2010

Genome-Scale Genetics: Lessons from Founder Populations

Eimear Elizabeth Kenny

GENOME-SCALE GENETICS:

LESSONS FROM FOUNDER POPULATIONS

A Thesis presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Eimear Elizabeth Kenny

June 2010

GENOME-SCALE GENETICS: LESSONS FROM FOUNDER POPULATIONS

Eimear Elizabeth Kenny

The Rockefeller University, June 2010

The potential benefits of using population isolates in genetic mapping due to reduced genetic and environmental heterogeneity are offset by the challenges posed by these populations for traditional association methods. Population isolates often contain large amounts of direct and cryptic relatedness that confound baseline assumptions of independence among genotypes and phenotypes and require specialized approaches to account for this sample structure. We examined three such approaches for association testing: (i) scoring allele transmission to offspring within families (ii) incorporating a permutation-based association score between families into the test statistic and finally (iii) incorporation of a kinship matrix to capture the relatedness among all individuals into a mixed model to test for association. The mixed model approach had 88% power to rank the true SNP as among the top 10 genome-wide top 10 with 56% achieving genome-wide significance, a >80% improvement over other methods. We then used the mixed model method for genome scans relating to metabolic traits and electrocardiographic measures in 2,906 related individuals from the Island of Kosrae, Federated States of Micronesia, who were previously genotyped for over 330,000 SNPs. We re-analyzed data for 17 published and 8 previously unpublished metabolic and electrocardiographic traits. We replicate seven genome-wide significant associations with known loci of plasma cholesterol, high density lipoprotein, low density lipoprotein, triglycerides, thyroid stimulating hormone and C-reactive protein, with only one detected in the previous analysis of the same traits. We further report novel associations for height (rs17629022, $p<2.1E-8$), homocysteine (rs7481043, $p<1.3E-8$) and uric acid (rs2186571, $p<1.8E-34$), the latter two near relevant candidate genes. We demonstrated the increased power of mixed-models for handling hidden and direct relatedness in isolated cohorts and discovered three novel associations with height, homocysteine and urate levels. Our experiences in association testing in an isolated population can serve as a model for other studies of similar cohorts

Dedicated to my parents, Stephen and Barbara Kenny, for their unceasing and unconditional love, guidance and support.

Acknowledgments

I want to express my deep gratitude to the many people who have helped get to where I am today. First and foremost, I would like to thank my two advisors: **Jan Breslow** and **Itsik Pe'er**. It has been a true pleasure to spend time in both of their labs and I am very grateful for the training they have given me. I would like to say a big thanks for all the ways they have gone out of their way to help and support me over the past four years.

I am thankful to every member of the Breslow and Pe'er labs, past and present, for their help and friendship. From the Pe'er lab, I want to give a big shout out to everyone in CSB 507 and say thanks for the past three years of being the nicest and most helpful officemates imaginable. I want to particularly thank **Sasha Gusev,** with whom I worked closely on the GERMLINE analysis for his unflagging patience and enthusiasm. Thanks also to **Snehit Prabhu, Piero Palamara, Yufeng Shen** and **Ninad Dewal** for hours and hours of exciting conversations and hypothesis formations. Guys, if we can conquer Koko, we can conquer anything! Many thanks to the students who worked with me during the past three years, in particular **Rossella Melchiotti** and **Minseung Kim**, two very talented individuals who always gave 200% – I really enjoyed working with you and appreciated your efforts. Thanks also to my new labmates **Arthi Ramachandran, Anat Kreimer** and **Vlada Marakov.** From the Breslow lab, I want to thank all by Breslow labmates for fun discussions across lab benches, in particular, **Martha Noel** and **Ralph Burkhardt.**

I would also like to thank the many heads and members of collaboration labs with whom I worked closely, in particular **Jeffery Freidman** (Rockefeller University)**, Dana Pe'er** (Columbia University), **David Altshuler** (MIT), **Effie Sehayek** (Cleveland Clinic) and **Chris Newton-Cheh** (MGH). .I want to thank **Jenni Lowe** for introducing me to statistical genetics and getting me started. The Tri-Institutions have so many useful resources and so many kind people who have helped me, in particular **Connie Zhao** (RU Genomics).

My graduate career would not have run as smoothly without the help of everyone behind the scenes at the Dean's office at the Rockefeller Graduate School and at the Columbia Computer Science Department. A big thanks in particular to **Cris, Michelle** and **Marta** for always going out of their way to be helpful. Thanks also to **Sid**, **Emily**, **Kristen**, **Daisy** and **Elias**.

I thank my committee members: **Fred Cross**, **Jeffery Friedman** and **Jason Mezey** for their advice, assistance and encouragment over the past few years. I also want to thank my external committee member, **Laurie Ozelius** (MSSM). I took a long windy road to graduate school, and along the way I was mentored and encouraged by some wonderful

advisors. **Richie Porter** (Trinity College Dublin), my undergraduate research advisor, gave me my first ever lab project and has been hugely supportive over the years. **Uttam Rajbhandary** (MIT), who advised me during my post-undergraduate research, instilled in me scientific discipline. **Paul Sternberg** (Caltech), my bioinformatics research advisor, allowed me great freedom to pursue scientific ideas and gave me huge encouragement as a scientist.

I never would have made it without the support and love from my family and friends. I especially want to thank my parents **Stephen** and **Barbara Kenny** for their constant love and understanding. Thank you for always providing that safety net of support that allows me to reach for the stars! I want to thank also my sister **Niamh**, thanks for keeping me in wine and cheese and for keeping me grounded and smiling. Thanks also to my brother **Stephen** and his girlfriend **Helen** (and best of luck for her own thesis writing). Thanks for a lifetime of support from my lovely aunts, uncles and cousins (some of whom have visited me in New York), in particular; **Radene, Horace, Meave, Bobby, Donal, Mary, Karen, Danny, Maureen, Patrick, Kathleen, Mairead, Robin, Cait, Martin, Finbarr, Paula, Neil, Ellen, Blaithnaid, Orla, Maria, David, Amelia, Eva** and **Antonia.** Many thanks to my friends far away, who keep my spirits up over the phone and whom I miss; **Sarah** (Boston), **Brian** (Boston), **Pankaj** (Californai), **Lisa** (Paris) and **Janice** (Dublin). I am very grateful to all my classmates in the Tri-Institutional CBM and the Rockefeller graduate programs, in particular **Byron** and **Amrita.** Finally, I want to extend a huge thank you to my "New York Family", **Sarah**, **Erica**, **Dave**, **Aine, Shaheen, Anna** and **Helgi**, I truly would not have made it without you!

Table of Contents

List of Figures

.

List of Tables

List of Abbreviations

ABCG5	ATP-binding cassette, sub-family G (WHITE), member 5
ABCG8	ATP-binding cassette, sub-family G (WHITE), member 8
AJ	Ashkenazi Jews
APOA5	Apolipoprotein A5
APOCIII	Apolipoprotein CIII
APOE	Apolipoprotein E
ASN	HapMap panel; combined Japanese and Han Chinese
ATG16L1	ATG16 autophagy related 16-like 1
BEC	Before the Common Era
BMI	Body mass index
CD	Crohn Disease
CETP	Cholesteryl ester transfer protein
CEU	HapMap panel; CEPH-Utah
CHB	HapMap panel; Han Chinese
CLV	Cornell voltage
CMP	Campesterol:Cholesterol
CRP	C-reactive protein
CRP	C-reactive protein
dbGAP	database of Genotypes and Phenotypes
DBP	Diastolic blood pressure
EMMAX	Efficient Mixed Model Association eXpedited
FBAT	Family-Based Association Testing
FBAT+Wald	Family-Based Association Testing + Wald test
FBS	Fasting blood sugar
FDR	False Discovery Rate
FLT	Folic acid
FOXE1	Forkhead box E1
GERMLINE	Genetic Error-tolerant Regional Matching with LINear-time Extension
GFAP	Glial fibrillary acidic protein
GFR	Glomular filtration rate
GWAS	Genome wide association study
gws	genome wide significance
HapMap	The International HapMap Project
HBA1C	Haemoglobin A1C
HDL	High density lipoprotein cholesterol
HGT	Height
HOMO	Homocysteine

IBD	Identity-by-descent
IBD	Inflammatory Bowel Syndrome
IBS	Identity-by-state
IL23R	interleukin 23 receptor
INS	Insulin sensitivity
JHapMap	Jewish HapMap
JPT	HapMap panel; Japanese
LAT	Lathosterol:Cholesterol
LDL	Low density lipoprotein cholesterol
LEP	Leptin hormone levels
NOD2	Nucleotide-binding oligomerization domain containing 2
NOX4	NADPH oxidase 4
OAT4	solute carrier family 22, member 11 (aka SLC22A11)
PCF	Percent body fat
PLINK/QFAM-total	Family-Based Association Tests for Quantitative Traits, within- and between-mode
PPS	Plasma Plant Sterols
PRD	PR interval
QRS	QRS interval
RAM	Random-access Memory
RRD	Resting rate interval
SBP	Systolic blood pressure
SIT	Sitosterol:Cholesterol
SLC22A11	solute carrier family 22, member 11 (aka OAT4)
SLC22A12	solute carrier family 22, member 12 (aka URAT1)
SLV	Sokolow-Lyon voltage
SNP	Single Nucleotide Polymorphism
SOLAR	Sequential Oligogenic Linkage Analysis Routines
TC	Total cholesterol
TG	Triglycerides
TSH	Thyroid stimulating hormone
URAT1	Solute carrier family 22, member 12 (aka SLC22A12)
URT	Uric acid
WST	Waist circumference
WT	Weight
YRB	HapMap panel; Yoruban-Nigeria

Chapter 1 : Introduction

Genome-scale human genetics in 2010

Rapid technological advances have accelerated the pace of human genetics in the past decade. Landmark accomplishments include the initial sequencing of the human genome in 2001 (Venter, 2001, Lander, 2001) and the completion of a "finished-grade" human genome reference sequence in 2004 (The Human Genome Project, 2004). Another landmark achievement was the release, in 2005, by the HapMap Consortium, of the first phase of an international effort to discover and expand the catalogue of human single nucleotide polymorphisms (SNPs). As a result, by August 2006, there were more than ten million SNPs in the National Center for Biotechnology SNP database (dbSNP), and examination of the HapMap SNP's, derived from four populations across three continents, revealed ~2M of them to be commonly polymorphic (The HapMap Consortium, 2005).

These advances made possible the development of genome-scale genotyping platforms ("SNP chips"), which could, at a reasonable cost, assay 100K-1M common polymorphic genotypes across hundreds to thousands of individuals (Hirschhorn, 2005). These "SNP chips" were utilized for the design of genome-wide association studies (GWAS), in

which a dense set of SNPs across the genome is used to survey the most common genetic variation for a role in disease or to identify the heritable quantitative traits that are risk factors for disease. The underlying rationale for GWAS is the 'common disease, common variant' hypothesis, positing that common diseases are attributable in part to alleles present in > 5% of the population (Reich, 2001). To achieve unbiased association signals, GWAS are mainly performed in cohorts of unrelated individuals, following a case/control design, a departure from decades of traditional variant discovery that relied on heritability within families. The first GWAS emerged in April 2005 (Klein, 2005), and to date, GWAS have reproducibly implicated over 700 genomic regions that modify the risk for over 120 complex disease and health-related phenotypes (Kruglyak, 2008, Manolio, 2008).

Taken together, this ongoing data bonanza has opened the window to myriad previously unknown variation in the human genome, and a current challenge in human genetics is to identify the specific variation that increases susceptibility to diseases.

Complex disease gene mapping

Human geneticists seek to understand the inherited basis of human disease and the underlying biology that contributes to disease pathophysiology. The goal is to gain clinical insights that could eventually improve treatment, produce useful diagnostic or

predictive tests, and to gain biological insights that could help elucidate the functional mechanisms underlying disease. Compared to rare Mendelian (or "monogenic") diseases, complex (or "polygenic") diseases (such as cancer, coronary artery disease and diabetes), or their underlying quantitative phenotypes (such as lipid levels and blood pressure), occur fairly commonly in human populations (see **Figure 1.1)**. A substantial portion of individual differences in complex disease susceptibility are known to be due to genetic factors, however effects of age, gender, lifestyle, other environmental covariates, and the interactions between them, are also known to play a role.

Detecting variants effecting complex disease

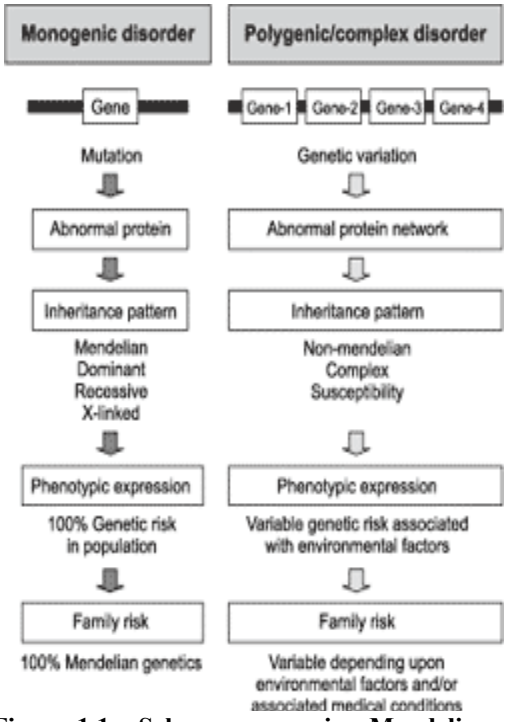


Figure 1.1: Schema comparing Mendelian ("monogenic") and complex ("polygenic") traits.

The small genetic contribution of individual genes, combined with the confounding effects of non-genetic factors on phenotype measures, have meant that loci affecting complex diseases have been historically difficult to detect, with only a handful of examples prior to 2005 (for example, the *APOE*ε2/ε4 polymorphisms affecting cholesterol levels and Alzheimer's disease (Zannis, 1981). However, with the advent of GWAS undertaken in a large number

of samples, many novel and convincingly associated loci have been mapped to numerous common polygenic disease and quantitative traits, for example, at least 32 loci for Crohn's disease and 95 loci for plasma lipid levels (McCarthy, 2008, Burnett, 2008, pers. comm. Sekar Kathiresan).

Examination of the emerging loci have revealed some interesting biological insights. On one hand, GWAS have implicated many genes that had prior evidence from candidate gene studies to play a role in their respective traits. For example, of 23 loci found to be associated with lipid levels in one study, 11 implicate genes encoding apolipoproteins, lipases and other key proteins in lipid metabolism pathways (Mohlke, 2008). Further, >20% of loci discovered in GWAS for a variety of complex traits are known to also harbor mutations that cause monogenic diseases (Hirschhorn, 2009). On the other hand, GWAS have highlighted pathways whose relevance to a particular disease or trait was previously unsuspected, such as autophagy and interleukin-23-related pathways for Crohn's disease and the complement system for age-related macular degeneration, and this clustering into pathways is highly non-random (Raychaudhuri, 2009).

Most of the time, the causal variant driving a GWAS signal is not the SNP directly assayed, but an unknown variant in the vicinity of the measured SNP, which is tagged by that SNP by virtue of residing on the same haplotypic background (being in "linkage disequilibrium" (LD)). Therefore, detecting a GWAS signal is merely the first step

toward discovering the underlying causal variant and various approaches can be taken to fine map the signal. One conventional approach is sequencing a region of arbitrary length around the signal peak across enough individuals to capture the causal allele. However, the larger the sequenced interval and number of individuals, the more variants discovered, making it difficult to distinguish the driver allele from the large number of passenger variants. This problem is exacerbated in cases where multiple, seemingly independent signals reside in close proximity along the genome (Haiman, 2007, Yeager, 2007, Gudmundsson, 2007, Fellay, 2007, Duerr, 2006, Barrett, 2008, Romeo, 2008, Graham, 2007, Graham, 2008, Abelson, 2009, Plenge, 2007). Studies to identify the culprit causal variant(s) underlying the local peaks in association signals have shown mixed success, with only handful of causal variants identified so far (Burnett, 2008, Moffatt, 2007, Burkhardt, 2008) and many re-sequencing efforts proving unsuccessful (Lowe, 2007, Burfoot, 2008, Hafler, 2009). One possible explanation may be that the GWAS signal is driven not by a single common causal variant, but by multiple effects of rare variant in a synthetic association with common alleles assayed in the GWAS (Dickson, 2010).

A limited fraction of heritability is captured by common variants

Thus far, almost all GWAS discovered signals, both individually and in combination, have exhibited modest to small effects on their associated traits (with a few notable exceptions, such as age-related macular degeneration (Klein, 2005) and drug induced

liver injury (Daly, 2009). In fact, most traits examined so far in GWAS explain <5% of the fraction of heritability of that trait (Hirschhorn, 2009). A recent very well-powered GWAS of > 100K individuals for lipid levels, that expanded on a previous study with 30K individuals, yielded ~65 extra loci, but added only 1-2% to the total fraction of heritability explained for the trait (Kathiresan, 2009, pers. comm. Sekar Kathiresan). Further, there has been a scarcity of detected reproducible associations for some examined traits, especially neurological disorders, despite numerous GWAS efforts (Liu, 2010, McMahon, 2010, Garriock, 2010).

Many explanations to account for the source of this missing heritability have been suggested, including rare and/or structural variants, poorly captured by current GWAS platforms, gene-gene interactions, high signal-to-noise ratio in current GWAS strategies, the affects of environment in study populations, and"hidden" epigenetic effects (Manolio, 2009). In response, the community is developing broader strategies for measuring such effects, for example, genome-wide analysis of epigenetic markers (Laird, 2010), a new generation of SNP chips (aided by the 1,000 Genomes Project) and tiling arrays targeting common and rare markers and structural variants, respectively, re-sequencing and sequencing-based population studies. Consensus is currently lacking, however, on which approaches and priorities for research will be most fruitful for pinning down the causes of the missing heritability.

Designing improved causal variant discovery approaches

In designing improved approaches for discovering causal variants that contribute to complex disease, a number of considerations must be taken into account, such as: optimizing power for discovery, the expected impact of findings to the field and finite research dollars. The ideal strategy, therefore, would balance cost-effective, genome-scale analysis with generality to different populations and multiple traits. In addition, the goal would be to redirect emphasis from discovering common, ancient variation (the target of current GWAS in outbred populations) to exploring rare or unique, newly arisen variants. To this end, renewed focus has been placed on study populations that play to these strengths and have been historically used for traditional genetic approaches, namely founder populations (Manolio, 2009).

Founder populations in gene mapping

Populations that have a small number of founders and/or are geographically or culturally isolated from other populations ("founder" or "isolated" populations) have a long history in genetic mapping studies of inherited disorders, particularly for the identification of rare, monogenic disease mutations (Arcos-Burgos, 2002, Heutink, 2002). Genetic isolates, such as the Finnish, Icelandic, Costa Rican, Old Order Amish, Jewish, Hutterite and Sardinian communities, have been successful in mapping not only rare monogenic,

but complex traits (Ober, 2001, Angius, 2002, Peltonen, 1999, Gulcher, 2001, Freimer, 1996, Zamel, 1996, Abney, 2000, Aulchenko, 2004, van der Walt, 2005).

Founder populations share a specific set of features as compared to outbred populations, namely reduced genetic diversity, reduced genetic and environmental heterogeneity, and often large, multi-generational pedigrees. The resulting reduction in allele heterogeneity has contributed to the success of genetic linkage and positional cloning approaches in isolated, founder populations (Peltonen, 1999, Gudmundsson, 2002, Lindqvist, 2000). These same properties that make isolated populations valuable for Mendelian trait genetics may be exploited for gene mapping of complex traits. Here we investigate the advantages and limitations of exploiting homozygosity in isolated populations for analysis of alleles affecting complex traits (Peltonen, 2000, Heutink, 2002, Gulcher, 2001).

Advantages of using founder populations in complex trait mapping

Genome-wide association studies (GWAS), performed in outbred populations, have thus far identified many common variation contributing to complex diseases (Hindorff, 2009). One limitation in these studies is that rare variants that are not in linkage disequilibrium with the common variants assayed are usually not detected (Bodmer, 2008). However, when the populations analyzed in a GWAS have a substantial number of individuals that

share a recent common ancestor, not only common variants (Lowe 2009, Wang 2009), but sometimes also rare variants appearing at a higher frequency, or within a discernable haplotype structure, can be identified (Sabatti, 2009, Pollin, 2008, Kallio 2009). Further, while multiple rare mutations may segregate in an outbred population, founding events and subsequent population bottlenecks may reduce the allelic diversity such that a single mutation dominates the allelic spectrum in an isolated population.

Founder populations typically exhibit a large amount of direct and cryptic relatedness, as compared to outbred cohorts, and this high degree of relatedness may result in increased homogeneity in phenotypes. In the case of small population isolates, well-documented multigenerational pedigrees can often be constructed, facilitating longitudinal analysis of both phenotypes and genotypes. Population isolates may also share similar environments, decreasing the signal-to-noise ratio for mapping. Finally, isolated populations often exhibit relatively high amounts of inbreeding producing an increased incidence of recessive phenotypes which are difficult to detect in outbred populations (Peltonen, 2000).

Technical challenges

Inbreeding and the historical lack of random mating in founder populations violate assumptions of independence between genotypes and phenotypes confounding classical

association tests which measure the frequencies of alleles in disease cases versus controls. Specialized tests may account for any "known" relationships in the population, however hidden (relationships unreported or incorrectly specified) or cryptic relatedness (unreported or unknown distant relationships) may confound even specialized test statistics. Further, since consanguinity may be present, two alleles in a random individual may also be correlated. The hidden correlation in the data can cause an overdispersion in the null distribution of the naïve test scores for association and consequently, false positive associations. Hence, the major challenge for performing GWAS in isolated populations is to account for this non-random intra- and inter-individual correlation.

Most GWAS have been thus far focused on outbred populations, and >75% utilize populations of European descent exclusively (Rosenberg 2010). This can lead to additional analytically challenges when studying non-European and/or founder populations. Several studies have found that markers typically used in GWAS are in various ways non-random (Nielsen 2004, Clark 2005). The focus on SNPs with high minor allele frequencies in populations of European ancestry in the initial selection of the markers for use on current GWAS platforms, in turn affects the relative proportion of the genome suited to mapping in different populations. This problem is exacerbated in founder populations, where the reduced genotypic heterogeneity further decreases the number of viable markers.

A recent trend for increasing power in GWAS is to adopt a genotype-imputation strategy (Halperin 2009). The approach relies on the assumption that two haplotypes, that are identical for a set of nearby markers, are likely to share the intervening chromosomal stretch identically by descent. Therefore, if one of the two haplotypes is genotyped more densely than the second haplotype, genotypes at the unmeasured positions on the second haplotype can be predicted ("imputed"). This approach requires the availability of a densely genotyped reference panel, such as the HapMap project (Frazer 2007). However, there is reduced portability of this approach to populations that are, compared to available reference panels, genetically distinct, such as founder populations. This leads to a reduction in imputation accuracy, and a consequent reduction in statistical power for imputation based association mapping.

To take full advantage of the opportunities afforded for detecting causal variants in founder populations, there is a clear need for improved methods to overcome these technical challenges. To tackle this problem, we studied two distinctly different founder populations; a small, island-based, extreme isolate (the Kosrae islanders) and a large, urban-based, less extreme founder population (the Ashkenazi Jews).

A small, extreme founder population: Kosrae

Kosrae is a Pacific Island member of the Federated States of Micronesia. It is located in the Pacific Ocean 370 mi (590 km) north of the equator, between Guam and the Hawaiian Islands, at 5°17' north and 162°58' east, longitude and latitude. Its land area is very small, approximately 42 mi^2 (110 km^2) and the current day population of the island is ~7,700 according to the most recent 2001 census.

The history of Kosrae

Although no historical records exist, it is believed that Kosrae was originally settled by a small number of founders (estimated to be 50) between 1,500-2,500 years ago, and subsequent estimates of the founder population size from current day genetic data from Kosrae support this (pers. comm. Dr Itsik Pe'er and Sasha Gusev). Recent work has suggested that the island settlers were Polynesian, originally of aboriginal Taiwanese descent (Friedlaender, 2008, Kayser, 2008). Kosrae was first sighted by Westerners in

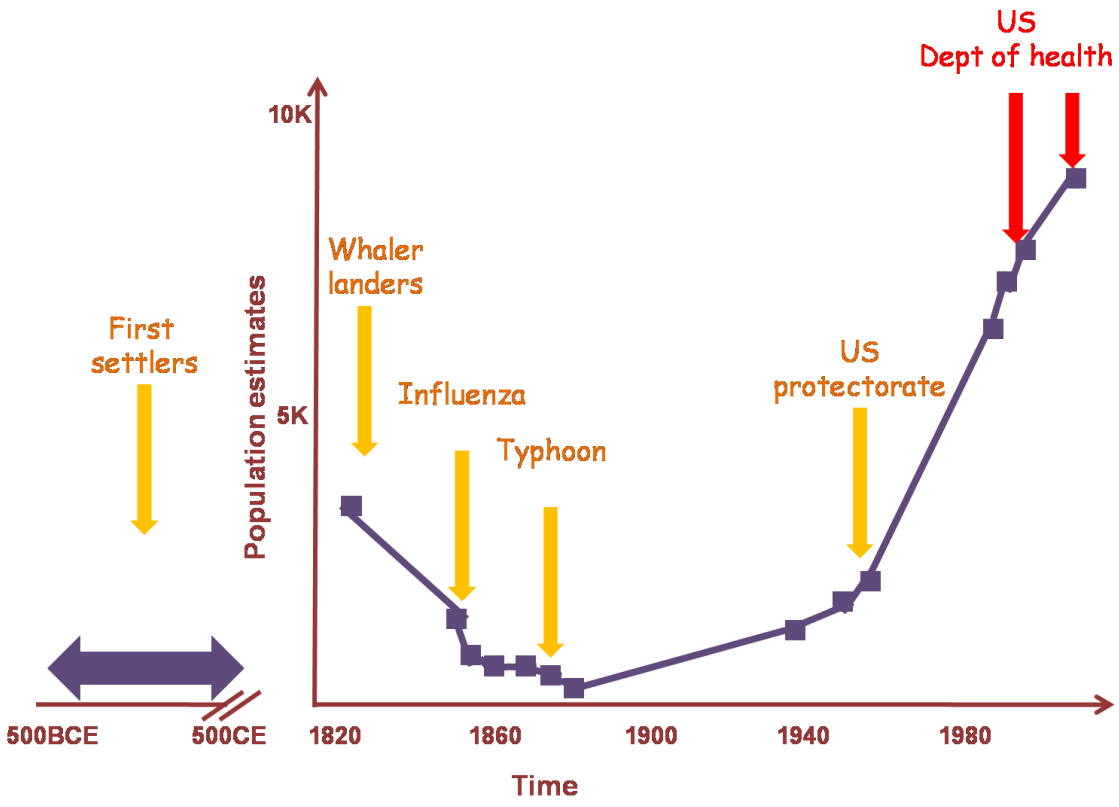


Figure 1.2: Estimates of population size for the past 2,500 years on the Microneasian Island of Kosrae

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 also indicate that there was admixture between native Kosraean females and 4-6 male Caucasian whalers from New England and Europe who visited the island in the mid to late 19th century. However, current day estimates of Caucasian admixture on the island indicate that there is an average <1% Caucasian genome across all the islanders (Bonnen, 2010).

Spain originally claimed Kosrae, but the island was purchased by Germany in 1899 following the Spanish-American War. After World War II (WWII), the Federated States of Micronesia, 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 diet shift on Kosrae

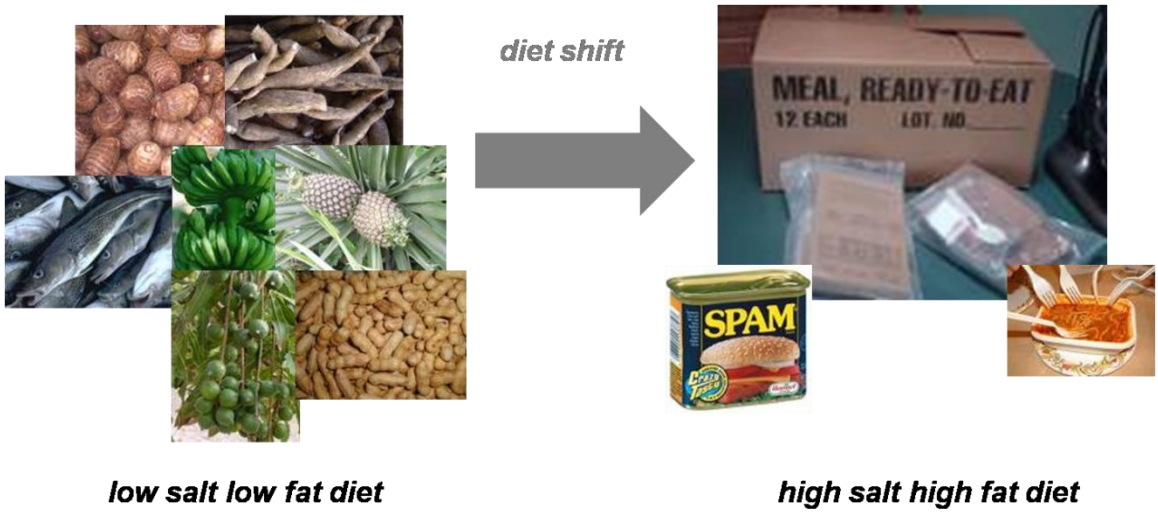


Figure 1.3: Representation of the diet of indigenous Kosraens pre- (to the left) and post- (to the right) WWII. Pre-WWII, the indigenous islanders diet was high in starch (from tubers and fruit) and lean proteins (from fish) and polyunsaturated fats (from nuts). Post-WWII, with the supply of ready-to-eat meals from the US, the diet shifted to a high salt, high fat diet.

A number of visitors to the island in the 19th and early 20th century noted the natives to be thin. Prior to WWII, the Kosraens consumed a diet consisting mostly of fish, fruits and

vegetables. The designation of Kosrae as a US protectorate led to a drastic life style change on the island. Kosraens 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 exposed to a Western diet and lifestyle, such as the Samoans of Polynesia and the Navajo Native Americans of southwestern United States (Schumacher 2008, Lee 2008).

The current health status of the islanders

In response to the apparent growing epidemic of obesity on Kosrae, the United States Department of Health conducted a survey of the Island in two screenings in 1994 and 2001. They collected information on a range of clinical traits that underlie the Metabolic Syndrome—a co-occurrence of metabolic abnormalities that groups together multiple traits that comprise serious risk factors to type 2 diabetes (5.8% prevalence in the US), coronary artery disease (4.8% prevalence and the leading cause of death in the US),

	Metabolic Syndrome †	Obesity‡	Diabetes*
Kosrae	42%	61%	20%
US	27% (47 million)	30%	7%

Table 1.1: Incidence of Metabolic Syndrome, Obesity and Diabetes on Kosrea as compared to the US. Estimates are based on guidelines from [1,2,3]Third Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation , and Treatment of High Blood Cholesterol in Adults 2003-2006.

stroke (1.6%) (Ervin 2009). These outcomes therefore have a significant negative impact on population health in developed countries.

Prevalence estimates of metabolic syndrome were calculated for both 1994 and 2001 health surveys and were found to be very high, rising from 27% in 1994 to 42% in 2001**,** a surge of more than 50% in a seven year span. Comparisons to other regions of the world show that Kosrae has one of the highest recorded rates of metabolic syndrome; roughly 3-, 4- and 2-fold higher than estimates for Koreans, Japanese and the United States, respectively (see **Table 1.1**) (Cameron 2004).

A large, less extreme founder population: Ashkenazi Jews

The Ashkenazi Jewish people (AJ) represent a founder population with a distinct well-chronicled history. Although the AJ and Kosraen populations were both founded at a similar time in history, the two populations differ in many of their features. The AJ population is linked not by geographical isolation, but rather by: religion, language, customs and the traditional practice of endogamy. These practices have resulted in a distinct AJ genetic identity that has persevered despite thousands of years of migration and the physical separation of different AJ communities.

The history of Ashkenazi Jews

The Jewish people originated in the Middle East around the first millenium before the Common Era (BCE) and have had a long history of migration throughout the Middle East, North Africa and Europe, and more recently the Americas, Australia and South Africa. According to their area of long term residence, the modern Jewish people are divided into three groups; Sephardic Jews (who resided in Spain and Portugal until the 15th century), Middle Eastern Jews (who lived in contemporary Palestine and Israel, as well as Iraq, Iran and other Middle Eastern countries) and Ashkenazi Jews (who lived in Western and Eastern Europe).

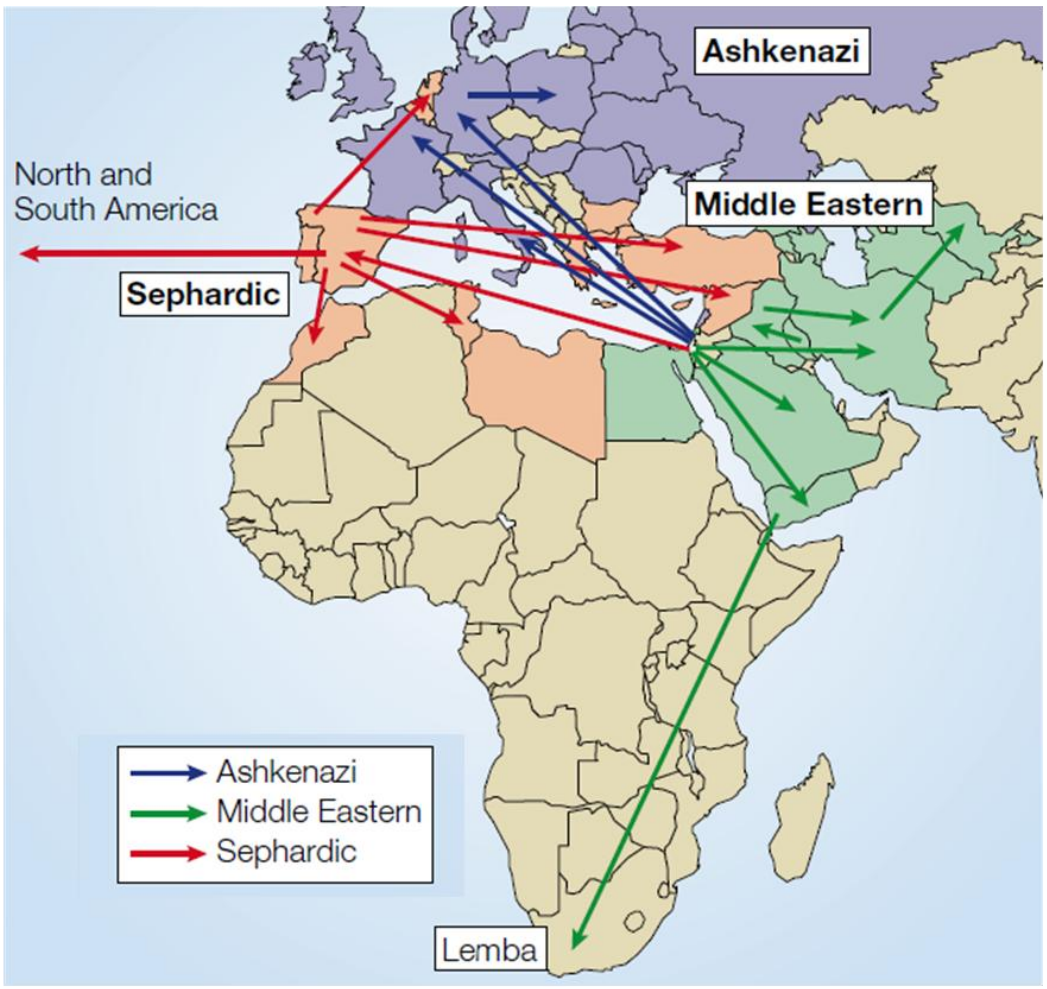
The Ashkenazi Jews (AJ) moved into Europe and then over the Alps into the cities of the Rhineland around the 9th century, where they developed their own language (Yiddish). During the 12th and 13th centuries, the AJ were expelled from the Rhineland, and resettled in Eastern Europe, mainly Poland, Lithuania and Russia. In the late nineteenth and early twentieth centurys, there were large AJ migrations westward (to America and Western Europe) and southward (to Australia and South Africa). Two important events in the twentieth century brought about the depletion of Jewry in Europe; the Jewish Holocaust of the WWII, which lead to the death of 6 million Jewish people and migration out of Europe, and the foundation of the state of Israel and the subsequent migration of Jews from all over the world to Israel.

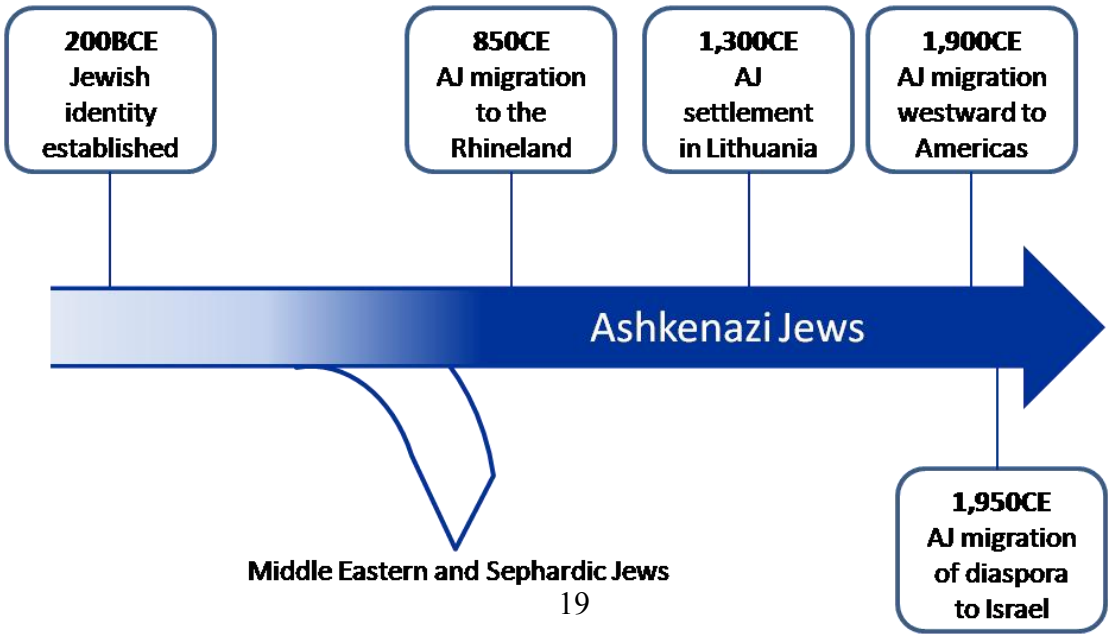
Contemporary Ashkenazi Jews comprise ~13 million people, with the largest demographic centers being the United States (US), primarily the New York metropolis, and Israel where 5-6 and 3-4 million Jews reside, respectively. Around 90% of New York based Jews are of Ashkenazi descent, making Ashkenazi Jews the largest founder population in the United States.

Ashkenazi genetic identity

By Jewish religious law, Jewish identity follows matrilineal inheritance. However, studies of mitochondrial and Y-chromosomal polymorphisms provide strong evidence for both maternal and paternal transmission, supporting historical evidence for many generations of endogamy in this population (Ostrer, 2001). Further, studies of maternally inherited mitochondrial DNA have suggested that most Jewish communities were founded by relatively few women (Thomas, 2002), and that almost half the modern AJ population may be descendent from just four women (Behar, 2006).

Since the middle of the 20th century however, many Ashkenazi Jews have intermarried, both with members of other Jewish communities and with people of other nations and faiths, while some Jews have also adopted children from other ethnic groups or parts of


Ashkenazi
North and
South America
Middle Eastern
Sephardic
Ashkenazi
Middle Eastern
Sephardic
Lemba


200BCE
Jewish
identity
established
850CE
AJ migration
to the
Rhineland
1,300CE
AJ
settlement
in Lithuania
1,900CE
AJ migration
westward to
Americas
Ashkenazi Jews
Middle Eastern and Sephardic Jews
1,950CE
AJ migration
of diaspora
to Israel

the world and raised them as Jews. Conversion to Judaism, rare for nearly 2,000 years, has become more common. Nonetheless, it has been shown that self-described Jewish ancestry is a major determinant of population genetic structure in European populations (Price 2008) and a recent study has shown that AJ individuals with different amount of AJ ancestry can form distinct identifiable groups based on genetic data alone (Need 2009).

Contemporary Ashkenazi Jewish populations have been studied extensively by the medical genetic community where at least 40 genetic conditions have been dissected (Ostrer 2001). Some of the mutations are found in multiple Jewish groups, such as the common mutation in the *BRAC1* gene which increases susceptibly to cancer, which is likely to have originated before the dispersion of Jews (Bar-Sade 1998). However, in many cases the mutations are likely to have risen during migration and/or bottleneck events, where one or two prevalent founder mutations have occurred, for example, mutations in the *GBA* and the *LDLR* genes which cause Gaucher disease and hypercholesterolaemia, respectively (Diaz 2000, Durst 2001).

Crohn's disease in Ashkenazi Jews

Crohn's disease (CD) is a chronic inflammatory disorder of the gastrointestinal tract, which is thought to result from the effect of environmental factors in a genetically

predisposed host. An epidemiological feature of CD is that it has highest prevalence among individuals of AJ descent, occurring two to four times more frequently than in non-Jewish Caucasian populations. Three coding variants of the NOD2 gene have been reported as independent disease-predisposing mutations for CD. Confirmation studies have shown, however, that these susceptibility loci have similar allele frequencies and magnitudes of the effect and are unlikely to contribute to the excess prevalence of the disease in the AJ population. Therefore, involvement of other, yet unknown, genetic variants unique to this population is hypothesized.

Improving methods for genome-scale genetics in founder populations

The focus of my thesis was the development of new, optimally-powered strategies for analysis in founder populations. I started with the genome-scale genetic analysis of an extreme founder population, the Pacific Island of Kosrae. Analysis of this population (ongoing when I joined the Breslow lab in the summer of 2006) using existing methods had demonstrated reduced power in this cohort as compared to a similarly sized outbred cohort (Lowe, 2009).

The primary goal of my thesis was to develop improved strategies for causal variant detection on Kosrae. We had previously shown abundant long genetic segments shared identically by descent on the island (Bonnen, 2006). Taking advantage of these long haplotypes, we developed a method for mapping causal haplotypes under a GWAS local signal peak (see **Chapter 2**) for dissecting multiple signals at the same locus, and

elucidated a putative causal variant. Next, we evaluated four specialized methods for performing GWAS in the context of relatedness using extensive genome-scale simulations (see **Chapter 3**), and demonstrated that mixed-models are best powered for these populations, Finally, leveraging these optimized strategies we performed a GWAS of 30 traits pertaining to metabolic disorders on Kosrae (see **Chapter 4**), reporting 10 positive controls and 3 novel findings, one of which was refined by a factor of 4 with haplotypic mapping.

The secondary goal, of my thesis was to determine whether analysis methods developed on Kosrae, could be utilized for the analysis of a less extreme founder populations, such as the Ashkenazi Jews. To demonstrate this, we performed a combined-analysis of seven Ashkenazi cohorts (see **Chapter 5**) and used mixed-models to perform a GWAS for Crohn's disease. We report two positive controls and one novel finding, where haplotypic mapping at one of the loci dissected two independent signals.

Chapter 2 : Materials and Methods

Kosrae study population collection

The data collection and phenotype definitions on the island of Kosrae is explained in detail in Lowe et al 2009 and Kenny et al 2010 (Lowe 2009, Kenny 2010). 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. Briefly, we surveyed 3,148 highly-related individuals from the Pacific Island of Kosrae in three separate screenings carried out in 1994 , 2001 and 2003 which represent >75% of the adult population on the island. 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, alcohol intake, medications (the island has a limited formulary), self-reported ethnicity, age, and health status. The family data was used to derive the parity status of the women (0 to 5 and ≥6 children). Finally, village of residence was recorded, where all individuals resided in one of five villages (Lelu, Malem, Tafunsek, Utwe and Welung).

Measuring phenotypes pertaining to Metabolic Syndrome and cardiovascular conductance on Kosrae

In all three screens, fasting blood was collected and centrifuged. Plasma and buffy coats were frozen and shipped to Rockefeller University, NY, for serological assays and DNA extraction. Phenotypes were measured for 30 metabolic and electrocardiographic traits; height (HGT), weight (WGT), waist circumference (WST), body mass index (BMI), percent body fat (PCF), leptin hormone levels (LEP), fasting blood sugar (FBS), insulin sensitivity (INS), Non-Insulin-Dependent Diabetes Mellitus (NIDDM), systolic blood pressure (SBP), diastolic blood pressure (DBP), total choleseterol (TG), low density lipoprotein cholesterol (LDL), high density lipoprotein cholesterol (HDL), triglyceride (TG), Lipoprotein(a) (LPA), Campesterol (CAMP), Sitosterol (SITO), Lathosterol (LATH), C-reactive protein (CRP), thyroid stimulating hormone (TSH), glomular filtration rate (GFR), homocysteine (HOMO), folic acid (FLT), uric acid (URT), QT interval (QT), PR interval (PRD), QRS interval (QRS), Sokolow-Lyon voltage (SLV), Cornell voltage (CLV) and RR (inverse heart rate) interval (RRD).

Anthropomorphic, Obesity, Diabetes and Hypertension:

HGT (cm's or inches converted to cm's), WGT (lb's or lb's converted to kg's) and WST (cm's of inches converted to cm's) were measured for males age 24-80 and females age

22-80 who were considered to have reached their full adult height and weight. Individuals measured for HGT in more than one screening were omitted if their measurements differed by more than 4 cm's. BMI was calculated as WGT (kg) divided by HGT squared (m^2). Obesity was defined as a BMI ≥30. PCF was measured using custom equipment (the "BodPod", Tanita manufacturing company, Tokyo). FBS (mg/dL) was measured by fingerstick with a glucometer and pre-NIDDM and NIDDM were defined according to the American Diabetes Association (ADA) recommendations as impaired fasting glucose (FBG 100-125 mg/dL) or impaired glucose tolerance (GTT 140-199 mg/dL) and FBS ≥ 126 mg/dL or OGTT2 ≥ 200 mg/dL, respectively. Individuals under treatment with insulin or oral hypoglycemic drugs were excluded. The average of three seated SBP (mmHg) and DBP (mmHg) measures, taken manually with a stethoscope and sphygmomanometer, for individuals aged 20-75, was used in the analysis. Treatment for hypertension was taken into account by calculating an adjusted residual based on (Levy 2000).

A venipuncture was done for determination of serum levels of steroid hormones, lipids and other biomarkers. LEP levels (ng/mL) were determined using commercial radioimmunoassay (ELISA) in the Freidman Lab by Jeffery Winick, The Rockefeller University, NY, (Diagnostic Systems Laboratories, Webster, TX and Linco Research Inc, St Charles, MO). 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 1985).

Lipid levels

TC (mg/dL) and TG (mg/dL) were measured either in the Breslow Lab, The Rockefeller Univeristy, NY, with commercially available enzymatic kits (Boehringer Manheim, Indianapolis, IN, Sigma Diagnostics, St. Louis, MO) or in the Rogosin Institute, NY, on a Roche COBAS Integra clinical chemistry analyzer using an enzymatic method, for the 1994 and 2001/2003 screens respectively. For the 1994 screen, LDL and HDL 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 and HDL (APOB (mg/dL), and APOA-I (mg/dL), 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 1997). For the 2001/2003 screen, HDL (mg/dL) was also measured in the Rogosin Institute on a Roche COBAS Integra clinical chemistry analyzer using an enzymatic method, and LDL (mg/dL) was calculated using the Friedewald formula (Friedewald 1972):

$$\mathrm{LDL\ (mg/dL)} = \mathrm{TC(mg/dL)} - \mathrm{HDL(mg/dL)} - \mathrm{TG(mg/dL)}/5$$

For individuals with a triglyceride value >400 mg/dL the LDL value was set as missing.

Plasma plant sterol levels (CAMP (mg/dL) , SITO (mg/dL) and LATH (mg/dL)) were measured by gas liquid chromatography (Heinemann 1993) in the Seyhayek Lab, The Cleveland Clinic, OH, and plasma levels of CAMP, SITO and LATH were expressed as the ratio of sterol to total plasma cholesterol levels. LPA levels (mg/dL) were measured in the Rogosin Institute, NY, using the Polymedco assay (which is not sensitive to the Lp(a) Kringle IV type 2 repeat size polymorphism).

Biomarkers

TSH levels (μIU/ml) were measured in the Rogosin Institute, NY. CRP (mg/dL) was measured using a chemiluminescent immunometric assay (Immulite 2000 High Sensitivity CRP, DPC, Los Angeles, CA) by the Rogosin Institute, NY. Positive outliers (>3 SD corresponding to >27.4 mg/L) were considered acute inflammatory reactions and excluded (n=28). All values below the functional sensitivity level of the assay (0.02) were Winsorized to 0.02 (n=36). Direct measures of GFR were not made for Kosrae subjects, and instead were based on plasma Creatinine (mg/dL) measured in the Rogosin Institute, NY, according to the widely used Modification of Diet in Renal Disease (MDRD) Study formula adjusted for a Chinese cohort:

$$\mathrm{GFR} = 229 \times \mathrm{Creatinine(mg/dL)}^{-1.154} \times \mathrm{Age(years)}^{-0.203} \ (\times 0.742 \text{ if female})$$

Individuals with GFR < 60 mg/dL (n=52) were considered "affected" for kidney disease and excluded. HOMO, URT, FLT levels were all measured in the Rogosin Institute, NY. Individuals taking diuretics (n=27) were excluded from the analysis of URT.

Electrocardiographic measures:

During the 2001 screen, twelve-lead electrocardiograms were recorded at a paper speed of 50 mm per second in 1756 individuals using a Universal ECG (QRS Diagnostic LLC, Plymouth, MN). De-identified paper electrocardiograms were scanned at high resolution using a Hewlett Packard 2840 All-in-one into digital images. Quantitative measures of QT interval (ms), RR (inverse heart rate) interval (ms), QRS duration (ms), PR interval (ms), P wave duration (ms), SLV (mV), CLV (mV) and PQ interval (ms) were obtained using digital calipers in Rigel 1.7.3 (AMPS LLC, NY) from two cardiac cycles of lead II and one cycle each from V1-V5 (approximately orthogonal vector components). PR interval was measured from the onset of the P wave to the onset of the QRS complex and PQ interval from the offset of the P wave to the onset of the QRS complex. SLV was determined from the sum of the amplitude of the S wave in lead V1 and the R wave in lead V5 and CLV from the sum of the amplitude of R wave in a VL and S wave in V3. The QT interval is measured from the earliest onset of the QRS complex to the end of the T wave (in leads II and V5) and the RR interval is the R wave to R wave interval, and is the inverse of the heart rate.

Trait covariate and normality adjustments

Statistical analysis was carried out using the JMP package (SAS Institute Inc., Cary, NC). All phenotyped and genotyped individuals between the ages of 18-88 were included in the analysis except where indicated above. All traits were separated by gender, adjusted for the effect of age (an adjusted residual based on age binned in 5-year intervals), outlier's $\geq$3 standard deviations from the mean were removed and Z-scores were calculated for each group. Pregnant women and non-Kosraen individuals were excluded. For monozygotic twins and individuals screened twice/thrice, the average Z-score was used. For some traits, the effects of additional covariates (BMI, smoking, village etc) were evaluated using either univariate logistic regression or multivariate logistic regression analysis to evaluate the effects of single or multiple covariates together, respectively, and the trait adjusted accordingly. Z-scores may be log or square-root transformed for normality. Finally, Z-scores were pooled across groups for downstream analysis.

Genotyping and quality control

2,906 study participants were successfully genotyped on the Affymetrix 500K platform; data were generated at Affymetrix, South San Francisco, CA. Genotypes were called with the BRLMM algorithm and a minimum call rate of 95% was achieved. 446,802 SNPs

passed quality control filters and between ~122K-91K SNP's that were monomorphic or very rare (minor allele frequency (MAF) < 0.010 for each phenotype were also excluded. The final dataset yielded between 354,901-323,902 SNPs per phenotype with MAF > 0.01 for the analysis.

Pedigree construction

A first pass reconstruction of the extended pedigree of the 3,000-strong cohort included denser sampling of a further ~1,000 related non-genotyped individuals to fill out the pedigree and careful cross-referencing of patient records. The extended Kosraen pedigree spans eight generations including 4,300 people (living or deceased), where ~90% of the individuals are contained in a single very large kindred. Genome-wide SNP data were subjected to identity-by-state analyses and identity-by-descent estimation performed in PLINK (Purcell 2007) to correct and validate the pedigree structure. Full details of the pedigree construction are described elsewhere (Lowe 2009). The CraneFoot pedigree drawing software was used to visualize subsets of the extended pedigree.

Kosrae trait heritability estimates

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. 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 based on the relationship between the known genetic correlations of relative pairs and their phenotypic similarity.

Comparison of methods for association on Kosrae using simulated data

Data simulation:

The performance of candidate association methods was evaluated by analyzing simulated datasets constructed from real BMI phenotype data in which no genome-wide significant associations were found (Lowe 2009). For simulation, 1,000 SNPs were randomly selected from Kosraen 500k Affymetrix genotypes and 770 remained after filtering MAF>0.01. Using these SNPs, we generated a total of 770 genome-scale (~350K SNP's each) simulated datasets, in which each phenotype (BMI) was "spiked" to represent a dependency with one of 770 SNPs that contains an effect explaining an additional 2% of phenotypic variance. Each simulated dataset, comprising 2,906 Kosraen individuals, was deconstructed into a pedigree of 586 sibships of size ≥ 2 plus 240 individuals who were less than first cousins (a total of 2,007 individuals) that could be handled by all four methods. The four methods (FBAT, FBAT/Wald, Plink/QFAM-total and EMMAX) were used to perform association for each of the 770 sibship-structured, simulated datasets. Finally, in order to assess the full power of the EMMAX and FBAT/Wald methods, the

simulated datasets were also re-analyzed by these methods, this time including an extra 310 individuals who did not fit into the sibship pedigrees (n= 2,317).

Association methods:

We performed comparative analysis for three different association approaches: a within-family test vs. a within- and quasi-between- family test vs. a within- and between- family test. FBAT (version 2.0.2c) (Lange 2007) was chosen to represent a within-family test, QFAM-total procedure in PLINK software (plink version 1.05) (Purcell 2007) represented a within and partial between family model that utilizes a special permutation procedure to account for family structure. And we selected EMMAX (pre-release beta version) for representing a within- and between-family test, which handles all the relatedness of the cohort (Kang 2010). As both FBAT and QFAM/PLINK do not support association mapping in complex pedigree structures, the pedigree was broken down into sibships (siblings without parents) as described previously (Lowe 2009) and these sibships were used for testing each association method. We tested each representative tool under additive models and with moderate parameters to optimize its behavior of association mapping. For FBAT, we set *minsize* to 3, which is the required parameter of minimum size of family to include for analysis, without loss of any data and biallelic test under an additive model was performed with the default settings. QFAM was run under the PLINK framework with the parameter of *qfam-total,* and the *aperm* parameter was

also included for adaptive permutation. EMMAX analysis was performed by constructing an IBS-kinship matrix and running association on default settings.

Association performance:

We evaluated the performance of each method by two metrics. First, a rank-based score measured the p-value rank of the ground truth SNP in its own dataset. We compared the aggregate of rank of ground truth SNP for the three different association methods and considered the one that assigns more high ranks to ground truth SNPs to be the more powerful method. The second metric was a p-value-based score, which measures the p-value of the ground truth SNP in its respective dataset. With the sum of ground truth SNPs exceeding a particular p-value threshold, we alternatively compared the power of each method.

Association mapping of Kosraen traits using EMMAX

EMMAX was selected for association study of 31 Kosraen traits where low frequency genotypes (MAF>0.01) were removed. A kinship matrix which captured the relatedness between all pairs of phenotyped individuals for each trait was incorporated into the EMMAX mixed model to test for association. Minimal score inflation was seen in the nominal p-values for all traits.

Since there is a high degree of linkage disequilibrium in our data, we determined a genome-wide significance threshold by first extrapolating an approximate testing burden from the trait data. The median minima p-value from 14 null traits was $1.6x10^{-6}$, indicating an estimated testing burden of ~310,000 (actual number of tests was ~386,000). The genome-wide significance threshold for to account for multiple testing for each trait was determined by Bonferroni correction based on the estimated testing burden, i.e. $0.05/310{,}000 = p \leq 1.6x10^{-7}$. A study-wide significance threshold that also accounts for testing multiple traits in the same study was determined conservatively using the Bonferroni correction: $1.6x10^{-7}/31 = p \leq 5.16x10^{-9}$.

For very strongly associated signal peaks $\leq 10^{-20}$ (such as those for URT, PPS, LPA), we derive a more conservative empirical genome-wide significance threshold. The empirical genome-wide significance threshold was calculated by the following steps: (i) phenotypes were randomly permuted between individuals 1,000 times, accounting for family structure, in order to generate 1,000 permuted genotype/phenotype datasets (ii) genome-wide association analysis was performed for each permuted dataset, and the minimum p-value recorded, (iii) the first percentile of the 1,000 genome-wide permuted minima was set as the empirical genome-wide significance threshold (1.4×10^{-10}). We considered a nominal p-value observed in the real data to be genome-wide significant if it exceeded the empirical genome-wide significance threshold.

Identity-by-descent haplotype mapping in Kosrae

Identifying local haplotypes:

Genotypes for the entire Kosraean population were phased using the BEAGLE-trio algorithm in one run according to the recommended parameters: 1 sample, 10 iterations (Browning 2008). An 11 Mb region underlying the genome scan signal peak (chr2:36-47MB) was excised for each individual and analyzed for pairwise IBD matching by the GERMLINE algorithm (Gusev 2008). GERMLINE first identifies short "seed" regions of 30 SNPs that are identical across subsets of the population. Pairs of matching individuals at a seed are then extended across flanking regions that are nearly-identical. We registered haplotypes whenever pairwise comparison revealed a window of allele-call identity at least 500 kb in length with up to 1% mismatch allowed for genotyping error

Haplotype clustering criteria:

Moving from the pairwise IBD segments generated by GERMLINE to a comprehensive set of unique haplotypes required a novel clustering approach. We clustered in two ways: first we identified sets of individuals that may have any amount of common sharing within a region of interest; then we mapped the specific sharing boundary positions

among individuals of interest. The first clustering methodology borrowed a graph-theory approach for finding connected components. We constructed a graph where each individual is represented as a vertex and a shared IBD segment across two individuals in the region of interest forms a simple undirected edge between their respective vertices. A connected component is a sub-graph in which any two nodes are connected by a path; in our IBD graph, a maximally connected component represented all individuals that shared a common haplotype. We used this approach to seek out sets of individuals sharing common haplotypes whose mean phenotype deviated significantly from the expected, designated "affected individuals". We devised a second clustering methodology to identified the specific physical boundaries of the unique region that is shared among multiple individuals from the pairwise sharing data. Conceptually, this algorithm iteratively layered pairwise shared segments between common individuals while keeping track of the physical positions where each individual initiates or ends sharing with the haplotype. Upon analyzing all pairwise shared segments, the algorithm generated, for each marker along the region of interest, the sets of individuals that share a common haplotype at that marker. These results were used to identify any region which was shared exclusively by affected individuals.

PPS signal peak fine-mapping

An 11 Mb region underlying the genome scan signal peak (chr2:36-47MB) was excised for each individual and analyzed for pairwise IBD matching by the GERMLINE

algorithm (Gusev 2008). All 13 exons of the ABCG5 and ABCG8 genes including their splicing donor/acceptor flanking sequences were PCR amplified and sequenced using the Applied Biosystems 3730xl sequencer. Sequencing chromatographs were analyzed by using the FinchTV 1.4.0 and DNASTAR Lasergene 8 software.

URT signal peak fine-mapping

93,629 distinct haplotype clusters were identified after filtering for haplotype clusters of ≥0.01 frequency ≥500kb in length. The haplotype clusters were associated to uric acid levels using EMMAX (see above) and a 13 genome-wide significant haplotypes spannning chr11:36.2-94.2MB emerged ($p<8.52x10^{-8}$), with a chr11: 63.4-65.2 sub-region being most significantly associated ($p<9.06x10^{-46}$) .

Ashkenazi Jewish study population collection

Genome-scale genotype data was collected for 3,573 individuals of self reported Ashkenazi Jewish (AJ) descent from seven different centers in New York, NY and Tel Aviv, Israel. The data consisted of called genotypes assayed on either the Illumina 300k, 550K or IM or Affymetrix 500k or 6.0 platforms. The combined sample contains 966, 449, 554, 101 and 297 individuals who have been diagnosed with Crohn's disease, Parkinson's disease, Schizophrenia, Dystonia and Non-Insulin dependent Diabetes,

respectively. The sample also contains 1,680 “control” individuals for each of these diseases.

Constructing a combined AJ dataset for genome mapping

Combining all AJ datasets:

All datasets were converted to PLINK format and low quality genotypes (missingness >0.02) or individuals with missing genotypes (missingness >0.02) were excluded (--missing flag). Also excluded were genotypes with minor allele frequency <0.01 (--geno flag). Ambiguous SNP’s (those that are A/T or C/G) were removed from individuals assayed on Affymetrix platform. If gender information was missing, it was inferred using PLINK (--impute-sex flag) where the X chromosome SNPs were provided. Duplicate samples within and between datasets were detected (n=9) using PLINK (--genome flag) and one of the pairs was removed. Homozygous heterozygote SNP’s were detected and set to missing. Finally the eight datasets were merged in PLINK (--merge flag) to yield a combined dataset of partial genotyping across the Affymetrix and Illumina platforms for 3,564 individuals. AJ ancestry for all samples was assessed by principal components analysis across ~23K unlinked SNPs in the sample, including Jewish Hapmap AJ samples (n=34), Hapmap CEU samples (n=164), and other non-AJ Jewish or middle eastern samples from the Jewish HapMap and Human Genome Diversity panel (Ostrer

2010). ~600 individuals were determined to have significant non-AJ ancestry and were excluded from the analysis.

Constructing an AJ reference panel for imputation:

We constructed an AJ reference panel for 100 AJ individuals who had been typed on both the Affymetrix 6.0 and Illumina 1M platforms, resulting in 1,778,360 genotyped SNP's for these individuals. Genotypes with >0.02 missingnes, homozygous heterozyotes and low frequency SNPs (MAF <0.02) were excluded. Ambiguous SNP's (A/T or C/G) on the Affymetrix platform were also excluded. We attempted to recover these by BEAGLE (Browning 2008) imputation, accepting on high quality imputed genotypes (r^2>0.9). Finally, 2 individuals who were discordant >0.008 were also excluded from the reference panel. The final panel consisted of 1,328,536 genotypes in 98 individuals.

Imputing missing genotypes in the combined AJ dataset using the reference panel

Missing genotypes in the combined AJ dataset were imputed from the constructed AJ reference panel using BEAGLE. Genotypes that were either poorly imputed (r^2<0.9) or had low frequency (MAF <0.01) were excluded from the subsequent analysis. The final dataset contained 1,068,161 SNP's in 2,860 individuals.

Mapping of AJ traits using EMMAX

Association for Crohn's disease (CD) was performed using EMMAX as described above. In each case, the case's for CD (n=781) were compared to a combination of non-disease controls and disease non-case controls (n=2,079 and n=2,411, respectively). To control for strong genetic effects, the LRRK2 region was excluded from the CD analysis. To account for batch effects between different platforms and/or datasets the association between every pair of datasets (partitioned by case and control) were analyzed and significant SNP's ($<10^{-4}$) were excluded (n=7,227).

Computation

All analysis of the simulated datasets and real data were performed on a 3.0 GHz Intel Xeon dual core 64-bit cluster containing 100 nodes (with 4 processors each), where each node had 8-16GB of RAM.

Chapter 3 Systematic haplotype analysis resolves a complex PPS locus

Introduction

Rapid technological advances in genomics over the last few years have focused on identifying common genetic variants affecting complex disease risk. To date genome-wide association studies (GWAS) have reproducibly implicated over 700 genomic regions that modify the risk for over 120 complex diseases and health-related phenotypes (Kruglyak, 2008, Manolio, 2008) . However, identifying the culprit causal variant(s) underlying these local peaks in association signals remains a challenging task with only a handful of causal variants identified so far (Burnett, 2008, Moffatt, 2007, Burkhardt, 2008). One conventional approach is sequencing a region of arbitrary length around the signal peak across enough individuals to capture the causal allele. However, the larger the sequenced interval and number of individuals, the more variants will be discovered making it difficult to distinguish the driver allele from the large number of passenger variants. This problem is exacerbated in cases where multiple, seemingly independent signals reside in close proximity along the genome. Recent examples of multiple alleles include the 8q24 region for prostate cancer(Haiman, 2007, Yeager, 2007, Gudmundsson, 2007), 6p21 region for HIV-1 viral setpoint (Fellay, 2007), the IL23R, 6p25 and 17q21 regions for Crohn's disease (Duerr, 2006, Barrett, 2008), the PNPLA3 region and nonalcoholic fatty liver disease (Romeo, 2008) the IRF5, STAT4 and TNFAIP3 regions

for systemic lupus erythematosis (Graham, 2007, Graham, 2008, Abelson, 2009) and the 6q23 region for rheumatoid arthritis (Plenge, 2007).

We introduce a method for exposing a causal variant in a region of a genome scan signal that begins with the identification of the full set of haplotypes underlying the signal peak. We assume that a causal derived allele enters a population as a mutation on the background of an ancestral chromosome or haplotype and, under an infinite sites model, that the allele is likely to have mutated only once on that unique ancestral haplotype (Gabriel, 2002, Watterson, 1975). Haplotypes are subject to decay over time by recombination events that occur with each generational meiosis, however mutations that have occurred relatively recently may still be observable as common haplotypes. If individuals in a population harbor a causal derived allele inherited from a recent progenitor, they likely also share a very long segment of the ancestral DNA around the allele. We previously developed the Genetic Error-tolerant Regional Matching with LINear-time Extension (GERMLINE) algorithm (Gusev 2008) which performs pairwise identity-by-descent matching to identify long shared genomic segments in a population. Here, we combine GERMLINE with a novel clustering algorithm which groups similar shared haplotypes to identify any common co-inherited haplotype that might associate to a signal peak. By selecting long co-inherited haplotypes our approach avoids the confounding effects of long range linkage disequilibrium faced by methods based on shorter haplotypes (de Bakker, 2005, Durrant, 2004, Horvath, 2004, Schaid, 2002,

Verzilli, 2006) and a reduced multiple test burden relative to other similar methods (Lin, 2004, Tregouet, 2009, Laramie, 2007, Purcell, 2007, Albrechtsen, 2009).

We tested our method in an extreme isolate founder population where we had previously shown abundant long haplotype segments shared between individuals (Gusev, 2008, Bonnen, 2006). The Pacific Island of Kosrae, in the Federated States of Micronesia, was likely settled 2,500-1,500 years ago by individuals of Asian ancestry and Islanders have a high incidence of obesity and diabetes (Gray, 2009, Friedlaender, 2008, Kayser, 2008, Shmulewitz, 2006, Lowe, 2009). We focused on plasma plant sterol (PPS) levels, which were ascertained as part of a broader study to examine genetic causes of the metabolic syndrome on the island (Lowe, 2009).

Plant sterols are a dietary source of neutral sterols that are structurally similar to cholesterol (Ostlund, 2002, Heinemann, 1993). Studies of the rare disorder phytosterolemia, which is characterized by extremely high PPS and severe premature atherosclerosis, have implicated mutations in the ATP binding cassette subfamily G members 5 or 8 (ABCG5 or ABCG8) genes (Berge, 2000, Hubacek, 2001). The ABCG5/ABCG8 genes form an obligate heterodimer that has been shown to be involved in intestinal absorption and biliary excretion of neutral sterols, including plant sterols and cholesterol (Graf, 2003). As such, PPS are considered to be a biomarker for dietary cholesterol absorption and changes in the balance of cholesterol absorption and secretion have been suggested to be atherogenic (Rudkowska, 2008, Sudhop, 2002). Further, it is unclear whether moderately elevated PPS may themselves affect cardiovascular risk

(Silbernagel, 2009, Fassbender, 2008, Chan, 2006). We had previously reported that 13.8% of Kosraens are carriers of an ABCG8 exon 2 frameshift mutation leading to a nonsense ABCG8 codon (nonsense ABCG8 mutation) which results in a premature truncation and non-functional protein. The ABCG8 nonsense mutation was shown to effect a 30%-50% increase in plasma levels of campesterol and sitosterol, the two most abundant plant sterols in plasma (Sehayek, 2004).

Here, we validated the ABCG8 exon 2 nonsense mutation effect in a larger Kosraean cohort of ~3,000 individuals and performed an association analysis of PPS levels, which implicated a second strong signal at the same locus. We analyzed identity-by-descent (IBD) shared genetic segments using the GERMLINE software to dissect the full set of long-range haplotypes in this region and identified a distinctive 526 kb haplotype, independent of the ABCG8 nonsense mutation, which is carried by 1.8% of Kosraens and associated with a striking 100% increase in PPS levels. Sequencing of the ABCG5/ABCG8 genes in carriers of the 526 kb haplotype revealed a novel ABCG5 D450H missense mutation, a plausible putative causal variant in this haplotype. These findings exemplify the power of haplotype analysis in resolving the effect of multiple variants of the same locus.

Results

Analysis of multiple signals on Chr2p21 for PPS

We performed a GWAS in 1,423 related individuals (who were phenotyped for PPS levels) on the Pacific Island of Kosrae with genotypes from the Affymetrix 500k array and incorporating the ABCG8 nonsense mutation we had separately genotyped. The analysis revealed 48 SNPs, all on chromosome 2p21, that surpassed an empirical permutation-based threshold (see methods for details) of genome-wide significance at a nominal p-value of 1×10^{-10} (**Figure 3.1(A)**). The strongest signal confirmed the association of the ABCG8 nonsense mutation ($p<5\times10^{-39}$) to PPS levels and validated our previous finding in a subset of this cohort (Sehayek, 2004). Closer examination of the signal at chromosome 2p21 revealed a broad signal peak with genome-wide significant signals for PPS extending up to 10 Mb upstream and 1 Mb downstream of the ABCG5/ABCG8 locus (**Figure 3.1(B)**). To determine whether this extended signal was entirely accounted for by the ABCG8 nonsense mutation, we reanalyzed the association of the PPS phenotype conditioned on the ABCG8 nonsense mutation genotype (**Figure 3.1(C)**).

Multiple strong signals that exceeded the genome-wide significance threshold up- and downstream of the ABCG8 locus persisted and suggested the presence of additional independent signals at this locus. The strongest remaining signal was associated with

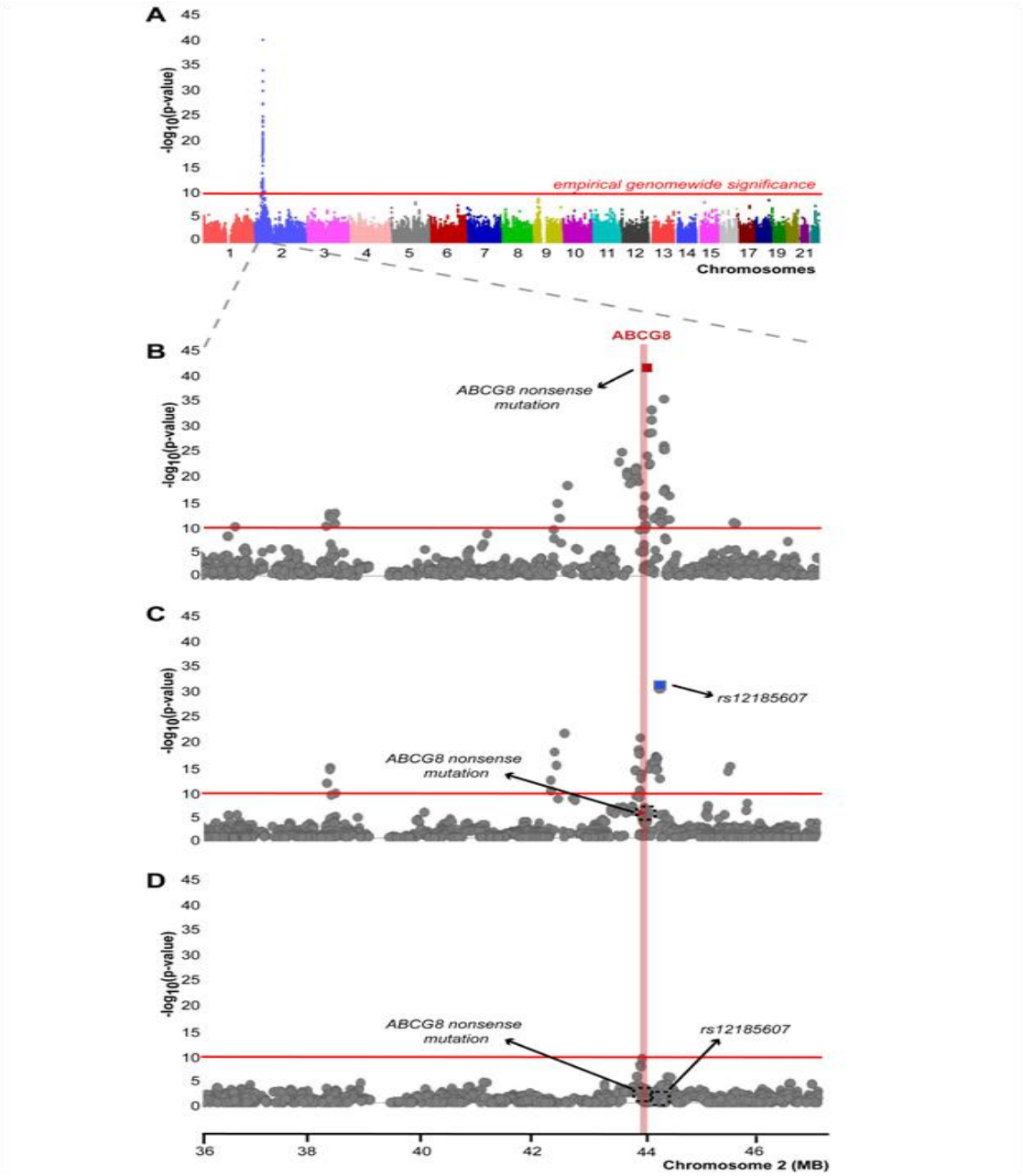


Figure 3.1 Stepwise conditional analysis of a genome-wide plasma plant sterol (PPS) levels: (A) Unconditioned genome-wide association of PPS. Association analysis was performed as described in the methods section. The red line marks the threshold of empirical genome-wide significance. (B) Unconditioned association of PPS in a 11Mb interval on chromosome 2 surrounding the genome-wide peak signal and the ABCG5/ABCG8 genes. Maroon square indicates the ABCG8 nonsense mutation with the best signal. Red shaded area marks the ABCG8 locus. (C) Association of PPS in the same chromosome 2 interval after conditioning for the ABCG8 nonsense mutation. Blue square indicates the SNP (rs12185607) with the best signal. (D) Association of PPS in the same chromosome 2 interval after conditioning for both the ABCG8 nonsense mutation and rs12185607 genotypes.

		rs12185607	
		GG	**GT**
ABCG8 nonsense mutation	**GG**	2512	64
	GC	321	2
	CC	5	0

Table 3.1 Linkage equilibrium between the ABCG8 nonsense mutation and rs12185607/G allele

rs12185607 ($p<3\times10^{-31}$) ~250 kb downstream of the ABCG5/ABCG8 locus. To examine the possibility of additional independent signals, we analyzed the association of PPS phenotype conditioned on both the ABCG8 nonsense mutation and rs12185607.

Conditioning of PPS on both variants abolished statistically significant signals across the genome, as shown in **Figure 3.1(D)**, and specifically at the chromosome 2p21 interval. We observed that the effects of the ABCG8 nonsense mutation and rs12185607/G were in the same direction and the minor alleles were the risk alleles in both cases. Given the carrier rates of 11.1% and 1.8% of the ABCG8 nonsense mutation and the rs121185607/G minor allele, respectively, we expected to find 5-6 Islanders carrying both alleles. We observed only two individuals (p=0.076 assuming independence) who were carriers of both minor alleles. Therefore, our findings suggested the presence of two independent variants effecting PPS at the chromosome 2p21 locus.

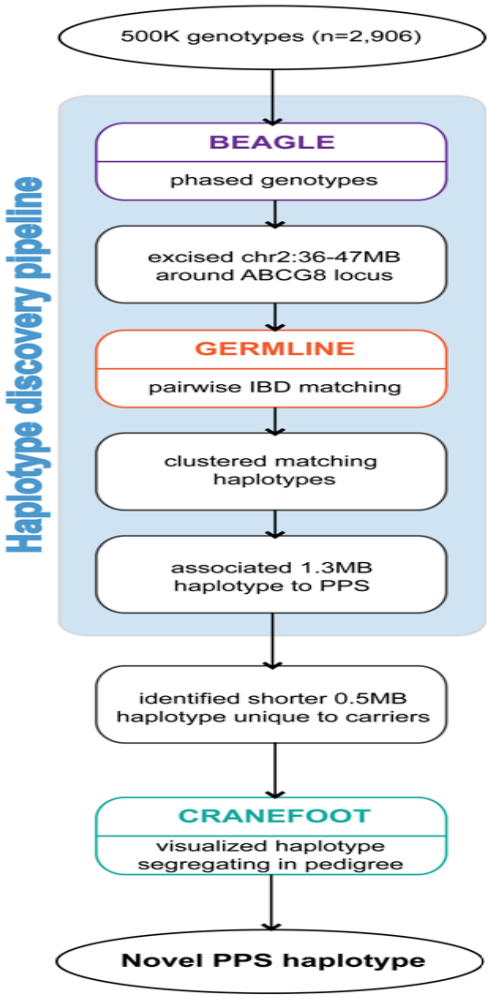


Figure 3.2 **Schema of the analysis pipeline for novel PPS haplotype discovery, refinement and validation.** BEAGLE-trio phased haplotypes for the chromosome 2 region (36-47MB) were excised and analyzed with GERMLINE to detect matching pairwise IBD segments. Overlapping segments that matched across the population were clustered and these the mean phenotype of cluster members was assessed. One cluster containing a 1.3MB haplotype strongly associated to high PPS levels. Careful comparison of haplotypes in that cluster revealed a unique ~526kb shared segment carried by 52 individuals. The CRANEFOOT software was used to visualize the ~526kb haplotype segregating in three closely related kindreds.

Systematic identification of a novel PPS haplotype at chr2p21

To capture the second causal allele on chromosome 2p21 we devised an analysis pipeline to systematically identify all common haplotypes in the region (see **Figure 3.2**). The 11 Mb signal region on chromosome 2p21 (chr2:36-47MB) around the ABCG5/ABCG8 locus was excised and common haplotypes were identified by IBD matching using GERMLINE (Gusev, 2009). The IBD segments were at least 500 kb in length and allowed for up to 1% mismatching due to genotyping error. Next, we clustered groups of similar haplotypes using similarity measures and binned together sets of individuals that had common sharing within the 2p21 region. These efforts identified 3,681,003 pairs of individuals that shared one or more IBD matching segments >500

kb in length. These matched pairs clustered to 233 bins of common sharing with >99% sequence identity (see **Figure 3.3**). We found a single cluster containing 44 individuals that shared a 1.3 Mb haplotype strongly associated with PPS levels ($p<4x10^{-65}$). This haplotype was shared by eight additional individuals who were not phenotyped for PPS levels. We mapped the specific sharing boundaries of the haplotype to a 526 kb sub-region (chr2:43.8-44.3 Mb) that was unique to this group of 52 individuals (see **Figure 3.4(A)**). Fourteen individuals (six phenotyped), who were carriers of rs12185607/G but were not on the background of the 526 kb haplotype, had significantly lower PPS levels ($p=5.2 \times 10^{-4}$).

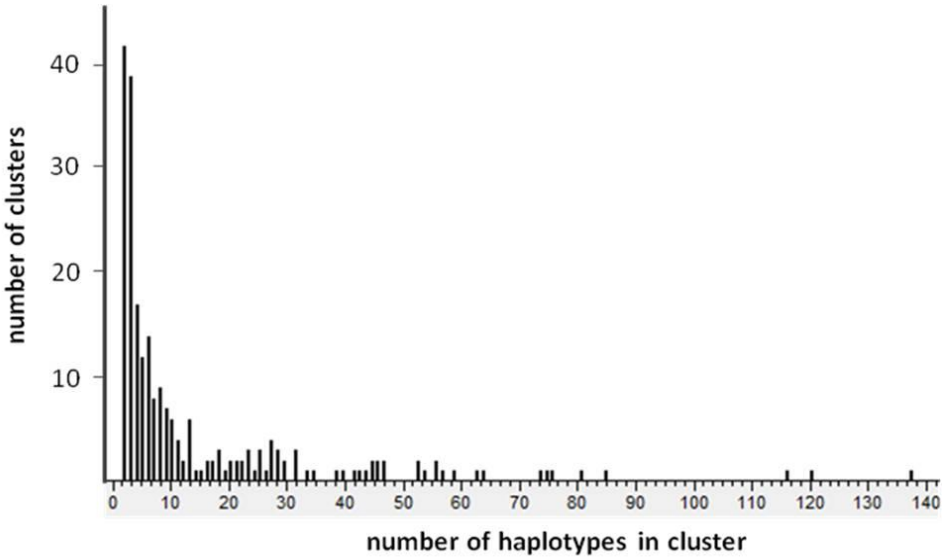


Figure 3.3 **Histogram of haplotype cluster sizes.** The x-axis shows the haplotypes binned by the number of haplotypes in each cluster and the number of clusters for each bin is shown on the y-axis.

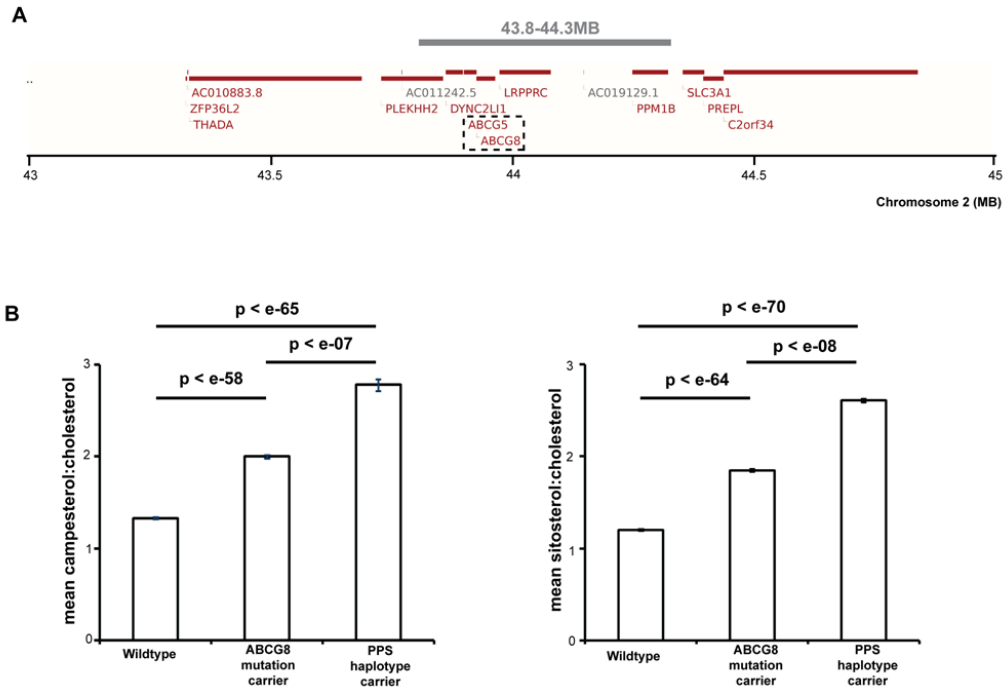


Figure 3.4**(A) Annotated genes on the background of the 526 kb haplotype.** The novel PPS haplotype (indicated as a grey bar) extends over a 526kb region on chromosome 2p21 that includes the ABCG8 and ABCG5 candidate genes (indicated by a black dashed square) and also fully or partially 5 other annotated loci (Ensembl Homo sapiens version 54.36p (NCBI 36) Chromosome 2 : 43,000,000 .. 45,000,000 [54]).(B) Effect of the ABCG8 nonsense mutation and 526 kb haplotype carrier state on PPS. Fasting plasma levels of campesterol and sitosterol were determined as described in the Methods section. Values shown are Mean ± SEM.

Clustering of the novel PPS haplotype in three Kosraean pedigrees

We investigated the family structure of the 52 carriers of the 526 kb haplotype. This haplotype segregated in three kindreds containing 110 genotyped individuals from the same village, and was present in one unrelated individual (see **Figure 3.5**). There were two carriers of both the 526 kb haplotype and the ABCG8 nonsense mutation, each

inherited from a different parent. We performed admixture analysis, using the ANCESTRYMAP algorithm (Patterson, 2004) trained on four Hapmap phase III reference panels (CEU, CHB, JPT, GIH) to examine the ancestral origin of the 526 kb haplotype, and found no evidence of admixture in the region.

The three kindreds segregating the PPS haplotype came from the same village, opening the possibility that a shared local environment or other non-genetic factor might cause increased PPS levels in the whole village and, hence, false positive association signals among this sub-population. To control for any confounding local environment in the village, we coded PPS haplotype carrier status as an allele (1=carrier, 0=non-carrier) and performed association analysis with PPS levels. 89 phenotyped villagers who fell into 39 nuclear families were analyzed with PLINK/QFAM and the PPS haplotype remained significantly associated with PPS levels in this sub-group with a permuted p-value$<1.6 \times 10^{-4}$.

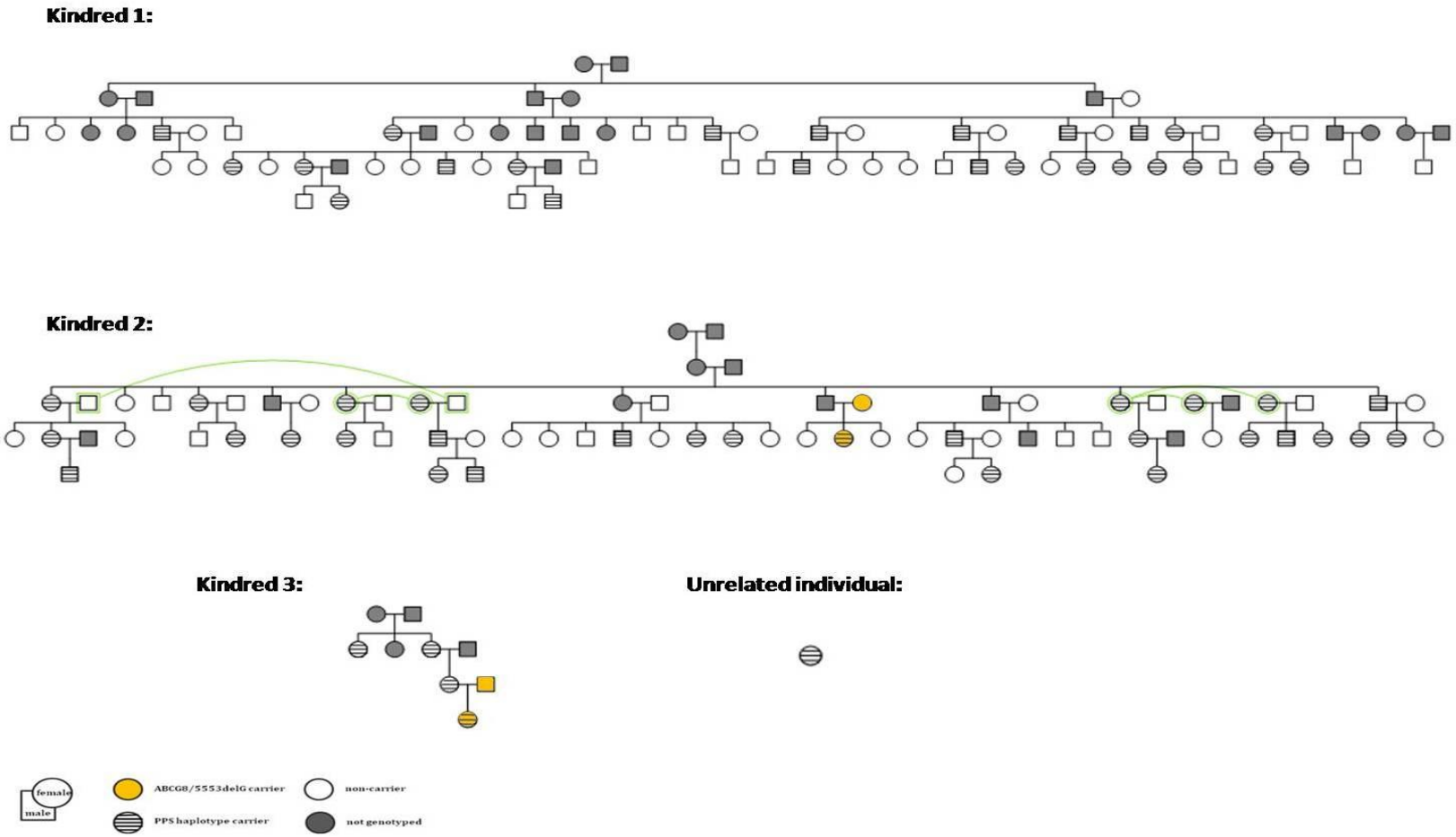


Figure 3.5 **Segregation of PPS haplotype in three kindreds from one village.** The pedigree structure of 146 individual forming three kindreds and one unrelated individual from the same village on Kosrae. PPS haplotype carriers are shown in horizontal strips. Non-carriers are shown in white. 36 non-genotyped individuals (shown in solid grey) are include to preserve the kindred structure.

Phenotypic effect of the 526 kb haplotype

Previous analysis in a subset of the Kosraen cohort revealed that carriers of the ABCG8 nonsense mutation are characterized by a 30-50% increase in plasma campesterol and sitosterol levels, but no alteration in plasma cholesterol levels (Sehayek, 2004). Here, we have examined the effect of the ABCG8 nonsense mutation and the 526 kb haplotype on PPS and plasma lipid levels in the entire cohort. As shown in **Figure 3.4(B)**, carriers of the ABCG8 nonsense mutation had a 50-54% increase in mean plasma campesterol (2.00 ± 0.87 vs. 1.33 ± 0.52) and sitosterol (1.85 ± 0.81 vs. 1.20 ± 0.48) levels compared to non-carriers (in both cases $p<0.0001$). Also shown in **Figure 3.3(B)**, carriers of the 526 kb haplotype had a 109-117% increase in mean plasma campesterol (2.78 ± 1.01 vs. 1.33 ± 0.52) and sitosterol (2.61 ± 1.00 vs. 1.20 ± 0.48) levels compared to non-carriers (in both cases $p<0.0001$).

	Total Population	Non-carriers	ABCG8 mutation carriers	PPS haplotype carriers	Non-carriers vs ABCG8 mutation carriers	Non-carriers vs PPS haplotype carriers	Non-carriers vs ABCG8 carriers vs PPS carriers
Lath:Chol	1.50 ± 0.66 (1900)	1.52 ± 0.67 (1687)	1.33 ± 0.55 (213)	0.84 ± 0.38 (44)	1.08×10^{-4}	3.68×10^{-11}	<0.0001
TC (mg/dL)	166.09 ± 34.43 (2819)	165.48 ± 34.35 (2452)	170.2 ± 34.48 (315)	167.55 ± 34.92 (52)	0.022	N.S.	N.S
LDL-c (mg/dL)	102.1 ± 29.54 (1899)	101.63 ± 29.3 (1645)	105.42 ± 30.63 (210)	100.84 ± 30.73 (44)	N.S.	N.S.	N.S
ApoB (mg/dL)	86.97 ± 21.32 (1884)	86.65 ± 21.31 (1621)	88.53 ± 21.64 (219)	89.7 ± 19.8 (35)	N.S.	N.S.	N.S
HDL-c (mg/dL)	38.28 ± 11.04 (1899)	38.09 ± 10.95 (1645)	38.77 ± 11.36 (210)	42.00 ± 9.93 (44)	N.S.	N.S.	N.S
ApoA1 (mg/dL)	116.87 ± 24.39 (1875)	116.64 ± 24.07 (1621)	118.07 ± 26.5 (219)	122.21 ± 25.3 (35)	N.S.	N.S.	N.S
TG (mg/dL)	99.94 ± 46.16 (2817)	100.94 ± 45.91 (2450)	95.58 ± 39.02 (315)	93.26 ± 36.32 (52)	0.05	N.S.	N.S

Table 3.2 **Lathosterol:cholesterol and other plasma lipid levels.** ANOVA analysis of lathosterol:cholesterol and plasma lipid levels comparing carriers of the ABCG8 mutation and 526 kb haplotype and individuals who are wildtype for both mutations. Values are shown as the mean ± standard deviation, with the number of individuals shown in brackets. **N.S.:** not significant, **Lath:Chol:** lathosterol/cholesterol ratio, **TC:** total cholesterol, **LDL-c:** low density lipoprotein cholesterol, **HDL-c:** high density lipoprotein cholesterol, **TG:** triglycerides, **ApoA1:** Apolipoprotein-A1, **ApoB:** Apolipoprotein B.

The effects of the ABCG8 nonsense mutation and the 526 kb haplotype on plasma lipid levels were next examined. As shown in **Table 3.1**, the only significant finding was a modest increase in total cholesterol levels in the ABCG8 nonsense mutation carriers. The effect of these two variants in the ABCG5/ABCG8 region on plasma lathosterol levels was also assessed. Plasma lathosterol levels have been shown to be proportional to whole body cholesterol synthesis. Also shown in **Table 3.1**, carriers of the ABCG8 nonsense mutation had a 12.5% decrease in plasma lathosterol (standardized to cholesterol and shown as a ratio of lathosterol:cholesterol) (1.33 ± 0.55) and carriers of the 526 kb haplotype a 45% decrease in plasma lathosterol (0.84 ± 0.38) compared to non-carriers (1.52 ± 0.67) (in both cases $p<0.0001$).

Sequencing reveals a novel putative ABCG5 causal variant

The 526 kb haplotype region on chromosome 2p21 encompasses both ABCG5 and ABCG8 and six other annotated genes (see **Figure 3.6**). ABCG5 and ABCG8 are obvious candidate genes to explain the haplotype effect. We first sequenced all 13 exons and adjacent intronic splice donor/acceptor site flanking sequences of these genes in four pairs of 526 kb haplotype carriers and non-carrier related controls. As shown in **Figure 3.6**, this effort revealed a novel ABCG5 exon 10 missense mutation resulting in a coding

change from negatively charged aspartic acid to positively charged histidine at residue 450 (D450H). This aspartic acid is highly conserved in vertebrates, suggesting a functional role that might be subverted by the histidine substitution (see **Figure 3.6**). Finally, to confirm association of the 526 kb haplotype to the ABCG5 missense mutation we sequenced exon 10 of ABCG5 in the remaining 48 carriers, 10 non-carrier controls,

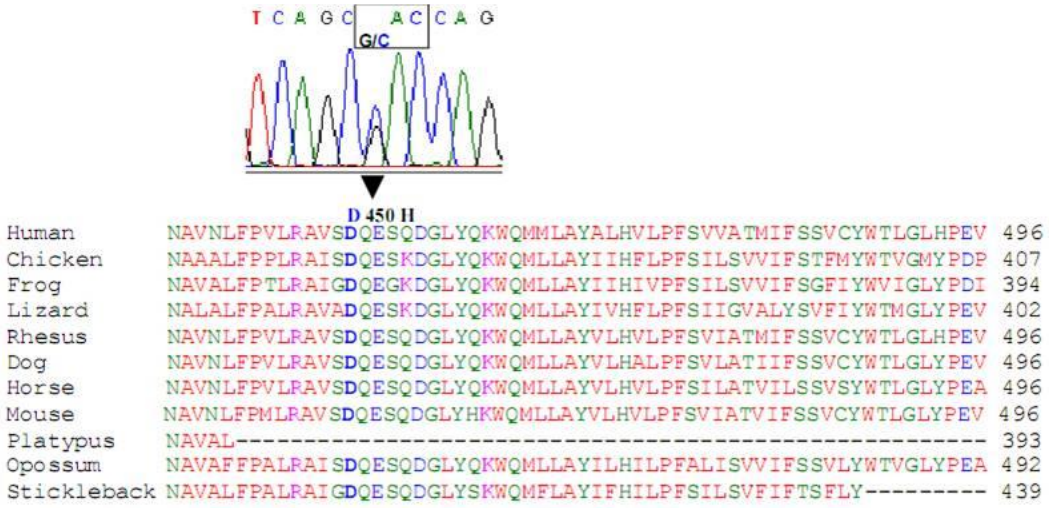


Figure 3.6 **D450H missense mutation in exon 10 of the ABCG5 gene.** The sequencing chromatograph showing the G to C mutation that results in a coding change from aspartic acid (D) to histidine (H) at codon 450 in exon 10 of the ABCG5 gene. The aspartic acid is highly conserved in vertebrates.

and 13 carriers of rs12185607/G but not the 526 kb haplotype. The ABCG5 missense mutation was only present in the 526 kb haplotype carriers; thus establishing complete linkage disequilibrium between the 526 kb haplotype and the ABCG5 missense mutation.

Discussion

We present the first implementation of a novel approach for leveraging common co-inherited haplotypes under the local peak of a genome scan to expedite the search for the culprit causal variant. We analyzed a strong genetic trait, plasma plant sterol levels, in an extreme isolate population known to segregate a defined ABCG8 nonsense mutation on chromosome 2p21. Our initial GWAS analysis revealed evidence for multiple independent signals in the chr2p21 region. Systematic dissection of common haplotypes in that region using the GERMLINE software identified a single common haplotype that harbored a second PPS signal. Re-sequencing the exons of two candidate genes in the haplotype, ABCG5 and ABCG8, revealed a plausible causal genetic variant.

The advantage of the IBD approach is that long shared segments bearing causal variants are unlikely to be shared by chance, thereby increasing their power for detection over traditional association mapping. Furthermore, this approach exactly identifies carriers and can be used to dissect multiple signals at the same locus. In the current study it was possible to detect a relatively rare (minor allele frequency < 1%) causal genetic variant. Since IBD haplotypes were constructed from SNP-chip data, this approach can complement a traditional association mapping method. However, the method is only suited to discover rare, relatively newly arisen or bottlenecked causal mutations that retain enough surrounding shared DNA to be detectable. If a causal mutation arose more

distantly and surrounding haplotypes are too short the method will not have any mapping power. For example, the ABCG8 nonsense mutation was entirely described by a short ~4kb haplotype and was not detectable by this method. The IBD approach to search for a causal allele affecting PPS was aided by the high degree of relatedness and founder effects on Kosrae that resulted in abundant, long haplotypes in our study population. However, recently tracks of relatedness in regions of the genome have been seen in less extreme founder populations (Kong, 2008), and even in apparently unrelated populations (Frazer, 2007). This has motivated several groups to develop IBD methods as a means of mapping disease-related alleles (Gusev, 2009, Purcell, 2007, Albrechtsen, 2009).

The unexpectedly large phenotypic effect of the 526 kb haplotype on PPS levels raises intriguing questions regarding the underlying mechanism. Given the evolutionary conservation of the aspartic acid residue in position 450 of ABCG5 and the nonconservative nature of the amino acid substitution, this variation is highly likely to be functional. It is unclear, however, how a missense mutation in ABCG5 elevates PPS levels ~100%, whereas an ABCG8 nonsense mutation, presumably yielding no full length protein, only elevates PPS levels ~50% (see **Figure 3.3**). It has been shown that ABCG5 functions by heterodimerization with ABCG8 (Graf, 2003). Therefore, we hypothesize that the greater effect of the missense mutation is due to one of the following: (i) imbalanced expression of a non-functional ABCG5-mutant allele, (ii) a

gain-of-function dominant-negative ABCG5 mutation, or (iii) linkage disequilibrium of a non-functional ABCG5-mutant with other genetic variant(s) at the 526 kb haplotype.

Recent well-powered meta-analyses of pooled GWAS in individuals largely of Caucasian ancestry found the ABCG5/ABCG8 locus associated with plasma total and LDL-cholesterol levels (Kathiresan, 2009, Aulchenko, 2009). In the current study we found a modest association of the ABCG8 nonsense mutation with total cholesterol levels ($p<0.02$), but no such association for the 526 kb haplotype containing the ABCG5 missense mutation. This is a paradox as the greater potency of the latter mutation is verified both by a larger increase in PPS levels and a greater decrease in cholesterol synthesis as shown by more of a decrease in plasma lathosterol levels. It is possible that the frequency of the 526 kb haplotype is too low for an effect on plasma cholesterol to be seen. It is also possible, but unlikely, that the 526 kb haplotype contains other variants besides the ABCG5 missense mutation that obscure effects on total and LDL cholesterol, while preserving affects on PPS and plasma lathosterol levels. Finally, the number of ABCG8 nonsense mutation carriers are relatively few in this study, especially compared to population based studies, and the borderline association seen with total cholesterol levels may be a false positive.

Mutations in ABCG5 and ABCG8 are rare in the general population. Homozygotes or compound heterozygotes have the disorder phytosterolemia at an estimated frequency of

<1:1,000,000 (there are only 40 reported cases). It is therefore interesting that we observed by sampling ~3,000 adults on Kosrae a combined carrier rate of 1:8 for the two mutations. The reason for such clustering is not clear, but may suggest a beneficial effect of these mutations or simply a chance observation owing to founder effects.

In summary, we have described a novel approach to identify and dissect the effects of large haplotypes under the signal of a single locus detected by GWAS. This approach allowed identification of a novel 526 kb haplotype that modifies PPS and dissected its effect from an ABCG8 mutant that was known to segregate on the island of Kosrae.

Chapter 4 Comparison of Association Methods in Isolated Populations

Introduction

Isolated populations have a long history in genetic mapping studies of inherited disorders with advantages including reduced environmental, phenotypic and genotypic heterogeneity as compared to outbred populations (Ober, 2001 ,Angius, 2002 ,Peltonen, 1999 ,Gulcher, 2001 ,Freimer, 1996 ,Zamel, 1996 ,Abney, 2000 ,Aulchenko, 2004 ,van der Walt, 2005). In particular, the reduction of genotypic heterogeneity observed in isolated populations due to founder effects, bottlenecks and genetic drift, may allow otherwise rare mutant alleles to rise to higher frequency in these populations, while at the same time narrowing the spectrum of candidate mutations (Ober 2001, Angius 2002). The increased chance for homozygosity has been a key factor in identifying mutations responsible for rare monogenic diseases in isolated populations (Peltonen 1999, Gulcher 2001, Freimer 1996). Here we investigate the advantages and limitations of exploiting homozygosity in isolated populations for analysis of alleles effecting complex traits (Zamel 1996, Abney 2000, Aulchenko 2004). Genome-wide association studies (GWAS) performed in outbred populations have thus far identified many common variation contributing to complex diseases (Manolio 2008). One limitation in these studies is that rare variants that are not in linkage disequilibrium with the common variants assayed are

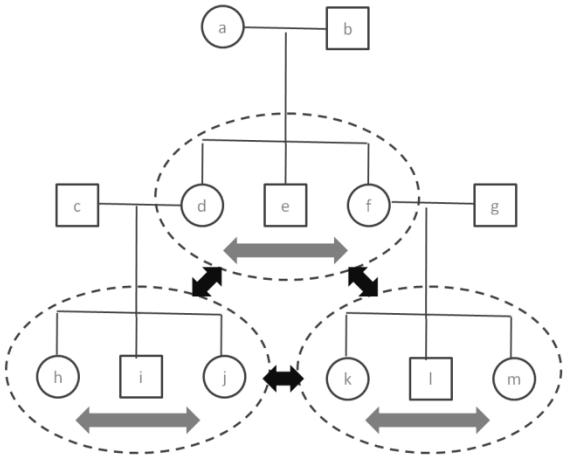
usually not detected (Bodmer, 2008). However, when the populations analyzed in a GWAS have a substantial number of individuals that share a recent common ancestor, not only common variants (Lowe, 2009, Wang, 2009), but sometimes also rare variants affecting complex disorders can be identified (Sabatti, 2009, Pollin, 2008, Kallio, 2009).

Genetic analysis of isolated populations pose unique challenges for traditional GWAS methods that have been mainly focused on outbred populations (Bourgain, 2005). The crux of the difference is that in isolated populations the likelihood of any two individuals in the population to be related is not negligible. The resulting direct and cryptic relatedness confounds assumptions of independence between genotypes of different individuals, as well as between their heritable phenotypes. Isolated populations therefore contain a large amount of cross-individual correlations which pose problems for most mapping methods. Further, consanguinity may be present so that two alleles in a random individual may also be correlated. The hidden correlation in the data can cause an overdispersion of the naïve test scores for association and consequently, lead to false positive associations. Hence, the major challenge for performing GWAS in isolated populations is to account for this non-random intra- and inter-individual correlation.

While standard association tests assume independence of genotypes and phenotypes across samples, several specialized approaches for association do account for underlying relatedness expected in isolated populations. In this work, we set out to select and

evaluate these different methods for maping complex traits in an isolated founder population. Some of these methods rely on knowledge of the underlying family structure in the population. These approaches overcome the confounding effects of non-random correlation via deconstruction of the population into family units and analysis of association independently within each unit (within family variance) (Lange, 2004 ,Abecasis, 2000 ,Almasy, 1998). Some of these "family-based" methods have been extended by adding the variance between families to the within family variance (Devlin, 1999, Purcell, 2007, Abecasis, 2001,Abecasis, 2002 ,Won, 2009). A different type of "population-based" approach does not utilize prior knowledge of family structure, but rather explicitly models the relatedness between all pairs of individuals based on their genotypes, and incorporates this variance into a mixed model for association (Yu, 2006 ,Abecasis, 2001 ,Kang, 2008, Aulchenko, 2007, Amin, 2007, Kang, 2010, Zhang, 2010). Such models have recently been extended to genome-scale human studies and have been shown to effectively control for population structure (Zhang, 2010, Kang, 2010). Here our emphasis is on evaluation of such methods in the context of extensive relatedness in study samples.

A



B

	a	b	c	d	e	f	g	h	I	j	k	l	m
a	1	0	0	0.5	0.5	0.5	0	0.25	0.25	0.25	0.25	0.25	0.25
b	0	1	0	0.5	0.5	0.5	0	0.25	0.25	0.25	0.25	0.25	0.25
c	0	0	1	0	0	0	0	0.5	0.5	0.5	0	0	0
d	0.5	0.5	0	1	0.5	0.5	0	0.5	0.5	0.5	0.25	0.01	0.25
e	0.5	0.5	0	0.5	1	0.5	0	0.02	0.01	0.25	0.25	0.25	0.25
f	0.5	0.5	0	0.5	0.02	1	0	0.25	0.25	0.25	0.5	0.5	0.5
g	0	0	0	0	0.01	0	1	0	0	0	0.5	0.5	0.5
h	0.25	0.25	0.5	0.5	0.25	0.25	0	1	0.5	1	0.25	0.5	0.5
I	0.25	0.25	0.5	0.5	0.25	0.25	0	0.5	1	0.5	0.5	0.5	0
j	0.25	0.25	0.5	0.5	0.25	0.25	0	1	0.5	1	0.5	0.5	0
k	0.25	0.25	0	0.25	0.25	0.5	0.5	0.5	0.5	0.5	1	1	0
l	0.25	0.25	0	0.01	0.25	0.5	0.5	0.5	0.5	0.5	1	1	0
m	0.25	0.25	0	0.25	0.25	0.5	0.5	0.5	0	0.5	0	0	1

Figure 4.1: Schema of the sibship- and kinship matrix- based approaches to association studies in related cohorts. (**A**) The extended Kosrae pedigree is broken into sibships without parents (indicated by the dotted gray circle). Parent-child or cousin relationships may exist between different sibships. Tests of association may be performed within sibships (gray arrows) and between sibships (black arrows). (**B**) A kinship matrix of all pairs of the same individual.

We compared the performance of the four representative methods that account differently for relatedness while testing for genome-wide association: (i) focusing on allele transmission to offspring within families (Lange, 2004), (ii) measuring association within as well as between families (Devlin, 1999 ,Purcell, 2007, Won, 2009), and (iii) capturing the relatedness between all individuals in the population to construct a mixed model to test for association (Kang, 2010). Our simulations showed the mixed model method to have increased statistical power to detect association, offering a 1.8- to 2-fold improvement over the family-based approaches.

Results

Computational performance

Five representative programs for performing association in the context of relatedness (FBAT, FBAT+Wald, PLINK/QFAM-total, EMMAX and SOLAR – see description below) were assessed for speed and maximum memory allocation to determine their suitability for a genome-scale association analysis across thousands of individuals. The average computation time for association per marker for each program is given in **Table 4.1**. FBAT, FBAT+Wald, EMMAX and PLINK/QFAM-total ran at speeds < 0.06

	FBAT	FBAT+Wald	PLINK/QFAM-total	EMMAX	SOLAR
Speed (sec/SNP)	0.013	0.021	0.057	0.039	8.74
Memory (MB ram)	1,660	1,660	760	790	3,200

Table 4.1: Computation time (average computation time per second per SNP) and maximum memory allocation (megabytes of RAM) for each association methods (FBAT, PLINK/QFAM-total, EMMAX and SOLAR).

seconds/SNP, with FBAT performing > 2-fold faster than the other three methods, compared to the SOLAR method which was > 100-fold slower. The slow speed of the SOLAR method prohibited its further inclusion in the study. The maximum virtual memory allocation required by the remaining four methods did not exceed 1.7GB of RAM during the runtime of each program.

Strategy for accuracy and efficacy comparison on known-answer data

We evaluated the performance of different approaches to genome-wide association in the context of relatedness by comparing representative tools for each approach. The family-based association test (FBAT) is a method that considers allele transmission to offspring within families (Lange, 2004). The FBAT group has recently suggested an extension of their method that combines the within-family FBAT score with a rank-based, between-family score, derived from a Wald test of the whole cohort, that is robust to

overdispersion, into one single test statistic (FBAT+Wald), which we also tested (Won, 2009). Plink family-based association test for quantitative traits, within- and between-mode (PLINK/QFAM-total), uses a linear regression-based between/within family approach similar to a model described previously (Abecasis, 2000, Fulker, 1999). In the PLINK/QFAM-total framework, confounding effects of family structure are controlled by independent within- and between-family permutation strategies for estimating exact significance levels, where the empirical within-and between family permuted p-values are combined to a single score. We corrected for any residual overinflation of the test statistic by standard genomic control adjustment (Devlin, 1999, Purcell, 2007). Finally, efficient mixed model association expedited (EMMAX) is a method that first uses high-density genotypes to empirically estimate levels of relatedness between every pair of individuals, which are captured in a kinship matrix. The kinship matrix is then incorporated into a linear mixed model to adjust for correlation in the phenotypic distribution during association mapping (Kang, 2010). We used all four methods to perform genome-wide association analysis in a highly related, population-based cohort from the Island of Kosrae, Micronesia, that is comparable in size (n = 2,906) to published single study case/control cohorts.

In order to assess the empirical power for each association method, we considered data for over 350,000 SNPs, with a minor allele frequency greater than 0.01, genotyped on an Affymetrix 500K platform in the Kosrae samples. We repeatedly analyzed association of

these SNPs to a null, moderately heritable (h^2 = 0.42 on Kosrae) phenotype, body mass index (BMI), which we modified afresh multiple times. Each modified phenotype included a simulated genotype/phenotype association explaining 2% of the phenotypic variance to a different random SNP, additively combined to the true trait value. This effect size was chosen to echo milder effects detected and detectable in larger-sample studies that are now available. These datasets were constructed by selecting 1,000 SNPs randomly across the genome (770 common SNPs after filtering, see methods), where each phenotype was altered to reflect association to a different SNP in the random subset. In total, 770 genome-scale datasets containing modified SNP-phenotype association were generated and analyzed by all four methods.

Following extensive pedigree construction based on genetic and oral information, more than 90% of the genotyped individuals on Kosrae form a single extended pedigree spanning 5+ generations and containing numerous consanguineous offspring and multiple marriage loops (Lowe, 2009). To the best of our knowledge, large pedigrees with such complexity cannot be handled whole by methods that have a within-family component. Therefore, we broke our pedigree into smaller units of sibships-without-parents for the purposes of method comparison. This resulted in 586 sibships consisting of two or more individuals who share a mother and father (**Figure 1A**). Any genotyped parents are considered only in the context of the parents' sibship. Of the individuals not included in any sibship (n = 612), a subset was identified in which any two members of the subset

were related to the degree of first cousins or less, as determined by genome-wide identity-by-descent sharing, resulting in 240 additional "sibships" of size 1. To fairly compare FBAT and PLINK/QFAM-total to EMMAX and FBAT+Wald, we first analyzed only sibship individuals. However, to also quantify the benefits of using the entire dataset, we repeated the analysis examining the entire cohort with EMMAX and the sibships for the FBAT and entire cohort for the Wald components of the FBAT+Wald method.

Empirical power results

The empirical power for each method was recorded in two ways: the reported (i) genome-wide rank (**Figure 4.2A**) and (ii) p-value (**Figure 4.2B**) of the ground-truth SNP according to each method across repeated simulation iterations. While these criteria are equivalent when reported p-values are uniformly distributed (as required by the definition of a p-value), inclusion of both criteria facilitates evaluation of potential bias in such distributions. Comparison of the within-family only (FBAT) vs. combined within- and between-family (PLINK/QFAM-total and FBAT+Wald) vs. mixed model (EMMAX) tests demonstrated that the greatest power, as measured by rank order of the true effect, was obtained by the use of the mixed model method. EMMAX had 88% power to rank the true SNP in the top 10 genome-wide, a 1.7-, 2- and 2.1-fold improvement over PLINK/QFAM-total, FBAT and FBAT+Wald, respectively (**Figure 4.2A**). In addition,

56% of truly associated SNPs achieved genome-wide significance (gws) for the mixed model method compared to < 16% for the other methods (**Figure 4.2B** – see discussion). Although, the inclusion of the entire cohort (an additional ~300 individuals that did not fall into sibships) in the FBAT+Wald method increased the number of truly associated SNPs that achieved gws to 21%, this was less than half of the gws SNPs EMMAX detected. The improvement in power was also observed in comparison of the PLINK/QFAM-total naïve test scores (i.e. not adjusted for genomic control) with EMMAX, indicating that the dampening effect of the lambda correction does not account for the total power difference between the two methods (**Figure 4.2B**). Finally, the EMMAX analysis of the entire cohort showed a further 1.2-fold increase in power for the top 10 ranked ground truth SNPs, with an additional 10% of SNPs surpassing genome-wide significance compared to the EMMAX analysis of the sibships.

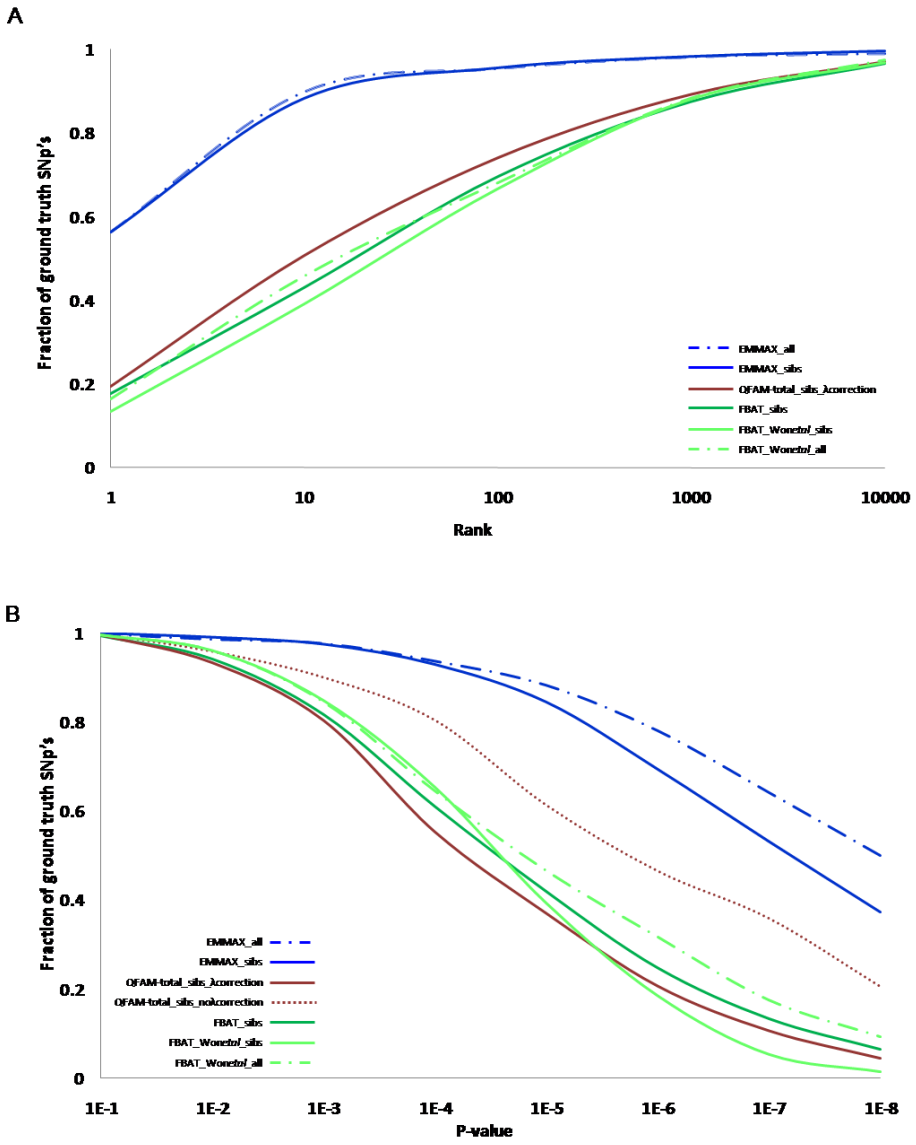


Figure 4.2: Empirical estimation of power of association for four representative association methods for related cohorts. (**A**) The aggregate rank of the ground truth SNP across 770 simulated dataset for association performed with FBAT, FBAT+Wald, PLINK/QFAM-total + genomic control and EMMAX with sibship structured (n = 2,007) and FBAT+Wald and EMMAX with all individuals (n = 2,317). (**B**) The aggregate p-value's of the ground truth SNP for association performed with FBAT, FBAT+Wald, Plink/QFAM-total + genomic control, Plink/QFAM-total without genomic control and EMMAX with sibship structured and FBAT+Wald and EMMAX with all individuals.

Discussion

To determine which association method would be best powered to analyze a highly related isolate population, we compared four methods utilizing different approaches for capturing underlying population structure. One of the methods measured variance within family units and three of the methods added the variance between families using different strategies to control for the over-inflation of the naïve score that resulted from the sample structure. Two of the approaches uniformly rescaled the markers by adding a rank-based p-value to a score or adjusting the score for overinflation by genomic control. These approaches reduce false positives, but also lead to decreased power for true positives. The mixed-model approach modified the test statistic in a marker specific way, which in turn changed the ranks of associated results, so that false positives are reduced, but true positives did not change. In our evaluation of four representative methods for mapping disease alleles in highly related isolated populations, we found that the mixed-model method significantly outperformed the other three methods. The power difference was indeed in part due to the over-conservative uniform adjustments, but they did not account for all the power difference between the two methods. The kinship matrix incorporated in the mixed-model likely captured more between-family variance than the other methods. Further, since the mixed-model method was not restricted to families, an analysis across the whole sample increased the power.

Handling hidden and direct relatedness in the context of a mixed-model has both theoretical and practical advantages over other approaches for association testing in closely-knit, isolated cohorts. Theoretically, it handles relatives distant and close, demographic trends and complex pedigrees (variation both within and between families is utilized) and the test statistic produced is uniformly distributed, with no artificial overdispersion of the p-value. Practically, the ability to include all samples in the analysis gave a > 2-fold boost in power and our empirical results support the use of mixed-models for association in populations such as our Kosrae samples.

It is worthwhile reviewing our work in the context of recent ideas and results in association analysis. Specifically, the emerging picture is that despite comprehensive scans of association with common SNPs of multiple complex phenotypes, only a limited fraction of their heritable component has been identified (Frazer, 2009, Hindorff, 2009, Frazer, 2009, Hindorff, 2009, Frazer, 2009, Hindorff, 2009). Expanding the search for associated alleles to variants that are rare in the general population provides another piece of this heritable component, with multiple gene sequencing studies implicating such variants with moderate to high penetrance (Fearnhead, 2004, Cohen, 2004, Kotowski, 2006, Romeo, 2007, Marini, 2008, Fearnhead, 2004, Cohen, 2004, Kotowski, 2006 ,Romeo, 2007, Marini, 2008, Fearnhead, 2004, Cohen, 2004, Kotowski, 2006, Romeo, 2007, Marini, 2008). Our experiences in association testing in an isolated population can serve as a model for other studies of similar cohorts. Methodologically, mixed-models

optimally extract association information from such samples. As GWAS cohorts increase in size to the 6-digit range, and inclusion of somewhat related individuals becomes a practical modus operandi, results in this work are expected to become relevant to a wide range of populations.

Chapter 5 Association mapping of 30 metabolic traits on Kosrae

Introduction

It has been shown that the limited genetic diversity and reduced allelic heterogeneity observed in isolated founder populations facilitates discovery of loci contributing to both rare monogenic traits and complex disease (Peltonen, 2000, Gulcher, 2001, Heutink, 2002). A strong founder effect, severe isolation and substantial inbreeding have dramatically reduced genetic diversity in the indigenous population form the island of Kosrae, Federated States of Micronesia(Lowe, 2009, Bonnen, 2006). The Islanders also exhibit a high prevalence of diabetes, obesity and other metabolic disorders (Shmulewitz, 2001). Here we present a genome-wide association study of 30 traits pertaining to metabolic disorder and electrocardiographic measures on the island.

Mapping in large, highly related, multi-generational cohorts is analytically challenging, as extensive relatedness between subjects violates assumptions of independence upon which traditional association tests are based. However, rigorous power analysis of several specialized methods designed to accommodate complex familial relationships revealed

that a mixed model approach optimally extracted association information from such samples, demonstrating that population isolates have similar power to non-isolate populations (see **Chapter 4**). Therefore, we used the mixed model method for genome scans in 2,906 related individuals from the Island of Kosrae, Federated States of Micronesia, who were previously genotyped for over 350,000 SNPs.

We re-analyzed data for 17 phenotypes previously studied in this cohort (Lowe, 2009, Smith, 2009), along with 8 additional phenotypes. As positive controls, we observed nine genome-wide significant associations with known loci of measured levels of plasma cholesterol, high density lipoprotein, low density lipoprotein, triglycerides, thyroid stimulating hormone, homocysteine, C-reactive protein and uric acid, with only one detected in previous analysis of the same traits(Lowe, 2009). In addition, we refined the broad signal peak for uric acid levels (chr11:59.1-65.3MB) by analysis of identity-by-descent (IBD) shared genetic segments and dissected the full set of long-range shared haplotypes in this region. We identified a single 3% carrier frequency haplotype that accounted for the entire signal in that region, and replicated one of the two previously known signals, refining that signal by a factor of four. Finally, we show a region of novel association for height (rs17629022, $p<2.1x10^{-8}$).

Results

Sample Ascertainment

We performed a population-based screen of native Kosraens over three separate visits to the Island in 1994, 2001 and 2003 (Shmulewitz, 2006, Shmulewitz, 2001, Lowe, 2009) – see **Table 5.1**. Self –reported family relationships were recorded for use in constructing pedigrees and blood was collected for DNA extraction and genotyping. Genetic accuracy of the Kosrae pedigree was assessed using pairwise identity-by-descent (IBD) estimates generated in PLINK (Purcell, 2007) (see **Materials and Methods**).The total number of unique individuals from the three screens that were successfully genotyped was 2,906, representing >75% of the adult population on the Island.

	N total	**N unique**	**%Males/Females**	**Mean age**	**Median age**	**Age range**
1994	1,935	903	49.6/50.4	43	39	20-86
2001	1,968	889	40.8/59.2	28	24	16-80
2003	84	33	51.5/48.5	24	22	16-53
Multiple exams	-	1,081	35.3/64.7	47	46	17-89
Total	3,987	2,906	41.6/58.4	40	38	16-89

Table 5.1 Study participants successfully genotyped for the Affymetrix 500k assay. Screenings tood place in 1994,2001 and 2003. For each screen, the total and unique number of individuals examined is shown, as some participants were examined in multiple screens. For subjects examined more than once, age (years) is reported form the most recent exam.

Phenotypes pertaining to Metabolic Syndrome and Electrocardiographic measures

A rich, 30-strong phenotypic dataset was collected that included multiple anthropomorphic and Metabolic Syndrome related phenotypes: height(HGT), weight (WGT), waist circumference (WST), bodpod percentage fat assessment (PCF), body mass index (BMI), leptin hormone levels (LEP), diastolic blood pressure (DBP), systolic blood pressure (SBP), total cholesterol levels (TC), low-density liprotein cholesterol levels (LDL), high-density lipoprotein cholesterol levels (HDL), triglyceride levels (TG), campesterol levels (CMP), sitosterol levels (SIT), lathosterol levels (LAT) (CMP, SIT and LAT are collectively called plasma plant sterols (PPS)), fasting plasma glucose (FPG), haemoglobin A1C levels (HBA1C) and insulin sensitivity (INS). In addition, a number of metabolic hormone and inflammation markers were collected: C-reactive protein (CRP), thyroid stimulating hormone (TSH), homocysteine (HOMO), folic acid (FLT), uric acid (URT) and creatinine levels, from which glomular filtration rates were derived (GFR). Finally, a set of electrocardiographic measurements were taken on the Island and a number of electrocardiographic conductance metrics were derived: QRS duration (QRS), PR interval (PR), QT interval (QT), resting heart rate (RR), Sokolow-Lyon voltage (SLV) and Cornell voltage (CLV). The heritability of each trait was calculated in two ways: under a classical polygenic model using nuclear families in SOLAR (Almasy, 1998) and by calculating the fraction of phenotypic variance explained by the empirically estimated kinship matrix in EMMAX (Kang, 2010). The latter is

Trait[1]		All[2]	Males[2]	Female[2]	Covariates[3]	h^2(SOLAR)	h^2(EMMAX)
Total cholesterol (mg/dL)		165.98±33.9 *2,717*	165.16±35.4 *1,122*	116.56±32.79 *1,595*	g,a	0.425	0.503
Low density lipoprotein cholesterol[4] (mg/dL)	LDL	101.71±28.93 *1,870*	102.37±31.16 *711*	101.39±27.66 *1,159*	g,a	0.414	0.459
	APOB	87.1±21.07 *1,834*	89.72±21.19 *766*	85.21±20.79 *1,068*	g,a	0.452	0.432
High density lipoprotein cholesterol[4] (mg/dL)	HDL	38.19±5.87 *1,873*	33.33±9.16 *709*	41.18±10.49 *1,164*	g,a	0.391	0.412
	APOA-1	116.87±24.20 *1,834*	108.23±21.76 *765*	123.13±23.78 *1.069*	g,a	0.398	0.401
Triglycerides (mg/dL)		103.03±46.23 *2,718*	112.79±53.38 *1,122*	96.4±39.73 *1,596*	g,a	0.274	0.352
Campesterol (mg/dL)[6]		1.36±0.57 *2,135*	1.34±0.67 *974*	1.37±0.74 *1,161*	g,a	0.423	0.451
Sitosterol (mg/dL)[6]		1.23±0.54 *2,135*	1.20±0.72 *974*	1.25±0.61 *1,161*	g,a	0.412	0.414
Lathosterol (mg/dL)[6]		1.19±0.58 *2,135*	1.18±0.64 *974*	1.19±0.59 *1,161*	g,a	0.442	0.501
C-reactive protein (mg/L)		0.282±0.349 *1,852*	0.268±0.345 *697*	0.293±0.356 *1,155*	g,a,b	0.245	0.346
Glomular filtration rate (mg/dL)		101.95±24.33 *1,872*	96.81±25.23 *708*	105.27±24.17 *1,164*	g,a	0.170	0.205
Uric acid (mg/dL)		5.46±1.47 *1,844*	6.3±1.45 *694*	4.96±1.23 *1,150*	g,a	0.420	0.475
Folic acid (ng/mL)		9.79±3.36 *1,857*	8.98±3.04 *699*	10.3±3.52 *1,158*	g,a	0.210	0.286
Homocysteine (mg/L)		5.76±2.01 *1,845*	6.82±2.13 *696*	5.14±1.71 *1,149*	g,a	0.160	0.206
Thyroid stimulating hormone (μIU/mL)		1.60±0.91 *1,834*	1.66±0.94 *705*	1.57±0.91 *1,129*	g	0.272	0.509
Height (inches)		62.21±3.21 *2,354*	65.12±2.1 *950*	60.10±2.17 *1,404*	g,a	0.790	0.844

Weight (lb)	172.23±34.12 *2,321*	180.92±34.24 *949*	165.34±33.12 *1,372*	g,a	0.520	0.586
Waist circumference (inches)	37.43±4.99 *2,356*	37.2±4.5 *955*	37,81±5.13 *1,401*	g,a	0.430	0.360
Body mass index (kg/m2)	31.46±5.87 *2,358*	30.28±5.34 *956*	32.29±6.15 *1,402*	g,a	0.473	0.521
Percent body fat (%)	30.37±9.81 *1,578*	20.97±7.65 *598*	35.87±6.54 *980*	g,a	0.414	0.485
Leptin hormone levels (ng/mL)	24.11±20.16 *2,808*	11.46±11.01 *1,166*	31.18±20.34 *1,642*	g,a	0.196	0.251
Fasting blood sugar (mg/dL)	85.34±8.12 *1,456*	85.45±8.41 *624*	85.27±7.81 *832*	g,a,b	0.188	0.180
HBA1C (%)	6.03±2.03 *2,034*	6.14±2.05 *768*	5.97±2.01 *1,266*	g,a	0.204	0.223
Systolic blood pressure (mmHg)	117.23±17.02 *2,402*	120.34±12.61 *967*	115.34±17.32 *1,435*	g,a,b	0.243	0.298
Diastolic blood pressure (mmHg)	77.25±10.45 *2,402*	79.6±10.13 *967*	75.04±10.12 *1,435*	g,a,b	0.289	0.346

Table 5.2 Phenotype characteristics. For each trait; the mean±standard deviation of the raw phenotype, and the number of all phenotyped individuals (All), the mean±standard deviation of the raw phenotype, and the number of male phenotyped individuals (Male), the mean±standard deviation of the raw phenotype, and the number of female phenotyped individuals (Female), any covariates that were adjusted for (Covariates), heritabilites as calculated using a polygenic model in SOLAR (h^2(SOLAR)) and by the fraction of heritability explained by the phenotypic variance in the relatedness metrix in EMMAX (h^2(EMMAX)).

[1]Trait: the trait (units of measurement)
[2] Mean±standard deviation, *number of phenotyped samples used in the analysis*
[3]Covariates; g=gender, a=age, b=bmi
[4]For the 1994 screen, LDL and HDL 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 and HDL (APOB (mg/dL), and APOA-I (mg/dL), respectively) were measured using a standard double antibody immunoassay. Both APOB and APOA-I are highly correlated with their respective cholesterol fractions, making them excellent surrogates. For 1,081 individuals who were screened in both 1994 and 2001, the average of their Z-scores for the both measurements was used.
[5]The distribution of Lipoprotein(a) levels was tri-modal, and raw values with no normalization were used in the association mapping
[6]Plasma plant sterols (Campesterol, Sitosterol and Lathosterol) are give as a ratio with Total Cholesterol

designated "psuedoheritability" since the estimated pairwise relatedness does not correspond exactly to the kinship coefficients. Nonetheless, the EMMAX heritability estimates are concordant with the heritability estimates from the nuclear families (r^2>0.9). The details of the phenotypes, including population means and covariate adjustments heritability estimates for the metabolic traits, are given in **Table 5.2** Electrocardiographic measures are also given (Lowe, 2009, Smith, 2009).

Results from the genome-wide association analysis of 30 metabolic and electrocardiographic traits.

Based on our observations from the simulated data (see **Chapter 4**), we selected the mixed model method as the most powerful method for analyzing our study population. We first built a kinship matrix of pairwise identity-by-state metrics based on high-density markers for the entire pedigree, which was then incorporated into the mixed model for association mapping (Kang, 2010). We had previously published the results of the genome-wide association analysis of 17 of these traits using the PLINK/QFAM-total framework (Lowe, 2009, Smith, 2009). Here, we re-analyzed data for these 17 published phenotypes using the EMMAX framework, along with another 13 previously unreported phenotypes from the island.

.

Symbol	Trait	#SNPs	λ
TC	Total cholesterol	354,880	1.0044
LDL	Low density lipoprotein cholesterol	354,901	0.9785
HDL	High density lipoprotein cholesterol	354,005	0.9714
TG	Triglycerides	354,250	1.0066
CMP[1]	Campesterol:Cholesterol	323,674	1.0103
SIT[1]	Sitosterol:Cholesterol	323,437	1.0029
LAT[1]	Lathosterol:Cholesterol	323,391	1.0034
CRP	C-reactive protein	329,883	1.0220
GFR	Glomular filtration rate	329,414	0.9894
URT	Uric acid	323,483	1.0064
FLT	Folic acid	323,869	0.9945
HOMO	Homocysteine	323,775	1.0123
TSH	Thyroid stimulating hormone	354,558	1.0170
HGT	Height	354,661	0.9756
WT	Weight	354,213	0.9690
WST	Waist circumference	354,290	0.9876
BMI	Body mass index	352,213	0.9599
PCF	Percent body fat	323,923	0.9947
LEP	Leptin hormone levels	323,797	0.9872
FBS	Fasting blood sugar	348,036	1.0135
HBA1C	Haemoglobin A1C	324,031	1.0122
INS	Insulin sensitivity	340,724	0.9975
SBP	Systolic blood pressure	353,640	1.0292
DBP	Diastolic blood pressure	353,640	1.0088
RRD	Resting rate interval	340,312	1.0092
PRD	PR interval	346,566	1.0188
QRS	QRS interval	346,710	1.0159
CLV	Cornell voltage	346,710	1.0082
SLV	Sokolow-Lyon voltage	346,710	1.0160

Table 5.3 Number of SNP's tested for each trait and estimate of p-value score inflation. For each trait; the number of >0.01 MAF SNP's per phenotype (#SNP's) and genomic control correction factor (λ_{median}) from the p-value distribution from the EMMAX association mapping (Devlin, 1999).

[1]The λ for the three plasma plant sterol traits was calculated after the removal of the SNP's comprising the broad peak around the top SNP (Chr2:36-47MB), see chapter 3 for details.

We observe minimal score inflation using the EMMAX method where the inflation factor λ is estimated to be in the range from 1.03 to 0.96 for all traits (see **Table 5.3**). This is in contrast with previously reported λ values that are significantly greater than 1 for a subset of the traits analyzed using PLINK/QFAM -total (λ range: 2.05 – 1.10)(Lowe, 2009). We noted that genome-wide analysis with λ=1.1 inflation of p-values results in a 4.8-fold

increase in the expected number of false-positive results attaining genome-wide significance, with practical implication for the feasibility of follow-up**Table 5.4** provides the top SNP per region emerging genome- and study-wide significant from the EMMAX analysis of all 30 traits. We determined a genome-wide significance threshold based on ~310K effective independent tests (~350k-320k actual tests) to be $1.6x10^{-7}$ (see **Materials and Methods**).

Trait	N[1]	CHR:POS[2]	Top SNP	A1[3]	MAF	effect size		% var[7]	p-value	Gene region[8]
						β[5]	s.e.[6]			
TC	2,753									
	19:45.4	*rs4420638*	*G/A*	*0.182*	*0.274*	*0.033*	*2.42*	*$1.47x10^{-17}$*	*APOE*	
LDL	2,775									
	19:45.4	*rs4420638(+1)*	*G/A*	*0.182*	*0.316*	*0.033*	*3.24*	*$2.57x10^{-23}$*	*APOE*	
	5:74.7	rs3846663	T/C	0.197	0.210	0.029	1.93	$8.81x10^{-7}$	HMGCR	
HDL	2,774									
	16:57.0	rs1800775	A/C	0.394	0.183	0.026	1.82	$7.03x10^{-9}$	CETP	
	19:45.4	rs4420638	G/A	0.182	-0.198	0.033	1.16	$6.30x10^{-7}$	APOE	
TG	2,754									
	11:116.7	*rs7396835(+6)*	*T/C*	*0.372*	*0.214*	*0.026*	*2.36*	*$2.42x10^{-12}$*	*APOA5/APOCIII*	
CRP	1,872									
	19:45.4	*rs4420638*	*G/A*	*0.182*	*-0.318*	*0.041*	*3.05*	*$6.16x10^{-13}$*	*APOE*	
	1:159.7	rs3093077(+4)	C/A	0.247	0.245	0.037	2.25	$7.10x10^{-9}$	CRP	
HOMO	1,870									
	11:89.3	rs1836883	T/C	0.321	0.210	0.036	1.75	$1.29x10^{-8}$	NOX4	
TSH	1,858									
	9:100.5	*rs1877431(+19)*	*A/G*	*0.224*	*-0.299*	*0.037*	*3.46*	*$6.32x10^{-14}$*	*FOXE1*	
HGT	2,364									
	17:43.0	rs17629022	C/T	0.064	0.481	0.058	2.83	$2.00x10^{-8}$	GFAP	
URT	1,861									
	11:63.9	*rs2186571(+58)*	*A/G*	*0.028*	*-1.433*	*0.101*	*9.79*	*$1.77x10^{-34}$*	*URAT1[9]*	
QT	1,432									
	21:34.7	*rs2070359(+1)*	*A/C*	*0.373*	*-0.226*	*0.032*	*2.35*	*$5.17x10^{-9}$*	*KCNE1*	
HBA1C	1,860									
	16:89.0	rs2968474	A/C	0.224	-0.309	0.021	3.05	$7.20x10^{-9}$	CBFA2T3	
PPS	2.034									
	2:44	*rs12185607*	*A/C*	*0.11*	*1.34*	*0.121*	*10.32*	*3.21×10^{-31}*	*ABCG5/ABCG8*	

Table 5.4 Genome- and study-wide significant SNPs from the analysis of 30 traits. This table lists the top SNP in a genetic region for the association of 25 traits pertaining to metabolic syndrome and electrocardiographic conductance that surpasses genome-wide significance ($p \leq 1.6 \times 10^{-7}$). SNP's that also surpassed study-wide significance ($p \leq 6.4 \times 10^{-9}$) are indicated in italics.

[1] N; the number of phenotyped individuals for the trait
[2]POS(MB): chromosomal physical position in megabases; Genome Reference Consortium build 37 (GRCh37)
[3]A1; the minor allele
[4]A2: the major allele
[5]β; the effect size (linear mixed model regression coefficient)
[6]s.e.; the standard error of the effect size
[7]%var; the percent variance explained (oneway ANOVA)
[8] Candidate genes in the region of the top SNP
[9] rs2186571 is > 400kb upstream of URAT1 (see results section for explanation)

Observation of positive controls in the Kosare dataset

As positive controls, we observed associations to ten genomic regions for ten traits exceeding our genome-wide significance threshold. Top SNPs are in or near to known loci for plasma cholesterol (*APOE*; rs4420638, $p \leq 1.47x10^{-17}$), high density lipoprotein (*CETP*; rs1800775, $p \leq 7.03x10^{-9}$), low density lipoprotein (*APOE*; rs4420638, $p \leq 1.47x10^{-23}$), triglycerides (*APOC3/A5*; rs7396835, $p \leq 2.42x10^{-12}$), thyroid stimulating hormone (*FOXE1*; rs1877431, $p \leq 6.32x10^{-14}$), homocysteine (*NOX4;* rs1836883, $p \leq 1.3x10^{-8}$), C-reactive protein (*APOE*; rs4420638, $p \leq 6.16x10^{-13}$ and *CRP*; rs3093077, $p \leq 7.10x10^{-9}$), QT interval (*KCNE1*; rs2070359, $p \leq 5.2x10^{-9}$) and the plasma plant sterols (*ABCG5/ABCG8*; rs12185607, 3.21×10^{-31}) (Kathiresan, 2009, Kathiresan, 2008, Elliott, 2009, Willer, 2008, Sandhu, 2008, Gudmundsson, 2009, Pare, 2009). Nine of these regions are shown in detail in **Figure 5.1A** (for details regarding the plasma plant sterol region, see C**hapter 2**). Furthermore, six of these signal peaks surpassed our study-wide significance threshold of $6.4x10^{-9}$ (see Material and Methods). Only one of these positive controls had been detected at genome-wide significance in the previous analysis of the same traits: triglycerides (*APOC3/A5*; rs7396835, $p \leq 1.2x10^{-9}$) (Lowe, 2009). In addition we observed two other known positive controls significant at FDR <0.3: low density lipoproteins (*HMGCR*; rs3846663, $p \leq 8.77x10^{-7}$) and high density lipoprotein (*APOE*; rs4420638, $p \leq 6.33x10^{-7}$)(Kathiresan, 2009, Aulchenko, 2009).

Novel findings from the Kosrae association mapping

As novel findings, we observed three interesting genome-wide significant signals, one strong signal for uric acid levels that resides >500kb upstream of two independent known associations (rs2186571, $p<1.8x10^{-34}$), and signals for height and HBA1C which represented a novel association in the Kosraen cohort (rs17629022, $p<2.1x10^{-8}$; and rs2968474, $p<7.2x10^{-9}$, respectively; see **Figures 5.1B** and **5.1C**).

The novel SNP rs17629022 for height has a minor allele frequency of 0.06 and a strong effect size of 0.48 standard deviations per allele (for normalize z-scores, $\mu=0$, $\sigma=\pm1$). This corresponds to an average 1.28" (95% confidence intervals: 0.97"-1.66") and 4.23" (95% confidence interval: 2.48"-5.97") average increase in height in carriers (n=271) and minor allele homozygotes (n=13), respectively, compared to major allele homozygotes. The SNP resides on chromosome 17q21 in a gene rich region containing *GAP*, *CCDC103* and *FAM187A*.

The novel SNP rs2968474 for HBA1C has a minor allele frequency of 0.224 and a strong effect size of -0.309 standard deviations per allele. In work by my collegue, Sasha Gusev, the signal for this SNP was fine mapped to a 200kb, 11 SNP haplotype in the region ($p<1.66x10^{-23}$), where the HBA1C haplotype carriers (n=284) and homozygotes (n=8) have decreased HBA1C levels, 5.00%±0.57% and 4.13%±0.51%, respectively, compared to non-carriers (n=1,282), 5.33%±0.53%. Further, low pass sequencing in the region provide preliminary evidence for both coding mutations in genes on the background of

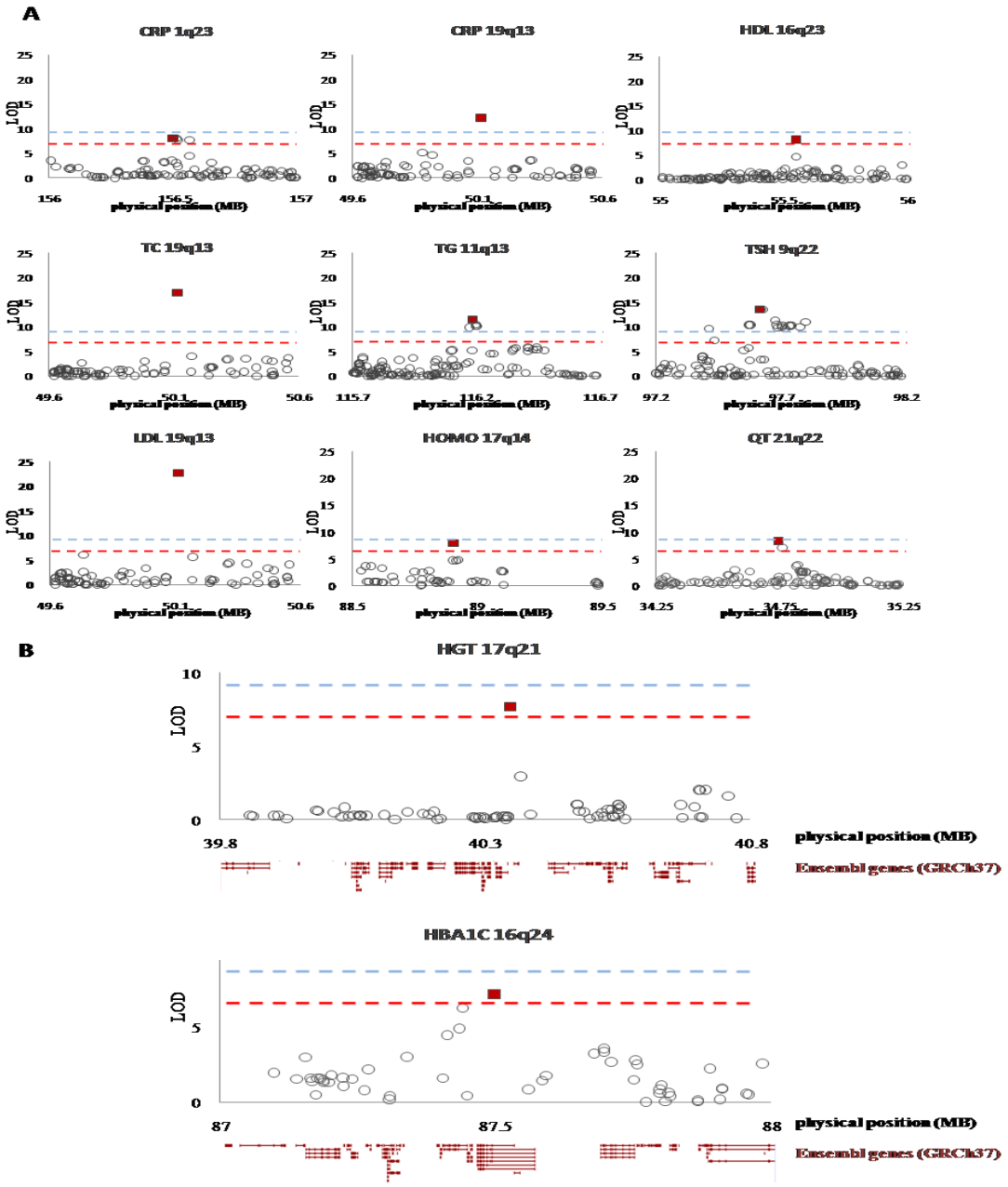


Figure 5.1 Eleven regions of genome-wide significant associations. (A) We observe nine genome-wide significant associations to known loci for plasma cholesterol, high density lipoprotein, low density lipoprotein, triglycerides, thyroid stimulating hormone, homocysteine, C-reactive protein and QT interval (the tenth positive control, ABCG8/G5 region for PPS is described in **Chapter 2**). (B) Region of novel association for height and haemoglobin A1C. The y-axis show's $-\log_{10}$p-value (LOD) and the x-axis shows the chromosomal physical position in megabases. The maroon square represents the location of the top scoring SNP. The red and pale blue dashed lines indicate the genome- and study-wide significance thresholds, respectively.

the haplotype and structural variation in carriers. This work is currently under preparation for publication (pers. comm. Sasha Gusev). Analysis of the uric acid trait revealed a broad peak in the genome scan on chromosome 11q13.1 (59.2-65.3 MB) containing 54 SNPs that surpassed the study-wide significance threshold (**Figure 5.3C**). The top SNP, rs2186571, represents a novel association to uric acid. Further, the very low p-value of the top SNP $p<1.8x10^{-34}$, signifies an effective internal replication of the signal, since, for an effect this strong, a genome-wide significant signal would be observed in any randomly partitioned half of the study population. rs2186571/A has a minor allele frequency of 0.028 and a strong effect size of -1.14 standard deviations per allele, accounting for ~10% of uric acid level variance on Kosrae, where carriers of rs2186571/A have a mean uric acid level of 3.6 mg/dL compared to the major allele homozygotes level of 5.5 mg/dL. However, rs2186571 resides ~500kb upstream of two variants on chromosome 11q13.1 recently found to be associated with uric acid levels (Kolz, 2009). To determine whether the strong uric acid signal represented a novel observation or might tag either of the previously observed signals, we performed a haplotype-based fine-mapping of the signal region and conditional analysis.

Fine mapping the interval of uric acid association

Haplotype mapping has been previously described in **Chapter 2**. Briefly, common shared long-range haplotypes were detected via pair-wise IBD genetic segment matching

between all individuals in the pedigree using the GERMLINE software (Gusev, 2009). The IBD segments were at least 3cM in length and allowed for up to 1% mismatching due to genotyping error. IBD segments were then clustered to groups of similar haplotypes, with >500kb overlapping sequence and sharing >99% sequence identity, as described in detail previously (Kenny, 2009). Clusters of shared haplotypes were then independently mapped to uric acid levels which identified a single strong mapping cluster, where the specific sharing boundaries of the haplotype mapped to a ~2MB sub-region (chr11:63.7-65.5) of the uric acid GWAS signal peak (see **Figure5.2(A)**). This "uric acid haplotype" (rs4980499 to rs1074156) is unique to a group of 96 individuals (carrier frequency 0.03), is strongly associated to uric acid ($p<9.07x10^{-46}$), and has an effect size of -2.12 standard deviations per allele (accounting for ~20% of uric acid level variance on Kosrae; 95% confidence intervals 16.8%-21.6%). Conditioning the uric acid levels for the uric acid haplotype effect abolishes signal at the rs218571 SNP ($p<0.83$), whereas, a significant signal remains ($p<2.1x10^{-5}$) at the uric acid haplotype when the levels are adjusted for the effect of rs218571 (see **Table 5.5**). Genome-wide scan of haplotype conditioned uric acid levels reveals no remaining genome-wide significant signals (see **Figure 5.2(B)**). Therefore, the uric acid haplotype likely captures the full signal tagged by the top GWAS SNP.

We next determined whether the uric acid haplotype represented a novel variant or a refinement of one or both loci that had previously shown uric acid variants, rs17300741

and rs505802. The two known variants reside in/near organic ion transporter 4 (*OAT4*) and urate transporter 1 (*URAT1*), respectively, which are known candidate genes for renal urate anion exchange and are thought to regulate blood uric acid levels (Taniguchi, 2008, Ekaratanawong, 2004). Neither variant was directly assayed in the Kosrae genotype panel. However, two assayed variants on the 500k Affymetrix chip, rs12362644 and rs10897518 that flank rs17300741 and rs505802, respectively, were observed to be in complete linkage disequilibrium (r^2=1) with each variant in all HapMap Asian panels (CHB, CHD, JPT, GIH). We therefore concluded that rs12362644 and rs10897518 genotyped on Kosrae would be strong tags for the unobserved rs17300741 and rs505802 variants, respectively. Indeed, signals for association were observed for rs12362644 ($p < 8.16x10^{-5}$) and rs10897518 ($p < 3.41x10^{-17}$), the latter being study-wide significant.

Analysis of the uric acid haplotype conditioned on rs12362644 and rs10897518 abolished the signal at these two variants ($p<0.41$ and $p<0.93$, respectively). However, the strong signal at the uric acid haplotype persisted ($p< 8.48x10^{-23}$). Conversely, when uric acid levels are conditioned on the uric acid haplotype, moderate signal for rs12362644 (tagging rs17300741) remained $2.66x10^{-3}$; however, signal at rs1087518 (tagging rs505802) was abolished ($p<0.94$) (see **Table 5.5**), indicating that, while there is some correlation between rs12362644 and the uric acid haplotype, the haplotype effect is not explained by the effect of rs12362644. Instead, this effect is the result of the variant tagged by rs10897518, which is carried on the uric acid haplotype background.

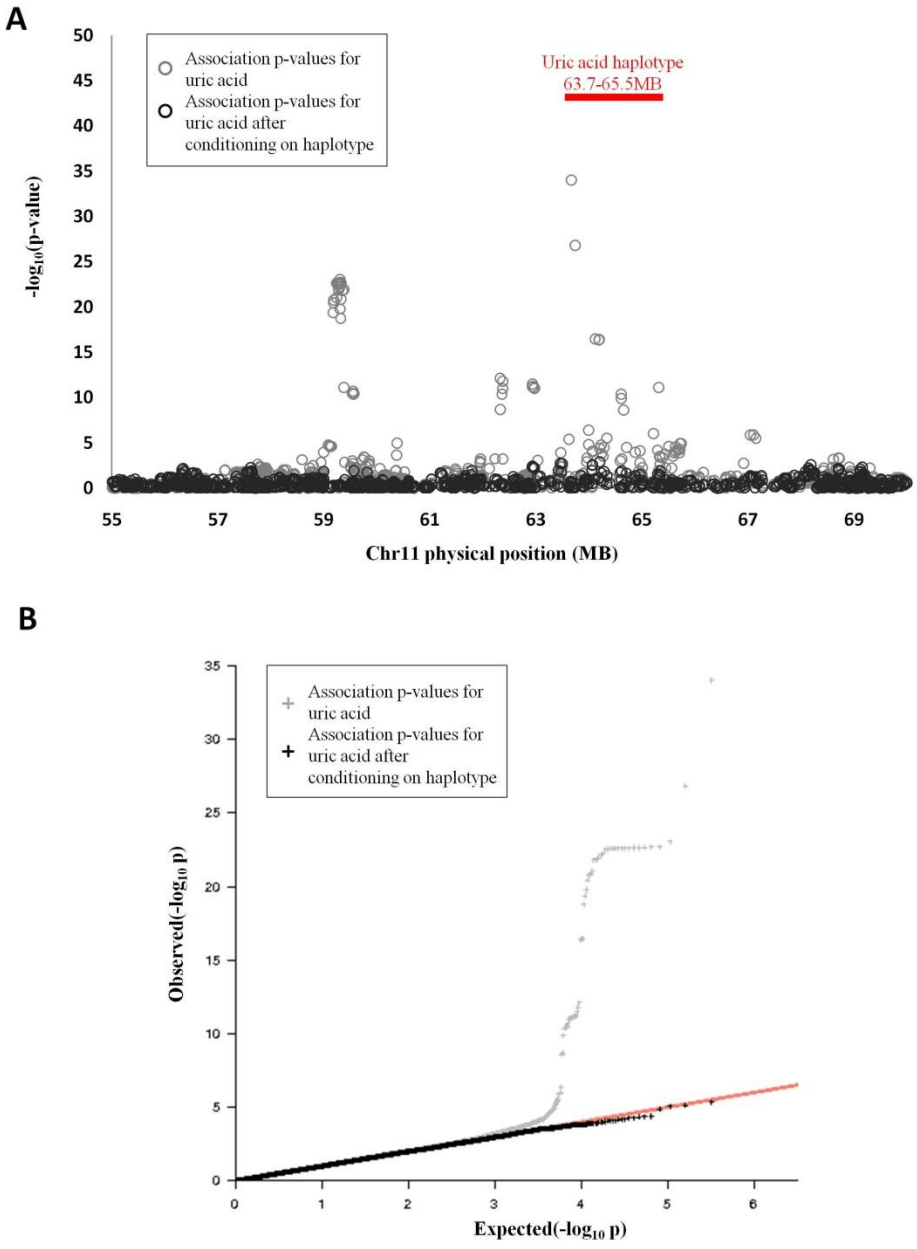


Figure 5.2 Conditioning on the URT haplotype. (A) Chromosome 11q13 region of association to uric acid levels before and after conditioning uric acid levels on the uric acid haplotype. (B) QQ plot showing genome-wide association to uric acid levels and to uric acid levels conditioned on the uric acid haplotype.

	URT Haplotype	rs12362644 (tagging rs17300741)	rs10897518 (tagging rs505802)
naïve score	9.07×10^{-46}	8.16×10^{-5}	3.41×10^{-17}
Adjusted allele(s)			
URT Haplotype	-	2.66×10^{-3}	-
rs12362644	4.71×10^{-44}	-	4.94×10^{-16}
rs10897518	8.56×10^{-24}	1.06×10^{-3}	-
rs12362644 and rs10897518	8.48×10^{-23}	-	-

Table 5.5 Conditional analysis. Conditional analysis of the URT haplotype, rs12362644 (tagging rs17300741) and rs10897508 (tagging rs505802) at the uric acid locus on chromosome 11q13.1

Moreover, comparing the 0.06 frequency of rs1087518 on Kosrae to the lower 0.015 frequency of the uric acid haplotype indicated a refinement of the signal by a factor of 4, dramatically reducing the sequence and sample search space for pinpointing the underlying causal variant.

Comparison of effect sizes for known associated loci

We observe that for common variants that are polymorphic in Kosrae, this cohort often achieves significant association, with significantly stronger statistical support than its sample size would suggest. A good example for that is the LDL phenotype, where ~2,800 phenotyped islanders are sufficient to detect positive controls signals – known

associations on chromosomes 19 (APOE) and 5 (HMGCR) are significant at levels $<3\times10^{-22}$ and $<9\times10^{-7}$, respectively. We extrapolated from reports in cohorts from European populations (Kathiresan, 2008, Aulchenko, 2009) that >10000 samples are needed to expose these associations at this significance level. Comparing the association regression coefficients (β) shows a larger increase on the normalized phenotype in Kosrae.

To more formally examine the genetics of common traits on Kosrae as compared to outbred populations, we assessed the effect sizes of known loci observed in our study. Specifically, we identified 53 established associations in large, mainly Caucasian studies across seven traits (BMI, CRP, HDL, HGT, LDL, TG and URT), where the associations had sufficiently high effect size and/or allele frequency to be detectable at nominal significance in the ~3,000-strong Kosrae study (see **Table 5.6**)

				Outbred				Kosrae			
Trait	SNP	Gene/Region	Allele[1]	Freq	Beta	P-value	On 500K?[2]	Freq	Beta	P-value	Study[3]
BMI	rs9939609	FTO	A	0.45	0.1	2.00E-20	same	0.12	0.13	3.00E-02	Frayling(2007)
BMI	rs17782313	MC4R	C	0.28	0.05	2.80E-15	same	0.08	-0.04	5.70E-01	Loos(2008)
CRP	rs4420638	APOE	G	0.18	-0.21	5.00E-27	same	0.21	-0.32	6.50E-13	Ridker(2008)
CRP	rs7310409	HNF1A	A	0.39	-0.15	6.80E-17	rs2393791	0.44	-0.139	2.84E-04	Ridker(2008)
CRP	rs7553007	CRP	A	0.33	-0.2	1.10E-26	same	0.38	0.111	3.90E-03	Ridker(2008)
CRP	rs4129267	IL6R	A	0.4	-0.1	2.00E-08	rs4537545	0.36	-0.097	8.56E-03	Ridker(2008)
HDL	rs1800775	CETP	C	0.43	-0.18	1.00E-73	same	0.58	-0.169	7.58E-09	Kathiresan (2008)
HDL	rs328	LPL	G	0.13	0.17	9.00E-23	rs10503669	0.98	0.247	1.03E-02	Kathiresan (2008)
HDL	rs7395662	MADD-FOLH1	G	0.61	-0.073	6.00E-11	same	0.43	-0.05	1.21E-01	Aulchenko(2009)
HDL	rs4939883	LIPG	A	0.83	0.07	2.40E-07	same	0.64	0.084	2.16E-01	Aulchenko(2009)
HDL	rs2271293	CTCF-PRMT8	G	0.87	-0.129	8.30E-16	same	0.6	-0.06	2.29E-01	Aulchenko(2009)
HDL	rs3890182	ABCA1	A	0.09	-0.1	3.00E-10	same	0.37	-0.03	2.60E-01	Kathiresan (2008)
HDL	rs2156552	LIPG	A	0.2	-0.07	2.00E-07	same	0.03	0.07	2.80E-01	Kathiresan (2008)
HDL	rs4846914	GALNT2	G	0.42	-0.07	2.00E-13	same	0.58	0.0009	7.45E-01	Kathiresan (2008)
HGT	rs6763931	ZBTB38	A	0.48	0.07	1.40E-27	rs6440003	0.25	0.105	3.17E-03	Gudbjartsson (2008)
HGT	rs724016	ZBTB38	G	0.48	0.37	8.30E-22	rs6440003	0.25	0.105	3.17E-03	Lettre(2008)
HGT	rs798544	GNA12	G	0.72	0.06	6.50E-15	same	0.47	0.078	1.23E-02	Gudbjartsson (2008)
HGT	rs3748069	GPR126	A	0.73	0.07	4.50E-14	rs7755109	0.25	0.082	3.76E-02	Gudbjartsson (2008)
HGT	rs2282978	CDK6-GATAD1	C	0.37	0.06	9.80E-09	same	0.06	-0.141	3.80E-02	Gudbjartsson (2008)
HGT	rs4533267	ADAMTS17	A	0.28	0.06	3.30E-08	same	0.4	-0.0588	6.90E-02	Gudbjartsson (2008)
HGT	rs6060369	GDF5-UQCC	C	0.36	0.44	1.40E-16	same	0.31	-0.06	7.20E-02	Lettre(2008)
HGT	rs1812175	HHIP	C	0.81	0.08	9.70E-12	same	0.63	0.05	1.13E-01	Gudbjartsson (2008)
HGT	rs7153027	TRIP11-ATXN3	A	0.61	0.06	1.10E-10	same	0.77	-0.059	1.41E-01	Gudbjartsson (2008)
HGT	rs2562784	SH3GL3-ADAMTSL3	G	0.17	0.34	6.40E-08	same	0.42	0.043	1.89E-01	Lettre(2008)
HGT	rs7846385	PXMP3-ZFHX4	C	0.34	0.05	4.70E-08	same	0.16	-0.05	2.58E-01	Gudbjartsson (2008)
HGT	rs1042725	HMGA2	T	0.49	-0.48	2.70E-20	same	0.55	0.022	5.02E-01	Lettre(2008)
HGT	rs4743034	ZNF462	A	0.24	0.05	2.10E-08	rs12411264	0.14	0.03	5.57E-01	Gudbjartsson (2008)
HGT	rs10513137	ZBTB38	A	0.26	0.03	8.00E-08	same	0.09	0.03	5.60E-01	Kim(2009)
HGT	rs3760318	CRLF3-ATAD5	C	0.64	0.06	1.80E-09	rs7225461	0.56	-0.0158	6.44E-01	Gudbjartsson (2008)
HGT	rs6918981	HMGA1	G	0.21	0.02	2.00E-08	same	0.14	0.018	6.90E-01	Kim(2009)
HGT	rs4794665	NOG-DGKE	A	0.53	0.04	9.90E-08	same	0.26	-0.008	8.13E-01	Gudbjartsson (2008)

HGT	rs6830062	LCORL-NCAPG	T	0.84	0.06	1.30E-10	same	0.94	0.016	8.34E-01	Gudbjartsson (2008)
HGT	rs967417	BMP2	C	0.57	0.04	1.50E-08	same	0.13	-0.006	9.14E-01	Gudbjartsson (2008)
LDL	rs4420638	APOE	G	0.18	0.19	1.00E-60	same	0.21	0.365	1.13E-23	Kathiresan (2008)
LDL	rs12654264	HMGCR	T	0.42	0.1	1.00E-20	same	0.42	0.133	1.66E-05	Kathiresan (2008)
LDL	rs157580	APOE	G	0.33	-0.111	2.10E-19	same	0.32	0.0326	3.49E-01	Aulchenko(2009)
LDL	rs16996148	CLIP2-PBX4	T	0.06	-0.1	3.00E-08	same	0.17	-0.05	5.63E-01	Kathiresan (2008)
LDL	rs693	APOB	A	0.49	0.12	1.00E-21	same	0.09	-0.026	6.43E-01	Kathiresan (2008)
LDL	rs174570	FADS2/3	G	0.83	0.11	4.40E-13	same	0.80	-0.0003	9.50E-01	Aulchenko(2009)
TG	rs16996148	CILP2-PBX4	T	0.06	-0.1	4.00E-09	same	0.17	-0.179	2.11E-05	Kathiresan (2008)
TG	rs780094	GCKR	T	0.38	0.13	3.00E-14	same	0.27	-0.125	1.15E-03	Kathiresan (2008)
TG	rs17321515	TRIB1	G	0.4	-0.08	4.00E-17	same	0.7	-0.0928	7.04E-03	Kathiresan (2008)
TG	rs328	LPL	G	0.13	-0.19	1.00E-28	rs10503669	0.98	0.143	1.51E-01	Kathiresan (2008)
TG	rs157580	APOE	G	0.33	-0.069	1.20E-08	same	0.32	-0.034	3.30E-01	Aulchenko(2009)
TG	rs12130333	DOCK7	T	0.24	-0.11	2.00E-08	same	0.04	-0.07	3.48E-01	Kathiresan (2008)
TG	rs693	APOB	A	0.49	0.08	2.00E-07	same	0.09	-0.05	3.63E-01	Kathiresan (2008)
TG	rs4846914	GALNT2	G	0.42	0.08	7.00E-15	same	0.58	-0.022	4.60E-01	Kathiresan (2008)
TG	rs10889353	DOCK7	C	0.32	-0.085	8.20E-11	same	0.23	0.023	5.06E-01	Aulchenko(2009)
TG	rs17145738	MLXIPL	T	0.12	-0.14	7.00E-22	same	0.14	-0.029	8.20E-01	Kathiresan (2008)
URT	rs505802	SLC22A12	T	0.7	-0.06	2.00E-09	rs10897518	0.97	-0.64	3.14E-17	Kolz(2009)
URT	rs17300741	SLC22A11	A	0.51	0.06	7.00E-14	rs12362644	0.71	0.162	8.16E-05	Kolz(2009)
URT	rs780094	GCKR	T	0.42	0.05	1.00E-09	same	0.18	-0.082	9.73E-02	Kolz(2009)
URT	rs1165205	SLC17A3	T	0.46	-0.09	4.00E-27	same	0.23	0.04	3.50E-01	Dehghan(2008)

Table 5.6*:* Comparison of effects and allele frequencies in Caucasians and Kosraens for known associated loci.

[1]Allele: the effected allele

[2]On 500K?: if the SNP from the outbred population study is on the 500K Affymetrix chip used in the Kosrae study "same" is indicated. If not, then a proxy SNP (r^2>0.95 between the proxy SNP and the outbred population SNP) that is on the 500K Affymetrix chip is listed.

[3]Study: The outbred population studies used for this analysis are given in the references (Frayling, 2007)(Loos, 2008)(Ridker, 2008)(Kathiresan, 2008)(Aulchenko, 2009)(Lettre, 2008)(Gudbjartsson, 2008)(Kim, 2010)(Kolz, 2009)(Dehghan, 2008)

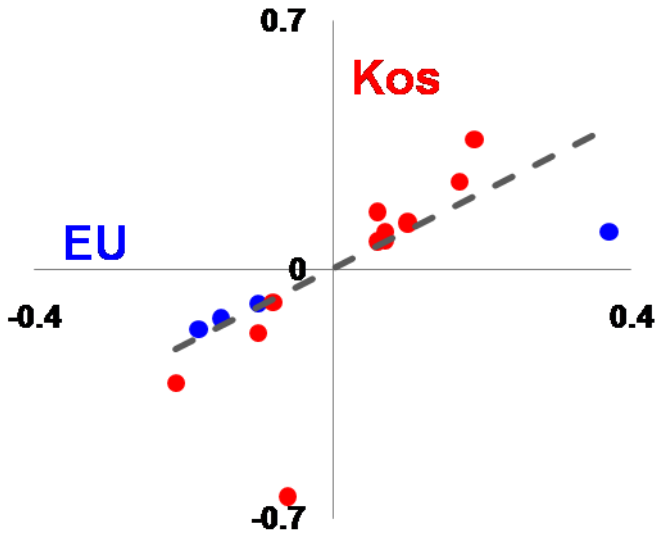


Figure 5.3 Comparison of effect sizes for known loci in outbred populations on Kosrae. Effect sizes (the linear mixed model regression coefficient) on Kosrae (Kos on the y-axis) and in outbred populations (EU on the x-axis) for 16 genome-wide significant SNP's from outbred population studies, where the signals are nominally significant on Kosrae and in the same direction, are plotted. A fitted line shows the cases where the effect size is stronger in the outbred population (blue) and on Kosrae (red)

(Lettre, 2008, Kathiresan, 2008, Frayling, 2007, Loos, 2008, Ridker, 2008, Aulchenko, 2009, Gudbjartsson, 2008, Kim, 2010, Kolz, 2009, Dehghan, 2008). For SNP's not directly typed on the Affymetrix array, association results are reported for a proxy on the Affymetrix chip with strong correlation (r^2>0.95) to the original SNP in both HapMap Caucasian (CEU) and Asian (CHB, CHD, JPT, GIH) panels. Of the 53 SNP's, only 21 were detectable on Kosrae at nominal significance, indicating that more than half the associated SNP's were tagging causal variants that were either too rare to be detected or not present on the Island. Sixteen of the 21 detectable SNP's on Kosrae had effect sizes in the same direction as those SNP's in the Caucasian cohorts ($p<0.01$) and 5 were in the opposite direction. Of the former set of 16 SNP's, 12 had stronger effect sizes on Kosrae, which is a greater number than expected by chance ($p<0.027$)

Discussion

We describe a GWAS in a population-based cohort with extensive family structure and explore the value of genetics studies in a extreme populations isolate. Our goal was to take advantage of the population genetic features such a population offer, while maximizing power to detect association using mixed models. The analysis and re-analysis of 30 biomedical traits from the island using the more powerful mixed-model approach yielding 13 significant hits, only one of which had been detected by other methods. More importantly, were the discovery of a putative novel association for height, HBA1C and a novel long shared haplotype that accounts for ~20% variance in uric acid levels on the Island.

These three novel associations exemplify both the limitations and opportunities for performing genetic mapping in isolated populations. We have demonstrated association for height in a region not previously observed, despite multiple large-scale genetic studies in Caucasians (Hirschhorn, 2009, Lettre, 2008, Weedon, 2008). It is possible that the underlying mutation is of stronger effect or higher frequency on the island, making it easier to detect. Alternatively, the causal variant may be Asian specific, or even private to the island. If the latter case is true, then the signal is unlikely to be confirmed by replication. It should also be noted that the signal for the height SNP did not reach study-wide significance and could represent a false positive.

On the other hand, we have observed a moderate signal for HBA1C (described elsewhere – pers. comm. Sasha Gusev) and an extremely strong signal for uric acid level, both also present in outbred, Caucasian populations, which we could refine by a IBD-based haplotype fine-mapping facilitated by the abundance of long shared haplotypes in isolated populations (Bonnen, 2006 ,Kolz, 2009). Further, since the uric acid haplotype tags only one of the two Caucasian variants in the same region previously known, our findings support independence between the two reported signals in the region. Finally, the strong effect size of the haplotype is beyond the combined effects of variants affecting uric acid level previous discovered (Kolz, 2009, Dehghan, 2008).

In conclusion, it is worthwhile reviewing our work in the context of recent ideas and results in association analysis. Specifically, the emerging picture is that despite comprehensive scans of association with common SNPs of multiple complex phenotypes, only a limited fraction of their heritable component has been identified(Frazer, 2009, Hindorff, 2009). Expanding the search for associated alleles to variants that are rare in the general population provides another piece of this heritable component, with multiple gene sequencing studies implicating such variants with moderate to high penetrance (Fearnhead, 2004 ,Cohen, 2004 ,Kotowski, 2006 ,Romeo, 2007, Marini, 2008). Our experiences in association testing in an isolated population can serve as a model for other studies of similar cohorts. Methodologically, mixed-models optimally extract association information from such samples. More importantly, in terms of study design and choice of cohort, we observe some of the previously theorized advantages and limitations of isolated populations. We showed that some variants that are associated to traits in other

populations do not replicate on Kosrae and this was due to bottleneck effects rather than lack of power. On the other hand, observed associations show strong effects and statistical support for a cohort of this size, potentially due to increased genetic and environmental homogeneity. Finally, as GWAS cohorts increase in size to the 6-digit range, and inclusion of somewhat related individuals becomes a practical modus operandi, results in this work are expected to become relevant to a wide range of populations.

Chapter 6 Genotypic and Haplotypic mapping of Crohn's Disease in Ashkenazi Jews

Introduction

Crohn's disease (CD) is a chronic inflammatory disorder of the gastrointestinal tract, which is thought to result from the effect of environmental factors in a genetically predisposed host. An epidemiological feature of CD is that it has highest prevalence among individuals of Ashkenazi Jewish (AJ) descent, occurring two to four times more frequently than in non-Jewish Caucasian populations. Three coding variants of the *NOD2* gene have been reported as independent disease-predisposing mutations for CD. Confirmation studies have shown, however, that these susceptibility loci have similar allele frequencies and magnitudes of the effect and are unlikely to contribute to the excess prevalence of the disease in the AJ population. Therefore, involvement of other, yet unknown, genetic variants unique to this population is hypothesized (Sugimura, 2003).

Here we analyzed CD in seven Ashkenazi Jewish cohorts, under a combined-analysis strategy. We generated a population specific reference panel for this dataset and impute m. As positive controls, we observed two genome-wide and six strongly significant

associations with known loci for Crohn's disease. In addition, we refined a broad signal peak for Crohn's disease at the *NOD2* locus on chromosome 16 using analysis of identity-by-descent (IBD) shared genetic segments. We identified a second 13% carrier frequency haplotype ~1MB upstream of the top SNP in the region, which is independent of the top SNP.

Results

Sample Ascertainment

We performed a population-based screen of Ashkenazi Jewish (AJ) individuals by combining six New York-based and one Israel-based screenings - see Table 6.1. Self–reported family relationships were recorded and only individuals who self-identified as AJ (or had at least one AJ grandparent) were used in the study. The cohorts comprised samples from; the NIDDK IBD genetics consortium, only persons with Jewish ancestry, obtained from dbGAP (Study accession: phs000130.v1.p1) -- NY1; Yale University (pers. comm.

	Nt_{otal}	$N_{CD\ cases}$	$N_{non\text{-}CD\ cases}$	$N_{controls}$	Platform
NY1	828	397	-	431	Illumina 300k
NY2	59	59	-	-	Illumina 1M
NY3	136	136	-	-	Illumina 550k
NY4	173	113	-	60	Affymetrix 500k
NY5	1,067	261	806	-	Affymetrix 6.0
NY6	651	-	-	651	Affymetrix 6.0
TA1	397	-	200	197	Illumina 300k
Total	**3,311**	**966**	**1,009**	**1,339**	

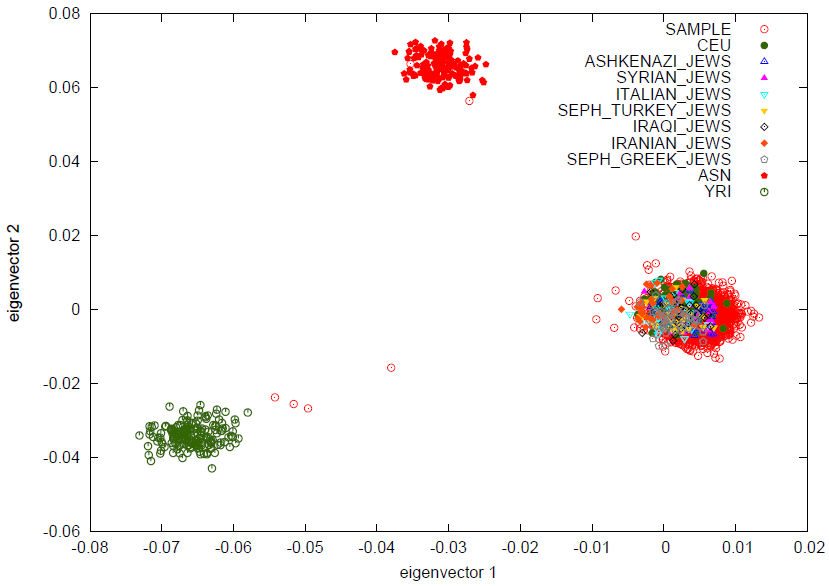
Table 6.1 Study participants from nine groups successfully genotyped on either an Illumina or Affymetrix platform. Screenings in nine centers based in New York (NY1-7) or Israel (TA1). For each screen, the total number of individuals examined is shown (N_{total}), in addition to any Crohn's disease cases (NC_{D_cases}), non-Crohn's disease cases, which are a mix of Parkinson's disease, Schizophrenics, Type-II Diabetes and Dystonia cases, ($N_{non\text{-}CD\_cases}$), and non-diseased controls ($N_{controls}$). . For each cohort, the platform used for genotyping is shown (Platform).

Judy Cho) -- NY2; the Children's Hospital of Philadelphia (CHOP) (pers. comm. Hakon Hakonarson) – NY3; Mount Sinai Medical School (pers. comm. Inga Peters and Laurie Ozelius) – NY4-5; Albert Einstein College of Medicine (pers. comm. Nir Barzilai) – NY6 and the Hebrew University of Jerusalem (pers. comm. Ariel Darvasi) –TA1. The total number of AJ individuals from the seven screens was 3,311, where 966 were Crohn's disease cases and 2,345 were either non-Crohn's disease cases or non-diseased controls (1,009 and 1,339, respectively).

Confirming Ashkenazi ancestry of study participants

To confirm the AJ ancestry of the all the samples in the combined cohort we performed a principal component analysis (PCA) plotting the samples with the three continental HapMap reference panels (CEU, YRI and ASN (combined JPT+CHB) and seven panels from the Jewish HapMap (JHapMap) consortium consisting of one Ashkenazi Jews, one European Jewish, three Middle Eastern and two Sephardic Jewish panels. **Figure 6.1** shows the first and second eigenvectors of the PCA. **Figure 6.1(A)** shows the separation the three continental groups, where all except eight of the AJ samples cluster with the CEU Hapmap reference panel as expected**. Figure 6.1(B)** shows the same graph with the outlier YRB and ASN samples removed. ~80% of individuals (n=2,661) in the AJ sample form a distinct cluster with the Ashkenazi Jews from JHapMap along the first principal components axis (PC1) as compared to both non-AJ Jews and individuals with

A



B

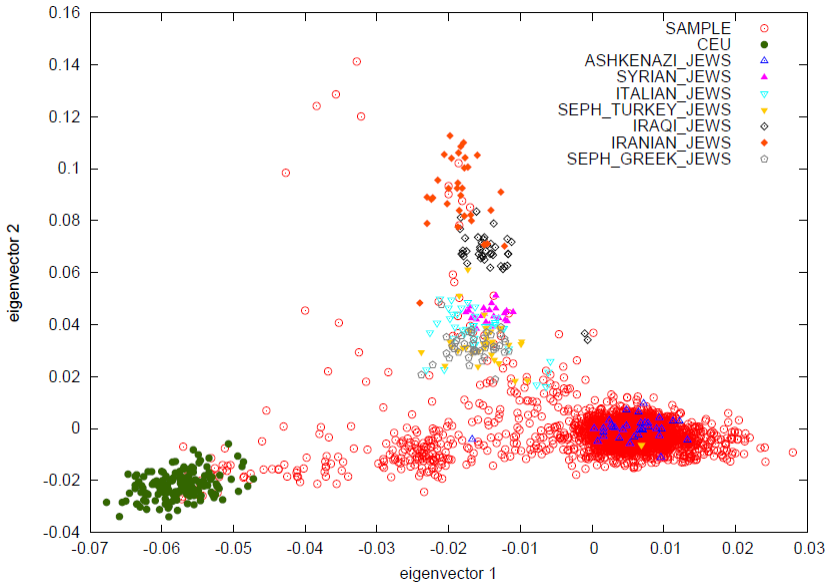


Figure 6.1 PCA analysis of AJ samples. (A) A principal component analysis graphs showing the first (x-axis) and second (y-axis) eigenvectors plotting all 3,490 AJ samples (SAMPLE). Also included and color-coded in the graphs are the four HapMap reference samples; CEPH-Utah (CEU), Yoruban-Nigeria (YRB) and the combined (Han Chinese and Japanese) Asian samples (ASN); and seven Jewish samples from the JewishHapMap project, consisting of Ashkenazi Jews (ASHKENAZI), one European Jewish (ITALIAN), three Middle Eastern (SYRIAN, IRAQI and IRANIAN) and two Sephardic Jewish cohorts (SEPH_TURKEY and SEPH_GREEK). **(B)** The same graph excluding the YRB and ASN outlier groups.

no Jewish ancestry (CEU), where the AJ cluster has a higher score than any other cluster. Further, we could identify ~300 samples who were positioned between the AJ cluster and the non-Jewish cluster (CEU) on PC1 (see **Figure 6.2(A)**). Upon examining the distribution of PC1 values in these samples, three distinct modes were defined; Group 1 (PC1 -0.005-0.018), Group 2 (PC1 -0.015- -0.006) and Group 3 (PC1 -0.03- -0.016) (see figure 6.2(B)). We postulated, based on previous PCA analysis of AJ individuals (Need, 2009) that groups 2 and 3 might represent individuals with 75% (one non-AJ grandparent) and 50% (one non-AJ parent or two non-AJ grandparents) AJ ancestry, respectively. So as not to exclude these individuals with partial AJ ancestry, we devised a GWAS strategy, whereby we would perform association mapping in within each group independently to control for admixture effects, and combined the p-values from each group under a meta-analysis design to construct a single test score. Further, we noted that group 3 (50% AJ ancestry) contained ~2-fold more CD cases than controls and thus

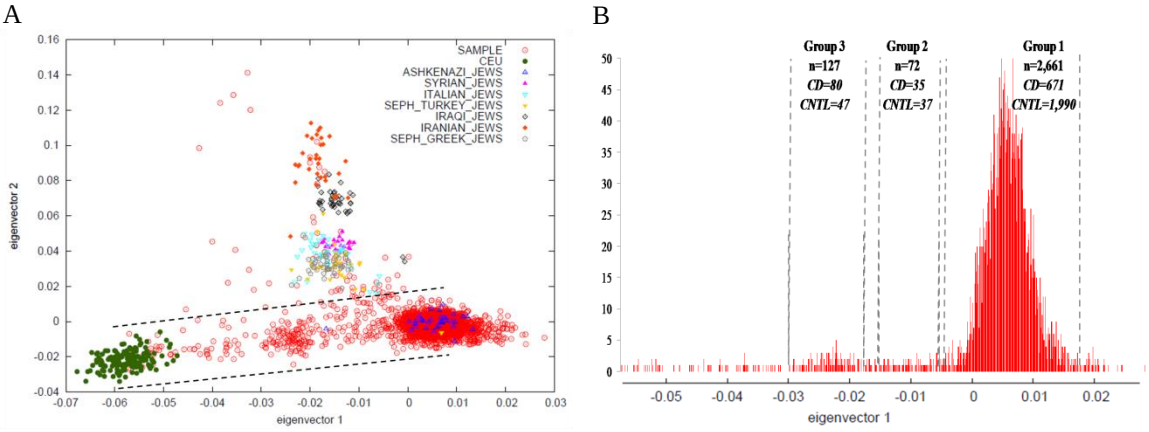


Figure 6.2 Clustering along PC1 for people with AJ (to the right) and without and Jewish (to the left) ancestry. Samples that clustered along PC1 between the AJ cluster (bottom right) and CEU cluster (bottom left) are shown between the dotted grey lines (A). A histogram of PC1 values for the included samples show three distinct modes (Groups 1-3).

.

would be under powered for detecting CD driven signal. To optimize the power in group 3, and under the assumption that group 3 comprised individuals that were 50% AJ/50% non-Jewish ancestry, we added 100 randomly drawn controls from group 1 and 100 randomly select non-Jewish controls from the HapMap CEU population to group three. The final count of CD cases/Controls in each group was; group 1 (100% AJ) 671/1,890, group 2 (75% AJ) 35/37 and group 3 (50% AJ) 80/247.

Constructing an Ashkenazi genotype reference panel

Due to low imputation quality for AJ samples imputed from any of the standard HapMap reference panels, we opted to construct a custom AJ reference panel for imputation. We selected a random 100 AJ samples from the NY6 cohort who had been genotyped on the Affymetrix 6.0 platform and also typed those individuals on the Illumina 1M platform.

	Total	Discordant
SNP's	268,907	15,956
SNP's X Ind's	24,062,937	31,354
Total concordance		0.9987

Table 6.2 SNP concordance between platforms.

The raw data consisted of 100 individuals typed for 1,778,360 unique SNP's across both platforms. To test for difference in the quality of genotyping on the two platforms, we examined the concordance of the overlapping SNP's. Of the 268,907 SNP's present on both platforms, >99% were concordant across all individuals between platforms (excluding SNP's with 1 or 2 missing genotypes on either platform and 'ambiguous SNP' that are A/T or C/G on the Affymetrix platform). Two individuals who were <99.5% concordance were excluded from the reference panel

(see **Figure 6.3(A)**). In addition, 263/268,907 were <0.9 concordant between the platform, which extrapolates to an estimated <0.0001 SNP discordance between the two platforms (see **Figure 6.3(B)**). To ensure the highest quality of the AJ reference panel we

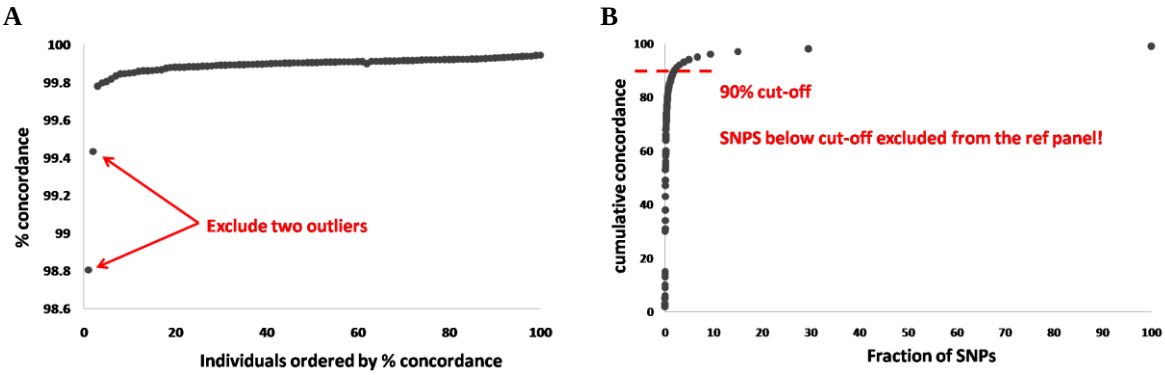


Figure 6.3 Individual (A) and SNP (B) concordance between the Affymetrix and Illumina platforms.

also excluded SNPs that were homozygous heterozygote, had a minor allele frequency (MAF) <0.02 and a genotype missingness of >0.02. For those SNP's that had ≤ 0.02 SNP missingness, we ran imputed the missing markers using BEAGLE (Browning, 2007) across the reference panel. The final reference panel consisted of 98 individuals, and 1,328, 536 SNP's at 0% missingness.

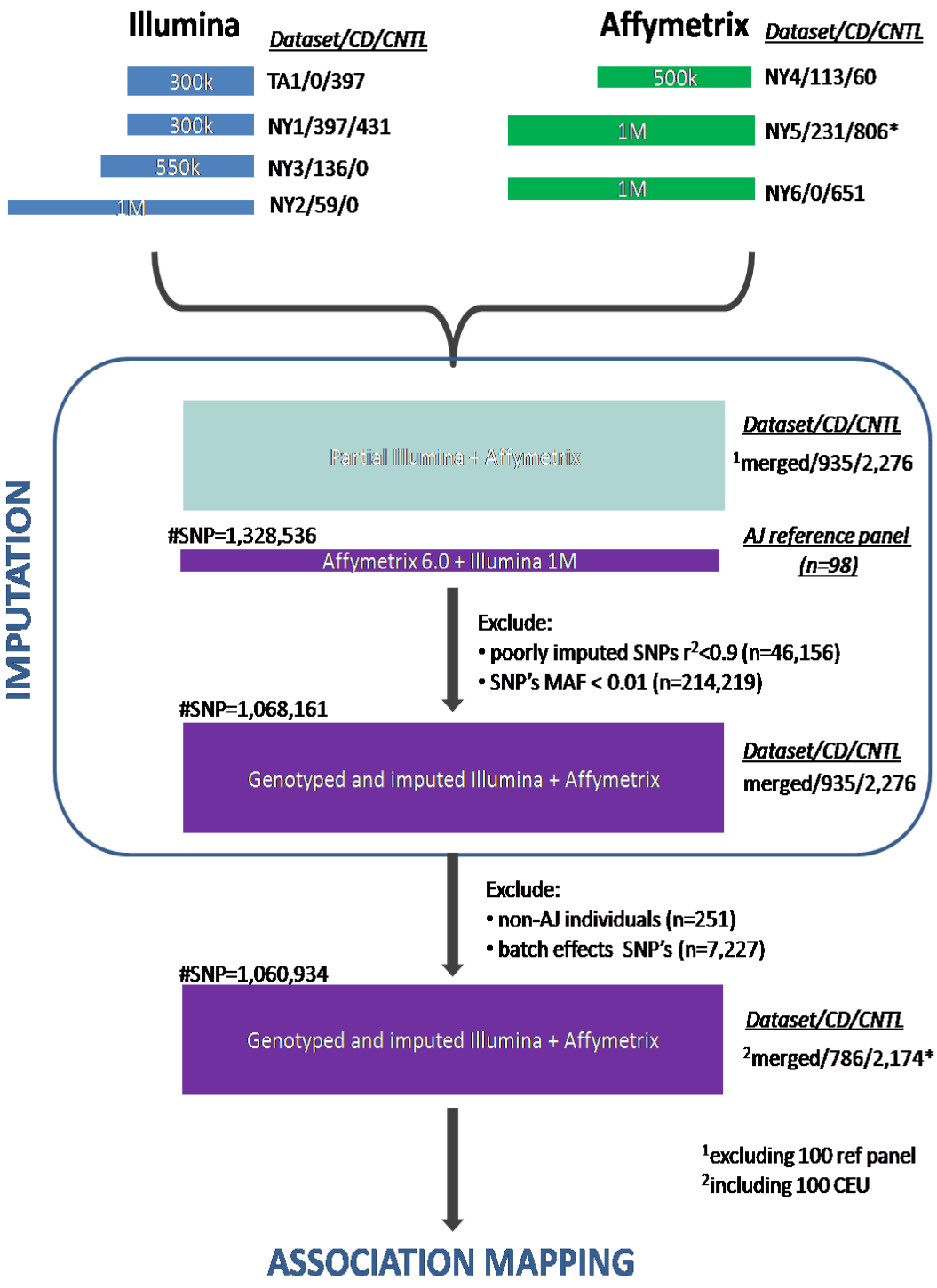


Figure 6.4 Schema of combined analysis quality control pipeline.

Imputation of genotypes from the Ashkenazi reference panel

The AJ reference panel was then used as a training set to impute missing genotypes for the combined NY1-6 and TA1 datasets using BEAGLE (Browning, 2007). SNP's that were either low in frequency (MAF <0.01) or had a poor imputation quality score (r^2<0.9) were excluded (see **Figure 6.4**). The final combined dataset of genotyped and imputed SNP's contained 1,068,161 SNP's across 3,211 AJ samples (935 cases and 2,276 controls).

Association mapping of CD in Ashkenazi Jews

Prior to performing association mapping, non-AJ samples as determined by PCA analysis (see above) were excluded from the sample (n=251). In addition, to control for batch effects across the seven cohorts, markers that were association $p<10^{-5}$ in a pairwise cohort-cohort association (stratified by CD case status), were also excluded (n=7,227), **Figure 6.4**. The final filtered dataset used for association mapping comprised 1,060,934 genotyped and imputed markers across 2,960 individuals, divided into three groups according to AJ ancestry (see above).

For association mapping, we first built a kinship matrix of pairwise identity-by-state metrics based the high-density markers for the entire sample, which was incorporated

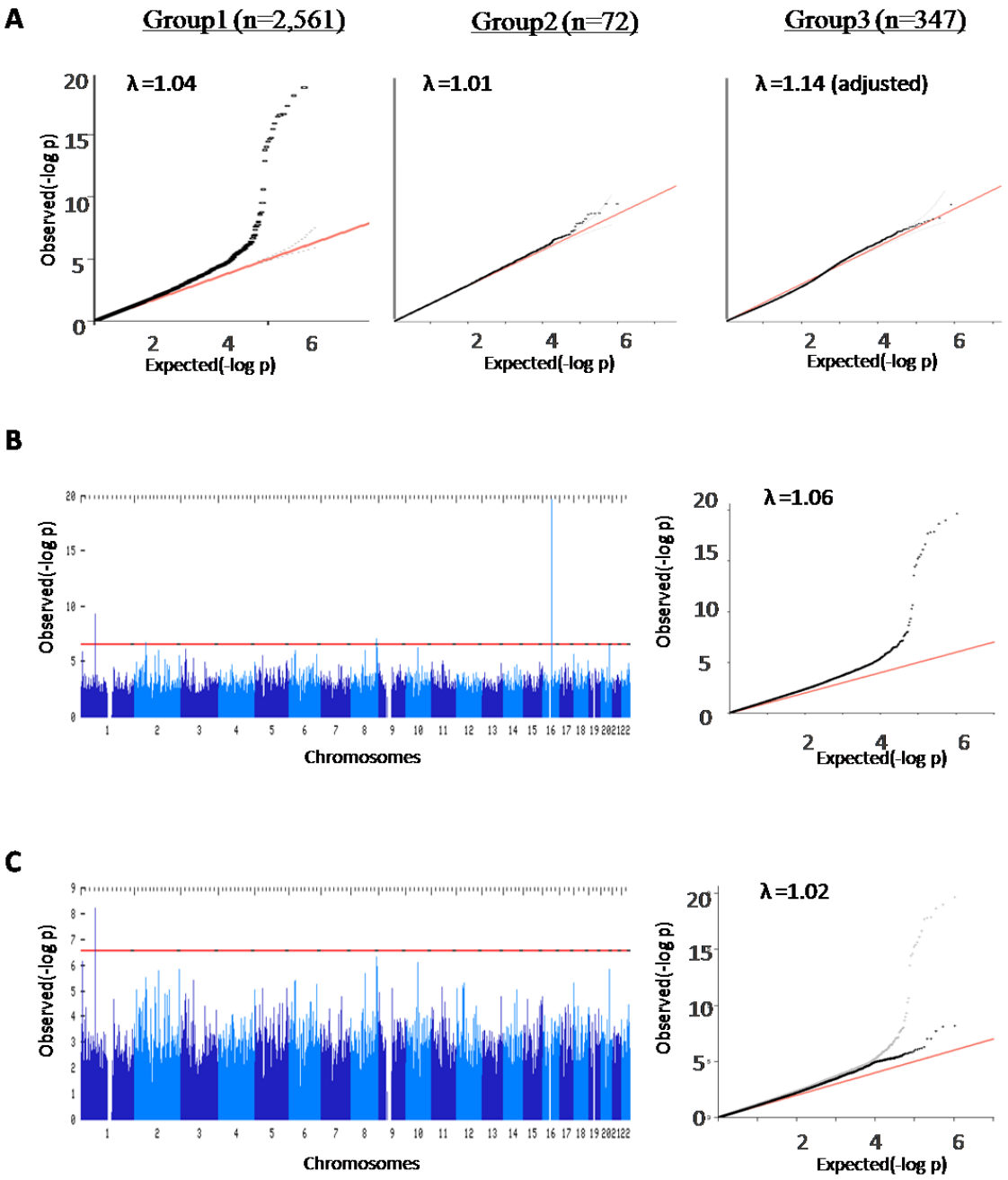


Figure 6.5 Association mapping of Crohn Disease in Ashekenazi Jews. (A) QQ-plots of the 100%, 75% and 50% AJ ancestry groups (groups 1,2 and 3, respectively). The number of individuals in each group is shown and the inflation factor for the p-value distribution is given. In the case of group 3, the p-values were genomic control adjusted for overinflation. (B) A manhattan plot (left) and qq-plot (right) of the combined association score from all three groups. (C) The same as B, but with 526 markers around the region of NOD2 signal on chromosome 16 removed before association mapping. The qq-plot shows the p-value distribution before (gray) and after (black) NOD2 removal and the inflation factor is for the latter.

into the mixed model for association mapping for the three groups (Kang, 2010). The p-values were combined across the three groups using METAL, which takes direction of effect into account and weights the groups by sample size (Willer, 2008). **Figure 6.5(A)** shows the qq-plots for 100%, 75% and 50% AJ ancestry groups (groups 1,2 and 3, respectively). In the case of group 3, the p-values were quite overinflated (λ=1.14) and were corrected by genomic control to approximate normality (Devlin, 1999).

Observation of positive controls in Ashkenazi Jews

Figure 6.5(B) shown the combined score from all three groups, and as positive controls, we observe two genomic regions exceeding our genome wide significance threshold of $5x10^{-8}$. Top SNPs are in known loci for Crohn Disease, *NOD2* (16q12*;* rs2076756; $p<2.32x10^{-20}$) and *IL23R* (1p31*;* rs11209026; $p<9.42x10^{-9}$) (Barrett, 2008). In addition, we see 6 other signals for known loci $p<10^{-4}$; *ATG16L1* (2q37; rs2241880; $p<2.88x10^{-6}$), *IL3* (5q31; rs3091338; $p<4.86x10^{-5}$), *ZNF365*(10q21; rs1076165; $p<9.31x10^{-6}$), intergenic region >300kb upstream of *PTGER4* (5p13.1; rs9292777; $6.92x10^{-5}$), *JAK2* (9p24.1;rs2230724; $p<8.11x10^{-5}$) and *NKX2-3* (10q24.2; rs11190141; $9.8x10^{-5}$) (Barrett, 2008)(Parkes, 2007). Finally we observed an interesting novel region approaching genome wide significance on chromosome 8q24 rs2922500 (chr8:134,676,639; $p<8.36x10^{-8}$). The SNP is in an intergenic region and resides in a known in large deletion spanning chr8:134675707-134680644.

Haplotypic fine mapping of the NOD2 locus

We performed haplotypic mapping on the broad peak at the *NOD2*. Haplotype mapping has been previously described in **Chapter 2**. Briefly, common shared long-range haplotypes were detected via pair-wise IBD genetic segment matching between all individuals in the pedigree using the GERMLINE software (Gusev, 2009). The IBD segments were at least 3cM in length and allowed for up to 1% mismatching due to genotyping error. IBD segments were then clustered to groups of similar haplotypes, with >500kb overlapping sequence and sharing >99% sequence identity, as described in detail previously (Kenny, 2009). Clusters of shared haplotypes were then independently mapped to Crohn's disease which identified a single strong mapping cluster, where the specific sharing boundaries of the 253kb haplotype mapped to a sub-region (chr16:48.1-48.4) of the Crohn's GWAS signal peak, ~1MB upstream of the top SNP in the region. This novel "CD haplotype" (rs1420682 to rs11076503) is unique to a group of 396 individuals (carrier frequency 0.13), is strongly associated to CD ($p<8.1x10^{-17}$), and has an odds ratio of 1.47. A logistical regression of CD with the CD haplotype as a covariate has little effect on the signal for the top *NOD2* SNP ($p<1.45x10^{-17}$). Likewise, a logistical regression of CD with the top *NOD2* SNP (rs2076756) as a covariate has little effect on the CD haplotype ($p<3.45x10^{-15}$), indicating that the two signals are independently associated to CD. A genome-wide scan of Crohn's disease using both the CD haplotype and rs2076756 revealed no remaining genome-wide significant signal in the region ($p<10^{-4}$), indicating that rs2076756 and the CD haplotype likely capture the full signal in the region.

Discussion

We describe a GWAS for Crohn's disease across seven Ashkenazi Jewish cohorts using a combined analysis strategy and a population specific reference panel for imputation. The analysis yielded 8 positive controls, 2 of which were genome-wide significant, and an interesting novel region of borderline significance on chromosome 8q24. More importantly, haplotypic fine mapping of the strong signal peak on chromosome 16 revealed a novel haplotype ~1MB upstream of the signal that was independent on the top SNP.

Chapter 7 : Discussion

Implementing improved analytical strategies for founder populations

The goal of this thesis was to explore the opportunities for analyzing founder populations to detect causal variants affecting complex disease. The idea was to leverage the unique features of such populations for exploring genetic variations, in particular rare variations, while overcoming the technical challenges inherent to their genetic architecture. We first tested specialized strategies in a small, highly related, founder population. We then demonstrated the extensibility of these approaches by successfully applying them in a larger, less extreme, founder population.

Improved methodology for genotypic mapping

GWAS of samples containing extensive relatedness, which is usually the case for founder populations, have been traditionally analyzed using family-based methods (Gudmundsson, 2007, Ober, 2001, Lowe, 2009). However, extensive analysis at a single locus level has previously shown that family-based tests have lower power than population-based methods, such as mixed-models, in the presence multiple levels of relatedness (Yu, 2006), and we have observed this on Kosrae (Lowe, 2009). The statistical advantage of mixed-models comes at the cost of higher model complexity,

resulting in a lack, until recently, of mixed model methods capable of handling large-scale human datasets in a computationally efficient manner.

Here we have demonstrated the improved power for mixed models in the context of the extensive, multi-generational relatedness found on the Island of Kosrae. We focused specifically on a range of p-values not yet explored and showed that mixed models gave a >2-fold boost in power over other methods. Importantly for the reduction of false positives, the test statistic produced is uniformly distributed, with no artificial overdispersion of the p-value. The improvement in power is due in part to the utilization of variation both within and between samples (captured in the kinship matrix) and also the ability to include all samples in the analysis. Our empirical results on Kosrae show similar power to like-sized outbred populations and support the use of mixed-models for association in founder populations.

Novel methodology for haplotypic mapping

Much of the work for association analysis has been focused on the effects of single SNPs. We provide a more extensive approach that considers long range haplotypes, identical-by-descent, across multiple samples, as genetic markers. Such recently arising haplotypes are likely to carry rare mutations, which are the focus of sequencing endeavors (see below). As proof of the pudding, we provide examples for three novel signals, mapped by

association to long range haplotypes on Kosrae, with two casual mutations confirmed by sequencing.

This approach was aided by a high degree of relatedness and founder effects on Kosrae that resulted in abundant, long haplotypes in our study population (Bonnen, 2006). However, recently, tracks of relatedness in regions of the genome have also been seen in less extreme founder population (Kong, 2008), and even in apparently unrelated populations (Frazer, 2007). We therefore tested the haplotype mapping approach in a canonical founder population, with less extreme founding effects than are seen on Kosrea, the Ashkenazi Jewish population. We demonstrated the extensibility of our method in this less founder population by mapping a novel haplotype for Crohn's disease.

A key advantage of haplotypic mapping compared to genotypic mapping, is the additional positional information, where the physical haplotype boundaries, as well as the specific carriers, are explicitly identified – a useful roadmap of follow up sequencing. We also demonstrated in three of the cases that our method can be used to dissect multiple signals at the same locus. In addition, since the long shared haplotypes were constructed from SNP-chip data, this approach can complement a traditional association mapping method. However, the method is limited to resolving very long haplotypes < 20 generations old. If a causal mutation arose more distantly and surrounding haplotypes are too short the method will not have any mapping power.

New biological insights

Overall, a GWAS, using mixed models, of 30 metabolic traits on Kosrae yielded 13 positive controls and 4 novel loci for height, haemoglobin A1C, uric acid and plasma plant sterol levels, 3 of the latter being refined by haplotype mapping. The same approaches applied in an Ashkenazi Jewish cohort for a single trait, produced two positive controls, a novel hit region and a novel independent associated haplotype at one of the positive control loci.

Effect sizes in founder populations

To examine the genetics of common traits on Kosrae as compared to outbred populations, we assessed the effect sizes observed on Kosrae for known loci discovered in studies using outbred, mainly Caucasian populations. We showed that some variants that are associated to traits in other populations do not replicate on Kosrae, due to bottleneck effects rather than lack of power. On the other hand, observed associations show strong effects and statistical support for a cohort of this size (with the caveat that the number of markers we were able to compare is small). While this empirical observation is open to interpretation, and somewhat anecdotal, it does indicate the benefit of increased genetic and environmental homogeneity for variant discovery.

Detection of rare variation

Using the haplotypic mapping method, it was possible to detect a relatively rare (minor allele frequency < 1%) causal genetic variants. Expanding the search for associated alleles to variants that are rare in the general population provides another piece of the heritable component of disease, with multiple gene sequencing studies implicating such variants with moderate to high penetrance (Fearnhead, 2004, Cohen, 2004, Kotowski, 2006, Romeo, 2007, Marini, 2008). The new era of high throughput sequencing returns the emphasis to rare variation, playing to the strength of isolated communities.

Limitations of founder populations studies

Replication of results in the hands of a different researcher is a cornerstone of good science. Genetic association signals for alleles that are common and shared across populations and continents have been successfully replicated. In the Kosraen study, such replication is expected to be more challenging, especially for rare, newly arisen variants, as some variants are expected to be limited to Pacific islanders or East Asians that are under-represented in large association studies. We rely on increased participation of these groups in such studies, facilitating replication in the future. Furthermore, as linkage studies have succeeded to establish over the last two decades, there are multiple options to support a genetic effect of a variant, even when it is a family-specific mutation. These criteria should hold for the "extended family" of Kosrae. Mutations can be supported by

evidence from annotation, other variants in the same gene, or pathways with observed enrichment of causal variation and functional follow-up in the lab.

Imputation of unobserved genotypes of common alleles has been previously used in general, outbred populations. This approach relies on the ancient shared origin of an imputed sample and reference individuals, forfeiting the option to impute the low-frequency alleles that have arisen since their common ancestor. The strategy requires the availability of a densely genotyped reference panel, such as the HapMap project (Frazer, 2007). However, there is reduced portability of this approach to populations that are, compared to available reference panels, genetically distinct, such as founder populations. This leads to a reduction in imputation accuracy, and a consequent reduction in statistical power for imputation based association mapping. We have observed poor imputation power for both founder populations used in this study, especially on Kosrae. This problem is expected to be somewhat alleviated as more populations are added to the existing referencing panels, such as the Jewish HapMap (pers.comm. Harry Ostrer). In addition, the abundance of homozygosity (Morrow, 2008) and long-range haplotypes (Kong, 2008) found in population isolates will facilitate the next generation of tools for the detection of rare disease-associated haplotypes. For example, methods for imputation based on IBD segments are expected to capture many recently arising mutations, thus leveraging the power of sequencing technologies to observe new variation.

Future directions

Hand in hand with the explosion of genotype data in the past few years, many exciting technical advances have been made in the area of sequencing. Recently, there has been a fundamental shift away from traditional Sanger sequencing towards so-called "next-generation sequencing" (NGS) technologies for genome analysis. These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods (Metzker, 2010). The production of large numbers of low-cost reads make the NGS platforms useful for many applications, such as cataloguing transcriptomes (RNA-sequencing) or epigenetic marks (ChIP-sequencing, Methyl-sequencing), and re-sequencing target areas of a genome. The broadest application of NGS is in the sequencing of human whole-genomes to enhance our understanding of how genetic differences affect health and disease.

Many of the initial whole-genome sequencing studies have focused on proof-of-technology, and presently, high-quality personal genomes exist from representatives of the major continental groups (Wheeler, 2008, Ahn, 2009, Bentley, 2008, Kim, 2009, Levy, 2007, McKernan, 2009, Wang, 2008). The major advance offered by NGS is the ability to produce enormous volume of data cheaply, the lowest published cost being ~$5K for a whole human genome (Drmanac, 2010), down from $1 in 2004 to a small fraction of a cent (<$0.000002) per base pair today.

Current approaches for sequence-based mapping of human disease included whole-exome capture in a small number of individuals. This approach has been successful in identifying causal variants in monogenic disease cases (Ng, 2010). However, such studies ignore non-coding regions and do not yet scale to large populations (Galvan, 2010). For complex traits, a popular approach is the targeted re-sequencing of candidate loci under local signal peaks detected in GWAS across many individual (Cohen, 2004). However, this strategy scales poorly for multiple traits and across a large number of loci and pursuing such a strategy genome-wide is still resource intensive, despite a considerable drop in sequencing costs. We suggest an alternative paradigm for population-based sequencing that plays to the strengths of a small, extreme isolate population, such as Kosrae.

A population-based paradigm for sequencing in founder populations

Our strategy leverages the current multi-trait GWAS data and the inherent relatedness of the cohort with the resultant enrichment of IBD long genetic segments on the island. We devise an approach to sequence a small number of individuals, forming a population specific reference panel of the island. In conjunction, we can use the GWAS SNP's to construct a genomewide, population-wide catalogue of long segments that are IBD across individuals. Finally, we can propagate any reference panel variants to non-sequenced

individuals through imputation based on the IBD segments for mapping across multiple traits.

Our intial estimates show that sequencing ~80 individuals selected for most IBD sharing on Kosrae will capture > 80% of the total variation on the island. In other words, a modest 0.25TB sized reference panel has the potential to yield >7TB of imputed sequence. This means that in Kosrae, we can cover the entire genome sequence with resources that are sufficient for only targeted, candidate gene sequencing of similar sample sizes in most populations, as they lack these characteristics of isolation. Analysis of this genetic data in conjunction with phenotypic data in Kosrae holds the potential pinpoint novel causal variants for multiple traits.

Conclusion

This study of founder populations has highlighted the usefulness of understanding the population-genetic properties of a particular study population for recognizing the features and limitations of disease mapping in that population. As the field of human genetics moves forward to explore new and expanded sources of variation, models of these population-genetic features offers a context with which to interpret the data. Because the full genetic history of the human population is unknown, population-genetic models can be used instead to generate plausible theories. Such models begin from the perspective that the factors that affect the genealogical descent of a disease mutation, such as

bottlenecks, migrations and the recombination landscape, ultimately affect the distribution of the variant across contemporary individuals and populations. Thus, the properties of risk variants simulated under such models can be used to evaluate strategies for detecting them. Further, the models themselves will be calibrated and improved as deeper and richer human genetic data becomes available.

References

http://www.1000genomes.org

http://www.hapmap.org

The Human Genome Project (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.

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

Abecasis,G.R., Cardon,L.R., Cookson,W.O. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279-292.

Abecasis,G.R., Cardon,L.R., Cookson,W.O., Sham,P.C., Cherny,S.S. (2001) Association analysis in a variance components framework. *Genet. Epidemiol.*, **21 Suppl 1**, S341-S346.

Abelson,A.K., Delgado-Vega,A.M., Kozyrev,S.V., Sanchez,E., Velazquez-Cruz,R., Eriksson,N., Wojcik,J., Linga Reddy,M.V., Lima,G., D'Alfonso,S., *et al.* (2009) STAT4 associates with systemic lupus erythematosus through two independent effects that correlate with gene expression and act additively with IRF5 to increase risk. *Ann. Rheum. Dis.*, **68**, 1746-1753.

Abney,M., McPeek,M.S., Ober,C. (2000) Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet.*, **66**, 629-650.

Abney,M., McPeek,M.S., Ober,C. (2001) Broad and narrow heritabilities of quantitative traits in a founder population. *Am. J. Hum. Genet.*, **68**, 1302-1307.

Ahn,S.M., Kim,T.H., Lee,S., Kim,D., Ghang,H., Kim,D.S., Kim,B.C., Kim,S.Y., Kim,W.Y., Kim,C., *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.*, **19**, 1622-1629.

Albrechtsen,A., Sand,K.T., Moltke,I., van Overseem,H.T., Nielsen,F.C., Nielsen,R. (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.*, **33**, 266-274.

Almasy,L., Blangero,J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198-1211.

Angius,A., Bebbere,D., Petretto,E., Falchi,M., Forabosco,P., Maestrale,B., Casu,G., Persico,I., Melis,P.M., Pirastu,M. (2002) Not all isolates are equal: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian sub-populations. *Hum. Genet.*, **111**, 9-15.

Arcos-Burgos,M., Muenke,M. (2002) Genetics of population isolates. *Clin. Genet.*, **61**, 233-247.

Aulchenko,Y.S., Vaessen,N., Heutink,P., Pullen,J., Snijders,P.J., Hofman,A., Sandkuijl,L.A., Houwing-Duistermaat,J.J., Edwards,M., Bennett,S., *et al.* (2003) A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the Netherlands. *Diabetes*, **52**, 3001-3004.

Aulchenko,Y.S., Heutink,P., Mackay,I., Bertoli-Avella,A.M., Pullen,J., Vaessen,N., Rademaker,T.A., Sandkuijl,L.A., Cardon,L., Oostra,B., van Duijn,C.M. (2004) Linkage disequilibrium in young genetically isolated Dutch population. *Eur. J. Hum. Genet.*, **12** , 527-534.

Aulchenko,Y.S., de Koning,D.J., Haley,C. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577-585.

Aulchenko,Y.S., Ripatti,S., Lindqvist,I., Boomsma,D., Heid,I.M., Pramstaller,P.P., Penninx,B.W., Janssens,A.C., Wilson,J.F., Spector,T., *et al.* (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.*, **41**, 47-55.

Bar-Sade,R.B., Kruglikova,A., Modan,B., Gak,E., Hirsh-Yechezkel,G., Theodor,L., Novikov,I., Gershoni-Baruch,R., Risel,S., Papa,M.Z., *et al.* (1998) The 185delAG BRCA1 mutation originated before the dispersion of Jews in the diaspora and is not limited to Ashkenazim. *Hum. Mol. Genet.*, **7**, 801-805.

Barrett,J.C., Hansoul,S., Nicolae,D.L., Cho,J.H., Duerr,R.H., Rioux,J.D., Brant,S.R., Silverberg,M.S., Taylor,K.D., Barmada,M.M., *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955-962.

Behar,D.M., Metspalu,E., Kivisild,T., Achilli,A., Hadid,Y., Tzur,S., Pereira,L., Amorim,A., Quintana-Murci,L., Majamaa,K., *et al.* (2006) The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. *Am. J. Hum. Genet.*, **78**, 487-497.

Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R., *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59.

Berge,K.E., Tian,H., Graf,G.A., Yu,L., Grishin,N.V., Schultz,J., Kwiterovich,P., Shan,B., Barnes,R., Hobbs,H.H. (2000) Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science*, **290**, 1771-1775.

Bodmer,W., Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695-701.

Bonnen,P.E., Pe'er,I., Plenge,R.M., Salit,J., Lowe,J.K., Shapero,M.H., Lifton,R.P., Breslow,J.L., Daly,M.J., Reich,D.E., *et al.* (2006) Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat. Genet.*, **38**, 214-217.

Bonnen,P.E., Lowe,J.K., Altshuler,D.M., Breslow,J.L., Stoffel,M., Friedman,J.M., Pe'er,I. (2010) European admixture on the Micronesian island of Kosrae: lessons from complete genetic information. *Eur. J. Hum. Genet.*, **18**, 309-316.

Bourgain,C., Genin,E. (2005) Complex trait mapping in isolated populations: Are specific statistical methods required? *Eur. J. Hum. Genet.*, **13**, 698-706.

Browning,S.R. (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.*, **124**, 439-450.

Burfoot,R.K., Jensen,C.J., Field,J., Stankovich,J., Varney,M.D., Johnson,L.J., Butzkueven,H., Booth,D., Bahlo,M., Tait,B.D., *et al.* (2008) SNP mapping and candidate gene sequencing in the class I region of the HLA complex: searching for multiple sclerosis susceptibility genes in Tasmanians. *Tissue Antigens*, **71**, 42-50.

Burkhardt,R., Kenny,E.E., Lowe,J.K., Birkeland,A., Josowitz,R., Noel,M., Salit,J., Maller,J.B., Pe'er,I., Daly,M.J., *et al.* (2008) Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arterioscler. Thromb. Vasc. Biol.*, **28**, 2078-2084.

Burnett,J.R., Hooper,A.J. (2008) Common and Rare Gene Variants Affecting Plasma LDL Cholesterol. *Clin. Biochem. Rev.*, **29**, 11-26.

Cameron,A.J., Shaw,J.E., Zimmet,P.Z. (2004) The metabolic syndrome: prevalence in worldwide populations. *Endocrinol. Metab Clin. North Am.*, **33**, 351-75, table.

Chan,Y.M., Varady,K.A., Lin,Y., Trautwein,E., Mensink,R.P., Plat,J., Jones,P.J. (2006) Plasma concentrations of plant sterols: physiology and relationship with coronary heart disease. *Nutr. Rev.*, **64**, 385-402.

Chasman,D.I., Pare,G., Mora,S., Hopewell,J.C., Peloso,G., Clarke,R., Cupples,L.A., Hamsten,A., Kathiresan,S., Malarstig,A., *et al.* (2009) Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS. Genet.*, **5**, e1000730.

Clark,A.G., Hubisz,M.J., Bustamante,C.D., Williamson,S.H., Nielsen,R. (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.*, **15**, 1496-1502.

Cohen,J.C., Kiss,R.S., Pertsemlidis,A., Marcel,Y.L., McPherson,R., Hobbs,H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869-872.

Daly,A.K., Donaldson,P.T., Bhatnagar,P., Shen,Y., Pe'er,I., Floratos,A., Daly,M.J., Goldstein,D.B., John,S., Nelson,M.R., *et al.* (2009) HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.*, **41**, 816-819.

de Bakker,P.I., Yelensky,R., Pe'er,I., Gabriel,S.B., Daly,M.J., Altshuler,D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217-1223.

Dehghan,A., Kottgen,A., Yang,Q., Hwang,S.J., Kao,W.L., Rivadeneira,F., Boerwinkle,E., Levy,D., Hofman,A., Astor,B.C., *et al.* (2008) Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet*, **372**, 1953-1961.

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

Diaz,G.A., Gelb,B.D., Risch,N., Nygaard,T.G., Frisch,A., Cohen,I.J., Miranda,C.S., Amaral,O., Maire,I., Poenaru,L., *et al.* (2000) Gaucher disease: the origins of the Ashkenazi Jewish N370S and 84GG acid beta-glucosidase mutations. *Am. J. Hum. Genet.* , **66**, 1821-1832.

Dickson,S.P., Wang,K., Krantz,I., Hakonarson,H., Goldstein,D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS. Biol.*, **8**, e1000294.

Drmanac,R., Sparks,A.B., Callow,M.J., Halpern,A.L., Burns,N.L., Kermani,B.G., Carnevali,P., Nazarenko,I., Nilsen,G.B., Yeung,G., *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78-81.

Duerr,R.H., Taylor,K.D., Brant,S.R., Rioux,J.D., Silverberg,M.S., Daly,M.J., Steinhart,A.H., Abraham,C., Regueiro,M., Griffiths,A., *et al.* (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461-1463.

Durrant,C., Zondervan,K.T., Cardon,L.R., Hunt,S., Deloukas,P., Morris,A.P. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.*, **75**, 35-43.

Durst,R., Colombo,R., Shpitzen,S., Avi,L.B., Friedlander,Y., Wexler,R., Raal,F.J., Marais,D.A., Defesche,J.C., Mandelshtam,M.Y., *et al.* (2001) Recent origin and spread of a common Lithuanian mutation, G197del LDLR, causing familial hypercholesterolemia: positive selection is not always necessary to account for disease incidence among Ashkenazi Jews. *Am. J. Hum. Genet.*, **68**, 1172-1188.

Elliott,P., Chambers,J.C., Zhang,W., Clarke,R., Hopewell,J.C., Peden,J.F., Erdmann,J., Braund,P., Engert,J.C., Bennett,D., *et al.* (2009) Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA*, **302**, 37-48.

Ervin,R.B. (2009) Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. *Natl. Health Stat. Report.*, 1-7.

Fassbender,K., Lutjohann,D., Dik,M.G., Bremmer,M., Konig,J., Walter,S., Liu,Y., Letiembre,M., von,B.K., Jonker,C. (2008) Moderately elevated plant sterol levels are associated with reduced cardiovascular risk--the LASA study. *Atherosclerosis*, **196**, 283-288.

Fearnhead,N.S., Wilding,J.L., Winney,B., Tonks,S., Bartlett,S., Bicknell,D.C., Tomlinson,I.P., Mortensen,N.J., Bodmer,W.F. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl. Acad. Sci. U. S. A*, **101**, 15992-15997.

Fellay,J., Shianna,K.V., Ge,D., Colombo,S., Ledergerber,B., Weale,M., Zhang,K., Gumbs,C., Castagna,A., Cossarizza,A., *et al.* (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science*, **317**, 944-947.

Frayling,T.M., Timpson,N.J., Weedon,M.N., Zeggini,E., Freathy,R.M., Lindgren,C.M., Perry,J.R., Elliott,K.S., Lango,H., Rayner,N.W., *et al.* (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889-894.

Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M., *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-861.

Frazer,K.A., Murray,S.S., Schork,N.J., Topol,E.J. (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241-251.

Freimer,N.B., Reus,V.I., Escamilla,M., Spesny,M., Smith,L., Service,S., Gallegos,A., Meza,L., Batki,S., Vinogradov,S., *et al.* (1996) An approach to investigating linkage for bipolar disorder using large Costa Rican pedigrees. *Am. J. Med. Genet.*, **67**, 254-263.

Friedlaender,J.S., Friedlaender,F.R., Reed,F.A., Kidd,K.K., Kidd,J.R., Chambers,G.K., Lea,R.A., Loo,J.H., Koki,G., Hodgson,J.A., *et al.* (2008) The genetic structure of Pacific Islanders. *PLoS. Genet.*, **4**, e19.

Fulker,D.W., Cherny,S.S., Sham,P.C., Hewitt,J.K. (1999) Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.*, **64**, 259-267.

Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M., *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225-2229.

Galvan,A., Ioannidis,J.P., Dragani,T.A. (2010) Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet.*, **26**, 132-141.

Graf,G.A., Yu,L., Li,W.P., Gerard,R., Tuma,P.L., Cohen,J.C., Hobbs,H.H. (2003) ABCG5 and ABCG8 are obligate heterodimers for protein trafficking and biliary cholesterol excretion. *J. Biol. Chem.*, **278**, 48275-48282.

Graham,R.R., Kyogoku,C., Sigurdsson,S., Vlasova,I.A., Davies,L.R., Baechler,E.C., Plenge,R.M., Koeuth,T., Ortmann,W.A., Hom,G., *et al.* (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. U. S. A*, **104**, 6758-6763.

Graham,R.R., Cotsapas,C., Davies,L., Hackett,R., Lessard,C.J., Leon,J.M., Burtt,N.P., Guiducci,C., Parkin,M., Gates,C., *et al.* (2008) Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat. Genet.*, **40**, 1059-1061.

Gray,R.D., Drummond,A.J., Greenhill,S.J. (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, **323**, 479-483.

Gudbjartsson,D.F., Walters,G.B., Thorleifsson,G., Stefansson,H., Halldorsson,B.V., Zusmanovich,P., Sulem,P., Thorlacius,S., Gylfason,A., Steinberg,S., *et al.* (2008) Many sequence variants affecting diversity of adult human height. *Nat. Genet.*, **40**, 609-615.

Gudmundsson,G., Matthiasson,S.E., Arason,H., Johannsson,H., Runarsson,F., Bjarnason,H., Helgadottir,K., Thorisdottir,S., Ingadottir,G., Lindpaintner,K., *et al.* (2002) Localization of a gene for peripheral arterial occlusive disease to chromosome 1p31. *Am. J. Hum. Genet.*, **70**, 586-592.

Gudmundsson,J., Sulem,P., Manolescu,A., Amundadottir,L.T., Gudbjartsson,D., Helgason,A., Rafnar,T., Bergthorsson,J.T., Agnarsson,B.A., Baker,A., *et al.* (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.*, **39**, 631-637.

Gudmundsson,J., Sulem,P., Gudbjartsson,D.F., Jonasson,J.G., Sigurdsson,A., Bergthorsson,J.T., He,H., Blondal,T., Geller,F., Jakobsdottir,M., *et al.* (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat. Genet.*, **41**, 460-464.

Gulcher,J., Kong,A., Stefansson,K. (2001) The genealogic approach to human genetics of disease. *Cancer J.*, **7**, 61-68.

Gulcher,J.R., Kong,A., Stefansson,K. (2001) The role of linkage studies for common diseases. *Curr. Opin. Genet. Dev.*, **11**, 264-267.

Gusev,A., Lowe,J.K., Stoffel,M., Daly,M.J., Altshuler,D., Breslow,J.L., Friedman,J.M., Pe'er,I. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, **19**, 318-326.

Hafler,J.P., Maier,L.M., Cooper,J.D., Plagnol,V., Hinks,A., Simmonds,M.J., Stevens,H.E., Walker,N.M., Healy,B., Howson,J.M., *et al.* (2009) CD226 Gly307Ser association with multiple autoimmune diseases. *Genes Immun.*, **10**, 5-10.

Haiman,C.A., Patterson,N., Freedman,M.L., Myers,S.R., Pike,M.C., Waliszewska,A., Neubauer,J., Tandon,A., Schirmer,C., McDonald,G.J., *et al.* (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.*, **39**, 638-644.

Halperin,E., Stephan,D.A. (2009) SNP imputation in association studies. *Nat. Biotechnol.*, **27**, 349-351.

Heinemann,T., Axtmann,G., von,B.K. (1993) Comparison of intestinal absorption of cholesterol with different plant sterols in man. *Eur. J. Clin. Invest*, **23**, 827-831.

Heutink,P., Oostra,B.A. (2002) Gene finding in genetically isolated populations. *Hum. Mol. Genet.*, **11**, 2507-2515.

Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S., Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A*, **106**, 9362-9367.

Hirschhorn,J.N., Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95-108.

Hirschhorn,J.N. (2009) Genomewide association studies--illuminating biologic pathways. *N. Engl. J. Med.*, **360**, 1699-1701.

Horvath,S., Xu,X., Lake,S.L., Silverman,E.K., Weiss,S.T., Laird,N.M. (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet. Epidemiol.*, **26**, 61-69.

Hubacek,J.A., Berge,K.E., Cohen,J.C., Hobbs,H.H. (2001) Mutations in ATP-cassette binding proteins G5 (ABCG5) and G8 (ABCG8) causing sitosterolemia. *Hum. Mutat.*, **18**, 359-360.

Jakkula,E., Rehnstrom,K., Varilo,T., Pietilainen,O.P., Paunio,T., Pedersen,N.L., deFaire,U., Jarvelin,M.R., Saharinen,J., Freimer,N., *et al.* (2008) The genome-wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.*, **83**, 787-794.

Kallio,S.P., Jakkula,E., Purcell,S., Suvela,M., Koivisto,K., Tienari,P.J., Elovaara,I., Pirttila,T., Reunanen,M., Bronnikov,D., *et al.* (2009) Use of a genetic isolate to identify rare disease variants: C7 on 5p associated with MS. *Hum. Mol. Genet.*, **18**, 1670-1683.

Kang,H.M., Zaitlen,N.A., Wade,C.M., Kirby,A., Heckerman,D., Daly,M.J., Eskin,E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709-1723.

Kang,H.M., Sul,J.H., Service,S.K., Zaitlen,N.A., Kong,S.Y., Freimer,N.B., Sabatti,C., Eskin,E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348-354.

Kathiresan,S., Melander,O., Guiducci,C., Surti,A., Burtt,N.P., Rieder,M.J., Cooper,G.M., Roos,C., Voight,B.F., Havulinna,A.S., *et al.* (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.*, **40**, 189-197.

Kathiresan,S., Willer,C.J., Peloso,G.M., Demissie,S., Musunuru,K., Schadt,E.E., Kaplan,L., Bennett,D., Li,Y., Tanaka,T., *et al.* (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, **41**, 56-65.

Kayser,M., Lao,O., Saar,K., Brauer,S., Wang,X., Nurnberg,P., Trent,R.J., Stoneking,M. (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am. J. Hum. Genet.*, **82**, 194-198.

Keebler,M.E., Sanders,C.L., Surti,A., Guiducci,C., Burtt,N.P., Kathiresan,S. (2009) Association of blood lipids with common DNA sequence variants at 19 genetic loci in the

multiethnic United States National Health and Nutrition Examination Survey III. *Circ. Cardiovasc. Genet.*, **2**, 238-243.

Kenny,E.E., Gusev,A., Riegel,K., Lutjohann,D., Lowe,J.K., Salit,J., Maller,J.B., Stoffel,M., Daly,M.J., Altshuler,D.M., *et al.* (2009) Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. *Proc. Natl. Acad. Sci. U. S. A*, **106**, 13886-13891.

Kenny,E.E., Kim M, Gusev,A., *et al.* (2010) Increased power of mixed-models facilitates association mapping of 10 loci for metabolic traits in an isolated population. *Hum. Mol. Genet.* (submitted)

Kim,J.I., Ju,Y.S., Park,H., Kim,S., Lee,S., Yi,J.H., Mudge,J., Miller,N.A., Hong,D., Bell,C.J., *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011-1015.

Kim,J.J., Lee,H.I., Park,T., Kim,K., Lee,J.E., Cho,N.H., Shin,C., Cho,Y.S., Lee,J.Y., Han,B.G., *et al.* (2010) Identification of 15 loci influencing height in a Korean population. *J. Hum. Genet.*, **55**, 27-31.

Klein,R.J., Zeiss,C., Chew,E.Y., Tsai,J.Y., Sackler,R.S., Haynes,C., Henning,A.K., SanGiovanni,J.P., Mane,S.M., Mayne,S.T., *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385-389.

Kolz,M., Johnson,T., Sanna,S., Teumer,A., Vitart,V., Perola,M., Mangino,M., Albrecht,E., Wallace,C., Farrall,M., *et al.* (2009) Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS. Genet.*, **5**, e1000504.

Kong,A., Masson,G., Frigge,M.L., Gylfason,A., Zusmanovich,P., Thorleifsson,G., Olason,P.I., Ingason,A., Steinberg,S., Rafnar,T., *et al.* (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.*, **40**, 1068-1075.

Kotowski,I.K., Pertsemlidis,A., Luke,A., Cooper,R.S., Vega,G.L., Cohen,J.C., Hobbs,H.H. (2006) A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.*, **78**, 410-422.

Kruglyak,L. (2008) The road to genome-wide association studies. *Nat. Rev. Genet.*, **9**, 314-318.

Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191-203.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Lange,C., van,S.K., Andrew,T., Lyon,H., Demeo,D.L., Raby,B., Murphy,A., Silverman,E.K., MacGregor,A., Weiss,S.T., Laird,N.M. (2004) A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article17.

Lange,C., DeMeo,D., Silverman,E.K., Weiss,S.T., Laird,N.M. (2004) PBAT: tools for family-based association studies. *Am. J. Hum. Genet.*, **74**, 367-369.

Laramie,J.M., Wilk,J.B., DeStefano,A.L., Myers,R.H. (2007) HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. *Bioinformatics.*, **23**, 2190-2192.

Lee,C.M., Huxley,R.R., Woodward,M., Zimmet,P., Shaw,J., Cho,N.H., Kim,H.R., Viali,S., Tominaga,M., Vistisen,D., *et al.* (2008) Comparisons of metabolic syndrome definitions in four populations of the Asia-Pacific region. *Metab Syndr. Relat Disord.*, **6**, 37-46.

Lee,M.H., Lu,K., Hazard,S., Yu,H., Shulenin,S., Hidaka,H., Kojima,H., Allikmets,R., Sakuma,N., Pegoraro,R., *et al.* (2001) Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. *Nat. Genet.*, **27**, 79-83.

Lettre,G., Jackson,A.U., Gieger,C., Schumacher,F.R., Berndt,S.I., Sanna,S., Eyheramendy,S., Voight,B.F., Butler,J.L., Guiducci,C., *et al.* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.*, **40**, 584-591.

Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G., *et al.* (2007) The diploid genome sequence of an individual human. *PLoS. Biol.*, **5**, e254.

Lin,S., Chakravarti,A., Cutler,D.J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.*, **36**, 1181-1188.

Lindqvist,A.K., Steinsson,K., Johanneson,B., Kristjansdottir,H., Arnasson,A., Grondal,G., Jonasson,I., Magnusson,V., Sturfelt,G., Truedsson,L., *et al.* (2000) A susceptibility locus for human systemic lupus erythematosus (hSLE1) on chromosome 2q. *J. Autoimmun.*, **14**, 169-178.

Liu,Y., Blackwood,D.H., Caesar,S., de Geus,E.J., Farmer,A., Ferreira,M.A., Ferrier,I.N., Fraser,C., Gordon-Smith,K., Green,E.K., *et al.* (2010) Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Mol. Psychiatry*.

Loos,R.J., Lindgren,C.M., Li,S., Wheeler,E., Zhao,J.H., Prokopenko,I., Inouye,M., Freathy,R.M., Attwood,A.P., Beckmann,J.S., *et al.* (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.*, **40**, 768-775.

Lowe,C.E., Cooper,J.D., Brusko,T., Walker,N.M., Smyth,D.J., Bailey,R., Bourget,K., Plagnol,V., Field,S., Atkinson,M., *et al.* (2007) Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.*, **39**, 1074-1082.

Lowe,J.K., Maller,J.B., Pe'er,I., Neale,B.M., Salit,J., Kenny,E.E., Shea,J.L., Burkhardt,R., Smith,J.G., Ji,W., *et al.* (2009) Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS. Genet.*, **5**, e1000365.

Mackay,T.F. (2001) The genetic architecture of quantitative traits. *Annu. Rev. Genet.*, **35**, 303-339.

Manolio,T.A., Brooks,L.D., Collins,F.S. (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest*, **118**, 1590-1605.

Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.

Marini,N.J., Gin,J., Ziegle,J., Keho,K.H., Ginzinger,D., Gilbert,D.A., Rine,J. (2008) The prevalence of folate-remedial MTHFR enzyme variants in humans. *Proc. Natl. Acad. Sci. U. S. A*, **105**, 8055-8060.

McCarthy,M.I., Hirschhorn,J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.*, **17**, R156-R165.

McKernan,K.J., Peckham,H.E., Costa,G.L., McLaughlin,S.F., Fu,Y., Tsung,E.F., Clouser,C.R., Duncan,C., Ichikawa,J.K., Lee,C.C., *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527-1541.

McMahon,F.J., Akula,N., Schulze,T.G., Muglia,P., Tozzi,F., Detera-Wadleigh,S.D., Steele,C.J., Breuer,R., Strohmaier,J., Wendland,J.R., *et al.* (2010) Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. *Nat. Genet.*, **42**, 128-131.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31-46.

Moffatt,M.F., Kabesch,M., Liang,L., Dixon,A.L., Strachan,D., Heath,S., Depner,M., von,B.A., Bufe,A., Rietschel,E., *et al.* (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, **448**, 470-473.

Mohlke,K.L., Boehnke,M., Abecasis,G.R. (2008) Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Hum. Mol. Genet.*, **17**, R102-R108.

Morrow,E.M., Yoo,S.Y., Flavell,S.W., Kim,T.K., Lin,Y., Hill,R.S., Mukaddes,N.M., Balkhy,S., Gascon,G., Hashmi,A., *et al.* (2008) Identifying autism loci and genes by tracing recent shared ancestry. *Science*, **321**, 218-223.

Need,A.C., Kasperaviciute,D., Cirulli,E.T., Goldstein,D.B. (2009) A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. *Genome Biol.*, **10**, R7.

Newman,D.L., Hoffjan,S., Bourgain,C., Abney,M., Nicolae,R.I., Profits,E.T., Grow,M.A., Walker,K., Steiner,L., Parry,R., *et al.* (2004) Are common disease susceptibility alleles the same in outbred and founder populations? *Eur. J. Hum. Genet.*, **12**, 584-590.

Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30-35.

Nielsen,R. (2004) Population genetic analysis of ascertained SNP data. *Hum. Genomics*, **1**, 218-224.

Ober,C., Abney,M., McPeek,M.S. (2001) The genetic dissection of complex traits in a founder population. *Am. J. Hum. Genet.*, **69**, 1068-1079.

Ostlund,R.E., Jr. (2002) Phytosterols in human nutrition. *Annu. Rev. Nutr.*, **22**, 533-549.

Ostrer,H. (2001) A genetic profile of contemporary Jewish populations. *Nat. Rev. Genet.*, **2**, 891-898.

Parkes,M., Barrett,J.C., Prescott,N.J., Tremelling,M., Anderson,C.A., Fisher,S.A., Roberts,R.G., Nimmo,E.R., Cummings,F.R., Soars,D., *et al.* (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.*, **39**, 830-832.

Patterson,N., Hattangadi,N., Lane,B., Lohmueller,K.E., Hafler,D.A., Oksenberg,J.R., Hauser,S.L., Smith,M.W., O'Brien,S.J., Altshuler,D., *et al.* (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.*, **74**, 979-1000.

Peltonen,L., Jalanko,A., Varilo,T. (1999) Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.*, **8**, 1913-1923.

Peltonen,L., Palotie,A., Lange,K. (2000) Use of population isolates for mapping complex traits. *Nat. Rev. Genet.*, **1**, 182-190.

Plenge,R.M., Cotsapas,C., Davies,L., Price,A.L., de Bakker,P.I., Maller,J., Pe'er,I., Burtt,N.P., Blumenstiel,B., DeFelice,M., *et al.* (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.*, **39**, 1477-1482.

Pollin,T.I., Damcott,C.M., Shen,H., Ott,S.H., Shelton,J., Horenstein,R.B., Post,W., McLenithan,J.C., Bielak,L.F., Peyser,P.A., *et al.* (2008) A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science*, **322**, 1702-1705.

Price,A.L., Butler,J., Patterson,N., Capelli,C., Pascali,V.L., Scarnicci,F., Ruiz-Linares,A., Groop,L., Saetta,A.A., Korkolopoulou,P., *et al.* (2008) Discerning the ancestry of European Americans in genetic association studies. *PLoS. Genet.*, **4**, e236.

Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J., Sham,P.C. (2007) PLINK: a tool set for whole-

genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559-575.

Raychaudhuri,S., Plenge,R.M., Rossin,E.J., Ng,A.C., Purcell,S.M., Sklar,P., Scolnick,E.M., Xavier,R.J., Altshuler,D., Daly,M.J. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS. Genet.*, **5**, e1000534.

Ridker,P.M., Pare,G., Parker,A., Zee,R.Y., Danik,J.S., Buring,J.E., Kwiatkowski,D., Cook,N.R., Miletich,J.P., Chasman,D.I. (2008) Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am. J. Hum. Genet.*, **82**, 1185-1192.

Romeo,S., Pennacchio,L.A., Fu,Y., Boerwinkle,E., Tybjaerg-Hansen,A., Hobbs,H.H., Cohen,J.C. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513-516.

Romeo,S., Kozlitina,J., Xing,C., Pertsemlidis,A., Cox,D., Pennacchio,L.A., Boerwinkle,E., Cohen,J.C., Hobbs,H.H. (2008) Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.*, **40**, 1461-1465.

Rosenberg,N.A., Huang,L., Jewett,E.M., Szpiech,Z.A., Jankovic,I., Boehnke,M. (2010) Genome-wide association studies in diverse populations. *Nat. Rev. Genet.*, **11**, 356-366.

Rudkowska,I., Jones,P.J. (2008) Polymorphisms in ABCG5/G8 transporters linked to hypercholesterolemia and gallstone disease. *Nutr. Rev.*, **66**, 343-348.

Sabatti,C., Service,S.K., Hartikainen,A.L., Pouta,A., Ripatti,S., Brodsky,J., Jones,C.G., Zaitlen,N.A., Varilo,T., Kaakinen,M., *et al.* (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.*, **41**, 35-46.

Sandhu,M.S., Waterworth,D.M., Debenham,S.L., Wheeler,E., Papadakis,K., Zhao,J.H., Song,K., Yuan,X., Johnson,T., Ashford,S., *et al.* (2008) LDL-cholesterol concentrations: a genome-wide association study. *Lancet*, **371**, 483-491.

Sarin,S., Prabhu,S., O'Meara,M.M., Pe'er,I., Hobert,O. (2008) Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nat. Methods*, **5**, 865-867.

Schaid,D.J., Rowland,C.M., Tines,D.E., Jacobson,R.M., Poland,G.A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425-434.

Schork,N.J., Murray,S.S., Frazer,K.A., Topol,E.J. (2009) Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, **19**, 212-219.

Schumacher,C., Ferucci,E.D., Lanier,A.P., Slattery,M.L., Schraer,C.D., Raymer,T.W., Dillard,D., Murtaugh,M.A., Tom-Orme,L. (2008) Metabolic syndrome: prevalence among American Indian and Alaska native people living in the southwestern United States and in Alaska. *Metab Syndr. Relat Disord.*, **6**, 267-273.

Sehayek,E., Yu,H.J., von,B.K., Lutjohann,D., Stoffel,M., Duncan,E.M., Garcia-Naveda,L., Salit,J., Blundell,M.L., Friedman,J.M., Breslow,J.L. (2004) Phytosterolemia on the island of Kosrae: founder effect for a novel ABCG8 mutation results in high carrier rate and increased plasma plant sterol levels. *J. Lipid Res.*, **45**, 1608-1613.

Shmulewitz,D., Auerbach,S.B., Lehner,T., Blundell,M.L., Winick,J.D., Youngman,L.D., Skilling,V., Heath,S.C., Ott,J., Stoffel,M., *et al.* (2001) Epidemiology and factor analysis of obesity, type II diabetes, hypertension, and dyslipidemia (syndrome X) on the Island of Kosrae, Federated States of Micronesia. *Hum. Hered.*, **51**, 8-19.

Shmulewitz,D., Heath,S.C., Blundell,M.L., Han,Z., Sharma,R., Salit,J., Auerbach,S.B., Signorini,S., Breslow,J.L., Stoffel,M., Friedman,J.M. (2006) Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and dyslipidemia on the island of Kosrae, Federated States of Micronesia. *Proc. Natl. Acad. Sci. U. S. A*, **103**, 3502-3509.

Shyn,S.I., Shi,J., Kraft,J.B., Potash,J.B., Knowles,J.A., Weissman,M.M., Garriock,H.A., Yokoyama,J.S., McGrath,P.J., Peters,E.J., *et al.* (2009) Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol. Psychiatry*.

Silbernagel,G., Fauler,G., Renner,W., Landl,E.M., Hoffmann,M.M., Winkelmann,B.R., Boehm,B.O., Marz,W. (2009) The relationships of cholesterol metabolism and plasma plant sterols with the severity of coronary artery disease. *J. Lipid Res.*, **50**, 334-341.

Smith,J.G., Lowe,J.K., Kovvali,S., Maller,J.B., Salit,J., Daly,M.J., Stoffel,M., Altshuler,D.M., Friedman,J.M., Breslow,J.L., Newton-Cheh,C. (2009) Genome-wide association study of electrocardiographic conduction measures in an isolated founder population: Kosrae. *Heart Rhythm.*, **6**, 634-641.

Sudhop,T., Gottwald,B.M., von,B.K. (2002) Serum plant sterols as a potential risk factor for coronary heart disease. *Metabolism*, **51**, 1519-1521.

Sugimura,K., Taylor,K.D., Lin,Y.C., Hang,T., Wang,D., Tang,Y.M., Fischel-Ghodsian,N., Targan,S.R., Rotter,J.I., Yang,H. (2003) A novel NOD2/CARD15 haplotype conferring risk for Crohn disease in Ashkenazi Jews. *Am. J. Hum. Genet.*, **72**, 509-518.

Thomas,M.G., Weale,M.E., Jones,A.L., Richards,M., Smith,A., Redhead,N., Torroni,A., Scozzari,R., Gratrix,F., Tarekegn,A., *et al.* (2002) Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am. J. Hum. Genet.*, **70**, 1411-1420.

Tregouet,D.A., Konig,I.R., Erdmann,J., Munteanu,A., Braund,P.S., Hall,A.S., Grosshennig,A., Linsel-Nitschke,P., Perret,C., DeSuremain,M., *et al.* (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.*, **41**, 283-285.

van der Walt,J.M., Scott,W.K., Slifer,S., Gaskell,P.C., Martin,E.R., Welsh-Bohmer,K., Creason,M., Crunk,A., Fuzzell,D., McFarland,L., *et al.* (2005) Maternal lineages and Alzheimer disease risk in the Old Order Amish. *Hum. Genet.*, **118**, 115-122.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.

Verzilli,C.J., Stallard,N., Whittaker,J.C. (2006) Bayesian graphical models for genomewide association studies. *Am. J. Hum. Genet.*, **79**, 100-112.

Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J., *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60-65.

Wang,Y., O'Connell,J.R., McArdle,P.F., Wade,J.B., Dorff,S.E., Shah,S.J., Shi,X., Pan,L., Rampersaud,E., Shen,H., *et al.* (2009) From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc. Natl. Acad. Sci. U. S. A*, **106**, 226-231.

Watterson,G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256-276.

Weiss,L.A., Veenstra-Vanderweele,J., Newman,D.L., Kim,S.J., Dytch,H., McPeek,M.S., Cheng,S., Ober,C., Cook,E.H., Jr., Abney,M. (2004) Genome-wide association study identifies ITGB3 as a QTL for whole blood serotonin. *Eur. J. Hum. Genet.*, **12**, 949-954.

Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.J., Makhijani,V., Roth,G.T., *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872-876.

Willer,C.J., Sanna,S., Jackson,A.U., Scuteri,A., Bonnycastle,L.L., Clarke,R., Heath,S.C., Timpson,N.J., Najjar,S.S., Stringham,H.M., *et al.* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* , **40**, 161-169.

Won,S., Wilk,J.B., Mathias,R.A., O'Donnell,C.J., Silverman,E.K., Barnes,K., O'Connor,G.T., Weiss,S.T., Lange,C. (2009) On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS. Genet.*, **5**, e1000741.

Yeager,M., Orr,N., Hayes,R.B., Jacobs,K.B., Kraft,P., Wacholder,S., Minichiello,M.J., Fearnhead,P., Yu,K., Chatterjee,N., *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645-649.

Yu,J., Pressoir,G., Briggs,W.H., Vroh,B., I, Yamasaki,M., Doebley,J.F., McMullen,M.D., Gaut,B.S., Nielsen,D.M., Holland,J.B., *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203-208.

Zamel,N., McClean,P.A., Sandell,P.R., Siminovitch,K.A., Slutsky,A.S. (1996) Asthma on Tristan da Cunha: looking for the genetic link. The University of Toronto Genetics of Asthma Research Group. *Am. J. Respir. Crit Care Med.*, **153**, 1902-1906.

Zannis,V.I., Just,P.W., Breslow,J.L. (1981) Human apolipoprotein E isoprotein subclasses are genetically determined. *Am. J. Hum. Genet.*, **33**, 11-24.

Zhang,Z., Ersoz,E., Lai,C.Q., Todhunter,R.J., Tiwari,H.K., Gore,M.A., Bradbury,P.J., Yu,J., Arnett,D.K., Ordovas,J.M., Buckler,E.S. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, **42**, 355-360.