Crohn's Disease Risk Alleles on the *NOD2* Locus Have Been Maintained by Natural Selection on Standing Variation

Shigeki Nakagome,[†,1,2] Shuhei Mano,[3,4] Lukasz Kozlowski,[5] Janusz M. Bujnicki,[5,6] Hiroki Shibata,[7] Yasuaki Fukumaki,[7] Judith R. Kidd,[8] Kenneth K. Kidd,[8] Shoji Kawamura,[1] and Hiroki Oota*[,1,2]

[1]Department of Integrated Biosciences, Graduate School of Frontier Sciences, University of Tokyo, Kashiwanoha, Kashiwa, Chiba, Japan

[2]Department of Anatomy, Kitasato University School of Medicine, Kitasato, Minami-ku, Sagamihara, Kanagawa, Japan

[3]Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Midori-cho, Tachikawa, Tokyo, Japan

[4]Japan Science and Technology Agency, Honcho, Kawaguchi-shi, Saitama, Japan

[5]International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

[6]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland

[7]Division of Human Molecular Genetics, Research Center for Genetics Information, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan

[8]Department of Genetics, Yale University School of Medicine, New Haven, CT

†Present address: Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Midori-cho, Tachikawa, Tokyo, Japan

***Corresponding author:** E-mail address: hiroki_oota@med.kitasato-u.ac.jp.

Associate editor: Noah Rosenberg

Abstract

Risk alleles for complex diseases are widely spread throughout human populations. However, little is known about the geographic distribution and frequencies of risk alleles, which may contribute to differences in disease susceptibility and prevalence among populations. Here, we focus on Crohn's disease (CD) as a model for the evolutionary study of complex disease alleles. Recent genome-wide association studies and classical linkage analyses have identified more than 70 susceptible genomic regions for CD in Europeans, but only a few have been confirmed in non-European populations. Our analysis of eight European-specific susceptibility genes using HapMap data shows that at the *NOD2* locus the CD-risk alleles are linked with a haplotype specific to CEU at a frequency that is significantly higher compared with the entire genome. We subsequently examined nine global populations and found that the CD-risk alleles spread through hitchhiking with a high-frequency haplotype (H1) exclusive to Europeans. To examine the neutrality of *NOD2*, we performed phylogenetic network analyses, coalescent simulation, protein structural prediction, characterization of mutation patterns, and estimations of population growth and time to most recent common ancestor (TMRCA). We found that while H1 was significantly prevalent in European populations, the H1 TMRCA predated human migration out of Africa. H1 is likely to have undergone negative selection because 1) the root of H1 genealogy is defined by a preexisting amino acid substitution that causes serious conformational changes to the NOD2 protein, 2) the haplotype has almost become extinct in Africa, and 3) the haplotype has not been affected by the recent European expansion reflected in the other haplotypes. Nevertheless, H1 has survived in European populations, suggesting that the haplotype is advantageous to this group. We propose that several CD-risk alleles, which destabilize and disrupt the NOD2 protein, have been maintained by natural selection on standing variation because the deleterious haplotype of *NOD2* is advantageous in diploid individuals due to heterozygote advantage and/or intergenic interactions.

Key words: Crohn's disease, *NOD2*, hitchhiking effect, natural selection, standing variation, mildly deleterious mutation.

Introduction

It is of great interest how complex disease-associated mutations are maintained in the human gene pool (Pritchard 2001; Di Rienzo and Hudson 2005; Di Rienzo 2006). Many known complex disease alleles are subsets of genetic variants that are spread over geographically separated populations (Myles et al. 2008). The frequencies of mildly deleterious mutations that contribute to complex diseases are predicted to fluctuate in the manner of neutral evolution when the effective population size is small (Ohta 1973). Population genetics studies on genome-wide polymorphism data have revealed that the frequencies of deleterious mutations vary due to random genetic drift following human demographic history, the tract of which is typically shown in the Out-of-Africa expansion model (Lohmueller et al. 2008). However, it remains unknown whether such complex disease risk alleles can be explained solely by stochastic fluctuations following demographic and migration events.

Crohn's disease (CD) is a highly heritable and complex disease characterized by chronic inflammation of the gastrointestinal tract (Mathew 2008). Genome-wide association studies (GWAS) have successfully detected more than 70 CD-susceptible genomic regions (Franke et al. 2010). The prevalence of CD is considerably higher in people of Europeans descent (2.2 and 13 million people in Europe and North America, respectively) than those descending from other geographical regions (Economou and Pappas 2008; Yazdanyar et al. 2010). Most of the CD-susceptibility alleles have been identified in populations of European ancestry by GWAS (Duerr et al. 2006; Franke et al. 2007; Libioulle et al. 2007; Raelson et al. 2007; Rioux et al. 2007; Barrett et al. 2008; Kugathasan et al. 2008) and early linkage analyses (Hugot et al. 2001; Ogura et al. 2001). However, significant associations between these genomic regions and CD have not been detected in non-European populations (Croucher et al. 2003; Zaahl et al. 2005; Nakagome et al. 2010), suggesting that there is a population-specific susceptibility to CD.

The question remains whether the population specificity of CD can be attributed to the neutrality of the CD-susceptible genomic regions. Uneven distributions of particular alleles and haplotypes are thought to be signals of natural selection, which may be a recent adaptation to the environment (Oota et al. 2004; Han et al. 2007). A classic example of a disease-causative allele maintained by environmental adaptation is that associated with sickle cell, which causes severe anemia, but its frequency is highly maintained in particular geographical regions because it confers resistance to malaria (Pasvol et al. 1978). However, few previous studies based on HapMap and CEPH-Human Genome Diversity Panel data have revealed a signal of natural selection that has operated on the CD-susceptible genomic regions using conventional tests, such as extended haplotype homozygosity (EHH) (Sabeti et al. 2002), integrated haplotype score (iHS) (Voight et al. 2006; Pickrell et al. 2009) for positive selection (typically, selective sweep), or Tajima's D (TD) (Tajima 1989) for balancing selection.

Recent theoretical and empirical studies have suggested the importance of "soft sweeps," including polygenic adaptation and positive selection on preexisting (i.e., standing) variation, which is more difficult to detect using conventional tests than "hard sweeps" or positive selection on newborn (novel) variation (Hermisson and Pennings 2005; Pritchard and Di Rienzo 2010; Pritchard et al. 2010; Hancock et al. 2011). An example of a soft sweep can be illustrated by artificial selection through domestication. Although a hard sweep on a novel variant would leave a large homogeneous region surrounding the advantageous allele, selection on a previously neutral or nearly neutral standing genetic variant, such as those artificially selected during domestication, is more likely to exist on multiple haplotypes and, as such, would not leave a selective footprint such as that found from hard sweeps (Innan and Kim 2004). Similar to artificial selection, when natural selection acts on a standing variation in a particular population, all haplotypes harboring the advantageous allele would increase in the population and thus the allele would be found on multiple haplotypes in both selected and nonselected populations depending upon its frequency and its probability of recombining into alternative backgrounds. Here, we examine CD-susceptibility loci as a model for determining the mechanism that maintains alleles involved in complex diseases. We tested eight CD-associated loci that are mostly European specific, confirmed previously using Japanese samples (Nakagome et al. 2010), to determine whether these genomic regions evolved under neutrality based on the Out-of-Africa model. We then genotyped and resequenced nine global populations at the *NOD2* locus and propose a new complex mechanism whereby mildly deleterious mutations have been maintained in humans by natural selection on standing variation.

Materials and Methods

Preliminary Analysis of Eight CD-Susceptibility Loci Using HapMap Data

We compiled a list of the CD-associated SNP sites and CD-risk alleles reported in the National Human Genome Research Institute database (http://www.genome.gov/gwastudies/) (Hindorff et al. 2009) and previous GWAS (supplementary table 1, Supplementary Material online). To compare the frequencies of haplotypes for which the CD-risk alleles were linked among the HapMap populations (The International HapMap Consortium 2005; The International HapMap Consortium 2007), we first selected the SNP sites that were shared among all populations (YRI, CEU, CHB, and JPT). These background SNPs were located in 5′ and 3′ flanking regions and gene regions and selected based on their allele frequencies and spatial intervals (supplementary appendix I, Supplementary Material online). In regards to the allele frequency, background SNPs were chosen when the expected heterozygosity was estimated to be greater than 0.2 in at least three of the four HapMap populations. This is because the SNPs shared among all HapMap populations at intermediate frequencies are likely to be old and predate the divergence of human populations (before the Out-of-Africa migration 100,000 years ago), and such SNPs have greater potential to produce high r^2-values than low-frequency alleles (Nordborg and Tavare 2002; VanLiere and Rosenberg 2008). In regards to spatial interval, the SNP sites were chosen when the interval between them was 2–16 kb. This is because linkage disequilibrium (LD) blocks generally extend up to 50 kb in non-African populations and is less than 10 kb in African populations due to bottlenecks and recent population growth in non-Africans (Tishkoff and Williams 2002; Kidd et al. 2004). We investigated LD blocks defined by pairwise r^2-values among the background SNPs using HAPLOT (initial $r^2 > 0.4$, average $r^2 > 0.3$, minimum $r^2 > 0.1$) (Gu et al. 2005) and excluded SNPs located in flanking regions when they were outside the blocks in all HapMap populations (supplementary fig. 1*a*, Supplementary Material online).

We also estimated the haplotypes consisting of the background SNPs and combined the haplotypes found at less than 3% in any populations (YRI, CEU, and CHB + JPT) into residuals. We then determined which haplotypes were linked with the CD-risk alleles (CD-haplotypes).

The CD-risk alleles selected may not be causal variants but index SNPs chosen as markers. However, we can expect that truly causal variants are located on the CD-haplotypes because of LD. To avoid any uncertainty from phasing about the residuals, we focused on the CD-haplotypes with frequencies of more than 3% in a population (supplementary fig. 1*b*, Supplementary Material online).

In order to determine whether the observed geographic differences in the CD-haplotype frequencies were the results of outliers in the genome, we divided genome-wide phased haplotypes (Public Release #21a) into "windows" defined by the total number of background SNPs (minor allele frequency, MAF $\geq$ 5%) and their intervals. Based on the allele frequency for each SNP site, we screened SNP sites with a MAF of less than 5%. We chose background SNPs with intervals that were defined as a range of lengths (minimum lengths are listed in supplementary appendix I, Supplementary Material online and maximum lengths were the minimum length + 10 kb). Reconstructions of haplotypes were started from the most 5′ SNP of the background SNPs in each chromosome according to the interval. If the number of background SNPs reached the number analyzed for each susceptibility locus (*NOD2*, 12 SNPs; *IL23R*, 16 SNPs; *TNFSF15*, 6 SNPs; supplementary fig. 1*b*, Supplementary Material online), then we determined the haplotypes in this window and calculated their frequencies in each population. We started a new haplotype reconstruction from the next background SNP site to the last SNP site in the previous window and repeated the analysis until we reached the last window located at the 3′-end. Finally, we obtained data sets of haplotype frequencies defined by background SNPs across the entire genome. When no background SNP was obtained within the given intervals, the windows were removed, and the new windows were defined from the next background SNP site.

To visualize genome-wide patterns of geographic differences in haplotype frequencies, we generated 2D histograms of pairwise joint haplotype frequency distributions (HFDs) for the eight CD loci: CEU versus YRI, CHB + JPT versus YRI, and CHB + JPT versus CEU (from top to bottom in supplementary fig. 2, Supplementary Material online). To investigate the significance of haplotype frequency in CEU in more detail, we performed a genome-wide haplotype analysis for the *NOD2* locus with the windows defined by LD and also by the number of SNPs and their intervals described above (supplementary fig. 3, Supplementary Material online). Pairwise r^2-values were calculated among background SNPs, and the empirical distribution was generated with windows screened with the average r^2-values greater than 0.477 (observed *NOD2* value in CEU).

DNA Samples

SNP genotyping and sequencing was done on 192 individuals from four African populations and 358 individuals from five non-African populations. Detailed information about the African and European samples can be found in ALFRED (http://alfred.med.yale.edu) using the following unique IDs: African populations consisted of Biaka Pygmy (N = 70; UID "PO000005F"), Hausa (N = 38; UID "PO000097"), Mbuti Pygmy (N = 39; UID "PO000006J"), and Chagga (N = 45; UID "PO000324J"). European populations consisted of Adygei (N = 54; UID "PO000017I"), Irish (N = 112; UID "PO000057M"), Russians (N = 48; UID "PO000019K"), and Danes (N = 49; UID "PO000007H"). In this study, Japanese (N = 95), the DNA of which was purchased from the Japan Health Science Foundation, were considered as "East Asians" since we focused on the relationship between Europeans and Africans. The Ethics Committee for Human Subjects at the University of Tokyo, Kitasato University School of Medicine and the Yale University School of Medicine approved all sampling protocols.

SNP Genotyping

We conducted genotyping for the same 12 SNPs as done in the preliminary analysis based on the HapMap database (supplementary appendix I, Supplementary Material online) in order to further characterize geographic haplotype patterns of the *NOD2* locus. In addition, one CD-associated SNP site (rs2066842: SNP#7) was also genotyped (supplementary fig. 4, Supplementary Material online). All SNPs were genotyped in the 550 individuals from the global samples mentioned previously. SNP typing was performed using TaqMan Genotyping Master Mix and TaqMan SNP genotyping assays (Applied Biosystems, Tokyo) on the LightCycler 480 System II (Roche Diagnostics, Tokyo). For all the SNPs in all the populations, we examined whether the Hardy–Weinberg equilibrium was significantly rejected at $P < 0.01$ using HAPLOT (Gu et al. 2005) before subsequent analyses. Furthermore, the LD structures among the 12 common SNP sites were defined in each population with the same criteria described above (supplementary fig. 5*a*, Supplementary Material online). The ancestral alleles for all the SNP sites were inferred by comparing to chimpanzee and orangutan reference sequences (101 bp centered around the SNP site).

Sequence Analysis

Polymerase chain reaction (PCR)-amplification was conducted using the primers listed in supplementary appendix II, Supplementary Material online. We analyzed a total of 3,016 bp sequences from the *NOD2* locus (supplementary fig. 4, Supplementary Material online), including 1,021 bp from exon 4, 1,642 bp from introns 7 to 9, and 353 bp from introns 10 to 11. We also sequenced two other CD loci: 640 bp from introns 7 to 8 of *IL23R* and 513 bp of 10q21. PCR products were purified by precipitation with 30% polyethylene glycol 6000. The DNA sequencing reaction was performed using the BigDye terminator cycle-sequencing kit version 3.1 according to the manufacturer's protocol (Applied Biosystems, Tokyo). The samples were then purified by ethanol precipitation with 3.4 mM EDTA (pH 8.1) and 81.1 mM sodium acetate. All nucleotide sequence data were obtained by ABI3130 and ABI3730 DNA Analyzers (Applied Biosystems, Tokyo). Base calling and detection of heterozygotes were performed using Sequencing

Analysis Software v5.2 (Applied Biosystems, Tokyo) followed by visual inspection for SNPs. All variants were confirmed by reading both strands, and all singletons were systematically reamplified and resequenced. Sequences were aligned with Clustal W, which was installed in MEGA (Tamura et al. 2007). An ancestral allele for each SNP site was inferred by alignment with chimpanzee and orangutan reference sequences. Except for one frameshift mutation for CD-risk allele rs2066847 (SNP#35 in supplementary fig. 4, Supplementary Material online), insertion/deletion polymorphisms were excluded from subsequent analyses.

Summary statistics of sequence diversity, such as nucleotide diversity (π), Tajima's D (TD), Fu and Li's D (FLD), and Fu and Li's F (FLF), were calculated using DnaSP version 5.0 (Librado and Rozas 2009). The significance of these statistics was obtained from the null distribution of 1,000 coalescent simulations under the constant size model using DnaSP.

Haplotype Phase Estimation

Phased haplotypes were estimated on the basis of the Bayesian statistical method using PHASE v.2.1 (Stephens et al. 2001) with population label information. We obtained 25 SNP sites from the sequencing analysis and 13 SNP sites from the SNP genotyping (12 background SNPs and SNP#7, supplementary fig. 4, Supplementary Material online). We reconstructed individual phases comprising the 38 total SNP sites spanning 57 kb in the *NOD2* locus by applying the algorithm three times using different seeds. We adopted the haplotypes for which consistent results were obtained in at least two of the three runs performed with different seeds. For the frameshift mutation (SNP#35), the ancestral (no insertion) and the derived (C insertion) states were treated as 0 and 1, respectively.

Phylogenetic Network Analysis

Phylogenetic networks of the *NOD2* haplotypes were constructed for each population and combined geographic populations (Africans, Europeans, and East Asians), using the median-joining algorithm of Network 4.5.1.6 (Bandelt et al. 1999) for the 38 SNP sites. To determine the distribution of the mutations in the genealogy of the *NOD2* locus, we initially defined the backbone, or internal branches, of the networks using the 12 SNP sites common to both African and non-African populations (SNP#1–6, 14–15, 17, 34, 37–38 in supplementary fig. 4, Supplementary Material online). We constructed the networks based on the major haplotypes with frequencies of more than 3% in a population in order to remove ambiguous phasing for the minor 12 SNP haplotypes, which may be generated from artificial phasing or very recent recombinants. On the basis of the backbone network of global populations shown in supplementary figure 5*b* (Supplementary Material online), we constructed the backbone networks of African, East Asian, and European populations (only the European backbone network is shown in supplementary fig. 5*c*, Supplementary Material online). Subsequently, we added the European-specific mutations, including both CD-associated alleles (SNP#7, 9, 13, 26, 30, 31, and 35) and non–CD-associated alleles, into the network and obtained complete networks for the 38 SNP haplotypes (data not shown for the African and East Asian networks). The root was inferred by assuming that the chimpanzee and orangutan allelic state at each SNP was ancestral.

Coalescent Simulation for Haplotype Analysis

We generated neutral patterns of genetic variation using the coalescent simulator "ms" (Hudson 2002) under the Out-of-Africa model (supplementary table 2, Supplementary Material online) (Gutenkunst et al. 2009) to determine the statistical significance of the observed geographic distribution for the CD-haplotype in the *NOD2* locus among global human populations. This model does not necessarily follow the true history of human populations. However, we confirmed that the pairwise HFDs generated from coalescent simulations fitted approximately with those from genome-wide patterns of haplotypes shown in supplementary figure 2, Supplementary Material online, under given window sizes (data not shown), suggesting that the simulated conditions were appropriate to evaluate neutral patterns of haplotype frequencies.

Distributions of haplotype frequencies in CEU or Europeans were generated from our coalescent simulations in order to compare the observed frequency with neutral expectations. The coalescent simulations were performed by specifying scaled mutation and recombination rates as well as demographic parameters. We adjusted the simulated data to the real genomic region of *NOD2* by setting the genomic region sizes and the genetic maps (cM) (Public Release #21a) between the first (rs4785223) and final (rs751919) SNP in the *NOD2* locus. Similar to the window analysis, we selected the background SNPs from the most 5′ SNP in the simulated data sets, which had an MAF $\geq$ 5% in any one of the African, European, or East Asian populations. The intervals between background SNPs were initially defined by more than 57 kb/12 SNPs (the genomic region size/the number of SNPs) (supplementary appendix I, Supplementary Material online). Perfect matching to the interval conditions resulted in insufficient simulation. If none of the background SNPs was obtained under the initial condition, the interval was gradually narrowed to 80%, 60%, 40%, and 20% of the initial condition. We discarded data sets when we did not find the next background SNP under all interval conditions. We repeated the simulations for 10,000 data sets and reconstructed the haplotypes of all chromosomes, assuming that the HapMap populations included YRI ($2N = 120$), CEU ($2N = 120$), and CHB + JPT ($2N = 180$) or that the global human populations included Africans ($2N = 384$), Europeans ($2N = 526$), and East Asians ($2N = 190$). We then generated the histograms of simulated distributions for the CEU-specific haplotypes or the European haplotypes that were observed in Africans ($\leq$0.5%) and East Asians (0%). We calculated the empirical P values by comparing the fraction of haplotypes greater than the observed frequency in CEU (33%) or Europeans (18%) to the total haplotypes.

We conducted comprehensive coalescent simulations based on 2^{13} combinations of demographic parameters

reported as the lowest and highest values of each confidence interval to further confirm the results from the coalescent simulation under maximum likelihood parameters (supplementary table 2, Supplementary Material online) (Gutenkunst et al. 2009). We repeated the simulations under each demographic condition for 100 data sets and investigated the distribution of European haplotype frequencies among the global human populations (supplementary fig. 6, Supplementary Material online).

Prediction of NOD2 Protein Structure and the Effects of Single Amino Acid Residue Substitutions

The GeneSilico metaserver (Kurowski and Bujnicki 2003) was used to predict protein secondary structure, boundaries between protein domains, and to carry out fold recognition by aligning the sequence of individual domains with previously solved crystal structures. If multiple templates were suggested by different methods with similar scores, the structure with the top score from HHSEARCH (Soding 2005) was used. Regions of intrinsic conformational disorder were predicted using the MetaDisorder method (http://iimcb.genesilico.pl/metadisorder; L.K. and J.M.B. manuscript in preparation). The metaserver identified the following structured elements in the NOD2 amino acid sequence that were separated by regions of intrinsic disorder: CARD domain I (residues 25–110; best template PDB code: 2b1w), CARD domain II (130–210; template: 2nz7), partially structured helical linker (211–274, no template), the NOD module (275–720; template 1z6t), and a leucine-rich repeats (LRRs) domain (730–1040; template: 3goz). The NOD module includes the following domains: NACHT-family AAA ATPase/nucleotide binding domain (NBD) (275–446), AAA-associated helical domain (447–535), winged helix domain (556–632), and SH domain (633–720).

A 3D model of the entire NOD2 amino acid sequence was generated using the FRankenstein's Monster method (Kosinski et al. 2003). The mutual position of the three main structural units in the model is arbitrary and the atomic details are only predictions. Nonetheless, the model is expected to be accurate on the level of interresidue contacts within individual domains and hence can be used to aid in the prediction of the effect of residue substitutions resulting from mutations.

We predicted the effects of amino acid substitutions on protein stability and function for SNP#7, 13, 26, 27, 33, and 35 based on the following methods: SNPs3D (Yue et al. 2006), CUPSAT (Parthiban et al. 2006), SIFT (Ng and Henikoff 2001), MutPred (Li et al. 2009), PopMusic (Gilis and Rooman 2000), MUpro (Cheng et al. 2006), PhD-SNP (Capriotti et al. 2006), I-Mutant2.0 (Capriotti et al. 2005), and PolyPhen (Adzhubei et al. 2010). The model structure was used as an input wherever possible; otherwise the NOD2 amino acid sequence was used. The degree of agreement among the majority of predictors was interpreted as an indication of the robustness of the consensus prediction. Predictions supported by more than 50% of methods were regarded as likely to be true, and their effect on the structure and function was additionally inferred from visual inspection of a predicted change in the modeled structure. In particular, the effect of the substitution on residue packing, electrostatic, hydrophobic, and polar interactions, the propensity for order/disorder, and secondary structure formation was taken into account. To further investigate the functional effects of the frameshift mutation (SNP#35), we predicted the ligand-binding regions around SNP#35 using the program Q-SiteFinder (Laurie and Jackson 2005). Previous studies have shown that these regions are essential for interactions with bacterial lipopolysaccharides (LPS) (Hugot et al. 2001; Ogura et al. 2001). We visualized the NOD2 protein structure on the highlighted regions around all the amino acid changes using the PyMOL Molecular Graphics System, Version 1.3 (Schrödinger, LLC).

Calculation of EHH and iHS

We examined the decay of EHH (Sabeti et al. 2002) for each of the 15 SNP alleles (SNP#1–7, 9, 14–15, 17, 31, 34, and 37–38) on the basis of the phased haplotype data sets in CEU (Public Release #22). We split all CEU chromosomes into two subsets of chromosomes, including those carrying the allele present on the H1 (H1 allele) and those carrying the non-H1 allele. For each subset, the EHH values were successively calculated between the core SNP site (each of the 15 SNP sites) and every upstream and downstream SNP site within 1.25 Mb. We plotted these values at given SNP sites and joined them with straight lines (supplementary fig. 7*a*, Supplementary Material online, blue for the non-H1 allele and red for the H1 allele). We further investigated the iHS using Haplotter (http://haplotter.uchicago.edu/) (Voight et al. 2006) on 2.5 Mb regions centering at the CD-associated SNP sites (SNP#7, 9, and 31) that define the root of H1 genealogy (supplementary fig. 7*b*, Supplementary Material online).

Comparison of the Number of Population-Specific Mutations

We determined the number of population-specific mutations (Africans, Europeans, or East Asians) for the three sequenced regions (*NOD2*, *IL23R*, and 10q21). These numbers were converted into a ratio of European- or East Asian-specific mutations to African-specific mutations (EUR/AFR or EAS/AFR). We used two publicly available data sets, the Seattle SNPs database (NHLBI Program for Genomic Application) and the National Institute of Environmental Health Sciences (NIEHS) Environmental Genome Project SNPs database (NIEHS Environmental Genome Project) to generate an empirical distribution of the ratio. We collected a total of 6.9 and 6.2 Mb sequences from 307 genes in the Seattle SNPs database (Panel 1, 24 African Americans and 23 European CEPH; Panel 2, 24 HapMap YRI and 23 HapMap CEU) and 253 genes in the NIEHS database (Panel 2, 15 African Americans, 12 HapMap YRI, 22 Europeans, 22 Hispanics, and 24 Asians). For the NIEHS database, we considered African Americans and HapMap YRI as Africans, and Europeans and Hispanics as Europeans. We determined

the number of polymorphic sites specific to Africans or Europeans and calculated EUR/AFR for each gene. To make a histogram of the empirical distribution, we divided the ranges of the ratios into 0.2 bins. The fraction of the ratio greater than the observed ratio is considered as empirical *P* values.

Estimation of Population Growth Rates, Ages of Mutations, and Time to Most Recent Common Ancestor

We conducted coalescent simulations using the program GENETREE (Griffiths 2007) in order to generate maximum likelihood estimates for the scaled population mutation rate (θ_{ML}), growth parameter (β_{ML}), time to most recent common ancestor (TMRCA), and ages of mutations for each of the haplogroups in Europeans (H1, H2, and H3) as well as for all of the *NOD2* sequences in global populations. We obtained a genealogy of each haplogroup that was deduced from the path of mutations to the MRCA under the infinitely many-site model using the 3,016 bp of sequence data. For the global populations (supplementary fig. 8, Supplementary Material online), we only used the 1,642 bp sequences from introns 7 to 9 since the LDs among the sequenced regions (exon 4, introns 7–9, and introns 10–11) were weak in the African population (supplementary fig. 5*a*, Supplementary Material online). Given the gene tree, we computed maximum likelihood values for the population mutation rate (θ_{ML}) under the exponential growth model where we specified the range of θ_{ML} and β_{ML}, respectively (Griffiths and Tavare 1994b). The likelihood surfaces with respect to θ_{ML} or β_{ML} were generated by the average likelihood values over 100,000 simulation runs. The θ_{ML} was then used to calculate the effective population size (N_e) on the basis of the mutation rate (1.421×10^{-9}/bp/year), which was calculated from the average number of substitutions between all human and chimpanzee sequences assuming a human–chimpanzee divergence time of 6.5 Ma (Patterson et al. 2006).

The empirical distributions of the TMRCA and ages of mutations were obtained based on the likelihood estimated from each simulation run conditional in the topology of the gene tree, θ_{ML}, and β_{ML}. Coalescent dates were translated to chronological time using a 25-year mean intergeneration interval. Based on the TMRCA (generations), β_{ML}, and N_e values, the effective population size at the TMRCA, N_t, was calculated by:

$$N_t = N_e e^{-\beta_{ML} t},$$

where t was a scaled time with TMRCA/$2N_e$ (generations) (Balding et al. 2007). Then, the growth rate per generation (α) was obtained from:

$$\alpha = \left(\frac{N_e}{N_t}\right)^{\frac{1}{\text{TMRCA(generations)}}}.$$

We depicted the history of effective population size changes and mutation accumulations from N_t at TMRCA to N_e at the present.

Results

We initially examined eight CD-susceptibility loci based on HapMap data. The associations to CD of the eight loci have been reproduced in populations of European ancestry in more than two independent and previously reported investigations (supplementary table 1, Supplementary Material online) and are absent in East Asians (Nakagome et al. 2010). In our preliminary analysis, we compared the frequencies of the haplotypes linked with the CD-risk alleles in the HapMap populations (YRI, CEU, CHB + JPT). We found that most of the CD-risk alleles for each locus were located on multiple haplotypes (CD-haplotypes) in a population, and these CD-haplotypes were usually shared among the HapMap populations (supplementary fig. 1*b*, Supplementary Material online). To further investigate the geographic differences of shared and population-specific haplotype frequencies, we generated 2D histograms of pairwise HFDs (supplementary fig. 2, Supplementary Material online). For the HFDs in CEU versus YRI or CHB + JPT versus YRI, the shared haplotype frequencies tended to be higher in CEU or CHB + JPT than YRI. In contrast, the HFDs were relatively correlated between CHB + JPT and CEU. These patterns of HFDs were mostly identical to those from coalescent simulations under the Out-of-Africa model (supplementary table 2, Supplementary Material online), representing the effects of human demography such as bottlenecks and recent population expansion in non-Africans (data not shown). We then plotted the observed differences in CD-haplotype frequencies on the bins in the HFDs (shown as gray squares in supplementary fig. 2, Supplementary Material online). As expected, most of the frequencies of the CD-haplotypes that were shared between populations or were specific to a population could be explained by demographic effects, indicating that these differences were negligible in terms of entire genomic regions.

We found that except for the *NOD2* locus, all the geographic variation of CD-risk allele frequencies for all the loci can be explained by genetic drift and demography (supplementary figs. 1 and 2 and table 1, Supplementary Material online). However, the three CD-risk alleles of *NOD2* were present only in CEU (supplementary table 1, Supplementary Material online) and were exclusively linked to a CEU-specific haplotype (supplementary fig. 1*b*, Supplementary Material online, dark green [33%]). To determine whether this pattern was unusual in the human genome, we generated an empirical distribution of the CEU-specific haplotype frequency derived from genome-wide patterns of the 12 SNP haplotypes summarized in supplementary figure 2*a*, Supplementary Material online. We found that the observed frequency (33%: 40 chromosomes) appeared significantly in the tail of the entire distribution ($P = 0.00017$) (fig. 1*a*). This result was also supported by a window analysis including LD ($P = 0.00046$) and a coalescent simulation ($P = 0.00007$) (supplementary fig. 3, Supplementary Material online). These results suggest that the CD-risk alleles on the *NOD2* locus are involved in the unusual prevalence of the CEU-specific haplotype, which is difficult to explain by genetic drift and demography.

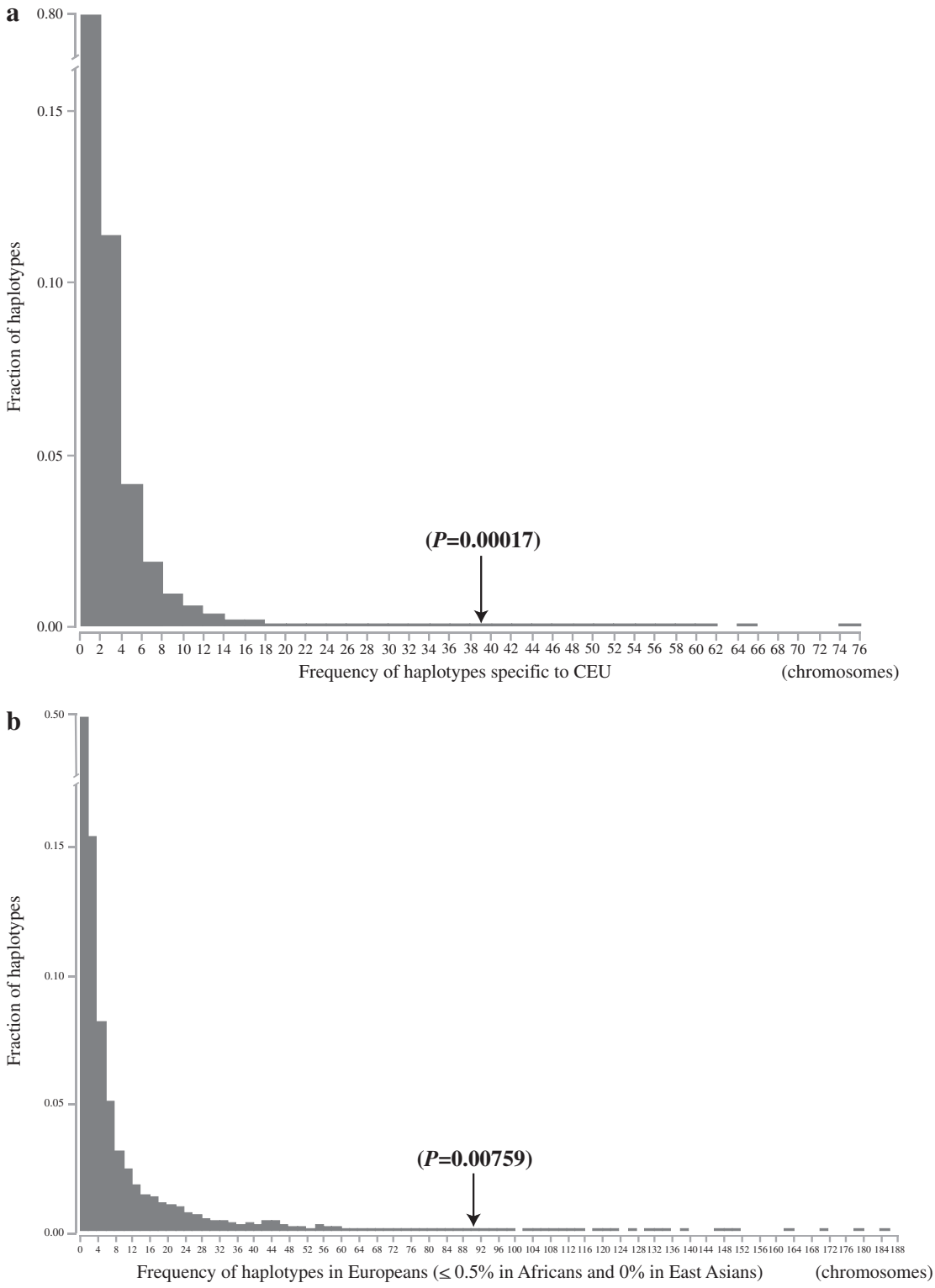


Fig. 1. Histograms of the empirical distribution for (*a*) CEU-specific haplotype frequencies generated from a window analysis (12 SNPs) of genome-wide phased haplotypes based on HapMap data and (*b*) European haplotype frequencies observed in Africans (≤0.5%) and in East Asians (0%) generated from a coalescent simulation involving the *NOD2* locus. The horizontal axis indicates the number of chromosomes, including the haplotypes binned with "2-chromosome," while the vertical axis indicates the fraction of haplotypes included in each bin. The black arrows represent the observed frequency at the *NOD2* locus in (*a*) CEU from the HapMap populations and (*b*) European populations from the global human population.

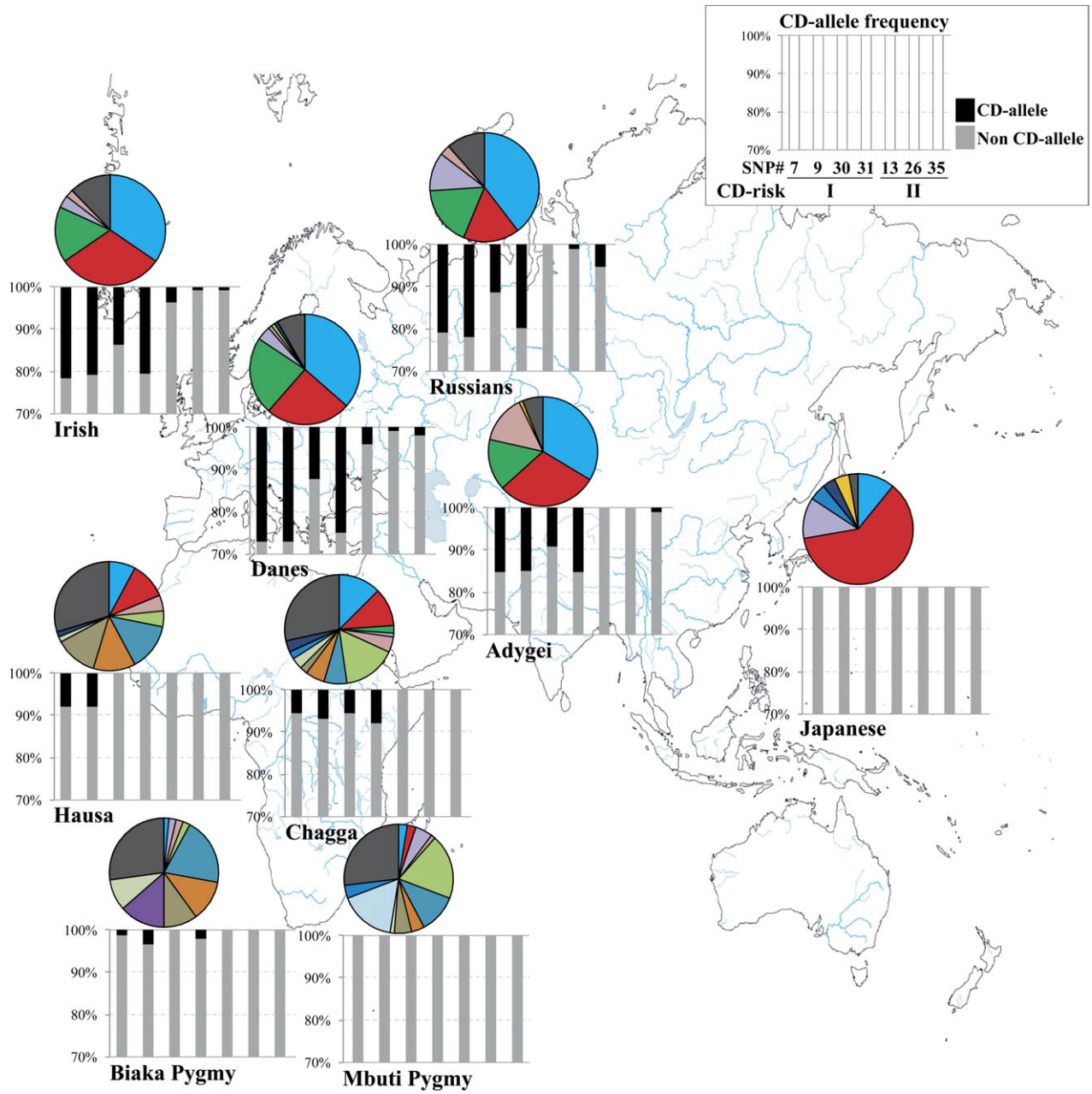


Fig. 2. Geographic frequency distributions of the 12 SNP haplotypes (pie charts) and the seven CD-associated alleles (bar graphs) observed in global human populations (see supplementary table 3, Supplementary Material online).

Distributions of CD-Risk Alleles and Haplotypes in Global Human Populations

To determine whether the CEU specificity of the CD-risk alleles on *NOD2* is common in European populations, we conducted genotyping and resequencing of the *NOD2* locus (supplementary fig. 4, Supplementary Material online), including seven CD-associated SNPs, in the global samples ($N = 550$) consisting of one East Asian (Japanese), four African (Biaka Pygmy, Hausa, Mbuti Pygmy, and Chagga), and four European (Adygei, Irish, Russians, and Danes) populations (fig. 2; supplementary table 3, Supplementary Material online). Three CD-risk alleles cause amino acid changes (SNP#7, Pro > *Ser*; SNP#13 Arg > *Trp*; SNP#26, Gly > *Arg*), and one CD-risk allele is a frameshift mutation (SNP#35 insertion of C) that generates a truncated NOD2 protein (1,007 of 1,040 amino acid residues, 1,007fs). The other risk alleles include a synonymous mutation (SNP#9) and occur in the intronic region (SNP#30 and SNP#31) (supplementary fig. 4, Supplementary Material online). We found that the SNP#13 Trp, SNP#26 Arg, and SNP#35 1,007fs alleles were present at very low frequencies (0.9–5.2%) in only European populations, while the rest of alleles (SNP#7 Ser; SNP#9, 30, and 31) were prevalent in Europeans (14.9–27.1%), absent in East Asians, and rare (<8%) in Africans (fig. 2, supplementary table 3, Supplementary Material online). Furthermore, our haplotype analyses showed that the CEU-specific haplotype (H1) was common in the four European populations (Adygei, 15%; Irish, 16%; Russians, 18%; and Danes, 23%), but

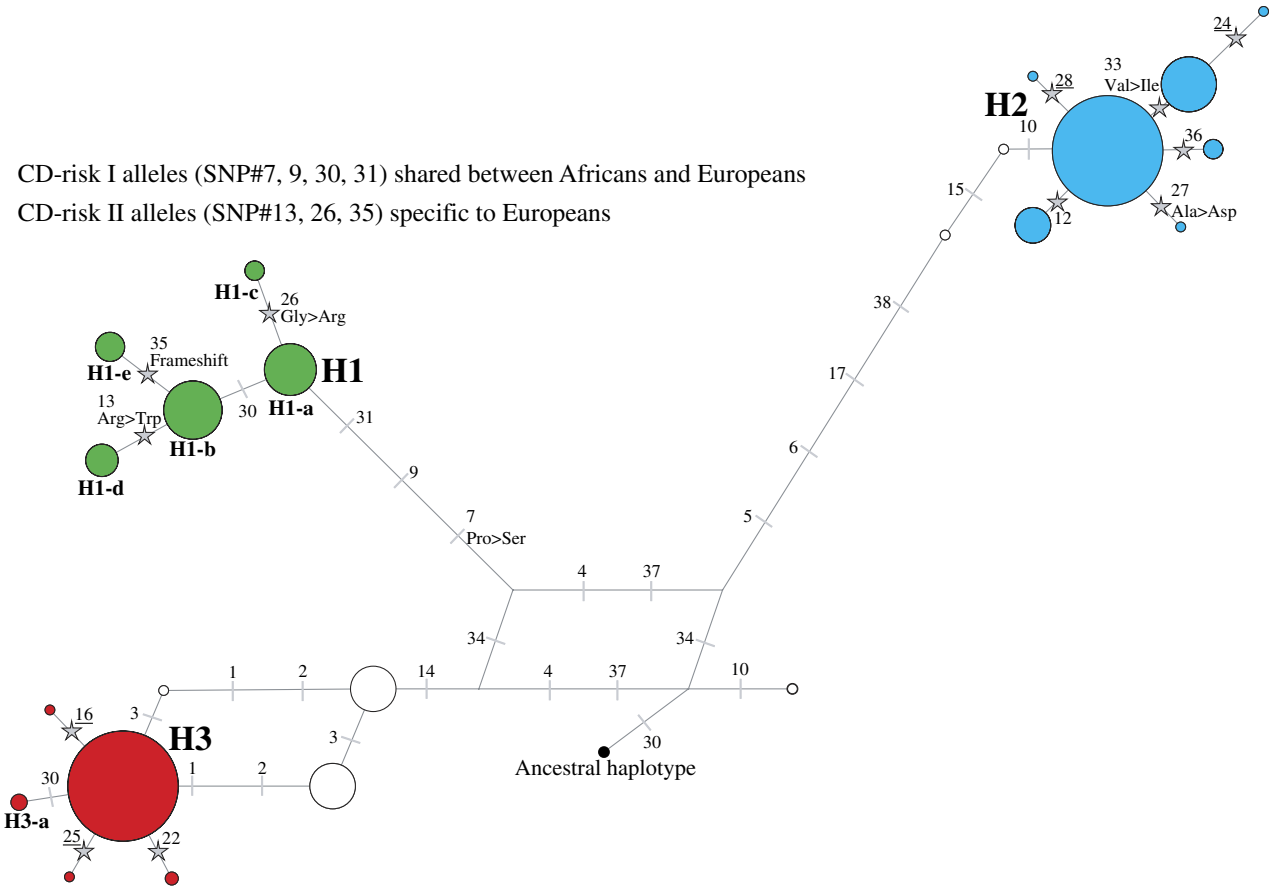


FIG. 3. Haplotype network in Europeans based on 38 SNP sites. The size of the circle corresponds to the frequency of the haplotype. The numbers on the branches correspond to the SNP# shown in supplementary figure 4, Supplementary Material online. The stars on the branches indicate European-specific alleles. Underlined SNP site numbers indicate the sites that the phase inference failed to assign to one of two chromosomes in an individual. The alternative phase is described in supplementary table 5, Supplementary Material online.

absent in East Asians and Africans, except for the Chagga (2.3%) (fig. 2; supplementary table 3, Supplementary Material online). Thus, some CD-risk alleles and the CEU-specific haplotype were exclusive to European populations. The CD-risk alleles were separated into two categories: SNP#7, 9, 30, and 31 were common in Europeans but present in Africans at low frequencies (hereafter referred to as CD-risk I) and SNP#13, 26, and 35 were specific to Europeans (CD-risk II).

A Phylogenetic Analysis of *NOD2*

Using the genotyping and resequencing data for *NOD2*, we constructed a phylogenetic network of Europeans (fig. 3, supplementary figs. 5*b* and *c*, Supplementary Material online). There were three major haplogroups (H1, dark green; H2, light blue; and H3, red) that constituted about 80% of the haplotype configuration in European populations. We found that H1 consisted completely of the seven CD-risk alleles, with the root and the internal branches defined by the CD-risk I alleles (SNP#7, 9, 30, and 31) and the external branches defined by the CD-risk II alleles (SNP#13, 26, and 35) (fig. 3). In contrast, H2 and H3 were shared among Africans, Europeans, and East Asians (fig. 2, supplementary table 3, Supplementary Material online) but they harbored many non-CD alleles specific to Europeans (H2 haplogroup: SNP#12, 24, 27, 28, 33, and 36; H3 haplogroup: SNP#16, 22, and 25). Some of these alleles result in amino acid changes (SNP#27 Ala > Asp; SNP#33 Val > Ile) (supplementary fig. 4, Supplementary Material online). These patterns suggest that H2 and H3 originated in Africa and the European-specific mutations appeared on those haplotypes after the Out-of-Africa expansion. To determine whether demographic history accounts for the observed H1 distribution, we conducted a coalescent simulation on the basis of the demographic parameters assumed under the "Out-of-Africa" model (supplementary table 2, Supplementary Material online). The empirical histogram indicated the haplotype frequency in Europeans, conditional on the frequencies in African ($\leq$0.5%) and in East Asian (0%) populations; the observed frequency at the *NOD2* locus in Europeans appeared significantly in the tail of the entire distribution ($P = 0.00759$ in fig. 1*b*). This result was also supported by the extensive demographic conditions ($P = 0.0173$ in supplementary fig. 6, Supplementary Material online), suggesting that the frequency and geographic distribution of H1 is unlikely to be explained by only demographic (evolutionary neutral) events. There are two possible scenarios for the distribution of H1. H1 either 1) originated and rapidly expanded in Europe after humans migrated out of Africa or

2) originated in Africa and was maintained in Europe but nearly lost in Africa. Both cases support the non-neutral evolution of *NOD2*, with the former (1) being a case of selective sweep from a novel mutation (recent sweep model) and the latter (2) an example of selection on preexisting variation (standing variation model). In either case, it is evident that the CD-risk I alleles, except for SNP#30, defining the root of H1 genealogy have spread by hitchhiking with H1.

Prediction of the Functional Effects of Each Amino Acid Change

We generated 3D models of the NOD2 protein, which consists of two N-terminal caspase activation and recruitment domains (CARDs), a centrally located NBD, and C-terminal tandem LRRs domain. We then used nine different methods to predict the effect of the nonsynonymous and frameshift mutations of the CD-associated SNP sites on protein stability and function (fig. 4).

The European-specific nonsynonymous substitutions (SNP#13 Arg702*Trp*; SNP#26 Gly908*Arg*) that appeared on the external branches of H1 genealogy were located in the SH and the LRRs domain, respectively. The Arg702*Trp* (SNP#13) and the Gly908*Arg* (SNP#26) substitutions are predicted to strongly destabilize a helix in the SH domain and a turn in the LRRs domain, respectively, both of which are likely to decrease protein stability. The frameshift mutation (SNP#35, 1007fs), which also constituted an external branch of H1, was located in the LRRs domain as well. Although this mutation was unlikely to destabilize the overall protein structure, the insertion of a "C" generates a truncated LRRs domain (by 33 amino acids) (fig. 4). The LRRs domain is predicted to include a putative ligand-binding region and the truncation is likely to interfere with the binding ability of this domain. Three of the CD-risk I alleles (SNP#7, 9, and 31) define the root of H1 genealogy in Europeans (fig. 3). The r^2-values between SNP#7 and 9/between SNP#7 and 31 were 1.0/1.0 for Adygei, 1.0/0.9 for Irish, 1.0/0.8 for Russians, and 1.0/0.9 for Danes populations. SNP#7 (Pro268*Ser*) affects a residue in the flexible hinge between the CARDs and NBD. This amino acid change was predicted to have a small destabilizing effect, although the introduction of the Ser allele was likely to decrease the tendency of the region to be disordered. We infer that the Pro268*Ser* (SNP#7) substitution mildly affects the conformational flexibility of the linker.

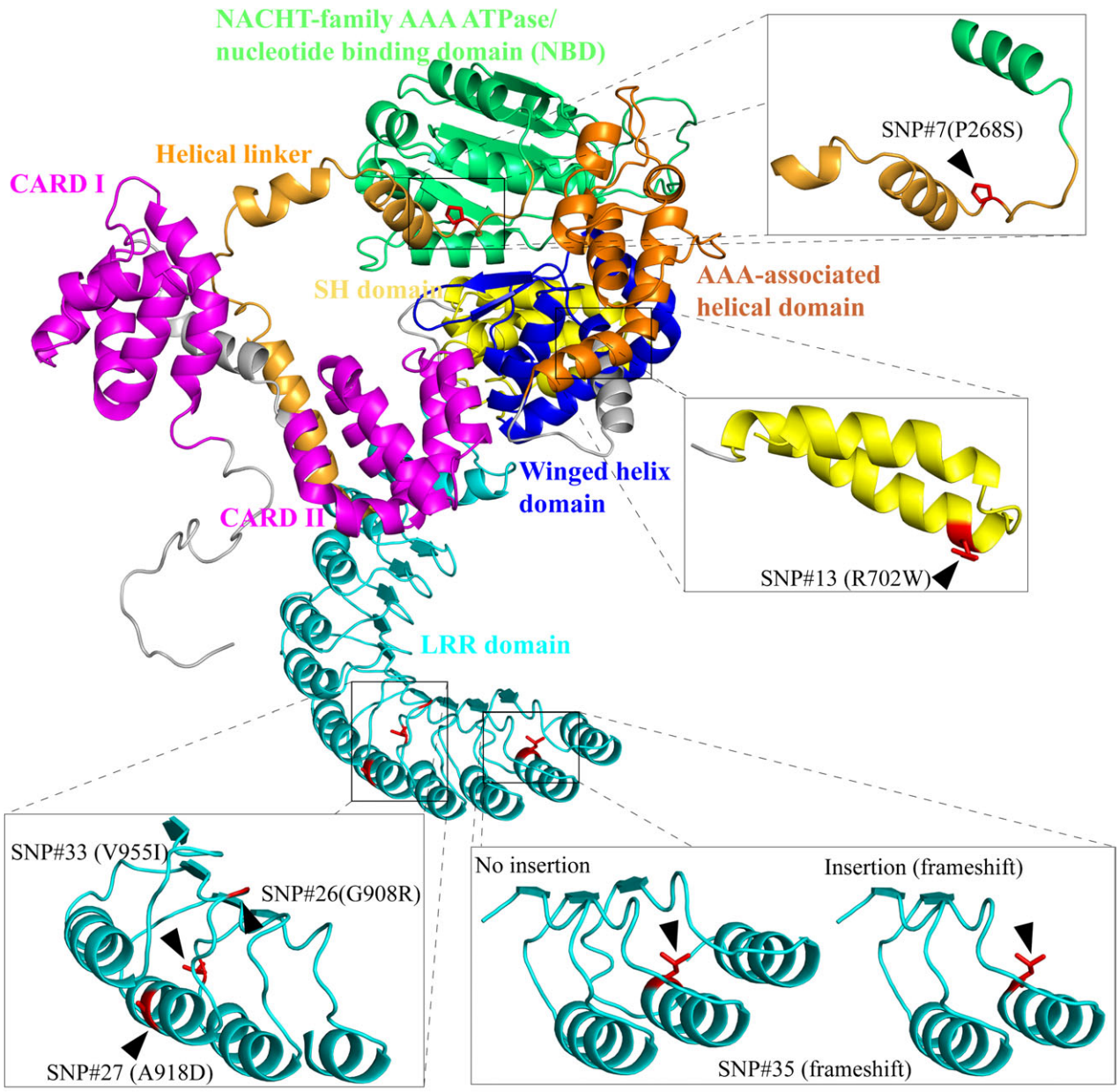
The European-specific substitutions (SNP#27, Ala918Asp; SNP#33, Val955Ile) that appeared on the H2 (fig. 3) were not CD-risk alleles and were located on the LRRs domain. The substitution Ala918Asp (SNP#21) was predicted to have a strong destabilizing effect on the structure of the LRRs domain, as it replaces a buried Ala residue within a helix with a negatively charged Asp. The substitution Val955Ile (SNP#27) may mildly destabilize a turn between a helix and a strand, but it is likely to have a benign structural effect. Thus, the CD- and non–CD-risk alleles that cause amino acid substitutions and are exclusive to Europeans can result in serious structural changes that may be mildly deleterious.

An Excess of Polymorphisms in *NOD2*

If the geographic distribution of H1 can be explained by natural selection that has operated on the *NOD2* locus, it remains unknown whether the H1 haplogroup that the CD-risk alleles have hitchhiked with has survived through recent or older selection. We set each of the 12 SNPs (SNP#1–6, 14–15, 17, 34, and 37–38) and three CD-risk I SNPs (SNP#7, 9, and 31) as a core SNP site and calculated the EHH in 2.5 Mb regions using HapMap data sets (supplementary fig. 7*a*, Supplementary Material online). However, the decay of EHH did not show any evidence of recent selective sweeps, which was also confirmed by iHS (SNP#7, −1.010; SNP#9, −1.119; and SNP#31, −1.139; supplementary fig. 7*b*, Supplementary Material online). Thus, no data supported the recent sweep model for H1. The three CD-risk I alleles, which define the H1 lineage, are present at low frequencies in African populations (fig. 2, supplementary table 3, Supplementary Material online) and most are independently located on non-H1 haplotypes (r^2-values between SNP#7 and 31: Biaka pygmy, 0.0; Hausa, 0.0; Chagga, 0.7; supplementary table 4, Supplementary Material online). Thus, these CD-risk alleles were present before the Out-of-Africa migration and natural selection may have acted on preexisting (standing) variation in H1. Moreover, we observed an excess of polymorphisms in *NOD2*, with two times more European-specific SNPs (12 sites; 0.20–6.47%; average frequency 1.36%) than African-specific SNPs (6 sites; 0.81–3.51%; average frequency 2.30%) (table 1, supplementary table 5, Supplementary Material online). We investigated the genome-wide distribution for the ratio of European-specific to African-specific mutations using genome-wide data from the Seattle SNPs and NIEHS databases and found that only four loci, including *NOD2*, had ratios greater than 2.0, which was significantly high compared to wide-range genomic regions ($P = 0.004$) (fig. 5). To characterize the excess of polymorphisms in detail, we calculated the FLD and FLF statistics based on the sequence data. Both the D and F values were negative only for *NOD2* from Europeans ($P = 0.008$ and $P = 0.028$, respectively) due to the excess of singletons in the external branches compared with total variations (table 1). These results suggest that the European-specific polymorphism pattern on *NOD2* is a significant departure from neutrality, showing that while natural selection has not operated very recently, it did so after the divergence between Africans and Europeans.

The Growth Parameters and the TMRCA of the Three Haplogroups

The phylogenetic network shows a star-like topology, which indicates demographic expansion (Di Rienzo and Wilson 1991), that is much more remarkable in the H2 and H3 haplogroups than in the H1 haplogroup (fig. 3). This indicates that the excess of polymorphisms is mainly due to H2 and H3. To investigate the demographic effect on each haplotype, we estimated the growth parameters (β_{ML}) and the TMRCA for the three lineages (haplogroups) (table 2). Both H2 and H3 showed signals of population growth (H2, $\beta_{ML} = 6.5$; H3, $\beta_{ML} = 22.5$). In contrast, H1



SNP#	Substitution	Structural context				Disease?					Destabilization?					Possible effect of substitutions (interpretation of a consensus from all predictions)
						Sequence-based							3D-based			
		Domain	Secondary structure	Solvent accessibility (%)	Disorder	Polyphen	PhD-SNP	MutPred	SNPs3D	SIFT	Mupro (delta G)	Mupro (svm)	I-Mutant2.0	PopMusic	CUPSAT	
7	P268S	Linker	Coil	23	D	0	0	+/-	0	+/-	-	-	-	-	-	affects a disordered region in a linker between CARD and NBD
13	R702W	Linker	Helix	19	/	-	0	-	-	-	-	-	-	-	-	destabilizes a helix
26	G908R	LRR	Turn	39	/	+/-	-	+/-	0	+/-	-	-	-	-	-	destabilizes a turn
27	A918D	LRR	Helix	3	/	+/-	-	+/-	*N.A.*	+/-	-	-	-	-	+	destabilizes a helix
33	V955I	LRR	Turn	62	/	0	0	+/-	0	+/-	-	-	-	-	-	destabilizes a turn, but the effect may be mild
35	1007fi	LRR	Helix	*N.A.*	/	*N.A.*	*N.A.*	*N.A.*	*N.A.*	*N.A.*	*N.A.*	*N.A.*	*N.A.*	*N.A.*	*N.A.*	frame shift mutation which removes last helix from LRR domain, indirect evidence that this helix is crucial for LPS binding

N.A.: not available (SNP#35 is a insertion polymorphism and we cannot predict the effects, because all programs are able to predict only single mutations (replacement of one amino acid by another on the level of protein).
Methods predicting disease effects ("Disease?"): '0' marginal effect, '+/-' may cause damage, '-' likely to cause disease.
Methods predicting (de)stabilization ("Destabilization?"): '+' increases stability, '-' decreases stability.

FIG. 4. A 3D model of the NOD2 protein and the positions of amino acid changes (SNP#7, 13, 26, 27, 33, and 35) identified in the global human population. The NOD2 protein is composed of two N-terminal CARDs (magenta), a helical linker (golden yellow), a centrally located NOD module including NBD (green), an AAA-associated helical domain (orange), a winged helix domain (dark blue), an SH domain (yellow), and 10 C-terminal tandem LRRs (light blue). Black triangles indicate the positions of the six amino acid substitutions and the frameshift, which are marked as red sticks in the predicted structure. The table below the 3D model includes the SNP#, type of amino acid substitution or insertion, the structural context of the substitution (location of domain and secondary structure, solvent accessibility, disorder region), predicted structural change on the basis of sequence and 3D structure, and the possible effect of each substitution.

Table 1. Summary Statistics of Sequence Diversity in Africans, Europeans, and East Asians.

	Africans					Europeans					East Asians				
Genes (length of sequences)	S (Spop)[a]	π[b]	TD[c]	FLD[d]	FLF[e]	S (Spop)	π	TD	FLD	FLF	S (Spop)	π	TD	FLD	FLF
NOD2 (3,016 bp)															
Exon 4 (1,021 bp)	4 (2)	0.390	−0.585	0.842	0.429	4 (2)	0.880	0.849	0.820	0.996	1 (0)	0.022	0.291	0.457	0.477
Intron 7 to exon 9 (1,642 bp)	9 (4)	0.580	−0.665	1.231	0.644	11 (8)	0.720	−0.589	−2.770**	−2.354*	4 (1)	0.290	−0.548	0.893	0.492
Intron 10 to intron 11 (353 bp)	0 (0)	N.A.[f]	N.A.	N.A.	N.A.	1 (1)[g]	0.004	0.791	0.415	0.036	0 (0)	N.A.	N.A.	N.A.	N.A.
10q21 (513 bp)	4 (4)	0.130	−1.456*	−0.525	−1.006	2 (2)	0.050	−1.089	−1.407	−1.557	1 (1)	0.040	−0.897	0.456	0.046
IL23R (640 bp)	4 (1)	0.390	−0.984	−0.524	−0.813	3 (0)	0.520	−0.354	0.713	0.426	0 (0)	N.A.	N.A.	N.A.	N.A.

[a] Total number of segregating sites (that of segregating sites specific to a population).
[b] Nucleotide diversity in a population ($\times 10^{-3}$).
[c] TD, Tajima's D.
[d] FLD, Fu and Li's D.
[e] FLF, Fu and Li's F.
[f] N.A. means that summary statistics were not calculated because there was no polymorphism in a population.
[g] The CD allele (frameshift mutation; SNP#35) specific to Europeans is not included in this analysis.
*$P < 0.05$, **$P < 0.01$: the significance of observed values estimated by coalescent simulation (1,000) assuming constant population size.

showed no influence from population growth ($\beta_{ML} = 0.13$). The H1 TMRCA was estimated to be 160,619 years ago (ya), which was older than those for H2 (89,485 ya) and H3 (43,771 ya) (table 2 and fig. 6*a*). This estimate was consistent with the age of the CD-risk allele SNP#31 from the complete *NOD2* genealogy of all populations of 173,550 ya (supplementary fig. 8, Supplementary Material online). Thus, it is likely that H1 already existed before or around the Out-of-Africa migration. Moreover, the appearance of European-specific mutations on H2 and H3 corresponds to the population size expansion, while the relatively older mutations (CD-risk alleles: SNP#13, 26, 30, and 35) have remained in H1 (fig. 6*b*). These results suggest that the survival of H1 in Europeans is not due to population growth and the CD-risk alleles have subsequently accumulated on H1, supporting the standing variation model that the putatively advantageous mutation preexisted on H1.

Discussion

We initially analyzed eight CD-susceptibility loci using HapMap data in order to determine whether European-specific susceptibility is due to the exclusive distribution of the CD-risk alleles in the European population and demographic history. The SNPs that we chose in the haplotype analysis are common in all of the populations analyzed and constitute the internal parts of the gene genealogy in humans (Fu and Li 1993). Hence, the differences in their frequencies among populations reflect recombination, mutation, demography, and natural selection that have occurred subsequent to human dispersal out of Africa (Tishkoff and Williams 2002). We focused on the *NOD2* locus based on the results from our preliminary analysis. The CD-risk alleles and the CD-haplotype (H1: that is a haplotype including CD-risk alleles) showed a CEU-specific distribution. This haplotype frequency pattern was not observed for the other CD-susceptibility loci (supplementary fig. 1*b*, Supplementary Material online) and was very rare across the entire genome (fig. 1*a*, supplementary fig. 3, Supplementary Material online). Therefore, it is difficult to explain the frequency and geographic distribution of the CD-haplotype as being due to neutral processes, such as spatial migrations in Europe (Novembre et al. 2008; Novembre and Stephens 2008), but it may be explained by a unique/complicated selective scenario.

Accumulation of CD-Risk Alleles on H1

Our sequencing and SNP typing using global populations demonstrated that four CD-risk I alleles (SNP#7, 9, 30, and 31) were shared between Africans and Europeans, and three CD-risk II alleles (SNP# 13, 26, and 35) were specific to European populations (fig. 2). These alleles were not found in East Asians, consistent with previous studies (Croucher et al. 2003; Nakagome et al. 2010). The CD-risk alleles genotyped in this study may be merely surrogate evidence of CD association for undetected causal variants, and the lack of any associations in East Asians could be due to the lack of representation of index SNPs in DNA arrays or the smaller sample size in the association studies. Although the LD

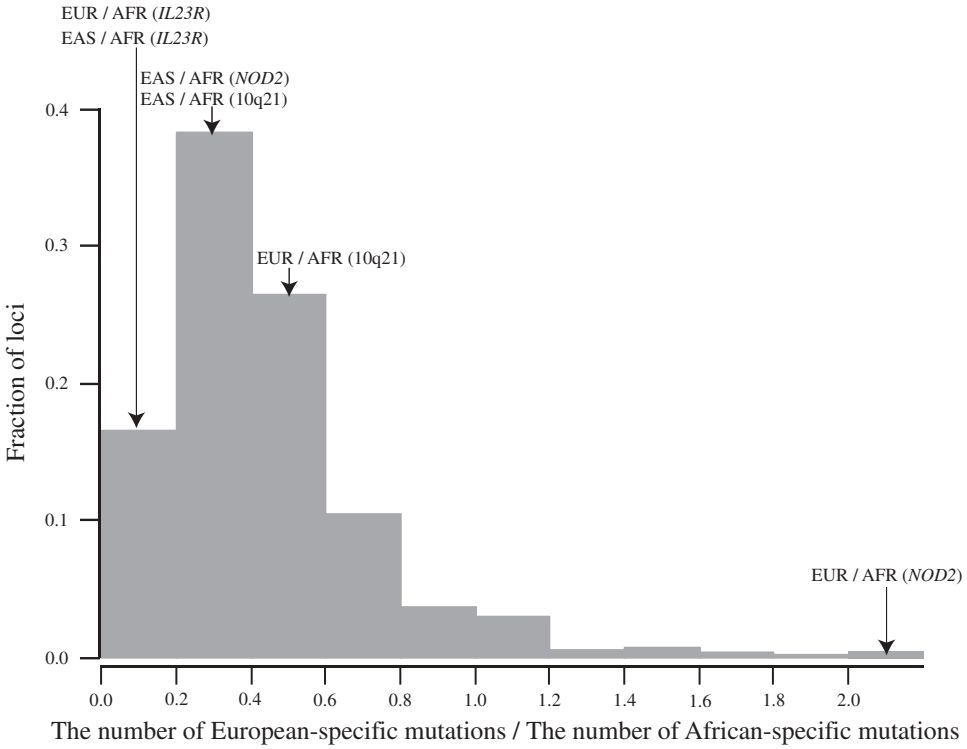


Fig. 5. Histogram of the empirical distribution of the ratio of the number of European-specific mutations (EUR) to that of African-specific mutations (AFR) generated from the 555 loci (see "Comparison of the number of population-specific mutations" in the Materials and Methods section). The horizontal axis indicates the values of EUR/AFR binned with 0.2, and the vertical axis indicates the fraction of the number of loci included in each bin. The three loci with EUR/AFR > 2.0 are included in the rightmost bin (2.00 in *HSPA1A*, 2.36 in *HSPA6*, and 2.67 in *MT2A*). The observed ratios in *NOD2*, *IL23R*, and 10q21 are shown by arrows on the histogram (EAS: the number of East Asian-specific mutations). The EUR/AFR in *NOD2* significantly deviates from the distribution ($P = 0.004$). The 10q21 and *IL23R* loci had more AFR than EUR and EAS (table 1), suggesting that our data are not artifacts of the limited resequencing length.

structures defined by the 12 common SNP sites in *NOD2* were similar between Europeans and East Asians (supplementary fig. 5*a*, Supplementary Material online), it is necessary to expand GWAS and deep sequencing analyses to non-European populations to address population-specific susceptibility (Rosenberg et al. 2010). However, the association of all CD-risk alleles has been reproduced in multiple GWAS (supplementary fig. 4, Supplementary Material online), and they are established index (or possibly causal) alleles in Europeans. The four of seven CD-risk alleles (CD-risk I: SNP#7, 9, 30, and 31) are standing variants (fig. 2, supplementary table 4, Supplementary Material online), and their odds ratios are 1.48 (SNP#9), 1.46 (SNP#30), and 1.54 (SNP#31) (Kugathasan et al. 2008; Glas et al. 2010). European-specific CD-risk alleles (CD-risk II) that result in amino acid changes (SNP#13 Arg702*Trp* and 26 Gly908*Arg*) and protein truncation (SNP# 35, 1,007fs) have also been identified by linkage analysis (Hugot et al. 2001; Ogura et al. 2001). They are likely to be highly heritable risk alleles, with odds ratios of 2.20 (SNP#13), 2.99 (SNP#26), and 4.09 (SNP#35) (Economou et al. 2004). These odds ratios are higher than those for other complex disease alleles, which usually range from 1.2 to 1.5 (Bodmer and Bonilla 2008). The external branches of H1 genealogy include higher risk CD alleles specific to Europeans, when compared with those in the root or internal branches (fig. 3). Gene genealogy suggests that these high-risk variants subsequently appeared on H1 after the pre-existing CD-risk alleles hitchhiked with H1 in Europeans.

Table 2. GENETREE Analysis for H1, H2, and H3 Haplogroups in Europeans.

Haplogroups	θ_{ML}[a]	β_{ML}[b]	N_e (N_t)[c]	TMRCA (95% CI)[d]
H1	0.79	0.13	1,843 (1,469)	160,619 (±122,835)
H2	2.00	6.50	4,666 (386)	89,485 (±31,806)
H3	2.30	22.50	5,366 (137)	43,771 (±11,401)

[a] Maximum-likelihood estimates of scaled population mutation rate, θ.
[b] Maximum-likelihood estimates of the growth parameter, β.
[c] N_e is calculated from $\theta_{ML} = 4N_e\mu$, and N_t represents the effective population size at the TMRCA.
[d] Time to MRCA estimated from GENETREE using the generation time (25 years).

Our computational analyses suggest that the amino acid substitutions (SNP#13 and 26) have strongly destabilizing effects on the NOD2 protein structure (fig. 4), consistent with previous studies demonstrating the functional deficiency of NF-κB activity by in vitro assays (Bonen et al. 2003; Chamaillard et al. 2003). Furthermore, the frameshift mutation (SNP#35) is predicted to disrupt the binding ability of the LRRs domain (fig. 4). The LRRs domain is essential for bacterial response to LPS (Inohara and Nunez 2001). The important part of LRRs domain is located in the C-terminus, and the frameshift mutation results in a truncation of the protein from 1,040 to 1,007 amino acids that completely abolishes responsiveness to LPS (Ogura et al.

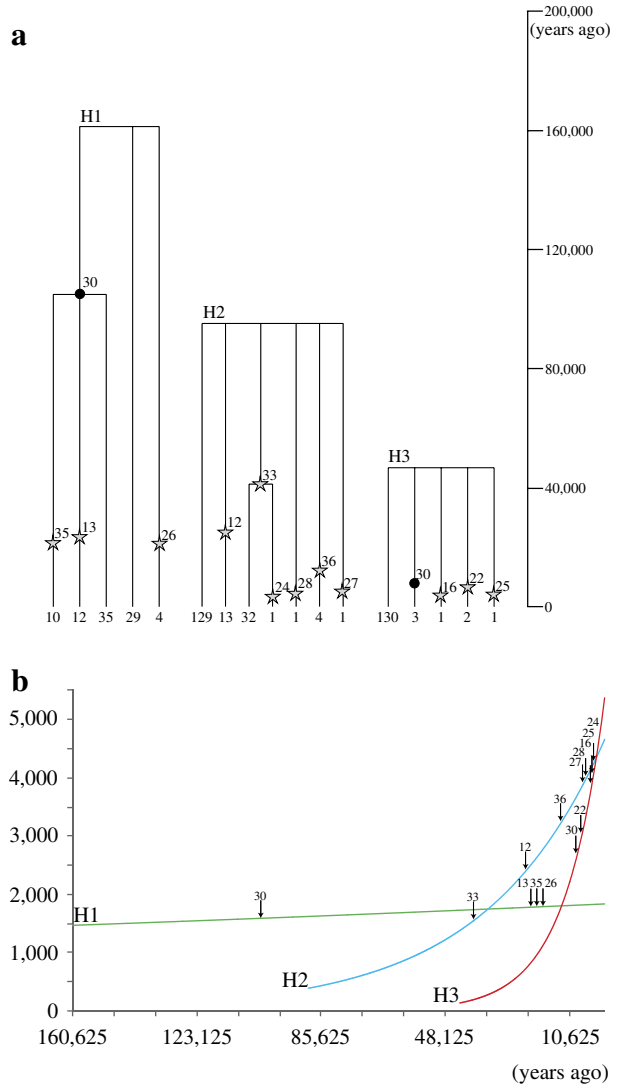


Fig. 6. (*a*) Gene trees for H1, H2, and H3 haplogroups in Europeans. The heights of the trees are proportional to the time to the most common recent ancestor (table 2). Black circles indicate mutations shared among Africans and Europeans, whereas stars indicate European-specific mutations. Numbers on the black circles or stars correspond to the SNP# listed in supplementary figure 4, Supplementary Material online. (*b*) Changes in population size and accumulation of mutations on H1 (dark green), H2 (light blue), and H3 (red) from the MRCA to the present in European populations. The horizontal axis indicates the time (years ago) and the vertical axis indicates the population size. Black arrows on the lines represent the number of years ago that the mutations appeared on H1, H2, and H3.

2001). Thus, these CD-risk alleles (CD-risk II: SNP#13, 26, and 35) have potential functional effects on the NOD2 protein and might be causal variants in Europeans. Although the destabilizing effect and the disruption of binding regions should be deleterious, the CD-risk alleles have remained in European populations (figs. 2 and 6). The results of the computational analyses also suggest that the Pro268*Ser* (SNP#7) substitution, a standing variant, affects the conformational flexibility of the linker between CARDs and NBD (fig. 4) and may influence the mutual mobility of these domains, which is likely to have a functional effect. In spite of its predicted effects, this allele along with the other CD-risk alleles of standing variation (SNP#9 and 31) define the root of H1 genealogy and exist as common alleles in European populations (figs. 2 and 3). Therefore, it is likely that the CD-risk alleles were not only originally on H1 but that more causative alleles accumulated on it, suggesting some advantage of H1 to overcome its deleterious effects.

Natural Selection on the H1 in European Populations

The results of our analyses lead us to hypothesize a scenario of natural selection on a standing variation, including a positive selection known as a soft sweep, at the *NOD2* locus. Under such a selective scenario we may, although not necessarily, observe that the advantageous allele is present on multiple haplotypes, whereas a particular haplotype involving the allele is increased in Europeans. Our data meet these conditions. First, our coalescent simulation shows that the frequency distribution of H1 significantly deviates from neutral expectations, suggesting H1 has experienced positive selection in Europeans (fig. 1*b*). Second, the CD-risk alleles define the root and the internal branch of H1 haplogroup genealogy (CD-risk I: SNP#7, 9, 30, and 31) and are found in Africans, albeit at low frequencies, and independently on non-H1 haplotypes (supplementary table 4, Supplementary Material online), suggesting the CD-risk I alleles are standing variants.

If H1 experienced positive selection, its frequency should have rapidly increased in Europeans. However, H1 has almost become extinct in Africans (fig. 2) and has not been affected by European demographic expansion (table 2 and fig. 6). Taken together, it is likely that H1 has experienced negative selection rather than positive selection. Given such a discrepancy, TMRCA estimates often provide an alternate scenario. Our estimation of TMRCA shows the deep coalescent time of the H1 haplogroup before human migration out of Africa (table 2 and fig. 6). Although it has been suggested that the method we used to determine the TMRCAs always yields results with large variances (Griffiths and Tavare 1994a; Tang et al. 2002), the TMRCAs for Europeans reveal that H1 is relatively older than H2 and H3, both of which existed in Africa (fig. 2, supplementary table 3, Supplementary Material online). Combining the best estimate of the H1 haplogroup age and the observation that CD-risk alleles and H1 are present in Africa reveals that the CD-risk I alleles and H1 were present in humans prior to human dispersal out of Africa. The existence of H1 since before the Out-of-Africa migration indicates that its constant frequency may be advantageous in European populations. From these results, we speculate that the best scenario of natural selection at the *NOD2* locus would be a "balancing selection on a standing variation."

Our data support a pattern of balancing selection. The relatively high nucleotide diversity in Europeans compared with Africans and East Asians (table 1) represents the abundance of intermediate frequency alleles, indicating that H1

has been maintained by balancing selection (Hudson and Kaplan 1988; Kreitman and Di Rienzo 2004; Andres et al. 2009) rather than directional selection. In contrast, the negative (but not significant) TD (table 1) in Europeans agrees with the expansion and accumulation of rare variants on H2 and H3 but not on H1 (table 2 and fig. 6). The ages of European-specific mutations range from old (>10,000 ya; agricultural expansion) to recent (<10,000 ya) (fig. 6*b*), and hence the excess of European-specific mutations could be the result of joint effects of population expansion on H2 and H3 and of balancing selection on H1.

Adaptive Function of H1

In the soft sweep model, it is assumed that standing variation becomes selectively favored and sweeps up in frequency, but before the change in selection, the variation was neutral or mildly deleterious (Pritchard et al. 2010). Our analysis suggests that H1 was likely to have experienced negative selection before humans migrated out of Africa (figs. 2, 4, and 6). Given that natural selection operates on diploids (Morton et al. 1956), there are two possibilities to explain the adaptation mechanism of the NOD2-H1 type. Either H1 with H2 (or H3) (heterozygotes) is more advantageous than homozygotes of H3 (or H2), indicating a "heterozygote advantage" (Pasvol et al. 1978), or the intergenic interactions under H1 work better than those under H2 and/or H3, indicating "polygenic adaptation" (Pritchard and Di Rienzo 2010; Pritchard et al. 2010). NOD2 is known to induce the host response to microbial pathogens or viral infection through interactions with other proteins involved in the NF-κB pathway (Ting et al. 2010). The three domains (CARDs, NBD, and LRRs) of the NOD2 protein function independently in the pathway. LRR is required to act as an intermediate between the bacterial products and NOD2 by binding directly or indirectly to the cognate peptidoglycan components. Subsequent to this activation step, CARD associates with the receptor interacting protein-2 (RIP2) through homotypic interactions, which activates the downstream cascade (i.e., IκB kinase complex) of the NF-κB pathway. This mechanism of NOD2 signaling implies that the conformational flexibility predicted from Pro268*Ser* (SNP#7) may result in some functional benefits for the NOD2 protein, which in some environments consequently overcomes the deleterious effects of the NOD2-H1 type itself. Thus, we speculate that balancing selection (or a complex effect similar to balancing selection) at the *NOD2* locus has operated in Europeans. Determination of the functional effect of this substitution would provide the evidence necessary to address this hypothesis.

Conclusion

H1 seems to have originated in Africa and been continuously maintained in European populations since the Out-of-Africa migration because it has an advantageous effect in European populations. This is an example of natural selection on standing variation (Hermisson and Pennings 2005), which is more difficult to detect than the classical selective sweep (Innan and Kim 2004). Nevertheless, our analysis of *NOD2* shows that the high-risk CD alleles have been maintained by natural selection (most likely balancing selection) on the standing variation whereby a particular deleterious haplotype is advantageous in diploid individuals either due to heterozygotic advantage and/or intergenic interactions. Such complexity is a remarkable feature of this disease, and we have shown a novel mechanism for maintaining mildly deleterious mutations with substantial frequency in a particular geographic population that does not involve genetic drift or population demography. While we have focused on *NOD2*, other loci are also of interest. It has recently been reported that another CD-associated gene has experienced a selective sweep (Huff et al. 2012). Further investigation, such as deep sequencing analysis, would be required in order to identify causative variants in non-Europeans and understand the population-specific susceptibility to CD and other complex diseases. In terms of preventive medicine, we suggest that it is crucial to survey local populations to identify risk alleles of complex diseases that are specific to geographical regions.

Supplementary Material

Supplementary figures 1–8, tables 1–5, and appendices I and II are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Takafumi Katsumura, Dr Yutaka Suzuki, and Dr Sumio Sugano in the Graduate School of Frontier Sciences, University of Tokyo, for his assistance of protein structure analysis and their technical assistance in DNA sequencing, respectively. We also thank three anonymous reviewers for many helpful comments and suggestions. S.N. was supported by a Grant-in-Aid for the Japan Society for the Promotion of Science (JSPS) Research fellow (21-7453). L.K. was supported by the 7th Framework Programme of the European Union (grant HEALTH-PROT, contract number 229676). J.M.B. was supported by the Polish Ministry of Science and Higher Education (grant POIG.02.03.00-00-003/09). H.S. was supported by Priority Area "Comparative Genomics" (20017021) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by a Grant-in-Aid for Scientific Research (C) from JSPS (21590360). J.R.K. and K.K.K. were supported in part by GM057672. S.K. was supported by a Grant-in-Aid for Scientific Research (A) from JSPS (22247036). H.O. was supported by a Grant-in-Aid for Scientific Research (B) from JSPS (21370108).

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* 7:248–249.

Andres AM, Hubisz MJ, Indap A, et al. (12 co-authors). 2009. Targets of balancing selection in the human genome. *Mol Biol Evol.* 26:2755–2764.

Balding DJ, Bishop M, Cannings C. 2007. Handbook of statistical genetics. 3rd ed. Chichester (UK): John Wiley & Sons.
Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16:37–48.
Barrett JC, Hansoul S, Nicolae DL, et al. (61 co-authors). 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet.* 40:955–962.
Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 40:695–701.
Bonen DK, Ogura Y, Nicolae DL, et al. (12 co-authors). 2003. Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. *Gastroenterology* 124:140–146.
Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22:2729–2734.
Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33:W306–W310.
Chamaillard M, Philpott D, Girardin SE, et al. (13 co-authors). 2003. Gene-environment interaction modulated by allelic heterogeneity in inflammatory diseases. *Proc Natl Acad Sci U S A.* 100:3455–3460.
Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62:1125–1132.
Croucher PJ, Mascheretti S, Hampe J, et al. (12 co-authors). 2003. Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. *Eur J Hum Genet.* 11:6–16.
Di Rienzo A. 2006. Population genetics models of common diseases. *Curr Opin Genet Dev.* 16:630–636.
Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* 21:596–601.
Di Rienzo A, Wilson AC. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci U S A.* 88:1597–1601.
Duerr RH, Taylor KD, Brant SR, et al. (23 co-authors). 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314:1461–1463.
Economou M, Pappas G. 2008. New global map of Crohn's disease: genetic, environmental, and socioeconomic correlations. *Inflamm Bowel Dis.* 14:709–720.
Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. 2004. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol.* 99:2393–2404.
Franke A, Hampe J, Rosenstiel P, et al. (25 co-authors). 2007. Systematic association mapping identifies NELL1 as a novel IBD disease gene. *PLoS One.* 2:e691.
Franke A, McGovern DP, Barrett JC, et al. (96 co-authors). 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* 42:1118–1125.
Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
Gilis D, Rooman M. 2000. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng.* 13:849–856.
Glas J, Seiderer J, Tillack C, et al. (16 co-authors). 2010. The NOD2 single nucleotide polymorphisms rs2066843 and rs2076756 are novel and common Crohn's disease susceptibility gene variants. *PLoS One.* 5:e14466.
Griffiths RC. 2007. GENETREE version 9.0. Available from: http://www.stats.ox.ac.uk/~griff/software.html
Griffiths RC, Tavare S. 1994a. Simulating probability distributions in the coalescent. *Theor Popul Biol.* 46:131–159.
Griffiths RC, Tavare S. 1994b. Ancestral inference in population genetics. *Stat Sci.* 9:307–319.
Gu S, Pakstis AJ, Kidd KK. 2005. HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* 21:3938–3939.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.
Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. 2007. Evidence of positive selection on a class I ADH locus. *Am J Hum Genet.* 80:441–456.
Hancock AM, Clark VJ, Qian Y, Di Rienzo A. 2011. Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol Biol Evol.* 28:601–614.
Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169:2335–2352.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 106:9362–9367.
Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* 120:831–840.
Huff CD, Witherspoon D, Zhan Y, et al. (18 co-authors). 2012. Crohn's disease and genetic hitchhiking at IBD5. *Mol Biol Evol.* 29(1):101–111. doi:10.1093/molbev/msr151
Hugot JP, Chamaillard M, Zouali H, et al. (20 co-authors). 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411:599–603.
Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci U S A.* 101:10667–10672.
Inohara N, Nunez G. 2001. The NOD: a signaling module that regulates apoptosis and host defense against pathogens. *Oncogene* 20:6473–6481.
The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1299–1320.
The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
Kidd KK, Pakstis AJ, Speed WC, Kidd JR. 2004. Understanding human DNA sequence variation. *J Hered.* 95:406–420.
Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM. 2003. A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins* 53(Suppl 6):369–379.
Kreitman M, Di Rienzo A. 2004. Balancing claims for balancing selection. *Trends Genet.* 20:300–304.
Kugathasan S, Baldassano RN, Bradfield JP, et al. (30 co-authors). 2008. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet.* 40:1211–1215.
Kurowski MA, Bujnicki JM. 2003. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* 31:3305–3307.
Laurie AT, Jackson RM. 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21:1908–1916.
Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of

molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25:2744–2750.
Libioulle C, Louis E, Hansoul S, et al. (23 co-authors). 2007. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 3:e58.
Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.
Lohmueller KE, Indap AR, Schmidt S, et al. (12 co-authors). 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451:994–997.
Mathew CG. 2008. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Rev Genet.* 9:9–14.
Morton NE, Crow JF, Muller HJ. 1956. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc Natl Acad Sci U S A.* 42:855–863.
Myles S, Davison D, Barrett J, Stoneking M, Timpson N. 2008. Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics.* 1:22.
Nakagome S, Takeyama Y, Mano S, Sakisaka S, Matsui T, Kawamura S, Oota H. 2010. Population-specific susceptibility to Crohn's disease and ulcerative colitis; dominant and recessive relative risks in the Japanese population. *Ann Hum Genet.* 74:126–136.
Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res.* 11:863–874.
NHLBI Program for Genomic Application. [cited 2010 Feb 13]. Available from: http://pga.gs.washington.edu
NIEHS Environmental Genome Project. [cited 2010 Feb 13]. Available from: http://egp.gs.washington.edu
Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18:83–90.
Novembre J, Johnson T, Bryc K, et al. (12 co-authors). 2008. Genes mirror geography within Europe. *Nature* 456:98–101.
Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet.* 40:646–649.
Ogura Y, Bonen DK, Inohara N, et al. (17 co-authors). 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 411:603–606.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.
Oota H, Pakstis AJ, Bonne-Tamir B, et al. (14 co-authors). 2004. The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Ann Hum Genet.* 68:93–109.
Parthiban V, Gromiha MM, Schomburg D. 2006. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* 34:W239–W242.
Pasvol G, Weatherall DJ, Wilson RJ. 1978. Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria. *Nature* 274:701–703.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.
Pickrell JK, Coop G, Novembre J, et al. (11 co-authors). 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.
Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 69:124–137.
Pritchard JK, Di Rienzo A. 2010. Adaptation—not by sweeps alone. *Nat Rev Genet.* 11:665–667.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:R208–R215.
Raelson JV, Little RD, Ruether A, et al. (25 co-authors). 2007. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc Natl Acad Sci U S A.* 104:14747–14752.
Rioux JD, Xavier RJ, Taylor KD, et al. (25 co-authors). 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet.* 39:596–604.
Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. 2010. Genome-wide association studies in diverse populations. *Nat Rev Genet.* 11:356–366.
Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
Soding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.
Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW. 2002. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* 161: 447–459.
Ting JP, Duncan JA, Lei Y. 2010. How the noninflammasome NLRs function in the innate immune system. *Science* 327:286–290.
Tishkoff SA, Williams SM. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet.* 3:611–621.
VanLiere JM, Rosenberg NA. 2008. Mathematical properties of the r2 measure of linkage disequilibrium. *Theor Popul Biol.* 74:130–137.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.
Yazdanyar S, Kamstrup PR, Tybjaerg-Hansen A, Nordestgaard BG. 2010. Penetrance of NOD2/CARD15 genetic variants in the general population. *CMAJ.* 182:661–665.
Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics.* 7:166.
Zaahl MG, Winter T, Warnich L, Kotze MJ. 2005. Analysis of the three common mutations in the CARD15 gene (R702W, G908R and 1007fs) in South African colored patients with inflammatory bowel disease. *Mol Cell Probes.* 19:278–281.