Current Biology, Vol. 14, 1531–1539, September 7, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.cub.2004.08.051

Positive Selection on *MMP3* Regulation Has Shaped Heart Disease Risk

Matthew V. Rockman,[1,*] Matthew W. Hahn,[1,2] Nicole Soranzo,[3] Dagan A. Loisel,[1] David B. Goldstein,[3] and Gregory A. Wray[1]
[1]Department of Biology
Duke University
Box 90338
Durham, North Carolina 27708
[2]Center for Population Biology
University of California, Davis
Davis, California 95616
[3]Department of Biology
University College London
Gower Sreet
London WC1E 6BT
United Kingdom

Summary

Background: The evolutionary forces of mutation, natural selection, and genetic drift shape the pattern of phenotypic variation in nature, but the roles of these forces in defining the distributions of particular traits have been hard to disentangle. To better understand the mechanisms contributing to common variation in humans, we investigated the evolutionary history of a functional polymorphism in the upstream regulatory region of the *MMP3* gene. This single base pair insertion/deletion variant, which results in a run of either 5 or 6 thymidines 1608 bp from the transcription start site, alters transcription factor binding and influences levels of *MMP3* mRNA and protein. The polymorphism contributes to variation in arterial traits and to the risk of coronary heart disease and its progression.

Results: Phylogenetic and population genetic analysis of primate sequences indicate that the binding site region is rapidly evolving and has been a hot spot for mutation for tens of millions of years. We also find evidence for the action of positive selection, beginning approximately 24,000 years ago, increasing the frequency of the high-expression allele in Europe but not elsewhere. Positive selection is evident in statistical tests of differentiation among populations and haplotype diversity within populations. Europeans have greater arterial elasticity and suffer dramatically fewer coronary heart disease events than they would have had this selection not occurred.

Conclusions: Locally elevated mutation rates and strong positive selection on a *cis*-regulatory variant have shaped contemporary phenotypic variation and public health.

Introduction

The genetic basis of human variation must ultimately be due to mutation, natural selection, and demographic factors such as migration and genetic drift in small populations. How these factors conspire to generate the observed distribution of phenotypes—whether variation is disproportionately due to mutational hot spots, whether patterns of variation among populations are due primarily to migration and drift or due to local selective regimes, and whether functional variation is the weakly deleterious relict of small ancient human populations or the positively selected raw material for adaptation—remains unanswered, despite its centrality to ambitions of mapping and characterizing the genes underlying complex traits. We focused our attention on a functional variant in the heart disease gene *MMP3*, encoding the matrix metalloproteinase stromelysin, to determine whether the phenotypic variation resulting from this polymorphism represents ordinary neutral variation or whether it is instead a mutational hot spot or a target for selection. The polymorphism is one of the best-characterized functional variants in humans, both at the biochemical and organismal level, and its study therefore offers a rare opportunity to bridge the disconnect between molecular and phenotypic evolution.

Biochemical and clinical studies have established the functional importance of the *MMP3* 5T/6T polymorphism, 1608 bp from the gene's transcription start site (dbSNP rs3025058). The polymorphism falls in a region bound by at least three different protein complexes, which contain the zinc-finger transcription factor ZNF148 and the NF-κB dimers p50/p50 and p50/p65 [1, 2]. The transcriptional effects of these proteins are mediated by protein-protein interactions between TCF20, which binds an adjacent element, and JUN, which binds in the proximal promoter, near the transcription start site [3]. The 5T allele binds with lower affinity than the 6T allele to at least one transcription complex, now identified as the p50/p50 NF-κB dimer [1, 4]. In reporter assays in cultured cells, the 5T allele drives higher levels of expression than the 6T allele [4, 5], implying that p50/p50 acts as a transcriptional repressor, perhaps by interfering with ZNF148, whose overexpression upregulates *MMP3* transcription [2], or by competing with p65/p50 heterodimers. The effect of the polymorphism seen in reporter assays has been confirmed by a strong association between 5T genotype and levels of both *MMP3* transcript and protein in vivo [6, 7].

Allelic differences in *MMP3* transcription translate into measurable differences in phenotypes, both in health and in disease. *MMP3* breaks down extracellular matrix components and plays a critical role in vascular tissue remodeling, mediating the balance between matrix accumulation and degradation. This balance is best characterized with respect to arterial phenotypes, where significant associations between 5T genotype and characteristics of the carotid and coronary arteries of healthy individuals, including elasticity and thickness of the arterial walls, have been found repeatedly [7–10]. Unsurprisingly, the *MMP3* polymorphism is implicated in heart disease. The low expression allele, on the matrix accumulation side of the balance, is implicated in risk of coronary artery disease, characterized by the buildup

*Correspondence: mrockman@duke.edu

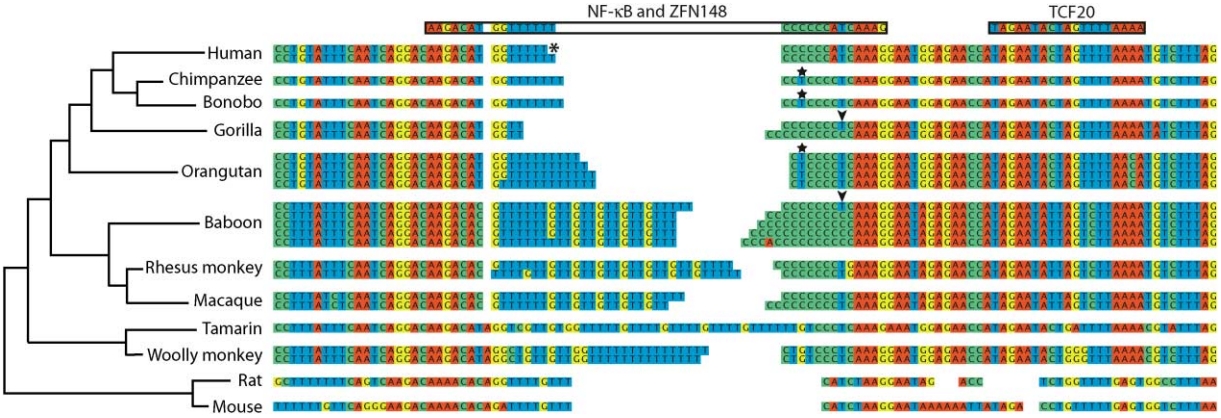

Figure 1. Alignment of the Functional Region around the –1608 SNP (*) Shows Its Rapid Evolution

Each unique allele is shown. A shared polymorphism is indicated by arrowheads, and a mutation likely to affect ZNF148 binding is indicated by stars. The transcription factor binding sites characterized biochemically in humans [1, 2, 4, 58, 59] are indicated above the alignment. Relationships among the species are shown by the phylogeny to the left of the alignment (branch lengths are arbitrary).

of atherosclerotic plaque, and in its progression and recurrence (as restenosis) after treatment [4, 5, 11–18]. Yet, because *MMP3* is expressed by macrophages within the atherosclerotic plaques typical of coronary artery disease [19], the high expression allele, by degrading matrix, promotes plaque instability or rupture, and thus myocardial infarction and aneurysm [5, 20–22]. The importance of the MMP3 balance in heart disease is corroborated by results from an atherosclerotic mouse model: inactivation of *MMP3* in these mice increases atherosclerotic plaque accumulation while reducing aneurysm [23].

We investigated the evolution of the *MMP3* polymorphic region, in both phylogenetic and population genetic contexts, to understand how mutation, selection, and demography contribute to phenotypic variation in health and disease. We show that both genomic variation in mutation rates and geographic variation in selective regimes have influenced *MMP3* regulation and shaped common variation in complex traits.

Results

Rapid Evolution and Elevated Polymorphism Among Primates

We first characterized the region around the polymorphic site in nonhuman primates, including chimpanzee (22 chromosomes), bonobo (8), gorilla (2), orangutan (10), baboon (74), rhesus monkey (2), pigtailed macaque (2), mustached tamarin (2), and woolly monkey (2). While humans are polymorphic for a run of five or six Ts, chimpanzees and bonobos possess seven Ts, gorillas two, and orangutans nine to twelve. This poly-T tract is the remnant of an ancient, complicated GT_n repeat that survives in the Old and New World Monkeys (Figure 1). Differences among species are also evident in the tract of cytosines that follows the Ts. Moreover, seven of the ten sampled primate species exhibit intraspecific polymorphism in this T_nC_n region. The rate of insertion and deletion events in the region, in a sample of 22 great ape and baboon chromosomes (see Supplemental Experimental Procedures), is greater than the rate of such events in the remainder of the 1.8 kb *cis*-regulatory region by a factor of 70, and the nucleotide substitution rate is elevated by a factor of 3.

The elevated level of variation observed in this functionally important region may be attributed to either increased maintenance of variation (balancing selection) or increased input of variation (hypermutation). The unusual pattern of shared states, such as a C/T polymorphism segregating at a homologous site (Figure 1, arrowheads) in both gorilla and baboon, is consistent with variation being maintained over long periods of time by balancing selection. To test this model, we generated a phylogeny of 1.8 kb promoter haplotypes representing multiple alleles from the great apes and baboons and determined that the intraspecific variation is of recent origin; the shared polymorphisms do not occur on shared haplotypes, even over distances of tens of nucleotides, implying recurrent mutation and not long term maintenance (Figures 1 and 2A).

Empirical support for elevated mutation rates comes from studies of somatic mutation. DNA mismatch-repair defects causing microsatellite instability are known to result in elevated somatic mutation rates precisely in the T_nC_n region of the *MMP3* promoter [24]. While the normal segregating variation at this locus in humans involves only one base pair, the phylogenetic and somatic mutation data indicate that this locus behaves as a cryptic microsatellite with respect to mutational processes.

Because several different protein complexes bind to this region, the interspecific differences and polymorphisms are likely to result in species- and allele-specific transcriptional regulation. The C to T substitution in chimpanzee and bonobo (and independently in orangutan) at the position corresponding to human –1605 (Figure 1, stars) is likely to abrogate binding of the activator ZNF148, which depends on the run of Cs for binding [2]. The ZNF148 protein sequence is identical between human and chimpanzee (Chimpanzee Genome Sequencing Consortium, 13 Nov. 2003 assembly). Empirical support for important differences among species

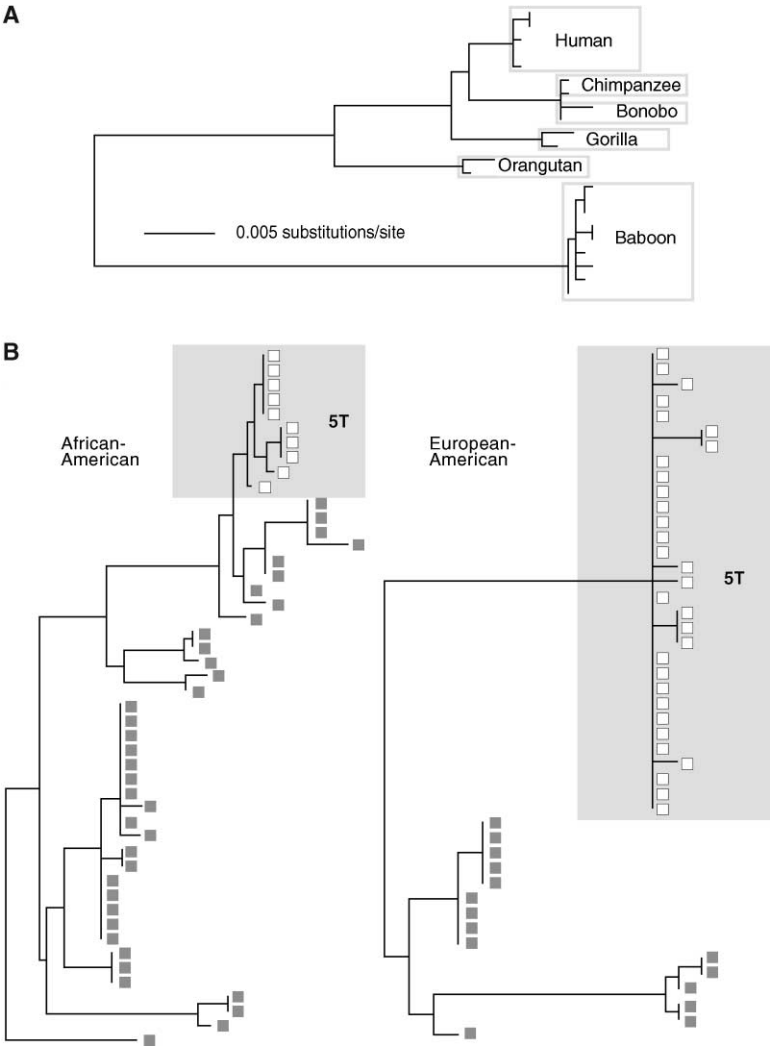

Figure 2. Evolution of the *MMP3* Promoter

(A) Maximum likelihood phylogeny of 1.8 kb *MMP3* promoter haplotypes.

(B) Genealogies of 11.9 kb human *MMP3* haplotypes, rooted by chimpanzee sequence. Clades characterized by the 5T allele are boxed. The derived position of the 5T clade indicates that the 5T allele arose from a 6T allele.

comes from a study of gene expression in cultured primary fibroblasts, which found that *MMP3* is expressed at dramatically lower levels (>6 fold) in bonobos than in humans [25].

Due to the rapid evolution of the T_n tract, the ancestral state of the human promoter polymorphism cannot be confidently inferred from the primate sequences. Human haplotype data from the SeattleSNPs database, however, indicate that the 5T allele arose from a 6T allele; the 5T haplotypes are nested withing several clades of 6T-bearing haplotypes (Figure 2B). The 5T allele thus represents an evolutionary loss of the repressive interaction between the DNA and the p50/p50 NF-κB dimer.

Differentiation Among Populations

We next investigated whether the functional variation in humans can be attributed to positive selection, or whether the mutation causing the loss of the NF-κB binding site is neutral or weakly deleterious, subject to gradual elimination from the population. We genotyped the polymorphism in six human populations (Cameroon, China, Ethiopia, India, Southern Italy, and Papua New Guinea) in order to investigate the pattern of genetic differentiation among populations. Under neutrality, differentiation reflects the demographic history of genetic drift and migration, a history common to all autosomal loci. Loci under selection, however, will show deviations from the neutral pattern: elevated differentiation when selection differs among populations and reduced differentiation when balancing selection maintains similar allele frequencies among populations [26, 27]. We compared genetic differentiation (measured with F_{ST}, which ranges from 0 to 1) at the 5T/6T site to that at 18 mutually unlinked single nucleotide polymorphisms (SNPs), each chosen to be more than 200 kb from any known gene and thus unlikely to be affected by selection. The presumed neutral polymorphisms provide an empirical estimate of the F_{ST} expected in the absence of selection. An empirical distribution based on candidate neutral loci is necessary both because there is no robust theoretical expectation and because the fraction of loci under selection is unknown, rendering random loci unsuitable for estimating the tails of a neutral distribution [28]. Additionally, analysis of random loci has shown that F_{ST}s are on average lower in coding than noncoding regions, consistent with the action of stronger purifying selection on coding

sequences [27]; these findings underscore the value of estimating the neutral distribution from noncoding candidate neutral SNPs and also point to the conservatism of such an approach. The observed *MMP3* frequencies fit the allele frequency criteria employed in our selection of neutral SNPs (see Experimental Procedures), suggesting that our SNP ascertainment strategy does not bias our results. As a test statistic, we considered the difference between the F_{ST} at *MMP3* and the 18-locus F_{ST} for the same pair of populations. We estimated the neutral distribution of the test statistic by bootstrap [28].

The pairwise F_{ST}s at the *MMP3 cis*-regulatory polymorphism (Table S1) are within the ranges attributable to drift, but the value for the Southern Italy-Cameroon comparison approaches significance ($F_{ST} = 0.392$, $p = 0.055$). Our Southern Italian sample has the highest 5T allele frequency among the six populations, 0.42, yet it is lower than that observed in European populations from further to the north, including Czech (0.48 [29]), Northern Italian (0.50 [17]), German (0.51 [14]), British (0.52 [12]), and Swedish (0.54 [29]) populations. We therefore added neutral marker data from a British population to test whether the 5T frequency in northern Europe is higher than that expected under neutrality. The *MMP3* 5T/6T site exhibits F_{ST}s significantly higher than at neutral sites in pairwise comparisons between England and Cameroon ($F_{ST} = 0.365$, $p = 0.001$) and between England and India ($F_{ST} = 0.215$, $p = 0.015$). These elevated F_{ST}s are improbable if the 5T allele is neutral or deleterious in northern Europe; instead, they suggest a role for positive selection raising the local frequency of the derived 5T allele.

Haplotypic Signature of Positive Selection

In order to test the positive selection hypothesis, we looked at the pattern of haplotypic variation within populations. We analyzed haplotypes of the *MMP3* locus sampled from healthy unrelated European- and African-Americans in the SeattleSNPs database [30] (23 and 24 individuals, respectively). The haplotypes span 11,903 bp, encompassing the full genomic extent of the *MMP3* transcript and 2.3 kb of flanking sequence from each end. The SeattleSNPs data were generated by complete sequencing of the 11.9 kb region from each sample.

Positive selection among Europeans would result in a rapid increase in frequency of the 5T allele, with a concomitant decrease in linked variation [31]. The European-American sample exhibits such a pattern, with significant deviations from neutral equilibrium indicating a deficit of haplotypes (Depaulis and Veuille's $K = 13$, $p = 0.006$, and Fu's $F_s = 2.57$, $p = 0.008$ [32, 33]) and a reduction in haplotype heterozygosity (Depaulis and Veuille's $H = 0.76$, $p < 0.00001$). The pattern is due largely to an excess of a single common 5T haplotype: 22 of the 46 sampled haplotypes are identical, despite the presence of 35 polymorphisms in the whole sample. Hudson's haplotype test [34] uses coalescent simulations to give the probability of seeing 22 identical haplotypes when 35 mutations are present in a genealogy of 46 samples; this test was highly significant ($p = 0.0095$), supporting the model of local adaptation through a partial or ongoing selective sweep among Europeans. In contrast, the African-American sample exhibits no deviations from neutral expectations in any of the statistical tests.

The observed departures from the neutral equilibrium model may be due to selection or to nonequilibrium demography in the European-American sample. Europeans are known to have experienced a recent population expansion, but this demographic effect will result in an excess of haplotypes in a sample, contrary to the observed pattern [33]. A dramatic bottleneck could produce a deficit of haplotypes, but such an event should affect both 5T and 6T haplotypes; at *MMP3*, only the 5T allelic class shows reduced nucleotide diversity and overrepresentation of a single haplotype.

Another way to determine whether the unusual pattern of variation at *MMP3* is due to selection or demography is to consider other loci. Because demography is shared among loci, while selective effects are unique to each locus, other loci sampled from the same individuals represent an empirical control for demographic effects. For a random sample of 50 loci in the SeattleSNPs database, we calculated the P values for Hudson's haplotype test; because recombination varies among the loci, we incorporated locus-specific recombination rates into the simulations. The P value for *MMP3* is lower than that found for 48 of the 50 loci. Thus, the striking patterns at *MMP3* do not appear to be genome-wide effects of demographic history but are instead locus-specific phenomena. It is important to note that the SeattleSNPs loci, unlike the loci included in our analysis of F_{ST}, are not candidate neutral loci. The distribution of their P values is therefore a conservative control for demography, because the tails of the distribution are likely to contain genes whose departures from expectation are due to selection. The SeattleSNPs loci are candidate genes for inflammation disorders, and most of the genes have been implicated by linkage or association in variation in disease susceptibility, which makes them probable targets of selection. Indeed, a systematic literature survey reveals independent evidence for positive selection on at least eight of the 50 genes, including those with the lowest and third-lowest P values, *IL4* and *IL13* [28, 35]. All of the evidence taken together, therefore, points to natural selection acting to increase the frequency of the 5T mutation within Europe.

Allele Age

The number of mutations arising on the 5T haplotype in Europe provides us with a means of estimating the age of the most recent common ancestor (MRCA) of the European-American 5T haplotypes [36]. The average branch length from the inferred MRCA of the sampled 5T haplotypes is 2.98×10^{-5} substitutions per site. From our primate *MMP3* data, we estimate the local mutation rate (excluding the hypermutable region) to be 1.26×10^{-9} per site per year, yielding an MRCA age of 23,700 years, roughly coincident with the last glacial maximum in Europe. The 95% confidence interval [37] encompasses ages from 10,800 to 36,600 years.

From the distributions of mutations on the European-American 5T genealogy, and from the global occurrence of the mutation, we infer that the 5T mutation predated a

change in selective regime that increased its frequency. Under such a model, some of the mutations on 5T haplotypes predate the inferred selection, and the ages given above are thus overestimates of the age of the selection event. Indeed, if we assume that only the singleton mutations arose after the selection event, then the age is estimated to be 8,600 years (2,200–17,200), which places it in the context of the neolithic agricultural revolution.

Discussion

Role of Mutation in Generating Phenotypic Variation

Although elevated mutation rates are known to underlie a number of rare pathological conditions [38], the role of locally elevated mutation in generating common variation is less appreciated. A large fraction of functional *cis*-regulatory polymorphism in humans is due to microsatellite and minisatellite variation [39]. Here, we have found that phenotypically penetrant single nucleotide polymorphisms may also be due to elevated spontaneous mutation rates relating to cryptic microsatellite-like DNA structure. Because the mutable *MMP3* promoter region interacts with several protein complexes in humans, its rapid evolution may imply correspondingly rapid evolution of DNA-protein interactions and thus transcriptional phenotypes. The prevalence of hypermutable sites in the genetic basis of phenotypic diversity is an important and underexplored parameter in understanding the forces that generate and maintain variation in populations and the genetic basis for evolutionary parallelisms [40, 41].

The rapid evolution of the polymorphic region also counsels against over-reliance on evolutionary conservation as a guide to the discovery of functional *cis*-regulatory DNA, despite the power of that approach [42, 43]. While conserved noncoding sequence is likely to represent functionally important regulatory DNA, the highly conserved fraction of regulatory DNA may not coincide with the fraction of regulatory DNA exploited by positive natural selection. Indeed, as selection acts on the heritable phenotypic variation generated by mutation, regions of elevated mutation rate may represent a disproportionate component of the genetic basis of selectable variation. The *cis*-regulatory variants underlying common variation may be those least likely to be evolutionarily conserved.

Phenotypic Consequences of Positive Selection

In much of the world, the phenotypic variation attributable to the 5T/6T polymorphism appears to have been shaped largely by demographic factors such as genetic drift. In Europe, however, the evidence suggests that positive selection played a role in increasing the frequency of the 5T allele. The shift in 5T frequencies shaped the physiological and morphological characteristics of European arteries in ways that can be directly estimated from data on associations between the *MMP3* promoter genotype and phenotypes in healthy individuals [7–10]. For example, age-associated large artery stiffening was strongly correlated with *MMP3* promoter genotype in a study of 203 healthy European-Australians [7]. In the study population, heterozygotes had the most elastic arteries. The relationship holds in both men and women and remains significant after controlling for known covariates. Although the phenotypic variance within each genotypic class is large, as for any complex trait, point estimates of the population mean arterial stiffness as a function of *MMP3* allele frequency can illustrate the predicted effect of the inferred selection event (Figure 3A). The European-derived study population is characterized by arteries that are on average more elastic than would be expected in the absence of positive selection, other things being equal.

Direct estimates of the effects of the polymorphism on disease risks suggest that they may also have been dramatically shifted by selection. Association studies have found that the derived 5T allele retards the progression of coronary artery disease while increasing the risk of myocardial infarction. To illustrate the potential magnitude of the public health consequences of the inferred selection, we estimated the fraction of coronary heart disease events—defined as sudden coronary death, myocardial infarction, or coronary artery surgery—prevented by the selected increase in 5T frequency. We used the genotypic relative risks estimated in the single most comprehensive study to date, a prospective study of middle-aged British men entailing almost 24,000 person-years of follow-up [16]. On balance, the derived allele is protective against coronary heart disease events in this population, although among smokers there is a strong interaction effect that increases myocardial infarction risk among 5T homozygotes; this genotype by environment interaction has itself been replicated [20]. Based on the genotype-specific relative risks, we estimate that these middle-aged British men would have suffered 43% more coronary heart disease events had the positive selection event not occurred (Figure 3B). If the relative risks were applicable to the general British population, one consequence of the partial selective sweep would be a reduction by more than 50,000 in the annual mortality from coronary heart disease in the UK [44].

What is the selective agent acting to increase the frequency of the 5T allele in Europe? Most likely, the arterial phenotypes we have considered are pleiotropic side effects of selection on some other consequence, currently unidentified, of upregulated *MMP3* expression [45]. MMP3 has many functions, including the degradation of collagens, fibronectin, laminins, and elastin, and the activation of other MMPs, which play critical roles in cell migration, proliferation, apoptosis, wound healing, and morphogenesis. Models to account for selection on the 5T allele must appeal to characteristics of the selective regime unique to Europe during and subsequent to the last Ice Age. Although coronary heart disease is considered a recent phenomenon, dependent on contemporary diets and behaviors, it is possible that the diet of Ice Age Europeans, rich in the atherogenic fats of large mammals, could have contributed to early onset coronary heart disease and hence a selective advantage for the high expression 5T allele. Investigations into the relationship between 5T allele frequencies and such environmental variables as diet may help identify the causes of selection.

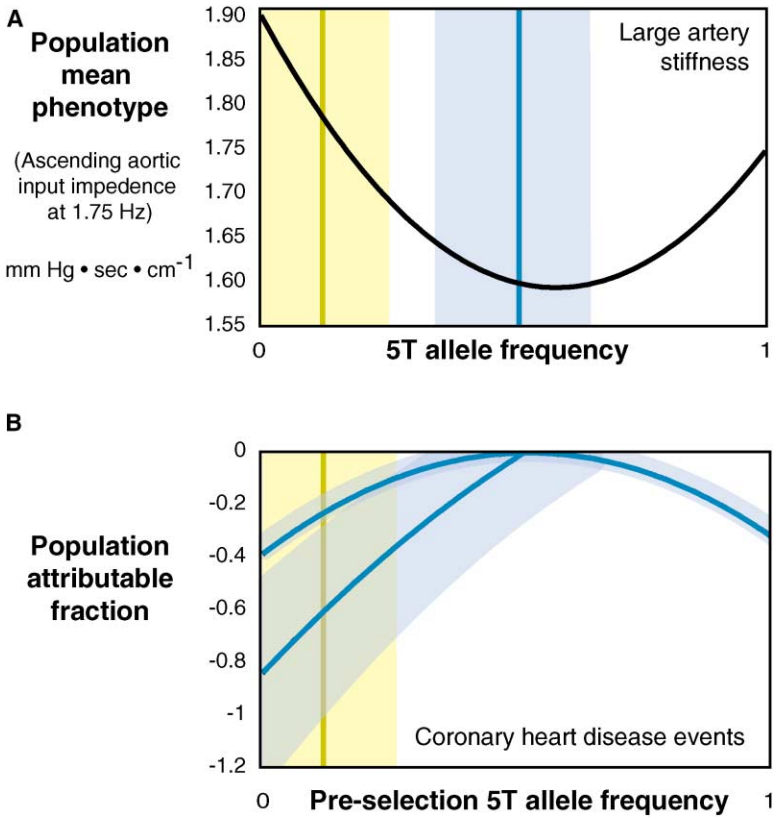

Figure 3. Consequences of Selection

(A) Population mean age-associated large artery stiffness [7] as a function of 5T allele frequency. The blue and yellow zones encompasses the known frequencies in European and non-European populations, respectively. The blue line indicates the British allele frequency (0.52), and the yellow line indicates the average non-European frequency (0.125), a proxy for the expected frequency in the absence of selection.

(B) Fraction of coronary heart disease events among middle-aged British men attributable to the selection event, as a function of the unknown preselection 5T frequency. Negative numbers indicate that the selective shift decreased the incidence of events. The postselection 5T frequency in England is represented by the blue lines, with attributable risk plotted separately for smokers (upper curve) and nonsmokers (lower curve). For example, the y-coordinate of intersection of the lower curve and the yellow line represents the fraction of coronary heart disease events due to the sweep among middle-aged male English nonsmokers, given a preselection 5T frequency of 0.125. The blue zone represents the range of curves for different European 5T frequencies.

Conclusion

Our results imply that a mutation that eliminates a transcription factor binding site and leads to an increase in gene expression has been favored by natural selection. These results cast light on three long-standing debates that converge in the study of human evolutionary functional genetics: the debate in the field of molecular evolution over the relative roles of neutral and advantageous mutations in contributing to polymorphism and divergence, that in medical genetics over the roles of advantageous, neutral, and deleterious mutations in contributing to variation in disease risks, and the debate in both fields over the relative importance of protein-coding versus *cis*-regulatory variation [43]. This study is one of an increasing number of studies to find positive selection on *cis*-regulatory mutations shaping phenotypic variation and disease risk in humans [28, 46–50], contrary to the notion that important mutations will tend to be deleterious mutations affecting protein structure. As evidence for positive selection corroborates inferences about the functional importance of nucleotide variants, we anticipate that in the future evolutionary analyses will form an essential complement to genetic association studies [27, 51].

Experimental Procedures

Primate Sequences

We PCR amplified and sequenced a 360 bp fragment of the *MMP3* promoter, using primers 5′-GGCTCCACTGTTTCTTCCTG and 5′-AAGATGCCCACACAGGTGAT, from both chromosomes of eleven chimpanzees (*Pan troglodytes*, subspecies *verus, vellerosus, schweinfurthii*), four bonobos (*Pan paniscus*), one gorilla (*Gorilla gorilla*), five orangutans (*Pongo pygmaeus*), 37 baboons (seven savannah baboons, *Papio cynocephalus*, from a wild population in Amboseli National Park, Kenya, and 30 guinea baboons, *P. papio*, from the inbred, captive population at the Brookfield Zoo), one pigtailed macaque (*Macaca nemestrina*), one rhesus monkey (*Macaca mulatta*), one common woolly monkey (*Lagothrix lagotricha*), and one red-chested mustached tamarin (*Saguinus labiatus*). Excluding the guinea baboons, the sampled animals are unrelated. Some *Pan* samples were gifts of A. Stone. Amboseli baboon samples were gifts of Jeanne Altmann and Susan Alberts. Brookfield baboon samples were provided by the Chicago Zoological Society. The other samples were purchased from the Coriell Institute (PR00002, PR000253, PR00251, NG06209, NG12256, NG06939, NG05253, NG05251, NA04272, NA03448, NA03450, NS03621, NS03657, NS03659, NG08452, NG07109, NG05356, and NG05308). PCR products were sequenced directly. In cases of ambiguous phasing, PCR products were cloned and multiple clones sequenced. Mouse and rat sequences were retrieved from the Rat Net and Mouse Net tracks of the UCSC Genome Browser (http://www.genome.ucsc.edu).

We also PCR amplified, cloned, and sequenced a 1.8 kb region of the *MMP3* promoter using primers 5′-GGCTCCACTGTTTCTTCCTG, 5′-CCTGAACAAGGTTCATGCTG. In this manner, we collected nine baboon haplotypes, (six from Amboseli, three from Brookfield), two orangutan, two gorilla, two bonobo, and two chimpanzee. We sequenced multiple identical clones of each haplotype as a check on PCR artifact. The sequenced region includes the 5′UTR and 35 bases of coding sequence.

Human Genotyping and F_{ST} Analysis

We obtained DNA from 45 unrelated individuals in each of seven populations: Southern Italy, Cameroon, Ethiopia, China (Singaporean Chinese), India (Uttar Pradesh), Papua New Guinea (Madang Coastal), and England (York). Human DNA samples were collected in the Goldstein lab with informed consent or were anonymized legacy collections provided to the Goldstein laboratory by collaborators from other academic research universities. Due to conditions on the use of the York sample, we did not type *MMP3* in that popula-

tion, but used 0.52 as the English 5T allele frequency; this is the frequency found in the study with the largest sample [12] and is similar to and intermediate between the frequencies found in two other studies, 0.49 [52] and 0.55 [16].

The –1608 region was PCR amplified using published primers and conditions [8], with the reverse primer fluorescently labeled. PCR products were then run on an ABI 3700 capillary gel machine and scored using Genotyper software (ABI). The 6T allele yielded a fragment of length 130 bp, versus 129 for the 5T allele. We sequenced 10% of samples to validate the genotyping.

The 18 neutral markers have been described previously [28]. SNPs were selected based on data from the SNP Consortium (http://snp.cshl.org/) to have a minor allele frequency of >0.3 in a European-derived population and >0.05 in African-American and Asian populations. The data are available at the Goldstein lab website (http://popgen.biol.ucl.ac.uk/).

F_{ST} was estimated, and its significance determined, as in [28]. Because pairwise F_{ST} values are not independent, a Bonferroni correction for multiple tests is conservative. Nevertheless, the England-Cameroon F_{ST} ($p < 0.001$) remains significant at the level implied by such a correction.

Haplotype Analyses

The haplotype genealogies in Figure 2B are based on the SeattleSNPs data (NHBLI Program for Genomic Applications, UWFHCRC, Seattle, WA; http://pga.gs.washington.edu; Oct. 28, 2003), with haplotypes estimated from genotypic data using PHASE [53] version 1.0. Trees were generated using neighbor-joining on counts of pairwise differences in PAUP*. These phased haplotypes were also used for the haplotype tests described in the text. Two singleton mutations occurred in 5T/6T heterozygotes; we adopted a conservative approach of placing these on the 5T haplotypes. If these singletons actually reside on the 6T haplotypes, our haplotype test P values would be more extreme, and the age of the selective sweep would be more recent. Calculations of *H, K*, and *F* were performed in DNAsp [54]. We performed these analyses under the assumption of no recombination, which is conservative with respect to these tests and is consistent with the absence of clear recombinant haplotypes in the 5T sample. Coalescent simulations to perform Hudson's haplotype test were implemented using Hudson's *ms* [55]. The locus-specific population recombination parameters are derived from the deCode genetic map [56]. We used an effective population size of 7,500, which is the usual human autosomal N_e estimate of $\sim$10,500 [57] rescaled to reflect the observation that the SeattleSNPs European-American sample heterozygosity (π) is 72.5% of the African-American sample π [30].

Supplemental Data

Supplemental Data including additional details of the experimental procedures and a table showing MMP3 allele frequencies and pairwise F_{ST} values are available at http://www.current-biology.com/cgi/content/full/14/17/1531/DC1.

Acknowledgments

We thank the Chicago Zoological Society for samples from *P. papio* baboons at Brookfield Zoo, Susan Alberts and Jeanne Altmann for samples from *P. cynocephalus* of Amboseli, Kenya, and Anne C. Stone for *Pan* samples. The Amboseli samples were collected under funding from the National Science Foundation and the Chicago Zoological Society, with permission from the Office of the President of the Republic of Kenya, and with support and assistance from the Institute for Primate Research, National Museums of Kenya, and the Kenya Wildlife Service. We thank Corbin Jones, Chuck Langley, and the Duke PopBio group for valuable comments. We gratefully acknowledge the support of the NSF (M.V.R., M.W.H., G.A.W.), NASA (G.A.W.), and the Leverhulme Trust (D.B.G. and N.S.). D.B.G. is a Royal Society/Wolfson Research Merit Award Holder.

Received: June 15, 2004
Revised: July 19, 2004
Accepted: July 19, 2004
Published: September 7, 2004

References

1. Borghaei, R.C., Rawlings, P.L., Jr., Javadi, M., and Woloshin, J. (2004). NF-κB binds to a polymorphic repressor element in the MMP-3 promoter. Biochem. Biophys. Res. Commun. *316*, 182–188.
2. Ye, S., Whatling, C., Watkins, H., and Henney, A. (1999). Human stromelysin gene promoter activity is modulated by transcription factor ZBP-89. FEBS Lett. *450*, 268–272.
3. Kirstein, M., Sanz, L., Quinones, S., Moscat, J., Diaz-Meco, M.T., and Saus, J. (1996). Cross-talk between different enhancer elements during mitogenic induction of the human stromelysin-1 gene. J. Biol. Chem. *271*, 18231–18236.
4. Ye, S., Eriksson, P., Hamsten, A., Kurkinen, M., Humphries, S.E., and Henney, A.M. (1996). Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression. J. Biol. Chem. *271*, 13055–13060.
5. Beyzade, S., Zhang, S., Wong, Y.K., Day, I.N., Eriksson, P., and Ye, S. (2003). Influences of matrix metalloproteinase-3 gene variation on extent of coronary atherosclerosis and risk of myocardial infarction. J. Am. Coll. Cardiol. *41*, 2130–2137.
6. Lichtinghagen, R., Bahr, M.J., Wehmeier, M., Michels, D., Haberkorn, C.I., Arndt, B., Flemming, P., Manns, M.P., and Boeker, K.H. (2003). Expression and coordinated regulation of matrix metalloproteinases in chronic hepatitis C and hepatitis C virus-induced liver cirrhosis. Clin. Sci. (Lond.) *105*, 373–382.
7. Medley, T.L., Kingwell, B.A., Gatzka, C.D., Pillay, P., and Cole, T.J. (2003). Matrix metalloproteinase-3 genotype contributes to age-related aortic stiffening through modulation of gene and protein expression. Circ. Res. *92*, 1254–1261.
8. Gnasso, A., Motti, C., Irace, C., Carallo, C., Liberatoscioli, L., Bernardini, S., Massoud, R., Mattioli, P.L., Federici, G., and Cortese, C. (2000). Genetic variation in human stromelysin gene promoter and common carotid geometry in healthy male subjects. Arterioscler. Thromb. Vasc. Biol. *20*, 1600–1605.
9. Rauramaa, R., Vaisanen, S.B., Luong, L.A., Schmidt-Trucksass, A., Penttila, I.M., Bouchard, C., Toyry, J., and Humphries, S.E. (2000). Stromelysin-1 and interleukin-6 gene promoter polymorphisms are determinants of asymptomatic carotid artery atherosclerosis. Arterioscler. Thromb. Vasc. Biol. *20*, 2657–2662.
10. Rundek, T., Elkind, M.S., Pittman, J., Boden-Albala, B., Martin, S., Humphries, S.E., Juo, S.H., and Sacco, R.L. (2002). Carotid intima-media thickness is associated with allelic variants of stromelysin-1, interleukin-6, and hepatic lipase genes: the Northern Manhattan Prospective Cohort Study. Stroke *33*, 1420–1423.
11. Humphries, S.E., Luong, L.A., Talmud, P.J., Frick, M.H., Kesaniemi, Y.A., Pasternack, A., Taskinen, M.R., and Syvanne, M. (1998). The 5A/6A polymorphism in the promoter of the stromelysin-1 (MMP-3) gene predicts progression of angiographically determined coronary artery disease in men in the LOCAT gemfibrozil study. Lopid Coronary Angiography Trial. Atherosclerosis *139*, 49–56.
12. Ye, S., Watts, G.F., Mandalia, S., Humphries, S.E., and Henney, A.M. (1995). Preliminary report: genetic variation in the human stromelysin promoter is associated with progression of coronary atherosclerosis. Br. Heart J. *73*, 209–215.
13. de Maat, M.P., Jukema, J.W., Ye, S., Zwinderman, A.H., Moghaddam, P.H., Beekman, M., Kastelein, J.J., van Boven, A.J., Bruschke, A.V., Humphries, S.E., et al. (1999). Effect of the stromelysin-1 promoter on efficacy of pravastatin in coronary atherosclerosis and restenosis. Am. J. Cardiol. *83*, 852–856.
14. Schwarz, A., Haberbosch, W., Tillmanns, H., and Gardemann, A. (2002). The stromelysin-1 5A/6A promoter polymorphism is a disease marker for the extent of coronary heart disease. Dis. Markers *18*, 121–128.
15. Humphries, S., Bauters, C., Meirhaeghe, A., Luong, L., Bertrand, M., and Amouyel, P. (2002). The 5A6A polymorphism in the promoter of the stromelysin-1 (MMP3) gene as a risk factor for restenosis. Eur. Heart J. *23*, 721–725.
16. Humphries, S.E., Martin, S., Cooper, J., and Miller, G. (2002). Interaction between smoking and the stromelysin-1 (MMP3)

gene 5A/6A promoter polymorphism and risk of coronary heart disease in healthy men. Ann. Hum. Genet. *66*, 343–352.
17. Ghilardi, G., Biondi, M.L., DeMonti, M., Turri, O., Guagnellini, E., and Scorza, R. (2002). Matrix metalloproteinase-1 and matrix metalloproteinase-3 gene promoter polymorphisms are associated with carotid artery stenosis. Stroke *33*, 2408–2412.
18. Hirashiki, A., Yamada, Y., Murase, Y., Suzuki, Y., Kataoka, H., Morimoto, Y., Tajika, T., Murohara, T., and Yokota, M. (2003). Association of gene polymorphisms with coronary artery disease in low- or high-risk subjects defined by conventional risk factors. J. Am. Coll. Cardiol. *42*, 1429–1437.
19. Henney, A.M., Wakeley, P.R., Davies, M.J., Foster, K., Hembry, R., Murphy, G., and Humphries, S. (1991). Localization of stromelysin gene expression in atherosclerotic plaques by in situ hybridization. Proc. Natl. Acad. Sci. USA *88*, 8154–8158.
20. Liu, P.Y., Chen, J.H., Li, Y.H., Wu, H.L., and Shi, G.Y. (2003). Synergistic effect of stromelysin-1 (matrix metallo-proteinase-3) promoter 5A/6A polymorphism with smoking on the onset of young acute myocardial infarction. Thromb. Haemost. *90*, 132–139.
21. Lamblin, N., Bauters, C., Hermant, X., Lablanche, J.M., Helbecque, N., and Amouyel, P. (2002). Polymorphisms in the promoter regions of MMP-2, MMP-3, MMP-9 and MMP-12 genes as determinants of aneurysmal coronary artery disease. J. Am. Coll. Cardiol. *40*, 43–48.
22. Terashima, M., Akita, H., Kanazawa, K., Inoue, N., Yamada, S., Ito, K., Matsuda, Y., Takai, E., Iwai, C., Kurogane, H., et al. (1999). Stromelysin promoter 5A/6A polymorphism is associated with acute myocardial infarction. Circulation *99*, 2717–2719.
23. Silence, J., Lupu, F., Collen, D., and Lijnen, H.R. (2001). Persistence of atherosclerotic plaque but reduced aneurysm formation in mice with stromelysin-1 (MMP-3) gene inactivation. Arterioscler. Thromb. Vasc. Biol. *21*, 1440–1445.
24. Moran, A., Iniesta, P., de Juan, C., Gonzalez-Quevedo, R., Sanchez-Pernaute, A., Diaz-Rubio, E., Ramon y Cajal, S., Torres, A., Balibrea, J.L., and Benito, M. (2002). Stromelysin-1 promoter mutations impair gelatinase B activation in high microsatellite instability sporadic colorectal tumors. Cancer Res. *62*, 3855–3860.
25. Karaman, M.W., Houck, M.L., Chemnick, L.G., Nagpal, S., Chawannakul, D., Sudano, D., Pike, B.L., Ho, V.V., Ryder, O.A., and Hacia, J.G. (2003). Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. Genome Res. *13*, 1619–1630.
26. Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics *74*, 175–195.
27. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
28. Rockman, M.V., Hahn, M.W., Soranzo, N., Goldstein, D.B., and Wray, G.A. (2003). Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. *13*, 2118–2123.
29. Lei, H., Zaloudik, J., and Vorechovsky, I. (2002). Lack of association of the -1171 (5A) allele of the MMP3 promoter with breast cancer. Clin. Chem. *48*, 798–799.
30. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. *74*, 106–120.
31. Maynard Smith, J., and Haigh, J. (1974). The hitchhiking effect of a favorable gene. Genet. Res. *23*, 23–35.
32. Depaulis, F., and Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. *15*, 1788–1790.
33. Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics *147*, 915–925.
34. Hudson, R.R., Bailey, K., Skarecky, D., Kwiatowski, J., and Ayala, F.J. (1994). Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics *136*, 1329–1340.
35. Zhou, G., Zhai, Y., Dong, X., Zhang, X., He, F., Zhou, K., Zhu, Y., Wei, H., Yao, Z., Zhong, S., et al. (2004). Haplotype structure and evidence for positive selection at the human IL13 locus. Mol. Biol. Evol. *21*, 29–35.
36. Stumpf, M.P., and Goldstein, D.B. (2001). Genealogical and evolutionary inference with the human Y chromosome. Science *291*, 1738–1742.
37. Thomas, M.G., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N., and Goldstein, D.B. (1998). Origins of Old Testament priests. Nature *394*, 138–140.
38. Cooper, D.N. (1999). Human Gene Evolution (Oxford: Bios Scientific).
39. Rockman, M.V., and Wray, G.A. (2002). Abundant raw material for *cis*-regulatory evolution in humans. Mol. Biol. Evol. *19*, 1991–2004.
40. Kashi, Y., King, D., and Soller, M. (1997). Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. *13*, 74–78.
41. Colosimo, P.F., Peichel, C.L., Nereng, K., Blackman, B.K., Shapiro, M.D., Schluter, D., and Kingsley, D.M. (2004). The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PloS Biol. 2(5): e109 DOI:10.1371/journal.pbio.0020109.
42. Hardison, R.C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. *16*, 369–372.
43. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. (2003). The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. *20*, 1377–1419.
44. Peterson, S., Peto, V., and Rayner, M. (2003). Coronary Heart Disease Statistics (Oxford: British Heart Foundation).
45. Gould, S.J., and Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc. R. Soc. Lond. B. Biol. Sci. *205*, 581–598.
46. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.D., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. *74*, 1111–1120.
47. Hamblin, M.T., and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. *66*, 1669–1679.
48. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.
49. Bamshad, M.J., Mummidi, S., Gonzalez, E., Ahuja, S.S., Dunn, D.M., Watkins, W.S., Wooding, S., Stone, A.C., Jorde, L.B., Weiss, R.B., et al. (2002). A strong signature of balancing selection in the 5′ *cis*-regulatory region of CCR5. Proc. Natl. Acad. Sci. USA *99*, 10539–10544.
50. Nakajima, T., Wooding, S., Sakagami, T., Emi, M., Tokunaga, K., Tamiya, G., Ishigami, T., Umemura, S., Munkhbat, B., Jin, F., et al. (2004). Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am. J. Hum. Genet. *74*, 898–916.
51. Payseur, B.A., Cutter, A.D., and Nachman, M.W. (2002). Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. *19*, 1143–1153.
52. Satsangi, J., Chapman, R.W., Haldar, N., Donaldson, P., Mitchell, S., Simmons, J., Norris, S., Marshall, S.E., Bell, J.I., Jewell, D.P., et al. (2001). A functional polymorphism of the stromelysin gene (MMP-3) influences susceptibility to primary sclerosing cholangitis. Gastroenterology *121*, 124–130.
53. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
54. Rozas, J., and Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics *15*, 174–175.

55. Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.
56. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241–247.
57. Yu, N., Jensen-Seaman, M.I., Chemnick, L., Kidd, J.R., Deinard, A.S., Ryder, O., Kidd, K.K., and Li, W.H. (2003). Low nucleotide diversity in chimpanzees and bonobos. Genetics *164*, 1511–1518.
58. Sanz, L., Berra, E., Municio, M.M., Dominguez, I., Lozano, J., Johansen, T., Moscat, J., and Diaz-Meco, M.T. (1994). Zeta PKC plays a critical role during stromelysin promoter activation by platelet-derived growth factor through a novel palindromic element. J. Biol. Chem. *269*, 10044–10049.
59. Borghaei, R.C., Sullivan, C., and Mochan, E. (1999). Identification of a cytokine-induced repressor of interleukin-1 stimulated expression of stromelysin 1 (MMP-3). J. Biol. Chem. *274*, 2126–2131.

Accession Numbers

Aligned haplotypes were submitted to GenBank with accession numbers AY541459–541475. Allele frequency data were submitted to dbSNP under the reference number for the 5T/6T SNP, rs3025058.