The doctoral Qualifying Exam and Master’s Comprehensive Exam begins with an oral presentation by the student of a research article from the primary literature. The paper serves as a starting point for an oral examination of the student’s understanding of the paper and knowledge of the field.

Students will be required to demonstrate expertise in their chosen areas of concentration, which may include deeper knowledge than covered in the required coursework. In addition to mastery in their specific area, students will be required to demonstrate broad knowledge of the field of Human Genetics as a whole. Students will be responsible for showcasing their ability to utilize knowledge and skills obtained through coursework and research experiences to inform their interpretation, synthesis, and evaluation of the contemporary peer-reviewed literature.

The following list of knowledge requirements that students are expected to master may be used to assist in preparing for the qualifying exam. Likewise, course objectives for the core Human Genetics curriculum are provided below and may serve as an outline of topics to assist in students’ preparations. Material covered in additional courses relevant to an individual student’s academic experience and trajectory may also be germane. Lastly, example qualifying exam questions pertaining to specific example papers are provided. Please be aware that some questions raised during the oral examination may not have clear or objective “right” answers, but may instead reflect current uncertainties in the field of human genetics. In these cases, students may be expected to provide thoughtful discussion regarding prevailing hypotheses or themes, demonstrate knowledge of ambiguities in the scientific evidence or limitations of our current knowledge, and propose new experiments that could yield insight into the question at hand.

Knowledge requirements:

- **DNA, Genes, Chromosomes, and Genomes**
  - definition, structure, function, and organization of genes, chromosomes, and genomes
  - mitosis, meiosis, and recombination
  - central dogma and mechanisms of DNA replication, transcription, and translation
  - gene expression and regulation during development, normal, and abnormal cellular functions
  - genetic variation at the nucleotide, gene, chromosome, and genome levels

- **Genetic Control of Phenotypes**
  - modes of inheritance (autosomal/sex-linked, dominant/additive/recessive, mitochondrial, etc.) and corresponding patterns of phenotypic variation in human pedigrees
- basic principles and examples of inborn errors of metabolism and classic genetic syndromes
- the multifactorial nature of complex traits and the principles of multifactorial inheritance
- molecular techniques for detecting genetic mutations and polymorphisms
- use of model organisms to study human genetic diseases
- principals of cancer genetics, the two-hit hypothesis, and somatic mutations

- Disease Gene Discovery and Gene Mapping
  - genetic epidemiology: modeling genotype-phenotype relationships, and identifying the genetic contributors of a phenotype
  - how multifactorial inheritance impacts disease gene mapping in humans
  - rationale, advantages, and limitations of linkage analysis study designs
  - rationale, advantages, and limitations of genetic association study designs
  - interpretation of classical statistical tests commonly used in genetic studies
  - approaches for studying the role of copy number variation (CNV) in complex disease
  - applications of next-generation sequencing (NGS) in the studying human disease
  - the role of rare variants in common complex disease and how rare variants are tested for association
  - approaches for finding functional variations
  - examples of complex disease genes that have been successfully identified
  - definition of multiple testing and common methods for correction
  - use of animal models to define the pathophysiology of disease

- Population Genetics
  - impact of fundamental principles of population genetics including Hardy-Weinberg Equilibrium, its assumptions, and the effects of violations of these assumptions
  - how evolutionary principles (like bottlenecks, founder effects, and population isolates) can affect or be exploited for human genetic studies
  - how allele frequencies differ in populations and the impact on genetic studies
  - linkage disequilibrium, how it is described and measured, how it relates to haplotypes, and how it impacts association studies

- Genetics in Society
  - how novel scientific discoveries are evaluated in a clinical context and applied appropriately to the care of patients, the function of genetic counseling
  - issues in genetic testing
  - how legal and ethical issues affect research, including informed consent

- Molecular Techniques
  - definition, implementation, advantages, and limitations of PCR, genotyping methods, gene expression methods, sequencing methods

Course objectives:

**HUGEN 2022: Human Population Genetics**

- apply the Law of Hardy-Weinberg Equilibrium and its assumptions to calculate allele and genotype frequencies
- predict the consequences of genetic inheritance and recombination in populations including the concepts of linkage and linkage disequilibrium
- interpret the qualitative effects of violations of Hardy-Weinberg Equilibrium and solve simple quantitative problems demonstrating these effects
- express the fundamental goals and principles of genetic epidemiology by modeling genotype-phenotype relationships, quantitative traits, and heritability
- interpret results from gene discovery methods such as linkage analysis and large-scale genetic association studies, and critically evaluate the strengths, limitations, and appropriate applications of these methods.

HUGEN 2031: Chromosomes and Human Disease

- compare and contrast the strengths and limitations of the various cytogenetic assays
- describe chromosomal alterations, their causes, and their clinical implications
- describe the role of chromosomal alterations in development of cancer and their use in cancer diagnostics
- apply principles of effective written and oral communication to cytogenetics topics
- critique published cytogenetics literature
- write a research proposal related to cytogenetics or chromosomes
- interpret cytogenetic nomenclature
- describe a variety of cytogenetic concepts and methodologies
  - the cellular basis of chromosome segregation, chromosome structure, meiosis and mitosis, numerical and structural chromosome abnormalities in clinical disorders, including chromosomal syndromes, chromosome breakage syndromes, cancer, sex determination and sex chromosome abnormalities, genomic imprinting, trinucleotide repeat disorders, classical and molecular cytogenetic methods including copy number arrays and nomenclature, the role of cytogenetic techniques in diagnosis of disorders, the history of cytogenetic methods, and ethical issues related to cytogenetic testing and results.

HUGEN 2034: Biochemical and Molecular Genetics of Complex Disease

- Describe the epidemiology and biochemical and molecular bases of selected common diseases
- Describe the underlying genetic architecture and environmental risk factors that influence genetic susceptibility to a variety of common, complex diseases ranging from cardiovascular disease to neurological disorders
- Describe and discuss the public health impact of a variety of common, complex diseases, including potential pharmacogenomics applications.

HUGEN 2040: Molecular Genetics of Human Inherited Disease

- Explain the structural components and functions of genes
- Evaluate the utility of molecular testing methods to diagnose inherited diseases, and calculate key test measures (specificity, sensitivity, etc.)
- Differentiate Mendelian disorders based on clinical manifestations
- Distinguish inherited diseases based on the molecular mechanisms
- Compare different treatment approaches to inherited diseases
- Recognize unique human populations, which harbor alleles for inherited diseases
- Compile clinical, molecular, diagnostic and treatment information on inherited diseases
Example qualifying exam (Ph.D.) and comprehensive exam (M.S.) questions pertaining to Guissart et al. (2018) American Journal of Human Genetics.

1. Why do many human genes have two paralogs in zebrafish? Why are both often retained? How does the presence of two paralogs affect the design of zebrafish models of human disease?

2. In Figure 4C the authors present an immunoblot for RORA in skin fibroblasts of patients and a control. In the samples from patients 2 and 6, is the detected protein a product of the WT allele, mutant allele, or both? Explain your reasoning.

3. The authors speculate that the Gly92Ala and Lys94Arg mutations might prevent the access of WT RORA to its target sites. Is this speculation consistent with the findings in figure 4 and the authors’ conclusion that these mutations act through a toxic gain of function mechanism?

4. Discuss the mechanisms by which one gene can produce different transcripts. Which of these mechanisms are illustrated by the human RORA gene and which by the zebrafish roraa gene?

5. What method would you use to test if the number or strength of roraa binding sites are altered by certain missense mutations? How would you test if binding at those sites result in an increase or a decrease of transcription?

6. Would you expect individual 11, who has the largest deletion, to show any different symptoms in addition to those shared with the others? Why or why not? Are there any particular traits or biomarkers you might choose to investigate?

7. What is the difference between a gain-of-function mutation and a dominant-negative one? Give an example of a human genetic disorder that is caused by a dominant negative mutation.

8. Explain array CGH, whole exome sequencing, and Sanger sequencing. What information do you get from each technique and to what type of research question would you apply each technique?

9. Cultured fibroblasts from patient #6 (p.Arg340Profs817) show no change in protein level via Western. Authors suggest that the transcript, which is a deletion with frame shift and early stop, is not degraded
by NMD. What is NMD? Do you think their explanation is likely, and why? How could they prove this? Can you suggest other reasons why the protein level might not be lower in this sample?

10. The authors conclude that mutations in the DNA binding domain (DBD) generate gain-of-function and are more deleterious than mutations in the ligand binding domain that generate loss of function (haploinsufficiency). Do you agree? Why or why not? What further experiments could you perform to prove their hypothesis and distinguish whether location of the mutation is important?

11. The specific *RORA* variants reported in this paper had not been discovered in previous GWAS studies of ASD or IQ. What is a likely reason for this?

12. The authors concluded that ectopic expression of each of the four WT *RORA* mRNAs did not lead to cerebellar phenotypes, but Figure S9 shows increased cerebellar area in zebrafish embryos injected with 200 pg of RORA1 mRNA. The p-value is 0.03. Does this result contradict their conclusion? Consider both the statistical significance and the scientific significance of this result.

13. What is the “multiple comparisons problem” and what strategies can be used to accommodate or resolve this problem? Does the multiple comparisons problem impact this paper? Why or why not?

14. Consider the deletion segregating in Family 10 that shows a dominant mode of inheritance with respect to the clinical phenotype. What do you speculate about the variant’s impact on fitness? Do you expect that it is adaptive, maladaptive, or neutral? How then do you expect the frequency of the variant will fluctuate across future generations due to selection? What if instead the variant showed a recessive mode of inheritance – how would selection differ?
Dual Molecular Effects of Dominant RORA Mutations Cause Two Variants of Syndromic Intellectual Disability with Either Autism or Cerebellar Ataxia

Claire Guissart,1,3,7 Xenia Latypova,2,3,4,7 Paul Rollier,5,3,7 Tahir N. Khan,3,7 Hannah Stamberger,6,7,8 Kirsty McWalter,9 Megan T. Cho,9 Susanne Kjaergaard,10 Sarah Weckhuysen,6,7,8 Gaetan Lesca,11,12 Thomas Besnard,5,4 Katrin Öunap,13 Lynn Schema,14 Andreas G. Chiocchetti,15 Marie McDonald,16 Julitta de Bellescize,17 Marie Vincent,5,4 Hilde Van Esch,18 Shannon Sattler,19 Irman Forghani,20 Isabelle Thiffault,21,22,23 Christine M. Freitag,15 Deborah Sara Barbouth,20 Maxime Cadieux-Dion,21 Rebecca Willaert,9 Maria J. Guillen Sacoto,9 Nicole P. Safina,23,24,25 Christèle Dubourg,26 Lauren Grote,23,24,25 Wilfrid Carré,26 Carol Saunders,21,22,23 Sander Paujasalu,13

(Raw_text continues on next page)

Introduction

Nuclear receptors appeared in the metazoan lineage as an adaptation to multicellular organization requiring distant cellular signaling through non-peptidic growth and differentiation factors.1 The nuclear receptor superfamily is a group of transcription factors regulated by small hydrophobic hormones, such as retinoic acid, thyroid hormone, and steroids.2 Mutations in nuclear receptors cause a diverse range of disorders, including central nervous system (CNS) pathologies, cancer, and metabolic disorders. For example, haploinsufficiency of RORB (MIM: 601972), encoding the nuclear receptor RORB, results in behavioral and cognitive impairment and epilepsy,3 while biallelic mutations in RORC (MIM: 602943), encoding RORY, result in immunodeficiency (MIM: 6166224). Nuclear Receptor Subfamily 0, Group B, Member 1 (NR0B1 [MIM: 300473]) is an orphan member of the nuclear receptor superfamily and has been implicated in sex reversal (MIM: 3000185) and congenital adrenal hypoplasia (MIM: 3002006). RORα (RORA) is most closely related to Retinoic Acid Receptor (RAR) yet functions differently; RAR acts as a ligand responsive heterodimer with retinoid X receptor (RXR). However, RORα isoforms 1 and 2 constitutively activate transcription and bind DNA as monomers at responsive elements which consist of 6-bp

1EA7402 Institut Universitaire de Recherche Clinique, and Laboratoire de Génétique Moléculaire, CHU and Université de Montpellier, 34093 Montpellier, France; 2Service de Génétique Médicale, CHU Nantes, 9 quai Moncousu, 44093 Nantes Cedex 1, France; 3Center for Human Disease Modeling, Duke University Medical Center, Durham, NC 27710, USA; 4Institut du thorax, INSERM, CNRS, UNIV Nantes, 44007 Nantes, France; 5Service de Génétique Clinique, Centre Référence “Déficience Intelectuelles de causes rares” (CRDi), Centre de référence anomalies du développement CLAD-Ouest, CHU Rennes, 35203 Rennes, France; 6Division of Neurology, University Hospital Antwerp (UZA), 2610 Antwerp, Belgium; 7Neurogenetics Group, Center for Molecular Neurology, VIB, 2650 Antwerp, Belgium; 8Laboratory of Neurogenetics, Institute Born-Bunge, University of Antwerp, 2650 Antwerp, Belgium; 9Genelix, 207 Perry Parkway, Gaithersburg, MD 20877, USA; 10Chromosome Laboratory, Department of Clinical Genetics, Copenhagen University Hospital, Righospitalet, 2100 Copenhagen, Denmark; 11Service de génétique, Groupement Hospitalier Est, Hospices Civils de Lyon, Lyon, France; 12INSERM U1028, CNRS UMR5292, Centre de Recherche en Neurosciences de Lyon, Université Claude Bernard Lyon 1, Lyon, France; 13Department of Clinical Genetics, United Laboratories, Tartu University Hospital and Institute of Clinical Medicine, University of Tartu, 2 L.Puusepa street, Tartu 51014, Estonia; 14University of Minnesota-Fairview, Minneapolis, MN 55454, USA; 15Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, JW Goethe
contiguous gene syndrome, with the smallest deletion have been reported in affected individuals as part of a humans, microdeletions overlapping 15q22.2 microdeletion share epileptic seizures, mild intellectual disability (ID), and dysmorphic features, with variable ataxia. Here, we report 16 affected individuals from 13 syndromic ID-affected families with intergenic or intragenic deletions, truncating mutations, or missense changes in RORA. We modeled these genetic findings in zebrafish larvae through endogenous roraa ablation or heterologous expression of RORA, followed by relevant cerebellar phenotyping. In zebrafish models, we recapitulated the neuroanatomical features of affected humans as well as the staggerer mouse, and we show that loss of roraa leads to reduction of both the Purkinje and granule compartments of the cerebellum. Further, our in vivo data indicate that missense variants in the DNA binding domain confer a dominant toxic effect, while a missense change in the ligand binding domain results in a loss-of-function effect. Together, our data highlight how different mutation effects at the RORA locus can produce overlapping but distinct phenotypic outcomes.

Subjects and Methods

Genetic Studies and Ethics Statement

Human genetic studies conducted in research laboratories were approved by local ethics committees from participating centers (Antwerp, Belgium; Lyon, France; Frankfurt, Germany; Copenhagen, Denmark; Montpellier, France; Tartu, Estonia; Minneapolis and Kansas City, USA). Written informed consent was obtained from all study participants. All 16 affected individuals underwent extensive clinical examination by at least one expert clinical geneticist. Routine genetic testing was performed whenever clinically relevant, including copy-number variation (CNV) analysis by high-resolution array-based comparative genomic hybridization (aCGH) using (1) 180k CytoSure ISCA v2 array (Oxford Gene Technology; individuals 10A–D), (2) 180k SurePrint G3 CGH microarray as described (Agilent; individual 11), or (3) 400k SurePrint G3 CGH microarray, as described (Agilent, individuals 12 and 13). Affected individuals with a negative aCGH result underwent whole-exome sequencing (WES) on an Illumina HiSeq platform according to the following paradigms: (1) trio-based clinical diagnostic WES (individuals 1, 2, 6, 7, and 9), (2) trio-based WES in a research laboratory (individuals 3 and 4), or (3) WES of an affected individual followed by single site testing in parental DNA samples (individuals 5 and 8; see Table S1 for further details). Point mutations were confirmed by Sanger sequencing of DNA sample from all available family members, when possible.

Establishment and Culture of Primary Fibroblasts

We conducted biochemical studies in primary cultures of skin fibroblasts from individuals 2, 3, and 6 and from two control individuals (WT1 and WT2). Fibroblasts were grown in RPMI 1640 medium, containing 5% fetal calf serum (FCS, ThermoFisher), 1% Ultroser G
RT-PCR Studies in Primary Fibroblasts
Total RNA was extracted using the RNeasy Plus Mini kit (QIAGEN) and reverse transcribed with the Moloney Murine Leukemia Virus Reverse Transcriptase (ThermoFisher). PCR amplification was performed with primers located in exons 3 and 4 (Table S2) and the Promega MasterMix (Promega). PCR products were separated on a 1.5% agarose gel and fragments were cut from the gel. After extraction from gel slices, the PCR products were purified with ExoSAP-IT (ThermoFisher) and sequenced by standard cycle-sequencing reactions with Big Dye terminators (ThermoFisher) with the PCR forward and reverse primers in an ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems). Mutation detection analysis was performed using 4Peaks.

RORA Immunoblotting
Total protein lysate was extracted from primary fibroblasts using 1 × Laemmli buffer. Proteins were separated on 10% SDS-PAGE gels and transferred to PVDF membrane (Westran Clear Signal Whatman; Dominique Dutcher, Brumath). The membranes were incubated overnight with 1:200 diluted anti-RORα (sc-6062, Santa Cruz biotechnology) primary antibody in 5% skim milk. Protein levels of the housekeeping protein GAPDH were assayed for internal control of protein loading with 1:1,000 diluted GAPDH antibody (sc-25778, Santa Cruz Biotechnology).

Three-Dimensional (3D) Protein Modeling
Missense changes located in the DNA-binding domain of RORA isoform a were solved by 3D modeling. Wild-type (WT) and mutated RORA DNA binding domain homology models were generated according to the crystal structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1 using Modeler software with standard parameters.

RORA Expression Analysis in Neuronal Cell Types
There are four RefSeq transcripts of RORA (GenBank): NM_134261.2 (RORA 1), NM_134260.2 (RORA 2), NM_002943.3 (RORA 3), and NM_134262.2 (RORA 4). Relative mRNA levels of the four RORA transcripts were assessed in three different commercial human cDNAs by quantitative (q)RT-PCR: (1) cerebellar cDNA from a 26-year-old male (Ambio), (2) whole-brain cDNA pooled from two males of Northern European descent aged 43–55 years, Human multiple tissue cDNA (MTC) panel I (Clontech), and (3) Human Universal QUICK-Clone II (Clontech) using transcript-specific primers (Table S2) and SYBR Green PCR Master Mix (Thermo Fisher) on an ABI 7900HT real-time PCR system. We determined relative gene expression levels in quadruplicate samples according to the ΔΔCt method. After normalization to β-actin levels, −log2 of ΔΔCt for each of the four transcripts was added together to determine total expression levels from the RORA locus for each cDNA. Relative RORA transcript expression levels expressed in −log2 of ΔΔCt were obtained by calculating the percentage of each transcript compared to total RORA expression.

Zebrafish Lines and Husbandry
All zebrafish work was performed in accordance with protocols approved by the Duke University Institutional Animal Care and Use Committee. Zebrafish embryos were obtained by natural matings of WT (ZDR strain, Aquatica BioTech) or transgenic (neural: eGFP) adults and maintained on a 14 hr/10 hr light-dark cycle. Embryos were reared in embryo media (0.3 g/L NaCl, 75 mg/L CaSO4, 37.5 mg/L NaHCO3, 0.003% methylene blue) at 28°C until processing for phenotypic analyses at 3 days post-fertilization (dpf).

CRISPR/Cas9 Genome Editing of roraa in Zebrafish
We designed two guide (g) RNAs targeting roraa (GRCz10: ENSDARG00000031768) using the CHOPCHOP v2 tool and synthesized them in vitro with the GeneArt precision gRNA synthesis kit (Table S2; Thermo Fisher) as described. We targeted the roraa locus by microinjection into the cell of zebrafish embryos with 1 nL of cocktail containing 100 pg gRNA and 200 pg Cas9 protein (PNA Bio) at the 1-cell stage. We harvested individual embryos (n = 8) at 1 dpf for DNA extraction to assess targeting efficiency. We PCR-amplified the region flanking the targeted site (Table S2), denatured the resulting product, and reannealed it slowly to form heteroduplexes (95°C for 5 min, ramped down to 85°C at 1°C/s and then to 25°C at 0.1°C/s). We performed polyacrylamide gel electrophoresis (PAGE) on a 20% precast 1 mm gel (Thermo Fisher) to visualize heteroduplexes. To estimate mosaicism of F0 mutants, PCR products were cloned into a TOPO-TA vector (Thermo Fisher) and individual colonies (n = 24) were sequenced (n = 3 larvae/gRNA).

Transient roraa Suppression, In Vivo Complementation, and Heterologous Expression Experiments
We designed splice blocking (sb) morpholinos (MO) targeting either the splice donor site of exon 2 (e2i2) or exon 3 (e3i3) of roraa (GeneTools, LLC; Table S2) and injected 1 nL of MO into the yolk of zebrafish embryos at 1- to 4-cell stage. To confirm expression of the two annotated roraa transcripts (GRCz10: ENSDART00000148537.2 and ENSDART0000121449.2) and to determine MO efficiency, we harvested uninjected control and MO-injected larvae in Trizol (Thermo Fisher) at 3 dpf, extracted total RNA, and conducted first-strand cDNA synthesis with the Quant iTect Reverse Transcription kit (QIAGEN). The targeted region of roraa was PCR amplified using primers complementary to sites in flanking exons (Table S2) and migrated by electrophoresis on a 1% agarose gel; bands were excised, gel purified using QIAquick gel purification kit (Qiagen) and sequenced. The optimal MO dose for in vivo complementation experiments was determined by injection of three concentrations of MO (2, 3, 4 ng of e2i2; or 6, 7, 8 ng of e3i3). To generate mRNA for injections, we purchased Gateway-compatible open reading frame (ORF) clones from Genecopoeia (rora 1, RORA 2, and RORA 3) or Thermofisher (RORA 4) and transferred the ORFs to pCS2 + vector by LR cloning II-mediated recombination (Thermo Fisher). We performed site-directed mutagenesis of RORA 4 according to the QuikChange protocol (Table S2; Agilent), using described methodology; sequences were validated by Sanger sequencing. Linearized pCS2 + vectors containing WT or mutant ORFs were transcribed in vitro with the mMessage mMACHINE SP6 Transcription kit (Ambion).

Whole-Mount Immunostaining
We stained axonal tracts of the cerebellum with monoclonal anti-acetylated tubulin antibody produced in mouse (Sigma-Aldrich, T7451, 1:1,000), as described. We fixed larvae in Dent’s solution (80% methanol and 20% DMSO) and carried out primary antibody detection overnight and secondary detection for 1 hr with
Alexa Fluor 488 goat anti-mouse IgG (A11001, Invitrogen; 1:1,000). We visualized the Purkinje compartment of the zebrafish cerebellum with anti-zebrin II antibody (a gift from Dr. Richard Hawkes) in neurod:egfp transgenic larvae. Briefly, zebrafish larvae were fixed in 4% paraformaldehyde and incubated overnight in mouse anti-zebrin II antibody (1:100); secondary antibody was applied for 1 hr (Alexa Fluor 594 goat anti-mouse IgG, A11005, Invitrogen; 1:1,000). Fluorescent signal was imaged manually on dorsally positioned larvae using an AxioZoom.V16 microscope and Axiocam 503 monochromatic camera, using Zen Pro 2012 software (Zeiss). Cerebellar structures of interest or optic tecta were measured using ImageJ. Total cerebellar area was measured on acetylated tubulin-stained larvae by outlining structures with fluorescent signal; regions comprised of Purkinje cells were measured on zebrin II-stained regions; region comprised of granule cells were measured on GFP-positive regions. Statistical analyses were performed using a two-tailed parametric t test (GraphPad software).

Results

Identification of Point Mutations or Copy-Number Variants Disrupting RORA
As part of our ongoing studies to understand the molecular basis of neurodevelopmental disorders, we identified a total of 16 individuals with rare variants suspected to alter RORA function (Figure 1; Tables 1, S3, and S4). The first individual under investigation was a female with severe syndromic ID, multifocal seizures, mild cerebellar hypoplasia, and hypotonia (individual 6, Table 1; Figure 2A). Upon performing WES, we identified a de novo frameshifting mutation in RORA (GenBank: NM_134261.2; c.1019delG [p.Arg340Profs*17]) that was not present in the genome aggregation database (gnomAD; >246,000 chromosomes) or the NHLBI Exome Variant Server (EVS; >13,000 alleles). Information exchange on community data-sharing platforms including GeneMatcher, the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER), and the Broad Institute matchbox repository facilitated the identification of an additional 15 affected individuals with RORA variants who displayed overlapping phenotypes. These variants were present de novo in 11 simplex families with affected individuals and segregated in one three-generation pedigree under a dominant paradigm, bolstering the candidacy of this locus. Of note, segregation analysis was impossible in individual 1’s family, since the child had been adopted.

Of the 16 affected individuals, 4 harbored likely pathogenic SNVs predicted to alter RORA protein sequence or
dose. In addition to individual 6, three affected individuals had de novo SNVs or small indels in RORA; these included one frameshifting mutation, one nonsense mutation, and one mosaic canonical splice acceptor site change (Figure 1A; Tables 1 and S3). These three RORA variants are predicted to result in protein sequence with either complete (c.804_805delGT [p.Ser269Hisfs*13]) or partial (c.1019delG [p.Arg340Profs*17] and c.1498C>T [p.Arg500*]) truncation of the ligand binding domain at the C terminus and/or nonsense-mediated mRNA decay (NMD). Notably, analyses of fibroblast protein lysates from individual 6 (p.Arg340Profs*17) showed that RORA protein levels were not decreased compared to control subjects, suggesting that related mRNA harboring the premature stop codon would not be eliminated by NMD. According to Alamut software (compiling five different prediction tools: SpliceSite Finder-like, MaxEntScan, NNSPLICE, GeneSplicer, and Human Splicing Finder), the splice acceptor site mutation, c.425/G>C>A, is predicted to cause in-frame exon skipping and deletion of 77 amino acids (p.Ala142-Leu273del); however, cell lines were unavailable from the affected individual to confirm this prediction; mosaicism was estimated at about 20% from exome and Sanger sequencing data.

An additional seven affected individuals had non-recurrent CNVs impacting the RORA locus detected by high-resolution aCGH (Tables 1 and S4; Figures 1A–1C and S1). We identified one ~63 kb duplication interrupting RORA exons 3–6, which is predicted to result in a premature termination of the protein (p.Gln315Leufs*51), one intragenic CNV that resulted in a ~27 kb deletion of RORA exon 3 and its flanking intronic regions to produce a putative frameshifting truncation, and one ~3.7 Mb intergenic deletion impacting 17 genes on 15q22.3q22.2 including RORA. Segregation in each of the three families showed that all three CNVs occurred de novo. Furthermore, we detected a ~1.5 Mb intergenic deletion in four individuals segregating neurological phenotypes in an autosomal-dominant pedigree (Figure 1B); this CNV encompasses nine genes, including RORA.

Finally, an additional five affected individuals had missense mutations predicted to be deleterious (Tables 1 and S3; Figure S2). All five alleles were absent from all available public databases queried (gnomAD, ExAC, EVS). Two of the five variants, c.1225A>C (p.Ser409Arg) and c.1385G>A (p.Arg462Gln), map to the ligand binding domain; both residues are conserved in vertebrate orthologs of RORα but not in the human paralogous nuclear receptor proteins, indicating that these positions are specific to the ROR sub-family. The other three variants—c.269C>G (p.Cys90Ser), c.275G>C (p.Gly92Ala), and c.281A>G (p.Lys94Arg)—are located in the conserved zinc-finger DNA-binding domain of RORα. The three residues Cys90, Gly92, and Lys94, together with Cys93, belong to the P box motif that is part of the alpha-helix of the first zinc-finger domain of the nuclear receptors and interacts directly with DNA (Figure S3A). Because the c.281A>G variant lies in the donor splice site of exon 3, we examined whether it impacts mRNA splicing in primary skin fibroblasts from individual 3; RT-PCR showed neither abnormal sized product nor semiquantitative differences in cDNA amounts (Figures S4A and S4B). Additionally, western blotting of lysates from primary fibroblasts harboring the c.275G>C change (individual 2) showed a modest increase of RORα levels when compared to a matched control (Figure S4C).

Clinical Features of Individuals with RORA Variants
The 16 affected individuals in our cohort display complex phenotypes with regards to their cognition, motor function, and electrophysiology (Table 1). The predominant phenotype was ID. However, cognitive function was variable among the 16 individuals and ranged from mild to moderate (13/16) to severe (2/16), while one individual had a low intelligence quotient (IQ) without ID. We noted IQ regression in one affected individual (individual 4) who had cognitive decline at ~10 years...
<table>
<thead>
<tr>
<th>Individual ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>female</td>
<td>female</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>GEOGRAPHIC ORIGIN</td>
<td>USA</td>
<td>Estonia</td>
<td>France</td>
<td>USA</td>
<td>USA</td>
<td>France</td>
<td>USA</td>
</tr>
<tr>
<td>AGE AT LAST INVESTIGATION (YEARS)</td>
<td>4</td>
<td>3.5</td>
<td>6</td>
<td>28</td>
<td>3</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>MUTATION TYPE</td>
<td>missense</td>
<td>missense</td>
<td>missense</td>
<td>splice site</td>
<td>frameshift</td>
<td>frameshift</td>
<td>missense</td>
</tr>
<tr>
<td>MUTATION (GRCh37)</td>
<td>c.269G&gt;C</td>
<td>c.275G&gt;C</td>
<td>c.281A&gt;G</td>
<td>c.425-1G&gt;A</td>
<td>c.804_805delGT</td>
<td>c.1019delG</td>
<td>c.1225A&gt;C</td>
</tr>
<tr>
<td>MODE OF INHERITANCE</td>
<td>unknown (adopted)</td>
<td>de novo</td>
<td>de novo</td>
<td>de novo (mosaic, 20% in blood)</td>
<td>de novo</td>
<td>de novo</td>
<td>de novo</td>
</tr>
<tr>
<td>FUNCTIONAL EFFECT</td>
<td>NA</td>
<td>dominant toxic</td>
<td>dominant toxic</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

**GROWTH PARAMETERS**

| BIRTH WEIGHT (GRAMS/SD) | NA | 3,500/0 | 3,310/0 | 4,238/+2 | 4,054/+1.4 | 2,655/~1.5 | 3,487/0 |
| BIRTH LENGTH (CM/SD) | NA | 52/~1 | 49/0 | NA | 54.6/~1.9 | 46/~2 | 52/~1 |
| BIRTH HEAD CIRCUMFERENCE (CM/SD) | NA | 38/~2.5 | 34.5/0 | NA | NA | 33/~1 | 34/~0.5 |
| HEIGHT AT AGE LAST INVESTIGATION (CM/SD) | 102/0 | 102/~0.5 | 104/~2 | 167.8/~0.5 | 89/~1.5 | 132/~1 | 120.8/~2 (8 y 5 mo) |
| WEIGHT AT AGE LAST INVESTIGATION (KG/SD) | 17.2/~0.5 | 17.8/~1 | 16/~1.5 | 69.4/~1 | 14.5/~0.3 | 26.2/~1.3 | 31/~1.3 (9 y 3 mo) |
| HEAD CIRCUMFERENCE AT AGE LAST INVESTIGATION (CM/SD) | 49/0 | 50.6/0 | 50/0 | 58.4/~1.9 | NA | 53/0 | 51.7/~0.3 (8 y 4 mo) |
| DEGREE OF DEVELOPMENTAL DELAY OR ID | mild | mild-moderate | severe | mild (regression at 10 years) | mild | severe | mild |
| AGE OF WALKING | 2 years | 3 years | 6 years | 14 months | 20 months | 3 years | 15 months |
| AGE OF FIRST WORDS | 2 years | delayed | before 1 year | 14-15 months | 12-13 months | 5 years | >2 years |
| CURRENT LANGUAGE ABILITY | speaking with sentences | speaking with sentences | no phrase | normal | around 20 words | no phrase | delayed |
| BEHAVIORAL ANOMALIES | no | no | no | no | ASD | ASD | ASD |

*Continued on next page*
<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>9</th>
<th>10D</th>
<th>10C</th>
<th>10B</th>
<th>10A</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>male</td>
<td>male</td>
<td>female</td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>male</td>
<td>male</td>
<td>male</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>USA</td>
<td>Belgium</td>
<td>Belgium</td>
<td>Belgium</td>
<td>Belgium</td>
<td>France</td>
<td>Denmark</td>
<td>Denmark</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>25</td>
<td>50</td>
<td>44</td>
<td>80</td>
<td>15</td>
<td>8</td>
<td>6,5</td>
<td></td>
</tr>
<tr>
<td>missense</td>
<td>nonsense</td>
<td>15q22 deletion including RORA</td>
<td>15q22 deletion including RORA</td>
<td>15q22 deletion including RORA</td>
<td>15q22 deletion including RORA</td>
<td>15q22 deletion including RORA</td>
<td>15q22 deletion including RORA</td>
<td>15q22 intragenic deletion of RORA</td>
<td>15q22 disruptive duplication of RORA</td>
</tr>
<tr>
<td>c.1385G&gt;A</td>
<td>c.1498C&gt;T</td>
<td>15q22.2 (59,641, 986-61,104, 231) x1</td>
<td>15q22.2 (59,641, 986-61, 104,231) x1</td>
<td>15q22.2 (59,641, 986-61, 104,231) x1</td>
<td>15q22.2 (59,641, 986-61, 104,231) x1</td>
<td>15q22.2 (59,641, 986-61, 104,231) x1</td>
<td>15q22.2 (58,622,268 -62,320,616) x1</td>
<td>15q22.2 (60,809, 984-60,837, 029) x1</td>
<td>15q22.2 (60,797, 691-60,860, 668) x3</td>
</tr>
<tr>
<td>p.Arg462Gln</td>
<td>p.Arg500*</td>
<td>de novo</td>
<td>dominant; inherited from mother (ID 10C)</td>
<td>dominant; inherited from mother (ID 10A)</td>
<td>dominant; inherited from mother (ID 10A)</td>
<td>dominant</td>
<td>de novo</td>
<td>de novo</td>
<td>de novo</td>
</tr>
<tr>
<td>loss-of-function</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>NA</td>
<td>3,340/−0.4</td>
<td>NC/+0.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>3,200/−0.5</td>
<td>4050/+1</td>
<td>NC; unremarkable</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>48.3/−0.7</td>
<td>NC/+0.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>49.5/0</td>
<td>56/+2.5</td>
<td>NC; unremarkable</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>46.5/+2 (6 mo)</td>
<td>NC/+0.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>34/−0.5</td>
<td>NA</td>
<td>NC; unremarkable</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>150.4/−0.7 (13 y)</td>
<td>166/+0.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>170/0</td>
<td>NC/+2.5</td>
<td>NC/−1</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>44.4/−0.2 (13 y)</td>
<td>74.3/+1.3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>58/+0.7</td>
<td>NC/+1</td>
<td>NC/−1</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>58/+2.6 (13 y)</td>
<td>54/−0.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>56/+0.7</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>no ID (IQ 85)</td>
<td>moderate</td>
<td>mild</td>
<td>mild</td>
<td>mild</td>
<td>mild</td>
<td>moderate</td>
<td>mild</td>
<td>mild</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>24 months</td>
<td>16 months</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>21 months</td>
<td>normal</td>
<td>mild delay</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>delayed</td>
<td>14 months</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2 years</td>
<td>delayed</td>
<td>delayed</td>
<td></td>
</tr>
<tr>
<td>normal</td>
<td>rudimentary sentences</td>
<td>delayed</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>normal</td>
<td>delayed</td>
<td>delayed</td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>no</td>
<td>behavioral problems</td>
<td>no</td>
<td>borderline personality disorder</td>
<td>no</td>
<td>hyperactivity</td>
<td>no</td>
<td>ASD</td>
<td></td>
</tr>
</tbody>
</table>

(Continued on next page)
of age. Developmental milestones were also delayed. The mean age of walking was delayed in 8/12 individuals; 11/12 walked by 3 years of age. Further, 11/13 individuals had speech delay and/or poor verbal communication abilities, and 2 were not speaking in sentences by the age of 6. We diagnosed epilepsy in 11/16 individuals (Table 1); the predominant seizure semiology was that of a generalized epilepsy with absences, drop attacks, and tonic-clonic seizure sub-types. None of the patients with RORA intragenic mutation had overt dysmorphic facial features. Of the nine individuals who underwent brain MRI, six had normal results and three were diagnosed with cerebellar hypoplasia, which predominantly affected the vermis (individuals 2 and 3; Figures 2B and 2C). These individuals developed early-onset ataxia and hypotonia by age 1.

**Modeling RORA Disruption in Zebrafish**

To corroborate the Rora<sup>−/−</sup> mouse mutant phenotype data<sup>8,23</sup> and to determine the effect of missense variants identified in affected individuals, we developed zebrafish models of RORA ablation and expression. We and others have shown previously that zebrafish is a robust model of neuroanatomical phenotypes observed in humans. In particular,
assay of cerebellar defects in zebrafish has provided crucial insights toward understanding underlying pathomechanism, especially given the high conservation of granule and Purkinje cell types between mammals and teleost species. First, we aimed to model RORA disruption in vivo by targeting and knocking down the relevant zebrafish ortholog. Through reciprocal BLAST of the zebrafish genome and four annotated human RORA transcripts (GenBank: NM_134261, NM_134260, NM_002943, and NM_134262; Figure S5A), we identified two zebrafish orthologs: roraa (Ensembl ID: ENSDARG00000031768; GRCz10; 88%, 91%, 91%, and 91% identity to proteins encoded by RORA1, RORA2, RORA3, RORA4, respectively) and rorab (Ensembl ID: ENSDARG0000001910; 69%, 72%, 72%, and 73% identity to RORA1, RORA2, RORA3, RORA4, respectively). Next, we considered endogenous expression data to determine the most appropriate D. rerio transcript(s) to modulate in vivo. RNA in situ hybridization studies in zebrafish larvae have documented roraa expression patterns in the developing cerebellum as well as other anterior structures (optic tecta, hindbrain, and retina). However, rorab expression is restricted to the hindbrain. Considering both the amino acid conservation and also the spatiotemporal expression of roraa in the
Figure 3. Disruption of rora in Zebrafish Larvae Results in Cerebellar Hypoplasia Driven by Purkinje and Granule Cell Loss
(A) Schematic of neuroanatomical structures painted with anti-acetylated tubulin antibody at 3 days post fertilization (dpf); IT, intertectal.
(B) Representative dorsal images of acetylated tubulin immunostained larvae show that rora ablation causes cerebellar defects in CRISPR/Cas9 F0 mutants and morphants. Cerebellar size was measured as indicated by the dashed red outline on inset panels.
(C) Quantification of cerebellar area in larval batches is shown for two guide (g)RNAs targeting either rora exon 5 (gRNA 1) or exon 8 (gRNA 2).
(D) rora morphants (injected with 3 ng morpholino; MO) display a cerebellar phenotype that can be rescued by four different wild-type RORA mRNA transcripts: co-injection of rora e2i2 splice-blocking MO with RORA splice variants (RORA 1, GenBank: NM_134261, RORA 2: NM_134260, RORA 3: NM_002943 and RORA 4: NM_134262).
(E) Schematic of cerebellar cell types assessed in 3 dpf larvae using either a neurod:egfp transgene (green) or anti-zebrin II immunostaining (red). Orientation is indicated with A, anterior; P, posterior; L, left; R, right.

(legend continued on next page)
developing cerebellum, we deemed rora as the most relevant D. rerio locus to test the activity of RORA mutations in humans.

Haploinsufficiency or knockout of Rora in mice lead to abnormal cerebellar layer morphology.\(^8^,\)\(^21^\) Therefore, to assess the effect of loss of RORA function consistent with deletion or truncating mutations identified in affected individuals, we compared the size of the developing cerebellum between rora knock-down and uninjected controls. The zebrafish rora locus has two annotated transcripts (GRcz10: [roraa-201] ENSDART00000121449.2 and [roraa-202] ENSDART00000148537.2), the latter of which has an incomplete 5’ coding sequence; RT-PCR using cDNA originating from 3 dpf embryos confirmed that both transcripts are detectable in this time window (Figures S6A and S6B). Next, we generated CRISPR/Cas9-based zebrafish F0 mutant models by targeting exon 5 or exon 8 of rora-201 (corresponding to exon 4 or exon 7 of rora-202) to produce mosaic mutants with >90% mosaicism (Figure S7). Immunostaining of the central nervous system (CNS) with anti-acetylated tubulin antibody and measurements of neuroanatomical structures showed that rora F0 mutants display a reduced cerebellar area and smaller optic tecta area compared to either control larvae or larvae injected with gRNA alone (p < 0.0001, for gRNA 1 and gRNA 2, n = 33–43 and 41–45 larvae/batch, respectively, repeated, masked scoring; Figures 3A–3C and S8A–S8C).

To assess phenotype specificity and to determine the effect of RORA missense variants, we performed MO-mediated transient suppression of rora. We designed two sb MOs targeting the splice donor site of exons 2 and 3 of rora-201, which we injected into embryo batches, generated cDNA at 3 dpf, and performed RT-PCR to determine efficiency; subsequent Sanger sequencing confirmed a frameshift deletion of rora exon 2 or exon 3 (Figures S6C and S6D). Next, we injected increasing concentrations of either MO (2, 3, 4 ng e2i2; or 6, 7, 8 ng e3i3). Consistent with our CRISPR F0 mutant data, we observed a dose-dependent reduction in cerebellar size for both reagents (p < 0.0001, 30 to 94 larvae/condition, Figure S6E). Co-injection of e2i2 sb MO with each of the four RefSeq annotated WT human RORA mRNAs rescued cerebellar and optic tecta defects, indicating MO specificity (p = 0.0060, 0.0387, < 0.0001, < 0.0001 for RORA 1, 2, 3, and 4, respectively, versus MO, n = 31–60 larvae/batch, repeated, Figures 3B, 3D, S8A, and S8D). In parallel, heterologous expression of each of the four WT RORA mRNAs did not lead to cerebellar phenotypes that differed from uninjected controls (n = 40–52 larvae/batch, Figure S9). To investigate further the relative expression of the four annotated RORA transcripts and to identify the most relevant isoform for in vivo modeling in zebrafish, we performed qRT-PCR on commercial human adult cDNA; we identified RORA 4 as the most abundantly expressed transcript in the human CNS (Figures S5A and S5B), potentially explaining the observation that WT RORA 4 mRNA produced the most significant restoration of cerebellar size when co-injected with MO.

**Altered Purkinje and Granule Layers in rora Zebrasfish Models**

Similar to mammals, Purkinje and granule cell crosstalk is vital for the function of the zebrafish cerebellum.\(^29^\) Cerebellar cell subpopulations have been characterized previously in zebrafish with lineage-specific markers and transgenic lines. In neurod:egfp transgenic zebrafish, granule cells are marked with GFP.\(^29^\) Furthermore, zebrin II is a specific marker of the Purkinje compartment,\(^27^\) which develops between 2.3 and 4 dpf.\(^31^\) To assess whether rora suppression affects granule or Purkinje cell layers of zebrafish cerebellum, we conducted whole-mount zebrin II immunostaining on neurod:egfp transgenic zebrafish larvae at 3 dpf. F0 rora mutants presented with a significantly decreased size of Purkinje and granule cell layers compared to controls (p < 0.0001 for both cell types; Purkinje cells, n = 26–55; granule cells, n = 11–55, repeated; Figures 3E–3H and S10). Importantly, gRNA injected alone was indistinguishable from controls (Figures 3G and 3H). Injection of e2i2 MO recapitulated these observations (p < 0.0001 or p < 0.001 for Purkinje [n = 29–55] and granule [n = 30–55] cells, respectively, in morphants versus controls; repeated; Figures 3E–3H and S10). Further, co-injection of e2i2 MO with WT human RORA 4 mRNA rescued the measured area of each of the granule and Purkinje compartments (p < 0.0001 or p < 0.001 versus MO alone for Purkinje or granule cells, respectively; Figures 3E–3H and S10). Together, our data confirm that in F0 mutant or transient knockdown zebrafish models, rora disruption leads to defects of Purkinje and granule cell layers, consistent with the Rora\(^-\) mouse model.\(^34^–\)\(^36^\)

**In Vivo Complementation Studies Indicate Dual Direction of RORA Allele Effect**

To determine the pathogenicity of RORA missense variants (c.275G>C [p.Gly92Ala], c.281A>G [p.Lys94Arg], and c.1385G>A [p.Arg462Gln]), we assessed the effect of mutant RORA 4 mRNA in the presence or absence of MO.

---

(F) Representative dorsal images show that reduction of Purkinje and granule cells contributes to cerebellar defects induced by rora targeting. Transgenic neurod:egfp larvae were fixed and immunostained with anti-zebrin II antibody (red), and the area comprised of cerebellar cell subpopulations have been characterized previously in zebrafish with lineage-specific markers and transgenic lines. In neurod:egfp transgenic zebrafish, granule cells are marked with GFP. Furthermore, zebrin II is a specific marker of the Purkinje compartment, which develops between 2.3 and 4 dpf. To assess whether rora suppression affects granule or Purkinje cell layers of zebrafish cerebellum, we conducted whole-mount zebrin II immunostaining on neurod:egfp transgenic zebrafish larvae at 3 dpf. F0 rora mutants presented with a significantly decreased size of Purkinje and granule cell layers compared to controls (p < 0.0001 for both cell types; Purkinje cells, n = 26–55; granule cells, n = 11–55, repeated; Figures 3E–3H and S10). Importantly, gRNA injected alone was indistinguishable from controls (Figures 3G and 3H). Injection of e2i2 MO recapitulated these observations (p < 0.0001 or p < 0.001 for Purkinje [n = 29–55] and granule [n = 30–55] cells, respectively, in morphants versus controls; repeated; Figures 3E–3H and S10). Further, co-injection of e2i2 MO with WT human RORA 4 mRNA rescued the measured area of each of the granule and Purkinje compartments (p < 0.0001 or p < 0.001 versus MO alone for Purkinje or granule cells, respectively; Figures 3E–3H and S10). Together, our data confirm that in F0 mutant or transient knockdown zebrafish models, rora disruption leads to defects of Purkinje and granule cell layers, consistent with the Rora\(^-\) mouse model. To determine the pathogenicity of RORA missense variants (c.275G>C [p.Gly92Ala], c.281A>G [p.Lys94Arg], and c.1385G>A [p.Arg462Gln]), we assessed the effect of mutant RORA 4 mRNA in the presence or absence of MO.

(G) Quantification of granule layer cells (GFP-positive).

(H) Quantification of Purkinje cells (zebrin II-positive). Both cell populations are reduced significantly in rora F0 mutants as well as morphant zebrafish.

Scale bar in (B) and (F): 100 μm. AU, arbitrary units. Stars indicate p value compared to uninjected controls (CRISPR/Cas9 and MO) or to morphants (MO + RORA). ****p < 0.0001, ***p < 0.001, **p < 0.01, *p < 0.05; ns, not significant. Error bars in (C), (D), (G), and (H) represent the 5th and 95th percentiles.

---

754 The American Journal of Human Genetics 102, 744–759, May 3, 2018
Figure 4.  RORA Missense Variant Located in the DNA Binding Domain and Ligand Binding Domain Confer a Dominant Toxic and Loss-of-Function Effect, Respectively

(A) Complementation of roraa morpholino (MO) with variant mRNA results in severe gross morphological defects or larval death before 3 days post-fertilization (dpf). Left, representative embryo classes are shown; right, quantification of outcomes. p.Gly92Ala (G92A), p.Lys94Arg (K94R) are located in the DNA binding domain; p.Arg462Gln (R462Q) localizes to the ligand-binding domain; T476A (rs190933482) is a negative control for the assay with an allele frequency of 0.001091 in gnomAD.

(B) Quantification of cerebellar area of larval batches with ectopic expression of RORA mRNA. G92A mRNA confers a significant reduction in cerebellar size.

(C) Representative dorsal images of 3 dpf larvae injected with gradually modified doses of WT and G92A RNA (top, amount in pg) and stained with anti-acetylated tubulin antibody.

(D) Quantification of cerebellar measurements from the batches shown in (C).

(legend continued on next page)
First, we tested a loss-of-function hypothesis by co-injecting \textit{roraa} e2i2 MO with WT or variant \textit{RORA} 4 mRNA. In vivo complementation of \textit{roraa} MO with p.Gly92Ala or p.Lys94Arg encoding mRNA resulted in early developmental defects including a reduction in the size of anterior structures and tail extension failure (65% and 69%, respectively) and a high mortality rate (19% and 10%, respectively), while co-injection of MO with p.Arg462Gln, WT, or a potentially benign variant p.Thr476Ala (rs190933482, minor allele frequency 0.001 in gnomAD) induced no appreciable morphological defects (Figure 4A). The gross morphological phenotype observed by co-injection of MO with mutant mRNAs located in the DNA binding domain was more severe than the morphant phenotype and suggested either a dominant-negative or a gain-of-function effect. To investigate these possibilities, we injected 200 pg of each mutant mRNA alone and again observed notable mortality at 3 dpf (2% for controls [1 dead larva/54] versus 31% for p.Gly92Ala [14/45]; 9% for p.Lys94Arg [4/46]; and 7% for p.Arg462Gln [3/41]). Of the larvae that survived to 3 dpf, we performed anti-acetylated immunostaining and measurement of the cerebellar area; p.Gly92Ala encoding mRNA induced a reduction in the mean cerebellar area compared to WT (p < 0.0001, 31–53 larvae/batch, Figure 4B). To determine whether WT mRNA could rescue the phenotypes induced by p.Gly92Ala, we titrated increasing amounts of p.Gly92Ala mRNA together with \textit{RORA} 4 WT mRNA and observed a dose-dependent effect that correlated with the injected dose of p.Gly92Ala (Figures 4C and 4D). These data suggested a dominant toxic effect as the likely mechanism for the p.Gly92Ala variant. Although expression of p.Lys94Arg induced morphological defects similar to p.Gly92Ala and we observed a broadened distribution of cerebellar measurements at the highest dose of mutant mRNA that was consistent with p.Gly92Ala, these results did not reach statistical significance (Figures 4E and 4F). Finally, in vivo complementation of \textit{roraa} MO with p.Arg462Gln encoding mRNA did not rescue the size of the cerebellum (Figures 4G and 4H, p < 0.001, n = 40–41), suggesting a loss-of-function effect as the possible disease mechanism for this change impacting the ligand binding domain (Figures 4G and 4H).

\section*{Discussion}

Here, we describe a cohort of 16 affected individuals who harbor 13 different rare variants disrupting \textit{RORA}. We observe a clinical spectrum of neurodevelopmental delay with at least two different presentations: (1) a cognitive and motor phenotype and (2) a cognitive and behavioral phenotype. The first sub-phenotype is characterized by a moderate to severe ID with a marked ataxic component, severe cerebellar vermis hypoplasia, and epilepsy with predominant generalized seizures as already reported in \textit{RORB}, whereas the hallmark of the second sub-phenotype is ASD with mild ID or normal cognition frequently associated with epilepsy. In addition to these two groups, three individuals exhibit only mild ID without behavioral problems.

Although the size of our cohort limits the delineation of an unambiguous genotype-phenotype correlation, we can infer pathomechanisms related to specific phenotypes from our in vivo complementation data. Individuals 2 and 3, who harbor the two dominant toxic mutations of the RORα DNA binding domain, display severe ID and motor phenotypes likely due to cerebellar hypoplasia. A similar severe phenotype is coincident with the truncating deletion (individual 6; p.Arg462Gln[17], raising the possibility that this mutation could also result in a dominant toxic effect by encoding a truncated protein; immunoblotting studies from fibroblast protein lysates showed RORα protein levels to be similar to that of controls, arguing against haploinsufficiency (Figure S4C). Individual 6 also displayed autistic traits, but these phenotypes occurred secondary to established ID, potentially excluding her from the Autism Diagnosis Interview (ADI) criteria for idiopathic ASD.

By contrast, we note variability in cognitive function and behavioral phenotypes in individuals with likely haploinsufficiency. One individual displays an ASD phenotype in the absence of ID (individual 8), and in vivo complementation testing in zebrafish indicate that the p.Arg462Gln mutation confers a loss-of-function effect. Further, 6/12 individuals from simplex families who likely have a reduction in protein dosage display mild to moderate ID but with normal behavior, while the remaining 5/12 affected individuals from simplex families present with both ID and autistic features. Notably, there is phenotypic variability among the individuals within multiplex family 10: all four individuals display mild ID, but only two display behavioral anomalies. However, we cannot exclude the possibility of trans effects from elsewhere in the genome, especially from \textit{DISC1} affected by the CNVs harbored by individuals 10A–C or individual 11 (Figure S1; Table S4). Phenotypic variability is not uncommon for neurodevelopmental disorders, and here

\begin{itemize}
  \item [(E)] Representative dorsal images of 3 dpf larvae injected with gradually modified doses of WT and K94R RNA (top, amount in pg) and stained with anti-acetylated tubulin antibody.
  \item [(F)] Quantification of cerebellar measurements from the batches shown in (E).
  \item [(G)] Representative dorsal images of 3 dpf larvae injected either with e2i2 splice-blocking MO or MO co-injected with WT \textit{RORA} mRNA, R462Q or population control T476A variant \textit{RORA} mRNA.
  \item [(H)] Quantification of cerebellar measurements from the batches shown in (E).
\end{itemize}
we account, in part, for variability due to allelism at the RORA locus by elucidating the direction of allele effect for missense changes.

The sub-group of our cohort who show a constellation of cognitive and motor defects are reminiscent of the homozygous staggerer (sg/sg) mouse, an early reported animal model of Rora ablation. The staggerer mutation consists of an intragenic CNV that results in a 122-bp frameshifting deletion that truncates the ligand binding domain, leading to the loss of RORz activity. This phenotype is similar to Rora<sup>−/−</sup> mice in which the most obvious symptom is an ataxic gait associated with defective Purkinje cell development leading to an abnormal cerebellar size. Heterozygous Rora<sup>+</sup> mice present with a comparable phenotype, although they display a late onset of neuronal loss and reduced phenotypic severity. Further studies will be required to understand the precise molecular mechanisms of p.Gly92Ala and p.Lys94Arg to account for their apparent toxic effects. We speculate that these changes in the DNA binding domain might hamper access of WT RORz to its natural target sites (Figure S3B), thereby leading to a phenotype resembling to the staggerer and Rora<sup>−/−</sup> homozygous mutants.

The autistic signs observed in two individuals with truncating mutations (individuals 6 and 13) and two individuals with missense mutations altering the ligand binding domain of RORA (individuals 7 and 8) are in agreement with recent reports suggesting that RORA is a candidate gene for ASD. ChIP-on-chip analysis has revealed that RORz can be recruited to the promoter regions of 2,544 genes across the human genome, with a significant enrichment in biological functions including neuronal differentiation, adhesion, and survival, synaptogenesis, synaptic transmission and plasticity, and axonogenesis, as well as higher-level functions such as development of the cortex and cerebellum, cognition, memory, and spatial learning. Independent ChIP-quantitative PCR analyses confirmed binding of RORA to promoter regions of selected ASD-associated genes, including A2BP1, CYP19A1, ITPR1, NLGN1, and NTRK2, whose expression levels are also decreased in RORA-repressed human neuronal cells and in prefrontal cortex tissues from individuals with ASD. Additionally, two RORA polymorphisms (rs11639084 and rs4774388) have been associated with ASD risk in Iranian individuals. Consistent with these data, treatment with a synthetic RORAγ agonist, SR1078, reduced repetitive behavior in the BTBR mouse model of autism, suggesting that RORA upregulation could be a viable therapeutic option for ASD.

In summary, our data implicate a diverse series of disruptive mutations in RORA with neurological phenotypes hallmarked by ID and either severe motor phenotypes or behavioral anomalies. Through combined clinical, genetic, and functional studies, we expand the genetic basis of rare neurodevelopmental syndromes and show how in vivo modeling can reveal dual molecular mutational effects.

Supplemental Data
Supplemental Data include ten figures and four tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.02.021.

Acknowledgments
We are grateful to Dr. R. Hawkes and Dr. C. Armstrong (University of Calgary) for providing the zebrin II antibody. We also thank Mr. Z. Kupchinsky for zebrafish husbandry, Mr. D. Morrow for assisting with reagents for the in vivo modeling study, and members of the Center for Human Disease Modeling for helpful comments. We are also grateful to Ms. J.A. Rosenfeld (Mokry) for her precious help in our search for similar affected individuals and for her review of the manuscript. This work was supported by funds from the Agence Nationale pour la Recherche/E-rare Joint-Transnational-Call 2011 (2011-RARE-004-01 “Euro-SCAR”) to M.K.; from the National Human Genome Research Institute with supplemental funding provided by the National Heart, Lung, and Blood Institute under the Trans-Omics for Precision Medicine (TOPMed) program and the National Eye Institute (to the Broad Center for Mendelian Genomics [UM1 HG008900]); from the NIH grant T32 HD07466 (to M.H.W.); US NIH grant R01 MH106826 to E.E.D.; and from the Estonian Research Council, grant PUT355 (to S.P. and K.Ö.). We also acknowledge the Deciphering Developmental Disorders (DDD) Study.

K.M., M.T.C., R.W., M.J.G.S., and K.R. are employees of GeneDx, Inc., a wholly owned subsidiary of OPKO Health, Inc. N.K. is a paid consultant for and holds significant stock of Rescindo Therapeutics, Inc.

Received: December 18, 2017
Accepted: February 26, 2018
Published: April 12, 2018

Web Resources
1000 Genomes, http://www.internationalgenome.org/
Clustal: Multiple Sequence Alignment, http://www.clustal.org/
Database of Genomic Variants (DGV), http://dgv.tcag.ca/dgv/app/home
DECIPHER, https://decipher.sanger.ac.uk/
ExAC Browser, http://exac.broadinstitute.org/
gnomAD Browser, http://gnomad.broadinstitute.org/
ImageJ Fiji, http://fiji.sc/Fiji
seqr, https://sequinstitute.org/
The Human Protein Atlas, http://www.proteinatlas.org/
UCSC Genome Browser, http://genome.ucsc.edu
UniProt, http://www.uniprot.org/
ZFIN, http://zfin.org/
References


Example qualifying exam (Ph.D.) and comprehensive exam (M.S.) questions pertaining to Minster et al. (2016) Nature Genetics 48:1049-1054.

1. The Samoan population has a high prevalence of obesity and includes many individuals with extremely high BMI. Explain the rationale for why geneticists often study (1) high-risk populations and (2) individuals with extreme phenotypes.

2. In Samoans, the heritability of BMI is 45%. How does this compare to other populations? How is this heritability estimate related, if at all, to the high BMI observed in the Samoan population?

3. The authors were careful to adjust their statistical model for population substructure and inferred relatedness. What is the reason behind this?

4. The investigators used sequence data from 96 individuals in order to impute unmeasured variants in the entire cohort. Explain the purpose of imputation and describe the logic behind it.

5. The authors state that because of high linkage disequilibrium in the region, conditional analyses were not able to distinguish between the top variants on statistical grounds. Explain the reason for this.

6. Authors used the p-value threshold of $5 \times 10^{-8}$ for statistical significance. This is considerably smaller than the typical p-value threshold of 0.05. Explain why investigators used the stricter p-value threshold?

7. In the LocusZoom plot, some of the points form horizontal bands? Explain why this phenomenon may occur and what, if anything, can be inferred about the linkage disequilibrium and allele frequencies of these SNPs.

8. The investigators report that they have identified evidence of positive selection at the missense variant in Samoans. What is positive selection?
A thrifty variant in CREBRLF strongly influences body mass index in Samoans

Ryan L Minster, Nicola L Hawley, Chi-Ting Su, Guangyun Sun, Erin E Kershaw, Hong Cheng, Olive D Buhule, Jerome Lin, Muagututia Sefuiva Reupena, Satupa’itea Viali, John Tuitele, Take Naseri, Zsolt Urban, Ranjan Deka, Daniel E Weeks & Stephen T McGarvey

Samoans are a unique founder population with a high prevalence of obesity, making them well suited for identifying new genetic contributors to obesity. We conducted a genome-wide association study (GWAS) in 3,072 Samoans, discovered a variant, rs12513649, strongly associated with body mass index (BMI) ($P = 5.3 \times 10^{-14}$), and replicated the association in 2,102 additional Samoans ($P = 1.2 \times 10^{-9}$). Targeted sequencing identified a strongly associated missense variant, rs373863828 (p.Arg457Gln), in CREBRLF (meta $P = 1.4 \times 10^{-29}$). Although this variant is extremely rare in other populations, it is common in Samoans (frequency of 0.259), with an effect size much larger than that of any other known common BMI risk variant (1.36–1.45 kg/m$^2$ per copy of the risk-associated allele). In comparison to wild-type CREBRLF, the Arg457Gln variant when overexpressed selectively decreased energy use and increased fat storage in an adipocyte cell model. These data, in combination with evidence of positive selection of the allele encoding p.Arg457Gln, support a ‘thrifty’ variant hypothesis as a factor in human obesity.

Obesity is essentially a disorder of energy homeostasis and has strong genetic and environmental components. As diets have modernized and physical activity has decreased, the prevalence of overweight and obesity in Samoa has escalated to be among the highest in the world. Over half of men and 84% of women in Samoa were overweight or obese by Polynesian cutoffs (BMI >26 kg/m$^2$); by 2010, prevalence had increased to 80% and 91%, respectively. Although the contribution of environmental factors to this trend is clear, the estimated 45% heritability of BMI in Samoans remains largely unexplained. Genetic susceptibility to obesity in the contemporary obesogenic environment may have resulted from putative selective advantages from efficient energy metabolism acquired during 3,000 years of Polynesian island discoveries, settlement, and population dynamics and/or from genetic drift due to founder effects, small population sizes, and population bottlenecks.

To discover genes influencing BMI, we genotyped 659,492 markers across the genome in our discovery sample of 3,072 Samoans recruited from 33 villages across Samoa using the Affymetrix 6.0 chip (Supplementary Fig. 1 and Supplementary Table 1). We adjusted for population substructure and inferred relatedness using an empirical kinship matrix and then tested for association with BMI using linear mixed models. Quantile–quantile plots indicated that $P$-value inflation was well controlled ($\lambda_{GC} = 1.07$) (Supplementary Fig. 2).

By far, the strongest association with BMI occurred at rs12513649 ($P = 5.3 \times 10^{-14}$) on chromosome 5q35.1 (Fig. 1a), and this association was strongly replicated ($P = 1.2 \times 10^{-9}$) in 2,102 adult Samoans from a 1990–1995 longitudinal study and a 2002–2003 family study, with participants of each study drawn from both American Samoa and Samoa (Table 1 and Supplementary Table 1). To fine-map the region encompassing this signal, we used the Affymetrix-based genotypes to select 96 individuals optimal for targeted sequencing of a 1.5-Mb region centered on rs12513649. The haplotypes generated from the sequencing data were used to impute genotypes for the rest of the discovery sample. Analyses of the imputed data highlighted two significantly associated variants in CREBRLF (encoding CREB3 regulatory factor), rs150207780 and rs373863828 (Fig. 1b). Because of high linkage disequilibrium (LD) in the region, conditional analyses were not able to distinguish between the top variants on statistical grounds (Supplementary Fig. 3). Annotation indicated that neither rs12513649, located between ATP6V0E1 and CREBRLF, nor rs150207780, located in intron 1 of CREBRLF, had any predicted regulatory function, drawing our attention to rs373863828, which was...
the only strongly associated missense variant among the 775 variants with \( P \leq 1 \times 10^{-5} \) in the targeted sequencing region. The rs373863828 missense variant (c.1370G>A, p.Arg457Gln) is located at a highly conserved position (GERP score 5.49) with a high probability of being damaging (SIFT, 0.03; PolyPhen-2, 0.996). The BMI-increasing A allele of rs373863828 has an overall frequency of 0.259 in Samoans but is unobserved or extremely rare in other populations, with an allele count in the Exome Aggregation Consortium of only 5 among 121,362 measured alleles (Table 1). Bayesian fine-mapping with PAINTOR\(^\text{13}\) strongly supported following up the missense variant. The two variants in the region with the highest posterior probability (PP) of being causal were rs373863828 (PP = 0.80) and rs150207780 (PP = 0.22); when Encyclopedia of DNA Elements (ENCODE) functional annotation was included, these probabilities increased to 0.92 and 0.34, respectively.

We then genotyped the missense variant rs373863828 in the discovery and replication samples, obtaining very significant evidence of association with BMI in adults (\( P = 7.0 \times 10^{-13} \) and \( P = 3.5 \times 10^{-9} \), respectively), with a combined meta-analysis \( P \) value of \( 1.4 \times 10^{-20} \) (Table 1). The meta-analysis showed no evidence of heterogeneity (\( P = 0\% \); \( Q = 1.12 \); \( P = 0.571 \)). In our discovery sample, each copy of the A allele increased BMI by 1.36 kg/m\(^2\) (Fig. 1c). In our adult replication sample, each copy of the A allele increased BMI by 1.45 kg/m\(^2\). There was a strong effect on BMI at this locus even after stratifying by sex and cohort (Supplementary Fig. 4), however sex–genotype interactions were not significant (discovery \( P = 0.060 \); replication \( P = 0.555 \)). There was also suggestive evidence (\( P = 1.1 \times 10^{-7} \)) that this variant increased BMI in our sample of 409 Samoan children (Table 1). The rs373863828 variant (encoding p.Arg457Gln) accounted for 1.93% of the variance in BMI in our discovery sample and 1.08% of the variance in BMI in our replication sample. In comparison, rs1558902, the main risk-associated variant in FTO, increases BMI by 0.39 kg/m\(^2\) per copy of the risk-associated allele and accounts for only 0.34% of the variance in BMI in Europeans\(^\text{14,15}\). In searches of the literature and databases (including GRASP\(^\text{16,17}\)), we were unable to identify any significant associations with BMI in the CREBRF region in other human studies.

In addition to BMI, the A allele of rs373863828 was also positively associated with obesity risk (odds ratio (OR) = 1.305 and 1.441 in the discovery and replication cohorts, respectively) as well as measures of total and regional adiposity, including percent body fat, abdominal circumference, and hip circumference, in both cohorts (Table 2 and Supplementary Table 2). The A allele was also positively associated with serum leptin levels in women (both cohorts) and men (replication cohort) before but not after adjusting for BMI. These data indicate that the association between the missense variant and BMI is indeed due to an association with adiposity.

Higher BMI and adiposity are usually associated with greater insulin resistance (higher fasting insulin levels and homeostatic model assessment–insulin resistance (HOMA-IR)), an atherogenic lipid profile (especially higher serum triglyceride and lower HDL cholesterol levels), and lower adiponectin levels. We therefore expected the BMI-increasing A allele of rs373863828 to also be associated with these metabolic variables. However, even though the A allele was consistently associated with higher BMI and adiposity in both the discovery and replication cohorts, the expected associations with the above obesity-related comorbidities were not observed and, in some cases, were even in the opposite direction to that expected (Table 2 and Supplementary Table 2). Notably, when considering all subjects, the risk of diabetes was actually lower (OR = 0.586 for the discovery cohort, \( P = 6.68 \times 10^{-9} \)) and lower adiponectin levels. We therefore expected the BMI-increasing A allele of rs373863828 to also be associated with these metabolic variables. However, even though the A allele was consistently associated with higher BMI and adiposity in both the discovery and replication cohorts, the expected associations with the above obesity-related comorbidities were not observed and, in some cases, were even in the opposite direction to that expected (Table 2 and Supplementary Table 2). Notably, when considering all subjects, the risk of diabetes was actually lower (OR = 0.586 for the discovery cohort, \( P = 6.68 \times 10^{-9} \)) and lower adiponectin levels (1.65 mg/dl (\( P = 9.5 \times 10^{-5} \)) and 1.54 mg/dl (\( P = 8.8 \times 10^{-5} \)) lower for each copy of the A allele, respectively). These effects became even more significant after adjusting for BMI (2.25 mg/dl, \( P = 6.9 \times 10^{-8} \) and 2.09 mg/dl, \( P = 7.6 \times 10^{-6} \)), suggesting an independent effect of the variant on glucose homeostasis and diabetes risk. Such effects are unlikely to be due to survival bias, as no correlation between age and genotype was observed (linear regression \( P = 0.849 \)). These effects seem to be independent of obesity-associated insulin resistance, as associations with fasting insulin levels and HOMA-IR were not consistently observed across the cohorts (associations were stronger only in the replication cohort before adjusting for BMI). Furthermore, although the variant was associated with lower total cholesterol levels in the discovery cohort, consistent effects on serum lipid or adiponectin levels were likewise not observed. Together, these data suggest that the missense variant does not promote, and may even protect against, obesity-associated comorbidities; however, additional studies will be required to confirm these findings and directly test this hypothesis.
Table 1  Association details for rs12513649 and rs373863828

<table>
<thead>
<tr>
<th>SNP rs ID</th>
<th>Discovery variant</th>
<th>Missense variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12513649</td>
<td>rs373863828</td>
<td></td>
</tr>
<tr>
<td>Chromosome</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Physical position (GRCh37.p13) (bp)</td>
<td>172,472,052</td>
<td>172,535,774</td>
</tr>
<tr>
<td>Effect allele</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>Other allele</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>Nearest gene upstream of the SNP</td>
<td>ATP6VOE1</td>
<td>CREBF</td>
</tr>
<tr>
<td>Distance to nearest upstream gene (bp)</td>
<td>10,152</td>
<td>0</td>
</tr>
<tr>
<td>Nearest gene downstream of the SNP</td>
<td>CREBF</td>
<td>CREBF</td>
</tr>
<tr>
<td>Distance to nearest downstream gene (bp)</td>
<td>11,302</td>
<td>0</td>
</tr>
<tr>
<td>Sample sizes (phenotyped and genotyped)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWAS Samoans from the 2010s (discovery)</td>
<td>3,072</td>
<td>3,066</td>
</tr>
<tr>
<td>Samoans from the 1990s (replication)</td>
<td>1,020</td>
<td>1,020</td>
</tr>
<tr>
<td>Samoans from the 2000s (replication)</td>
<td>1,082</td>
<td>1,083</td>
</tr>
<tr>
<td>Meta-analysis of the 1990s and 2000s samples</td>
<td>2,102</td>
<td>2,103</td>
</tr>
<tr>
<td>Meta-analysis of the 1990s, 2000s, and 2010s samples</td>
<td>5,174</td>
<td>5,169</td>
</tr>
<tr>
<td>Samoan children from the 2000s</td>
<td>409</td>
<td>409</td>
</tr>
</tbody>
</table>

| P values for log-transformed BMI |                   |                 |
| GWAS Samoans from the 2010s (discovery) | $5.3 \times 10^{-14}$ | $7.0 \times 10^{-13}$ |
| Samoans from the 1990s (replication) | $5.8 \times 10^{-4}$ | $8.0 \times 10^{-4}$ |
| Samoans from the 2000s (replication) | $3.0 \times 10^{-7}$ | $6.5 \times 10^{-7}$ |
| Meta-analysis of the 1990s and 2000s samples | $1.2 \times 10^{-9}$ | $3.5 \times 10^{-9}$ |
| Meta-analysis of the 1990s, 2000s, and 2010s samples | $4.0 \times 10^{-22}$ | $1.4 \times 10^{-20}$ |
| Samoan children from the 2000s | $4.1 \times 10^{-3}$ | $1.1 \times 10^{-3}$ |

| Effect allele frequencies |                   |                 |
| GWAS Samoans from the 2010s | 0.041 (0.005) | 0.039 (0.005) |
| Samoans from the 1990s (replication) | 0.029 (0.008) | 0.028 (0.008) |
| Samoans from the 2000s (replication) | 0.056 (0.011) | 0.054 (0.011) |
| Samoan children from the 2000s | 0.031 (0.011) | 0.035 (0.011) |

Although the majority of genes contributing to obesity do so by influencing the central regulation of energy balance, emerging evidence highlights the contribution of altered cellular metabolism to obesity. Therefore, we examined the impact of rs373863828 on cellular bioenergetics. To do so, we selected the established 3T3-L1 mouse adipocyte model for two reasons: (i) CREBF is widely expressed in virtually all tissues, including adipose tissue (Supplementary Fig. 5), suggesting a fundamental cellular function, and (ii) several CREB family proteins have been linked to mitochondrial function and metabolic phenotypes in adipocytes. Thus, this model is well suited to assess multiple potentially relevant metabolic phenotypes.

We first characterized the effects of adipogenic differentiation and ectopic overexpression of human wild-type or Arg457Gln CREBF on endogenous Crebrf expression in 3T3-L1 cells. Crebrf expression was induced during adipogenesis in conjunction with that of adipogenic markers (Cebpa, Pparg, and Adipoaq), suggesting a role for CREBF in this process (Supplementary Fig. 6). Indeed, comparable stable overexpression of the transcripts for human wild-type and Arg457Gln CREBF (Fig. 2a), without changing endogenous Crebrf levels (Fig. 2b), was sufficient to induce the expression of adipogenic markers (Fig. 2c–e) and promote lipid and triglyceride accumulation (Fig. 2f–h) in the absence of standard hormonal induction of adipogenesis. Although Arg457Gln CREBF resulted in slightly weaker induction of adipogenic markers than wild-type protein (Fig. 2c,e), it promoted significantly ($P < 0.02$) greater lipid and triglyceride accumulation (Fig. 2f–h). To determine whether this increased energy storage was associated with decreased energy use, we next assessed glycolysis, mitochondrial respiration, and ATP production. Consistent with published data, glycolysis was suppressed and mitochondrial respiration and ATP production were enhanced by hormonally induced adipogenic differentiation (Supplementary Fig. 7). Stable overexpression of wild-type CREBF increased whereas Arg457Gln CREBF decreased multiple measures of cellular energy use, including basal and maximal mitochondrial respiration, mitochondrial ATP production, and basal glycolysis (Fig. 2i). These data indicate that the Arg457Gln CREBF variant promotes more lipid storage while using less energy than wild-type CREBF.

In addition to having a role in cellular energy storage and use, the Drosophila melanogaster CREBF ortholog REPTOR has recently been implicated in both cellular and organismal adaptation to nutritional stress by mediating the downstream transcriptional response to the cellular energy sensor TORC1 (refs. 26,27). In support of this hypothesis, expression of CREBF orthologs is highly induced by starvation in all tissues of Drosophila as well as in human lymphoblasts. Moreover, REPTOR-knockout flies and Crebf-knockout mice have lower total energy storage and body weight, respectively. Similarly, we found that nutrient starvation of 3T3-L1 preadipocytes rapidly increased Crebf mRNA levels, which peaked by 4 h at levels 13-fold higher than those seen at 0 h ($P = 1.1 \times 10^{-16}$) and remained elevated by 5-fold at 24 h after the start of starvation ($P = 4.1 \times 10^{-14}$) (Fig. 3a). Treatment with rapamycin, a TORC1 inhibitor, also rapidly increased Crebf mRNA levels, but did so to a lesser extent than starvation (Fig. 3b), indicating that additional TORC1-independent signals converge on Crebf. Furthermore, overexpression of wild-type and...
Table 2 Association of rs373863828 with untransformed adiposity, metabolic, and lipid traits in the discovery sample

<table>
<thead>
<tr>
<th>Quantitative trait</th>
<th>n</th>
<th>β (s.e.)</th>
<th>P</th>
<th>Covariatesa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adiposity traits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>3,066</td>
<td>1.356 (0.183)</td>
<td>1.12 × 10⁻¹³</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>Body fat (%)</td>
<td>2,893</td>
<td>2.199 (0.345)</td>
<td>1.78 × 10⁻¹⁰</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>Abdominal circumference (cm)</td>
<td>3,057</td>
<td>2.842 (0.404)</td>
<td>2.05 × 10⁻¹²</td>
<td>A, A², S, A × S, A² × S</td>
</tr>
<tr>
<td>Hip circumference (cm)</td>
<td>3,058</td>
<td>2.361 (0.332)</td>
<td>1.19 × 10⁻¹²</td>
<td>A, A², S, A² × S</td>
</tr>
<tr>
<td>Abdominal–hip ratio</td>
<td>3,056</td>
<td>0.005 (0.002)</td>
<td>2.23 × 10⁻³</td>
<td>A, A², S, A × S, A² × S</td>
</tr>
<tr>
<td>Metabolic traits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting glucose (mg/dl)b</td>
<td>2,393</td>
<td>−1.652 (0.423)</td>
<td>9.52 × 10⁻⁵</td>
<td>A, A², S</td>
</tr>
<tr>
<td>Fasting insulin (µU/ml)b</td>
<td>2,392</td>
<td>1.342 (0.449)</td>
<td>2.83 × 10⁻³</td>
<td>A, S, A × S</td>
</tr>
<tr>
<td>HOMA-IR</td>
<td>2,392</td>
<td>0.241 (0.114)</td>
<td>0.035</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>Adiponectin (µg/ml)</td>
<td>2,858</td>
<td>−0.228 (0.083)</td>
<td>0.006</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>Leptin in men (ng/ml)c</td>
<td>1,151</td>
<td>0.719 (0.326)</td>
<td>0.027</td>
<td>A</td>
</tr>
<tr>
<td>Leptin in women (ng/ml)c</td>
<td>1,707</td>
<td>1.888 (0.525)</td>
<td>3.25 × 10⁻⁴</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>Metabolic traits adjusted for BMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting glucose (mg/dl)b</td>
<td>2,383</td>
<td>−2.248 (0.417)</td>
<td>6.89 × 10⁻⁸</td>
<td>A, A², S, B</td>
</tr>
<tr>
<td>Fasting insulin (µU/ml)b</td>
<td>2,382</td>
<td>0.225 (0.420)</td>
<td>0.592</td>
<td>A, A², S, B, A × S, A² × S</td>
</tr>
<tr>
<td>HOMA-IR</td>
<td>2,382</td>
<td>−0.034 (0.107)</td>
<td>0.754</td>
<td>A, B</td>
</tr>
<tr>
<td>Adiponectin (µg/ml)</td>
<td>2,844</td>
<td>−0.066 (0.080)</td>
<td>0.412</td>
<td>A, A², S, B, A × S</td>
</tr>
<tr>
<td>Leptin in men (ng/ml)c</td>
<td>1,143</td>
<td>−0.262 (0.210)</td>
<td>0.213</td>
<td>A, A², B</td>
</tr>
<tr>
<td>Leptin in women (ng/ml)c</td>
<td>1,701</td>
<td>−0.516 (0.366)</td>
<td>0.159</td>
<td>A, A², B</td>
</tr>
<tr>
<td>Serum lipid levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total cholesterol (mg/dl)</td>
<td>2,858</td>
<td>−3.203 (1.029)</td>
<td>1.84 × 10⁻³</td>
<td>A, A², S, A × S, A² × S</td>
</tr>
<tr>
<td>Triglycerides (mg/dl)</td>
<td>2,858</td>
<td>0.349 (2.769)</td>
<td>0.900</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>HDL cholesterol (mg/dl)</td>
<td>2,858</td>
<td>−0.322 (0.321)</td>
<td>0.317</td>
<td>A, A², S</td>
</tr>
<tr>
<td>LDL cholesterol (mg/dl)</td>
<td>2,851</td>
<td>−2.347 (0.945)</td>
<td>0.013</td>
<td>A, A², S, A² × S</td>
</tr>
<tr>
<td>Dichotomous traits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity (&gt;32 kg/m²)</td>
<td>3,066</td>
<td>1.305 (1.159–1.470)</td>
<td>1.12 × 10⁻⁶</td>
<td>A, A², S, A × S</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2,876</td>
<td>0.637 (0.536–0.758)</td>
<td>3.86 × 10⁻⁷</td>
<td>A</td>
</tr>
<tr>
<td>Diabetes adjusted for BMI</td>
<td>2,861</td>
<td>0.586 (0.489–0.702)</td>
<td>6.68 × 10⁻⁹</td>
<td>A, B</td>
</tr>
<tr>
<td>Hypertension</td>
<td>3,041</td>
<td>1.014 (0.898–1.145)</td>
<td>0.818</td>
<td>A, S</td>
</tr>
</tbody>
</table>

Boldface represents a P value < 2.17 × 10⁻⁵, s.e., standard error; OR, odds ratio; 95% CI, 95% confidence interval. A, age; A², age²; S, sex; A × S, age × sex interaction; A² × S = age² × sex interaction. Analysis was conducted only in non-dietics. Leptin was not analyzed in men and women together because the distributions were very different for the sexes.

Arg457Gln human CREBRF equivalently reduced the cell death rate to approximately one-third of that in controls within the first 6 h of nutrient starvation in 3T3-L1 preadipocytes (P = 5 × 10⁻⁸ and P = 4 × 10⁻⁸, respectively; Fig. 3c,d). These data indicate that CREBRF is a starvation-responsive factor and that wild-type and Arg457Gln CREBRF when overexpressed confer similar protection against cellular nutritional stress.

Complementing the functional evidence of ‘thriftness’, we identified evidence of positive selection at the missense variant in Samoan genomes. The core haplotype carrying the derived BMI-increasing allele exhibited long-range LD (corresponding to the single thick branch in Fig. 4b versus Fig. 4a) and had elevated extended haplotype homozygosity (EHH) relative to haplotypes carrying the ancestral allele (Fig. 4c). Haplotypes carrying the derived allele were longer than haplotypes carrying the ancestral allele (Fig. 4d). Evidence of positive selection was provided by an integrated haplotype score (iHS) of 2.94 (P = 0.003) and a number of segregation sites by length (ns²) score of 2.63 (P = 0.008) (Supplementary Fig. 8).

In 1962, James Neel posited the existence of a thrifty gene that provides a metabolic advantage in times of famine but promotes metabolic disease in times of nutritional excess. By carrying out a genome-wide association analysis of BMI in Samoans, we discovered and replicated a strong association with a missense variant in CREBF that has a much larger effect size than any other known common risk-associated variant for BMI. Functional evidence from an adipocyte model further demonstrated that CREBF with this missense variant promotes cellular energy conservation by increasing fat storage and decreasing energy use in comparison to the wild-type protein.
The potential importance of this variant in organisinal energy homeostasis is further supported by the ‘lean’ phenotype of mice\(^3\) and flies\(^26\) lacking the ortholog for this gene. These data, in combination with evidence of positive selection, support a thrifty variant hypothesis for human obesity and underscore the value of examining unique populations to identify new genetic contributions to complex traits.

However, many questions remain unanswered. More detailed studies in animal models and humans are required to define the systemic and tissue-specific (particularly central) contributions of the missense variant to overall energy balance. Such studies would also help confirm and clarify the mechanism by which this missense variant might protect against obesity-associated metabolic disease, which perhaps involves preferential promotion of more metabolically ‘safe’ or efficient energy storage and use. Studies that consider potential modifying and mediating environmental influences of this variant as well as gene–gene interactions might illuminate additional new factors contributing to these complex traits. Finally, additional anthropological genetic studies might determine the evolutionary origin of this variant or the potential role of drift in determining its frequency. Such research is urgently needed to inform decisions about how to use knowledge of this obesity risk variant to benefit Samoans at both individual and population health levels and to determine how this discovery might contribute to the understanding and treatment of more common obesity in general.


**METHODS.** Methods and any associated references are available in the online version of the paper.

**Accession codes.** The discovery data set is available from the database of Genotypes and Phenotypes (dbGaP) under accession phs000914.v1.p1.

**Note:** Any Supplementary Information and Source Data files are available in the online version of the paper.

**ACKNOWLEDGMENTS.** The authors would like to thank the Samoan participants of the study, and local village authorities and the many Samoan and other field workers over the years. We acknowledge the Samoan Ministry of Health and the Samoan Bureau of Statistics, and the American Samoan Department of Health for their support of this research. We also acknowledge S.S. Shiva and C.G. Corey at the University of Pittsburgh Center for Metabolism and Mitochondrial Biology for assistance with cellular bioenergetic profiling. This work was funded by NIH grants R01-HL093093 (S.T.M.), R01-AG09375 (S.T.M.), R01-HL52611 (I. Kamboh), R01-DK59642 (S.T.M.), P30-ES006906 (S.M. Ho), R01-DK53406. (R.D.), R01-HL090648 (Z.L.), and R01-DK090166 (E.E.K.) and by Brown University student research funds. Genotyping was performed in the Core Genotyping Laboratory at the University of Cincinnati, funded by NIH grant.
AUTHOR CONTRIBUTIONS

R.L.M. performed the genotype quality control and association analyses, with guidance from D.E.W. and assistance from O.D.B. and J.L. G.S. and R.D. facilitated fieldwork in Samoa and American Samoa. T.N. contributed to the discussion of the public health implications of the findings. All authors contributed to this work, discussed the results, and critically reviewed and revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

ONLINE METHODS

Participants. The participants in this study are derived from the populations of the Independent State of Samoa and the US territory of American Samoa. We used two samples in this study: a discovery sample of 3,072 phenotyped and genotyped Samoans and a replication sample of 2,103 phenotyped and genotyped Polynesians. Extensive quality control was conducted on the basis of a pipeline developed by Laurie et al. Genotyping was performed using Genome-Wide Human SNP 6.0 arrays (Affymetrix). Additional details for sample genotyping can be found in the Supplementary Note.

In the original GWAS study design, our goal of a discovery sample size of 2,500 (which we exceeded) was chosen so as to have high power to detect risk-associated SNPs with realistic effect sizes. Power was estimated as follows: we used Quanto34,35 to estimate the power to detect the rs9930506 SNP in FTO, which in the Sardinia study36 explained 1.34% of variance in BMI. If we assume that this SNP has the same allele frequencies and that BMI has the same overall mean values and standard deviation as in Scuteri et al.37, then at a significance level of \(1 \times 10^{-7}\), power is \(>80\%\) when the risk-associated SNP explains at least 1.1% of the variance (and power is \(90\%\) when the SNP explains 1.3% of the variance). If we instead test at a threshold of \(1 \times 10^{-7}\), power is \(>80\%\) if the SNP explains at least 1.5% of the variance.

Anthropometric and biochemical measurements. Height, weight, and BMI were measured as previously described37,38. Polynesian cutoffs were used to classify adults as normal weight, overweight, or obese on the basis of BMI of \(<26\text{ kg/m}^2\), \(26–32\text{ kg/m}^2\), and \(>32\text{ kg/m}^2\), respectively39. Obesity in children was categorized from BMI using the international age- and sex-specific classifications developed by Cole et al.40.

In the discovery sample, abdominal (at the level of the umbilicus) and hip circumferences were measured in duplicate, and the measures were averaged (Supplementary Table 1). Bioelectrical impedance measures of resistance and reactance (RJL BIA-101Q device, RJL Systems) were used to estimate percent body fat on the basis of Polynesian-specific equations41,42. Serum separated from whole-blood samples, collected after a 10-h overnight fast, was assayed for cholesterol (total, HDL, and LDL), triglycerides, glucose, and insulin. The assay techniques for these metabolic markers have been described previously41. Individuals were classified as having type 2 diabetes on the basis of fasting serum glucose levels \(\geq 126\text{ mg/dl}\) or the current use of diabetes medication43. Hypertensives either had systolic blood pressure \(\geq 140\text{ mm Hg}\) or diastolic blood pressure \(\geq 90\text{ mm Hg}\), or were currently taking hypertension medication. Additionally, serum levels of leptin and adiponectin were obtained by using commercially available radioimmunoassay kits (EMD Millipore). HOMA-IR was calculated as glucose (mg/dl) \(\times\) insulin (\(\mu\text{U/ml}\))/405, as recommended42.

Genotyping. Genotyping of the discovery sample was performed using Genome-Wide Human SNP 6.0 arrays (Affymetrix). Extensive quality control was conducted on the basis of a pipeline developed by Laurie et al.43. Additional details for sample genotyping and genotype quality control can be found in the Supplementary Note.

Statistical analysis. During quality control, significant relatedness was observed among the discovery sample participants, so empirical kinship coefficients were estimated using genotyped markers, in two iterations. In the first iteration, we selected 10,000 independent autosomal markers using PLINK44 and used them to generate empirical kinship coefficients with GenABEL45. Individuals with kinship coefficients less than 0.0625 (corresponding to first cousins) were considered unrelated. A maximal set of 1,891 unrelated individuals was then determined using previously published methods46. In the second iteration, the kinship matrix for all participants was estimated using a new set of 10,000 independent autosomal markers that had been selected using the set of unrelated individuals.

We tested for association between autosomal marker genotypes and BMI residuals while using the empirical kinship matrix to adjust for population substructure and subject relatedness. The tests were conducted using a score test as implemented in the mmscore function in GenABEL47. The statistics for association of X-chromosome genotypes with BMI residuals were calculated in GenABEL without adjusting for the empirical kinship estimates.

Meta-analysis of the adult samples was performed using METAL48 to generate two replication \(P\) values: one for the adult replication samples and one for the adult replication samples and the discovery sample together (Table 1). Additional details of the statistical analyses, including ancestry principal components (Supplementary Fig. 1 and Supplementary Video 1), can be found in the Supplementary Note.

Targeted sequencing. Before undertaking targeted sequencing, we first used SHAPEIT9–33 and IMPUTE2 (refs. 54–56) for imputation in our region of interest centered on rs12513649 with the December 2013 1000 Genomes Phase I integrated variant set release haplotype reference panel. The approach implicated only one strongly associated variant (with a predicted allele frequency of 0.075), but when we genotyped this variant in a pilot sample it turned out to be monomorphic (as it was in the subsequent targeted sequencing experiment). On the basis of this experience, as well as what we would expect given the unique population history of Samoans, we believe that the best way to perform accurate imputation in Samoans is by using a Samoan-specific reference panel. This idea is in agreement with recent recommendations for optimal fine-mapping in populations with unique ancestry not found in a cosmopolitan reference panel57. A panel of 1,295 Samoans from the discovery sample is currently undergoing whole-genome sequencing by the National Heart, Lung, and Blood Institute (NHLBI) TOPMed Consortium. Additional details for targeted sequencing can be found in the Supplementary Note.

Imputation. We prephased the targeted sequencing sample using SHAPEIT9–33 and then imputed into our discovery sample using IMPUTE2 (refs. 54–56). Association testing was carried out using ProBABEL, adjusting for relatedness with the empirical kinship matrix generated by GenABEL. Three variants had nearly equivalent \(P\) values (rs12513649, rs150207780, and rs373863828) because of nearly perfect LD between them \(\left(r^2 = 0.988\right)\), imputation was very good for rs150207780 and rs373863828 (IMPUTE2 info metric = 0.954 for both variants). To determine which of these variants might be the most likely causal candidate, we tested for association in the targeted sequencing region with conditioning on each of these variants as well as the next most significant variant (rs3095870; info metric = 0.957), using ProBABEL and adjusting for relatedness. As expected for variants in such high LD, the signals in the region were eliminated after conditioning (Supplementary Fig. 3).

Bayesian fine mapping. Details can be found in the Supplementary Note.

Confirmatory genotyping. Genotyping was attempted for both rs150207780 and rs373863828 using TaqMan technology in all discovery and replication sample participants. The assay for rs150207780 failed; genotyping was not attempted because this SNP showed no residual association signal in the analyses of the imputed data with conditioning on the missense variant rs373863828 (Supplementary Fig. 3). The replication plates included the 96 samples that had been sequenced in the targeted sequencing experiment. Laboratory personnel were blinded to the sequence-derived genotypes of these 96 samples, as well as to the phenotypes for all the samples. Association analysis was performed using the same regression models and meta-analysis as for the GWAS and replication analyses above. Effect size estimates were calculated using untransformed BMI separately for men and women from the discovery sample with age and age2 as covariates.

Association analyses of additional phenotypes. rs373863828 genotype was examined for association with the additional adiposity-related phenotypes listed in Table 2. Association was assessed in both the discovery sample (Table 2 and Supplementary Table 2a) and a mega-analysis of the adults from the replication sample (Supplementary Table 2b). Although meta-analysis of properly transformed phenotypes generates more accurate \(P\) values (as in Table 1), we chose instead to carry out mega-analyses here because...
we were primarily interested in estimating effect sizes on the natural scale for each trait. Sex-stratified analyses were also conducted in both samples (Supplementary Table 2). Diabetes was excluded from analyses of glucose, insulin, and HOMA-IR. Because the distributions of leptin levels varied greatly for women and men, a combined-sex analysis was not conducted for this trait. Residuals for quantitative traits were generated using linear regression. Age, age², sex, and the interactions between age and sex and between age² and sex were initially included in sex-combined models. For glucose, insulin, HOMA-IR, adiponectin, leptin, and diabetes status, a second set of models was used that included log-transformed BMI as a covariate. Sex and age × sex interactions were not included in the sex-stratified models. In the replication meta-analysis models, polity (Samoa or American Samoa) and cohort (1990s or 2000s) were initially included in the models as well. Stepwise regression was used to reduce the number of covariates for each trait separately. For quantitative traits, residuals were tested for association using the mscore function of GenABEL, adjusted for the empirical kinship matrix as above. Dichotomous traits were analyzed using the palogist function of ProABELE, while adjusting for covariates and empirical kinship. A Bonferroni-corrected $P$-value threshold of $2.17 \times 10^{-3}$ was used to assess significance; this threshold is conservative, as it adjusts for 23 tests even though some traits are correlated with each other. To assess a possible survivor effect as the cause of the association between the BMI-increasing allele and decreased fasting glucose levels and risk of diabetes, we conducted linear regression of age by genotype. In the discovery sample, in regard to the association of rs373863828 with BMI, fasting glucose, fasting insulin, obesity risk, and diabetes risk, addition of the first ten 'local' principal components from the discovery sample, in regard to the association of rs373863828 with BMI, fasting glucose, fasting insulin, obesity risk, and diabetes risk, addition of the first ten 'local' principal components from Supplementary Figure 1b into the statistical models had a negligible effect on the effect estimates and statistical significance (data not shown).

**Expression of CREBF in human and mouse tissues.** For human gene expression analysis, a Human Normal cDNA Array was obtained from Origene Technologies (HMRT103 and HBRT101). The human standard curve was prepared from Control Human Total RNA (Thermo Fisher Scientific, 4307281). For mouse gene expression analysis, mouse tissues were collected from 8–10 a.m. from littermate–matched, *ad libitum*–fed male C56BL/6J mice at 10 weeks of age (n = 6 mice/group). The mouse standard curve was prepared from pooled kidney RNA from the above mice. mRNA was prepared using the RNeasy Lipid Tissue Mini kit with on-column DNase treatment (Qiagen) followed by reverse transcription to cDNA using qScript cDNA Supermix (Quanta Biosciences). Gene expression was determined by qPCR (Quanta PerfeCta SYBR Green FastMix or PerfeCta qPCR FastMix) using an Eppendorf Realplex System. Human CREBF was amplified using species-specific primers (Supplementary Table 3). Mouse CREBF was amplified using a species-specific primer–probe set (Thermo Fisher Scientific, Mm00661538_m1). Mouse CREBF expression was normalized to species-specific peptidylprolyl isomerase A or cyclophilin A as the endogenous control gene (Thermo Fisher Scientific, 4333767T and Mm02342430_g1 for human and mouse, respectively). Mouse data are expressed as means plus s.e.m. Data are relative expression values, and so randomization, blinding, and statistical comparisons were not indicated. Gene expression analysis was performed in accordance with Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines. Mouse experiments were approved by the University of Pittsburgh Institutional Animal Care and Use Committee and conducted in conformity with the Public Health Service Policy for Care and Use of Laboratory Animals. Human samples from Origene Technologies conform to federal policies for the protection of human subjects (45 CDR 46) and are HIPAA compliant. Additional information and documentation can be obtained by contacting the company.

**Plasmid construction and mutagenesis.** Expression plasmids with ORFs for eGFP and human CREBF (NM_153607.2) were obtained from GeneCopoeia (EX-EFGE-M10, EX-E3374-M10). The backbone vector was pPERCEIVER-M10, which has a cytomegalovirus promoter and encodes a C-terminal Myc-(His)₆ tag. A rare missense variant, c.1447A>G, p.Thr483Ala (rs17854147), affecting a conserved residue was present in the CREBF ORF. To avoid using this potentially function-altering variant, we converted CREBF to the wild-type sequence and introduced the BMI risk-associated mutation c.1370G>A, p.Arg457Gln (rs373863828), using PCR mutagenesis. The segments obtained by PCR in each plasmid were verified by sequencing before large-scale plasmid purification for transfection.

**Cell culture and transfection, adipocyte differentiation, Oil Red O plate assays, microscopy, triglyceride assays, and quantitative RT–PCR.** These methods are described in detail in the Supplementary Note.

**Bioenergetic profiling.** OCR, a measure of mitochondrial respiration, and ECAR, a measure of glycolysis, were determined using an XF96 extracellular flux analyser (Seahorse Bioscience). Transfected 3T3-L1 cells were seeded in a 96-well XF96 cell culture microplate (Seahorse Bioscience) at a density of 7,000 cells per well in 200 µl of DMEM (4.5 g/l glucose) supplemented with 10% FBS (Sigma) 36 h before measurement. Six replicates per cell type were included in the experiments, and four wells were chosen evenly in the plate to correct for temperature variation. On the day of the assay, the growth medium was exchanged for assay medium (unbuffered DMEM with 4.5 g/l glucose). Oligomycin at a final concentration of 2.0 µM, FCCP (carbonyl cyanide-p-trifluoromethoxyphenylhydrazone) at 1.0 µM, 2-deoxyglucose at 100 mM, and rotenone at 15.0 µM were sequentially injected into each well in accordance with the manufacturer’s protocol. Basal mitochondrial respiration, maximal respiration, ATP production, and basal glycolysis were determined according to the manufacturer’s instructions. At the conclusion of the assay, cells in the analysis plate were lysed using Cellytic M (Sigma). Protein concentration was measured using the Bradford assay and used to normalize the bioenergetic profile data.

**Starvation and rapamycin treatment.** 3T3-L1 predipocytes were subjected to starvation for 0, 2, 4, 12, and 24 h by culturing cells in Hank’s balanced salt solution (HBSS). To investigate the response to refeeding starving cells, a set of cells undergoing 12 h of starvation was fed with fresh growth medium for an additional 12 h (Fig. 3a). For rapamycin stimulation, predipocytes were treated with 20 ng/ml rapamycin (Sigma), for 2, 4, 12, and 24 h. A set of cells kept in rapamycin for 12 h was cultured in fresh growth medium for the following 12 h (Fig. 3b). To quantify cell survival, 3T3-L1 cells and transfected cells were seeded in six-well plates at 86,000 cells per well. Two days later, the cells were starved in HBSS. At 0, 2, 4, 6, 12, and 24 h, the cells were collected and 100 µl of the cell suspension samples was added to an equal volume of trypsin blue (Life Technologies). The mixture was loaded into an automated cell counter (Celsius Mini, Nexcelom Bioscience), and viable cell numbers were measured. Cell death rates were calculated by subtracting the number of viable cells at 6 h from cell numbers at 0 h and dividing the result by the cell numbers at 6 h.

**Cell studies statistical analysis.** For the cell studies, adequate sample sizes were determined on the basis of publications using similar methods and pilot experiments. No blinding was used. Each experiment was performed twice with similar results unless otherwise stated in the corresponding figure legend. The data were initially evaluated by one-way ANOVA implemented in SPSS (IBM). The homogeneity of variances was examined using Levene’s test. Two-sided Bonferroni and Games–Howell *post-hoc* tests were used to compare data with equal and unequal variance, respectively. Alternatively, pairwise two-sided *t* tests for unequal variance were used. *P < 0.05* was considered to be statistically significant. SPSS analyses were verified using the same tests as implemented in R (ref. 60).

**Selection analyses.** On the basis of the genome-wide Affymetrix 6.0 SNP genotyping data, we used Primos61,62 to select 626 individuals from the discovery sample using a kinship threshold (0.039) halfway between the values expected for first and second cousins, so that first cousins and more closely related relatives were excluded. These ‘unrelated’ individuals were then haplotyped using SHAPEIT49–53 and were annotated with ancestral allele information using the selectionTools pipeline.64 Haploype bifuication diagrams and EHH plots were drawn using the rehh R package.64 The haplotype bifuication diagram visualizes the breakdown of LD as one moves away from the core allele at the focal SNP; each branch reflects the creation of new haplotypes, and the thickness of the line reflects the number of samples with the haplotype. EHH represents the
probability that two randomly chosen chromosomes are identical by descent from the focal SNP to the current position of interest. Selection at the core allele is expected to result in EHH values close to 1 in an extended region centered on the focal SNP. To measure the deviation, we used selscan to compute the iHS, which is defined as the log of the ratio of the integrated EHH for the derived allele over the integrated EHH for the ancestral allele. These values are then normalized in frequency bins across the whole genome (we used 25 bins). Note that selscan’s definition of iHS differs from earlier definitions where the ancestral allele was in the numerator of the ratio. In our case, a standard normal. We also used selscan to compute nSL scores (the number of segregating sites by length). The nSL is similar to the iHS, but instead of integrating over genetic distance the nSL uses the number of segregating sites as a measure of ‘distance.’ Thus, the nSL is more robust to demographic assumptions than the iHS, as it does not depend on a genetic map. As with the iHS, we normalized the nSL scores in 25 frequency bins across the whole genome and computed approximate two-sided P values assuming a standard normal distribution. The selscan program was run using its assumed default values. As we were focused on testing whether there is positive selection at the missense variant, we did not adjust the P values for multiple testing.

1. Why did authors choose to perform meta-analysis across different parent studies instead of combining data at the level of individual participant and analyzing the entire sample altogether? Discuss both practical and scientific reasons for this approach.

2. The authors performed three gene-based tests (T1, T0.1, and SKAT) for rare variants. What is the rationale for this strategy? Is there any benefit over single-SNP tests?

3. In the “Statistical Significance” sub-section under the “Statistical Analysis” heading, the authors describe the Bonferroni correction they applied in this paper. Would you characterize this approach as conservative or liberal? Would you expect that some of the results they call significant are actually false positives? Would you expect that some true positive signals were missed?

4. Table 1 categories variants on the Exome Array into five types: (1) nonsynonymous, (2) nonsense/splice, (3) synonymous, (4) intronic/intergenic, and (5) other. Why might authors have decided to group nonsense and splice site variants together as a single category? Do you agree with this decision? Why or why not? Same questions for intronic/intergenic variants.

5. Analyses were performed separately in participants of European ancestry (EA) vs. African ancestry (AA)? What is one rationale for this decision?

6. Minor allele frequency thresholds for SNP-based and gene-based tests were different in EA and AA. Explain the authors’ reason for this. Do you agree? Why or why not?

7. Authors compare their results to the previous discovery of low frequency variants in PCSK9, which lower LDL-C and protect against CHD. Specifically, they query whether the PCSK9 example represents a paradigm or a fairly unique scenario. How do authors interpret their results in light of the PCSK9 discovery? Do you agree with the authors’ interpretation?

8. Authors did not show results for individual samples (they only showed meta-analysis results). What type of plot would have been useful for showing results of the four novel variants across each of the samples? What could be gained from this type of plot?

9. Why is exomic coverage higher in whites compared to blacks?
10. In general, populations of African ancestry are expected to have more low frequency variants than populations of European ancestry. Is Table 1 inconsistent with this expectation? Explain why or why not.

11. The LDL-C phenotype used in this paper was fairly crude (i.e., derived from the Friedewald equation, and adjusted for statin use). How might this affect the statistical analysis, if at all?
Association of Low-Frequency and Rare Coding-Sequence Variants with Blood Lipids and Coronary Heart Disease in 56,000 Whites and Blacks


1Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA; 2Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, USA; 3Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; 4Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA; 5Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; 6School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA; 7Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA; 8Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; 9Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; 10Cardiovascular Division, Department of Medicine, University of Washington, Seattle, St. Louis, MO 63110, USA; 11Division of Statistical Genomics, Department of Genomics, Washington University School of Medicine, St. Louis, MO 63110, USA; 12Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; 13Department of Biostatistics, Biocomputing, and Systems Biology, The University of Texas Graduate School of Biomedical Sciences at Houston, Houston, TX 77030, USA; 14Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; 15Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam 3015CN, the Netherlands; 16Icelandic Heart Association, Kopavogur 201, Iceland; 17Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK; 18Department of Psychology, University of Edinburgh, Edinburgh EH4 2XU, UK; 19Center for Public Health Genomics and Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA; 20Wake Forest School of Medicine, Winston-Salem, NC 27106, USA; 21Faculty of Medicine, University of Iceland, Reykjavik 101, Iceland; 22Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA; 23National Heart, Lung, and Blood Institute (NHLBI) Framingham Heart Study, Framingham, MA 01702, USA; 24Division of Endocrinology, Diabetes and Metabolism, Department of Internal Medicine, The Ohio State University, Columbus, OH 43210, USA; 25Department of Epidemiology, College of Public Health, University of Iowa, Iowa City, IA 52242, USA; 26Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA; 27Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; 28Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN 55445, USA; 29Department of Epidemiology, Erasmus University Medical Center, Rotterdam 3015CN, the Netherlands; 30National Institute on Aging, NIH, Bethesda, MD 20892, USA; 31Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; 32School of Medicine, Yale University, New Haven, CT 06510, USA; 33The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 34The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 35Department of Clinical Sciences in Malmö, Lund University, Clinical Research Center, Malmö 20502, Sweden; 36Department of Clinical Sciences, Diabetes and Endocrinology, Lund University, University Hospital Malmö, Malmö 20502, Sweden; 37The Mindich Child Health and Development Institute, The Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 38Department of Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano, Milano 20123, Italy; 39Associazione per lo Studio della Trombosi in Cardiologia (ASTC), Pavia 27100, Italy; 40Divisione di Cardiologia, Ospedale Niguarda, Milano 20162, Italy; 41Department of Cardiovascular Medicine, The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; 42Department of Medicine, Verona School of Medicine, Verona 37134, Italy; 43Department of Obstetrics and Gynecology, Division of Urogynecology, Duke University, Durham, NC 27710, USA; 44Division of Cardiology, Department of Medicine, Duke University School of Medicine, Durham, NC 27710, USA; 45Division of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; 46Atherogenomics Laboratory, Division of Cardiology, University of Ottawa Heart Institute, Ottawa,
Low-frequency coding DNA sequence variants in the proprotein convertase subtilisin/kexin type 9 gene (PCSK9) lower plasma low-density lipoprotein cholesterol (LDL-C), protect against risk of coronary heart disease (CHD), and have prompted the development of a new class of therapeutics. It is uncertain whether the PCSK9 example represents a paradigm or an isolated exception. We used the “Exome Array” to genotype ~200,000 low-frequency and rare coding sequence variants across the genome in 56,538 individuals (42,208 European ancestry [EA] and 14,330 African ancestry [AA]) and tested these variants for association with LDL-C, high-density lipoprotein cholesterol (HDL-C), and triglycerides. Although we did not identify new genes associated with LDL-C, we did identify four low-frequency (frequencies between 0.1% and 2%) variants (ANGPTL8 rs145464906 [c.361C>T]; p.Gln121*, PAFAH1B2 rs186808413 [c.482C>T]; p.Ser161Leu], COL18A1 rs114139997 [c.331G>A]; p.Gly111Arg], and PCSK7 rs142953140 [c.1511G>A]; p.Arg504His]) with large effects on HDL-C and/or triglycerides. None of these four variants was associated with risk for CHD, suggesting that examples of low-frequency coding variants with robust effects on both lipids and CHD will be limited.

Introduction

Recently, a compelling new therapeutic target for lowering low-density lipoprotein cholesterol (LDL-C) emerged from human genetics: the proprotein convertase subtilisin/ kexin type 9 gene (PCSK9 [MIM 607786]).\(^1\)\(^2\) PCSK9 protein-coding sequence variants that are low in frequency (defined here as allele frequencies from 0.1% to 5%) have been associated with lower plasma LDL-C\(^2\) and reduced risk for coronary heart disease (CHD).\(^4\) With the identification of low-frequency variants that protected against CHD, many pharmaceutical companies have established drug development programs targeting PCSK9. These observations have raised the question of whether the PCSK9 example is a paradigm for complex diseases like CHD or an exception.\(^7\)

Low-frequency DNA sequence variants and rare alleles (defined here as <0.1% allele frequency) are poorly characterized by earlier generations of genome-wide association study (GWAS) genotyping arrays.\(^6\) Sequencing across the exome or genome can directly assay low-frequency and rare variants but such an approach is currently too costly to study tens of thousands of individuals. One proposed method for testing low-frequency and rare DNA variation is to first sequence the exome to discover variation and, subsequently, to genotype the discovered variants in a larger number of individuals from the same or similar populations to test for association with phenotype. Based on this principle, the Illumina HumanExome genotyping array (“the Exome Array” or “Exome Chip”) was designed based on coding sequence variants discovered from sequencing the exomes of ~12,000 individuals.

Here, we set out to address two questions: (1) are there novel low-frequency nonsynonymous and splice site variants associated with lipid levels in the population? and (2) if so, will these coding sequence variants also associate with risk of clinical CHD? To address these questions, we first genotyped and analyzed the Exome Array in 42,208 European ancestry (EA) individuals and 14,330 African ancestry (AA) individuals from 13 cohorts with blood levels of fasting low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and fasting triglycerides (TG). We subsequently tested validated lipid variants for association with CHD. We discovered four variants associated with HDL-C and/or TG, but these variants did not associate with CHD.

Subjects and Methods

Estimation of Proportion of Variation Captured by Exome Array

We identified EA (n = 3,173) and AA (n = 2,408) participants from the Atherosclerosis Risk in Communities (ARIC) Study with available exome sequence who were not among the ~12,000 individuals utilized to design the Exome Array. Separately by ancestry, we identified all variants with a minor allele frequency (MAF) >0.1% from exome sequencing of these independent ARIC individuals. These variants were compared with the variants available on the Illumina HumanExome v1.0 array.

Study Participants

Thirteen studies genotyped the Exome Array on a total of 56,538 participants (Table S1 available online). 42,208 individuals were of European ancestry from ARIC, AGES, CHS, FHS, RS, MESA, WHI, CARDIA, Health ABC, FamiHS, LBC1936, and Korcula. A total of 14,330 subjects were of African ancestry from ARIC, CHS, MESA, Health ABC, FamiHS, WHI, and CARDIA. All participants provided informed consent and each study was approved by their governing ethics committee.

ON K1Y 4W7, Canada; \(^{49}\) Merck Sharp & Dohme Corp., Rahway, NJ 07065, USA; \(^{50}\) Divisione di Cardiologia, Azienda Ospedaliero-Universitaria di Parma, Parma 43100, Italy; \(^{51}\) Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55445, USA; \(^{52}\) Jackson State University, Jackson, MS 39217, USA; \(^{53}\) Togoloo College, Togoloo, MS 39174, USA; \(^{54}\) Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; \(^{55}\) Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; \(^{56}\) Alzheimer Scotland Dementia Research Centre, University of Edinburgh, Edinburgh EH9 9JZ, UK; \(^{57}\) Centre for Population Health Sciences, The University of Edinburgh Medical School, Edinburgh EH8 9AG, UK; \(^{58}\) Croatian Centre for Global Health, Faculty of Medicine, University of Split, Split 21000, Croatia; \(^{59}\) Department of Cardiovascular Epidemiology and Population Genetics, National Center for Cardiovascular Investigation, Madrid 28049, Spain; \(^{60}\) IMDEA-Alimentacion, Madrid 28049, Spain; \(^{61}\) Nutrition and Genomics Laboratory, Jean Mayer-USDA Human Nutrition Research Center on Aging at Tufts University, Medford, MA 02155, USA; \(^{62}\) Division of Intramural Research, NHLBI, NIH, Bethesda, MD 20892, USA; \(^{63}\) Department of Public Health, Faculty of Medicine, University of Split, Split 21000, Croatia; \(^{64}\) Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA; \(^{65}\) Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101, USA; \(^{66}\) Department of Health Services, University of Washington, Seattle, WA 98195, USA; \(^{67}\) Cardiology Division, Massachusetts General Hospital, Boston, MA 02114, USA; \(^{68}\) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; \(^{69}\) These authors contributed equally to this work

*Correspondence: skathiresan@partners.org (S.K.), adrienne@bu.edu (L.A.C.)

http://dx.doi.org/10.1016/j.ajhg.2014.01.009. ©2014 by The American Society of Human Genetics. All rights reserved.
Genotyping and Quality Control

All study participants were genotyped on the HumanExome BeadChip v.1.0 (Illumina). ARIC, AGES, CHS, FHS, RS, MESA, CARDIA, Health ABC, FamHS, and JHS were jointly called in the University of Texas Health Science Center at Houston. LBC1936 and Korcula were called in GenomeStudio (Illumina) via the CHARGE Consortium joint calling cluster file. WHI used zCalls.9 Variant quality control (QC) was performed at the University of Texas Health Science Center and by the individual cohorts. QC involved checking concordance to GWAS data and excluding those individuals missing >5% genotypes, population clustering outliers, individuals with high inbreeding coefficients or heterozygote rates, individuals with gender mismatches, one individual from duplicate pairs, and individuals with an unexpectedly high proportion of identity-by-descent sharing, with consideration for family studies, based on high-quality variants. All contributing studies used an additive coding of variants to the minor allele observed in the jointly called data set.

Phenotypes

Association analyses were performed for three blood lipid traits in mg/dl: LDL-C, HDL-C, and TG. Fasting lipids values were used from the earliest available exam in each study. Studies with only nonfasting lipid levels contributed only to the HDL-C analysis. To account for lipid-lowering therapy in selected participants, we sought to estimate the untreated lipid value in each participant. Such an approach has been demonstrated to perform well in accounting for treatment effects in studies of quantitative traits.10 Statins are the most widely used treatment to lower plasma lipids and a statin at average dose reduces total cholesterol (TC) by 20% and LDL-C by 30%.11 Statins became routinely used after the publication of the seminal 4S randomized control trial in 1994.12

If the sample was collected after 1994, we accounted for lipid-lowering medication in the following manner: the treated TC value was divided by 0.8. No adjustment was done on data collected before 1994 unless specific information on statin use was available. LDL-C was calculated via the Friedewald equation (LDL-C = TC − HDL-C − (TG/5)) for those with TC < 400 mg/dl. If TG > 400 mg/dl, calculated LDL-C was set to missing. If TC was modified as described above for medication use, the modified total cholesterol was used to calculate LDL-C. If only measured LDL-C was available in a study, we accounted for lipid-lowering medication in the following manner: the treated LDL-C value was divided by 0.7. No adjustments for medication use were made for HDL-C and TG. TG was natural log transformed for analysis.

Statistical Analyses

Single-variant, conditional, and gene-based analyses were conducted with the seqMeta package in R. In brief, seqMeta implements efficient single-variant and gene-based meta-analysis. Each study computes association-specific statistics for each variant and trait assuming an additive model, and genotypic covariance matrices within predefined gene regions. Score statistics and genotypic covariance matrices are combined across studies to obtain an overall score statistic for each variant and a combined covariance matrix across studies. The combined score statistic is used for the single-variant results. For each gene, the score statistics for the variants within the gene along with the combined covariance matrix across those variants are used to construct the gene-based tests. An additive effect of each variant is assumed for individuals carrying more than one rare variant. All analyses were performed separately by race. We chose to meta-analyze EA and AA participants separately to minimize population biases and to increase the number of variants at specific MAF thresholds where there could be vastly different frequencies in the two ancestries.

Single-Variant Analysis

Though we defined low-frequency variants as having a MAF between 0.1% and 5%, we restricted single-variant tests to nonsynonymous and splice site variants with a MAF > 0.02% in EA individuals (nvariants = 54,837) and > 0.07% in AA individuals (nvariants = 90,218), corresponding to approximately 20 copies of the minor allele in each ancestry sample. Twenty copies of the minor allele provided stable estimates of the standard errors of the beta estimates and well-calibrated quantile-quantile plots.

Gene-Based Analysis

We constructed gene-based tests from the cohort-combined variant statistics and LD matrices. We only included nonsynonymous and splice site variants when producing the gene-based statistics. We performed three gene-based tests: (1) T1, where all variants with a MAF < 1% were collapsed into gene-based score;15 (2) T0.1, where all variants with a MAF < 0.1% were collapsed into gene-based score; and (3) sequence kernel association test (SKAT),16 at a MAF < 5%. SKAT is more powerful than T1 when there are both protective and deleterious variants with different magnitudes in the same gene whereas T1 and T0.1 are more powerful when the magnitudes and directions of effect for the individual variants in a gene are consistent. We filtered gene-based results with a cumulative MAF < 0.05% in the EA individuals and < 0.14% in the AA individuals for SKAT analyses, corresponding to approximately 40 copies of the minor allele. For the T1 and T0.1 tests, we used a cumulative MAF > 0.04% for the EA individuals and > 0.11%, corresponding to 30 copies of the minor allele across all cohorts. Simulations under the null (i.e., no associations) were performed to determine the thresholds to control type I error for the burden tests (results not shown). The gene-based T1 and T0.1 analyses were limited to genes with at least two variants contributing to the test. For EA, T1 analysis included 12,351 genes with a cumulative MAF > 0.04% and T0.1 with 8,357 genes; for AA participants, T1 analysis included 13,574 genes with a cumulative MAF > 0.11% and T0.1 with 7,579 genes.

All analyses included age, age², sex, and up to the first four principal components of ancestry as covariates. Cohort statistics were adjusted for related individuals, where appropriate. Beta coefficients were checked for consistency across the individual studies contributing to the meta-analysis. Hardy-Weinberg equilibrium p-values were calculated and cluster plots from the joint calling were verified for the reported associations (Figure S1). Fasting glucose at baseline (between 1987 and 1989) in ARIC EA and exam 5 (between 1991 and 1995) in HHS was tested for association with rs145464906 in ANGPTL8 by linear regression adjusting for age, sex, and PCs.

Population Structure

Each study identified components to correct for population structure in their study and used these components as covariates in all association analyses.

Conditional Analysis

The allele dosages (0–2 copies) of the top previously identified GWAS variants in each region were used as additional covariates in the association model.

Statistical Significance

For single-variant association, we set the significance threshold to \( p < 3 \times 10^{-8} \), corresponding to a Bonferroni correction for 1,485,864 tests (3 phenotypes \( \times \) 247,644 variants on the
array \times 2 \text{ ancestries}). \text{ For the gene-based association, we set the significance threshold to} < 2 \times 10^{-7}, \text{ corresponding to a Bonferroni correction for} 316,332 \text{ tests (3 phenotypes} \times 17,574 \text{ genes on the array} \times 3 \text{ tests} \times 2 \text{ ancestries).}

\textbf{Annotation}

We used dbNSFP v.2.0 to annotate the variants.\textsuperscript{17}

\textbf{Clinical CHD Association Analysis}

Individuals from AGES, ARIC, CHS, FHS, MESA, Health ABC, RS1, WHI, MDC-CVA, IPM, ATVB/VHS, Ottawa, Procardis, Duke, and Penn contributed to the CHD analysis in EA. Individuals from ARIC, CHS, MESA, Health ABC, WHI, and IPM contributed to the CHD analysis in AA. All individuals were genotyped on the HumanExome BeadChip. Counts of number of CHD events in carriers and noncarriers of the four reported lipid variants were obtained from each available study. A Cochran-Mantel-Haenszel test was performed to associate the four reported lipid variants with CHD by using the R metafor package across the contributing studies.

\textbf{Power Calculation}

Power for the CHD association was calculated via the case-control for discrete traits option in the Genetic Power Calculator.

\textbf{Mouse Studies Involving PCSK7}

We created an adeno-associated virus 8 (AAV8) vector encoding the human PCSK7 coding region (cDNA) driven by the liver-specific thyroxine-binding globulin (TBG) promoter. An AAV8 vector lacking any transgene was used as a control. Separate groups of 10- to 12-week-old wild-type C57BL/6j male mice (six to seven per group) were injected via the peritoneal route with 1 \times 10^{12} vector genomes/mouse of the relevant vectors. Mice were fed a regular chow diet (Purina LabDiet 5010) throughout the study. Plasma samples after 4 hr of fasting were taken from the mice immediately prior to vector administration, as well as 7 and 14 days after AAV injection for analysis of lipids. Lipid measurements were performed on a Roche Cobas-Mira autoanalyzer (Roche Diagnostic Systems) with Wako Chemicals reagents. Mice were sacrificed at 14 days after AAV administration. All animal procedures were performed according to the regulations of, and with the prior approval of, the University of Pennsylvania Animal Care and Use Committee (IACUC). We repeated this experiment in a separate design comparing plasma lipids at baseline and 14 days after injection of an AAV8 vector encoding the mouse Pcsk7 cDNA versus control in wild-type C57BL/6j male mice (data not shown).

\textbf{Results}

\textbf{Coverage of Exome Array}

We evaluated the coverage of Exome Array in samples independent from those used to design the array. Based on exome sequences from 3,173 EA and 2,408 AA independent participants, we estimate that the Illumina Exome Array captures 78% of nonsynonymous coding and splice site variation with >1:1,000 allele frequency in EA individuals and 71% in AA individuals. Therefore, the array provides good coverage for low-frequency DNA variation and because of its low cost can be assayed in a large number of individuals.

Exome-sequencing variants not captured by the Exome Array may have been dropped during the design of the array. To address this possibility, we compared the variants discovered from exome sequencing to proposed content of the array (before variants failed design). We estimate that 95% of the variation in EA participants (and 85% of the variation in AA) with frequency >1:1,000 is captured by the proposed content of the array.

\textbf{Study Participants and Genotypes}

Clinical characteristics of the 56,538 analyzed study subjects are summarized in Table S1. Of the 247,644 variants that passed quality control criteria, 85% (n = 209,756) are polymorphic in the EA participants and 82% (n = 202,255) are polymorphic in the AA participants. Approximately 90% (n = 188,480) of the polymorphic variants are annotated\textsuperscript{17} as nonsynonymous (nonsense or missense) or splice variants in the EA participants, and 89% (n = 180,908) of the polymorphic variants are annotated as such in the AA participants (Table 1).

\textbf{Single-Variant Association}

We tested each variant individually for association with blood lipid levels separately in each cohort and association summary statistics across cohorts were combined by fixed effects meta-analysis, separately within each ancestry group. There was no evidence of inflation in the association test statistics (Figure S2).

We replicated previously reported associations with common and low-frequency variants (Table 2) in both EA and AA participants. For example, we found that variants in PCSK9 (c.137G>T [p.Arg46Leu], [rs11591147] in EA and c.2037C>A [p.Cys679*], [rs28362286] in AA) were associated with lower LDL-C (−17 mg/dl; p = 3 \times 10^{-59} and −40 mg/dl; p = 2 \times 10^{-55}, respectively).

We discovered four associations of low-frequency variants with either HDL-C and/or TG that met our a priori significance threshold of p < 3 \times 10^{-8} and not previously reported in the literature. Two signals emerged from EA participants and two from AA participants (Table 3). In contrast to HDL-C and TG, we did not discover any new genes where low-frequency or rare DNA sequence variants significantly associated with LDL-C. Below, we describe each of the four new associations.

First, in EA participants, we found an association of blood HDL-C with a 0.1% premature stop codon at the chromosome 19 open reading frame 80 gene (C19orf80, aliases include ANGPTL8, lipasin, and betatrophin; c.361C>T [p.Gln121*]; rs145464906). Carriers of ANGPTL8 rs145464906 had 10 mg/dl higher HDL-C (p = 5 \times 10^{-11}), lower TG (−15%; p = 0.003), and nonsignificantly lower LDL-C (−5.8 mg/dl; p = 0.13) than did noncarriers of this variant (Table 4). ANGPTL8 rs145464906 is seen at a 0.01% frequency in AA participants and we were not able to reliably estimate its effect on lipid measures in this group.
ANGPTL8 rs145464906 is located in a genomic region previously associated with lipid levels in GWASs (rs737337 at the DOCK6 [MIM 614194] locus; c.2136A>G [p.(=)]). An analysis conditioning on rs737337 in DOCK6 revealed rs145464906 to be an independent association (p_{conditional} = 5.5 \times 10^{-11}). Another more common coding variant in ANGPTL8 (8% frequency; rs2278426; c.175C>T [p.Arg59Trp]) has been associated with low LDL-C and low HDL-C. This variant is not present on the exome array, but the correlation between rs2278426 and rs145464906 is low at 0.08 and conditional analyses confirm that rs2278426 and rs145464906 are independent association signals (data not shown).

Angptl8-null mice display lower blood TG and functional studies have suggested that ANGPTL8 may activate ANGPTL3 and thereby regulate lipoprotein metabolism. Of note, based on studies in mice, another proposed role for ANGPTL8 is as a hormone that promotes pancreatic β cell proliferation, expands β cell mass, and improves glucose tolerance. These studies raise the hypothesis that loss of ANGPTL8 function might worsen measures of glucose tolerance. We studied the association of ANGPTL8 rs145464906 with fasting plasma glucose levels among 10,642 EA participants in the ARIC study and 3,112 EA participants from FHS and observed no evidence for association in either study. In ARIC, plasma glucose did not differ between the carriers of the rs145464906 allele (n = 23) and noncarriers (n = 10,620) (beta = -3.3 mg/dl; SE = 5.8 mg/dl; p = 0.57). In FHS, carriers of rs145464906 (n = 13) had a mean fasting glucose of 88.2 mg/dl whereas noncarriers had mean glucose of 99.9 mg/dl (p = 0.13).

A second discovery among EA participants is a significant association between a 1% frequency variant in the Platelet-Activating Factor Acetylhydrolase 1b, Catalytic Subunit 2 gene (PAFAH1B2 [MIM 602508]; c.482C>T [p.Ser161Leu]; rs186808413) and HDL-C. Carriers of rs186808413 had 3 mg/dl higher HDL-C (p = 2.2 \times 10^{-10}), lower TG (p = 2.3 \times 10^{-8}), and marginally lower LDL-C (p = 0.01) than did noncarriers (Table 4). In AA participants, the direction of association between HDL-C and PAFAH1B2 rs186808413 (0.2% frequency in AA participants) was consistent with the EA results, but not significant (1 mg/dl; p = 0.58). PAFAH1B2 is located near the chromosome 11 gene cluster involving APOAI/C3/A4/A5, a region where several genes regulate TG. Conditional analysis adjusting for six previously studied variants in this region (rs964184, rs3135506, rs2266788, rs76353203, rs147210663, rs140621530) showed PAFAH1B2 rs186808413 to be an independent association signal (p_{conditional} = 2.1 \times 10^{-11}). Platelet-activating factor (PAF) is a lipid messenger functioning in many cellular processes. Plasma platelet-activating factor-acetyl hydrolase (PAF-AH) belongs to a subfamily of phospholipases A2 that remove the sn-2 acetyl group and inactive a number of oxidized lipids, but intracellular type I (PAF-AH-ib) is structurally distinct and may form a part of a signal transduction pathway.

Third, among AA participants, we found an association of blood TG with a 2% frequency variant in the Collagen, Type XVIII, Alpha 1 gene (COL18A1 [MIM 120328]; c.331G>A [p.Gly111Arg]; rs114139997), where carriers of rs114139997 had 16% lower TG than did noncarriers (p = 2 \times 10^{-16}). COL18A1 rs114139997 was also marginally associated with higher HDL-C (2 mg/dl; p = 4 \times 10^{-8}) (Table 4). The frequency of COL18A1 rs114139997 is 0.003% in EA participants and we were not able to reliably estimate the effect on lipid measures in this group. COL18A1 is a basement membrane proteoglycan with a

### Table 1. Number of Exome Array Variants Available for Analysis by Variant Type

<table>
<thead>
<tr>
<th>Variant Type</th>
<th>All Polymorphic Sites</th>
<th>MAF &lt; 0.1%</th>
<th>MAF 0.1%–5%</th>
<th>MAF &gt; 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>42,208 European Ancestry Samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>175,444</td>
<td>130,553</td>
<td>32,501</td>
<td>12,390</td>
</tr>
<tr>
<td>Nonsense/splice</td>
<td>13,036</td>
<td>10,694</td>
<td>1,686</td>
<td>656</td>
</tr>
<tr>
<td>Synonymous</td>
<td>5,444</td>
<td>3,710</td>
<td>978</td>
<td>756</td>
</tr>
<tr>
<td>Intronic/Intergenic</td>
<td>14,205</td>
<td>139</td>
<td>1,186</td>
<td>12,880</td>
</tr>
<tr>
<td>Other</td>
<td>1,627</td>
<td>156</td>
<td>199</td>
<td>1,272</td>
</tr>
<tr>
<td>14,330 African Ancestry Samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>169,140</td>
<td>91,083</td>
<td>61,809</td>
<td>16,248</td>
</tr>
<tr>
<td>Nonsense/splice</td>
<td>11,768</td>
<td>7,729</td>
<td>3,177</td>
<td>862</td>
</tr>
<tr>
<td>Synonymous</td>
<td>5,398</td>
<td>2,341</td>
<td>2,098</td>
<td>959</td>
</tr>
<tr>
<td>Intronic/Intergenic</td>
<td>14,169</td>
<td>56</td>
<td>650</td>
<td>13,463</td>
</tr>
<tr>
<td>Other</td>
<td>1,780</td>
<td>116</td>
<td>356</td>
<td>1,308</td>
</tr>
</tbody>
</table>

Variant type was determined based on dbNSFP v.2.0 annotations. Abbreviation is as follows: MAF, minor allele frequency. Other category includes variants labeled as downstream (n = 187), ncRNA_exonic (n = 111), ncRNA_intronic (n = 447), ncRNA_splicing (n = 1), ncRNA_UTR3 (n = 8), _UTR5 (n = 1), upstream (n = 181), upstream/downstream (n = 8), UTR3 (n = 518), UTR5 (n = 77).
recently established role in presenting the enzyme lipoprotein lipase to the luminal side of the vascular endothelium. 25 Mice deleted for Col18a1 (Col18a1−/−) were shown to have hypertriglyceridemia resulting from decreased lipoprotein lipase activity, and humans with homozygous deficiency of COL18A1 (Knobloch syndrome [MIM 267750]) have been observed to have higher blood TG than do normal individuals. 25

Finally, in AA participants, we also identified an association of blood HDL-C with a 0.2% frequency variant at the proprotein convertase subtilisin/kexin type 7 serine protease gene (PCSK7 [MIM 604874]; c.1511G>A [p.Arg504His]; rs142953140). Carriers of the PCSK7 rs142953140 mutation had 17 mg/dl higher HDL-C levels than did non-carriers (p = 5 × 10−20). PCSK7 rs142953140 was also associated with lower TG (−30%; p = 2 × 10−9) and nominally associated with lower LDL-C (−11.5 mg/dl; p = 0.02) among the AA participants (Table 4). PCSK7 is located near the chromosome 11 gene cluster involving APOA1/C3/A4/A5, a region where several genes regulate HDL-C and TG. Conditional analysis adjusting for six known variants in this region (rs964184, rs3135506, rs2266788, rs76353203, rs147210663, rs140621530) showed PCSK7 rs142953140 to be an independent association signal (pconditional = 3.4 × 10−20). PCSK7 rs142953140 was not polymorphic in individuals of European ancestry. PCSK7 belongs to the same family of serine proteases as does PCSK9, and a putative link between PCSK7 and lipoprotein metabolism was recently established in vitro. 26,27 In cultured cells, PCSK7 cleaved angiopoietin-like protein 4 (ANGPTL4) and activated its ability to inhibit lipoprotein lipase.

We studied the functional consequence of overexpression of human PCSK7 in mice and found that mice overexpressing PCSK7 in liver have lower HDL-C and higher TG compared to controls (Figure 1), suggesting that gain
of PCSK7 function decreases HDL-C and increases TG in vivo. Further studies are warranted to determine whether reduction of hepatic PCSK7, for example through Pcsk7 knockdown in mice, will result in the same lipid profile as that of human subjects with the PCSK7 rs142953140 mutation.

Gene-Based Association
In contrast to low-frequency variants, the alternative alleles at rare variants occur too infrequently to be analyzed individually and must be evaluated by aggregating into sets. Here, typically, the unit of analysis is a gene. We used two methods to perform gene-based analyses incorporating variants annotated as nonsynonymous or splice site: the sum of the number of minor alleles including variants with a MAF < 1\% or MAF < 0.1\% and the sequence kernel association test (SKAT), a test that considers effect sizes in both directions, including variants with a MAF < 5\%. Gene-based association results were well calibrated (Figure S3). We were able to replicate genes previously reported to be associated with blood lipid levels (Table 5). For example, multiple rare alleles at LDLR (MIM 606945) were strongly associated with blood LDL-C (p = 5 \times 10^{-11}). No new genetic associations were identified through the gene-based analyses.

Association with CHD
To assess whether the four newly associated low-frequency coding variants also relate to clinical cardiovascular disease, we tested association of each variant with CHD in 63,470 individuals from 15 EA studies and 13,772 individuals from 7 AA studies (Table S2). Of these participants, 14,201 EA and 2,380 AA individuals were affected with CHD.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Gene</th>
<th>rsID</th>
<th>Mutation (Substitution)</th>
<th>Beta$^a$</th>
<th>p Value</th>
<th>Conditional Variants</th>
<th>p\textsubscript{conditional}</th>
<th>EA MAF</th>
<th>AA MAF</th>
<th>p \textsubscript{conditional}</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL-C</td>
<td>ANGPTL8</td>
<td>rs145464906</td>
<td>c.361C&gt;T (p.Gln121*)</td>
<td>10 mg/dl</td>
<td>5.1 \times 10^{-11}</td>
<td>rs737337</td>
<td>5.5 \times 10^{-13}</td>
<td>0.1%</td>
<td>0.01%</td>
<td>0.1%</td>
</tr>
<tr>
<td>HDL-C</td>
<td>PAFAH1B2</td>
<td>rs186808413</td>
<td>c.482C&gt;T (p.Ser161Leu)</td>
<td>3 mg/dl</td>
<td>2.2 \times 10^{-10}</td>
<td>rs964184, rs3135506, rs2266788, rs76353203, rs14721066, rs140621530</td>
<td>2.1 \times 10^{-11}</td>
<td>1.1%</td>
<td>0.2%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

Results are based on 42,208 European ancestry (EA) individuals and 14,330 African ancestry (AA) individuals. Abbreviations are as follows: TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; MAF, minor allele frequency. $^a$Effect estimate for triglycerides is based on the geometric mean.

<table>
<thead>
<tr>
<th>Gene</th>
<th>rsID</th>
<th>Trait</th>
<th>EA Beta</th>
<th>EA p Value</th>
<th>AA Beta</th>
<th>AA p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANGPTL8</td>
<td>rs145464906</td>
<td>HDL-C</td>
<td>10 mg/dl</td>
<td>5.1 \times 10^{-11}</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TG</td>
<td>–15%</td>
<td>0.003</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LDL-C</td>
<td>–5.8 mg/dl</td>
<td>0.13</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PAFAH1B2</td>
<td>rs186808413</td>
<td>HDL-C</td>
<td>3 mg/dl</td>
<td>2.2 \times 10^{-10}</td>
<td>1 mg/dl</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TG</td>
<td>–10%</td>
<td>2.3 \times 10^{-09}</td>
<td>–4%</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LDL-C</td>
<td>–3 mg/dl</td>
<td>0.01</td>
<td>–2.6 md/dl</td>
<td>0.58</td>
</tr>
<tr>
<td>COL18A1</td>
<td>rs114139997</td>
<td>HDL-C</td>
<td>–</td>
<td>–</td>
<td>2.2 mg/dl</td>
<td>4.1 \times 10^{-4}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TG</td>
<td>–</td>
<td>–</td>
<td>–16%</td>
<td>1.6 \times 10^{-16}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LDL-C</td>
<td>–</td>
<td>–</td>
<td>–2.6 mg/dl</td>
<td>0.12</td>
</tr>
<tr>
<td>PCSK7</td>
<td>rs142953140</td>
<td>HDL-C</td>
<td>–</td>
<td>–</td>
<td>17 mg/dl</td>
<td>4.9 \times 10^{-20}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TG</td>
<td>–</td>
<td>–</td>
<td>–30%</td>
<td>1.5 \times 10^{-09}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LDL-C</td>
<td>–</td>
<td>–</td>
<td>–11.5 mg/dl</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Abbreviations are as follows: TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; EA, European ancestry; AA, African ancestry; Beta, effect size.
As a positive control, we confirmed that the PCSK9 c.137T (c.137G>T [p.Arg46Leu]; rs11591147) allele leading to lower LDL-C was associated with reduced risk for CHD among EA participants (OR = 0.84, 95% CI = 0.74–0.95, p = 0.007) and PCSK9 c.2037A (c.2037C>A [p.Cys679*]; rs28362286) allele leading to lower LDL-C was similarly associated with reduced risk for CHD among AA participants (OR = 0.40, 95% CI = 0.23–0.68, p = 5.4 × 10^{-3}).

In contrast to the PCSK9 variants, none of the four coding sequence variants at ANGPTL8, PAFAH1B2, COL18A1, or PCSK7 were associated with risk for CHD (Table 6).

**Discussion**

The contribution of low-frequency DNA sequence variation to complex phenotypes such as blood lipids and CHD has been the subject of much debate. Some have argued that low-frequency variants with large effects may be instrumental in discovering genes and that such discoveries can lead to novel therapeutic targets. By using the Exome Array in >42,000 EA individuals and >14,000 AA individuals, we discovered four low-frequency variants with large effects on lipids, but none of these variants associated with risk for CHD.

Several limitations of our study deserve mention. We were not able to evaluate extremely rare variants that may be unique to individuals; exome or whole-genome sequencing is needed to capture this type of variation. As such, it remains possible that a burden of such very rare mutations could contribute to plasma lipid variation. If so, the sample sizes required to yield new rare variant discoveries are likely to be extraordinarily large.

The Exome Array is constrained to the coding and splice site variation observed in the ~12,000 individuals who comprised the initial exome sequencing discovery set.

### Figure 1. Effects of Human PCSK7 Overexpression in Mouse Liver on Plasma Lipids

Plasma TC (A), HDL-C (B), and TG (C) were measured at baseline and at 7 and 14 days after injection of the human PCSK7 or control AAV8 vectors in C57BL/6j male mice. Error bars show standard deviations. *p < 0.05 and **p < 0.01, Student's unpaired t test.

<table>
<thead>
<tr>
<th>Gene</th>
<th>p Value</th>
<th>Best Test Beta</th>
<th>SE</th>
<th>CMAF</th>
<th>No. Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL-EA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCSK9</td>
<td>1.81 × 10^{-62}</td>
<td>SKAT – –</td>
<td>0.0626</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>LDLR</td>
<td>4.9 × 10^{-19}</td>
<td>T1 15.36</td>
<td>2.47</td>
<td>0.0027</td>
<td>28</td>
</tr>
<tr>
<td>HDL-EA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOC3</td>
<td>3.2 × 10^{-19}</td>
<td>T1 11.75</td>
<td>1.31</td>
<td>0.0035</td>
<td>4</td>
</tr>
<tr>
<td>ANGPTL4</td>
<td>5.41 × 10^{-30}</td>
<td>SKAT – –</td>
<td>0.0283</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>LIPG</td>
<td>2.55 × 10^{-29}</td>
<td>T1 0.41</td>
<td>0.05</td>
<td>0.0036</td>
<td>4</td>
</tr>
<tr>
<td>LPL</td>
<td>5.76 × 10^{-29}</td>
<td>SKAT – –</td>
<td>0.0162</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Triglycerides-EA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOC3</td>
<td>4.0 × 10^{-29}</td>
<td>T1 -0.41</td>
<td>0.05</td>
<td>0.0036</td>
<td>4</td>
</tr>
<tr>
<td>ANGPTL4</td>
<td>2.70 × 10^{-37}</td>
<td>SKAT – –</td>
<td>0.0286</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>LPL</td>
<td>1.56 × 10^{-31}</td>
<td>SKAT – –</td>
<td>0.037</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>LDL-AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCSK9</td>
<td>1.40 × 10^{-71}</td>
<td>SKAT – –</td>
<td>0.1295</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>APOE</td>
<td>2.65 × 10^{-12}</td>
<td>T1 -0.34</td>
<td>0.05</td>
<td>0.0055</td>
<td>4</td>
</tr>
<tr>
<td>HDL-AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOC3</td>
<td>9.5 × 10^{-12}</td>
<td>T1 10.14</td>
<td>1.49</td>
<td>0.0055</td>
<td>4</td>
</tr>
<tr>
<td>CETP</td>
<td>7.7 × 10^{-9}</td>
<td>T1 6.18</td>
<td>1.07</td>
<td>0.0065</td>
<td>7</td>
</tr>
<tr>
<td>Triglycerides-AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOC3</td>
<td>2.9 × 10^{-19}</td>
<td>T1 -0.34</td>
<td>0.05</td>
<td>0.0055</td>
<td>4</td>
</tr>
<tr>
<td>APOE</td>
<td>2.26 × 10^{-19}</td>
<td>SKAT – –</td>
<td>0.0205</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations are as follows: TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; EA, European ancestry; AA, African ancestry; Beta, effect size; SE, standard error.

*Beta is in mg/dl units for LDL and HDL and in percent change for triglycerides.

*CMAF, cumulative minor allele frequency of the variants contributing to the test.

*No. variants, number of variants contributing to the test.
Furthermore, ~20% of the content contributed for design failed to be converted into genotyping assays and thus these variants are not present on the Exome Array.

The lack of significant association of our reported low-frequency variants with CHD may be due to insufficient statistical power. For the CHD association testing among EA participants, we had >80% power at \( \alpha = 0.05 \) to detect an odds ratio outside the range (0.86, 1.14) with a 2% frequency variant, and an odds ratio outside the range (0.74, 1.29) with a 0.5% frequency variant. Among AA participants, we had >80% power at \( \alpha = 0.05 \) to detect an odds ratio outside the range (0.69, 1.35) with a 2% frequency variant, and an odds ratio outside the range (0.43, 1.73) with a 0.5% frequency variant.

Of note, all four of our reported loci were associated with HDL-C, which may not be a causal risk factor for clinical CHD, and this observation could explain the lack of association with CHD. Finally, we do not yet fully understand the biologic mechanism by which the genes reported here affect blood lipoproteins.

In summary, by using the Exome Array, we identified four low-frequency coding variants in ANGPTL8, PAFAH1B2, COL18A1, and PCSK7 that altered plasma HDL-C and/or TG but did not affect risk for CHD. These results suggest that the example of PCSK9 with low-frequency alleles that affect both plasma lipids and CHD is likely to be an exception rather than a paradigm.

### Supp. Data

Supplemental Data include descriptions of study samples, Supplemental Acknowledgments, three figures, and two tables and can be found with this article online at http://www.cell.com/AJHG/.

### Acknowledgments

The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO), and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). Infrastructure for the CHARGE Consortium is supported, in part, by the National Heart, Lung, and Blood Institute (grant HL105756). G.M.P. is supported by award number T32HL007208 from the NHLBI. S.K. is supported by a Research Scholar award from the Massachusetts General Hospital (MGH), the Howard Goodman Fellowship from MGH, the Donovan Family Foundation, 801HL107816, and a grant from Fondation Leducq. Exome Array genotyping in case-control studies of coronary heart disease was supported by NIH RC2 HL-102925 and an investigator-initiated research grant from Merck to S.K. N.O.S. is supported, in part, by a career development award from the NIH/NHLBI K08-HL114642. A.D. is supported by NWO grant (veni, 916.12.154) and the EUR Fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHLBI or NIH. A full listing of acknowledgements is provided in Supplemental Data.

Received: November 8, 2013  
Accepted: January 9, 2014  
Published: February 6, 2014

### Web Resources

- dbNSFP v.2.0, https://sites.google.com/site/jpopgen/dbNSFP
- Exome Array site information, ftp://share.sph.umich.edu/exomeChip/
- Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
- R metafor package, http://cran.r-project.org/web/packages/metafor/index.html
- seqMeta, http://cran.r-project.org/web/packages/seqMeta/
- UCSC Genome Browser, http://genome.ucsc.edu

### Supplementary Table 6

Association of Lipid Variants with Coronary Heart Disease

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mutation (Substitution)</th>
<th>rsID</th>
<th>Case Frequency</th>
<th>Control Frequency</th>
<th>Odds Ratio</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ancestry: EA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANGPTL8</td>
<td>c.361C&gt;T (p.Gln121*)</td>
<td>rs145464906</td>
<td>0.29%</td>
<td>0.27%</td>
<td>1.023</td>
<td>0.985</td>
</tr>
<tr>
<td>PAFAH1B2</td>
<td>c.482C&gt;T (p.Ser161Leu)</td>
<td>rs186808413</td>
<td>2.05%</td>
<td>2.23%</td>
<td>0.905</td>
<td>0.170</td>
</tr>
<tr>
<td><strong>Ancestry: AA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCSK7</td>
<td>c.1511G&gt;A (p.Arg504His)</td>
<td>rs142953140</td>
<td>0.53%</td>
<td>0.47%</td>
<td>1.258</td>
<td>0.592</td>
</tr>
<tr>
<td>COL18A1</td>
<td>c.331G&gt;A (p.Gly111Arg)</td>
<td>rs114139997</td>
<td>3.39%</td>
<td>3.38%</td>
<td>0.996</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Association with CHD was performed in a total of 63,470 EA individuals and 13,772 AA individuals. 14,201 EA and 2,380 AA individuals developed CHD.


1. Authors state that “it is still unclear whether these GWAS loci [that were implicated in GWASs of European-ancestry populations] can be generalized to other ethnic groups, such as African Americans.” Do you agree or disagree with this assertion? Explain why or why not.

2. Studies have shown that obesity is highly genetic, with heritability estimates ranging from 40-70%. Furthermore, the prevalence of worldwide obesity has increased sharply over recent decades. Are these two observations conflicting? Does the recent increase in obesity indicate recent changes in the genetic composition of the human population, worldwide?

3. Investigators performed genetic imputation for the WHI SHARE cohort. In brief, 1,962 women serving as a reference panel were genotyped for both (1) the Metabochip custom panel of SNPs at putative BMI loci, and (2) the (less-dense) Affymetrix 6.0 platform. Another 6,326 women were genotyped only for the Affymetrix platform. Imputation of the Metabochip SNPs was performed for the 6,326 women. Explain the logical basis of imputation, what the investigators stood to gain from this analysis, and whether you think this was a good approach? What could the investigators have done to assess the accuracy of the imputed genotypes?

4. Authors used the p-value threshold of $5.8 \times 10^{-5}$ for statistical significance. This is a considerably smaller p-value threshold than 0.05. Explain why investigators used the stricter p-value threshold.

5. Authors were careful to model population structure (i.e., principal components of ancestry). Given that their samples were all African American, do you think this approach was necessary?

6. The investigators claim that differences in the linkage disequilibrium patterns between European ancestry and African ancestry population can aid in narrowing the association signals. Explain the logic behind this approach. What is the purpose of narrowing the association signal? Were investigators successful in this effort?

7. The investigators repeated their analysis in sex-stratified samples. Why did they do this?
Fine Mapping and Identification of BMI Loci in African Americans

Jian Gong,¹,* Fredrick Schumacher,⁹ Unhee Lim,⁸ Lucia A. Hindorff,¹⁰ Jeff Haessler,¹ Steven Buyske,¹,⁵ Christopher S. Carlson,¹ Stephanie Rosse,¹ Petra Büžková,¹⁹ Myriam Fornage,²³ Myron Gross,²⁴ Nathan Pankratz,²⁴ James S. Pankow,²⁵ Pamela J. Schreiner,²⁵ Richard Cooper,¹³ Georg Ehret,¹¹,¹² C. Charles Gu,¹⁷ Denise Houston,¹⁸ Marguerite R. Irvin,¹⁶ Rebecca Jackson,⁶ Lew Kuller,⁷ Brian Henderson,⁹ Iona Cheng,²⁷ Lynne Wilkens,⁸ Mark Leppert,¹⁵ Cora E. Lewis,¹⁴ Rongling Li,¹⁰ Khanh-Dung H. Nguyen,¹¹ Robert Goodloe,²² Eric Farber-Eger,²² Jonathan Boston,²² Hollï H. Dilks,²² Marylyn D. Ritchie,²⁰ Jay Fowke,²¹,²² Loreall Pooler,⁹ Misa Graff,³ Lindsay Fernandez-Rhodes,³ Barbara Cochrane,²⁶ Eric Boerwinkle,²³ Charles Kooperberg,¹ Tara C. Matise,⁴ Loïc Le Marchand,⁸ Dana C. Crawford,²² Christopher A. Haiman,⁹ Kari E. North,²,³ and Ulrike Peters¹,⁸

Genome-wide association studies (GWASs) primarily performed in European-ancestry (EA) populations have identified numerous loci associated with body mass index (BMI). However, it is still unclear whether these GWAS loci can be generalized to other ethnic groups, such as African Americans (AAs). Furthermore, the putative functional variant or variants in these loci mostly remain under investigation. The overall lower linkage disequilibrium in AA compared to EA populations provides the opportunity to narrow in or fine-map these BMI-related loci. Therefore, we used the Metabochip to densely genotype and evaluate 21 BMI GWAS loci identified in EA studies in 29,151 AAs from the Population Architecture using Genomics and Epidemiology (PAGE) study. Eight of the 21 loci (SEC16B, TMEM18, ETV5, GNPDA2, TEAP2B, BDNF, FTO, and MC4R) were found to be associated with BMI in AAs at 5.8 × 10⁻⁵. Within seven out of these eight loci, we found that, on average, a substantially smaller number of variants was correlated (r² > 0.5) with the most significant SNP in AA than in EA populations (16 versus 55). Conditional analyses revealed GNPDA2 harboring a potential additional independent signal. Moreover, Metabochip-wide discovery analyses revealed two BMI-related loci, BRE (rs116612809, p = 3.6 × 10⁻⁸) and DHX34 (rs4802349, p = 1.2 × 10⁻⁷), which were significant when adjustment was made for the total number of SNPs tested across the chip. These results demonstrate that mapping in AAs is a powerful approach for both narrowing in on the underlying causal variants in known loci and discovering BMI-related loci.

Introduction

Obesity (MIM 601665) is a major risk factor for a number of chronic diseases, such as type 2 diabetes (MIM 125853), hyperlipidemia (MIM 144250), cardiovascular diseases, and several cancer types.¹,² Worldwide obesity prevalence has nearly doubled since 1980, and in 2008 more than 1.4 billion adults worldwide were obese. In the United States, more than one-third of adults (35.7%) were obese in 2010.

Studies have shown that obesity is highly heritable; heritability is estimated to fall in the range of 40%–70%.³,⁴ Genome-wide association studies (GWAS) have identified numerous loci associated with body mass index (BMI),⁵–⁷ a common measure of obesity. However, most of these studies were performed among European-ancestry (EA) populations. It is still unclear whether previously identified GWAS loci are population specific or whether they can be generalized to other ethnic groups, such as African Americans (AAs). Furthermore, the overall lower linkage disequilibrium (LD) patterns in AA compared to EA populations can offer opportunities to narrow in or fine-map BMI-related loci.⁸ This will help to reduce the number of variants for functional follow-up studies, which tend to

¹Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; ²Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27514, USA; ³Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA; ⁴Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA; ⁵Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854, USA; ⁶Department of Internal Medicine, Ohio State Medical Center, Columbus, OH 43210, USA; ⁷Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA; ⁸Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA; ⁹Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA; ¹⁰Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; ¹¹Center for Complex Disease Genomics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; ¹²Division of Cardiology, Geneva University Hospital, Geneva, OH 44041, Switzerland; ¹³Preventive Medicine and Epidemiology, Loyola University Chicago, Chicago, IL 60613, USA; ¹⁴Department of Medicine, University of Alabama, Birmingham, AL 35294, USA; ¹⁵Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA; ¹⁶Department of Biostatistics, University of Alabama, Birmingham, AL 35294, USA; ¹⁷Department of Biostatistics, Washington University, St. Louis, MO 63110, USA; ¹⁸Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA; ¹⁹Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; ²⁰Pennsylvania State University, State College, PA 16801, USA; ²¹Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, Nashville, TN 37203, USA; ²²Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN, 37203, USA; ²³The Human Genetics Center and Institute of Molecular Medicine, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; ²⁴Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, USA; ²⁵Division of Epidemiology & Community Health, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA; ²⁶Baylor College of Medicine, Houston, TX 77030, USA; ²⁷Cancer Prevention Institute of California, Fremont, CA 94538, USA

*Correspondence: jgong@fhcrc.org (J.G.), upeters@fhcrc.org (U.P.)

http://dx.doi.org/10.1016/j.ajhg.2013.08.012. ©2013 by The American Society of Human Genetics. All rights reserved.
be time and labor intensive. In addition, dense genotyping of the GWAS loci could aid the discovery of additional independent signals within the GWAS loci.

In this study, we densely genotyped 21 BMI loci identified in EA studies in 29,151 AAs from the Population Architecture using Genomics and Epidemiology (PAGE) consortium by using the Metabochip. We aimed to fine-map the 21 known BMI loci in the AA population and search for additional independent signals associated with BMI. For validated loci in AAs, we evaluated whether weaker LD patterns in AAs can help narrow in on the underlying potential causal variants. In addition, because Metabochip was developed to test putative association signals for BMI and many obesity-related metabolic and cardiovascular traits and to fine-map established loci, we also conducted a Metabochip-wide discovery-oriented analysis to search for potential BMI-associated loci.

Subjects and Methods

Study Population

The National Human Genome Research Institute funds the PAGE consortium to investigate the epidemiologic architecture of well-replicated genetic variants associated with human diseases or traits. PAGE consists of a coordinating center and four consortia, Epidemiologic Architecture for Genes Linked to Environment (EAGLE), which is used data from Vanderbilt University Medical Center's biorepository and links it to deidentified electronic medical records (BioVU); the Multiethnic Cohort Study (MEC); the Women's Health Initiative (WHI); and Causal Variants Across the Life Course (CALiCo), itself a consortium of five cohort studies—the Atherosclerosis Risk in Communities (ARIC) study, Coronary Artery Risk Development in Young Adults (CARDIA), the Cardiovascular Health Study (CHS), the Hispanic Community Health Study/Study of Latinos, and the Strong Heart Study.

This PAGE Metabochip study included AA participants from the ARIC, BioVU, CHS, CARDIA, MEC, and WHI studies and from extended collaborations to two additional studies – GenNet and the Hypertension Genetic Epidemiology Network (HyperGEN) (Table S1, in the Supplemental Data available with this article online). The detailed description of each study can be found in the Supplemental Data. We excluded underweight (BMI < 18.5 kg/m²) and extremely overweight (BMI > 70 kg/m²) individuals under the assumption that these extremes could be attributable to data-coding errors or an underlying rare condition outside the scope of this investigation. We also limited analysis to adults (defined as having an age > 20 years). The CARDIA participants are young, and the BMI < 18.5 exclusion criterion was not applied in this cohort. All studies were approved by institutional review boards at their respective sites, and all study participants provided informed consent.

Anthropometric Measurements

For individuals from the ARIC, CHS, CARDIA, HyperGEN, GenNet, and WHI studies, BMI was calculated from height and weight measured at the time of study enrollment. For individuals from BioVU, the median height and weight across all visit years were used in BMI calculations. For individuals from MEC, self-reported height and weight were used for calculations of baseline BMI. A validation study within MEC has shown high validity of self-reported height and weight. Specifically, this study showed that BMI was underestimated on the basis of self-reports versus measured weight, but the difference was small (< 1 BMI unit) and was comparable to the findings from national surveys.

Genotyping and Quality Control

Genotyping was performed with the Metabochip, whose design has been described elsewhere. In brief, the Metabochip, a custom Illumina Select genotyping array of nearly 200,000 SNP markers, is designed to cost-effectively analyze putative association signals identified through GWAS meta-analyses of many obesity-related metabolic and cardiovascular traits and to fine-map established loci. Metabochip SNPs were selected from the catalogs developed by the International HapMap and 1000 Genomes projects. More than 122,000 SNPs were included for fine mapping of 257 GWAS loci of 23 traits (including 21 BMI loci). For determination of the boundaries around each GWAS index SNP, all SNPs with an r² ≥ 0.5 with the index SNP were identified, and then initial boundaries were expanded by 0.02 cM in either direction through use of the HapMap-based genetic map. SNPs were then excluded if (1) the Illumina design score was < 0.5 or (2) SNPs within 15 bp in both directions of the SNP of interest could be found with an allele frequency of > 0.02 among Europeans (CEU). SNPs annotated as nonsynonymous, essential splice site, or stop codon were included regardless of allele frequency, design score, or nearby SNPs in the primer.

Twenty-one BMI GWAS loci identified at the time at which the Metabochip was designed were represented for signal fine mapping (Table S2). Samples were genotyped at the Human Genetics Center of the University of Texas, Houston (ARIC, CHS, CARDIA, GenNet, and HyperGEN), the Vanderbilt DNA Resources Core in Nashville (BioVU), the University of Southern California Epigenome Center (MEC), and the Translational Genomics Research Institute (WHI). Each center genotyped the same 90 HapMap YRI (Yoruba in Ibadan, Nigeria) samples to facilitate cross-study quality control (QC), as well as 2%–3% study-specific blinded replicates to assess genotyping quality. Genotypes were called separately for each study via GenomeStudio with the GenCall 2.0 algorithm. Study-specific cluster definitions (based on samples with call rate > 95%; ARIC, BioVU, CHS, CARDIA, MEC, and WHI) or cluster definitions provided by Illumina (GenNet and HyperGEN) were used for sample calling, and samples were kept in the analysis if the call rate was > 95%. We excluded SNPs with a GenTrain score < 0.6 (ARIC, BioVU, CHS, CARDIA, MEC, and WHI) or < 0.7 (GenNet and HyperGEN), a cluster separation score < 0.4, a call rate < 0.95, and a Hardy-Weinberg equilibrium p < 1 × 10⁻⁶. We utilized the common 90 YRI samples and excluded any SNP that had more than 1 Mendelian error (in 30 YRI trios), any SNP that had more than two replication errors with discordant calls when comparisons were made across studies in 90 YRI samples, and any SNP that had more than three discordant calls for 90 YRI genotyped in PAGE versus the HapMap database. SNPs were excluded from the meta-analyses if they were present in less than three studies.

For ARIC, BioVU, CHS, CARDIA, MEC, and WHI combined we identified related individuals by using PLINK to estimate identical-by-descent (IBD) statistics for all pairs. When apparent pairs of first-degree relatives were identified, we excluded from each pair the member with the lower call rate. We excluded from further analysis samples with an inbreeding coefficient (f) above 0.06.
SNPs with allele frequency $\frac{1}{4}$). No inflation was observed in any analysis (the inflation factor weighting to pool the study-specific association results as imputed). We used fixed-effect models with inverse variance weighting to pool the study-specific association results as accounted for. We used family data from GenNet and HyperGen data to account for relatedness. We then performed a haplotype-type analysis imputation on Metabochip SNPs for the remaining 6,326 SHARE subjects with Affymetrix 6.0 data. Details can be found elsewhere. In brief, we first merged genotypes for the 1,962 subjects genotyped on both the Affymetrix 6.0 platform and the Metabochip and constructed haplotypes (study-specific reference panel). We then phased the haplotypes for samples genotyped on the Affymetrix 6.0 platform only and performed a haplotype-to-haplotype imputation on Metabochip SNPs for the 6,326 target individuals to estimate genotypes (as allele dosages). We used MACH for phasing and Minimac for final imputation. To evaluate the quality of each imputed SNP, we calculated the dosage $r^2$. We excluded imputed SNPs with $r^2 < 0.5$ for SNPs with allele frequency < 1% and with $r^2 < 0.3$ for SNPs with allele frequency > 1%. Given the large reference panel and strict QC criteria, this resulted in high imputation quality.

### Statistical Analysis

In each study, we evaluated the association between natural-log-transformed BMI and each SNP. Because the distribution of BMI was not normal (it was skewed toward higher BMI), we used natural-log-transformed BMI, which reduces the influence of potential outlying observations on the analyses. Except for GenNet and HyperGen, linear regression models were used under the assumption of an additive genetic model and with the adjustment for age, sex, study site (as applicable), and ancestry principal components in each study. All models (except WHI) included the interaction term of sex and age so that possible effect modification by sex was accounted for. Family data from GenNet and HyperGen were analyzed with linear mixed models so that relatedness was accounted for. We used fixed-effect models with inverse variance weighting to pool the study-specific association results as implemented in METAL. We used Q-statistics and $I^2$ to measure heterogeneity across studies. To determine whether the 21 previously identified GWAS loci were significantly associated with BMI in AA, we used the $p$ value threshold $5.8 \times 10^{-5}$ as an approximate correction for an average of 866 SNPs at each locus (i.e., 0.05/866 SNPs). To identify additional independent signals in any of the loci that were significantly associated with BMI in our study, we performed conditional analyses. We performed linear regression models that included the most significant SNP (i.e., lead SNP) as a covariate and each of the other SNPs at the same locus (two SNPs in each model) to search for additional independent signals. If, after adjustment for the lead SNP, any SNP remained significant at the locus-specific Bonferroni-corrected significance level (i.e., 0.05/number of SNPs tested at a given locus), we defined the SNP as an additional independent signal. When testing all other SNPs on the Metabochip, we used a Bonferroni-adjusted significance level based on the total number of SNPs on the chip, $2.5 \times 10^{-7}$ (0.05/200,000 SNPs), to declare a BMI-associated SNP. No inflation was observed in any analysis (the inflation factor $\lambda = 1.00$) in our Metabochip-wide analysis. For sex-stratified analysis, GenNet and HyperGen were excluded because of the family-study design.

LD in the AA sample was calculated in 500 kb sliding windows via PLINK. Likewise, the Malmo Diet and Cancer Study on 2,143 controls from a Swedish population provided Metabochip LD and frequency information in Europeans to facilitate the LD pattern comparisons between AA and EA populations. We used LocusZoom plots to graphically display the fine-mapping results. SNP positions from NCBI build 37 were used, and recombination rates were estimated from 1000 Genomes Project data.

### Functional Annotation

To inform the discussion about the underlying potential functional variants, we made functional hypotheses for each of our most significantly BMI-associated variants by compiling a list of correlated SNPs ($r^2 > 0.5$) genotyped in our AA study populations and annotating each list for potential regulatory evidence consistent with enhancers, promoters, insulators, silencers and other effects related to gene expression. Because our lead SNPs and the SNPs in strong LD with our lead SNPs were in noncoding regions, we hypothesized that the underlying biology behind the signal was likely to impact gene expression through some unknown regulatory mechanism. For each list we aligned correlated SNPs with a combined browser view of all currently available ENCODE tracks in the UCSC Genome Browser and compared each allelic region for altered transcription factor binding site (TFBS) motifs by using JASPAR and ConSite. Given that methylation patterns are highly variable, it is useful to look for, in addition to BMI-relevant tissues, histone modifications in many cell lines to identify regions that are actively regulated, meaning that histone marks are present in some but not all cell lines. Thus, in addition to adipocytes, hepatocytes, and neurons, we used ENCODE’s histone modification tracks to query a wide variety of cell lines for all available histone marks to identify SNPs falling in various regulatory regions. The Dnase hypersensitivity track provided a more precise demarcation of open chromatin loci, and the ChiP-Seq TFBS track provided evidence for the binding of specific proteins. Although less specific than ChiP-Seq, JASPAR, ConSite, and HaploReg databases were used for querying a larger number of conserved TFBS and predicting alterations in predicted motifs between reference and alternative alleles. A 46-way PhastCons track in the UCSC Genome Browser was used as secondary evidence for a regulatory region, but lack of conservation did not rule out a functional candidate. The most likely functional-candidate SNPs for each locus were evaluated for statistical significance in association with BMI. The detailed list of functional annotation data sets we used is shown in Table S3.

### Results

This study consisted of 29,151 AAs from eight studies. Study participants had an average age of 51.2 years (Table S1). Approximately 80% of study participants were women. Within and across studies, men tended to have a lower mean BMI than women. The obesity rate (BMI $\geq 30$ kg/m$^2$) ranged from 16%–46% in men and 26%–64% in women. After quality control, we tested 18,187 genetic variants across 21 BMI loci and 177,663 variants across the Metabochip.
<table>
<thead>
<tr>
<th>Region</th>
<th>Lead AA SNP on Metabochip</th>
<th>GWAS Index SNP</th>
<th>Position</th>
<th>Number of SNPs</th>
<th>Candidate Gene</th>
<th>Candidate Gene with GWAS SNP in AAs</th>
<th>CA *</th>
<th>Rsq</th>
<th>Het P *</th>
<th>I²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1q25.2</td>
<td>rs543874</td>
<td>rs543874</td>
<td>177889480</td>
<td>765</td>
<td>SEC16B</td>
<td>-</td>
<td>0.75</td>
<td>0.81</td>
<td>NA</td>
<td>0.0</td>
</tr>
<tr>
<td>2p25.3</td>
<td>rs6548240</td>
<td>-</td>
<td>636929</td>
<td>1,123</td>
<td>TMEM18</td>
<td>-</td>
<td>0.87</td>
<td>0.83</td>
<td>NA</td>
<td>0.1</td>
</tr>
<tr>
<td>3q27.2</td>
<td>rs6548238</td>
<td>-</td>
<td>634905</td>
<td>-</td>
<td>TMEM18</td>
<td>0.86</td>
<td>0.99</td>
<td>NA</td>
<td>NA</td>
<td>0.1</td>
</tr>
<tr>
<td>4p12</td>
<td>rs9816226</td>
<td>185834499</td>
<td>-</td>
<td>-</td>
<td>ETV5</td>
<td>0.37</td>
<td>0.85</td>
<td>NA</td>
<td>NA</td>
<td>0.1</td>
</tr>
<tr>
<td>6p12.3</td>
<td>rs2774475</td>
<td>50784880</td>
<td>1,685</td>
<td>-</td>
<td>TFAP2B</td>
<td>0.23</td>
<td>0.50</td>
<td>NA</td>
<td>NA</td>
<td>0.1</td>
</tr>
<tr>
<td>11p14.1</td>
<td>rs1519480</td>
<td>27675712</td>
<td>688</td>
<td>-</td>
<td>BDNF</td>
<td>0.14</td>
<td>0.12</td>
<td>NA</td>
<td>NA</td>
<td>0.1</td>
</tr>
<tr>
<td>16q12.2</td>
<td>rs62048402</td>
<td>53803223</td>
<td>1,814</td>
<td>-</td>
<td>FTO</td>
<td>0.44</td>
<td>0.81</td>
<td>NA</td>
<td>NA</td>
<td>0.1</td>
</tr>
<tr>
<td>18q21.32</td>
<td>rs6657160</td>
<td>57829135</td>
<td>1275</td>
<td>-</td>
<td>MC4R</td>
<td>0.15</td>
<td>0.94</td>
<td>NA</td>
<td>NA</td>
<td>0.1</td>
</tr>
</tbody>
</table>

(Continued on next page)
Fine Mapping BMI Loci

Among the 21 BMI GWAS loci identified in EA studies, eight loci (SEC16B [MIM 612855], TMEM18 [MIM 613220], ETV5 [MIM 601600], GNPDA2 [MIM 613222], TFAP2B [MIM 601601], BDNF [MIM 113505], FTO [MIM 610966], and MC4R [MIM 155541]) displayed SNPs with significant evidence of association (Table 1; see also Table S2). The lead SNP (the most significant SNP in AAs) in each of these eight loci had a minor-allele frequency >0.05 and showed little evidence of heterogeneity. In these eight loci most GWAS index SNPs previously identified in EA GWASs (20 out of 23) had a consistent direction of the effects reported in original EA studies. Because the results of the FTO locus have been described previously, we focused on the other seven regions. Among those seven loci, the lead SNPs, rs543874 in SEC16B (p = 1.5 × 10^{-9}), rs7647305 in ETV5 (p = 3.2 × 10^{-5}), and rs10938397 in GNPDA2 (p = 1.7 × 10^{-7}), were consistent with the previously identified GWAS SNP (the most significant SNP highlighted in the previous GWAS) in EA populations. For rs543874 and rs10938397, the observed effects on BMI were slightly stronger in AAs than in EA individuals (change in BMI per coded allele: 1.1% in AAs and 0.9% in EA individuals for rs543874; 1.0% in AAs and 0.8% in EAs for rs10938397), whereas they were slightly weaker in AAs than in EA individuals for rs7647305 (change in BMI per coded allele: 0.7% in AAs and 0.9% in EA individuals). The minor-allele frequency (MAF) was higher for rs543874 and rs7647305 and lower for rs10938397 in AAs than in EA individuals (Table 1). In the other four loci, the lead SNPs in AAs differed from the GWAS SNPs from EA populations. The lead SNPs in all four loci (rs6548240 in TMEM18, rs2744475 in TFAP2B, rs1519480 in BDNF, and rs6567160 in MC4R) were modestly to strongly correlated (r² ranged from 0.5–1.0) with at least one of the GWAS SNPs on the basis of LD in EA populations. However, when LD was based on AA populations, the correlation was weaker, in several cases substantially weaker (Table 1). For the 13 BMI loci that did not replicate in our AA analysis, most GWAS index SNPs (13 out of 17) from previous GWASs involving EA individuals showed effects in the same direction when results from our AA samples were compared with results from the previous EA studies.

We investigated the question of whether LD patterns in AA studies can narrow previous association signals from EA studies and found that at seven out of the eight significant loci, AA LD patterns assisted with narrowing association signals (Table 2; see also Figure S1). One of the most extreme examples was for MC4R. Among EA individuals, 107 and 119 SNPs were correlated (r² > 0.5) with the lead SNP in our analysis (rs6567160) and with the GWAS index SNPs, respectively; these SNPs represent a region spanning 184 kb and 230 kb, respectively. However, among AAs only five SNPs were correlated with rs6567160 at r² > 0.5, and these SNPs represented a region spanning 71 kb. In SEC16B, ETV5, TFAP2B, BDNF, and FTO, the number of
SNPs correlated at $r^2 > 0.5$ with the lead SNP and the spanning region size were also substantially reduced and to a limited degree for TMEM18 (Table 2). Only for GNPD2A did the number of correlated SNPs and the spanning region size not reduce, but the number of correlated SNPs in Europeans was already small.

To further refine the regions associated with BMI in AAs, we performed conditional analyses for each of the eight significant loci by including the lead SNP in a locus as a covariate to search for additional independent signals. rs186117327 in GNPD2A was borderline significantly associated with BMI when the locus-specific significance level was taken into account (Table 3). No evidence of heterogeneity was observed across studies. In GNPD2A, rs186117327 was significantly associated with BMI in the marginal analysis ($p = 3.7 \times 10^{-7}$); this association became less significant but remained marginally so when adjustment was made for the lead SNP ($p = 1.7 \times 10^{-4}$). rs186117327 was weakly correlated with the lead SNP ($r^2 = 0.11$ in AAs; 0.16 in EA individuals).

### Identification of Two BMI-Related Loci in AAs by MetaboChip-wide Analysis

In the MetaboChip-wide analysis (excluding the SNPs in the 21 BMI-related loci), we identified five SNPs in two loci (these SNPs were rs116612809, rs114584581, rs74941130, and rs79329695 in 2p23.2/BRE [MIM 610497] and rs4802349 in 19q13.32/DHX34) as being associated with BMI in AAs at a MetaboChip-wide significance level ($p < 2.5 \times 10^{-7}$). Two of the SNPs in 2p23.2/BRE (rs116612809 and rs79329695) reached the conventional genome-wide significance level ($p < 5.0 \times 10^{-8}$) (Table 4 and Figure 1). Furthermore, with $p = 3.7 \times 10^{-7}$, rs57813622 at the 7p21.2/GNBAP1 [MIM 604070] locus approached the MetaboChip-wide significance level.

In 2p23.2/BRE, all four SNPs are rare variants in Europeans (MAF = 0.1% in the 1000 Genomes Project) but common in AAs (MAF = 10%). The most significant SNP, rs116612809, is located in the intronic region of BRE and is highly correlated with the three other significant SNPs ($r^2$ ranges from 0.98–1.00, Figure 2). In the conditional analysis for all other SNPs at the BRE locus, we did not observe evidence for an additional independent signal associated with BMI after conditioning on rs116612809, suggesting that all four SNPs point to the same functional variant. The 2p23.2/BRE was included on the MetaboChip as a part of a ~1.3-Mb-long region so that a GWAS locus associated with blood triglyceride (TG) concentrations could be fine-mapped. Within this region, the four BMI-associated SNPs were more than 500 kb away from the TG GWAS index SNPs (rs1260326, rs1260333, rs780093, and rs780094; $r^2$ ranged from 0.02–0.05 in AAs). To examine the association between rs116612809 and blood TG concentrations in our population, we used ARIC and WHI, for which the TG individual data were available ($n = 11,680$). We did not observe a significant
### Table 3. Conditional Analysis Showing a Locus with Evidence of Second Independent BMI Association Signals in African Americans

<table>
<thead>
<tr>
<th>Region</th>
<th>Gene</th>
<th>SNP</th>
<th>Position</th>
<th>Number of Tested SNPs</th>
<th>Coded Allele</th>
<th>CAF</th>
<th>Marginal Resultsa</th>
<th>Conditional Resultsb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Effect p Value</td>
<td>Eff ect p Value</td>
</tr>
<tr>
<td>4p12</td>
<td>GNPDA2</td>
<td>rs186117327</td>
<td>45,101,187</td>
<td>285</td>
<td>A</td>
<td>0.75</td>
<td>-0.0095 3.7 × 10⁻²</td>
<td>-0.0074 1.7 × 10⁻⁴</td>
</tr>
<tr>
<td>*</td>
<td></td>
<td>rs10938397</td>
<td>45,182,527</td>
<td>-</td>
<td>A</td>
<td>0.75</td>
<td>-0.0099 1.7 × 10⁻²</td>
<td>-0.0068 1.6 × 10⁻³</td>
</tr>
</tbody>
</table>

aMarginal results represent results when only the single variant was in the model.
bConditional analysis represents the result for the SNP when adjustment was made for the most significant lead SNP and vice versa.
cThe most significant SNP in a locus.
dHet p: heterogeneity-test p value.

### Table 4. Two BMI Loci Identified in African Americans on the Basis of Metabochip-wide Analysis

<table>
<thead>
<tr>
<th>Region</th>
<th>SNP</th>
<th>Position</th>
<th>Candidate Gene</th>
<th>Coded Allele</th>
<th>CAF in AAs</th>
<th>CAF in EA Individuals</th>
<th>Effect p Value</th>
<th>Rsqa</th>
<th>Het p</th>
<th>r² with Lead SNP in AAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2p23.2</td>
<td>rs116612809</td>
<td>28,301,171</td>
<td>BRE</td>
<td>A</td>
<td>0.90</td>
<td>0.001</td>
<td>-0.0151</td>
<td>3.6 × 10⁻⁶</td>
<td>0.99</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>rs114584581</td>
<td>28,304,380</td>
<td>BRE</td>
<td>A</td>
<td>0.10</td>
<td>0.001</td>
<td>0.0148</td>
<td>5.9 × 10⁻⁸</td>
<td>0.99</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>rs74941130</td>
<td>28,306,293</td>
<td>BRE</td>
<td>A</td>
<td>0.10</td>
<td>0.001</td>
<td>0.0148</td>
<td>6.9 × 10⁻⁸</td>
<td>0.99</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>rs79329695</td>
<td>28,319,874</td>
<td>BRE</td>
<td>A</td>
<td>0.10</td>
<td>0.001</td>
<td>0.0151</td>
<td>3.7 × 10⁻⁸</td>
<td>0.99</td>
<td>0.56</td>
</tr>
<tr>
<td>19q13.32</td>
<td>rs4802349</td>
<td>47,874,510</td>
<td>DHX34</td>
<td>A</td>
<td>0.48</td>
<td>0.12</td>
<td>-0.0087</td>
<td>1.2 × 10⁻⁷</td>
<td>1.00</td>
<td>0.46</td>
</tr>
</tbody>
</table>

aMeasurement of imputation accuracy, ranging from 0 (low) to 1 (high).
association \( (p = 0.71) \) with blood TG concentrations, which did not change after adjustment for BMI. Furthermore, the association between rs116612809 and BMI did not change after adjustment for TG (Table S4).

In 19q13.32/\( \text{DHX34} \), rs4802349 is located in an intron of \( \text{DHX34} \) and has a MAF of 0.48 in AAs and 0.12 in EA individuals. Only ten SNPs surrounding rs4802349 (\( \pm 200 \) kb) were genotyped; none was correlated with rs4802349 in AAs, and none was associated with BMI. Previously, rs4802349 was reported to be a putative association signal for high-density lipoprotein (HDL) cholesterol at a moderate \( p \) value of 0.005.\(^\text{26}\) We examined the association between rs4802349 and blood HDL concentrations with and without BMI adjustment in ARIC and WHI \( (n = 11,680) \). We observed that the \( p \) value changed from 0.25 to 0.04 after BMI adjustment. Furthermore, the association between rs4802349 and BMI did not change after adjustment for HDL (Table S4).

**Sex-Stratification Analyses**

We did not observe any additional loci when we stratified the analysis by sex (Tables S5 and S6 and Figures S2 and S3). Also, the results for the SNPs that were significant in the fine-mapping analyses and the two BMI loci results were consistent across sex \( (p\)-heterogeneity \( \geq 0.29, \text{Table S7}) \). However, as a result of the relatively small sample of men in this study, we have limited power to detect a difference-of-sex effect.

**Functional Annotation**

In silico analysis assessed whether each BMI-associated locus was in an “open,” transcriptionally permissive conformation, as would be expected of a functional locus. Using ENCODE data sets, we found each signal to be in a region consistent with regulatory evidence, such as active histone marks, open chromatin structure (DNase hypersensitivity), or regions experimentally shown to bind one or more transcription factors (Table S8). Bioinformatics analyses revealed that the lead SNP, rs6548240 in the \( \text{TMEM18} \) locus, was the strongest functional candidate. rs6548240 is in a region of open chromatin that binds multiple transcription factors, and elevated levels of active histone marks associated with promoters were detected in several cell lines (Figure S4). The functional properties of \( \text{TMEM18} \) are obscure, although a recent study indicated that \( \text{TMEM18} \) plays a regulatory role in adipocyte differentiation and biology.\(^\text{27}\) For \( \text{GNPDA2} \), the potential additional independent signal rs18617327 could be tagging rs7659184 \( (r^2 = 0.6 \) with rs186117327 in AAs), which falls in a region of open chromatin, and relative to the reference allele, the alternate allele reduces the binding affinity of GATA2 and EN1 transcription factors. In 2p23.2/\( \text{BRE} \), bioinformatics analyses indicated that the lead SNP, rs116612809, tags another intronic SNP, rs78003529 \( (r^2 = 0.65 \) in AAs), which falls in a region of open chromatin with elevated enhancer histone marks and is in a region that binds both Pol2 and c-Jun. Furthermore, evidence from scans of positional weight matrices (PWMs) suggests that the alternate allele of rs78003527 has a much higher binding affinity for Nkx3 and Nkx2. In 19q13.32/\( \text{DHX34} \), the \( \text{DHX34} \) signal rs4802349 and two SNPs highly correlated with it are located within putative regulatory regions. rs4802349 falls within histone-modification marks associated with enhancer activity and alters the Mtf1 transcription-factor binding motif. In addition, two other functional candidates, rs2547369 and rs2341878 \( (r^2 = 0.7 \) and 0.6, respectively in African populations) were highly correlated with this signal and fall in regions of open chromatin having methylation patterns associated with promoter or enhancer activity. Furthermore, rs25476369 \( (r^2 = 0.65 \) with rs4802349 in African populations) falls in a region that binds six transcription factors, whereas rs2341878 is of particular interest because it falls in a strong promoter region in a number of relevant tissues. The function of \( \text{GNPDA2} \), \( \text{BRE} \), and \( \text{DHX34} \) is unknown, which makes it difficult to link the underlying biological mechanisms of these genetic variants to BMI.

**Discussion**

In this study, encompassing close to 30,000 AAs, we used the Metabochip to systematically evaluate 21 BMI loci...
discovered among European descent populations in previous GWAs and to search for potential BMI loci. Eight of the 21 loci (SEC16B, TMEM18, ETV5, GNPDA2, TFAP2B, BDNF, FTO, and MC4R) were found to be associated with BMI in our AA study population. Further conditional analyses indicated that GNPDA2 contained an additional independent signal. Moreover, Metabochip-wide analyses revealed two BMI-associated loci: the BRE locus at genome-wide significance (p < 5 × 10^{-8}) and the DHX34 locus at Metabochip-wide significance (p < 2.5 × 10^{-7}).

Among the eight BMI loci that were significant in the AA population, the lead SNPs in SEC16B, ETV5, and GNPDA2 were the same as the GWAS SNPs identified in previous EA studies, which provided further support for the idea that the three SNPs (rs543874, rs7647305, and rs10938397) are proxies of causal variants influencing BMI. The most significant (lead) SNPs in the other five loci differed from the previously reported GWAS SNPs, although, as expected, all were moderately to highly correlated with the GWAS SNPs in European populations (r² ranged from 0.5 to 1.0). In TFAP2B and MC4R, the GWAS SNPs were not significantly associated with BMI in AAs, suggesting that the lead SNPs in these two loci are better proxies for the underlying functional variants. We showed that the weaker LD patterns in AA populations than in EA populations substantially reduced the number of functional-variant proxies in six of the eight BMI loci that were significant in AAs (and to a limited extent in TMEM18). Our results illustrate the important contribution of AAs to systematic fine-mapping of GWAS loci originally reported in EA populations. In addition, the lead SNPs that were either consistent or correlated with the GWAS SNPs at these replicated loci in EA populations might indicate that EA populations and AA populations share the underlying causal variants at these loci.

Bioinformatics analyses revealed that the lead SNP, rs6548240 in the TMEM18 locus, was the strongest functional candidate. Although rs6548240 is located in an intergenic region 31 kb downstream of TMEM18, ChIP-seq evidence indicates that CTCF looping might anchor this distant putative enhancer to the promoter of TMEM18. Taken together, these pieces of evidence make rs6548240 an interesting functional candidate for future laboratory follow-up. This example shows that combining fine mapping with bioinformatics analysis can help to narrow in on the putative functional variants for further follow-up studies.

There are multiple reasons that 13 BMI loci originally identified in EA populations were not significantly associated with BMI in this sample of AAs. Limited statistical power could be an important reason for this observation. Statistical power is impacted by the variance of BMI, MAF, effect size, and sample size. Compared to populations in some large European-focused studies, our AA population had larger variance in BMI (standard deviation: 6.2 kg/m² versus 4.2 kg/m²), which reduces the statistical power. Also compared with primarily very large European-focused studies with sample sizes from ~90,000 to ~250,000, our study was relatively small. Nonreplication might arise because of different causal variants between EA individuals and AAs, a weak AA LD pattern, which leads to weak correlation between causal variants and marker SNPs on the Metabochip, and limited statistical power. All of this might explain that we did not observe all loci significantly associated with BMI in AAs, and it emphasizes the need for a larger sample size. We used a uniform p value threshold of 5.8 × 10^{-5} (0.05/average number of SNPs per locus) as an approximate correction for the average of 866 SNPs across these loci. However, if we use Bonferroni correction (i.e., 0.05/the number of SNPs at a given locus) or correction by the effective number of SNPs at each locus after accounting for LD patterns, one more locus, SH2BI (MIM 608937), would be indicated to be significant in AAs (Table S2).

Conditional analysis indicated that the GNPDA2 locus contained a potential additional independent signal. The potential additional independent signal, rs186117327, is physically relatively close (80 kb) to the lead SNP,
rs10938397, but is only weakly correlated ($r^2 = 0.11$) with it.

We identified two BMI-related loci, 2p23.2/BRE and 19p13.32/DHX34, in the Metabochip-wide analysis. We evaluated whether African ancestry plays role in these two BMI-related loci among 8,310 AAs from a GWAS in WHI. We didn’t observe a significant effect of ancestry at these two loci ($p > 0.5$; Table S9). At 2p23.3/BRE, the most significant SNP, rs116612809, reached the conventional genome-wide significance level ($5 \times 10^{-8}$) and was surrounded (3–18 kb) by three highly correlated SNPs ($r^2 = 0.98–1.00$) showing very similar results. All four SNPs are common (MAF = 0.1), but they all are rare variants in EA populations (MAF = 0.001) and hence it is unlikely that these SNPs would be identified in EA GWASs unless they have much stronger effects than the moderate effects observed in our study or are tested in very large sample sets. The lead SNP in 2p23.3/BRE (rs116612809) is located within an intron of BRE, which is stress responsive and highly expressed in brain and reproductive organs. No previous study reported an association with BMI or any other trait; however, in support of our finding, the GIANT consortium reported that 82 of 200 SNPs within BRE had a p value < 0.05 (min. $p = 2.5 \times 10^{-4}$) for a BMI association in EA populations. Accordingly, our results showed that varying allele frequency in different ancestral groups significantly contributes to the statistical power and that studying different ancestral groups helps to identify potentially functional loci.

We found another potential BMI-related locus at 19q13.32/DHX34. The SNP rs4802349, located in an intron of DHX34, was marginally significantly associated with BMI at a Metabochip-wide significance level ($p = 1.2 \times 10^{-7}$), but because the finding did not reach the conventional genome-wide significance level of $5 \times 10^{-8}$, additional replication studies in AAs are warranted. DHX34 is a putative RNA helicase and has not been reported to be associated with any diseases or traits in humans before, except for a suggestive association with HDL cholesterol ($p = 0.005$). A recent GWAS in EA populations showed an association with BMI in a neighboring region, 19q13.32/ZC3H4; rs3810291 was the most significant SNP. However, rs4802349 is physically 300 kb away from rs3810291, and a recombination hotspot lies between ZC3H4 and DHX34 in AAs and Europeans (HapMap phase II YRI and CEU), resulting in a low correlation between both SNPs in African populations ($r^2 = 0.04$) and Europeans ($r^2 < 0.001$). This observation suggests that these two associations are independent from each other. Our results along with the evidence from functional annotation warrant additional validation studies for this locus.

In conclusion, we observed that eight BMI-associated GWAS loci identified from EA populations were significantly associated with BMI in AAs and identified a potential additional independent signal in one locus. In addition, we discovered two potential BMI-related loci through Metabochip-wide analysis; one of these loci which reached the conventional genome-wide significance level. Importantly, our study demonstrated that fine mapping in AA populations in combination with bioinformatics analyses is a valuable and effective way to narrow in on the underlying causal variants in GWAS loci discovered in EA populations and that studying minority populations can contribute to loci discovery.

Supplemental Data

Supplemental data include four figures, nine tables, and supplemental text, including acknowledgments and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

See Supplemental Data.

Web Resources

The URLs for data presented herein are as follows:

Center for Disease Control and Prevention (CDC), Adult Obesity Facts, http://www.cdc.gov/obesity/data/over.html
MACH, http://www.sph.umich.edu/csg/abecasis/mach
Minimac, http://genome.sph.umich.edu/wiki/Minimac
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
PAGEx Study, http://www.pagenxsite.org

References


670 The American Journal of Human Genetics 93, 661–671, October 3, 2013


Example qualifying exam (Ph.D.) and comprehensive exam (M.S.) questions pertaining to Jordan et al. (2012) *American Journal of Human Genetics.*

1. The burden test and variable threshold test (both of which collapse information on rare variants across a genomic region) were used to test rare variants in CARD14 exons for association with psoriasis. What is the rationale for such tests? Why did investigators not test each rare variant individually?

2. The common variant rs11652075 (see Figure 2A) was tested for genetic association with psoriasis in six individual cohorts, as well as across all six cohorts simultaneously using meta-analysis. The odds ratios in the individual cohorts ranged from moderately protective (OR=0.66) to weakly harmful (OR=1.09) and associations were statistically significant in some, but not all, cohorts. However, the odds ratio of the meta-analysis was protective (OR=0.87) and very significant (p=0.000002). Interpret the meta-analysis results in light of the findings from individual cohorts. Is the meta-analysis inconsistent with the individual cohorts? Why or why not?

3. Consider the “common disease, common variant” hypothesis. What is the premise of this hypothesis, and do the results from this paper support or refute this hypothesis an explanation for genetic nature of psoriasis?

4. Consider Table 2, which shows for each CARD14 coding variant (1) the bioinformatically-predicted effect on protein function, (2) the effect on NF-kB activation, and (3) the allele frequencies in cases and controls. Suppose that the bioinformatics prediction of variant’s function was “benign” and that this prediction was accurate. What would you expect for the variant’s effect on NF-kB activation? What would you expect for the relative allele frequencies in cases and controls?

Suppose instead that the bioinformatics prediction of a variant’s function was “damaging” and that this prediction was accurate. What would you expect for the variant’s effect on NF-kB activation? What would you expect for the relative allele frequencies in cases and controls?

Scanning Table 2, are there any coding variants for which functional prediction is inconsistent with effect of NF-kB activation, or relative allele frequencies in cases vs. controls, or both?

5. The effect of CARD14 variants on NF-kB activity was measured both with and without TNF-alpha stimulation. Explain the rationale for this.
Psoriasis is a common inflammatory disorder of the skin and other organs. We have determined that mutations in CARD14, encoding a nuclear factor of kappa light chain enhancer in B cells (NF-kb) activator within skin epidermis, account for PSORS2. Here, we describe fifteen additional rare missense variants in CARD14, their distribution in seven psoriasis cohorts (>6,000 cases and >4,000 controls), and their effects on NF-kb activation and the transcriptome of keratinocytes. There were more CARD14 rare variants in cases than in controls (burden test \( p = 0.0015 \)). Some variants were only seen in a single case, and these included putative pathogenic mutations (c.424G>A [p.Glu142Lys] and c.425A>G [p.Glu142Gly]) and the generalized-pustular-psoriasis mutation, c.413A>C (p.Glu138Ala); these three mutations lie within the coiled-coil domain of CARD14. The c.349G>A (p.Gly117Ser) familial-psoriasis mutation was present at a frequency of 0.0005 in cases of European ancestry. CARD14 variants led to a range of NF-kappaB activities; in particular, putative pathogenic mutations in caspase recruitment domain family, member 14 (CARD14) transfectants in keratinocytes differentiated probably pathogenic mutations from neutral variants such as polymorphisms. Over 20 susceptibility loci for psoriasis have been identified, and these include putative pathogenic mutations in CARD14. Two variants (c.511C>A [p.His171Asn] and c.536G>A [p.Arg179His]) associated with increased risk of psoriasis susceptibility locus 2 (PSORS2 [MIM 602723]) in chromosomal region 17q25. These mutations are responsible for the elusive psoriasis susceptibility locus 2 (PSORS2 [MIM 602723]) in chromosomal region 17q25. These results led us to hypothesize that additional rare and common variants in CARD14 might contribute to psoriasis and/or psoriatic arthritis in the general population.

Here, we identify and characterize 15 additional rare missense variants within CARD14 and determine their frequencies in a large cohort of approximately 6,000 psoriasis cases and 4,000 controls. Statistical analyses revealed an excess of rare variants in psoriasis cases relative to controls. The potential pathogenicity of variants was demonstrated by their ability to increase transcriptional activation by nuclear factor of kappa light chain enhancer in B cells (NF-kb) and to enhance production of a subset of psoriasis-associated transcripts. A common missense mutation (c.424G>A) in CARD14 increased NF-kB levels. Transcription profiling of wild-type and variant CARD14 transfectants in keratinocytes provided insights into the effects of CARD14 variants on the transcriptome and the pathogenesis of psoriasis.

Introduction

Psoriasis is a chronic, inflammatory disease of the skin and other organs. It affects approximately 2% of individuals of European descent, and in up to 30% of cases, it is associated with chronic inflammatory psoriatic arthritis. Genome-wide association studies (GWASs) have identified 20 susceptibility loci for psoriasis. However, with the exception of psoriasis susceptibility locus 1 (PSORS1 [MIM 177900]), for which the odds ratio (OR) is approximately 3.0, risk conferred by these loci is generally small (ORs ≤ 1.5). Moreover, less than 20% of disease variance has been explained. This implies that additional low-risk loci, genetic interactions, or rare variants of large effect account for the remaining variance.

In our accompanying paper, we identified rare, gain-of-function mutations in caspase recruitment domain family, member 14 (CARD14 [MIM 607211]) in two large multiplex families affected by Mendelian forms of psoriasis and psoriatic arthritis (see the accompanying paper in this issue of AJHG). We also identified a de novo mutation in CARD14 in a child with early-onset, severe pustular psoriasis (PSORP [MIM 614204]). These mutations are responsible for the elusive psoriasis susceptibility locus 2 (PSORS2 [MIM 602723]) in chromosomal region 17q25. These results led us to hypothesize that additional rare and common variants in CARD14 might contribute to psoriasis and/or psoriatic arthritis in the general population.

Here, we identify and characterize 15 additional rare missense variants within CARD14 and determine their frequencies in a large cohort of approximately 6,000 psoriasis cases and 4,000 controls. Statistical analyses revealed an excess of rare variants in psoriasis cases relative to controls. The potential pathogenicity of variants was demonstrated by their ability to increase transcriptional activation by nuclear factor of kappa light chain enhancer in B cells (NF-kb) and to enhance production of a subset of psoriasis-associated transcripts. A common missense mutation (c.424G>A) in CARD14 increased NF-kB levels. Transcription profiling of wild-type and variant CARD14 transfectants in keratinocytes provided insights into the effects of CARD14 variants on the transcriptome and the pathogenesis of psoriasis.

1Division of Human Genetics, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA; 2Department of Dermatology, University of Michigan, Ann Arbor, MI 48109, USA; 3Department of Dermatology, University of Utah School of Medicine, Salt Lake City, UT 84132, USA; 4Department of Dermatology, University of California, San Francisco, San Francisco, CA 94115, USA; 5Cardiovascular Research Institute and Department of Physiological Nursing, University of California, San Francisco, San Francisco, CA 94143, USA; 6Cardiovascular Research Institute, Department of Medicine, and Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143, USA; 7Sarah M. and Charles E. Seay Center for Musculoskeletal Research, Texas Scottish Rite Hospital for Children, Dallas, TX 75219, USA; 8National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892, USA; 9Laboratory for Investigative Dermatology, The Rockefeller University, New York, NY 10065, USA; 10Department of Medicine, Division of Rheumatology, Memorial University of Newfoundland, St. John’s, NL A1C 5S7, Canada; 11University of Toronto and Toronto Western Hospital, Toronto, ON MST 2S8, Canada; 12Psoriasis Research Institute, Baylor University Medical Center, Dallas, TX 75246, USA
*Correspondence: bowcock@genetics.wustl.edu
DOI 10.1016/j.ajhg.2012.03.013. ©2012 by The American Society of Human Genetics. All rights reserved.
variant within CARD14 was also associated with psoriasis, and that evidence for association increased when this locus was conditioned on the presence of PSORS1. Our findings indicate that a range of NF-kB responses in the skin are mediated by CARD14 and that a subset of rare CARD14 variants leads to psoriasis and psoriatic arthritis.

**Subjects and Methods**

**Subjects**

Cases and controls for sequencing and genotyping were recruited from multiple institutions. Samples were organized into cohorts as shown in Table 1. There were six cohorts of European ancestry and one of Asian ancestry. Referring to Table 1, European cases in cohort A were recruited from either Washington University in St. Louis or the Department of Dermatology at the University of California, San Francisco (UCSF). Controls in cohort A were unaffected individuals who were over 20 years of age and who had no family history of psoriasis; they were recruited from the Texas Scottish Rite Hospital for Children or from the Cardiovascular Research Institute and Center for Human Genetics at the University of California, San Francisco or they were CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) grandparents.

Cohort B samples were from the National Psoriasis Foundation Victor Henschel Tissue Repository (NPF). Cases and controls in cohort C were recruited from the Department of Dermatology at the University of Utah. For cohorts A-C, psoriasis was diagnosed by a dermatologist.

Cohort D samples were recruited from the Department of Dermatology at the University of Michigan. Cases had at least two psoriatic plaques or a single plaque occupying at least 1% of the total body surface outside the scalp. Individuals presenting with only palmoplantar psoriasis, inverse psoriasis, or sebopsoriasis were excluded. Controls were at least 18 years of age and had no personal or family history of psoriasis.

Cohort E samples were gathered from the University of Toronto and Toronto Western Hospital. Cohort F samples were gathered from the Department of Medicine, Division of Rheumatology, Memorial University of Newfoundland. Psoriasis was diagnosed by a dermatologist. When psoriatic arthritis was suspected, cases were evaluated according to clinical history and rheumatologic and radiologic evaluation. Control individuals showed no evidence of psoriasis, psoriatic arthritis, or any other autoimmune disease.

Asian samples in Cohort G were recruited from the Cardiovascular Research Institute and Center for Human Genetics at the University of California, San Francisco, from the University of Toronto and Toronto Western Hospital and the Department of Medicine, or from the NPF.

DNA was isolated from whole blood or saliva by standard methods. Protocols were approved by local institutional review boards. All subjects or their parents (if the subjects were minors) provided informed consent.

**Sanger Resequencing and Genotyping**

As a first pass, all coding exons of CARD14 (full-length, CARD14fl) were resequenced in 192 psoriasis cases and 96 controls of European ancestry. Exons in which rare missense mutations were identified were resequenced in 95 more controls of European ancestry.

Table 1. Psoriasis Cases and Controls Included in this Study

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Cohort Description</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>St. Louis/Dallas/UCSF</td>
<td>676</td>
<td>570</td>
</tr>
<tr>
<td>B</td>
<td>NPF (European)</td>
<td>486</td>
<td>154</td>
</tr>
<tr>
<td>C</td>
<td>Utah</td>
<td>931</td>
<td>236</td>
</tr>
<tr>
<td>D</td>
<td>Michigan</td>
<td>2,768</td>
<td>2,749</td>
</tr>
<tr>
<td>E</td>
<td>Toronto</td>
<td>981</td>
<td>483</td>
</tr>
<tr>
<td>F</td>
<td>Newfoundland</td>
<td>340</td>
<td>379</td>
</tr>
<tr>
<td>G</td>
<td>Asians</td>
<td>194</td>
<td>193</td>
</tr>
<tr>
<td>H</td>
<td>Combined European</td>
<td>6,182</td>
<td>4,571</td>
</tr>
<tr>
<td></td>
<td>(rows A–F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>NPF (all samples)</td>
<td>976</td>
<td>758</td>
</tr>
</tbody>
</table>

Individuals were recruited from multiple institutions and organized into cohorts, as shown in this table. Cohorts A-F are independent case/control cohorts of Northern European ancestry. Cohort G includes samples of Asian ancestry from the St. Louis/Dallas/UCSF (University of California, San Francisco) and the NPF (National Psoriasis Foundation Victor Henschel Tissue Repository) cohorts (cohorts A and B, respectively). Cohorts A-G include samples of known ancestry. Cohort H lists all cases and controls contributed by the NPF, including those of unknown ancestry. Demographic data were available only for approximately half of the NPF samples.

Exon 4 of CARD14 was resequenced in an additional 1,856 cases and 882 controls of European ancestry. Primers are available upon request. Samples were genotyped with the Sequenom MassARRAY at Washington University and by TaqMan at the University of Michigan (Table S2, available online). All samples from the NPF were genotyped, but only samples of known ethnicity were included in statistical analyses.

**Expression Plasmids**

Full-length CARD14sh (GenBank BC018142, coding for 740 amino acids) and CARD14cl (RefSeq NM_052819, coding for 434 amino acids) were cloned into pReceiver-M11 (Capital Biosciences). The CARD14sh construct was subjected to in vitro mutagenesis with the QuikChange Site-Directed Mutagenesis Kit (Stratagene). The numbering of all CARD14 mutations in this manuscript is based on RefSeq NM_024110.3. For rare, nonannotated missense variants, constructs were generated with the mutant allele. For polymorphisms (rare and common), constructs were generated as follows (the allele and amino acid used are listed in parentheses): rs11558260 (c.185A [p.Gln185]), rs1159200Asn (c.599A [p.Asn200]), rs146214639 (c.449G [p.Arg150]), rs144475004 (c.526C [p.His176]), rs2066964 (c.930C [p.Ser547]), rs34367357 (c.1042A [p.Ile585]), rs117918077 (c.2044T [p.Glu683]), rs146214639 (c.449G [p.Arg150]), rs151150961 (c.2140A [p.Ser714]). Full-length CARD14fl was not available for subcloning. As a result, constructs could not be created for rs144285237 (c.2919C>T [p.Arg973Gln]) or rs11652075 (c.2458C>T [p.Asp820Trp]).

**NF-kB Luciferase Reporter Assay**

The NF-kB luciferase reporter assay was performed with the pNFkB-luc system (Clontech) as described in our accompanying manuscript. The CARD14cl clone was used as a negative control in this assay because it lacks the sequence encoding the CARD domain, which is necessary for CARD14-induced NF-kB activation.
Expression Profiling and qRT-PCR
HEK 001 cells (human-papillomavirus-16-transformed keratinocytes) were transfected with wild-type CARD14sh or expression constructs encoding CARD14sh substitutions. Cells were cultured for 24 hr, and then RNA was isolated with the mirNeasy kit (QIAGEN). Global expression profiling of RNA from cells was performed with the HumanHT-12 v4 Expression BeadChip (Illumina). Experiments were conducted in compliance with MIAME (minimum information about a microarray experiment) guidelines. Raw and normalized expression data are deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) as series GSE36381. 100 ng to 1 μg of randomly primed total RNA was used for quantitative RT-PCR (qRT-PCR) according to standard procedures. Expression levels were normalized to 18S rRNA. Relative expression levels were calculated as follows: 2^(-ΔΔCt) (ΔCt = CT 18S − CT gene). We also normalized expression levels in transfected cells to levels of FLAG to correct for differences in transfection efficiency.

Clustering of Variants on the Basis of Expression Levels in Keratinocytes
On the basis of the NF-κB reporter assays, a subset of CARD14 variants were classified as (1) leading to enhanced basal NF-κB activation when they were compared to the effects of wild-type CARD14sh (these are c.349G>A [p.Gly117Ser], c.413A>C [p.Glu138Ala], c.424G>A [p.Glu142Lys], and c.425A>G [p.Glu142Gly]), (2) leading to downregulation of NF-κB activation (c.112C>T [p.Arg38Cys]), or (3) having no effect on NF-κB activation (these are c.185G>A [p.Arg62Gln] [rs115582620], c.930G>C [p.Arg547Ser] [rs2066964], c.449T>G [p.Leu150Arg] [rs146214639], c.854A>G [p.Asp285Gly], c.1778T>A [p.Ile593Asn], c.2044C>T [p.Arg628Trp] [rs1177918077], and c.2140G>A [p.Gly714Ser] [rs151150961]). After performing mRNA transcriptome analysis on cells that were transfected with either wild-type CARD14sh or CARD14 variants, we found that 1,531 transcripts had at least a 2-fold change (up or down) in expression as well as a significant p value in transfecants with the CARD14 variants causing enhanced NF-κB activation relative to transfecants with wild-type CARD14sh. The variants were then clustered with the use of the reduced probe set and the R (v2.10.1) randomForest package (v4.6-2). Random forests were generated with 500 trees and unweighted classes, and importance was calculated for a total of 10,000,000 forests. The probes were then ranked from highest to lowest mean decrease in Gini. A heat map of the clustering of the full set of variants was generated with the top 30 probes.

Pathway Analysis of the “CARD14 Pathogenic Keratinocyte Signature”
A list was generated of genes differentially expressed as a consequence of the introduction of CARD14 psoriasis-specific alterations into HEK 001 cells as described above. We obtained this list, termed the “CARD14 pathogenic keratinocyte signature,” by comparing the global transcriptomes of keratinocyte transfecants with overtly pathogenic CARD14 substitutions (p.Gly117Ser, p.Glu138Ala, p.Glu142Lys, and p.Glu142Gly) to those with nonpathogenic substitutions (p.Leu150Arg [rs146214639], p.Val191Leu, p.Asp285Gly, and wild-type CARD14sh). Expression data from each sample were quantile normalized, log_2 transformed, and fitted to linear models in R (R v2.13.1; Biobase v2.12.2; BeadArray v2.2.0; and limma v3.8.3). The contrast was defined as “pathogenic versus nonpathogenic,” and the t tests, fold changes, and false-discovery-rate-corrected p values were calculated with lmFit and eBayes. Given the overexpression of the constructs, larger sample sizes would be required for the detection of significant group-wise effects. However, by taking the genes with a nominal group-wise p value of 0.05 and ranking them by fold change, we generated a list of the top 200 upregulated and top 200 downregulated genes. This list was analyzed with Ingenuity pathway analysis (IPA).

Statistical Analysis
Analysis of variants was performed with PLINK. The variable threshold test and a straightforward burden test for association of rare variants with disease were performed with PLINK/SEQ version 0.05. For both tests, we included CARD14 variants with minor allele frequencies of less than or equal to 0.01 among our controls of northern European ancestry. This included the following 16 variants: c.112C>T [p.Arg38Cys], c.185G>A (p.Arg62Gln) [rs115582620], c.349G>A (p.Gly117Ser) (altered in family PS1), c.413A>C (p.Glu138Ala) (altered in generalized pustular psoriasis), c.424G>A (p.Glu142Lys), c.425A>G (p.Glu142Gly), c.449T>G (p.Leu150Arg) [rs146214639], c.511C>A (p.His171Asn), c.526G>C (p.Asp176His) [rs144475004], c.536G>A (p.Arg179His), c.571G>T [p.Val191Leu], c.599G>A (p.Ser200Asn), c.854A>G (p.Asp285Gly), c.1778T>A (p.Ile593Asn), c.2140G>A [p.Gly714Ser] [rs151150961], and c.2919C>G (p.Asp973Gly) [rs144285237]. Meta-analysis of polymorphisms and generation of forest plots were performed with R version 2.12.2 with the meta package.

Results
Rare-Variant Screening
In our accompanying manuscript, we identified rare gain-of-function CARD14 mutations that lead to psoriasis. These included the familial c.349G>A (p.Gly117Ser) mutation in family PS1, affected by multiple cases of psoriasis and psoriatic arthritis, and the de novo c.413A>C (p.Glu138Ala) germline mutation in a case of childhood generalized pustular psoriasis. To determine whether there were additional rare variants predisposing to psoriasis in CARD4, we resequenced all coding exons of CARD14 (full-length, CARD14fl) in over 192 psoriasis cases and 96 controls (see Subjects and Methods). This revealed ten rare missense mutations in CARD14 (Figure 1, Table 2, and Table S1): c.112C>T (p.Arg38Cys), c.185G>A (p.Arg62Gln) [rs115582620], c.425A>G (p.Glu142Gly), c.449T>G (p.Leu150Arg) [rs146214639], c.599G>A (p.Ser200Asn), c.854A>G (p.Asp285Gly), c.1778T>A (p.Ile593Asn), c.2044C>T (p.Arg628Trp) [rs1177918077], c.2140G>A [p.Gly714Ser] [rs151150961], and c.2919C>G (p.Asp973Gly) [rs144285237]. On the basis of the observation that several rare mutations were clustered in exon 4, which encodes part of the critical coiled-coil domain, we reasoned that it might be a mutation hotspot and resequenced an additional 1,800 cases and 900 controls for that exon. This revealed five additional rare variants within CARD14: c.424G>A (p.Glu142Lys), c.511C>A (p.His171Asn), c.526G>C (p.Arg179His) [rs144285237], c.449T>G (p.Leu150Arg) [rs146214639], and c.536G>A (p.Arg179His), c.599G>A (p.Ser200Asn), c.854A>G (p.Asp285Gly), c.1778T>A (p.Ile593Asn), c.2140G>A [p.Gly714Ser] [rs151150961], and c.2919C>G (p.Asp973Gly) [rs144285237].
(p.Asp176His) (rs144475004), c.536G>A (p.Arg179His), and c.571G>T (p.Val191Leu) (Figure 1, Table 2, and Table S1). A search in dbSNP135 revealed that 8 of the 15 identified rare variants have not been previously annotated; these eight are c.112C>T (p.Arg38Cys), c.424G>A (p.Glu142Lys), c.425A>G (p.Glu142Gly), c.511C>A (p.His171Asn), c.536G>A (p.Arg179His), c.571G>T (p.Val191Leu), c.599G>A (p.Ser200Asn), c.854A>G (p.Asp285Gly), and c.1778T>A (p.Ile593Asn). Furthermore, this search revealed that the c.349G>A (p.Gly117Ser) and c.413A>C (p.Glu138Ala) mutations have also not been previously annotated.

Figure 1. CARD14 Protein Domains and Locations of Amino Acid Substitutions
Missense variants identified in CARD14 by resequencing are shown relative to key protein domains. Red outlining indicates increased NF-kB activation by a variant (relative to wild-type); green outlining indicates reduced activation. Gold outlining indicates variants that showed increased NF-kB activation only in response to TNF-α stimulation. SNPs and variants also identified in dbSNP135, which includes data from the 1,000 Genomes Project and the National Heart, Lung, and Blood Institute (NHBLI) and National Human Genome Research Institute (NHGRI) Exome Project, are indicated with blue triangles. A polymorphism, rs114688446, was observed at the same site as p.Ser200Asn, but a different substitution, p.Ser200Ile, was detected. SNPs used for meta-analysis are indicated with gray triangles.

Population-Based Frequency Estimates of Rare Variants
We determined the frequencies of all rare CARD14 variants, including those encoding the familial p.Gly117Ser alteration (causing psoriasis and/or psoriatic arthritis) and the p.Glu138Ala alteration (causing generalized pustular psoriasis), by high-throughput genotyping seven independent case/control cohorts (>10,000 individuals; Tables 1 and 2 and Tables S1 and S2). This revealed two other unrelated psoriasis cases with the CARD14 c.349G>A (p.Gly117Ser) mutation; one woman was from Utah and had a history of psoriasis, which was diagnosed at 65 years old. This latter woman transmitted the mutation to her daughter, who developed psoriasis at age 32. This mutation was also detected in a male NPF control for whom additional data (e.g., ethnicity, age, and family history of psoriasis) were not available. None of these individuals harbored the SLC26A11 (solute carrier family 26, member 11, [SLC26A11 (MIM 610117)]) c.365A>G (p.Tyr122Cys) mutation, which cosegregated with c.349G>A (p.Gly117Ser) in family PS1,17 providing evidence that these variants can arise independently. Numbering of the SLC26A11 mutation is based on RefSeq NM_001166347.1. In total, the frequency of the CARD14 c.349G>A (p.Gly117Ser) mutation in cases of European ancestry was 0.0005. The CARD14 c.349+5G>A mutation, which segregated with psoriasis in a large multiply affected Taiwanese family,17 was not detected in any other cases or controls. However, given the relatively small number of Asian samples screened, we were not well powered to detect variants at low frequencies in that population. The mutation encoding p.Glu138Ala seen in the single pustular psoriasis case was also not seen in any other cases or controls.

Five rare CARD14 variants (encoding p.Arg38Cys, p.Glu142Lys, p.Glu142Lys, p.Val191Leu, and p.Asp285Gly; Figure 1 and Tables S1 and S2) were seen in only one case and...
in no controls. The variant encoding p.His171Asn was only seen in two psoriasis- and psoriatic-arthritis-affected cases from Newfoundland and in no controls. We performed a simple burden test and a variable threshold test\(^{19}\) to compare the distribution of rare variants in cases and controls. These tests provided evidence of an excess of rare \(\text{CARD14}\) variants in cases versus controls (burden test \(p = 0.0015\); variable threshold test \(p = 0.00015\)).

**Common-Variant Association Tests**

Resequencing validated several common missense polymorphisms described in dbSNP: rs2066964, rs34367357, and rs11652075. We genotyped these and 20 other previously described SNPs in our seven psoriasis case/control cohorts (Table S2) and looked for association with psoriasis. Most variants were present at similar frequencies in cases and controls. However, three missense SNPs (rs2066964, rs34367357, and rs11562075) within \(\text{CARD14}\) and rare variant c.599G>A (p.Ser200Asn) provided evidence of association with psoriasis in some of the cohorts (Table S2). A meta-analysis of these four missense variants was performed for the six cohorts of European ancestry (Figure 2 and Figure S1). This revealed further evidence of association between psoriasis and rs11562075 (c.2458C>T [p.Arg820Trp]) (fixed effects \(p = 2.1 \times 10^{-6}\), \(OR = 0.87 [0.83–0.92]\); random effects \(p = 0.031\), \(OR = 0.86 [0.75–0.99]\); c.2458C was the risk allele (Figures 2A and 2B). This SNP, as well as the same risk allele, was also associated with psoriasis in Asians (\(p = 0.0029\), \(OR = 0.64 [0.48–0.86]\)). The meta-analysis also revealed evidence of association with \(p\text{.Ser200Asn}\) (fixed and random effects \(p = 0.05\), \(OR = 1\)).

### Table 2. Characteristics and Frequencies of \(\text{CARD14}\) Coding Variants

<table>
<thead>
<tr>
<th>(\text{CARD14}) Exon</th>
<th>cDNA Mutation and Corresponding Protein Change</th>
<th>Protein Domain</th>
<th>PolyPhen2(^{27})-Predicted Effect on Protein Function</th>
<th>Effect on NF-(\kappa)B Activation (FC versus Wild-Type (\text{CARD14sh}))</th>
<th>Allele Frequency in Cases (Number Sampled)</th>
<th>Allele Frequency in Controls (Number Sampled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>c.112C&gt;T (p.Arg38Cys)</td>
<td>CARD</td>
<td>probably damaging</td>
<td>0.11</td>
<td>0.00019 (2,691)</td>
<td>0 (1,271)</td>
</tr>
<tr>
<td>2</td>
<td>c.185G&gt;A (p.Arg62Gln)</td>
<td>CARD</td>
<td>benign</td>
<td>1.06</td>
<td>0.0014 (3,284)</td>
<td>0.00084 (1,797)</td>
</tr>
<tr>
<td>3</td>
<td>c.349G&gt;A (p.Gly117Ser)</td>
<td>none</td>
<td>possibly damaging</td>
<td>3.71</td>
<td>0.00023 (6,630)</td>
<td>0 (4,731)</td>
</tr>
<tr>
<td>3</td>
<td>c.349&gt;5G&gt;A</td>
<td>none</td>
<td>NA</td>
<td>ND</td>
<td>0 (2,871)</td>
<td>0 (1,339)</td>
</tr>
<tr>
<td>4</td>
<td>c.413A&gt;C (p.Glu138Ala)</td>
<td>coiled-coil</td>
<td>probably damaging</td>
<td>8.95</td>
<td>0.00015 (3,488)</td>
<td>0 (1,902)</td>
</tr>
<tr>
<td>4</td>
<td>c.424G&gt;A (p.Glu142Lys)</td>
<td>coiled-coil</td>
<td>probably damaging</td>
<td>4.03</td>
<td>0.00012 (4,107)</td>
<td>0 (1,874)</td>
</tr>
<tr>
<td>4</td>
<td>c.425A&gt;G (p.Glu142Gly)</td>
<td>coiled-coil</td>
<td>probably damaging</td>
<td>5.00</td>
<td>0.00019 (2,848)</td>
<td>0 (1,451)</td>
</tr>
<tr>
<td>4</td>
<td>c.449T&gt;G (p.Leu150Arg)</td>
<td>coiled-coil</td>
<td>probably damaging</td>
<td>1.79</td>
<td>0.0025 (6,140)</td>
<td>0.0016 (4,614)</td>
</tr>
<tr>
<td>4</td>
<td>c.511C&gt;A (p.His171Asn)</td>
<td>coiled-coil</td>
<td>benign</td>
<td>0.68 (5.95 with TNF-(\alpha) stimulation)</td>
<td>0.00025 (4,077)</td>
<td>0 (1,858)</td>
</tr>
<tr>
<td>4</td>
<td>c.526G&gt;C (p.Asp176His) (rs144475004)</td>
<td>coiled-coil</td>
<td>probably damaging</td>
<td>2.78</td>
<td>0.00056 (3,575)</td>
<td>0.00062 (1,609)</td>
</tr>
<tr>
<td>4</td>
<td>c.536C&gt;A (p.Arg179His)</td>
<td>coiled-coil</td>
<td>probably damaging</td>
<td>1.38 (2.19 with TNF-(\alpha) stimulation)</td>
<td>0.00025 (4,061)</td>
<td>0.00027 (1,848)</td>
</tr>
<tr>
<td>4</td>
<td>c.571G&gt;T (p.Val191Leu)</td>
<td>coiled-coil</td>
<td>benign</td>
<td>1.02</td>
<td>0.00014 (3,575)</td>
<td>0 (1,613)</td>
</tr>
<tr>
<td>4</td>
<td>c.599G&gt;A (p.Ser200Asn)</td>
<td>coiled-coil</td>
<td>benign</td>
<td>0.67</td>
<td>0.011 (6,163)</td>
<td>0.0084 (4,624)</td>
</tr>
<tr>
<td>6</td>
<td>c.854A&gt;G (p.Asp285Gly)</td>
<td>none</td>
<td>possibly damaging</td>
<td>1.14</td>
<td>0.00019 (2,673)</td>
<td>0 (1,467)</td>
</tr>
<tr>
<td>13</td>
<td>c.1778T&gt;A (p.Ile593Asn) (rs146214639)</td>
<td>PDZ</td>
<td>probably damaging</td>
<td>1.30</td>
<td>0.00024 (2,049)</td>
<td>0.00048 (1,039)</td>
</tr>
<tr>
<td>15</td>
<td>c.2044C&gt;T (p.Arg682Trp) (rs117918077)</td>
<td>SH3</td>
<td>probably damaging</td>
<td>0.95</td>
<td>0.013 (2,169)</td>
<td>0.012 (1,042)</td>
</tr>
<tr>
<td>15</td>
<td>c.2140G&gt;A (p.Gly714Ser) (rs151150961)</td>
<td>SH3</td>
<td>benign</td>
<td>1.02</td>
<td>0.0021 (2,105)</td>
<td>0.0014 (1,038)</td>
</tr>
<tr>
<td>21</td>
<td>c.2919C&gt;G (p.Asp973Glu) (rs144285237)</td>
<td>GUK</td>
<td>benign</td>
<td>ND(^{b})</td>
<td>0.0024 (5,177)</td>
<td>0.0015 (4,099)</td>
</tr>
</tbody>
</table>
[1.3–1.82]); the rare c.599A allele increased psoriasis risk (Figure 2C). However, this would not be significant if it were adjusted for multiple testing. Because of the large effect of HLA-Cw*0602 from the major histocompatibility complex (MHC) class I region (PSORS1),12,13 we investigated its connection with the four CARD14 variants described above. In the Michigan and Utah psoriasis cohorts, rs11652075 was found to have a higher association with psoriasis when it was conditioned on HLA-Cw*0602 (p value of rs11652075 alone = 0.023 (Michigan), 0.017 (Utah); stratified on HLA-Cw*0602, p value = 0.0021 (Michigan), 0.0086 (Utah); Table S3). No such evidence was observed with the other variants.

Effect of Variants on CARD14 Function In Vitro

CARD14 encodes a 1,004 amino acid protein that activates NF-kB.16 In our companion paper, we observed that compared to wild-type CARD14, the familial p.Gly117Ser and de novo p.Glu138Ala substitutions lead to enhanced NF-kB activation (3.71- and 8.95-fold enhancement, respectively). To test the effect of rare variants described here on this activity, we again used an NF-kB luciferase reporter assay. Several rare variants (p.Glu142Lys, p.Glu142Gly, and p.Asp176His [rs144475004]) in the coiled-coiled domain increased NF-kB activation two to five times more than did wild-type CARD14sh (Table 2, Figure 3A). The p.Arg38Cys, p.His171Asn, and p.Ser200Asn substitutions led to less NF-kB activation than did CARD14sh, but compared to wild-type CARD14sh, other rare variants and the CARD14 missense polymorphisms (p.Arg62Gln [rs115582620], p.Arg547Ser [rs2066964], p.Arg682Trp [rs117918077]) did not significantly alter NF-kB activation levels (Table 2, Figure 3A). As discussed below, variants that increased NF-kB activation at least 2.5x more than that seen with CARD14sh also induced greater upregulation of psoriasis-associated genes. Relative to wild-type CARD14sh, variants that resulted in a more
modest increase in NF-kB activation, those that reduced NF-kB activation, and those that did not change it did not induce upregulation of those genes to the same degree. Therefore, when compared with the level of NF-kB activation caused by wild-type CARD14, a 2.5× or greater increase in NF-kB activation is predictive of putative pathogenic CARD14 amino acid substitutions.

Onset of psoriatic lesions is thought to be triggered by an inflammatory stimulus. We therefore examined the effects of wild-type and variant CARD14 on NF-kB activation after stimulation with tumor necrosis factor alpha (TNF-α). Compared with unstimulated CARD14sh, TNF-α-stimulated p.His171Asn and p.Arg179His resulted in a 5.95- and a 2.95-fold increase in NF-kB activation, respectively; compared with TNF-α-stimulated CARD14sh, TNF-α-stimulated p.His171Asn and p.Arg179His resulted in a 2.87- and a 1.06-fold increase in NF-kB activation, respectively (Table 2 and Figure 3B, discussed further below).

Effect of CARD14 Substitutions on Keratinocyte Gene Expression

We have shown that CARD14 is localized in keratinocytes and that the familial and pustular-psoriasis variants (p.Gly117Ser and p.Glu138Ala, respectively) lead to enhanced production of some chemokines and other transcripts that are upregulated in psoriatic skin.17 To evaluate the effect of the additional rare variants on transcription in keratinocytes, we transfected all altered constructs into the keratinocyte cell line HEK 001. The transcriptome of each transfectant was then evaluated after 24 hr by interrogation with Illumina bead arrays. A heat map with 30 probes (see Subjects and Methods, random forest classification) revealed clustering of the p.Glu142Lys and p.Glu142Gly transfectants with those with the p.Gly117Ser and p.Glu138Ala alterations (Figure 4). The p.Glu138Ala substitution clustered at the extreme end of other pathogenic variants and on its own branch of the tree. This might be expected given the severity of the disease in the child (who has generalized pustular psoriasis) in whom it was found.

Within this heat map, the following 13 genes were upregulated in keratinocytes transfected with pathogenic substitutions: superoxide dismutase 2, mitochondrial (SOD2 [MIM 147460]); interleukin 6 (interferon, beta 2 [IL6 [MIM 147620]]); colony stimulating factor 2 (granulocyte-macrophage [CSF2 [MIM 138960]]); interleukin 8 (IL8 [MIM 146930]); matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase [MMP9 [MIM 120361]]); BRF2, subunit of RNA polymerase III transcription initiation factor, BRF1-like (BRF2 [MIM 607013]); chemokine (C-C motif) ligand 20 (CCL20 [MIM 601960]); solute carrier family 7 (cationic amino acid transporter, y+ system), member 2 (SLC7A2 [MIM 601872]); oxidized low density lipoprotein (lectin-like) receptor 1 (OLR1 [MIM 602601]); interleukin 36, gamma (IL36G [MIM 605542]); guanylate binding protein 2, interferon-inducible (GBP2 [MIM 600412]); tumor necrosis factor, alpha-induced protein 2 (TNFAIP2 [MIM 603300]); and tumor necrosis factor (TNF [MIM 191160]). We used g:Profiler23 to look for specific functions of these genes. The most significant Gene Ontology24 terms associated with this group of genes included: immune-system process/development, hematopoietic- or lymphoid-organ...
development, and response to lipopolysaccharide/bacte-
rium/molecule of bacterial origin. IL6, MMP9, and BRF2
have also been implicated in cell migration and could
mediate immune cell infiltration into the skin. CSF2, IL8,
and SLC7A2 have been specifically implicated in myeloid
leukocyte activation. Psoriasis is a disease of keratinocyte
and immune cell proliferation, and genes involved in cell
proliferation include SOD2, IL6, CSF2, and IL8.
Other terms included response to wounding (SOD2, IL6, IL8,
CCL20, SLC7A2, and OLR1), nitric oxide biosynthetic
process (SOD2, IL6, and SLC7A2), and antiapoptosis
(SOD2, IL6, and CSF2). IL6 and CSF2 are implicated in
the regulation of the JAK-STAT pathway, and IL6, IL8,
and MMP9 are implicated in angiogenesis. Genes involved
in the response to biotic stimuli include SOD2, IL6, IL8,
MMP9, and CCL20. Psoriasis increases risk of cardiovas-
cular disease, and five of these genes have been impli-
cated in the development of the cardiovascular system
(SOD2, IL6, IL6, MMP9, and TNFAIP2). Upregulation of
these thirteen genes thus constitutes a pathogenic psori-
atic signature.

Two variants clustered together: c.526G>C (p.Asp176His)
(rs144475004) and c.536G>A (p.Arg179His). Both amino
acid substitutions were predicted to be damaging by Poly-
Phen2, and both lead to upregulation of NF-κB activation
(2.78- and 1.38-fold increases, respectively). In the case of
p.Arg179His, TNF-α was required for NF-κB activation to
achieve pathogenic levels; compared with unstimulated
CARD14sh and TNF-α-stimulated CARD14sh, TNF-α-stimu-
lated p.Arg179His produced a 2.95- and a 1.06-fold increase,
respectively. Both variants were seen in a small number of
cases and controls (4:2 and 2:1, respectively).

Three other constructs clustered in the same branch with
p.Asp176His (rs144475004) and p.Arg179His. These
included wild-type CARD14sh, c.2140G>A (p.Gly714Ser)
(rs151150961), and c.599G>A (p.Ser200Asn). However,
the latter three did not induce expression of the patho-
genic psoriatic signature described above to the same

Figure 4. Heat Map Generated after Transcriptome Analysis of Wild-Type and Mutant CARD14 Transfectants
Cells were transfected with CARD14cl, wild-type CARD14sh, or variant CARD14 expression constructs. After 24 hr, RNA was extracted
and used for the interrogation of Illumina bead arrays. A subset of differentially expressed genes was used for clustering of CARD14 vari-
ants (see Subjects and Methods).
degree as the overtly pathogenic variants. One noteworthy example of this is that the IL8 expression produced by these three constructs was much lower than that seen with the pathogenic variants. Moreover, clustering of these constructs was due to other transcripts that are not considered part of the pathogenic signature. Other variants exhibited reduced levels of genes in this psoriasis signature, even when compared to wild-type CARD14sh.

We confirmed altered expression of all thirteen transcripts (SOD2, IL6, CSF2, BRF2, MMP9, IL8, CCL20, SLC7A2, OLR1, IL36G, GBP2, TNFAIP1, and TNF) by qRT-PCR (Figure S2).

We also performed a group-wise comparison of the global expression profiles of the overtly pathogenic substitutions (p.Gly117Ser, p.Glu138Ala, p.Glu142Lys, and p.Glu142Gly) and several nonpathogenic variants (p.Leu150Arg, p.Val191Leu, p.Asp285Gly, and wild-type CARD14sh). The top 200 upregulated and top 200 downregulated genes were identified (see Subjects and Methods), and pathway analysis was performed with IPA (Table S4). A number of cytokine signaling pathways were significant (including IL-17 signaling, IL-6 signaling, and TNFR2 signaling). Also significant were communication between innate and adaptive immune cells, dendritic cell maturation, mTOR signaling, notch signaling, and atherosclerosis signaling pathways. This latter pathway is interesting given the association between psoriasis and other systemic comorbidities, including cardiovascular disease.25,26

A comparison of these results with a published psoriasis transcriptome28 revealed that a number of these pathways are significantly represented in both groups. For example, the atherosclerosis signaling pathway, the NF-kB signaling pathway, and many of the cytokine signaling pathways were significant in both the published psoriasis transcriptome and the CARD14 pathogenic keratinocyte signature (Table S4). These results indicate that some of the pathways upregulated in keratinocytes in which CARD14 harbors a pathogenic substitution are also upregulated in classic psoriatic skin. This suggests that altered keratinocyte activation might significantly contribute to the transcriptome signatures in classic psoriasis lesions.

**Discussion**

Here, we describe a spectrum of rare and common variation within CARD14, an activator of NF-κB16 in skin epidermis, and we demonstrate enrichment of rare variants in cases by using two independent statistical tests. The burden test, which performs a straightforward comparison of the number, or “burden,” of rare variants in cases and controls, provided a p value of 0.0015. The variable threshold test,19 which compares rare variants subject to a variable allele-frequency threshold in cases and controls, gave a p value of 0.0053. We also demonstrate that pathogenic alterations were enriched in the coiled-coil domain of CARD14. This domain is predicted to be involved in the oligomerization of CARD14 with other proteins and the formation of its active conformation.29,30 Interestingly, although some of the rare variants we identified have been annotated in dbSNP135, none of the putative pathogenic alterations are annotated. A coding polymorphism in CARD14, rs11652075 (p.Arg820Trp), and c.599G>A (p.Ser200Asn) were also associated with psoriasis in several large cohorts. In the two largest psoriasis cohorts, evidence of association between psoriasis and rs11652075 increased when rs11652075 was conditioned on PSORS1.

Two rare variants, c.424G>A (p.Glu142Lys) and c.425A>G (p.Glu142Gly), were identified in cases but not in controls and manifested as overtly causing disease. Compared with wild-type CARD14sh, they significantly enhanced NF-κB activation (4.03- and 5.00-fold enhancement, respectively), and they clustered with p.G117Ser and p.Glu138Ala after transfection into the keratinocyte line HEK 001 and global expression profiling. The c.424G>A (p.Glu142Lys) variant was identified in a Caucasian male who was diagnosed with psoriasis at 42 years of age and who responded well to treatment with UV light and a topical mixture of corticosteroid and a vitamin D analog. The c.425A>G (p.Glu142Gly) variant was found in a Caucasian male who was diagnosed with psoriasis in infancy and whose father also had psoriasis. He experienced a partial remission of psoriasis with methotrexate treatment. It is noteworthy that after these variants were stimulated with TNF-α, levels of NF-κB activation induced by these variants and the pustular-psoriasis substitution, p.Glu138Ala, decreased at the 24 hr mark. This suggests that at this time, downregulation of the NF-κB response might have been initiated in our cell-culture system, which merits further study. Both of these variants lie in the coiled-coil domain of CARD14, as does the de novo pustular-psoriasis substitution, p.Glu138Ala.

Compared with wild-type CARD14sh, a third variant, p.Asp176His (rs144475004), leads to enhanced NF-κB activation. However, its frequency was similar in cases and controls, and it didn’t increase NF-κB activity in vitro as much as other variants did. Hence, it might lie below the NF-κB-activation threshold required for disease. Alternatively, disease might require a specific stimulus or interaction with a second genetic factor. Other variants such as p.Arg38Cys and p.Ser200Asn exhibited significantly less NF-κB activation than did wild-type CARD14sh. Previous studies have shown that decreased activation of NF-κB, much like increased activation, can induce inflammation and epidermal hyperplasia.31,32 It might be interesting to examine clinical features, such as inflammation after skin wounding, of individuals with these variants. However, it should be noted that the p.Arg38Cys and p.Ser200Asn substitutions did not induce expression of the pathogenic psoriasis signature when transfected into keratinocytes. Thus, although we cannot completely rule out a role for these variants in some aspects of disease, they are neither...
overtly pathogenic nor likely to be causative for initiation of psoriasis.

Two variants, p.His171Asn and p.Arg179His, required stimulation with TNF-α to achieve maximal levels of NF-kB activation. The p.Arg179His substitution was observed in two unrelated cases from Toronto and one control. The cases included a female who was diagnosed with psoriasis at 40 years of age and a male who was diagnosed with psoriasis at 64 years of age and who had a family history of psoriasis. The female responded well to oral and topical steroids, but the male was not treated.

The p.His171Asn alteration was seen in two unrelated psoriasis- and psoriatic-arthritis-affected individuals from Newfoundland and was not seen in controls. One individual was diagnosed with psoriasis at 40 years of age and psoriatic arthritis at 41 years of age and had a family history of psoriasis. The second individual was diagnosed with psoriasis at 55 years of age after being diagnosed with psoriatic arthritis at 53 years of age. The identification of the c.511C>A (p.His171Asn) variant in only the Newfoundland population suggests that it arose as a result of a founder effect in this population, but confirming this will require further studies. Inspection of the genotypes of 13 polymorphisms in a 75 kb region harboring CARD14 revealed a shared haplotype that is common in cases and controls from the Newfoundland cohort and our other case/control cohorts from the United States (e.g., St. Louis/Dallas/UCSF, Michigan, and Utah). Therefore, further evidence of a founder effect for the p.His171Asn substitution seen in these two cases was not possible.

The altered transcriptome signature with pathologic substitutions included upregulation of psoriasis-specific transcripts SOD2, IL6, CSF2, IL8, MMP9, BRF2, CCL20, SLC7A2, OLR1, IL36G, GBP2, TNAIP2, and TNF. Expression of these molecules is expected to be an early event in the pathogenesis of psoriasis. Many of these genes have been implicated in immune-system development. However, BRF2 is implicated in the development of squamous cell carcinoma of the lung.33 This suggests that it might have global effects on the transcriptional profile of squamous cell epithelia in general and might help elicit a wound-healing or regenerative response in psoriatic keratinocytes.

Despite the dramatic effects of some CARD14 variants as keratinocyte transfectants, there was a wide range of phenotypes, even among individuals who carried the same substitution. This suggests that in many instances, the variable phenotypes are likely to be due to genetic background and/or environmental factors. For example, affected members of family PS1 all harbor the c.349G>A (p.Gly117Ser) mutation, but they have variable ages of onset (ranging from infancy to 83 years of age) and variable levels of disease severity, including the presence of psoriatic arthritis. Similarly, age of onset and response to treatment differed among individuals with the putative pathogenic variants from the coiled-coil domain (p.Glu142Lys, p.Glu142Gly, and p.Glu138Ala).

However, there might be some genotype-phenotype correlations because the pustular-psoriasis substitution, p.Glu138Ala, led to the most severe phenotype (in terms of both clinical presentation and increased NF-kB activation) relative to that produced by wild-type CARD14. The child with this alteration presented with a spectrum of plaque-type lesions, but she mostly presented with pustular lesions. This implies that some forms of plaque psoriasis might be pathogenetically linked to pustular psoriasis at the severe end of the disease spectrum. The child’s lesions also exhibited a pronounced infiltration of neutrophils. Interestingly, in keratinocyte transfectants with this CARD14 substitution, there was a higher level of IL8 than there was with other variants, which could lead to higher levels of neutrophil infiltration. The observation that the p.Glu138Ala alteration led to the most severe clinical phenotype and induced the greatest increase in NF-kB activation and upregulation of psoriasis-associated transcripts suggests that the phenotype of psoriasis could, in some cases, be predicted by the detection of pathogenically increased levels of NF-kB activation and signaling. How the CARD14 substitutions translate to variable levels of chemokine activation and how genetic background and environment trigger variable phenotypes are important areas for further study.

CARD14 missense variants rs11652075 (c.2458C>T [p.Arg820Trp]) and c.599G>A (p.Ser200Asn) were associated with psoriasis in cohorts of European ancestry. In the Asian cohort, the c.2458T>C polymorphism was also associated with psoriasis, but c.599G was monomorphic. CARD14 was not shown to be associated with psoriasis in a previous GWAS.8 However, none of the polymorphisms with evidence of association with psoriasis in this study were included in that GWAS and would not have been likely to exceed genome-wide significance if they had been. From the PSORS2 region of linkage, the SNP with the most significant p value in the GWAS was rs7216577 (p value = 0.00261).8 That SNP is located within an intron of SLC26A11 and might regulate levels of CARD14 mRNAs in the skin. However, rs11652075, although not on the Perlegen microarray in the CASP-GWAS,8 was part of the HapMap-based imputed dataset that was used for the association analyses in that study. Its association p value was of nominal significance (p = 0.039, OR = 1.10 for the C risk allele reported here) and would not have been reported because it was considerably below the threshold for genome-wide significant association.

PSORS2 was originally identified by linkage,21,34 and our studies with a common missense SNP (rs11652075) also indicate that it can be associated with psoriasis. Linkage and association have also been seen in some genes for other common diseases, e.g., NOD2 (nucleotide-binding oligomerization domain containing 2, [MIM 605956] in inflammatory bowel disease (IBD1 [MIM 266600])35–37 and CFH (complement factor H, [CFH (MIM 134370)]) in
Supplemental Data

Supplemental Data include two figures and three tables and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

This work was supported by grants from the National Institutes of Health (NIH): AR050266 and 5RCA1058681 (A.M.B.), T32AR007279 (E.D.O.R.), T32HL083822 and T32 GM07200 (C.E.J), AR060222 (M.A.L), T32HG000045 (C.E.J), and K08AR007763 (W.L.). J.T.E., R.N., and P.S. were supported by NIH grants R01 AR042742, R01 AR050511, and AR054966 and by the Babcock Memorial Trust. J.T.E. is supported by the Ann Arbor Veterans Affairs Hospital. R.G.M. is supported by the Intramural Research Programs of the National Institute of Arthritis and Musculoskeletal and Skin Diseases. Additional funding came from the Canadian Institute of Health Research, the Arthritis Society of Canada, the Atlantic Innovation Fund, the Canadian Institute of Health Research, the Arthritis Society, the Dana Foundation, and the Academia Sinica and National Science Council (National Clinical Core, National Genotyping Core) of Taiwan. The authors thank the many individuals with psoriasis and the controls who participated in this study. We thank Linus Schwantes-An, Weimin Duan, and Nancy Saccone for advice on statistical analyses. We also thank Mayte Suarez-Farinas for biostatistical assistance, Haoyan Chen, Olivia Lai, Fawnda Pellet, and M.J. Malloy for assembling the control population, Pui-Yan Kwok for sample procurement, Trilokraj Tejasvi for clinical evaluation of study subjects, Nikki Plass for patient scheduling and study logistics, Damaris Garcia for the management of samples, and Debbie Stone and Dawn Chapelle for patient care. Mike Lovett provided helpful comments on the manuscript. The authors are indebted to the National Psoriasis Foundation for continuing support during the course of this study.

Received: December 27, 2011
Revised: February 29, 2012
Accepted: March 13, 2012
Published online: April 19, 2012

Web Resources

The URLs for data presented herein are as follows:
1,000 Genomes Project, http://www.1000genomes.org
Ingenuity Pathway Analysis (IPA), http://www.ingenuity.com/
Microarray Gene Expression Data Society (MIAME), http://www.mged.org/Workgroups/MIAME/MIAME_checklist.html
Online Mendelian Inheritance in Man (OMIM), http://omim.org
PLINK, http://pngu.mgh.harvard.edu/purcell/plink/
PLINK/SEQ version 0.05, http://atgu.mgh.harvard.edu/plinkseq/PolyPhen-2.0, http://genetics.bwh.harvard.edu/pph2/
rmeta, http://cran.r-project.org/web/packages/rmeta/index.html
SIFT, http://sift.jcvi.org/

References


In two cohorts of European ancestry, evidence of association of psoriasis and rs11652075 increased when the PSORS1 variant HLA-Cw*0602 was included as a covariate, suggesting a genetic connection between PSORS1 and PSORS2. This variant resides in the CARD14 fl C-terminal GUK domain, which is predicted to relay external signals to the cellular milieu.16 This could explain the possible genetic connection because antigen stimulation via PSORS1 might increase the risk of psoriasis by upregulating signaling through the CARD14 pathway. This is consistent with the fact that CARD14 affects signaling downstream of antigen stimulation. However, it is important to note that the PSORS1 risk variant, HLA-Cw*060240, was not present in the affected members of the 17q-linked multiplex families, the pustular-psoriasis case described previously,17 or the cases with the p.Glu142Lys or p.Glu142Gly variants, indicating that rare CARD14 variants can be sufficient to lead to disease.

Our study illustrates some of the difficulties in searching for rare variants associated with common disease, even with an established gene. For example, it is not always easy to differentiate pathogenic rare variants from others when their numbers are very small. This also has an impact on the detection of gene-gene or gene-environment interactions. Moreover, it is sometimes necessary to recreate the cellular milieu (e.g., an inflammatory stimulus in the case of CARD14) when one attempts to differentiate disease-causing variants from neutral variants through functional studies. Nevertheless, our findings provide evidence that some rare CARD14 variants predispose to psoriasis and, possibly, to psoriatic arthritis, and they suggest that common variants in this region might also predispose to disease. They illustrate the challenges faced in identifying truly pathogenic rare variants in common disease and contribute to our understanding of the genetic basis of psoriasis.


Example qualifying exam (Ph.D.) and comprehensive exam (M.S.) questions pertaining to Albersen et al. (2012) PLoS ONE.

1. Why would investigators perform a GWAS of endometriosis when many previous candidate gene studies have been reported?

2. Surgically-confirmed cases of endometriosis were compared to controls. Given that endometriosis is a common gynecological condition and that diagnosis is typically delayed 7-10 years from onset, do you expect that some control participants may actually be undiagnosed cases? How would the presence of cryptic endometriosis cases in the control sample effect this study? Comment on how this scenario would impact the reported odds ratios.

3. Previous twin studies have estimated that the narrow-sense heritability of endometriosis is 51%, 75%, and 87%. Do you interpret these three estimates as consistent with each other? If so, explain why. If not, explain possible reasons for this inconsistency. What additional pieces of information could help you jointly interpret these results?

4. Genomic data were used to test for unknown relatedness among participants, and if detected, related samples were excluded from analysis. Why did the investigators do this?

5. Genomic data were used to determine genetic ancestry, and only samples of 95% or more European ancestry were included in statistical analysis. Why did the investigators do this?

6. Numerous genotyped and imputed SNPs in the chromosomal 1 region containing WNT4, CDC42, and HSPC157 showed evidence of genetic association (Figure 1). Based on these associated SNPs, is it possible to determine how many causal alleles are in this region? Are variants in all three genes causal? What additional information could help answers these questions?
Introduction

Endometriosis affects 5–10% of women in their reproductive years with symptoms including pelvic pain, dyspareunia, dysmenorrhea and infertility [1]. Although ectopic endometrium has been observed in female fetuses [2], symptoms of endometriosis usually don’t manifest until adolescence, and some patients with severe endometriosis remain asymptomatic. Definitive diagnosis is often delayed 7–10 years after the onset of symptoms severely impacting quality of life [3–5]. Family history of endometriosis has been reported in multiple studies to increased relative risk about a 5-fold OR = 1.19, Cl = 1.11–1.27). Consistent with previously suggested associations to WNT4 our study implicate a 150 kb region around WNT4 that also include LINC00339 and CDC42. A univariate analysis of documented infertility, age at menarche, and family history did not show allelic association with these SNP markers. Clinical data from patients in our study reveal an average delay in diagnosis of 8.4 years and confirm a strong correlation between endometriosis severity and infertility (n = 1182, P<0.001, OR = 2.18). This GWAS of endometriosis was conducted with high diagnostic certainty in cases, and with stringent handling of population substructure. Our findings broaden the understanding of the genetic factors that play a role in endometriosis.

Results and Discussion

GWAS and Replication

We conducted a discovery GWAS on surgically confirmed endometriosis patients and population controls using the Illumina OmniExpress BeadChip. SNPs were limited to the autosomes and endometriosis as summarized by Montgomery et al. [11] and Rahmioglu et al. [12], but the first strong evidence to date for genetic association are reported in two large Genome-Wide Association Studies (GWAS). In the first study Uno et al. [13] identified rs10965235 located in an intron of CDKN2BAS on chromosome 9p21 to be associated in a Japanese cohort, and in the second study Painter et al. [14] identified the intergenic SNP rs12700667 on chromosome 7p15.2 to be associated in a European cohort. A meta-analysis of the two studies extend this evidence and identify a total of seven loci associated with endometriosis [15]. To replicate and extend our understanding of the genetic factors that contribute to endometriosis we have undertaken a large two-stage GWAS in a European cohort.
SNPs with an Illumina Genotyp score $\geq 0.65$. We further eliminated SNPs with calibrate $< 0.98$, Hardy-Weinberg Equilibrium (hwe) $< 0.001$ and minor allele frequencies (MAF) $< 0.01$. After filtering 580,699 SNPs remained. Next, samples with calibrate $< 0.98$ were eliminated. The remaining samples were tested for unknown familial relationships using genome-wide identity-by-state (IBS), and samples closer than 3rd-degree ($r > 0.2$) were removed. We used ADMIXTURE (ver. 1.22) [16] to estimate individual ancestry proportions based on a subset of SNPs on the Illumina OmniExpress chip (see Materials and Methods) and restricted our analysis to samples with $\geq 95\%$ European ancestry. The calculated ancestral distribution of samples within Europe is shown in Figure S1. After applying quality, relatedness and ethnicity filters 1,514 case and 12,660 control samples were used for the discovery phase of the association analysis. The genomic inflation factor lambda ($\lambda$) was determined to be 1.18, indicating measurable population stratification across the samples. To account for the elevated $\lambda$, we performed a PCA adjusted association analysis that resulted in a $\lambda$-value of 1.05 shown in QQ-plots in Figure S2. We selected the top 100 SNPs with the lowest PCA-adjusted $P$-values (ranging between $8.20 \times 10^{-5}$ and $1.36 \times 10^{-7}$) for further association analysis in the replication stage (Table S1).

The replication samples included 505 cases and 1811 controls selected for the same criteria as the discovery set. The $\lambda$-value for the replication cohort was determined to be 1.01 suggesting no measurable population stratification. After applying the same SNP filters as above we analyzed the top 100 SNPs from the discovery GWAS in the replication set. A significance threshold for the study, allowing for multiple correction, was chosen at 4.51 $\times 10^{-7}$ (0.05/189,699; 189,699 being the number of independent SNPs in the panel of 580,699 filtered SNPs with $r^2 > 0.20$). A meta-analysis of the discovery and replication results was performed using Cochran-Mantel-Hanzel test and revealed 8 SNPs from 4 genomic regions that passed our genome-wide significance threshold including: LINC00339-WNT4 on 1p36.12 (rs2235529; $P_{meta} = 3.05 \times 10^{-9}$, OR = 1.30); RND3-RBM43 on 2q23.3 (rs6757804; $P_{meta} = 6.45 \times 10^{-8}$, OR = 1.20); RNF141-BID4 on 6p22.3 (rs6907340; $P_{meta} = 2.19 \times 10^{-7}$, OR = 1.20); and HNRNPA3P1-LOC100130539 on 10q11.21 (rs10508881; $P_{meta} = 4.08 \times 10^{-7}$, OR = 1.19) and shown in detail in Table 1. Table 1 also shows that three SNPs (rs2235529, rs1519761 and rs6757804) pass a conventional genome wide significance threshold of $P < 5 \times 10^{-5}$ in the combined analysis. There was no evidence of heterogeneity between the discovery and replication datasets as judged by the Breslow Day test and shown in Table 1. A second group of 15 SNPs from nine genomic regions that show suggestive replication provide additional candidate loci for endometriosis that merit further investigation. The most significant of these regions encompass IL33 on 9p24.1 (rs10975359; $P_{meta} = 9.26 \times 10^{-7}$, OR = 1.19). IL33 is a chemokine that has been linked to deep infiltrating endometriosis [17]. A summary of the discovery and replication GWAS, together with the meta and combined analysis for all 100 SNPs is presented in Table S1.

To further characterize the signals from the four most strongly associated regions and the IL33 region, we performed imputation using 1000-Genome and dataset utilizing IMPUTE2 (ver. 2.2.2) [18]. The results from the imputed dataset for the WNT4 region is shown in Table S2a. Figure 1 show a regional association plot around WNT4 and reveal that the associated region extends across 150 kb and include two new genes, HSPC157 (gene symbol: LINC00339) and CDC42, in addition to WNT4. The imputation results identify 25 additional SNPs that are strongly associated to endometriosis and reveal that the imputed SNP, rs10917151, is the most strongly associated SNP in the region ($P_{imputed} = 5.63 \times 10^{-10}$, OR = 1.3). WNT4 is important for steroidogenesis, ovarian follicle development and the development of the female reproductive tract and a very plausible candidate for endometriosis based on its biological functions [19]. Cell division cycle 42, CDC42, is a small GTPase of the Rho-subfamily, which regulates signaling pathways that control diverse cellular functions including cell morphology, migration, endocytosis and cell cycle progression. CDC42 is in part regulated by estrogen, expressed in endometrium, and has been shown to be differentially expressed in endometriosis [20]. HSPC157 is an alias for the Long Intergenic Non-protein Coding RNA 339 (gene symbol: LINC00339). HSPC157 has been found to be differentially expressed in endometriosis versus autologous uterine endometrium [21]. Based on this biological evidence both CDC42 and LINC00339 must also be considered candidates for endometriosis.

WNT4 has previously been associated with endometriosis, first by Uno et al. [2010] who noted that rs16826658, approximately 16 kb upstream of WNT4, showed a possible association to endometriosis ($P = 1.66 \times 10^{-5}$, OR = 1.20) in a Japanese population, and again by Painter et al. [2011] who reported that rs7210902, located approximately 22 kb upstream of WNT4, also showed evidence for association in a European cohort ($P = 9.0 \times 10^{-5}$, OR = 1.16). Our imputation analysis replicate the association of rs7210902 with endometriosis ($P = 6.4 \times 10^{-5}$, OR = 1.17) and confirm the involvement of the WNT4 region in the pathogenesis of endometriosis. We only found weak evidence of association with rs16826658 ($P = 0.05$, OR = 1.07), because the minor allele is very common and because of the different ethnic backgrounds between the studies. To further evaluate the signals from the WNT4 region, we performed a haplotype analysis of three key SNPs from our study (rs10917151, rs4654783 and rs2235529) together with rs16826658, and rs7210902 using the imputed data from our population. The haplotype-results are summarized in Table S3 and show that the risk for endometriosis is confined to a single haplotype anchored in rs10917151, rs4654783 and rs2235529 ($HAP_{1} = 7.26 \times 0.9$, OR$_{HAP-1} = 1.28$), and that this haplotype starts to deteriorate with the addition of rs16826658 and rs7210902 ($HAP_{12} = 8.33 \times 0.7$, OR$_{HAP-12} = 1.25$). This analysis suggests that the risk allele observed in the present study and the two previously published reports is located on the same ancestral haplotype. Imputation was also performed on the three other significant regions on chromosome 2, 6, and 10, and on the region surrounding IL33 on chromosome 9 (Table S2a-c, Figure S3).

Overlap with Other Reported Loci

Uno et al. [13] reported association between endometriosis and rs10965235, located in intron 19 of CDKN2BAS, in a Japanese population. The protective minor allele A, that has a minor allele frequency of 0.20 in the Japanese population, is not observed in the European population preventing a direct comparison between the two ethnic groups. In lieu of a direct comparison between the two ethnic groups we scanned 66 SNPs from a region of 200 kb surrounding rs10965235 in our European population. After correcting for multiple-testing ($P$-value $\leq 0.05/66 = 0.00076$) we didn’t find any evidence that CDKN2BAS is associated with endometriosis in the European population (Table S4a). In contrast, a second SNP (rs13271465) located on 8p22 between MTMR7 and SLC7A2, that Uno et al. identified as being associated with endometriosis ($P_{combined} = 9.84 \times 10^{-6}$, OR = 1.18), is present in both studies, show weak association ($P = 0.0057$, OR = 1.14) in our study, while none of the other 88 SNPs from the 200 kb region surrounding rs13271465 showed
any evidence for association with endometriosis (Table S4b). Uno et al. provided a supplementary list of 100 additional candidate SNPs from their study of which 60 are present in our analysis. Among the 60 SNPs only rs2473277 has a P-value, 0.001 in our study (P = 5.97 × 10⁻²⁶, OR = 1.17). SNP rs2473277 is located between LINC00339 and CDC42 at the left-most boundary of the WNT4 LD-block discussed above (Figure 1), and the risk allele of rs2473277 tag perfectly with the risk haplotype HAP-1 shown in Table S3 (data not shown).

Painter et al. [14] reported significant association of moderate and severe endometriosis with rs12700667 on 7p15.2 (Pall = 2.6 × 10⁻⁷, OR = 1.22), with flanking support from

<table>
<thead>
<tr>
<th>SNP</th>
<th>gene</th>
<th>Chr</th>
<th>Pos</th>
<th>allele</th>
<th>stage</th>
<th>Case MAF</th>
<th>Control MAF</th>
<th>OR</th>
<th>95%CI</th>
<th>Phet</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4654783</td>
<td>WNT4</td>
<td>1</td>
<td>22,439,520</td>
<td>a/g</td>
<td>Discovery</td>
<td>0.34</td>
<td>0.295</td>
<td>2.43E⁻⁰⁷</td>
<td>1.23</td>
<td>1.14–1.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.323</td>
<td>0.298</td>
<td>1.31E⁻⁰¹</td>
<td>1.13</td>
<td>0.97–1.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>1.40E⁻⁰⁷</td>
<td></td>
<td>1.21</td>
<td>1.13–1.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.336</td>
<td>0.295</td>
<td>1.17E⁻⁰⁷</td>
<td>1.21</td>
<td>1.13–1.29</td>
</tr>
<tr>
<td>rs2235529</td>
<td>WNT4</td>
<td>1</td>
<td>22,450,487</td>
<td>a/g</td>
<td>Discovery</td>
<td>0.188</td>
<td>0.153</td>
<td>1.36E⁻⁰⁷</td>
<td>1.28</td>
<td>1.16–1.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.182</td>
<td>0.142</td>
<td>1.38E⁻⁰³</td>
<td>1.36</td>
<td>1.13–1.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>3.05E⁻⁰⁹</td>
<td></td>
<td>1.3</td>
<td>1.19–1.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.186</td>
<td>0.151</td>
<td>8.65E⁻⁰⁹</td>
<td>1.29</td>
<td>1.18–1.40</td>
</tr>
<tr>
<td>rs1519754</td>
<td>RND3</td>
<td>RBM43</td>
<td>2</td>
<td>151,619,693</td>
<td>c/a</td>
<td>Discovery</td>
<td>0.446</td>
<td>0.403</td>
<td>5.67E⁻⁰⁵</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.45</td>
<td>0.405</td>
<td>9.63E⁻⁰³</td>
<td>1.2</td>
<td>1.04–1.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>1.75E⁻⁰⁷</td>
<td></td>
<td>1.2</td>
<td>1.14–1.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.447</td>
<td>0.403</td>
<td>1.15E⁻⁰⁷</td>
<td>1.2</td>
<td>1.12–1.28</td>
</tr>
<tr>
<td>rs6734792</td>
<td>RND3</td>
<td>RBM43</td>
<td>2</td>
<td>151,624,882</td>
<td>g/a</td>
<td>Discovery</td>
<td>0.448</td>
<td>0.404</td>
<td>3.52E⁻⁰⁵</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.453</td>
<td>0.406</td>
<td>7.48E⁻⁰³</td>
<td>1.2</td>
<td>1.05–1.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>8.18E⁻⁰⁸</td>
<td></td>
<td>1.2</td>
<td>1.12–1.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.449</td>
<td>0.404</td>
<td>5.19E⁻⁰⁸</td>
<td>1.2</td>
<td>1.12–1.28</td>
</tr>
<tr>
<td>rs1519761</td>
<td>RND3</td>
<td>RBM43</td>
<td>2</td>
<td>151,633,204</td>
<td>g/a</td>
<td>Discovery</td>
<td>0.445</td>
<td>0.401</td>
<td>3.54E⁻⁰⁵</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.452</td>
<td>0.403</td>
<td>5.85E⁻⁰³</td>
<td>1.2</td>
<td>1.05–1.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>7.30E⁻⁰⁸</td>
<td></td>
<td>1.2</td>
<td>1.12–1.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.447</td>
<td>0.401</td>
<td>4.70E⁻⁰⁸</td>
<td>1.2</td>
<td>1.13–1.29</td>
</tr>
<tr>
<td>rs6757804</td>
<td>RND3</td>
<td>RBM43</td>
<td>2</td>
<td>151,635,832</td>
<td>g/a</td>
<td>Discovery</td>
<td>0.445</td>
<td>0.401</td>
<td>3.43E⁻⁰⁵</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.452</td>
<td>0.403</td>
<td>5.44E⁻⁰³</td>
<td>1.2</td>
<td>1.06–1.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>6.45E⁻⁰⁸</td>
<td></td>
<td>1.2</td>
<td>1.13–1.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.446</td>
<td>0.401¹¹</td>
<td>4.05E⁻⁰⁸</td>
<td>1.2</td>
<td>1.13–1.29</td>
</tr>
<tr>
<td>rs6907340</td>
<td>RNF144B</td>
<td>ID4</td>
<td>6</td>
<td>19,803,768</td>
<td>a/g</td>
<td>Discovery</td>
<td>0.417</td>
<td>0.371</td>
<td>5.49E⁻⁰⁶</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.412</td>
<td>0.378</td>
<td>4.54E⁻⁰²</td>
<td>1.15</td>
<td>1.00–1.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>2.19E⁻⁰⁷</td>
<td></td>
<td>1.2</td>
<td>1.12–1.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.415</td>
<td>0.372</td>
<td>1.25E⁻⁰⁷</td>
<td>1.2</td>
<td>1.12–1.28</td>
</tr>
<tr>
<td>rs10508881</td>
<td>HNRNPA3P1</td>
<td>LOC100130539</td>
<td>10</td>
<td>44,541,565</td>
<td>a/g</td>
<td>Discovery</td>
<td>0.45</td>
<td>0.405</td>
<td>3.18E⁻⁰⁵</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>0.42</td>
<td>0.387</td>
<td>6.06E⁻⁰²</td>
<td>1.15</td>
<td>1.00–1.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meta⁵</td>
<td></td>
<td>4.08E⁻⁰⁷</td>
<td></td>
<td>1.19</td>
<td>1.11–1.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined Trend⁶</td>
<td>0.442</td>
<td>0.403</td>
<td>1.57E⁻⁰⁶</td>
<td>1.18</td>
<td>1.10–1.26</td>
</tr>
</tbody>
</table>

The discovery stage included 1,514 endometriosis cases and 12,660 population controls, and the replication stage included 505 cases and 1,811 controls.

The P-values were determined using the Cochrane-Armitage trend test. P-values for the Discovery set reflect PCA adjusted P trend values.

Odds-ratios (OR) and confidence intervals (CI) are calculated using the non-risk allele as the reference.

The Meta analysis was performed using Cochran-Mantel-Haenzel statistics.

P values of heterogeneities (Phet) across discovery and replication stages calculated using Breslow-Day Test.

Cochrane-Armitage trend test P-values based on the combined genotypes from the Discovery and Replication data.

Significance threshold is 4.59 × 10⁻⁷, and is determined by 0.05/108,699 where 108,699 is the number of independent SNPs in the panel with r² less than 0.20.

doi:10.1371/journal.pone.0058257.t001

PLOS ONE | www.plosone.org

March 2013 | Volume 8 | Issue 3 | e58257
rs7798431 located 41 kb away. The two SNPs were reported to be in strong linkage disequilibrium (LD), but unfortunately neither SNP is included in our study. A review of the Hapmap3 data from the region show that three SNPs in our study (rs12535837, rs10282436, rs10232819) are in moderate to strong LD with rs12700667 and rs7798431, but we find no evidence for association between endometriosis and any of these markers in our analyses of all endometriosis cases together nor do we find any evidence for association to the moderate and severe subset (Table S4c). A broader scan of the 200 kb region surrounding rs12700667, suggests weak association with rs4722551 (P = 0.000867) about 90 kb downstream of rs12700667 (Table S4c). Painter et al. provided a supplementary list of 73 candidate SNPs, but none of the thirty-nine SNPs present in our study reach a P-value threshold <0.001.

A recent meta-analysis published by Nyholt et al. [15] extend the findings by Uno et al. and Painter et al. and report a total of seven SNPs that pass a genome-wide significance (P<5 x 10^-8). A comparison of our results to each of the seven loci show evidence for association to endometriosis for three of the seven loci as detailed in Table S7.

A recent candidate gene study investigating the LCP6 variant (rs61764370) in the 3'-UTR of KRAS showed very strong association with endometriosis among 132 women (MAF-case = 0.311, MAF-control = 0.076) [22], and proposed that let-7 microRNA play a functional role in the development of endometriosis. We investigated this association by Taqman assay in a set of 1123 cases and 832 controls from our European population and found the allele frequencies to be identical in the two populations (MAF-case = 0.095, MAF-control = 0.093). Our result show that the LCP6 variant of KRAS does not provide any value as an indicator for endometriosis risk in a European population in agreement with results reported by Luong et al. [23].

ENDO1 is a susceptibility locus on chromosome 10q26 (OMIM phenotype number 131200) identified by linkage analysis [24], but we find no evidence for SNP association (P-value >0.0001) in this 16 Mb region (data not shown).

Clinical Stratification and Diagnostic Delay

Clinical features commonly used to characterize and stratify endometriosis include infertility, pelvic pain, severity, age-at-menarche and familiality. To determine the diagnostic delay in our patient-population we identified a group of women (n = 874) that reported both age at onset-of-symptoms (mean-age-onset = 19.04 years) and age at diagnosis (mean-age-diagnosis = 27.49), and observed an average diagnostic delay of 8.44 years, similar to previous studies. We then went on to examine if our samples showed any clinical correlations using logistic regression. The analysis revealed strong correlations between severity and infertility (P<0.001, OR = 2.19), and between severity and diagnostic delay (P<0.001, OR = 1.04) as shown.
in Table 2. To identify loci associated with the progression of endometriosis, we compared patients with mild endometriosis to patients with moderate or severe endometriosis in a two-stage GWAS. Stage one included 657 patients with mild endometriosis vs. 523 patients with moderate-or-severe endometriosis and a stage two replication set of 318 mild vs. 519 moderate-or-severe patients. A meta-analysis using the CMH test in this limited sample set, found no loci that pass the genome-wide significant threshold which suggest the SNPs identified in the primary study contribute to the general endometriosis risk rather than endometriosis progression. A separate analysis of the top five SNPs using logistic regression also showed no correlation with severity as shown in Table S5a, and Table S5b shows no noticeable increase in effect size when comparing moderate and severe disease against all controls.

Risk Analysis of Endometriosis

After removing markers with $r^2$ > 0.8 among the top 5 associated regions (incl. the IL13 locus), we conducted multivariate logistic regression using the combined set of 2,019 cases and 14,471 controls. All of the 5 SNPs rs101917151, rs6757804, rs6907340, rs10975519 and rs10508881 analyzed remained significant with OR of 1.3, 1.2, 1.18, 1.17 and 1.17. Each marker appear to be an independent risk factors for endometriosis. Comparison of the OR between the discovery and replication datasets, shown in Table 1, does not suggest any significant inflation of effect size in the discovery dataset (winner’s curse), but this conclusion remain tentative due to the difference in size between the two datasets.

Conclusion

A two-stage GWAS and a replication study involving 2,019 cases and 14,471 controls was performed which identified four novel loci strongly associated with endometriosis and confirmed the involvement of a region around WNT4 which previously have been suggested as being associated to endometriosis. Nine other regions identified in the study also hold promise as candidate loci for endometriosis. Utmost care was taken in the clinical classification of patients and only surgically-confirmed cases with >95% European ancestry were considered in this large GWAS of endometriosis. The study is well powered (>90%) to identify a marker at or above 10% minor allele frequency (MAF) with odds-ratio (OR) > 1.20, but we estimate the top 5 loci only explain about 1.5% of the phenotypic variance of endometriosis. Since the few risk loci we detected all have odds ratios < 1.30 it must be assumed that any new endometriosis loci that contribute to the “missing heritability” must be rare, recent, or show minimal effect. GWAS, by design, detects only very old founder effects. When a phenotype includes infertility, like endometriosis, a high mutation rate would be required to replenish the disease-causing alleles lost from the gene-pool due to infertility. One suitable avenue to investigate under that scenario is to use whole genome sequencing of high-risk families rather than SNP-based GWAS. Little is presently known about the pathophysiology of endometriosis, but we hope that a more detailed investigation of the loci presented in this paper will help elucidate the pathogenesis of endometriosis and clarify its genetic underpinnings.

Materials and Methods

Ethics Statement

All subjects and controls provided written informed consent in accordance with study protocols approved by Quorum Review IRB (Seattle, WA 98101).

Participant Recruitment

Patients included in the present study were invited to participate via an outreach program at www.endoendo.com, where our research initiative is described in more detail. Briefly, the “End to Endo” website provides general information regarding endometriosis and our research project, and invites women diagnosed with endometriosis to participate in our study.

Medical Review

The inclusion criteria in the endometriosis case population in the present study is surgically confirmed diagnosis of endometriosis with laparoscopy being the preferred method. Trained OB/GYN clinicians performed the medical record review and clinical assessment of each individual patient. Patients were considered to be affected if they had biopsy-proven lesions or if operative reports revealed unambiguous gross lesions. Patients were further categorized by severity, clinical history of pelvic pain, infertility, dyspareunia or dysmenorrhea and family history. Patients were grouped into one of three classes of severity: mild, moderate or severe, following the general guidelines set forth by ASRM [25]. Exclusion of endometriosis also requires surgical intervention and we made no attempt to exclude endometriosis in the population controls. Thus, in this analysis we are comparing cases with 100% prevalence of endometriosis to controls with the population prevalence of endometriosis (5–10%), which leads to a systematical underestimation of the true odds ratios and a decrease in statistical power to detect associations.

DNA Extraction

Saliva samples were collected using the Oragene 300 saliva collection kit (DNA Genotek; Ottawa, Ontario, Canada) and

<table>
<thead>
<tr>
<th>Clinical Feature</th>
<th>Moderate or Severe endometriosis (n = 842)</th>
<th>Mild endometriosis (n = 1177)</th>
<th>Category</th>
<th>OR</th>
<th>Beta</th>
<th>SE</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infertility (182)</td>
<td>525</td>
<td>657</td>
<td>Yes or No</td>
<td>2.19</td>
<td>0.78</td>
<td>0.12</td>
<td>8.52E-11</td>
</tr>
<tr>
<td>Family History (1881)</td>
<td>790</td>
<td>1091</td>
<td>Yes or No</td>
<td>0.89</td>
<td>-0.11</td>
<td>0.09</td>
<td>0.23</td>
</tr>
<tr>
<td>Age at Menarche (921)</td>
<td>405</td>
<td>516</td>
<td>&lt;12 or &gt;12 yrs</td>
<td>1.20</td>
<td>0.18</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>Diagnostic Delay (874)</td>
<td>383</td>
<td>491</td>
<td>0 to 35</td>
<td>1.04</td>
<td>0.04</td>
<td>0.01</td>
<td>2.18E-05</td>
</tr>
</tbody>
</table>

Clinical features were correlated to severity. Only patients that could be unambiguously categorized were included in the analysis with total counts provided in parenthesis next to the clinical feature. P-values (P) are calculated using Wald test. Beta is the regression coefficients and SE the standard error from logistic regression. doi:10.1371/journal.pone.0058257.t002
DNA was extracted using an automated extraction instrument, AutoPure LS (Qiagen; Valencia, CA), and manufacturer’s reagents and protocols. DNA quality was evaluated by calculation absorbance ratio OD$_{260}$/OD$_{280}$, and DNA quantification was measured using PicoGreen® (Life Technologies; Grand Island, NY).

**Microarray Genotyping**

The discovery set of 1514 cases and 12660 controls and replication set of 505 cases and 1011 controls were genotyped using the Illumina Human OmniExpress Chip (Illumina; San Diego, CA) according to protocols provided by the manufacture. Figure S4 show the genotype clusters for the top eight SNPs in our study. All SNPs reported in the present study passed visual inspection for cluster quality. It is our experience that technical replication does not affect genotype calls of SNPs with high quality clusters and due to cost we could not justify independent technical replication.

**Taqman Genotyping**

A Taqman® 7900 instrument (Life Technologies; Grand Island, NY) and manufacturer’s protocols were used to genotype rs1764570. Genotypes were determined using Taqman genotyping software SDS (v2.3) and the genotype cluster was visually inspected. Genotyping QC for rs1764570 passed standard criteria of call rate >95% and no deviation from HWE (p<0.001) was observed.

**Sample Quality Control**

Samples were excluded from the analysis if they missed any of the following quality thresholds:

a) Evidence of familial relationship closer than 3rd-degree (p>0.2) using genome-wide Identity-By-State (IBS) estimation implemented in PLINK
b) Samples with missing genotypes >0.02
c) Samples with non-European admixture >0.05 as determined by ADMIXTURE

**SNP Quality Control**

SNPs were excluded from the analysis if they missed any of the following quality thresholds:

a) SNPs with Illumina GenTrain Score <0.65
b) SNPs from copy number variant regions or regions with adjacent SNPs
c) SNPs failing Hardy-Weinberg Equilibrium (HWE) P≤10$^{-3}$
d) SNPs with minor allele frequency (MAF) ≤0.01 in the control population
e) SNP call rate ≤98%

**Admixture**

ADMIIXTURE (ver. 1.22) was used to estimate the individual ancestry proportion [16]. The software estimates the relative admixture proportions of a given number of a priori defined ancestral groups contributing to the genome of each individual. We used the POPRES dataset [26] as a reference group to create a supervised set of 9 ancestral clusters. Seven of them belong to the European subgroups along with African and Asian groups. Since POPRES dataset utilized Affymetrix 5.0 chip, we used 105,079 autosomal SNPs that overlapped with the Illumina OmniExpress dataset. Among the 105,079 SNPs we selected a subset of 33,067 SNPs that showed greater genetic variation (absolute difference in frequency) among the 9 reference groups. The pair-wise autosomal genetic distance determined by Fixation Index (F$_{st}$) using 33,067 SNPs was calculated for the 9 reference groups and show in Table S6 [27]. Subsequently, a conditional test was used to estimate the admixture proportions in the unknown samples as described by Alexander et al. (2009).

**Principal Component Analysis (PCA)**

PCA was applied to account for population stratification among the European subgroups. We selected the previously identified 33,067 SNPs to infer the axes of variation using EIGENSTRAT [28]. Only the top 10 eigenvectors were analyzed. Most of the variance among the European populations was observed in the first and second eigenvector. The first eigenvector accounts for the east-west European geographical variation while the second accounts for the north-south component. Only the top 10 eigenvectors showed population differences using Anova statistics (p<0.01). We then calculated the PCA adjusted Armitrage trend P-values using the top 10 eigenvectors as covariates.

**Power Analysis**

Power calculations were performed using QUANTO (ver. 1.2), using a log-additive model. The analysis included 2019 cases and 14471 controls with the following assumptions: Type I error = 0.05, a minor allele frequency =0.10 and the odds-ratio =1.2.

**Association Analysis**

After the quality of all data was confirmed for accuracy, genetic association was determined using the whole-genome association analysis toolset, PLINK (ver. 1.07) [29].

Differences in allele frequencies between endometriosis patients and population controls were tested for each SNP by a 1-degree-of-freedom Cochran-Armitrage Trend test.

The allelic odds ratios were calculated with a confidence interval of 95%. SNPs that passed the quality control parameters were used to calculate the genomic inflation factor ($\lambda$) as well as to generate Quantile-Quantile (QQ) plots (Figure S2), which were generated by ranking a set of –log$_{10}$ P-values and plotting them against their expected values. PCA adjusted Cochran-Armitrage trend test P-values were also determined. The combined/meta-analysis of discovery and replication dataset was performed using Cochran-Mantel-Hanzsel method. Breslow Day test was used to determine between-cluster heterogeneity in the odds ratio for the disease/SNP association. Multivariate Logistic regression was used to test for independence of SNP effects. Univariate Logistic regression was used to test for correlation of clinical factors to the severity of the disease.

Control samples include both male and female samples in approximately equal proportions. The allele frequencies for the 8 strongly associated SNPs and the 15 SNPs with suggested associations did not show any significant gender bias.

Haplotype-based association tests were calculated by 1-degree of freedom $\chi^2$-test, along with their respective odds ratios using PLINK.

The variance explained by logistic regression model is calculated using the Cox Snell and Nagelkerke pseudo $R^2$ method which is similar to the $R^2$ concept of linear regression [30].

**Imputation Analysis**

IMPUTE2 (ver. 2.2.2) was used for imputing SNPs against the 1000-Genome (version 3 of the Phase 1 integrated data). Samples
were pre-phased with IMPUTE2 using actual genotypes and then imputed for SNPs included in the 1000-Genome reference panel to form imputed haplotypes. Imputation was carried out within +/- 250 kb of the main marker of interest. Only SNPs that pass the confidence score of $> 0.9$ from imputation, call rate of 0.95 and with MAF $> 0.01$ are reported. The imputation was performed on the total dataset of 2,019 cases and 14,471 control subjects.

Software Used
PLINK (version 1.07; http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml).
QUANTO (version 1.2; http://hydra.usc.edu/gxE).
R (version 2.15.0; http://www.r-project.org/).
Impute2 (version 2.2.2; http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) and haplotype analysis was performed using PLINK (ver. 1.07).
LocusZoom (version 1.1; http://csg.sph.umich.edu/locuszoom/) was used for regional association plots.
EIGENSTRAT (version 3.0; http://genepath.med.harvard.edu/reich/Software.htm).

Supporting Information
Figure S1 PCA classification of Case and Control samples. A reference set of samples previously identified as European are shown in Panel A (reference). Samples selected for analysis include 580,699 SNPs and included only samples that passed our Ethnicity, SNP and Sample quality filters.

Figure S2 Quantile-quantile plots for the Discovery set of 1,514 endometriosis cases and 12,660 population controls before and after PCA-based adjustment. The unadjusted QQ plot in Panel A is showing a $\lambda = 1.18$. The adjusted QQ plot in Panel B show $\lambda = 1.05$. The association analysis include 580,699 SNPs and included only samples that passed our Ethnicity, SNP and Sample quality filters.

Figure S3 Regional association plots for the five top regions (Panel A–E).

Figure S4 Genotype clusters for the 8 most strongly associated SNPs.

References

Table S1  Top 100 SNPs from Discovery and Replication GWAS.

Table S2 a Genotyped and Imputed P-values for the LINC00339-WNT4 region on 1p36.12. b Genotyped and Imputed P-values for the RND3-RBM43 region on 2q23.3. c Genotyped and Imputed P-values for the RNF144B-ID4 region on 6p22.3. d Genotyped and Imputed P-values for the IL33-TPD52L3 region on 3p24.1. d Genotyped and Imputed P-values for the HNRNPA3P1-LOC100130539 region on 10q11.21.

Table S3  Haplotype analysis for the WNT4-region.

Table S4  No association with CDKN2BAS on 9p21 in Europeans. a Tentative replication of rs13271465 located on 8p22 between MTMR7 and SLC7A2. c SNP rs12700667 fails replication.

Table S5 a Severity of endometriosis is independent of the five top loci by logistic regression analysis. b Severity of endometriosis is independent of the top five loci by association analysis.

Table S6  Pair-wise autosomal genetic distance among ethnic groups as measured by the Fixation Index ($F_{ST}$).

Table S7 Support for endometriosis association in a Caucasian cohort found at 3 of 7 loci reported by Nyholt et al.

Acknowledgments
We would like to extend our sincere gratitude to all our study participants. It is because of them that this study has been possible. We also extend our thanks to our colleagues in Juneau’s Family Studies Group and our Laboratory for their untiring efforts on this endometriosis project.

Author Contributions
Conceived and designed the experiments: HA RC KW. Performed the experiments: HA RC. Analyzed the data: HA RC PF KW. Contributed reagents/materials/analysis tools: KW. Wrote the paper: HA RC KW.