Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae

Eimear E. Kenny[a,b], Alexander Gusev[b], Kaitlin Riegel[c], Dieter Lütjohann[d], Jennifer K. Lowe[a,e,f], Jacqueline Salit[a], Julian B. Maller[e,f,g], Markus Stoffel[h], Mark J. Daly[e,g], David M. Altshuler[e,f,i], Jeffrey M. Friedman[a,j], Jan L. Breslow[a,1], Itsik Pe'er[b], and Ephraim Sehayek[c]

[a]Rockefeller University, New York, NY 10065; [b]Computer Science Department, Columbia University, New York, NY 10027; [c]Genomic Medicine Institute, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, OH 44195; [d]Department of Clinical Pharmacology, University of Bonn, D-53012 Bonn, Germany; [e]Medical and Population Genetics, The Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142; [f]Department of Molecular Biology and [g]Center for Human Genetics Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114; [h]Institute of Molecular Systems Biology, Swiss Federal Institute of Technology (ETH), Wolfgang-Pauli-Strasse 16, 8093 Zurich, Switzerland; [i]Department of Medicine, Harvard Medical School, Boston, MA 02115; and [j]Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815

Contributed by Jan L. Breslow, July 1, 2009 (sent for review March 18, 2009)

Pinpointing culprit causal variants along signal peaks of genome-wide association studies (GWAS) is challenging. To overcome confounding effects of multiple independent variants at such a locus and narrow the interval for causal allele capture, we developed an approach that maps local shared haplotypes harboring a putative causal variant. We demonstrate our method in an extreme isolate founder population, the pacific Island of Kosrae. We analyzed plasma plant sterol (PPS) levels, a surrogate measure of cholesterol absorption from the intestine, where previous studies have implicated 2p21 mutations in the ATP binding cassette subfamily G members 5 or 8 (ABCG5 or ABCG8) genes. We have previously reported that 11.1% of the islanders are carriers of a frameshift ABCG8 mutation increasing PPS levels in carriers by 50%. GWAS adjusted for this mutation revealed genomewide significant signals along 11 Mb around it. To fine-map this signal, we detected pairwise identity-by-descent haplotypes using our tool GERMLINE and implemented a clustering algorithm to identify haplotypes shared across multiple samples with their unique shared boundaries. A single 526-kb haplotype mapped strongly to PPS levels, dramatically refining the mapped interval. This haplotype spans the ABCG5/ABCG8 genes, is carried by 1.8% of the islanders, and results in a striking 100% increase of PPS in carriers. Resequencing of ABCG5 in these carriers found a D450H missense mutation along the associated haplotype. These findings exemplify the power of haplotype analysis for mapping mutations in isolated populations and specifically for dissecting effects of multiple variants of the same locus.

genetics | genomewide association study | haplotype mapping | plasma plant sterols

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 270 genomic regions that modify the risk for over 70 complex disease and health-related phenotypes (1–3). However, identifying the culprit causal variant underlying these local peaks in association signals remains a challenging task with only a handful of causal variants identified so far (4–6). 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. Recent examples of multiple alleles include the 8q24 region for prostate cancer (7–9), 6p21 region for HIV-1 viral setpoint (10), the IL23R, 6p25, and 17q21 regions for Crohn's disease (11, 12), the PNPLA3 region and nonalcoholic fatty liver disease (13), the IRF5, STAT4, and TNFAIP3 regions for systemic lupus erythematosis (14–16), and the 6q23 region for rheumatoid arthritis (17).

We introduce a method for exposing a causal variant in a region of 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 (18, 19). 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 (20), which performs pairwise identity-by-descent matching to identify long shared genomic segments in a population. Here we combine GERMLINE with a clustering algorithm that 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 (21–25) and a reduced multiple test burden relative to other similar methods (26–30).

We tested our method in an extreme isolate founder population where we had previously shown abundant long haplotype segments shared between individuals (20, 31). 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 (32–36). We focused on plasma plant sterol (PPS) levels, which were ascertained as part of a broader study to examine genetic causes

Author contributions: E.E.K., A.G., J.L.B., I.P., and E.S. designed research; E.E.K., A.G., K.R., D.L., and E.S. performed research; E.E.K. and A.G. contributed new reagents/analytic tools; E.E.K., A.G., K.R., D.L., J.K.L., J.S., J.B.M., M.S., M.J.D., D.M.A., J.M.F., J.L.B., I.P., and E.S. analyzed data; and E.E.K., J.L.B., I.P., and E.S. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: breslow@rockefeller.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907336106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0907336106

of metabolic syndrome on the island (36). Plant sterols are a dietary source of neutral sterols that are structurally similar to cholesterol (37, 38). Studies of the rare disorder phytosterolemia, which is characterized by extremely high PPS levels and severe premature atherosclerosis, have implicated mutations in the ATP binding cassette subfamily G members 5 or 8 (ABCG5 or ABCG8) genes (39–41). 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 (42). As such, PPS levels 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 (43, 44). Further, it is unclear whether moderately elevated PPS levels may themselves affect cardiovascular risk (45–47). 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 nonfunctional 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 (48).

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 that 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 an 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 of PPS levels in 1,423 related individuals on the Pacific Island of Kosrae with genotypes from the Affymetrix 500-k 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 1.4×10^{-10} (Fig. 1*A*). The strongest signal confirmed the association of the ABCG8 nonsense mutation ($P < 5 \times 10^{-39}$) to PPS levels and validated our previous finding in a subset of this cohort (48). 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 (Fig. 1*B*). To determine whether this extended signal was entirely accounted for by the ABCG8 nonsense mutation, we reanalyzed the association of the PPS level's phenotype conditioned on the ABCG8 nonsense mutation genotype (Fig. 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 rs12185607 ($P < 3 \times 10^{-31}$) ≈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 Fig. 1*D*, and specifically at the chromosome

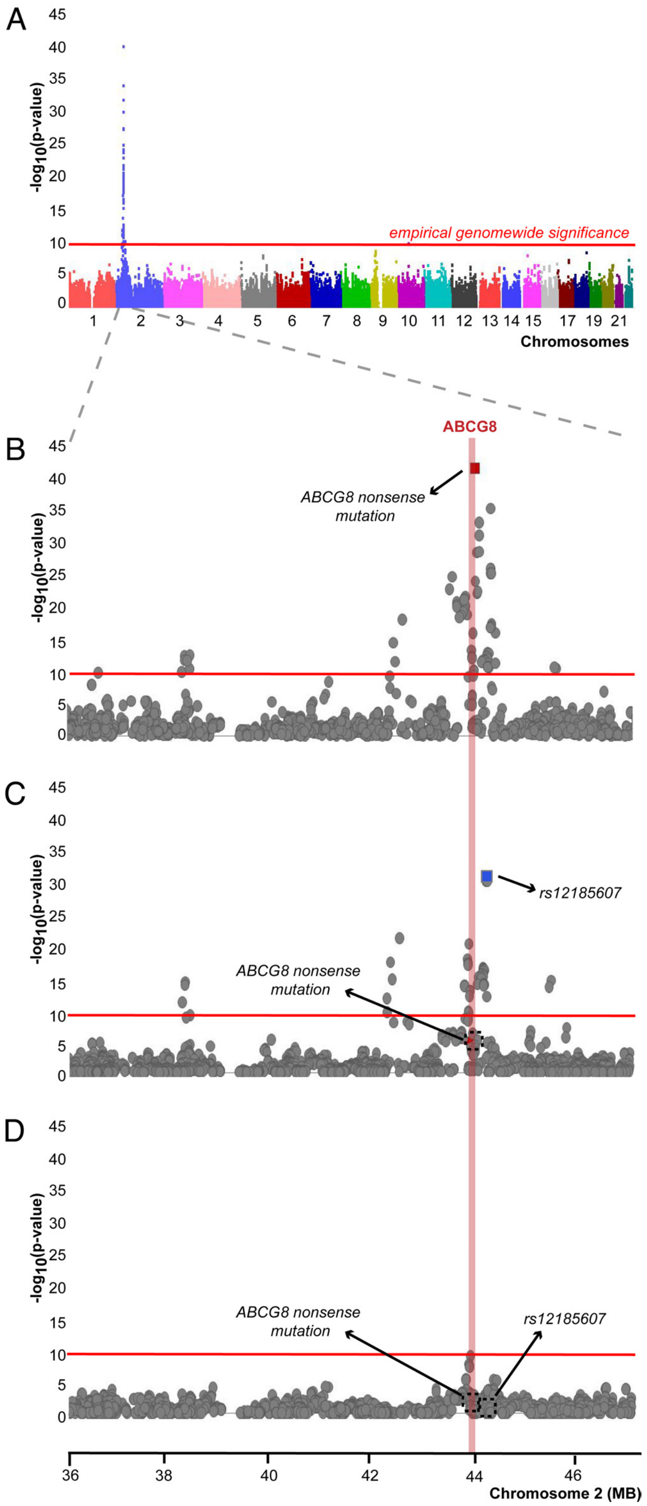

Fig. 1. Stepwise conditional analysis of PPS levels on the Island of Kosrae. (*A*) Unconditioned genome-wide association of PPS levels. Association analysis was performed as described in the methods section. The red line marks the threshold of empirical genome-wide significance. (*B*) Unconditioned association of PPS in an 11-Mb 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 levels 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 levels in the same chromosome 2 interval after conditioning for both the ABCG8 nonsense mutation and rs12185607 genotypes.

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 rs12185607/G minor allele, respectively, we expected to find five to six Islanders carrying both alleles. We observed only two individuals ($P = 0.076$ assuming independence) who were carriers of both minor alleles (Table S1). Therefore, our findings suggested the presence of two independent variants effecting PPS levels at the chromosome 2p21 locus.

Systematic Identification of a 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 (Fig. 2). The 11-Mb signal region on chromosome 2p21 (chr2: 36–47 Mb) around the ABCG5/ABCG8 locus was excised, and common haplotypes were identified by IBD matching using GERMLINE (20). 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 (Fig. S1). We found a single cluster containing 44 individuals that shared a 1.3-Mb haplotype strongly associated with PPS levels ($P < 4 \times 10^{-65}$). 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 subregion (chr2: 43.8–44.3 Mb) that was unique to this group of 52 individuals (Fig. 3*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 ($P = 5.2 \times 10^{-4}$).

Clustering of the 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 (Fig. S2). 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 (49) 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 nongenetic factor might cause increased PPS levels in the whole village and, hence, false positive association signals among this subpopulation. To control for any confounding local environment in the village, we coded PPS haplotype carrier status as an allele (1 = carrier, 0 = noncarrier) and performed association analysis with PPS levels. Eighty-nine 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 subgroup with a permuted $P < 1.6 \times 10^{-4}$.

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 (48). 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 Fig. 3*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 noncarriers (in both cases $P < 0.0001$).

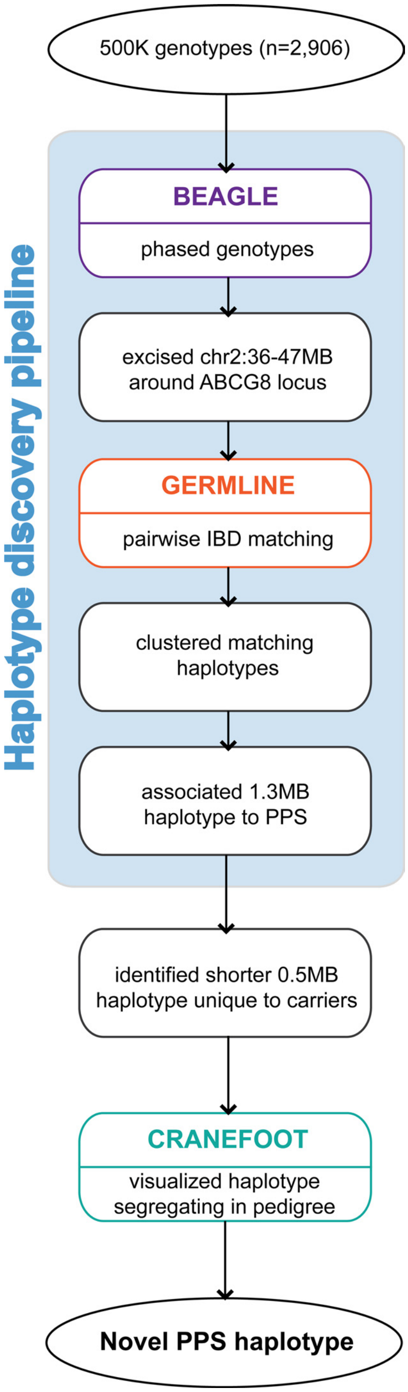
500K genotypes (n=2,906)
Haplotype discovery pipeline
BEAGLE
phased genotypes
excised chr2:36-47MB around ABCG8 locus
GERMLINE
pairwise IBD matching
clustered matching haplotypes
associated 1.3MB haplotype to PPS
identified shorter 0.5MB haplotype unique to carriers
CRANEFOOT
visualized haplotype segregating in pedigree
Novel PPS haplotype

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

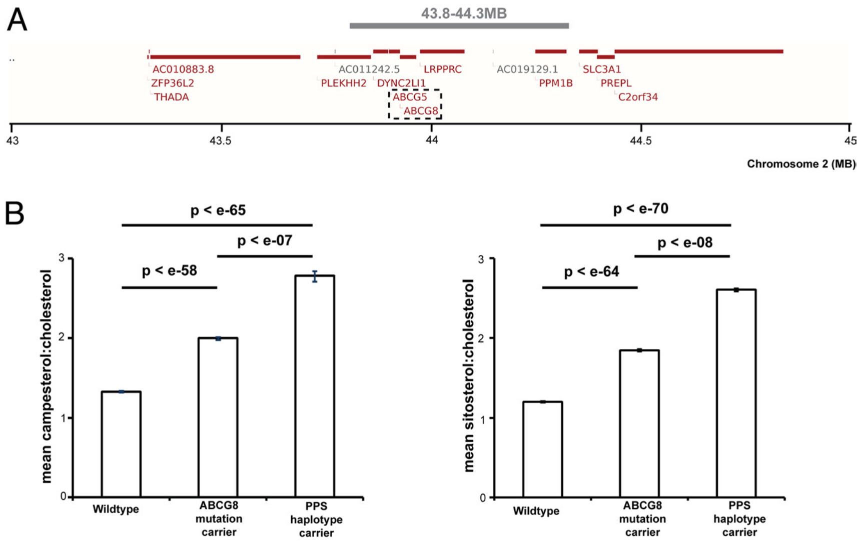

Fig. 3. Location of PPS haplotype and effect on PPS levels. (*A*) Annotated genes on the background of the 526-kb haplotype. The PPS haplotype (indicated as a gray bar) extends over a 526-kb region on chromosome 2p21 that includes the ABCG8 and ABCG5 candidate genes (indicated by a black dashed square) and also fully or partially five 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.

Also shown in Fig. 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 noncarriers (in both cases $P < 0.0001$).

The effects of the ABCG8 nonsense mutation and the 526-kb haplotype on plasma lipid levels were next examined. As shown in Table S2, 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 S2, carriers of the ABCG8 nonsense mutation had a 12.5% decrease in plasma lathosterol (1.33 ± 0.55), and carriers of the 526-kb haplotype had a 45% decrease in plasma lathosterol (0.84 ± 0.38) compared to noncarriers (1.52 ± 0.67) (in both cases $P < 0.0001$).

Sequencing Reveals a Putative ABCG5 Causal Variant. The 526-kb haplotype region on chromosome 2p21 encompasses both ABCG5 and ABCG8 and six other annotated genes (Fig. 3*A*). ABCG5 and ABCG8 are obvious candidate genes to explain the haplotype effect. We first sequenced all 13 exons and splicing donor/acceptor flanking sequences of these genes in four pairs of 526-kb haplotype carriers and noncarrier-related controls. As shown in Fig. S3, this effort revealed an 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 (Fig. S3). 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 noncarrier controls, 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 implementation of an 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. Resequencing 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 ≈4-kb haplotype and was not detectable by this method. The IBD approach to search for a causal allele affecting PPS levels 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 (50) and even in apparently unrelated populations (51). This has motivated several groups to develop IBD methods as a means of mapping disease-related alleles (20, 29, 30).

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% (Fig. 3*B*). It has been shown that ABCG5 functions by heterodimerization with ABCG8 (42). Therefore, we hypothesize that the greater effect of the missense mutation is due to one of the following: (*i*) Imbalanced expression of a nonfunctional ABCG5-mutant allele; (*ii*) a gain-of-function dominant-negative ABCG5 mutation; or (*iii*) linkage disequilibrium of a nonfunctional 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 (52, 53). 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 effects 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 fairly rare in the general population. Homozygotes or compound heterozygotes have the disorder phytosterolemia at an estimated frequency of 1:1,000,000, which implies a carrier rate of 1:500. 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 an approach to identify and dissect the effect of large haplotypes under the signal of a single locus detected by GWAS. This approach allowed identification of a 526-kb haplotype that modifies PPS and dissected its effect from an ABCG8 mutant that was known to segregate on the island of Kosrae.

Methods

Sample Collection, Phenotyping, and Genotyping. A full description of the screening and genotyping of the Kosraen cohort was described elsewhere (36). 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. Informed consent was obtained from each individual screened and so were self-reported family histories and lifestyle information. Fasting blood was collected and centrifuged. Plasma and buffy coats were frozen and shipped to Rockefeller University, NY, for serological assays and DNA extraction. The levels of plasma sterols (campesterol, sitosterol, lathosterol, and total cholesterol) were measured by gas liquid chromatography, and plasma levels of campesterol, sitosterol, and lathosterol were expressed as the ratio of sterol to total plasma cholesterol levels (48). Plasma LDL-cholesterol, HDL-cholesterol, ApoA1, ApoB, and triglycerides were measured as described previously (36). IRB approval was obtained from all participating institutions. Study participants (2,906) were successfully genotyped on the Affymetrix 500-K platform; data were generated at Affymetrix. Genotypes were called with the BRLMM algorithm, and a minimum call rate of 95% was achieved. SNPs (446,802) passed quality control filters, and an additional 109,443 SNP's that were monomorphic or very rare [minor allele frequency (MAF) < 0.01] were also excluded. The final data set yielded 337,359 SNPs with MAF > 0.01 for the analysis. The ABCG8 nonsense mutation had been previously genotyped in a subset of 1,090 islanders by using either *PfoI* RFLP or fluorescently labeled allele-specific primers (48). This effort revealed a 13.8% carrier rate for this mutation. We repeated the genotyping of these 1,090 individuals and an additional 1,214 Islanders using a TaqMan assay consisting of specific probes that targeted the ABCG8 nonsense mutation and the normal allele. This revealed an ABCG8 nonsense mutation carrier rate of 11.1% in the entire cohort.

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 nongenotyped individuals to fill out the pedigree and careful cross-referencing of patient records. Genome-wide SNP data were subjected to identity-by-state analyses and identity-by-descent estimation performed in PLINK (29) to correct and validate the pedigree structure. Full details of the pedigree construction are described elsewhere (36). The CraneFoot pedigree drawing software was used to visualize subsets of the extended pedigree (54).

Association Analyses. Due to the difficulties of analyzing a large cohort that spans multiple generations, we deconstructed the pedigree to groups of ''sibships'' of 2+ sibs and ''unrelateds'' who were less than first cousins. We performed genome-wide association for the PPS traits for 611 sibships (1,328 individuals) and 101 unrelateds via the PLINK/QFAM-Total algorithm after correcting for any age or gender biases (29). PLINK/QFAM-Total performs a linear regression for phenotype and genotype and uses a special permutation procedure to correct for family structure. However, the top nominal scores outputted by PLINK/QFAM-Total for the PPS trait were orders of magnitude beyond what could reasonably be permuted to derive empirical *P* values. Instead, we assessed an empirical genome-wide significance threshold for the nominal scores (55). The empirical genome-wide significance threshold was calculated by the following steps: (*i*) Phenotypes were randomly permuted between individuals 1,000 times, accounting for sibship structure, to generate 1,000 permuted genotype/phenotype data sets; (*ii*) genome-wide association analysis was performed for each permuted data set, and the minimum *P* value recorded; and (*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. Genome-wide association analysis that conditioned on the ABCG8 nonsense mutation or the ABCG8 nonsense mutation plus rs12185607 involved a preliminary regression step to adjust the phenotype by genotype status.

Identifying Haplotypes. Genotypes for the entire Kosraean population were phased using the BEAGLE algorithm in one run according to the recommended parameters: One sample, 10 iterations (56). An 11-Mb region underlying the genome scan signal peak (chr2: 36–47 Mb) was excised for each individual and analyzed for pairwise IBD matching by the GERMLINE algorithm (20). 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 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, and 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 (57). 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 subgraph 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 that was shared exclusively by affected individuals.

ABCG5/ABCG8 Sequencing. 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.

ACKNOWLEDGMENTS. We thank the Government and Department of Health of Kosrae for their partnership and especially the people of Kosrae for making this study possible.

1. Kruglyak L (2008) The road to genome-wide association studies. *Nat Rev Genet* 9:314–318.
2. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 118:1590–1605.
3. Hindorff LA, Junkins HA, Mehta JP, Manolio TA (2008) A catalog of published genome-wide association studies. Available at: www.genome.org/26525384. Accessed on 12.03.2008.
4. Burnett JR, AJ Hooper (2008) Common and rare gene variants affecting plasma LDL cholesterol. *Clin Biochem Rev* 29:11–26.
5. Moffatt MF, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448:470–473.
6. Burkhardt R, 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.
7. Haiman CA, et al. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* 39:638–644.
8. Yeager M, et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 39:645–649.
9. Gudmundsson J, et al. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* 39:631–637.
10. Fellay J, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science* 317:944–947.
11. Duerr RH, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314:1461–1463.
12. Barrett JC, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40:955–962.
13. Romeo S, et al. (2008) Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 40:1461–1465.
14. Graham RR, et al. (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci USA* 104:6758–6763.
15. Abelson AK, et al. (2008) STAT4 Associates with SLE through two independent effects that correlate with gene expression and act additively with IRF5 to increase risk. *Ann Rheum Dis* Dec 9. [Epub ahead of print].
16. Graham RR, et al. (2008) Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat Genet* 40:1059–1061.
17. Plenge RM, et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet* 39:1477–1482.
18. Gabriel SB, et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.
19. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7:256–276.
20. Gusev A, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19:318–326.
21. de Bakker PI, et al. (2005) Efficiency and power in genetic association studies. *Nat Genet* 37:1217–1223.
22. Durrant C, et al. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet* 75:35–43.
23. Horvath S, et al. (2004) Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics. *Genet Epidemiol* 26:61–69.
24. Schaid DJ, et al. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70:425–434.
25. Verzilli CJ, Stallard N, Whittaker JC (2006) Bayesian graphical models for genomewide association studies. *Am J Hum Genet* 79:100–112.
26. Lin S, Chakravarti A, Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* 36:1181–1188.
27. Tregouet DA, 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.
28. Laramie JM, et al. (2007) HaploBuild: An algorithm to construct non-contiguous associated haplotypes in family based genetic studies. *Bioinformatics* 23:2190–2192.
29. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
30. Albrechtsen A, et al. (2008) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* 33:266–274.
31. Bonnen PE, et al. (2006) Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet* 38:214–217.
32. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.
33. Friedlaender JS, et al. (2008) The genetic structure of Pacific Islanders. *PLoS Genet* 4:e19.
34. Kayser M, et al. (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet* 82:194–198.
35. Shmulewitz D, et al. (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 USA* 103:3502–3509.
36. Lowe JK, et al. (2009) Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS Genet* 5:e1000365.
37. Ostlund RE, Jr (2002) Phytosterols in human nutrition. *Annu Rev Nutr* 22:533–549.
38. Heinemann T, Axtmann G, von Bergmann K (1993) Comparison of intestinal absorption of cholesterol with different plant sterols in man. *Eur J Clin Invest* 23:827–831.
39. Berge KE, et al. (2000) Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science* 290:1771–1775.
40. Lee MH, et al. (2001) Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. *Nat Genet* 27:79–83.
41. Hubacek JA, et al. (2001) Mutations in ATP-cassette binding proteins G5 (ABCG5) and G8 (ABCG8) causing sitosterolemia. *Hum Mutat* 18:359–360.
42. Graf GA, et al. (2003) ABCG5 and ABCG8 are obligate heterodimers for protein trafficking and biliary cholesterol excretion. *J Biol Chem* 278:48275–48282.
43. Rudkowska I, Jones PJ (2008) Polymorphisms in ABCG5/G8 transporters linked to hypercholesterolemia and gallstone disease. *Nutr Rev* 66:343–348.
44. Sudhop T, Gottwald BM, von Bergmann K (2002) Serum plant sterols as a potential risk factor for coronary heart disease. *Metabolism* 51:1519–1521.
45. Silbernagel G, et al. (2009) The relationships of cholesterol metabolism and plasma plant sterols with the severity of coronary artery disease. *J Lipid Res* 50:334–341.
46. Fassbender K, et al. (2008) Moderately elevated plant sterol levels are associated with reduced cardiovascular risk—the LASA study. *Atherosclerosis* 196:283–288.
47. Chan YM, et al. (2006) Plasma concentrations of plant sterols: Physiology and relationship with coronary heart disease. *Nutr Rev* 64:385–402.
48. Sehayek E, et al. (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.
49. Patterson N, et al. (2004) Methods for high-density mapping of admixture mapping of disease genes. *Am J Hum Genet* 74:979–1000.
50. Kong A, et al. (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40:1068–1075.
51. Frazer KA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
52. Kathiresan S, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 41:56–65.
53. Aulchenko YS, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. 41:47–55.
54. Mäkinen VP, et al. (2005) High-throughput pedigree drawing. *Eur J Hum Genet* 13:987–989.
55. Abney M, Ober C, McPeek MS (2002) Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: Fasting serum-insulin level in the Hutterites. *Am J Hum Genet* 70:920–934.
56. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81:1084–1097.
57. Reingold O (2008) Undirected connectivity in log-space. *J ACM* 55(4) Article 17.