second set, mainly the markers used in the analyses were typed, with a few replacements and additions, so there were 365 markers in total. 353 of these were typed on all the pedigree, and 12 only on the first set. This yields an average density of 10.1, but as three markers were removed due to error rates, the largest gap was 19 cM. This minimal screening set should allow most of the QTLs to be found in a relatively quick and efficient manner though simulation studies have suggested that at times a10 cM map is too sparse to localize all genes (Almasy and Blangero 1998).

Initial Loki analysis

The following quantitative traits were analyzed, with models correcting for the significant covariates as calculated by regression analysis described above: BMI, WT, WST, HIP, LEP, FBS, INS, SBP, DBP, MAP, TC, TG, APOA-I, APOB, and HT. Some traits were modeled with corrections for other quantitative traits that were correlated physiologically, but not due to shared genes. The purpose of doing any correction is to restrict the variance of the trait to be solely due to the major genes as much as possible. Further corrections are discussed below, and all models are listed in Table 5.1.

Table 5.1: Trait models for genome scans

Trait	Covariates	models
LEP	1,3,4,5	p, wt, ht, map, map_wt, ht_map, ht_wt, ht_map_wt, all
BMI	1,2,3,4,5	cov, p, all
HIP	1,2,3,4,5	cov, p, all
WST	1,3,4	cov, p, all
WT	1,2,3,4,5	cov, p, all, ht
FBS	1,2,3,5	cov, wt, all
INS	3	cov, redo
SBP	1,2,3,4,5	cov, p, all
MAP	1,2,3,4,5	wst, p, lep, wst_lep, all
DBP	2, 4,5	wst, p, lep, wst_lep, all
APOA-I	1,4,5	cov, p, all
APOB	1,2,5	cov, redo, all
TG	1,2,4	cov, p ,all
TC	1,2,4,5	cov, p, all

"cov" is covariate corrections only; "redo" is the cov scan repeated; "p" is also the polygenic correction; "all" is also all the other quantitative traits, except those that are components of each other; other corrections abbreviated as usual. Covariate codes: 1 = sex; 2 = parity; 3 = smoking; 4 = village of residence; 5 = age. For example, for leptin, the first model listed is the sex + smoking + village + age + polygenic

Additionally, to look for genes that may be involved in the association of these traits (Syndrome X), the factors described above were analyzed, both with corrections simultaneous to the analysis, or on the factors derived from the residuals. As these factors were obtained by maximizing the variance due to phenotypic correlation, they may only map to genes that effect the traits that correlate the most to the factor, as opposed to coding for some combinatorial syndrome. To investigate if there are genes underlying these factors, it is probably best to cluster based on some genotypic association, such as those that use genetic covariances or condition phenotypes on shared genotype effects before doing the factor analysis (Comuzzie et al. 1997). Therefore, only LA results where the factors overlapped with component traits were reported.

The initial exploratory genome scans were done in a manner shown from the GAW12 analyses to be relatively quick and efficient, that is to analyze one or two chromosomes, grouped together to include about 30 markers, per analysis run. Each was run for 110,000 iterations, the first 10,000 discarded as "burn-in". This way, a complete scan for a trait on one Pentium III machine was done in about a week.

L-scores show how much more likely a region is to contain a QTL than by chance, as described above. In the GAW12 analysis, a cutoff of 20 was used to indicate a positive linkage signal. Here, a score of around 10 was considered to be suggestive for linkage. These data are much less complete than GAW12, and therefore the scores are expected to be both lower and less stable (so that a score of around 10, when the same run is redone, could come up as 20, and vice versa). As these signals were analyzed more stringently than those for GAW12, the assumption was that the false positives would eventually disappear, leaving more leeway to include borderline scores. The cutoff was not as strict as for GAW12 so if there were other reasons to "believe" a signal, for example if a few correlated traits mapped to the same region, a slightly lower score was accepted. A few traits had elevated background, scores for many chromosomes above 3 or 4, even close to 10, and for those a higher score was demanded.

Model corrections: polygenic, quantitative traits, major gene

After the initial scan for each trait, additional scans for each trait with model changes were performed. Complete scans were done to see if the same signals would come up (if the model changes affect the L-score, peak shape, or size), and if there were new ones, that these changes had a real effect on the ability to detect trait loci. These included including the polygenic effect, correcting each trait for other quantitative traits, and modeling a major gene effect.

As shown in a simulation study, it is often difficult to differentiate between a major gene with an additive gene effect and a polygenic effect (Snow and Wijsman 1998). Therefore, it is hard to model the polygenic effect well in Loki, as there is a strong negative correlation between variance due to a QTL and residual additive variance (possibly due to polygenes). This can be especially bad when the QTL gene frequency is not extreme, and, therefore, the inheritance model for the QTL looks more additive than dominant. For many of these traits, there is an effect due to polygenes expected, so while it is important to do these analyses, the results have to be interpreted appropriately. If the polygenic effect suggests no QTL or causes a signal to disappear, this could be true (that there are no QTL or that one is not real) or it could be due to the negative correlation between the variances. But as correcting for polygenes removes variance not explained by the major genes, making it easier to model the major genes, real signals analyzed under this model may be stronger and more reliable. Before doing the polygenic scans, SA was run for each trait to ensure that no genome scans were done for a trait with no evidence of a major gene. For those that showed no QTL 30% of the iterations or more of the genome scan without the polygenic effect were redone (see Table 5.1).

The next set of genome scans were models with different phenotype corrections. Correcting traits for correlated quantitative phenotypes has been shown to increase linkage scores and localization power, presumably as it removes background noise due to phenotypic (rather than genotypic) associations, leaving the residual variance more likely to be due to major genes (Arya et al. 2001). A genome scan of each trait corrected for almost all the others, leaving out ones that were part of the trait, was done. This way only that trait, with no interactions with other traits, was analyzed. To ensure that the signals obtained this way were not due to the correlation in the covariates, for a selection of trait / chromosome scores the trait was corrected prior to the analysis, and the residuals were analyzed. These results also have to be looked at carefully, as if a signal disappears, it could be that it is real; but there are gene-covariate and covariate-covariate interactions that are lost with pre-correction.

Additionally, it is possible that the signal obtained for one trait is really due to its correlation to another trait that maps to the same locus. For these traits, corrections were done both ways and a complete scan was carried out. Again, correcting for these correlated traits cleans up the data and can increase the scores for real loci. If both traits map to the same locus again, that suggests that both are real, but if one or the other disappears, that can be because it was due to the correlation only. As mentioned before, Syndrome X may have one gene controlling multiple traits, so all possible combinations need to be investigated. For most traits, one or two additional models based on this were done, with the exception of LEP, where eight different models were analyzed to investigate this fully.

The last correction tested was for a known major gene effect. The apolipoprotein E gene (APOE) on ch 19 is known to effect TC levels, so when a linkage for TC on ch 19 was obtained, it had to be shown that the signal was not due to APOE. Even if it is not due to APOE, by correcting out the variance attributable to that locus, again a larger portion of the remaining variance should be due to QTLs, leading to easier localization and increased scores. Loki models the major gene effect as any other covariate, that is by taking the average TC in each genotype class and standardizing TC levels based on the differences between the classes.

Joint analysis of positive signals

As suggested from the GAW12 analysis, with a sense of what could be real, all the positive chromosomes for each trait model were jointly analyzed with or without the polygenic effect. As the sample space is greatly increased with more markers, these analyses were run for longer, usually 100,000 iterations per chromosome, plus 10,000 for burn-in. There are a few possible results for this type of analysis. Loki, as a sampling method, chooses the best (most likely) configuration. Often, if there is a bit of linkage information on a chromosome, the signal is put there, as opposed to on the unlinked section. But when there are multiple chromosomes, and a limited number of QTL, it actually chooses the best chromosome(s), which results in some scores increasing and others decreasing or disappearing. This makes it easier to say that the scores that remain are more likely to be real, but if the signals are decreased, it could be because they are not real, or that in comparison to the others, they are less likely. This could be for various reasons, such as explaining less of the variance or because the model used is less close to the real one for that QTL. This choosing one QTL or the other could be more of a problem with the polygenic effect which often reduces the number of QTL in the inheritance model, which is one reason the number of QTL (k) was set to 5, as described below.

In an analysis that has more than one QTL in the model, there is no reason two signals can not be found at the same time; so it is troublesome if at each iteration only one comes up as opposed to both together. One concern is that the sampler is not mixing properly and is getting stuck on a local maximum. Mixing issues will be further discussed below, as this was seen in one of the multiple chromosome analysis for LEP. Another possibility is that there is heterogeneity, with subsets of the pedigree segregating different QTL, but when they are analyzed together, one overpowers the other or it switches between one or the other, resulting in a reduction of both real signals.

Another possibility with multi-chromosome runs is that running them together is akin to correcting for the major gene effect of each QTL, therefore reducing the variance that needs to be explained by the others, which makes it easier for the separate QTL to be found. Even if one of the chromosomes has a slight effect and the QTL disappears, including it in the model could clean up the variance enough to help the others to be localized. This could result in increased scores for all the real signals.

For each trait, one or two multiple chromosome runs were done, depending on how different the results obtained from the different models were. The exception again was LEP, where multiple chromosome runs for all models were carried out, with or without polygenic effect, and with or without ch 3. As mentioned above, sometimes chromosomes serve to correct the data, so though in the multiple runs ch 3 had no signal, its inclusion seemed to effect other scores.

As Loki is a sampling method, there is always the possibility that the sampler was not run long enough, or that it got stuck at a local maximum and did not mix properly. Therefore, a number of runs were replicated multiple times. There were two types of signals this was done for: one, those that were not consistent to see if this would clarify the signals (MAP_lep_wst on 9 and 12; LEP_map_wt on 9), and two, high signals (TC_all on 19, with and without APOE; HT_all on 10), to check the stability of the L-score and of the parameter estimates. Stability of the parameter estimates across different models for the same trait was also checked. To further check the mixing graphs of the runs were visually inspected to make sure that the signal was moving on and off the chromosome, which suggests adequate mixing. With the runs that did not mix well, a hybrid sampler that used the meiosis sampler 25% of the time, which runs much slower but mixes better was used. After that instance, for all runs, the meiosis sampler was used 5% of the iterations.

Complete pedigree analysis

The final test for acceptability was the addition of the second set of data. Knowing more of the genotypes instead of sampling from the possible configurations should give more reliable L-scores and better localization. All positive signals from the genome scans described above were reanalyzed, the same way they were done the first time. Then for each trait model, all signals that remained positive were analyzed jointly. Signals that disappeared could be because they were based on possible configurations that were impossible with the new data. It is possible that additional signals not seen at all in the original scans would appear with complete scans for the whole dataset, but as the first set had very high power to at least detect QTL and even borderline loci were reanalyzed, it would probably be a very small number of QTL.

L-graphs and parameter estimates

With a set of signals that are consistent and, therefore, seem to be real, the next thing to do is to inspect the L-score peaks, for height and shape, and to calculate the parameter estimates. L-scores are hard to use as a test of significance for a few reasons. Nominally, an L-score means that it is that many times more likely that there is a linked gene at that location than not, corrected for the whole genome and number of QTL in the model. Therefore, a conservative estimate of the genome-wide p-value is just 1 divided by the L-score, though it is unclear how to correct for multiple models and multiple genome scans. Additionally, the empirical distribution for this test statistic under the null hypothesis of no linkage has not yet been determined, therefore, there is no way of really knowing how often a score of 10, 20, or 100 is expected. Another problem is that the L-score is not stable across repeat runs, so it seems wrong to say that the same locus has a different significance depending on the run. The L-score height is also affected by covariate and genetic corrections, so a real locus could have a score ranging from 10 to 100 depending on the model or run. This is one reason it is important to do many tests and then to evaluate what seems to be happening.

Next it is necessary to look at peak localizations. Peaks that localize off the end of a chromosome are suspect, as often that is due to inability to localize the peak to the correct region of the chromosome. This could be due to many reasons, such as missing information, genotyping error, or model misspecification. Peaks that are split on a marker also suggest a problem either with the model or with the genotyping. In the GAW12 analysis, a split peak for a trait was seen where the gene-gene interaction could not be modeled, and with GAW9, a split peak was seen when there were three alleles at the QTL instead of two as modeled in Loki (Heath et al. 1997; Shmulewitz and Heath 2001). In general problems of localization can be due to mispecifying the model, but this may not be correctable, depending on how much information is available. If there are genotyping errors that lead to excess single or double recombinants, this could lead to mapping problems. Reporting recombinants (to check the genotypes) from Loki runs is currently under investigation. Last, with reliable peaks, sometimes one model gives a tighter (better localized) peak, and that would be the best model to use in further analysis.

After these analyses and inspections were carried out, a reliable set of QTL that explain a percentage of the variance in a majority of the traits of interest was obtained. Loki provides estimates of QTL parameters, such as allele frequencies at the QTL, dominant and additive effects, QTL effect size, % genetic variance, and % total variance attributable to the QTL. As these estimates were seen to be robust to analysis fluctuations (same over runs even when L-score changed and basically the same with minor model changes), they appear reliable; effect size and % variance for each QTL found are reported. As the alleles (normal or disease) are assigned after QTL localization, they can switch at each iteration, as well as the dominance / additive effect, so those estimates are not particularly informative. Therefore, a general gene effect, the square root of the weighted average of the dominance and additive variances weighted by the allele frequencies, is calculated.

Kosambi map function, sex-specific maps, and setting k to 5

Additionally, there were particular questions we wanted to ask, about the effect of other analysis parameters, that were tested on selected trait / chromosome combinations. These included using the Kosambi map setting (LEP_ht_map_wt on ch 5, LEP_ht on ch 6, LEP_all on ch 7, HT_lep on ch 15, APOA-I on ch 16, and HT_all on ch 10), using sex-specific maps (same as Kosambi except HT_all on 10), and changing k to 5 (four multiple runs for LEP, see Table 6.1). Loki actually expects to have map positions in Haldane cM, but all analyses were done using Kosambi cM. This was primarily because at the onset, there was no way in Loki to specify which map was being used. This should not really affect the analysis, as the relative distances between two markers should be close to the same, but the correct map function might affect localization. Therefore a number of analyses using the Kosambi map as the input map and Haldane as the output were done. As expected, this made little or no difference, so all the analyses that were done with the Kosambi map function are reliable.

The second map-related change tested was the use of sex-specific maps. As LA is dependent on the number of recombinants observed as compared to the number expected based on genetic distance, map misspecification, such as the unrealistic distances used in the sex-averaged map, may affect ability to localize QTL. It was shown in a simulation study that this loss of power is quite small, except when there is a big difference in map distance over a large distance, at least for correctly specified single-gene traits (Daw et al. 2000). But this could be worse with complex traits; therefore, it may be important to reanalyze positive scores using sex-specific maps. The problem is that the sex-specific maps publicly available were built from only 94 meioses and therefore have large confidence intervals, which reduces their power to localize genes. Ideally, population specific maps could be built from the dataset, but as long as there are undetected genotype errors this is very difficult. Additionally, there are regions on the male map that have very little recombination, which may cause a mixing problem. Loki calculates recombination fractions separately in the paternally and maternally segregating chromosomes when sex-specific maps are used. The same runs analyzed with the Kosambi map were done in this way.

One of the preset parameters is the mean number of QTL in the model, called k. Previously, Heath showed that setting k for values 1 to 10 had no effect on the analysis, so initially all the analyses were done with k set to 1 (Heath 1997). But this may have an effect when there are multiple chromosomes and therefore multiple QTL in one analysis run, as described above. To test this, some analyses were run with k set to 5, and this was continued from that point on.

The hierarchical analysis described above provides a list of proscribed steps to refine the results at each stage, to obtain a prioritized list of QTLs that affect variance of the traits encompassed by Syndrome X. In the next chapter, the results will be reported, as what QTL were found for each trait, and how each is reliable based on the stages it passed. In Chapter 7 the suggestions for doing a genome scan using Loki based on this will be discussed as well as reliability of the results, and how they compare to other similar studies.

Chapter 6: Results from genome scans using Loki

Introduction

The Loki analysis on the 14 quantitative traits associated with Syndrome X and stature was carried out as described, first by doing a series of complete genome scans to get an idea as to what might be happening for each trait and to identify the most informative model for each trait. Then these results were tested more rigorously by doing multiple chromosome runs and adding the second set of genotype data. Inspection of the L-graphs was used to further evaluate reliability. All this information is used together to decide what signals are real and to prioritize the list for further investigation. How this serves to crystallize what appears to be the most reliable score through the results of the various LEP runs is shown, and then the significant scores for each of the other traits is reported with some detail about model changes as well. Last, chromosome regions that are interesting because they show a number of correlated traits suggestively linked were reported as well.

Leptin analysis

To get a complete picture of how model changes affect analysis, nine complete scans for LEP were done. Results for each model are summarized in Table 6.1.

Table 6.1: Leptin genome scans and multiple chromosome analyses

Models	Ch 3	Ch 5	Ch 6	Ch 7	Ch 9
p	NS	6.5	NS	NS	NS
wt	NS	**25.7**	NS	9.1	**24**
wt_p_m	NS	**60.1**	NS	NS	**96.3**

ht	**18.9**	7.9	**37.4**	**20.3**	4.7
ht_m	NS	7.1	**20.1**	**21.2**	4.7
ht_p_m	NS	6.2	NS	NS	NS
map	NS	14.7	**41.1**	**72.1**	NS
map_m	NS	10.8	**28.2**	**92.3**	4.5
map_p_m	NS	17.3	**52.6**	**36.5**	NS
map_p_m	ND	6.8	10.0	6.7	NS
ht_map	NS	**28.8**	**58.6**	**40.7**	7.4
ht_map_m	NS	**54.8**	**25.7**	**61.0**	6.7
ht_map_p_m	NS	**49.0**	**29.3**	NS	11.2
ht_map_p_m	ND	11.0	NS	9.5	NS
ht_wt	NS	**90.0**	**38.6**	8.9	NS
ht_wt_m	15.1	NS	NS	NS	**30.3**
ht_wt_p_m	NS	**61.9**	NS	NS	**30.6**
map_wt	NS	**73.5**	6.3	**30.8**	**39.0**
map_wt_m	NS	**57.6**	NS	**39.6**	13.6
map_wt_m_k=5	NS	**31.6**	NS	11.1	NS
map_wt_p_m	NS	NS	**123.0**	NS	**20.3**
map_wt_p_m_k=5	NS	**72.7**	NS	NS	**18.8**
map_wt_p_m	ND	**67.9**	**44.6**	NS	**81.9**
map_wt_p_m_k=5	ND	**40.3**	**48.6**	NS	8.9
ht_map_wt	NS	**90.3**	**31.1**	**28.2**	**29.1**
ht_map_wt_m	NS	**66.7**	13.4	**44.0**	**26.6**
ht_map_wt_p_m	NS	**48.4**	NS	NS	13.6
ht_map_wt_p_m	ND	**49.9**	NS	NS	NS
all	12.2	**28.4**	5.8	**60.3**	NS

all_m	4.5	**26.6**	4.0	**40.6**	NS
all_m_k=5	ND	9.0	NS	14.9	NS
all_p_m	NS	**33.1**	NS	NS	NS
all_p_m	ND	**45.6**	NS	NS	NS
all_p_m_k=5	ND	**56.2**	NS	NS	NS

models as listed in table 4.1; "p" = polygenic effect; "m"= multiple chromosome analysis; "k=5" is setting mean number of loci to 5. "NS" = no signal, "ND" = not done. Scores in bold are 20 or above, $p \leq 0.05$

It is possible that individual models give unique scores that are real for that particular configuration, but first loci that come up more or less for all models are more interesting. Here, the polygenic model had no corrections for other traits and no real scores. Of the other 8 models, 2 gave suggestive scores (L between 10 and 20) on ch 3, 6 for ch 5 with the seventh having a suggestive signal, 5 gave signal on ch 6, 6 gave signal on ch 7, and 3 gave signal on ch 9, but there were 2 or 3 loci on 9, and the signal was not well localized. Next one model (map_wt) for ch 9 was done ten times to see if that would lead to some consistency, but it did not. Two runs gave no signal, five gave signal at one locus, and three gave signal at the other (data not shown). Already ch 3 and ch 9 appear to be less likely and ch 5, ch 6, and ch 7 are stronger possibilities.

Then multiple chromosome runs for each model were done, with and without polygenic effect, and a selection of these with or without ch 3. One run had mixing problems (map_wt) and will be discussed separately. None of the multiple runs confirmed the signal on ch 3, but for some of the models not corrected for WT (map_p, ht_map_p), inclusion of ch 3 strongly affected L-scores at other loci. As there was a signal for WT on ch 3, this suggests that perhaps including 3 serves to correct for the WT effect as efficiently as actually modeling the WT correction.

For ch 9, there was still a lot of inconsistency, both in some models giving signal, others not, and switching between two loci. Most of the models included a WT correction, suggesting that there may be something there that is related to LEP independent of WT, but the model is not specific enough to allow for localization. Ch 6 came up for 3 of the models, but 2 of the signals disappeared with the polygenic correction, leaving only one model with a signal. On ch 7, 5 of the signals came up again, but only 1 remained after correcting for the polygenic effect.

In contrast, ch 5 came up in 5 models with a suggestive signal for the sixth. This suggests that especially in the polygenic model, which often removes loci of small effect (correctly or not as described above), when there was a choice between loci, ch 5 was selected. To see if this happened because there was often only 1 QTL in the model, k was set to 5. This did increase the number of QTL in the model but did not effect which signals came up. So from this, the score on ch 5 looks reliably real, while ch 6 and ch 7 are slight possibilities, ch 9 may be real as well but is very hard to localize, and ch 3 has no signal for LEP.

Next the second set of typed individuals was added, and all nine models were analyzed for chromosomes 3, 5, 6, and 7 and 9 only for the models that gave positive signals (Table 6.2).

Table 6.2: Leptin analysis on complete pedigree

model	**Ch 3**	**Ch 5**	**Ch 6**	**Ch 7**	**Ch 9**
p	NS	6.7	NS	NS	ND
wt	NS	**33.4**	NS	10.0	NS
ht	**83.7**	**28.9**	NS	NS	ND
ht_p_m	8.8	**29.4**	ND	ND	ND
map	15.2	**50.7**	NS	NS	ND
map_p_m	7.1	**32.5**	NS	NS	ND
ht_map	**37.4**	**37.1**	NS	**25.5**	ND
ht_map_p_m	NS	**100.0**	NS	NS	ND
ht_wt	NS	**37.1**	NS	NS	ND
map_wt	NS	17.4	NS	11.2	NS
map_wt_p_m	NS	**104.2**	NS	NS	ND
ht_map_wt	NS	**42.7**	NS	8.3	NS
all	14.7	11.5	NS	11.8	ND
all_p_m	NS	**75.5**	8.6	NS	ND

models as listed in Table 5.1; "p" = polygenic effect; "m"= multiple chromosome analysis; "k=5" = setting mean number of loci to 5. "NS" = no signal, "ND" = not done. Scores in bold are 20 or above, $p \leq 0.05$

For the single chromosome analysis, results were as follows: ch 3, the two models without WT (ht, ht_map) came up positive with 2 additional suggestive scores; ch 9, no signal, ch 7, 1 signal (ht_map) and 4 suggestive; ch 6, no signal, ch 5, 6 high and 2 suggestive, essentially all models except the polygenic alone. This clearly suggests that ch 5 contains a real locus for LEP, while ch 3 perhaps has something related to WT. Five multiple runs with the polygenic corrections (ht_p_m, map_p_m, ht_map_p_m, map_wt_p_m, all_p_m) were done, and ch 5 always came up nice and high, the rest not at all, again suggesting that ch 3 served merely to correct for a correlated trait.

As seen in Table 6.3, this locus peaks at 183.5 cM closest to D5S211, has a gene effect size of up to 0.4 on the residual distribution (depending on model), and explains 20 to 100% (mode of 40%) of the genetic variance (depending on how many QTL in the model) and 10 to 25 % of the total variance. Figures 6.1 and 6.2 show how different models / runs affect the L-peak size, shape, and localization. Figure 6.1 compares the scores for all 9 models while Figure 6.2 compares the original run to the multiple polygenic run for each model. As seen in the GAW12 analysis, this suggests that model changes usually do not identify unique loci; rather, the closer the analyzed model is to the

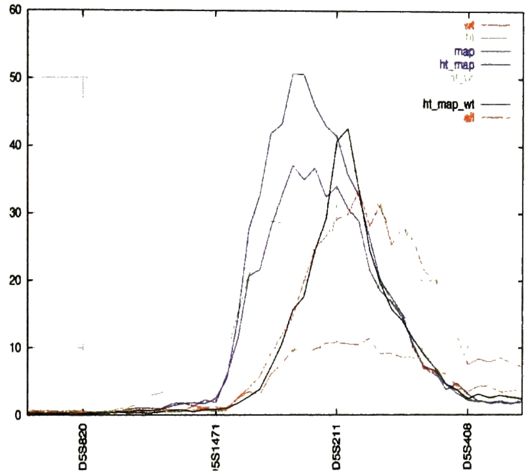

Figure 6.1: LEP on 5

comparison across models

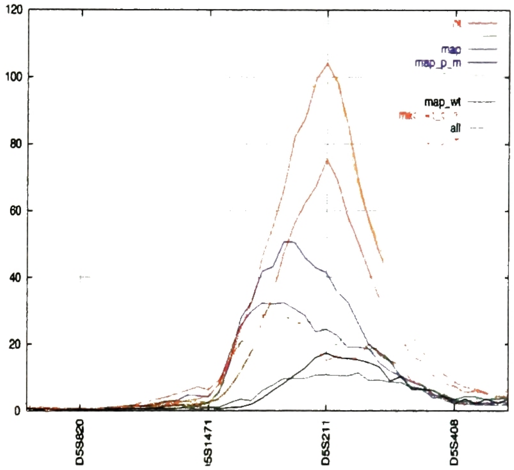

Figure 6.2: LEP on 5

comparison between single and multiple chromosome analvses real model, the cleaner and nicer the L-score graph is. This could help to chose the best model to narrow down the region in fine mapping studies.

Lastly, the map_wt multiple runs point out one problem with MCMC simulation

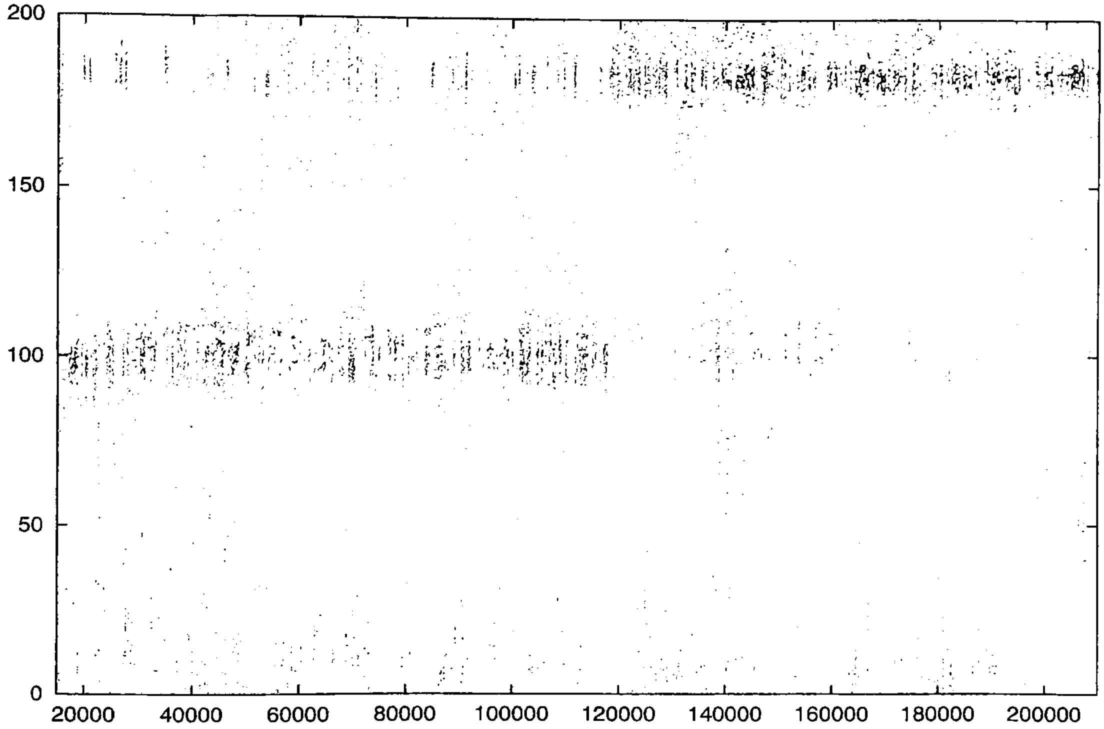

Figure 6.3: LEP for 5 and 6
X axis is iteration, Y axis is cM on ch. Signal at 100 cm is for ch 6, at 180 is for ch 5. Signal is first found on ch 6, and then only on ch 5, which suggests poor mixing.

methods, that of mixing. One way to check mixing is to visually inspect the graph of cM vs. iterations, to see if the signal goes on and off the chromosome location. In one run (polygenic without ch 3) which gave L-scores of 40 for ch 5 and of 49 for ch 6, the signal was seen only for ch 6 for the first 100,000 iterations, and then only ch 5 was found for the next 100,000 iterations (Figure 6.3). For a similar run done twice (polygenic with ch

3), the first time ch 6 and 9 were found, but not ch 5, though ch 5 had been found consistently in the individual runs. The second time, ch 5 and ch 9 were found, but not ch 6 at all (Figure 6.4). This suggests a mixing problem, which is something to be careful for, one of the reasons runs are done multiple times (discussed above). The other analyses looked fine by visual inspection so this could be somehow specific to this model, but it could also indicate a systematic problem. This model was reanalyzed using the hybrid sampler 25% of the time, which runs much slower but all those showed signal for ch 5 only, and appeared to mix very well, further suggesting that the problematic results observed was due to poor mixing. All runs subsequent to this were run with the hybrid sampler 5% of the time, and they all appeared to mix well. Which sampler to use when and at what payoff for speed is under constant investigation.

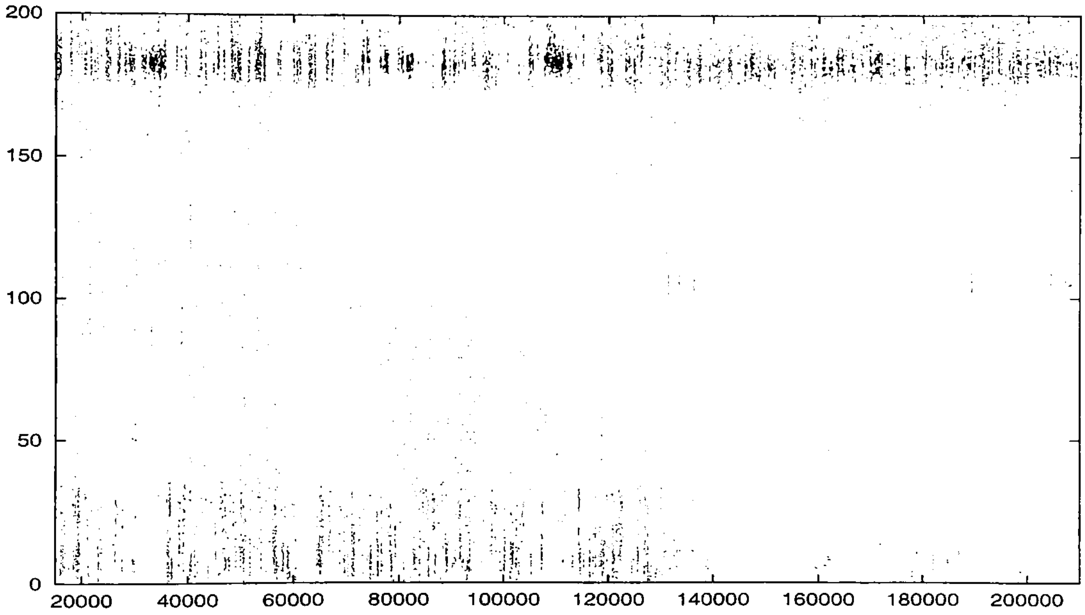

Figure 6.4: LEP on 5 and 6

X axis is iteration, Y axis is cM on ch. Signal at 180 is for ch 5, with no signal for ch 6. The signal can be seen to move off 180 cM, suggesting better mixing

For the other traits, a similar strategy was carried out, but on a lesser scale. For each trait, any analytical change that makes a point relevant to general analysis strategy is discussed and the final significant results are reported. Results are listed in Tables 6.3 and 6.4, with figures as mentioned. In this case, an L-score of 20 was used as a cut-off, nominally a genome-wide p-value of 0.05, though even a score of that size is of borderline reliability so those runs should be repeated many more times to ensure stability. A low score that is stable could be real, and low merely because the model is not precise enough or only a subset of the pedigree is segregating that QTL and would, therefore, merit further study but of lower priority.

Table 6.3: Significant results for all traits

Trait	Chr	Closest marker (peak cM)	Highest L-score	p-value	Effect size	% variance genetic	total	Allele frequency
LEP	5	D5S211 (183.5)	104.2	0.0096	0.2 – 0.4	20 - 100	10 - 25	0.2 / 0.8
BMI	20	D20S171 (96.5)	20.1	0.050	0.9 – 1.2	40 - 70	15 - 25	0.1 / 0.9
	18	D18S967 (16.5)	67.5	0.015	0.9 – 1.2	30 - 70	15 - 30	0.1 / 0.9
HIP	4	D4S1647 (105.5)	20.1	0.050	0.5 – 0.8	30 - 50	5 – 15	rare
	10	D10S1817 (135.5)	176.8	0.006	0.75 – 1.0	50 - 70	10 - 20	rare
WT	18	D18S64 (83.5)	20.0	0.050	3.0 – 4.0	20 - 40	5 - 15	0.1 / 0.9
FBS	1	D1S547 (261.5)	20.0	0.050	0.05	~100	15 - 25	common
SBP	1	D1S534 (146.5)	63.6	0.016	4.0 – 6.0	20 - 40	10 - 25	common
	20	GATA81E09 (32.5)	142.5	0.007	~ 6.5	30 – 50	15 - 20	common
MAP	9	GATA62F03 (6.5)	20.1	0.050	3.0 – 5.5	30 - 70	15 - 30	0.25 / 0.75
APOB	2	D2S1400 (32.5)	37.0	0.027	8.0 – 10.0	40 - 60	~ 20	0.15 / 0.85
TC	12	D12S2078 (152.5)	32.0	0.031	~ 15.0	40 - 80	20 - 25	common
	16	D16S2621 (139.5)	24.4	0.041	12.0 – 16.0	40 - 60	~ 25	common
	19	D19S714 (46.5)	68.2	0.015	~ 14.0	40 - 50	20 – 30	0.25 / 0.75
HT	15	D15S642 (118.5)	30.0	0.033	~ 0.8	15 - 30	15 - 20	common
	5	D5S1471 (172.5)	23. 6	0.042	0.6 – 0.8	10 - 20	8 - 16	0.2 / 0.8
	19	D19S714 (46.5)	28.5	0.035	0.6 – 1.0	10 - 40	~ 20	common
	1	D1S1728 (106.5)	79.4	0.013	0.4 – 0.8	10 - 30	5 - 20	0.2 / 0.8
	10	GATA121A08 (88.5)	314.3	0.003	0.5 – 0.7	10 – 30	5 - 20	rare

Table 6.4: Chromosome regions for significant results

Trait	**Chr**	**range (cm)**	**other studies**
LEP	5	170 - 200	1. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)
			2. National Heart, Lung, and Blood Institute (NHLBI) Family Heart Study, BMI (Feitosa et al. 2002)
			3. French Family study, BMI (Clement et al. 1996)
			4. French study, LEP and BMI (Hager et al. 1998)
BMI	20	85 - qter	1. University of Pennsylvania Family Study, % body fat (Lee et al. 1999)
			2. Quebec family Study, % body fat, BMI, INS, skinfolds, and fat mass (Lembertas et al. 1997)
			3. French study, BMI and skinfolds (Comuzzie and Allison 1998)
	18	pter - 40	1. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)
			2. Quebec Family Study, BMI, skinfolds, fat mass, and % body fat (Chagnon et al. 1997)
			3. For the HT/DYSL factor, Rochester Family Heart Study, APOA-II and APOC-II (Klos et al. 2001)
HIP	4	90 - 115	Nothing
	10	125 - 140	French study, LEP (Hager et al. 1998)
WT	18	75 - 105	1. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)

			2. Finnish study, obesity (Ohman et al. 2000)
			3. Icelandic study, Peripheral Arterial Occlusive Disease (PAOD) (Gudmendsson et al. 2002)
SBP	1	100 - 160	see overlap region in Table 6.5
	20	25 - 40	1. GENOA Network of the Family Blood Pressure Program in Rochester, MN, SBP (Krushkal et al. 1999)
			2. San Antonio Family Heart Study, pulse pressure (Atwood et al. 2001b)
			3. Framingham Heart Study, TG and TG/HDL-C ratio (Shearman et al. 2000)
MAP	9	pter - 45	1. Old Order Amish Study, DBP and SBP (Hsueh et al. 2000)
APOB	2	20 - 45	1. Dutch Family Study, Familial Combined Hyperlipidemia (FCHL) (Aouizerat et al. 1999)
			2. Pima Indians, TG (Imperatore et al. 2000)
			see overlap region with BP in Table 6.5
TC	12	140 - 170	Rochester Family Heart Study, APOA-I (Klos et al. 2001)
	16	125 - qter	Nothing
	19	35 - 50	1. Rochester Family Heart Study, TC (Klos et al. 2001)
			2. Pima Indians, TC (Imperatore et al. 2000)

			3. San Antonio Family Heart Study, LDL and LDL2 (Rainwater et al. 1999) 4. Washington University Medical Center Study of Familial Hypobetalipoproteinemia, TC (Yuan et al. 2000)
HT	1	100 - 120	Swedish population study, HT (Hirschhorn et al. 2001)
	5	160 - qter	Nothing
	10	80 - 95	Diabetes UK Warren 2 Consortium, HT (Wiltshire et al. 2002).
	15	110 - qter	Finnish population study, HT (Hirschhorn et al. 2001)
	19	30 - 55	Diabetes UK Warren 2 Consortium, HT (Wiltshire et al. 2002).
FBS	1	210 - 275	1. Pima Indians, OGTT (Pratley et al. 1998) 2. Finnish Study in Botnia, NIDDM (Lindgren et al. 2002) 3. Finland – United States Investigation of NIDDM Genetics (FUSION), INS and 2 hour INS (Watanabe et al. 2000) 4. French Caucasians, NIDDM (Vionnet et al. 2000)

Obesity

BMI: For BMI, the addition of new individuals made a big difference, as significant scores were found on ch 20 and ch 18. The ch 20 score was borderline (20.1) and when repeated twice decreased to 10 (Figure 6.5); but HIP also mapped suggestively to that locus (not shown) and it mapped directly onto a marker, which adds reliability. This locus peaks at 96.5, at D20S171, and has an effect size of around 1, which explains 40 to 70 % of genetic and 15 to 25 % total variance. The more robust score was found on ch 18 (67.5, Figure 6.6) but the peak is split directly on the marker, GATA88A12. This has been seen before when the QTL inheritance model was mispecified, either because it actually had 3 alleles, or there are interactions that could not be modeled using Loki. Another possibility is that there are genotyping errors that prevent proper localization due to observed recombinants. As mentioned before, reporting and checking recombinants is currently under investigation. Nevertheless, this locus at 16.5 cM / D18S967 appears real, and it has an effect size of close to 1.2, which explains 30 to 70 % of genetic and15

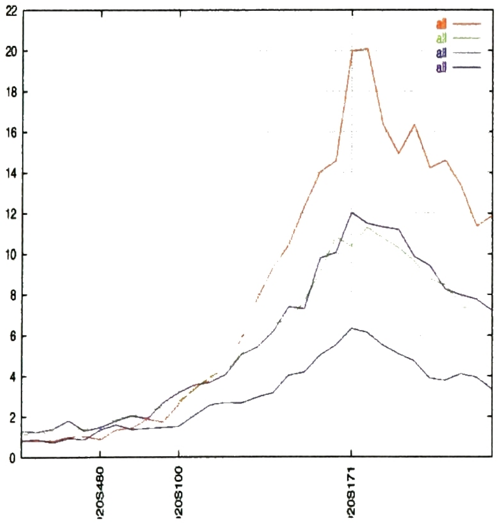

Figure 6.5: BMI on 20

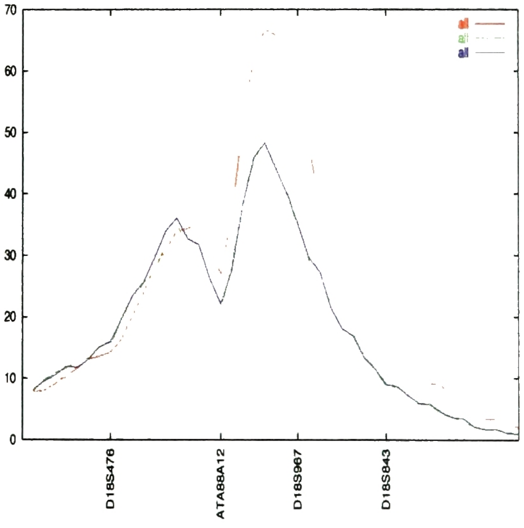

Figure 6.6: BMI on 18

to 30 % total variance. Additionally, for one model LEP mapped close by, as did the HT/DYSL factor (not shown).

HIP: For HIP, the additional data had an obvious effect as well, as it resulted in two loci. One is a borderline score (20.1) for ch 4 (Figure 6.7), at 105.5 / D4S1647, with an effect of 0.5 to 0.8, which explains 30 to 50 % genetic and 5 to 15 % total variance.

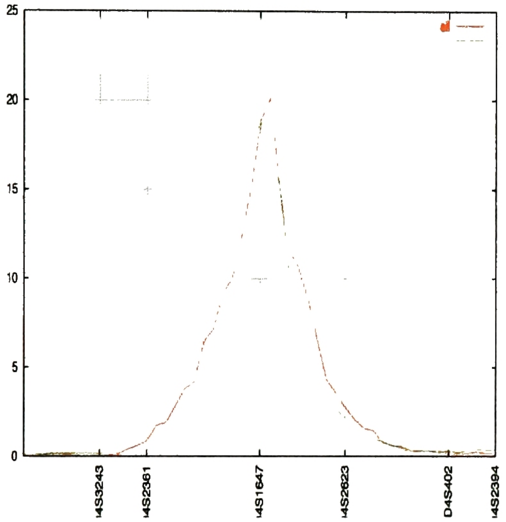

Figure 6.7: Hip on 4

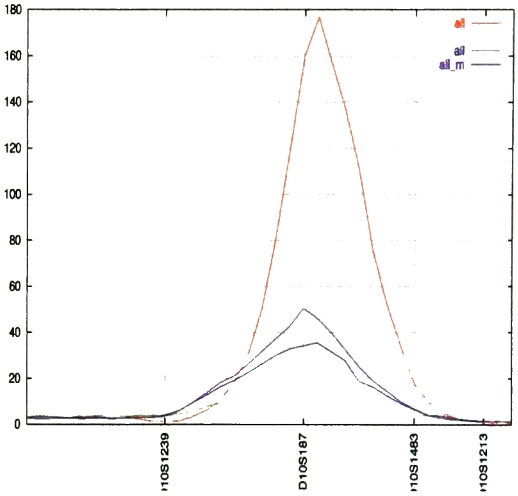

Figure 6.8: HIP on 10

The second is a relatively high score on ch 10 (Figure 6.8, highest 176.8, other runs around 40), at 135.5 / D10S1817, which overlapped with a slight WT signal (not shown). This has an effect size of 0.75 to 1, which explains 50 to 70 % genetic and 10 to 20 % total variance. Both scores localized on a marker.

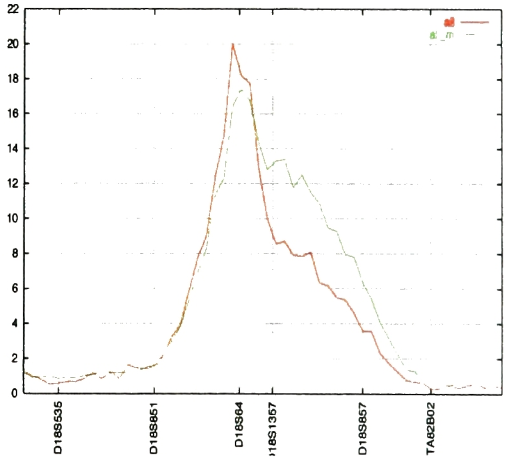

Figure 6.9: WT on 18

WT: The only result for WT is a

borderline peak (20.0) on ch 18 (Figure 6.9), but this signal may be retroactively more reliable as it peaks at 83.5, closest to D18S64, which is close to a candidate gene, MC4R. This has an effect size of 3 to 4, which explains 20 to 40 % of genetic and 5 to 15 % of total variance. It is possible that signal can only be picked up when the % variance reaches a threshold and here that may be the reason the score is consistent and on a marker, but low.

Diabetes

FBS: The only score for FBS is borderline (20) and not well localized (Figure 6.10) on ch 1. This has an effect size of 0.05 (ln FBS), which explains all the genetic and 15 — 25 % total variance. In general, FBS is the quantitative trait least likely to identify genes for its associated disease, as diabetics have severe derangement in their FBS levels, and therefore were excluded from the analysis, yet they are the ones that potentially have the most information. This analysis is really looking for genes that affect the normal distribution of FBS, and a qualitative analysis should be done for diabetes. If this is a real locus, as is suggested by many other similar studies (see Table 6.3 and discussion), the localization problems may be due to this issue.

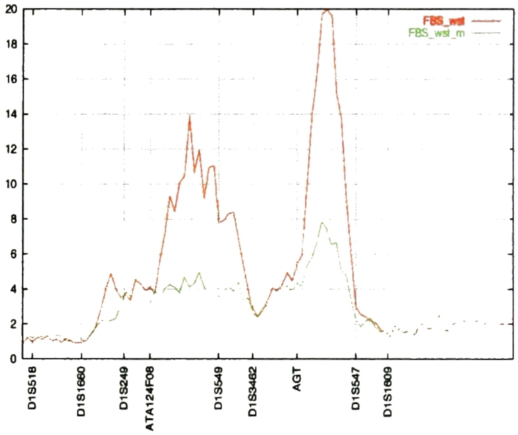

Figure 5.6: **FBS on 1**

Hypertension

SBP: SBP analysis resulted in two loci. One signal on ch 1 is quite difficult to localize, though it overlaps with slight scores for other blood pressure measures (Figure 6.11) and many other traits (Figure 6.23), which will be discussed later. The highest peak (63.6) is at 146.5 / D1S534, with an effect size of 4 to 6, that explains 20 to 40 % genetic and 10 to 25 % total variance. The signal on ch 20 is much higher (142.5) and nicer, localized to the marker at 32.5 / GATA81E09 (Figure 6.12), and this has an effect size of 6.5, which explains 30 to 50 % genetic and 15 to 20 % total variance. Seen here is the increase in power with the addition of markers in large gaps, as this was a new marker (GATA81E09) typed on the whole pedigree with the addition of the second set. This allowed for the localization of a QTL previously put off the end of the

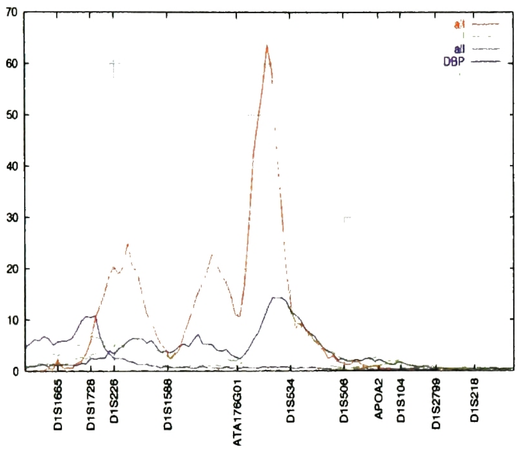

Figure 6.11: SBP on 1

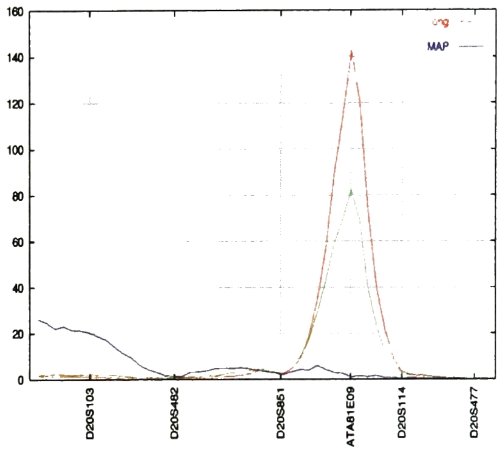

Figure 6.12: SBP on 20

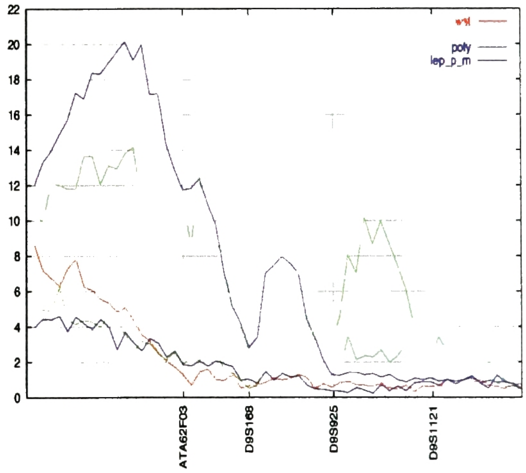

Figure 6.13: MAP on 9

chromosome, though MAP is still off the end.

MAP: The signal off the end of ch 20 for MAP (Figure 6.12) is probably due to the correlation to SBP and is not considered a signal for this trait. The other borderline signal (20.1) on ch 9 is hard to localize (Figure 6.13) at the end of the chromosome. It is interesting as it only comes up for the LEP model in the multiple chromosome scan, suggesting some sort of interactions that cannot be modeled using Loki. This QTL peaks at 6.5 / GATA62F03, has an effect size of 3 to 5.5, which explains 30 to 70 % genetic and 15 to 30 % total variance. DBP also maps to both ends of ch 9 (Figure 6.13), which further suggests that there may be some signal on 9 that cannot be localized using our dataset and analytical methods.

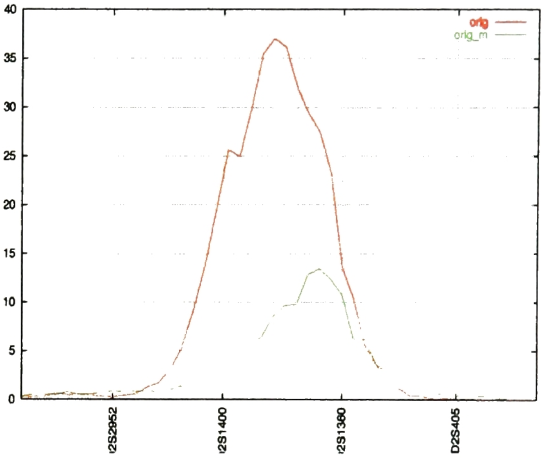

Figure 6.14: APOB on 2

Dyslipidemia

APOB: APOB results in one signal (37.0) on ch 2, which peaks between two markers, at 32.5 / D2S1400 (Figure 6.14). The gene effect is 8 to 10, which explains 40 to 60 % of genetic and 20 % of total variance.

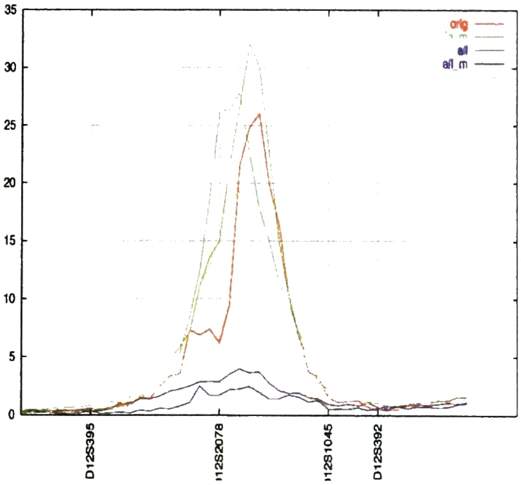

Figure 6.15: TC on 12

TC: TC was quite an interesting trait for a number of reasons. One is that two models gave different loci, all of which appear to be reliable. Secondly, for one model (all) there were two signals that came up together. Thirdly, correcting prior to the analysis for all the quantitative traits and analyzing the residuals increases the L-score. Fourthly, overlap between this trait and the DYSL factor is seen. Lastly, correcting for a major gene effect is seen as well. For the model without correction for other quantitative traits, there is a locus on ch 12 (L-score 32.0). This peaks between two markers, at 152.5 / D12S2078 (Figure 6.15), and has an effect size of about 15, which explains 40 to 80 % genetic and 20 to 25 % total variance. This directly overlaps with the DYSL factor.

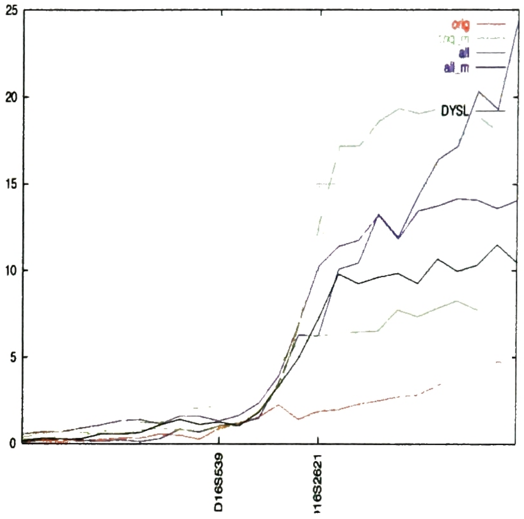

Figure 6.16: TC on 16

For the model with all corrections, two loci came up, on ch 16 (24.4) and ch 19 (68.2). The one on 16 is less reliable, as it is off the end of the chromosome, peaking at 139.5 / D16S2621 (Figure 6.16). This has an effect size of 12 to 16 that explains 40 to 60 % genetic and 25 % total variance. Often signals are put off the end of a chromosome because they cannot be mapped to the real region on the

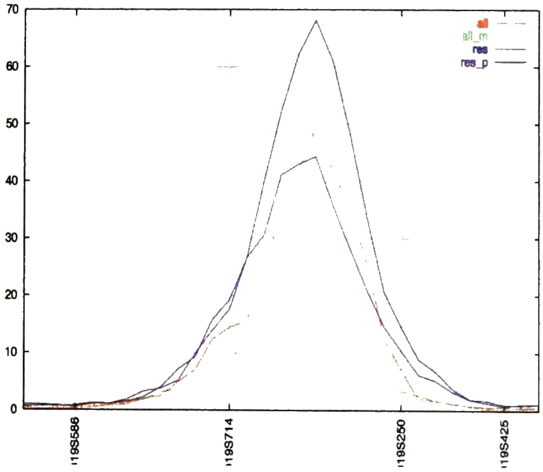

Figure 6.17: TC on 19

chromosome, such as with SBP and ch 20. There is a very slight signal at D16S772 (not shown), which suggests that perhaps that is really the locus. The DYSL factor also mapped directly under this peak.

The second locus seen is on ch 19, between two markers, at 46.5 / D19S714 (Figure 6.17), with an effect size of about 14, which explains 40 to 50 % genetic and 20 to 30 % total variance. APOB has a slight signal at this locus as well (not shown). As both signals are seen together, though they have the same estimates, it is not likely that it is the same QTL being localized to two different regions. To make sure the signals for models with all corrections were not due to correlation among covariates causing sticking, the residuals (TC_res in the figures) were analyzed on ch 16 and ch 19, and the scores either remained or increased, so actually the opposite is true. In this case, cleaning the data in advance simplifies the localization. To investigate correcting for APOE, a major gene on ch 19, a number of runs with or without this correction were done. The L-scores were more consistently found and more stable with this correction (not shown).

Height

HT is also an interesting trait as three different models give different loci most of which appear real. The model with no correction for other traits had one signal on ch 15 (L 30.0), which is not well localized and off the end of the chromosome, though it does peak

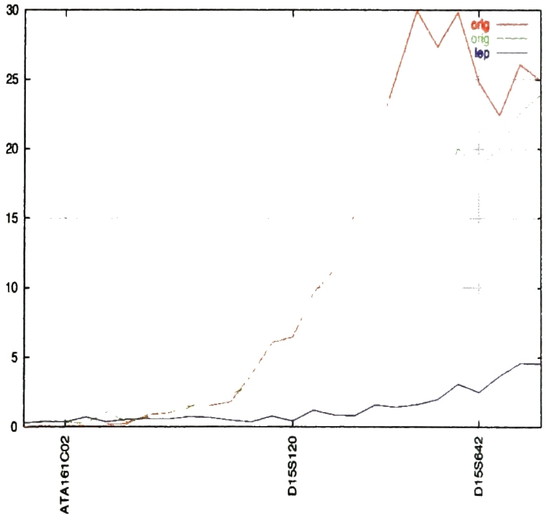

Figure 6.18: HT on 15

between the last two markers at 118.5 / D15S642 (Figure 6.18). This has an effect size of 0.8, which is 15 to 30 % genetic and 15 to 20 % total variance.

The model with correction for LEP resulted in 3 scores, all of which came up together a significant portion of the time. On ch 5, the peak (23.6) is split on the marker D5S211, but it peaks at 172.5 / D5S1471 (Figure 6.19). This one has an effect size of 0.6 to 0.8, which explains 10 to 20 % genetic and 10 to 15 % total variance. Possible causes of split peaks are discussed above, though here it is probably not due to genotyping error, as no split peaks were seen for any of the leptin models. On ch 19 (28.5), the peak is between two markers at 46.5 / D19S714 (the TC locus), with effect size 0.6 to 1, which

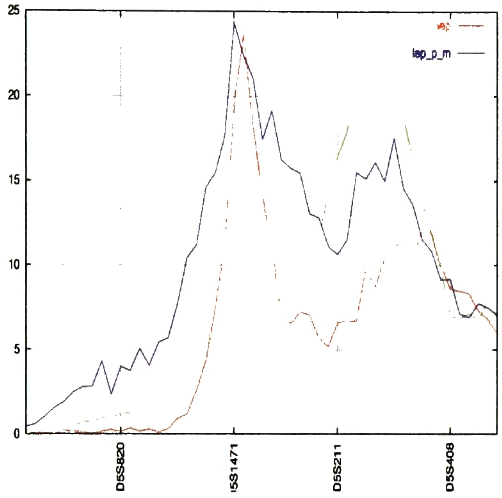

Figure 6.19: HT on 5

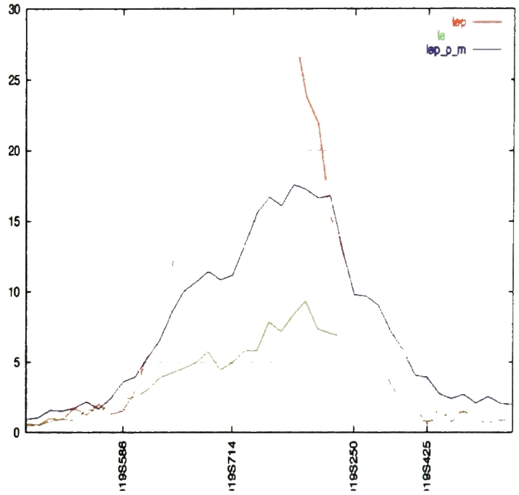

Figure 6.20: HT on 19

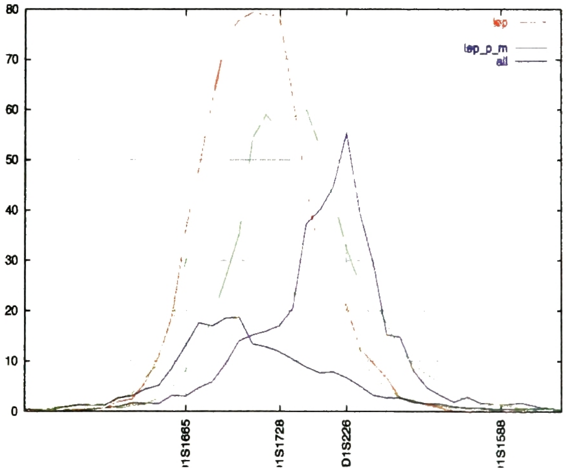

Figure 6.21: HT on 1

explains 10 to 40 % genetic and 20 % total variance (Figure 6.20). Though both these loci are consistent, neither are particularly high or well localized.

The locus on ch 1 is found with both this model and the one with all corrections. The signals are robust (79.4) but the localization is slightly different with the two models, so the best one should be used for further analyses (Figure 6.21). This peaks at 106.5 / D1S1728, has an effect size of 0.5 to 0.8, and explains 10 to 20 % genetic and 5 to 20 % total variance. The final signal, found only for the all model, is on ch 10, and it is the nicest score from the whole dataset, as it has the highest L-score (314.3), is consistently found, and maps tightly directly on top of a marker, GATA121A08 / 88.5 (Figure 6.22). This has an effect size of 0.5 to 0.7, which explains 10 to 30 % genetic and 5 to 20 % of total variance. This signal was found together with ch 1, again suggesting that though the estimates are quite similar, these are two distinct loci. As HT has traditionally been known as a polygenic trait, all the runs were analyzed with the polygenic effect, and they all remained, though some of the scores were slightly decreased.

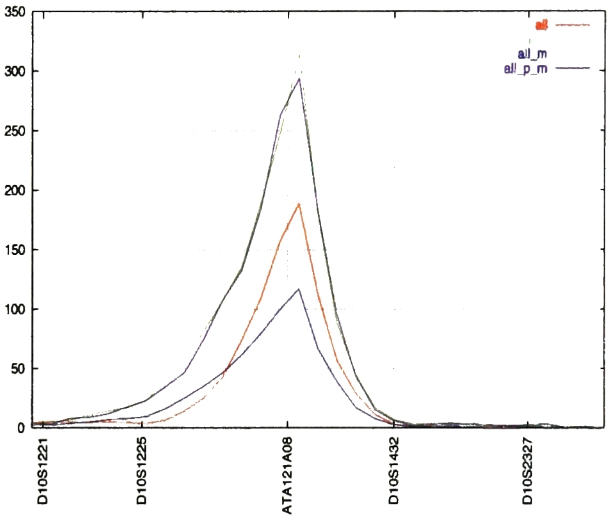

Figure 6.22: HT on 10

No compelling signals were found for WST, INS, DBP, APOA-I, and TG.

Chromosome regions with mainly borderline signals from multiple correlated traits are listed in Table 6.5 and can been seen on the figures as mentioned.

Table 6.5: Overlap regions

chr	Region (cM)	traits	Other studies
1	90 – 125	MAP, DBP, BP	1. Quebec Family Study, SBP (Rice et al. 2000)
	100 - 160	SBP	2. Icelandic study, PAOD (Gudmendsson et al. 2002)
			3. HyperGEN of NHLBI, renal function in hypertensives (DeWan et al. 2001)
	170 - 182	OB/DB, BP	1. Diabetes UK Warren 2 Repository, NIDDM (Wiltshire et al. 2001)
	170 – 212	BMI	2. FUSION study, insulin resistance and FBS (Watanabe et al. 2000)
			3. Caucasian study, BMI (Comuzzie and Allison 1998)
			4. University of Pennsylvania Family Study, % body fat (Lee et al. 1999)
			5. Finnish study, obesity (Ohman et al. 2000)
			6. Pima Indians, waist-to-thigh ratio, 24 hour ratio of carbohydrate oxidation to fat oxidation, 24 hour energy expenditure, and sleeping metabolic rate (Norman et al. 1998)
			7. Utah Caucasians, NIDDM (Elbein et al. 1999)
			8. Pima Indians, NIDDM (Hanson et al. 1998)

			9. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001) 10. NHLBI Family Heart Study, BMI (Feitosa et al. 2002) 11. Studies from FBS region in Table 6.4 with signal in this region as well: French Caucasians (Vionnet et al. 2000) and Botnia Study (Lindgren et al. 2002)
2	20 – 50, - 75	APOB, MAP, DBP, SBP, BP	1. Finnish study, coronary heart disease (Pajukanta et al. 2000) 2. GENOA Network of the Family Blood Pressure Program, SBP (Krushkal et al. 1999) 3. Old Order Amish Study, SBP (Hsueh et al. 2000) 4. Dutch Family Study, FCHL (Aouizerat et al. 1999) 5. Quebec Family Study, DBP (Rice et al. 2000) 6. Framingham Heart Study, DBP (Levy et al. 2000)
	90 – 130	MAP	1. San Antonio Family Heart Study, DBP and SBP (Atwood et al. 2001a) 2. San Antonio Family Heart Study, PP (Atwood et al. 2001b) 3. Old Order Amish Study, DBP (Hsueh et al. 2000) 4. HyperGEN of NHLBI, change in SBP (Pankow et al. 2000)

			5. Quebec Family Study, DBP and SBP (Rice et al. 2000)
7	50 – 90	FBS, LEP	1. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)
			2. GENNID Study of 4 American populations, Mexican American group, NIDDM (Ehm et al. 2000)
			3. NHLBI Family Heart Study, BMI (Feitosa et al. 2002)
			4. Old Order Amish Study, LEP (Hsueh et al. 2001)
	110 – 140	LEP, HT/DYSL	1. NHLBI Family Heart Study, BMI (Feitosa et al. 2002)
			2. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)
			3. San Antonio Family Diabetes Study, HDL-C (Duggirala et al. 2000)
	150 – qter	DYSL, OB/DB	1. San Antonio Family Diabetes Study, TG (Duggirala et al. 2000)
			2. Framingham Heart Study, TG and TG/HDL-C ratio (Shearman et al. 2000)
			3. Pima Indians, FBS and various insulin measures (Pratley et al. 1998)
9	105 – 135	OB/DB, HIP	1. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)
			2. GENNID Study, Caucasian American group, NIDDM (Ehm et al. 2000)
			3. FUSION study, insulin resistance (Watanabe et al. 2000)

13	20 – 60	WST, WT	1. Quebec Family Study, abdominal subcutaneous fat (Perusse et al. 2001)
			2. NHLBI Family Heart Study, BMI (Feitosa et al. 2002)
			3. French study, % body fat and skinfolds (Comuzzie and Allison 1998)
			4. FUSION study, BMI, C-Peptide, and INS (Watanabe et al. 2000)
16	100 – 130	OB/DB, FBS	1. Pima Indians, INS and FBS (Pratley et al. 1998)
			2. GENNID Study, African American group, NIDDM (Ehm et al. 2000)
			3. Finnish Study, NIDDM (Lindgren et al. 2002)

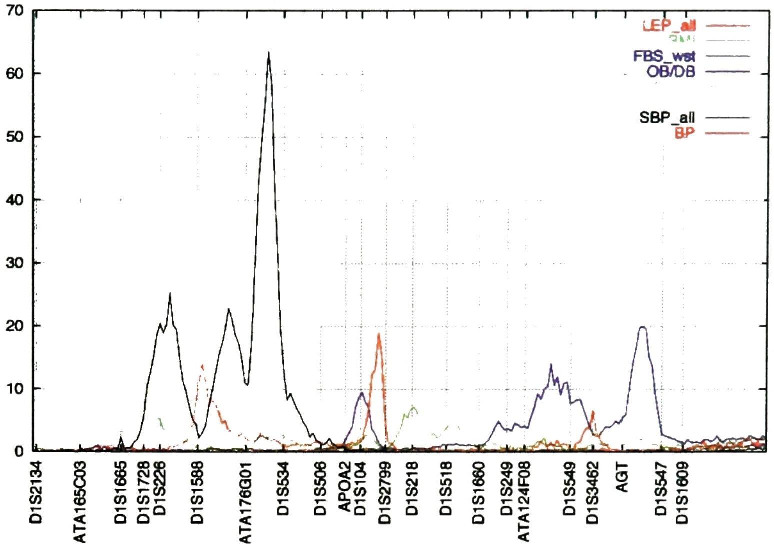

Figure 6.23: Chromosome 1

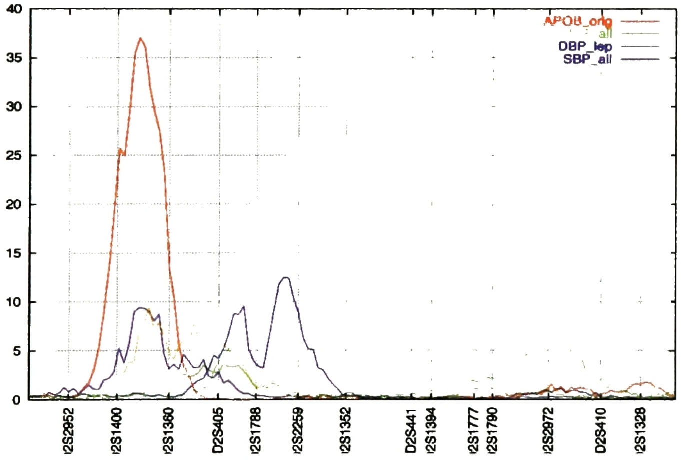

Figure 6.24: Chromosome 2

The overlap region on ch 1 includes the following regions: 90-125 cM, with signal for MAP, DBP, SBP, and residual BP factor; 125 — 163, for LEP then SBP; 170 — 182, for OB/DB and BP; 170 — 212, for BMI; 212 — 275, for FBS (Figure 6.23). The region on ch 2 includes: 20 — 50, - 75, with signal for APOB, MAP, DBP, SBP, BP, and residual BP; and 90 — 130, for MAP (Figure 6.24). The region on ch 7 includes: 50 —90, with signal for LEP and FBS; 110 — 140, for LEP, and residual HT/DYSL; 150 — end, for DYSL and residual OB/DB (Figure 6.25). The region on ch 9 includes 105 — 135, for OB/DB and hip (Figure 6.26). The region on ch 13 includes 20 — 60 for WST and WT (Figure 6.27). The overlap region on ch 16 includes 100 — 130 for FBS and OB/DB (Figure 6.28). These are interesting in that there is some sort of signal there, but no further mapping studies can be done until a better model or trait can be used.

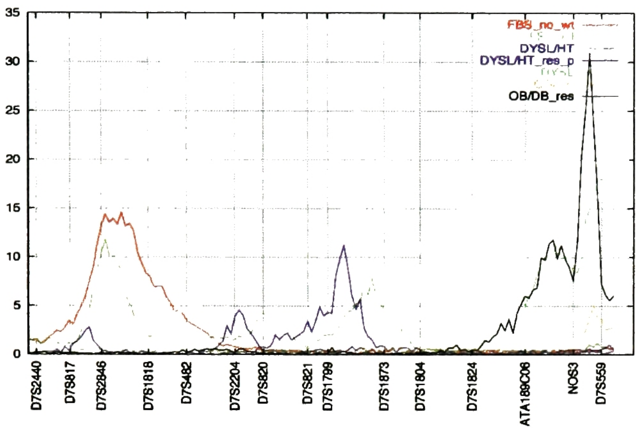

Figure 6.25: Chromosome 7

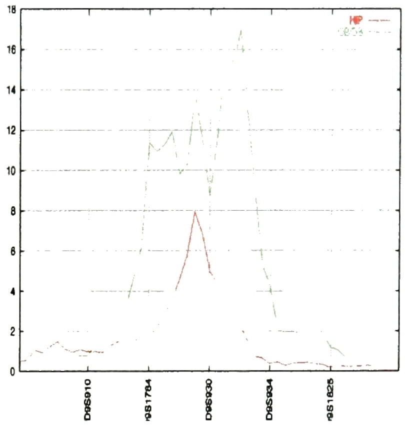

Figure 6.26: Chromosome 9

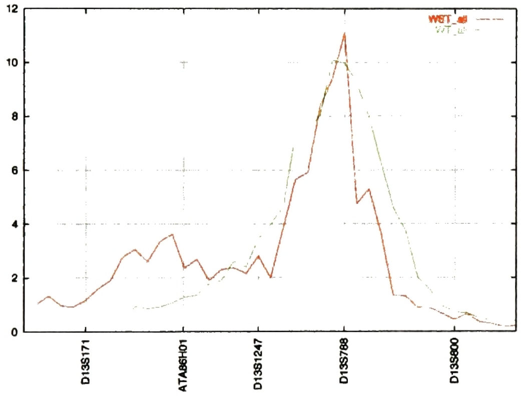

Figure 6.27: Chromosome 13

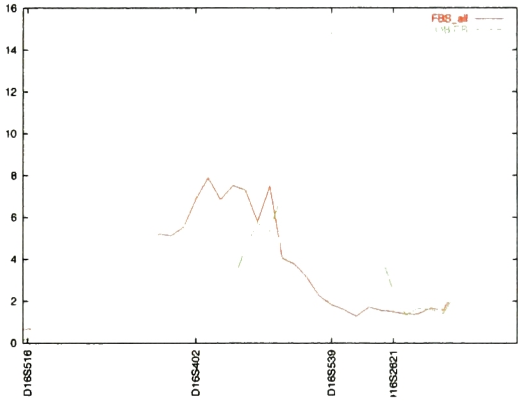

Figure 6.28: Chromosome 16

Finally, there were a few analysis parameters tested as well. Using the setting for the Kosambi map, all LEP scores replicated, as well as the HT score, while the APOA-I score was much lower, and the WT score was really low as well. As seen, this helped to reduce false positives, and therefore this setting should be used in future Loki analyses. As of now, there is no way to calculate L-scores for the sex-specific map analyses, but again the LEP scores and HT were found in these analyses. As it is not the best idea to do a scan with the current sex-specific maps, it may be an idea to recheck positive scores more this way. Lastly, as mentioned, setting k to 5 did not increase the ability to detect loci, but it did allow for more QTL in the model, without any penalty, and, therefore, there is no reason not to use that setting.

These results fulfill the dual purposes of clarifying how to carry out a genome scan and identifying QTL for the traits encompassed by Syndrome X. These will be further discussed in the next section.

Chapter 7: Discussion

Introduction

One current challenge in human genetics is to find genes that increase susceptibility to common complex disorders. One way to do this is to map genes that contribute to the variance in the underlying quantitative traits for these diseases. To do this powerfully, it is often necessary to study large complex pedigrees with multipoint analysis. This is quite problematic analytically, as the widely used linkage programs are either limited to a few markers (Elston-Stewart algorithm, such as in LINKAGE, FASTLINK) or to small pedigrees (Lander-Green algorithm, such as in Genehunter) due to computational limits. Therefore, it is suggested to use sampling methods, such as MCMC, to generate samples of possible configurations to identify the most likely configuration, instead of specifying every possibility exactly. This has been implemented in a LA program, Loki, developed by Simon Heath (Heath 1997), and it has been used in a small number of genome scans, on both simulated and real datasets (Daw et al. 1997; Heath et al. 1997; Yuan et al. 2000; Shmulewitz and Heath 2001). This method is very promising, but it not yet clear what the exact properties of Loki are and how to best use it for a genome scan.

The purpose of this study is to investigate Loki, to develop a strategy to use for stringent LA, and then to do a genome scan to identify QTL for traits encompassed by Syndrome X as well as height in the population of Kosrae. The testing was done first on simulated data from GAW12 with known answers so its performance could evaluated, and then the method was further adapted for real data from the Kosrae population. Once a working strategy has been developed, the results from this study were evaluated, to identify regions linked to obesity (BMI, HIP, WT, LEP), diabetes (FBS), hypertension (SBP, MAP), dyslipidemia (TC, APOB), and HT. First, how to do a scan will be discussed, focusing on what affects the analysis and how to interpret the output. Second, the loci that have been elucidated will be discussed in the context of similar studies.

General Strategy

The basic strategy as suggested from the GAW12 analysis is to do an initial scan, 1 to 2 chromosomes at a time, each for 100,000 iterations, with simple models, and then to test some model changes and multiple chromosome analyses. As this is a sampling scheme, to test the final results, the analyses should be repeated a few times, with more iterations, to make sure the sampler is mixing and the estimates are stable. For the real data, it appears necessary to do many more levels of testing to reduce false positives and to be sure about the real positives. Again, first initial runs with simple models were done to get a general idea as to what is going on. Next, the genome scans are redone with model changes. With a good idea as to what models seem to be most informative for each trait and some initial positive results, multiple chromosome runs should be performed multiple times. Lastly, the mixing should be inspected, as should the size and shape of the L-score peaks, to have confidence in the localization and to prioritize which QTLs seem most real.

Phenotypic and genotypic model corrections

Loki attempts to map genes that explain a percentage of the variance in a trait. The more variance that can be explained by environmental covariates or other genetic effects, the better the major gene being mapped explains the residual variance, and therefore it will be easier to detect and localize. As it may not always be obvious what to correct for, a number of models should be tested to see if there is any overlap of loci between models. Obvious corrections are for traits that are phenotypically correlated in the population under study or those that mapped to the same locus in the initial scan (such as HT / LEP on ch 5 and LEP / BP on ch 9). Also, correcting for all the quantitative traits worked well for some traits, perhaps for those that the other traits are not expected to directly influence, such as HT. But this may not work so well for traits that influence each other (perhaps like LEP and FBS) or are genetically related to the others. As the correlation could be genetic, it would also be interesting to find that a combination of traits maps to the same locus, which is why all the possible models and some sort of combinatorial phenotypes were analyzed. Of course, it is highly possible that there are important corrections that cannot be made due to lack of information and that could be why it was hard to either detect or precisely map QTL for some traits.

Results may vary across models, which aids in evaluation of the "realness" of the signals. If a signal is consistent across models, such as LEP for ch 5, that fact itself adds reliability. In this case, it is clear that the model affects the accuracy of localization more than the ability to detect a QTL. In other cases, there are really nice signals that only come up with certain models, such as HT on ch 10, suggesting in this case that those corrections are important for QTL detection. Another possibility is that certain corrections lead to a different form of the trait than others, which maps to a different location, so the different QTL are real. Important to remember is that a signal cannot be evaluated based on one run alone, rather there needs to be a complete picture obtained from various analyses.

Correcting for other genetic effects appears to be important as well. One way is to account for the effects of polygenes, which cleans up some of the genetic variance not attributable to the QTL being mapped, leading to increased scores, such as HT on ch 10. As discussed, at times the polygenic effect is hard to model using Loki, so if the signal is reduced with this effect, this does not automatically mean the signal is not real. But if the score remains or goes up, it is probably real. Another correction to do is a known major gene effect, such as APOE for TC, which gave more stable results. Another way to do this (if the major gene itself is not known) could be by jointly analyzing chromosomes with positive signals, where the slight signal on one makes the others easier to find, such as seen with ch 3 and ch 5 for LEP. This moves on to the next stage of analysis, which is multiple chromosome runs.

Multi-chromosome analyses

When all positive chromosomes for a trait model are jointly analyzed, this often results in some scores increasing, with others decreasing or disappearing. As discussed before, this could be for a variety of reasons, such as that the sampler is choosing the best locus, there is locus heterogeneity, or that the sampler is not mixing well. As most models are not constrained to only one QTL, it is not really a concern that one signal is being chosen at the expense of another. For both TC and HT, the multiple runs clearly chose more than one signal per iteration. When two signals are found but never together, it could be there are two distinct configurations in the data that lead to these different scores. This could mean that only one is real, but there is not enough information to completely determine which one. Another possibility is that there is heterogeneity, with a subset of the pedigree segregating one or the other QTL. Testing this is quite difficult in one large pedigree and is currently under investigation.

Multi-chromosome runs require careful inspection of the iterations to make sure the sampler is mixing, as discussed in the LEP results section. This highlights the mixing problem inherent in the sampling scheme and which is under constant investigation. Another effect of this problem is the inability to use markers closer than 5 cM in the same analysis, as the correlated segregation of alleles at linked loci may cause sticking. This method is not yet able to do finer mapping, so the resolution can be quite low. One way to really ensure that the answer is real is to repeat the analysis a few times from different starting points to see that the scores and estimates are stable.

L-graphs and parameter estimates

With a set of signals that are consistent and therefore seem to be real, the next thing to do is to inspect the L-score peaks for height and shape. L-scores are not necessarily used as a test of significance, but rather as a starting point for identifying potentially interesting linkage regions. As discussed above, this is mainly because the L-score height is affected by covariate and genetic corrections, so a locus could have a score ranging from 10 to 100 depending on model or run. This is one of the reasons it is so important to do many tests and then to evaluate the signals. When necessary, a cutoff of 10 was used to decide which signals to follow-up for multi runs or reanalysis with the complete pedigree, but in the final results signals of 20 and above were reported. Scores around 20 are borderline, and have to be looked at more carefully to decide if they could be real. Of course, higher scores mean that there is more evidence for linkage, and that does add reliability, but not on its own, as discussed.

The next thing to look at is peak localization. Peaks that localize off the ends of a chromosome are certainly suspect, such as TC on ch 16 and MAP on ch 9 and ch 20. On ch 20 the signal for SBP was able to move itself onto the chromosome with the addition of more information (both individuals and another marker), suggesting that the end of the chromosome is a default place for QTL that cannot be mapped well. Peaks that are split on a marker, such as BMI on ch 18 and HT on ch 5, suggest a problem either with the model (in general many problems of localization can be due to mispecifying the model) or with genotyping. This may be going on with BMI on ch 18, but it is less likely for ch 5 where LEP mapped there with no problem. Next, the width of the peak should be looked at, mainly to see how wide a region needs to be investigated for fine mapping. Sometimes one model gives a tighter (better localized) peak that would be the best model to use in further analysis, such as LEP on ch 5 and HT on ch 1. The nicest, most reliable, and easiest to follow-up peaks are those that map narrowly to a marker, such as HT on ch 10.

The parameter estimates are much more stable across runs than L-graphs. As this is a sampling method, the parameter values are hard to determine exactly, as they change by iteration, depending on the data configuration. As the trait is mapped based on summing over the genotypes at the QTL, which allele is disease and which is normal is not fixed. Actually, the alleles switch at each iteration, so which allele is which cannot be determined and the allele frequencies are symmetrical. Also, one allele / model (dominant or additive) combination is chosen by iteration, and this changes as well. Therefore, these estimates of effect are not informative; rather the combination of both, the gene effect size, is used, as discussed before. From this, the amount of variance of the trait explained by this QTL can also be calculated. Here, genes that explain as low as 5 or 10 % of the total variance were found, possibly because the gene explains a greater amount of the residual variance, which is actually what is being modeled, or because the Loki analyses have more power.

Miscellaneous changes

A number of other changes were evaluated, with the following suggestions. First, setting the input map to Kosambi centimorgans should be done. Second, it is not a bad idea to reanalyze positive scores using sex-specific map, as it may help with localization or to decrease false positives. But the caveat here is that these maps themselves are not so precise, so it would be better to build maps from the data if possible. Third, the hybrid sampler should be used for a small percent of iterations. Fourth, k should be set to some value higher than 1. Two other points to note is that the dataset is much more powerful when the entire pedigree is analyzed at once, and adding markers in gaps aids in localization.

Comparison to other linkage studies

Taking all these steps together, one should be able to make a prioritized list of QTLs worthy of further investigation, which is the real purpose of a genome scan, as QTLs on their own provide very little information. One way to further prioritize is to see if any of these loci have been found in similar studies and / or if they are close to obvious candidate genes. If any QTL is a replicate, that adds reliability, but again the opposite is not necessarily true. This study is different from others in many ways, and therefore one would not expect to get the same answers. In general, there are few genome scans that have the power of the Kosrae study. This could be due to computational difficulties, that multipoint analysis on pedigrees of this size and complexity cannot be done. Also, many studies do not have all the quantitative and covariate data that is available here. Therefore a lot of studies of these diseases have been done in less powerful ways (such as qualitative, no corrections, and two-point analysis), so the results are not directly comparable. Also, these metabolic traits have been shown to have different pathology in different racial or ethnic groups. As this study was done in the population on the South Pacific island of Kosrae, it could be some of the loci found are specific to that ethnic group, and therefore would not be replicated in most other studies. Another possibility is that the mutations came in with the Caucasian admixture, and then perhaps these could be found in other studies of Caucasians. It is also possible that there are loci common to all ethnic or racial groups.

Table 6.4 shows which other studies found linkage in the same region for the same or a correlated trait. Below each trait locus is discussed in turn, with comments on reliability based on everything discussed above.

Obesity

The LEP signal on ch 5, from170-200 cM, is very reliable, as it came up with all models, and in the multiple chromosome runs, especially the polygenic analyses. The peaks were high and nice, though the localization was a bit different by model. This region has signal for other obesity traits in the following studies: Quebec Family Study, with 156 families and 521 individuals, had a suggestive LOD score for abdominal subcutaneous fat (~ 1.5) at the end of the chromosome (Perusse et al. 2001); National Heart, Lung, and Blood Institute (NHLBI) Family Heart Study, with 401 families and

3027 individuals, for BMI in this region (Feitosa et al. 2002); French study, with 42 families and 88 pairs, for BMI (p=0.09) in this region (Clement et al. 1996); French study, with 158 families and 514 individuals, high score (~ 3) for LEP and a lower score for BMI close to this region (Hager et al. 1998).

The BMI score on ch 18, from pter-40, was high and consistent, but split on the peak, which should be resolved before further analyses. For one model, LEP mapped here as well. This locus was found for similar traits in the following studies: Quebec Family Study, for abdominal subcutaneous fat (~ 1.2) in this region (Perusse et al. 2001); Quebec Family Study, 124 families with 289 pairs, for BMI, skinfolds, fat mass, and % body fat (p=0.02-0.001) in this region (Chagnon et al. 1997); and for the overlap with the HT/DYSL factor, the Rochester Family Heart Study, with 232 families and1484 individuals, for APOA-II (~ 1.5) and APOC-II (~ 1.6) in this region (Klos et al. 2001). One candidate gene near this region is the MC5 receptor. The ch 20 signal, from 85-qter, was low but consistent, mapped directly on the marker, and overlapped with a low score for HIP, so it is a possibility. This locus was found for similar traits in the following studies: University of Pennsylvania Family Study, with 92 plus an additional 32 families, for % body fat (~ 2.5) in this region (Lee et al. 1999); Quebec Family Study with 152 pedigrees and 650 individuals, for % body fat, BMI, INS, skinfolds, and fat mass (p=0.02-0.0005) close to this region (Lembertas et al. 1997); French study with 428 sibpairs, for BMI and skinfolds (p=0.02-0.001) close to this region (Comuzzie and Allison 1998).

The HIP signal on ch 10, from 125-140, was tight on the marker, with consistent midsize (~40) scores with the hybrid sampler, and therefore appears to be quite reliable.

This locus was found in the French study for LEP (~ 2.5) in this region (Hager et al. 1998). The ch 4 signal, from 90-114, was on the marker but low, and there were no other linkage studies in the literature that found a similar locus, so this is a possibility.

The WT signal on ch 18, from 75-105, was low, but consistent and on a marker. This was found in the following other studies: Quebec Family Study, for abdominal subcutaneous fat (~ 1.2), close to this region (Perusse et al. 2001); Finnish study, with 100 plus an additional 93 sibpairs, for obesity in this region (Ohman et al. 2000); Icelandic study, with 116 families and 884 individuals, for Peripheral Arterial Occlusive Disease (PAOD) (~ 1.8) in this region (Gudmendsson et al. 2002). This is close to an obvious candidate gene, the MC4 receptor, and therefore is a possibility.

Diabetes

The FBS score on ch 1, from 200-270, is low, but it is a region that has been found in many other studies, including the following: Pima Indians in the Gila River Indian Community, with 109 families and 363 individuals, for OGTT (~ 1.4) in this region (Pratley et al. 1998); Finnish Study in Botnia, with 58 families and 223 individuals, for NIDDM (~ 1.0) in this region (Lindgren et al. 2002); Finland – United States Investigation of NIDDM Genetics (FUSION), with 580 families, for INS (~ 1.4) and 2 hour INS (~ 2.6) in this region (Watanabe et al. 2000); French Caucasians, with 143 families and 677 pairs, for NIDDM (~ 1.6-3.6) in this region (Vionnet et al. 2000). As this is not localizing well in these analyses, further studies should be done to model this trait (or diabetes) differently to get better localization before fine mapping.

Hypertension

The SBP signal on ch 20, from 25-40, was tight, consistent, and on the marker, so is robust. This was found in these studies as well: GENOA Network of the Family Blood Pressure Program in Rochester, MN, with 69 discordant sibpairs, for SBP (p=0.024) close to this region (Krushkal et al. 1999); San Antonio Family Heart Study, with 10 families and 440 indviduals, for pulse pressure (~ 1.6) close to this region (Atwood et al. 2001b); Framingham Heart Study, with 332 families and 1702 individuals, for TG and TG/HDL-C ratio (~ 1.6) at the beginning of the region (Shearman et al. 2000). The ch 1 signal, from 100-160, was poorly localized, and therefore needs to be modeled better prior to further analyses. Comparison to other studies is discussed in overlap section below.

The MAP signal for ch 9, from pter-45, is similarly difficult to localize, so should not be followed up as is, though DBP also slightly mapped to this region. This region was identified as influencing blood pressure in an Old Order Amish Study, with 28 families and 694 individuals, for DBP (~ 1.5) and SBP (~ 0.8) (Hsueh et al. 2000).

Dyslipidemia

The APOB signal on ch 2, from 18-45 is tight, but it mapped between two markers, and decreased in the multi-chromosome analyses. But this is a possibility, particularly as it is close to the APOB gene. This locus was found in the following studies: Dutch Family Study, with 35 families and 253 individuals, for Familial Combined Hyperlipidemia (FCHL) (~ 2.6) at beginning of the region (Aouizerat et al.

1999); Pima Indians, with 188 families and 544 individuals, for TG (~1.7) in this region (Imperatore et al. 2000).

The TC signal on ch 12, from 140-170, is tight, consistent, and close to a marker, and quite reliable. As it only came up with the less-corrected model, and it overlapped with the DYSL factor, perhaps this is a locus that influences more than one of these correlated traits or some sort of combinatorial trait. This locus was found in the Rochester Family Heart Study, for APOA-I (~ 2.0), close to this region (Klos et al. 2001). The ch 16 signal, from 125-qter, overlapped with signals for DYSL factor as well, but mapped off end of the chromosome. As previously suggested, the locus may actually be somewhere else on the chromosome, but there is some problem preventing correct localization, and, therefore, this is less reliable. No other studies found this locus. The third signal for TC, on ch 19, from 37-51, is tight and consistent, so though it maps between two markers, it appears to be quite reliable. This was also found in the following studies: Rochester Family Heart Study, for TC (~ 1.1) close to this region (Klos et al. 2001); Pima Indians in the Gila River Indian Community, with 550 families and around 1500 sibpairs, for TC (~ 4.0), which peaks at the pter but extends to whole chromosome (Imperatore et al. 2000); San Antonio Family Heart Study, with 10 families and 470 individuals, for LDL (~ 2.3) in this region, and LDL2 (~ 3.0) close to this region (Rainwater et al. 1999); Washington University Medical Center Study of Familial Hypobetalipoproteinemia, with 1 family and 38 individuals, for TC (~ 1.7) close to this region (Yuan et al. 2000).

Height

The best signal in the whole dataset is for HT on ch 10, from 81-94, as it is high, consistent, and tight, mapping directly on the marker. There are only three other genome scans published for HT, so much replication is not expected. Two of the scans actually use data collected for disease studies, which have data for HT, but were previously not analyzed (Hirschhorn et al. 2001; Wiltshire et al. 2002). One of these, which used data from the Diabetes UK Warren 2 Consortium, with 573 families and 1377 sibpairs. had a HT score (~ 1.9) which peaks close to this region (Wiltshire et al. 2002). The next nice score was on ch 1, from 100-120, which has a wider range because the two models peaked slightly apart. The second study that combined data from four other genome scans, including 483 families and 2327 individuals, had a score in this region from the Sweden population (LOD ~ 1) (Hirschhorn et al. 2001). The ch 5 signal, from 160-qter, is consistent but is split on the peak, and this should be resolved, if possible, before further studies. None of the other scans found this locus. The ch 19 signal, from 33-55, is less reliable as the multi-chromosome scores were reduced and it is also a split peak, though this was also found in the Diabetes UK Warren 2 Consortium, with a suggestive score (~ 1.6) in this region (Wiltshire et al. 2002). The final signal for HT, on ch 15, from 113-qter, was only found with one model, and poorly mapped to the end of the chromosome, and is less reliable. It was found in the Finland population (~ 1.3) (Hirschhorn et al. 2001).

Overlap regions

Table 6.5 shows which other studies found linkage in the same regions as there was overlap between a number of small scores for correlated traits. This is interesting, but there is not much that can be done without a good trait to map further. One idea is to try to work out some sort of combinatorial phenotype for the traits that map to this region. Another possibility is that perhaps a signal due to an unmeasured correlated trait, such as the ones found in the similar studies listed.

On ch 1 there are a series of adjacent areas that are interesting. First, from 90-160 there are signals for hypertension, which was also found in the following studies: Quebec Family Study, with about 180 families and 679 individuals, for SBP (~ 1.1-1.8) in this region (Rice et al. 2000); Icelandic study, for PAOD (~ 2.8) in this region (Gudmendsson et al. 2002); Hypertension Genetic Epidemiology Network (HyperGEN) of NHLBI in Caucasians, with 480 sibpairs, for renal function in hypertensives (~ 1.5-1.9) in this region (DeWan et al. 2001). Then there are signals for diabetes, obesity, and BP again in the 170-212 range. Here a LOD score of close to 2.9 for BMI was seen from the SOLAR analysis. This region was seen in these studies: Diabetes United Kingdom Warren 2 Repository, with 573 families, for NIDDM (~ 1.5) in this region (Wiltshire et al. 2001); FUSION study, for insulin resistance (~ 1.5) and FBS (~ 1.7) close to this region (Watanabe et al. 2000); Caucasians, large pedigree study, for BMI ($p=4.7 \times 10^{-5}$) in this region (Comuzzie and Allison 1998);University of Pennsylvania Family Study, for % body fat in this region (Lee et al. 1999); Finnish study, for obesity (~ 1.4) in this region (Ohman et al. 2000); Pima Indians, with 127 families, for waist-to-thigh ratio (~ 1.2), 24 hour ratio of carbohydrate oxidation to fat oxidation (~ 2.0), 24 hour energy expenditure

(~ 1.0), sleeping metabolic rate (~ 1.6), all over this region (Norman et al. 1998); Utah Caucasians, with 19 families and 468 individuals, for NIDDM (~2.3) close to this region (Elbein et al. 1999); Pima Indians, with 332 families and about 1700 pairs, for NIDDM (~ 2.4) in this region (Hanson et al. 1998); Quebec Family Study, for abdominal subcutaneous fat (~ 2.0), all over the chromosome (Perusse et al. 2001); NHLBI Family Heart Study, for BMI in this region (Feitosa et al. 2002); studies from FBS region in Table 6.4 with signal in this region as well: French Caucasians, ~ 2.5 (Vionnet et al. 2000) and Botnia Study (Lindgren et al. 2002).

On ch 2 there were two regions where hypertension traits mapped to, 20-50 and then to 75, which also had signal for dyslipidemia, as well as 90-130. The following studies had signals in the first region: Finnish study, with 181 families and 345 individuals, for coronary heart disease (CHD, ~1), in this region (Pajukanta et al. 2000); GENOA Network of the Family Blood Pressure Program, for SBP (p=0.0089) close to this region (Krushkal et al. 1999); Old Order Amish Study, for SBP (~ 1.0) close to this region (Hsueh et al. 2000); Dutch Family Study, for FCHL (~ 1.3) close to this region (Aouizerat et al. 1999); Quebec Family Study, for DBP (~ 1.1) in this region (Rice et al. 2000); Framingham Heart Study, with 332 families and 1702 individuals, for DBP in this region (Levy et al. 2000). These studies gave signal in the second region (some, as seen, gave signals in both): San Antonio Family Heart Study, with 10 families and 495 individuals, for DBP (~ 3.9) and SBP (~ 1.3) in this region (Atwood et al. 2001a); San Antonio Family Heart Study, for PP (~ 1.3) in this region (Atwood et al. 2001b); Old Order Amish Study, for DBP (~ 1.0) close to this region (Hsueh et al. 2000); HyperGEN of NHLBI, with 285 families and 636 individuals, for change in SBP (~ 1.5) close to this region (Pankow et al. 2000); Quebec Family Study, for DBP (~ 1.0-2.2) and SBP (~ 1.3) in this region, (Rice et al. 2000).

On ch 7 there were 3 semi-continuous regions with overlapping scores. The first, from 50-90, had signals for obesity and diabetes, and this was found in the following studies: Quebec Family Study, for abdominal subcutaneous fat (~ 1.6) in this region (Perusse et al. 2001); Genetics of NIDDM (GENNID) Study of 4 American populations (Mexican American, Caucasian, African American, and Japanese American) for NIDDM or impaired insulin homeostasis (~1.0) in this region for the Mexican American group with 53 families and 365 individuals (Ehm et al. 2000); NHLBI Family Heart Study, for BMI (~1.0) in this region (Feitosa et al. 2002); Old Order Amish Study, with 28 families and 672 individuals, for LEP (~ 1.3) in this region (Hsueh et al. 2001). The next region, from110-140 (end of this range maps close to the LEP gene), had signal for dyslipidemia and obesity, which was also found in these studies: NHLBI Family Heart Study, for BMI past the end of this region (Feitosa et al. 2002); Quebec Family Study, abdominal subcutaneous fat after end of this region (Perusse et al. 2001); San Antonio Family Diabetes Study, with 32 families and 579 individuals, for HDL-C (~ 1.8) at end of chromosome (Duggirala et al. 2000). The last region is from 150-qter, with scores for dyslipidemia, obesity, and diabetes, as found in these studies: San Antonio Family Diabetes Study, for TG (~ 1.9) at end of chromosome (Duggirala et al. 2000); Framingham Heart Study, for TG (~ 1.8) and TG/HDL-C ratio (~ 2.5) in this region (Shearman et al. 2000); Pima Indians for FBS (~ 1.2) and various insulin measures (~ 1.5-1.8) in this region (Pratley et al. 1998).

On ch 9, the region from 105 to 135 had signals for obesity and diabetes, as seen in these studies: Quebec Family Study, for abdominal subcutaneous fat (~ 2.3) at beginning of this region (Perusse et al. 2001); GENNID Study of 4 American populations, for NIDDM (~ 1.5) in this region, in the Caucasian American group, with 77 families and 497 individuals (Ehm et al. 2000); FUSION study, for insulin resistance (~ 2.3) in this region (Watanabe et al. 2000).

Ch 13 showed a region from 20-60 with signals for obesity, as seen in these studies: Quebec Family Study, for abdominal subcutaneous fat (~ 1.6) in this region (Perusse et al. 2001); NHLBI Family Heart Study, for BMI in this region (Feitosa et al. 2002); French study with 428 sibpairs, for % body fat and skinfolds ($p<0.04$), in this region (Comuzzie and Allison 1998); FUSION study, for BMI (~3.0), C-Peptide (~ 1.2), and INS measures (~ 1.2-3.0) close to this region (Watanabe et al. 2000).

Ch 16 had signal for diabetes and obesity from 100 to 130, which was seen in these following studies: Pima Indians, for INS and FBS (~ 1.3) in this region (Pratley et al. 1998); GENNID Study of 4 American populations, for NIDDM (~ 1.9) in this region, in the African American group, with 65families and 229 individuals (Ehm et al. 2000); Finnish Study in Botnia, for NIDDM (~ 2.0) in this region (Lindgren et al. 2002).

Future plans

Future directions for the study are dual as well, with suggestions to improve Loki, and to refine the linkage signals found. The main improvement to Loki would be more efficient mixing, both for general analysis and to be able to use markers closer together. This is a major issue and there is no simple answer. Additionally, a series of simulations to determine the empirical distribution of L-score under the null of no linkage for significance levels needs to be done. These simulation studies will also show how Loki behaves with different analysis variables. With the hybrid sampler, the L-scores appear more stable, so the L-scores and associated p-values will have more meaning. Often there are genotype errors that are unobserved and therefore left in. This is suggested to negatively affect all methods of LA (Sobel et al. 2002). As there are definitely unobserved errors, it would be interesting to see how this would affect Loki analysis. This could be done by introducing errors into the GAW12 dataset and reanalyzing. As a way to try to identify errors, it would be useful for Loki to report obligate recombinants. Another idea is to see how Loki handles locus heterogeneity and how one could check for it. Lastly, it has been shown that there are important gene-gene and gene-environment interactions that affect the traits under study, and it would be informative to be able to model these using Loki.

To further the Kosrae study, the first thing to do is complete the genome scans for all the pedigree data. It is possible that new QTL will be found, but as even the most borderline scores were reanalyzed with the new data, it is not likely. Next a scan for diabetes as a qualitative trait should be done. Furthermore, Loki is unable to analyze the X chromosome so other methods with smaller pedigrees or fewer markers will be tried. As this has less power, if no signal is found that does not necessarily mean nothing is there, but a stronger signal might be picked up. Lastly some of the data will be re-analyzed with other programs to see how Loki scores compare to others and to do some modeling not possible in Loki (such as interactions).

Then, the main focus would be to go from wide linkage regions to narrower regions, to eventually identify the genes involved. For this fine mapping, the first step would be to type denser markers (one every 1 cM) to see if that helps to narrow the region. The denser map, the more important it is to make sure the genotype data is error free, as the more significant each recombinant is. LA though may not be able to achieve high resolution with quantitative traits, as there is no way to define absolute recombinants between a locus and affection status as could be done for qualitative traits. Therefore linkage disequilibrium analysis could be performed, as it uses ancestral recombinants (as opposed to familial only) for finer resolution. In this case, markers as dense as possible, perhaps SNPs, should be typed. The challenges here include how to generate haplotypes for the large complex pedigree and then the best way to analyze these.

In summary, this study developed a strategy of how to use Loki to find genes that affect the levels of quantitative traits. This includes initial single chromosome scans, using the Kosambi map setting, setting the number of QTL to 5, and using the hybrid mixer 5% of the iterations. Then more stringent analyses should be carried out with model corrections, joint analysis of positive chromosomes, repeat runs with significant scores, and inspection of L-graphs. Additionally, this method is more powerful with genotype information on more of the individuals in the pedigree and with denser marker coverage. There have been a few other MCMC methods suggested and the main differences between these and Loki are the acceptance of new QTL, and generating genotypes for these QTL (Thomas et al. 1997; Stephens and Fisch 1998; Lee and Thomas 2000; Uimari and Sillanpaa 2001). As these are mostly for nuclear families, they still cannot perform multipoint analysis in large extended families as Loki can.

Through this, we identified possible QTLs for many of the traits involved in Syndrome X pathogenesis. As a prioritized list, these regions should be followed up: HT on ch 10 and ch 1, LEP on ch 5, TC on ch 19, SBP on ch 20, HIP on ch 10, and BMI on ch 18 (if split peak can be resolved). Then, APOB on ch 2, TC on ch 12, and HT on ch 19 and ch 5 (again, if both split peaks can be resolved) could be investigated. Slight possibilities are WT on ch 18, HIP on ch 4, BMI on ch 20, and maybe HT on ch 15. The signals for FBS on ch 1, SBP on ch 1, MAP on ch 9, and TC on ch 16 cannot be mapped further until there is better localization.

Conclusion

The linkage analysis program Loki uses reversible jump Markov chain Monte Carlo sampling to analyze genetic data. As this is a sampling method, Loki can perform multipoint analysis on large extended pedigrees. This is crucial to have power to map genes that affect susceptibility to multifactorial diseases, by analyzing the underlying quantitative traits. Loki does this well by modeling quantitative traits with covariate corrections and identifying quantitative trait loci that explain a proportion of the residual variance. When Loki is used in the manner described above, it is an excellent method of preliminary linkage analysis to elucidate regions deserving further investigation. This hierarchal strategy includes initial single chromosome scans, additional scans with model changes, multi-chromosome analysis, repeat runs, and inspection of L-score graphs.

Loki was used to map genes that affect susceptibility to obesity, diabetes, hypertension, and dyslipidemia, by analyzing the quantitative traits associated with these traits (LEP, BMI, HIP, WT, WST, FBS, INS, SBP, MAP, DBP, TC, TG, APOA-1, and APOB), as well as HT, in the population on the island of Kosrae, the Federated States of Micronesia. This is an extremely powerful dataset as it includes phenotype and covariate information on greater than 2200 individuals, with one large pedigree including 1564 individuals that were genotyped for a 10 cm genome scan and analyzed using Loki. This resulted in seven very good scores: HT on ch 10 (80 cM – 95 cM) and ch 1 (100-120), LEP on ch 5 (170-200), TC on ch 19 (35-50), SBP on ch 20 (25-40), HIP on ch 10 (125-140), and BMI on ch 18 (pter-40); four good signals: APOB on ch 2 (20-45), TC on ch 12 (140-170), and HT on ch 19 (30-55) and ch 5 (160-qter); and four possibilities: WT on ch

18 (75-105), HIP on ch 4 (90-115), BMI on ch 20 (85-qter), and HT on ch 15 (110-qter). These are promising regions that merit further study to try to identify the causal genes for these traits and their effect on disease pathogenesis.

References

Allison, D., J. Kaprio, M. Korkeila, M. Koskenvuo, M. Neale and K. Kayakawa (1996). The heritability of body mass index among an international sample of monozygotic twins reared apart. Int. J. Obes. Relat. Metab. Disord. **20**: 501-6.

Allison, D. B., B. S. Gorman and E. M. Kucera (1995). Unicorn: A program for transforming data to approximate normality. Educational & Psychological Measurement **55**: 625-9.

Almasy, L. and J. Blangero (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62**: 1198-21.

Almasy, L., J. D. Terwilliger, D. Nielsen, T. D. Dyer, D. Zaykin and J. Blangero (2001). GAW12: simulated genome scan, sequence, and family data for a common disease. Genet Epidemiol **21**: S332-8.

Altmuller, J., L. J. Palmer, G. Fischer, H. Scherb and M. Wjst (2001). Genomewide scans of complex humann diseases: true linkage is hard to find. Am J Hum Genet **69**: 936-50.

Amos, C. I., R. C. Elston, S. R. Srinivasan, A. F. Wilson, J. L. Cresanta, L. J. Ward and G. S. Berenson (1987). Linkage and segregation analyses of apolipoproteins A1 and B, and lipoprotein cholesterol levels in a large pedigree with excess coronary heart disease: the Bogalusa Heart Study. Genet. Epidemiol. **4**: 115-28.

Anonymous (1980). Lipid Research Clinics Program: The Prevalence Study. US Department of Health and Human Services, National Institutes of Health.

Anonymous (1989). Report of the Expert Panel on the detection, evaluation, and treatment of high blood cholesterol in adults. NIH Publication No. 89-2925.

Anonymous (1993). The Fifth Report of the joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. Arch. Intern. Med. **153**: 154-83.

Anonymous (1994). National Cholesterol Education Program. Second Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel). Circulation **89**: 1333-445.

Anonymous (1997). Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care **20**: 1183-97.

Aouizerat, B. E., H. Allayee, R. M. Cantor, R. C. Davis, C. D. Lanning, P. Wen, G. M. Dallinga-Thie, T. W. A. de Bruin and et. al. (1999). A genome scan for familial combined hyperlipidemia reveals evidence of linkage with a locus on chromosome 11. Am J Hum Genet **65**: 397-412.

Arya, H., R. Duggirala, J. T. Williams, L. Almasy and J. Blangero (2001). Power to localize the major gene for disease liability is increased after accounting for the effects of related quantitative phenotypes. Genet Epidemiol **21**: S774-S8.

Atwood, L. D., P. B. Samollow, J. E. Hixson, M. P. Stern and J. W. MacCluer (2001a). Genome-wide linkage analysis of blood pressure in Mexican Americans. Genet Epidemiol **20**: 373-82.

Atwood, L. D., P. B. Samollow, J. E. Hixson, M. P. Stern and J. W. MacCluer (2001b). Genome-wide linkage analysis of pulse pressure in Mexican Americans. Hypertension **37**: 425-8.

Bachorik, P. S., K. L. Lovejoy, M. D. Carroll and C. L. Johnson (1997). Apolipoprotein B and AI distributions in the United States, 1988-1991: results of the National Health and Nutrition Examination Survey III (NHANES III). Clin. Chem. **43**: 2364-78.

Ballinger, S. W., J. M. Shoffner, E. V. Hedaya, I. Trounce, M. A. Polak, D. A. Koontz and D. C. Wallace (1992). Maternally transmitted diabetes and deafness associated with a 10.4 kb mitochondrial DNA deletion. Nature Genetics **1**: 11.

Barrett-Conner, E. and K. Khaw (1984). Family history of heart attack as an independent predictor of death due to cardiovascular disease. Circulation **4**: 793-801.

Biron, P., J. Mongeau and D. Bertrand (1977). Familial resemblance of body weight and weight/height in 374 homes with adopted children. J. Pediatr. **91**: 555-8.

Blangero, J. and L. Almasy (1997). Multipoint oligogenic linkage analysis of quantitative traits. Genet. Epidemiol. **14**: 959-64.

Blangero, J., J. T. Williams and L. Almasy (2001). Variance component methods for detecting complex trait loci. Adv Genet **42**: 151-81.

Blangero, J., S. Williams-Blangero and M. C. Mahaney (1993). Multivariate genetic analysis of apo AI concentration and HDL subfractions: evidence for major locus pleiotropy. Genet. Epidemiol. **10**: 617-22.

Bouchard, C. (1997). Genetics of human obesity: recent results from linkage studies. J. Nutr. **127**: 1887s-90s.

Brenn, T. (1994). Genetic and environmental effects on coronary heart disease risk factors in northern Norway. The cardiovascular disease study in Finnmark. Ann. Hum. Genet. **58**: 369-79.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White and J. L. Weber (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet **63**: 861-9.

Burt, V., P. Whelton, E. Roccella, C. Brown, J. Cutler, M. Higgins, M. Horan and D. Labarthe (1995). Prevalence of hypertension in the US adult population. Results from the third national health and nutrition examination survey, 1988-1991. Hypertension **25**: 305-13.

Cannings, C., E. A. Thompson and M. H. Skolnick (1978). Probability functions on complex pedigrees. Adv Appl Prob **10**: 26-61.

Castelli, W., R. Garrison, P. Wilson, R. Abbott, S. Kalusdian and W. Kannel (1986). Incidence of coronary heart disease and lipoprotein cholesterol levels. the Framingham Study. JAMA(256): 2835-8.

Chagnon, Y. C., W. J. Chen, L. Perusse, M. Chagnon, A. Nadeau, W. O. Wilkison and C. Bouchard (1997). Linkage and association studies between the melanocortin receptors 4 and 5 genes and obesity-related phenotypes in the Quebec Family Study. Mol Med **3**: 663-73.

Chakraborty, R. and K. M. Weiss (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. U S A **85**: 9119-23.

Chen, H., O. Charlat, L. A. Tartaglia, E. A. Woolf, X. Weng, S. J. Ellis, N. D. Lakey, J. Culpepper, K. J. Moore, R. E. Breitbart, G. M. Duyk, R. I. Tepper and J. P Morgenstern (1996). Evidence that the diabetes gene encodes the leptin receptor: Identification of a mutation in the leptin receptor gene in *db/db* mice. Cell **84**: 491-5.

Cheng, L. S., D. Carmelli, S. C. Hunt and R. R. Williams (1995). Evidence for a major gene influencing 7-year increases in diastolic blood pressure with age. Am. J. Hum. Genet. **57**: 1169-77.

Cheng, L. S., G. Livshits, D. Carmelli, J. Wahrendorf and D. Brunner (1998). Segregation analysis reveals a major gene effect controlling systolic blood pressure and BMI in an Israeli population. Hum. Biol. **70**: 59-75.

Clement, K., A. Philippi, C. Jury, R. Pividal, J. Hager, F. Demenais, A. Basdevant, B. Guy-Grand and e. al. (1996). Candidate gene approach of familial morbid obesity: linkage analysis of the glucocorticoid receptor gene. Int J Obes Relat Metab Disord **20**: 507-12.

Colditz, G., E. Rimm, E. Giovannucci, M. Stampfer, B. Rosner and W. Willett (1991). A prospective study of parental history of myocardial infarction and coronary artery disease in men. Am. Journal of Cardiology **67**: 933-8.

Comuzzie, A. G. and D. B. Allison (1998). The search for human obesity genes. Science **280**: 1374-7.

Comuzzie, A. G., M. C. Mahaney, L. Almasy, T. D. Dyer and J. Blangero (1997). Exploiting pleiotropy to map genes for oligogenic phenotypes using extended pedigree data. Genet Epidemiol **14**: 975-80.

Contois, J., J. R. McNamara, C. Lammi-Keefe, P. W. Wilson, T. Massov and E. J. Schaefer (1996a). Reference intervals for plasma apolipoprotein A-1 determined with a standardized commercial immunoturbidimetric assay: results from the Framingham Offspring Study. Clin. Chem. **42**: 507-14.

Contois, J. H., J. R. McNamara, C. J. Lammi-Keefe, P. W. Wilson, T. Massov and E. J. Schaefer (1996b). Reference intervals for plasma apolipoprotein B determined with a standardized commercial immunoturbidimetric assay: results from the Framingham Offspring Study. Clin. Chem. **42**: 515-23.

Cordy, R. (1993). The Lelu Stone ruins (Kosrae, Micronesia) 1978-1981. Historical and Archaeological Research. Asian and Pacific Archaeology Series. University of Hawaii Press **10**.

Cusi, D., E. Fossali, A. Piazza, G. Tripodi, C. Barlassina, E. Pozzoli, G. Vezzoli, P. Stella, L. Soldati and G. Bianchi (1991). Heritability estimate of erythrocyte Na-K-Cl cotransport in normotensive and hypertensive families. Am. J. Hypertens. 1991 **4**: 725-34.

Dallinga-Thie, G. M., M. van Linde-Sibenius Trip, J. I. Rotter, R. M. Cantor, X. Bu, A. J. Lusis and T. W. de Bruin (1997). Complex genetic contribution of the Apo AI-CIII-AIV gene cluster to familial combined hyperlipidemia. Identification of different susceptibility haplotypes. J. Clin. Invest. 1997 **99**: 953-61.

Daw, E. W., S. C. Heath and E. M. Wijsman (1997). Multipoint oligogenic analysis of age of onset data with applications to large Alzheimer's disease pedigrees. Am J Hum Genet **61**: A273.

Daw, E. W., E. A. Thompson and E. Wijsman (2000). Bias in multipoint linkage analysis arising from map misspecification. Genet Epidemiol **19**: 366-80.

Defronzo, R. A. and E. Ferrannini (1991). Insulin resistance: a multifaceted syndrome responsible for NIDDM, obesity, dyslipidemia, and atherosclerotic cardiovascular disease. Diabetes Care **14**: 173-94.

Deurenberg, P., M. Yap and W. van Staveren (1998). Body mass index and percent body fat: a meta analysis among different ethnic groups. International J of Obesity **22**: 1164-71.

DeWan, A. T., D. K. Arnett, L. D. Atwood, M. A. Province, C. E. Lewis, S. C. Hunt and J. Eckfeldt (2001). A genome scan for renal function among hypertensives: The HyperGEN Study. Am J Hum Genet **68**: 136-44.

Duggirala, R., J. Blangero, L. Almasy, T. D. Dyer, K. L. Williams, R. J. Leach, P. O'Connell and M. P. Stern (2000). A major susceptibility locus influencing plasma triglyceride concentrations is located on chromosome 15q in Mexican Americans. Am J Hum Genet **66**: 1237-45.

Durnin, J. and J. Womersley (1974). Body fat assessed from total body density and its estimation from skinfold thickness: measurement on 481 men and women aged from 16 to 72 years. Br. J. Nutr. **32**: 77-97.

Edwards, K., M. Austin, B. Newman, E. Mayer, R. Krauss and J. Selby (1994). Multivariate analysis of the insulin resistance syndrome in women. Arterioscler Thromb. **14**: 1940-5.

Edwards, K., C. Burchfiel, D. Sharp, J. Curb, B. Rodriquez, W. Fujimoto, A. LaCroix, M. Vitiello and M. Austin (1998). Factors of the insulin resistance syndrome in nondiabetic and diabetic elderly Japanese-American men. Am J Epidemiol **147**: 441-7.

Edwards, K. L., B. Newman, E. Mayer, J. V. Selby, R. M. Krauss and M. A. Austin (1997). Heritability of factors of the insulin resistance syndrome in women twins. Genet. Epidemiol. **14**: 241-53.

Ehm, M. G., M. C. Karnoub, H. Sakul, K. Gottschalk, D. C. Holt, J. L. Weber, D. Vaske, D. Briley and e. al. (2000). Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet **66**: 1871-81.

Elbein, S., M. Hoffman, K. Teng, M. Leppert and S. Hasstedt (1999a). A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. Diabetes **48**: 1175-82.

Elbein, S. C., S. J. Hasstedt, K. Wegner and S. E. Kahn (1999b). Heritability of pancreatic beta-cell function among nondiabetic members of
Caucasian familial type 2 diabetic kindreds. J. Clin. Endocrinol. Metab. **84**: 1398-403.

Elston, R. C. (1998). Linkage and association. Genet. Epidemiol. **15**: 565-76.

Elston, R. C. and J. Stewart (1971). A general model for the genetic analysis of pedigree data. Hum Hered **21**: 523-42.

Feitosa, M. F., I. B. Borecki, S. S. Rich, D. K. Arnett, P. Sholinsky, R. H. Myers, M. Leppert and M. Province (2002). Quantitative-trait loci influencing body-mass index reside on chromosomes 7 and 13: The National Heart, Lung, and Blood Institute Family Heart Study. Am. J. Hum. Genet. **70**: 72-82.

Ferrannini, E., S. Haffner, B. Mitchell and M. Stern (1991). Hyperinsulinaemia: the key feature of a coardiovascular and metabolic syndrome. Diabetologia **34**: 416-22.

Flegal, K. M., M. D. Carroll, R. J. Kuczmarski and C. L. Johnson (1998). Overweight and obesity in the United States: prevalence and trends, 1960-1994. Int J Obes Relat Metab Disord **22**: 39-47.

Gauderman, W. J. and C. L. Faucett (1997). Detection of gene-environment interactions in joint segregation and linkage analysis. Am. J. Hum. Genet. **61**: 1189-99.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Pattn Anal Mach Intell **6**: 721-41.

Geyer, C. J. and E. A. Thompson (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. J Am Stat Assoc **90**: 909-20.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996). Introducing Markov chain Monte Carlo. Markov Chain Monte Carlo in Practice. D. J. Spiegelhalter. London, Chapman and Hall: 1-19.

Gray, R. S., R. R. Fabsitz, L. D. Cowan, E. T. Lee, B. V. Howard and P. J. Savage (1998). Risk factor clustering in the Insulin Resistance Syndrome: The Strong Heart Study. Am J Epidemiol **148**: 869-78.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**: 711-32.

Gu, C., I. Borecki, J. Gagnon, C. Bouchard, A. Leon, J. Sinner, J. Wilmore and D. Rao (1998). Familial resemblance for resting blood pressure with particular reference to racial differences: Preliminary analyses from the HERITAGE Family Study. Human Biology **70**: 77-90.

Gudmendsson, G., S. E. Matthiasson, H. Arason, H. Johannsson, F. Runarsson, R. Bjarnason, K. Helgadottir, S. Thorisdottir and e. al. (2002). Localization of a gene for peripheral arterial occlusive disease to chromosome 1p31. Am J Hum Genet **70**: 586-92.

Guo, S.-W. (2000). Genetic mapping of complex traits: promises, problems, and prospects. Theoretical Population Biology **57**: 1-11.

Haffner, S., R. Valdez, H. Hazuda, B. Mitchell, P. Morales and M. Stern (1992). Prospective analysis of the insulin-resistance syndrome (syndrome X). Diabetes **41**: 715-22.

Hager, J., C. Dina, S. Francke, S. Dubois, M. Houari, V. Vatin, E. Vaillant, N. Lorentz and e. al. (1998). A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. Nat. Genet. **20**: 304-8.

Hamann, R. (1992). Genetic and environmental determinants of non-insulin-dependent diabetes (NIDDM). Diabetes Metab. Rev. **8**: 287-338.

Hanson, R. L., M. G. Ehm, D. J. Pettitt, M. Prochazka, D. B. Thompson, D. Timberlake, T. Foroud, S. Kobes and e. al. (1998). An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. Am J Hum Genet **63**: 1130-8.

Hanson, R. L., R. C. Elston, D. J. Pettitt, P. H. Bennett and W. C. Knowler (1995). Segregation analysis of non-insulin-dependent diabetes mellitus in Pima Indians: evidence for a major-gene effect. Am. J. Hum. Genet. **57**: 160-70.

Hansson, J., C. Nelson-Williams, H. Suzuki, L. Schild, R. Shimkets, Y. Lu, C. Canessa, T. Iwasaki, B. Rossier and R. Lifton (1995). Hypertension caused by a truncated epithelial sodium channel γ subunit: genetic heterogeneity of Liddle syndorme. Nature Genetics **11**: 76-82.

Harris, M. (1995). Summary in Diabetes in America. NIH Publication **No.95 1468**.

Harris, M. I., K. M. Flegal, C. C. Cowie, M. S. Eberhardt, D. E. Goldstein, R. R. Little, H.-M. Wiedmeyer and D. D. Byrd-Holt (1998). Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. Diabetes Care **21**: 518-24.

Hartling, S. G., B. Dinesen, A. M. Kappelgard, O. K. Faber and C. Binder (1985). ELISA for human proinsulin. Clin Chim Acta **156**: 289-98.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**: 97-109.

Heath, S. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am.J. Human Genet. **61**: 748-60.

Heath, S., G. Snow, E. Thompson, C. Tseng and E. Wijsman (1997). MCMC Segregation and linkage analysis. Proceedings of Genetic Analysis Workshop 10. Genetic Epidemiology **14**: 1011-6.

Heath, S. C. (2002). Genetic linkage analysis using Markov chain Monte Carlo techniques. in press.

Hezel, F. (1983). The First Taint of Civilization: A History of the Caroline and Marshall Islands in Pre-Colonial Days, 1521-1885. University of Hawaii Press.

Higgins, M., M. Province, G. Heiss, J. Eckfeldt, R. C. Ellison, A. R. Folsom, D. C. Rao, J. M. Sprafka and R. Williams (1996). NHLBI Family Heart Study: objectives and design. Am J Epidemiol **143**(12): 1219-28.

Hirschhorn, J. N., C. M. Lindgren, M. J. Daly, A. Kirby, S. F. Schaffner, N. P. Burtt, D. Altshuler, A. Parker and e. al. (2001). Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. Am J Hum Genet **69**: 106-16.

Hodge, A. M., G. K. Dowse and P. Z. Zimmet (1993). Association of body mass index and waist-hip circumference ratio with cardiovascular disease risk factors in Micronesian Naruans. Int J Obes Relat Metab Disord **17**: 399-407.

Hong, Y., U. de Faire, D. Heller, G. McClearn and N. Pedersen (1994). Genetic and environmental influences on blood pressure in elderly twins. Hypertension **24**: 663-70.

Hopkins, P., R. Williams, H. Kuida, B. Stults, S. Hunt, G. Barlow and K. Ash (1988). Family history as an independent risk factor for incident coronary artery disease in a high-risk cohort in Utah. Am. Journal of Cardiology **62**: 703-7.

Horikawa, Y., N. Iwasaki, M. Hara, H. Furutal, Y. Hinokio, B. Cockburn, T. Lindner, K. Yamagata, M. Ogata, O. Tomonaga, H. Kuroki, T. Kasahara, Y. Iwamoto and G. Bell (1997). Mutation in hepatocyte nuclear factor-1-beta gene (TCF2) associated with MODY. Nature Genetics **17**: 384-5.

Hsueh, W., B. D. Mitchell, J. L. Schneider, P. L. St. Jean, T. I. Pollin, M. G. Ehm, M. J. Wagner, D. K. Burns and e. al. (2001). Genome-wide scan of obesity in the Old Order Amish. J Clin Endocrinol Metab **86**: 1199-205.

Hsueh, W., B. D. Mitchell, J. L. Schneider, M. J. Wagner, C. J. Bell, E. Nanthakumar and A. R. Shuldiner (2000). QYL influencing blood pressure maps to the region of PPH1 on chromosome 2q31-34 in Old Order Amish. Circulation **101**: 2810-6.

Imperatore, G., W. C. Knowler, D. J. Pettitt, S. Kobes, J. H. Fuller, P. H. Bennett and R. L. Hanson (2000). A locus influencing total serum cholesterol on chromosome 19p. Arterioscler Thromb Vasc Biol **20**: 2651-6.

Irwin, G. (1994). The prehistoric exploration and colonisation of the Pacific. Cambridge University Press.

Kadowaki, T., H. Kadowaki and Y. Mori (1994). A subtype of diabetes mellitus with a mutation of mitochondrial DNA. New Eng. J. Med. **330**: 962-8.

Klos, K. L., S. L. R. Kardia, R. E. Ferrell, S. T. Turner, E. Boerwinkle and C. F. Sing (2001). Genome-wide linkage analysis reveals evidence of multiple regions that influence variation in plasma lipid and apolipoprotein levels associated with risk of coronary heart disease. Arterioscler Thromb Vasc Biol. **21**: 971-8.

Kong, A. (1991). Analysis of pedigree data using methods combining peeling and Gibbs sampling. Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Fairfax Station, Interface Foundation of North America: 379-85.

Krushkal, J., R. Ferrell, S. C. Mockrin, S. T. Turner, C. F. Sing and E. Boerwinkle (1999). Genome-wide linkage analyses of systolic blood pressure using highly discordant siblings. Circulation **99**: 1407-10.

Lander, E. and P. Green (1987). Construction of multilocus genetic maps in humans. Proc Natl Acad Sci U S A **84**: 2363-7.

Lange, K. and T. M. Goradia (1987). An algorithm for automatic genotype elimination. Am J Hum Genet **40**: 250-6.

Lange, K. and E. Sobel (1991). A random walk method for computing genetic location scores. Am J Hum Genet **49**: 1320-34.

Lange, K. and D. E. Weeks (1989). Efficient computation of lod scores: genotype elimination, genotype redefinition, and hybrid maximium likelihood algorithms. Ann Hum Genet **53**: 67-83.

Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. J R Stat Soc B **50**: 157-224.

Lecomte, E., B. Herbeth, V. Nicaud, R. Rakotovao, Y. Artur and L. Tiret (1997). Segregation analysis of fat mass and fat-free mass with age- and sex-dependent effects: the Stanislas Family Study. Genet. Epidemiol. **14**: 51-62.

Lee, J. H., D. R. Reed, W. D. Li, W. Xu, E. J. Joo, R. L. Kilker, E. Nanthakumar, M. North, H. Sakul, C. Bell and R. A. Price (1999). Genome scan for human obesity and linkage to markers in 20q13. Am. J. Hum. Genet. **64**: 196-209.

Lee, J. K. and D. C. Thomas (2000). Performance of Markov chain-Monte Carlo approaches for mapping genes in oligogenic models with an unknow number of loci. Am J Hum Genet **67**: 1232-50.

Lembertas, A. V., L. Perusse, Y. C. Chagnon, J. S. Fisler, C. H. Warden, D. A. Purcell-Huynh, F. T. Dionne, J. Gagnon and e. al. (1997). Identification of an obesity quantitative trait locus on mouse chromosome 2 and evidence of linkage to body fat and inslin on the human homologous region 20q. J Clin Invest **100**: 115-21.

Levy, D., A. L. DeStefano, M. G. Larson, C. J. O'Donnell, R. P. Lifton, H. Gavras, L. A. Cupples and R. H. Myers (2000). Evidence for a gene influencing blood pressure on chromosome 17. Hypertension **36**: 477-83.

Leyva, F., I. Godsland, M. Ghatei, A. Proudler, S. Aldis, C. Walton, S. Bloom and J. Stevenson (1998). Hyperleptinemia as a component of a metabolic syndrome of cardiovascular risk. Arterioscler Thromb Vasc Biol. **18**: 928-33.

Lifton, R., R. Dluhy, M. Powers, G. Rich, S. Cook, S. Ulick and J. Lalouel (1992). A chimaeric 11 beta-hydoxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. Nature **355**: 262-5.

Lifton, R. P. (1995). Genetics determinants of human hypertension. Proc. Natl. Acad. Sci. USA **92**: 8545-51.

Lin, S. (1995). A scheme for constructing an irreducible Markov chain for pedigree data. Biometrics **51**: 318-22.

Lindgren, C. M., M. M. Mahtani, E. Widen, M. I. McCarthy, M. J. Daly, A. Kirby, M. P Reeve, L. Kruglyak and e. al. (2002). Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: The Botnia Study. Am J Hum Genet **70**: 509-16.

Liu, L., S. R. Choudhury, A. Okayama, T. Hayakawa, Y. Kita and H. Ueshima (1999). Changes in body mass index and its relationships to other cardiovascular risk factors among Japanese population: results from the 1980 and 1990 national cardiovascular surveys in Japan. J Epidemiol **9**: 163-74.

Mayer, E. J., B. Newman, M. A. Austin, D. Zhang, C. P. Quesenberry, K. Edwards and J. V. Selby (1996). Genetic and environmental influences on insulin levels and the insulin resistance syndrome: an analysis of women twins. Am. J. Epidemiol. **143**: 323-32.

McCarthy, D. and P. Zimmet (1994). Diabetes 1994 to 1020: Estimates and projections. International Diabetes Institute, Melbourne Australia.

McKeigue, P. M. (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am. J. Hum. Genet. **60**: 188-96.

McKeigue, P. M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am. J. Hum. Genet. 1 **63**: 241-51.

Meigs, J., R. D'Agostino, P. Wilson Sr., L. Cupples, D. Nathan and D. Singer (1997). Risk variable clustering in the insuling resistance syndrome. The Framingham Offspring Study. Diabetes **46**: 1594-600.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). Equations of state calculations by fast computing machines. J Chem Phys **21**: 1087-91.

Mitchell, B. D., C. M. Kammerer, J. Blangero, M. C. Mahaney, D. L. Rainwater, B. Dyke, J. E. Hixson, R. D. Henkel, R. M. Sharp, A. G. Comuzzie, J. L. VandeBerg, M. P. Stern and J. W. MacCluer (1996a). Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. Circulation **94**: 2159-70.

Mitchell, B. D., C. M. Kammerer, M. C. Mahaney, J. Blangero, A. G. Comuzzie, L. D. Atwood, S. M. Haffner, M. P. Stern and J. W. MacCluer (1996b). Genetic analysis of the IRS. Pleiotropic effects of genes influencing insulin levels on lipoprotein and obesity measures. Arterioscler. Thromb. Vasc. Biol. **16**: 281-8.

Moll, P. P., T. L. Burns and R. M. Lauer (1991). The genetic and environmental sources of body mass indes variability: The Muscatine ponderosity family study. Am J Hum Genet **49**: 1243-55.

Morgan, T. H. (1911). Random segregation versus coupling in mendelian inheritance. Science **34**: 384.

Muller, D., D. Elahi, R. Pratley, J. Tobin and R. Andres (1993). An epidemiological test of the hyperinsulinemia-hypertension hypothesis. J Clin Endocrinol Metab **76**: 544-8.

Muller, D. C., D. Elahi, J. D. Tobin and R. Andres (1996). The effect of age on insulin resistance and secretion: a review. Semin Nephrol **16**: 289-98.

Nabika, T., A. Bonnardeaux, J. M., C. Julier, X. Jeunemaitre, P. Corvol, M. Lathrop and F. Soubrie (1995). Evaluation of the SA locus in human hypertension. Hypertension 1995 **25**: 6-13.

Narkiewicz, K., R. Szczech, M. Winnicki, M. Chrostowska, R. Pawlowski, W. Lysiak-Szydlowska, I. Choe, M. Kato, W. I. Sivitz, B. Krupa-Wojciechowska and S. V.K. (1999). Heritability of plasma leptin levels: a twin study. J. Hypertens. **17**: 27-31.

Neel, J. V. (1962). Diabetes mellitus: A "thrifty" genotype rendered detrimental/by "progress"? Am. J. Hum. Genet. **14**: 353-62.

Neel, J. V., S. Julius, A. Weder, M. Yamada, S. L. R. Kardia and M. Haviland (1998). Syndrome X: Is it for real? Genet. Epidemiol. **15**: 19-32.

Nelson, R. G., M. L. Sievers, W. C. Knowler, B. A. Swinburn, D. J. Pettitt, M. F. Saad, I. M. Liebow, B. V. Howard and B. H. Bennett (1990). Low incidence of fatal coronary heart disease in Pima Indians despite high prevalence of non-insulin-dependent diabetes. Circulation **81**: 987-95.

Norman, R. A., P. A. Tataranni, R. Pratley, D. B. Thompson, R. L. Hanson, M. Prochazka, L. Baier, M. G. Ehm, H. Sakul, T. Foroud, W. T. Garvey, D. Burns, W. C. Knowler, P. H. Bennett, C. Bogardus and E. Ravussin (1998). Autosomal genomic scan for loci linked to obesity and energy metabolism in Pima Indians. Am J Hum Genet **62**(3): 659-68.

O'Connell, J. R. and D. E. Weeks (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet **63**: 259-66.

Ohman, M., L. Oksanen, J. Kaprio, M. Koskenvuo, P. Mustajoki, A. Rissanen, J. Salmi, K. Kontula and e. al. (2000). Genome-wide scan of obesity in Finnish sibpairs reveals linkage to chromosome Xq24. J Clin Endocrinol Metab **85**: 3183-90.

Ott, J. (1999). Analysis of Human Genetic Linkage. Baltimore, The John Hopkins University Press.

Pajukanta, P., M. Cargill, L. Viitanen, I. Nuotio, A. Kareinen, M. Perola, J. D. Terwilliger, E. Kempas and e. al. (2000). Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland. Am J Hum Genet **67**: 1481-93.

Pankow, J. S., K. M. Rose, A. Oberman, S. C. Hunt, L. D. Atwood, L. Djousse, M. Province and D. C. Rao (2000). Possible locus on chromosome 18q influencing postural systolic blood pressure changes. Hypertension **36**: 471-6.

Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann. Eugen. **6**: 133-8.

Perusse, L., T. Rice, Y. C. Chagnon, J. P. Despres, S. Lemieux, S. Roy, M. Lacaille, M. Ho-Kim, M. Chagnon, M. Province, D. C. Rao and C. Bouchard (2001). A genome-wide scan for abdominal fat assessed by computed tomography in the Quebec family study. Diabetes **50**: 614-21.

Perusse, L., T. Rice, J. P. Despres, J. Bergeron, M. A. Province, J. Gagnon, A. S. Leon, D. C. Rao, J. S. Skinner, J. H. Wilmore and C. Bouchard (1997). Familial resemblance of plasma lipids, lipoproteins and postheparin lipoprotein and hepatic lipases in the HERITAGE Family Study. Arterioscler. Thromb. Vasc. Biol. **17**: 3263-9.

Ploughman, L. M. and M. Boehnke (1989). Estimating the power of a proposed linkage study for a complex genetic trait. Am J Hum Genet **44**: 543-51.

Poulsen, P., K. O. Kyvik, A. Vaag and H. Beck-Nielsen (1999). Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia **42**: 139-45.

Pratley, R. E., D. B. Thompson, M. Prochazka, L. Baier, D. Mott, E. Ravussin, H. Sakul, M. G. Ehm and e. al. (1998). An autosomal genomic scan for loci linked to prediabetic phenotypes in Pima Indians. J Clin Invest **101**: 1757-64.

Price, R. A., M. A. Charles, D. J. Pettitt and W. C. Knowler (1993). Obesity in Pima Indians: large increases among post-World War II birth cohorts. Am J Phys Anthropol **92**: 473-9.

Price, R. A., K. Lunetta, R. Ness, M. A. Charles, M. F. Saad, E. Ravussin, P. H. Bennett, D. J. Pettitt and W. C. Knowler (1992). Obesity in Pima Indians. Distribution characteristics and possible thresholds for genetic studies. Int J Obes Relat Metab Disord **16**: 851-7.

Rainwater, D. L., L. Almasy, J. Blangero, S. A. Cole, J. L. VandeBerg, J. W. MacCluer and J. E. Hixson (1999). A genome search identifies major quantitative trait loci on human chromosomes 3 and 4 that influence cholestoerol concentrations. Arterioscler Thromb Vasc Biol **19**: 777-83.

Ravussin, E., M. E. Valencia, J. Esparza, P. H. Bennett and L. O. Schulz (1994). Effects of a traditional lifestyle on obesity in Pima Indians. Diabetes Care **17**: 1067-74.

Reaven, G. (1988). Role of insulin resistance in human disease. Diabetes **37**: 1595-607.

Reaven, G. M. (1993). Role of insulin resistance in human disease (Syndrome X): an expanded definition. Annui Rev Med **44**: 121-31.

Rice, T., T. Rankinen, M. Province, Y. C. Chagnon, L. Perusse, I. B. Borecki, C. Bouchard and D. C. Rao (2000). Genome-wide linkage analysis of systolic and diastolic blood pressure: The Quebec Family Study. Circulation **102**: 1956-63.

Rice, T., G. P. Vogler, T. S. Perry, P. M. Laskarzewski, M. A. Province and D. C. Rao (1990). Heterogeneity in the familial aggregation of fasting plasma glucose in five North American populations: the Lipid Research Clinics Family Study. Int. J. Epidemiol. **19**: 290-6.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. J R Stat Soc B **59**: 731-92.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. Markov Chain Monte Carlo in Practive. D. J. Spiegelhalter. London, Chapman and Hall: 45-57.

Rose, K. M., B. Newman, E. J. Mayer-Davis and J. V. Selby (1998). Genetic and behavioral determinants of waist-hip ratio and waist circumference in women twins. Obes. Res. **6**: 383-92.

Rotimi, C., A. Luke, Z. Li, J. Compton, R. Bowsher and R. Cooper (1997). Heritability of plasma leptin in a population sample of African-American families. Genet. Epidemiol. **14**: 255-63.

Schumacher, M. C., S. J. Hasstedt, S. C. Hunt, R. R. Williams and S. C. Elbein (1992). Major gene effect for insulin levels in familial NIDDM pedigrees. Diabetes **41**: 416-23.

Segal, H. (1989). Kosrae: The Sleeping Lady Awakens. Kosrae Tourist Division, Dept. of Conservation and Development. Kosrae state government, FSM.

Shearman, A. M., J. M. Ordovas, L. A. Cupples, E. J. Schaefer, M. D. Harmon, Y. Shao, J. D. Keen, A. L. DeStefano and e. al. (2000). Evidence for a gene influencing the TG/HDL-C ratio on chromosome 7q32.3-qter: a genome-wide scan in the Framingham Study. Hum Mol Genet **9**: 1315-20.

Sheehan, N., A. Possolo and E. A. Thompson (1989). Image processing procedures applied to the estimation of genotypes on pedigrees. Am J Hum Genet **45**: A248.

Sheehan, N. and A. Thomas (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrika **49**: 163-75.

Shimkets, R., D. Warnock, C. Bositis, C. Nelson-Williams, J. Hansson, M. Schambelan, J. J. Gill, S. Ulick, M. RV and J. Findling (1994). Liddle's syndrome: heritable human hypertension caused by mutations in the beta subunit of the epithelial sodium channel. Cell **79**: 407-14.

Shintani, T. T., S. Beckham, H. K. O'Connor, C. Hughes and A. Sato (1994). The Waianae Diet Program: a culturally sensitive, community-based obesity and clinical intervention program for the Native Hawaiian population. Hawaii Med J **53**: 136-41.

Shintani, T. T., C. K. Hughes, S. Beckham and H. K. O'Connor (1991). Obesity and cardiovascular risk intervention through the ad libitum feeding of traditional Hawaiian diet 1-3. Am. J. Clin. Nutr. **53**: 1647S-51S.

Shmulewitz, D., S. B. Auerbach, T. Lehner, M. Blundell, J. D. Winick, L. D. Youngman, V. Skilling, S. C. Heath, J. Ott, M. Stoffel, J. L. Breslow and J. M. Friedman (2001). Epidemiology and factor analysis of obesity, type II diabetes, hypertension, and dyslipidemia (Syndrome X) on the island of Kosrae, Federated States of Micronesia. Human Heredity **51**: 8-19.

Shmulewitz, D. and S. C. Heath (2001). Genome scans for Q1 and Q2 on general population replicates using Loki. Genet Epidemiol **21**: S686-91.

Snow, G. L. and E. M. Wijsman (1998). Pedigree Analysis Package (PAP) vs. MORGAN: Model selection and hypothesis testing on a large pedigree. Genet Epidemiol **15**: 355-69.

Sobel, E., J. C. Papp and K. Lange (2002). Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet **70**: 496-508.

Sorensen, T., C. Holst, A. Stunkard and L. Skovgaard (1992). Correlations of body mass index of adult adoptees and their biological and adoptive relatives. Int. J. Obes. Relat. Metab. Disord. **16**: 227-36.

Soutar, A. K. (1998). Update on low density lipoprotein receptor mutations. Curr. Opin. Lipidol. **9**: 141-7.

Stephens, D. A. and R. D. Fisch (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54**: 1334-47.

Stevens, J. (1986). Applied Multivariate Statistics for the Social Sciences. Hillsdale, NJ, Lawrence Erlbaum Associates.

Stoffers, A., J. Ferrer, W. Clarke and J. Habener (1997). Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. Nature Genetics **17**: 138-9.

Stroup-Benham, C. A., K. S. Markides, D. V. Espino and J. S. Goodwin (1999). Changes in blood pressure and risk factors for cardiovascular disease among older Mexican-Americans from 1982-1984 to 1993-1994. J Am Geriatr Soc **47**: 804-10.

Stunkard, A. J., T. T. Foch and Z. Hrubec (1986). A twin study of human obesity. JAMA **256**: 51-4.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J. of Experimental Zoology **14**: 43-59.

Suzuki, M., M. Ikebuchi, K. Shinozaki, Y. Hara, M. Tsushima, T. Matsuyama and Y. Harano (1996). Mechanism and clinical implication of insulin resistance syndrome. Diabetes **45 Suppl 3**: S52-4.

Szydlowski, M. and S. C. Heath (2000). Performance of hybrid MCMC samplers on a large complex pedigree. Am J Hum Genet **67**: A207.

Thomas, D. C. and W. J. Gauderman (1996). Gibbs sampling methods in genetics. Markov chain Monte Carlo in Practice. D. J. Spiegelhalter. London, Chapman and Hall.

Thomas, D. C., S. Richardson, J. Gauderman and J. Pitkaniemi (1997). A Bayesian approach to multipoint mapping in nuclear families. Genet Epidemiol **14**: 903-8.

Thompson, E. and R. Shaw (1992). Estimating polygenic models for multivariate data on large pedigrees. Genetics **131**: 971-8.

Thompson, E. A. (1991). Probabiliteis on complex pedigrees: the Gibbs sampler approach. Computer Science and Statistics: Proceedings of teh 23rd Symposium on the Interface. Fairfax Station, Interface Foundation of North America**:** 371-8.

Thompson, E. A. (1994a). Monte carlo estimation of multilocus autozygosity probabilities. Proceedings of the 1994 Interface Conference. A. Lehman. Fairfax Station, Interface Foundation of North America.

Thompson, E. A. (1994b). Monte carlo likelihood in genetic mapping. Stat Sci **9**(355-366).

Thompson, E. A. (1996). Likelihood and linkage: from Fisher to the future. Ann Statist **24**: 449-65.

Thompson, E. A. and S. C. Heath (1999). Estimation on conditional multilocus gene identity among relatives. Statistics in Molecular Biology and Genetics. IMS Lecture Notes- monograph series**:** 95-113.

Uimari, P. and M. Sillanpaa (2001). Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genet Epidemiol **21**: 224-42.

Van der Kooy, K. and J. Seidell (1993). Techniques for the measurement of visceral fat: a practical guide. Int. J. of Obesity **17**: 187-96.

Vionnet, N., E. H. Hani, S. Dupont, S. Gallina, S. Francke, S. Dotte, F. De Matos, E. Durand and e. al. (2000). Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet **67**: 1470-80.

Vionnet, N., M. Stoffel, J. Takeda, K. Yasuda, G. I. Bell, H. Zouali, S. Lesage, G. Velho, F. Iris, P. Passa, P. Froguel and D. Cohen (1992). Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. Nature **356**: 721-3.

Wannamethee, S., A. Shaper, P. Durrington and I. Perry (1998). Hypertension, serum insulin, obesity and the metabolic syndrome. Journal of Human Hypertension **12**: 735-41.

Watanabe, R. M., S. Ghosh, C. D. Langefeld, T. T. Valle, E. R. Hauser, V. L. Magnuson, K. L. Mohlke, K. Silander and e. al. (2000). The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. Am. J. Hum. Genet . **67**: 1186-200.

West, D. B., C. N. Boozer, D. L. Moody and R. L. Atkinson (1992). Obesity induced by a high fat diet in nine strains of inbred mice. Am. J. Physiol. **262**: R1025-R32.

Wijsman, E. and C. Amos (1997). Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet. Epidemiol. **14**: 719-35.

Williams, R. R., S. C. Hunt, P. N. Hopkins, L. L. Wu, S. J. Hasstedt, T. D. Berry, G. K. Barlow, B. M. Stults, M. C. Schumacher and E. H. Ludwig (1993). Genetic basis of familial dyslipidemia and hypertension: 15-year results from Utah. Am. J. Hypertens. **6**: 319S-27S.

Williams, R. R., S. C. Hunt, P. N. Hopkins, L. L. Wu and J. M. Lalouel (1994). Evidence for single gene contributions to hypertension and lipid disturbances: definition, genetics, and clinical significance. Clin. Genet. **46**: 80-7.

Wilson, P. W. (1994). Established risk factors and coronary artery disease: the Framingham Study. Am J Hypertens **7**: 7S-12S.

Wiltshire, S., T. M. Frayling, A. T. Hattersley, G. A. Hitman, M. Walker, J. C. Levy, S. O'Rahilly, C. J. Groves and e. al. (2002). Evidence for linkage of stature to chromosome 3p26 in a large U.K. family data set ascertained for type 2 diabetes. Am J Hum Genet **70**: 543-6.

Wiltshire, S., A. T. Hattersley, G. A. Hitman, M. Walker, J. C. Levy, M. Sampson, S. O'Rahilly, T. M. Frayling and e. al. (2001). A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (The Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. Am. J. Hum. Genet. **69**: 553-69.

Yamagata, H. Furuta, N. Oda, P Kaisaki, S. Menzel, N. Cox, S. Fajans, S. Signorini, M. Stoffel and G. Bell (1996a). Mutation in hepatocyte nuclear factor 4a gene in maturity-onset diabetes of the young (MODY). Nature **384**: 458-60.

Yamagata, K., N. Oda, P. Kaisaki, S. Menzel, H. Furuta, M. Vaxillaire, L. Southam, R. Cox, G. Lathrop, V. Boriraj, X. Chen, N. Cox, Y. Oda, H. Yano, M. Le Beau, S. Yamada, H. Nishigori, J. Takeda, S. Fajans, A. Hattersley, N. Iwasaki, T. Hansen, O. Pedersen, K. Polonsky and G. Bell (1996b). Mutations in the hepatocyte nuclear factor 1a gene in maturity onset diabetes of the young (MODY3). Nature **384**: 455-8.

Yuan, B., R. Neuman, S. H. Duan, J. L. Weber, P. Y. Kwok, N. L. Saccone, J. S. Wu, K. Liu and e. al. (2000). Linkage of a gene for familial hypobetaliaproteinemia to chromosome 3p21.1-22. Am J Hum Genet **66**: 1699-704.

Zhang, Y., P. Proenca, M. Maffei, M. Barone, L. Leopold and J. M. Friedman (1994). Positional cloning of the mouse *obese* gene and its human homologue. Nature **372**: 425-32.

Zimmet, P., P. Taft, A. Guinea, W. Guthrie, and K. Thoma (1977). The high prevalence of diabetes mellitus on a Central Pacific Island. Diabetologia **13:** 111-5.

Zimmet, P., M. Arblaster and K. Thoma (1978). The effect of westernization on native populations. Studies on a Micronesian community with a high diabetes prevalence. Australian & New Zealand Journal of Medicine **8**(2): 141-6.