Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes.


To cite this version:

HAL Id: pasteur-00169326
https://hal-pasteur.archives-ouvertes.fr/pasteur-00169326
Submitted on 10 Sep 2007
Deciphering the Ancient and Complex Evolutionary History of Human Arylamine N-Acetyltransferase Genes

Etienne Patin,1,5 Luis B. Barreiro,1 Pardis C. Sabeti,6 Frédéric Austerlitz,7 Francesca Luca,8 Antti Sajantila,9 Doron M. Behar,10 Ornella Semino,11 Anavaj Sakuntabhai,2 Nicole Guiso,1 Brigitte Gicquel,3 Ken McElreavey,4 Rosalind M. Harding,12 Evelyne Heyer,5 and Luís Quintana-Murci1

1Centre National de la Recherche Scientifique (CNRS) FRE 2849, Unit of Molecular Prevention and Therapy of Human Diseases, 2Unité des Maladies Infectieuses et Autoimmunes, 3Unité de Génétique Mycobactérienne, and 4Unit of Reproduction, Fertility and Populations, Institut Pasteur, and 5CNRS UMR 5145, Musée de l’Homme, Paris; 6Whitehead Institute/MIT Center for Genome Research, Cambridge, MA; 7University of Paris-Sud, Orsay, France; 8University of Calabria, Rende, Italy; 9University of Helsinki, Helsinki, Finland; 10Bruce Rappaport Faculty of Medicine and Research Institute, Technion and Rambam Medical Center, Haifa, Israel; 11University of Pavia, Pavia, Italy; and 12University of Oxford, Oxford, United Kingdom

The human N-acetyltransferase genes NAT1 and NAT2 encode two phase-II enzymes that metabolize various drugs and carcinogens. Functional variability at these genes has been associated with adverse drug reactions and cancer susceptibility. Mutations in NAT2 leading to the so-called slow-acetylation phenotype reach high frequencies worldwide, which questions the significance of altered acetylation in human adaptation. To investigate the role of population history and natural selection in shaping NATs variation, we characterized genetic diversity through the resequencing and genotyping of NAT1, NAT2, and the pseudogene NATP in a collection of 13 different populations with distinct ethnic backgrounds and demographic pasts. This combined study design allowed us to define a detailed map of linkage disequilibrium of the NATs region as well as to perform a number of sequence-based neutrality tests and the long-range haplotype (LRH) test. Our data revealed distinctive patterns of variability for the two genes: the reduced diversity observed at NAT1 is consistent with the action of purifying selection, whereas NAT2 functional variation contributes to high levels of diversity. In addition, the LRH test identified a particular NAT2 haplotype (NAT2*5B) under recent positive selection in western/central Eurasians. This haplotype harbors the mutation 341T→C and encodes the “slowest-acetylator” NAT2 enzyme, suggesting a general selective advantage for the slow-acetylator phenotype. Interestingly, the NAT2*5B haplotype, which seems to have conferred a selective advantage during the past ~6,500 years, exhibits today the strongest association with susceptibility to bladder cancer and adverse drug reactions. On the whole, the patterns observed for NAT2 well illustrate how geographically and temporally fluctuating xenobiotic environments may have influenced not only our genome variability but also our present-day susceptibility to disease.

The two human N-acetyltransferase genes, NAT1 (MIM 108345) and NAT2 (MIM 243400), represent one of the first and clearest examples of the importance of genetic variation among individuals and across populations in drug response (Weber 1987). The two homologous genes are situated within a 200-kb region in 8p22, together with the NATP pseudogene (fig. 1). Both genes encode phase-II enzymes named “arylamine N-acetyltransferases” (NATs), which catalyze the transfer of an acetyl group to different arylhydrazines and arylamine drugs (Blum et al. 1990). Both genes carry functional polymorphisms whose effects on enzymatic activity have been well studied (Hein et al. 2000). Whereas the variants associated with reduced activity attain only low frequencies in NAT1, they constitute common polymorphisms in NAT2 (Upton et al. 2001). Two main classes of NAT2 phenotypes are therefore observed: the “fast-acetylation” phenotype, which refers to the wild-type acetylation activity, and the “slow-acetylation” phenotype, which results in reduced protein activity. In addition, NAT1 and NAT2 metabolize numerous common carcinogens, and variation in these genes can result in varying susceptibility to cancer (for a review, see the work of Hein [2002]). For example, the slow-acetylator NAT2 phenotype has been associated with side effects to the commonly used antitubercular isoniazid (Huang et al. 2002) and with higher risk for bladder cancer (Cartwright et al. 1982; Garcia-Closas et al. 2005). Nevertheless, most NAT2 mutations leading to the slow phenotype are found at high frequencies worldwide, calling into question the role of altered acetylation in human adaptation. Moreover, the function of NATs in medi-
Subjects and Methods

**DNA Samples**

The resequencing panel consisted of 80 individuals (160 chromosomes) from eight populations representing major geographic regions; sub-Saharan African chromosomes were represented by Bakola Pygmies from Cameroon (20) and by Bantu speakers from Gabon (20); western Eurasian samples were represented by Ashkenazi Jews (20), Sardinians (12), French (20), and Saami from Finland (20); and eastern Eurasian samples were represented by Indians from Gujarat (20) and by Thai (28). One chimpanzee (*Pan troglodytes*) was also sequenced to define the ancestral state of each mutation. The genotyping panel consisted of 563 individuals (1,126 chromosomes) from 13 populations. Sub-Saharan African chromosomes were represented by Bakola Pygmies (80) and Baka Pygmies (60) from Cameroon, Ateke Bantu speakers from Gabon (100), and Somali (48); North African and western and central Eurasian samples were represented by Morrocan (88), Ashkenazi Jews (80), Sardinians (98), Swedes (100), Saami from Finland (96), and Turkmen from Uzbekistan (100); and eastern Eurasian samples were represented by Gujarati from India (100), Chinese from the Hunan and Zhejiang regions (88), and Thai (88). All individuals were healthy donors from whom informed consent was obtained.

**PCR and Sequence Determination**

Six different regions were PCR amplified, for a total of ~8.5 kb per chromosome (fig. 1): the entire coding exon of the *NAT1* gene (870 bp) and 1,735 bp of noncoding flanking parts (1,122 bp in 5' end and 613 bp in 3' end); the entire coding exon of the *NAT2* gene (870 bp) and 1,950 bp of noncoding flanking parts, including 1,603 bp surrounding its first non-coding exon; the pseudogene *NATP* (2,145 bp); and two intergenic noncoding regions at 10 kb and 100 kb (1,068 bp) from *NAT1* 5' end. Details about PCR and sequencing conditions are available on request. As a measure of quality control for the data, individuals presenting singletons or ambiguous polymorphisms were reamplified and resequenced. Sequences were analyzed using the GENALYS software (Takahashi et al. 2003).

**Selection and Genotyping of Polymorphisms**

The newly discovered sequence-based variation was used to determine the minimal number of SNPs able to distinguish the haplotypic diversity (haplotype-tagging SNPs [htSNPs]) of *NAT1* and *NAT2* loci in a given population. Five SNPs and one (TAA) microsatellite were typed in *NAT1* by either genotyping or sequencing, and seven SNPs were genotyped in the *NAT2* coding region. In addition, we genotyped nine intergenic SNPs selected because they were polymorphic in all human populations (fig. 1). These SNPs were chosen either from dbSNP (when dbSNPs met the previous criterion) or from the intergenic regions sequenced here. Genotyping was performed by either fluorescence polarization (VICTOR-2TM technology) or TaqMan (ABI Prism-7000 Sequence Detection System) assays.

**Sequence-Based Data Analysis**

Allele frequencies were determined by gene counting, and deviations from Hardy-Weinberg equilibrium were tested by Arlequin v.2.001 (Schneider et al. 2000). Haplotype reconstruction was performed using the Bayesian method implemented in PHASE v.2.1.1 (Stephens and Donnelly 2003), and htSNPs were defined using BEST v.1.0 (Sebastiani et al. 2003), after the exclusion of singletons because they could not be positioned with certainty on a given haplotypic context. With
the use of phased data, the neutral parameter $\theta_{ML}$ and the time since the most recent common ancestor ($T_{MRCA}$) were estimated by maximum likelihood with GENETREE (Griffiths and Tavaré 1994), under a standard coalescent model. Since this model assumes no recombination, for this particular analysis we had to exclude a few SNPs or rare recombinant haplotypes (in NAT1, the first four 5′ SNPs; in NATP, three singleton haplotypes; in NAT2, two singleton haplotypes). Time, scaled in $2N_e$ units, was converted into years by use of a 25-year generation time and an $N_e$ value obtained as $\theta_{ML}$ divided by $4N_e$. The mutation rate per gene per generation ($\mu$) was deduced from $D_{xy}$, the average number of nucleotide substitutions per site between human and chimpanzee (Nei 1987, equation 10.20), calculated by DnaSP v.4.0 (Rozas et al. 2003), with consideration that the two species diverged 200,000 generations ago. Simulations were performed to estimate the probability of a $T_{MRCA}$ greater than a given value, under a Wright-Fisher model. Fifty thousand simulations were performed using a version of the MS program modified to obtain $T_{MRCA}$ values (R. Hudson, personal communication).

Using DnaSP, we calculated the nucleotide diversity ($\pi$) and Watterson’s estimator of $\theta$ ($\theta_W$) (Watterson 1975), and we performed a number of statistical tests: Tajima’s $D$ (TD) (Tajima 1989), Fu and Li’s $F^*$ (Fu and Li 1993), Fay and Wu’s $H$ (Fay and Wu 2000), $K_s/K_t$ (Kimura 1968), the Hudson-Kreitman-Aguadé (HKA) test (Hudson et al. 1987), and the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991). A neutrality test based on the expected heterozygosity was also performed with the Bottleneck program (Cornuet and Luikart 1996) on the NAT1 3′ UTR microsatellite, by the use of coalescent simulations (10,000 runs) and with the assumption of different mutational models (stepwise mutation model and two-phased mutation model with 0%–40% of mutistep changes).

Genotyping-Based Data Analysis

Pairwise LD between the 21 genotyped SNPs was estimated after the exclusion, in each population, of SNPs with a minor-allele frequency (MAF) <0.10. Using DnaSP, we calculated the statistics $D^*$ (Lewontin 1964) and $r^2$ (Hill and Robertson 1968) and tested their statistical significance, using a Fisher’s exact test followed by Bonferroni corrections. To perform the LRH test, we selected two core regions (in NAT1, SNPs 445, 1088, 1095, and 1191; in NAT2, SNPs 341, 481, 590, 803, and 857) identified as haplotype blocks, following the criteria of Gabriel et al. (2002), and we assessed, for each core haplotype, its relative extended haplotype homozygosity (REHH) 200 kb apart. To test the significance of potentially selected core haplotypes, we first compared our sub-Saharan African and non-African data sets with coalescent simulations of 1-Mb regions, assuming a neutral model of evolution with recombination (Hudson 2002). Model parameters (including demography and recombination rate) were consistent with current estimates for African and non-African populations (Schaffner et al. 2005). Similarly, our sub-Saharan African and non-African data sets were compared with the empirical distribution of “core haplotype frequencies versus REHH” obtained from the screening of the entire chromosome 8 in Yoruban and European-descent populations, respectively (HapMap database).

To infer the population growth rate, $r$, and the age, $g$, of NAT2 nonsynonymous mutations, we used a joint maximum-likelihood estimation of these parameters, as described in Autertitz et al. (2003). We compared these results with coalescent-based estimations of the two parameters: the growth rate estimation of Slatkin and Bertorelle (2001) and the Reeve and Rannala (2002) age estimation using the DMLE+ v.2.2 software. One million iterations were performed for each estimation. The recombination parameter required for these analyses was estimated by comparing deCODE and Marshfield genetic and physical distances in the NATs region (UCSC Genome Bioinformatics). The coefficient of selection, $s$, of the NAT2 mutation 341T→C was estimated using the deterministic equation 3.29 of Wright (1969), which relates the frequency of an allele in generation $t + 1$ to its frequency in generation $t$. We stated the degree of dominance, $h$, to 0.0 (recessivity) and 0.5 (codominance). We assumed the frequency, $p_0$, of the C allele before selection to vary between 0.05 and 0.15 (corresponding to the allele frequency in Pygmies and eastern Eurasians). Making these assumptions, we calculated the $s$ values that would yield a frequency of 0.50 (the present-day frequency of the 341C allele in western Eurasians) from its initial $p_0$ frequency in $g$ generations.

Results

NATs Nucleotide Sequence Variation

The initial sequencing screening of the resequencing panel yielded a total of 111 mutations, including 68 transitions, 34 transversions, 8 insertions/deletions, and 1 triallelic microsatellite (table 1) (GenBank accession numbers DQ305496–DQ305975). In NAT1, we observed 2 nonsynonymous and 4 synonymous SNPs in its coding region and 26 SNPs and the triallelic (TAA) microsatellite in its flanking regions. In the NAT2 coding region, we found two synonymous and eight nonsynonymous mutations, three of which were newly identified (L241, T193M, and Y208H). These three variants were singletons and were restricted to sub-Saharan samples. In addition, 14 SNPs and 3 indels in NAT2 flanking regions were observed. In NATP, we identified 32 SNPs and 5 indels. For all the SNPs, only 1.54% of the tests departed significantly from Hardy-Weinberg equilibrium. However, these few tests would become nonsignificant after a correction for multiple testing.

The nucleotide diversity, $\pi$, in the NAT1 and NAT2 coding and flanking regions as well as in the pseudogene NATP is reported in table 2. Interestingly, the two duplicated NAT genes, which share 87.3% of nucleotide
Haplotype Diversity and TMRCA and associated with the fast-acetylator phenotype (cestral haplotype with all other populations (table 2). In the French, Gujarati, and Thai samples, as compared restricted to Eurasian populations. As for gent and low-frequency results confirmed the sequence data, in that the diver-
85%–100% of NAT1 haplotype frequencies are reported in tables 5 and 6. Two NAT1 haplotypes, NAT1*10 and NAT1*4, account for 85%–100% of NAT1 diversity. In addition, genotyping results confirmed the sequence data, in that the diver-
gent and low-frequency NAT1*11A haplotype is restricted to Eurasian populations. For as NAT2, the an-
cestral haplotype NAT2*4 and the remaining haplotypes associated with the fast-acetylator phenotype (NAT2*12 and NAT2*13) were most frequent in Bakola Pygmies and eastern Eurasians. Among the derived haplotypes associated with the slow-acetylator phenotype, NAT2*14 (191G→A) was African specific, NAT2*7 (857G→A) was observed mainly in eastern Eurasians, NAT2*5 (341T→C) was common in western and central Eurasians as well as in sub-Saharan Africans (Pygmies excepted), and NAT2*6 (590G→A) was found ubiqui-
tously at intermediate frequencies.

To investigate the tree topology and time depth of the three NAT loci, we next estimated the gene tree and the TMRCA of NAT1, NAT2, and NATP. The two divergent NAT1 lineages coalesced 2.01 ± 0.29 million years ago (MYA) (fig. 2), one of the highest estimated TMRCA values in the human genome (Excoffier 2002). By contrast, the TMRCA values of NAT2 (1.01 ± 0.27 MYA) and NATP (1.05 ± 0.24 MYA) were in agreement with neutral expectations, since most human neutral loci should co-
alesce ∼4N, generations ago (i.e., ∼1 MYA) (Takahata 1993).

Population Variation in LD Patterns

To determine global LD patterns in the 13 populations of our genotyping panel, we estimated D' and r2 for the NATs region (data not shown). Both NAT1 and NAT2 genes showed significant and strong intragenic LD levels: the proportion of SNP pairs in significant LD equaled 87.5% and 89.6% in Africans and non-Africans, re-
spectively, at the NAT1 locus and 73.7% and 84.0% at NAT2. The genomic structure of the entire 200-kb NATs

Table 2
Diversity Indices of NAT1, NAT2, and NATP Sequences

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>NAT1 CODING REGION (870 bp)</th>
<th>NAT2 CODING REGION (2,605 bp)</th>
<th>NATP PSEUDOGENE (2,145 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>θw (%)</td>
<td>π (%)</td>
<td>HD (%)</td>
</tr>
<tr>
<td>Bakola</td>
<td>0.065</td>
<td>0.023</td>
<td>0.100</td>
</tr>
<tr>
<td>Bantu</td>
<td>0.000</td>
<td>0.000</td>
<td>0.087</td>
</tr>
<tr>
<td>Ashkenazi</td>
<td>0.000</td>
<td>0.000</td>
<td>0.043</td>
</tr>
<tr>
<td>Sardinian</td>
<td>0.000</td>
<td>0.000</td>
<td>0.076</td>
</tr>
<tr>
<td>French</td>
<td>0.997</td>
<td>0.034</td>
<td>0.238</td>
</tr>
<tr>
<td>Saami</td>
<td>0.000</td>
<td>0.000</td>
<td>0.043</td>
</tr>
<tr>
<td>Gujarati</td>
<td>0.997</td>
<td>0.034</td>
<td>0.227</td>
</tr>
<tr>
<td>Thai</td>
<td>0.012</td>
<td>0.056</td>
<td>0.204</td>
</tr>
<tr>
<td>Total</td>
<td>0.012</td>
<td>0.021</td>
<td>0.073</td>
</tr>
</tbody>
</table>

NOTE.—HD = haplotype diversity.

Haplotype Diversity and TMRCA Estimates of NAT Loci

Haplotypes of the NAT genes were reconstructed first by use of the sequence data obtained from the resequ-
encing panel (tables 3 and 4). The most unusual observation was the haplotype diversity of the NAT1 locus, made of two highly divergent haplotype clusters separated by 17 mutations. One cluster contained most of NAT1 haplotype diversity (97.5%), whereas the other contained a unique haplotype, called “NAT1*11A,” that was observed in just three non-African individuals (one heterozygous French, one heterozygous Gujarati, and one homozygous Thai). Consequently, diversity esti-
mates of the NAT1 locus were inflated in these three populations, as attested by the much higher θw values in the French, Gujarati, and Thai samples, as compared with all other populations (table 2).

As for the genotyping panel, NAT1 and NAT2 haplo-
types frequencies are reported in tables 5 and 6. Two NAT1 haplotypes, NAT1*10 and NAT1*4, account for 85%–100% of NAT1 diversity. In addition, genotyping results confirmed the sequence data, in that the diver-
gent and low-frequency NAT1*11A haplotype is re-
stricted to Eurasian populations. As for NAT2, the an-
cestral haplotype NAT2*4 and the remaining haplotypes associated with the fast-acetylator phenotype (NAT2*12 and NAT2*13) were most frequent in Bakola Pygmies and east.
region was made of two independent haplotype blocks, one corresponding to NAT1 and the other to NATP and NAT2. Further, we observed strong population variation in LD levels when plotting the proportion of SNP pairs in significant LD against physical distance for each population separately (fig. 3). Sub-Saharan Africans showed lower LD levels than non-Africans, with the clear exception of Bakola Pygmies. Both western and eastern Eurasians exhibited similar LD patterns, excluding the Saami, who were, by far, the population with the highest degree of allelic association.

Tests of the Standard Neutral Model

The absence of LD between the two NAT genes in all populations enabled us to study the evolutionary forces independently shaping NAT1 and NAT2 diversity. For the NAT1 coding region, tests were not feasible in four of eight populations because of a complete absence of exonic variation (table 2). In the remaining populations, most Tajima’s, Fu and Li’s, and Fay and Wu’s tests gave significant negative values (table 7). When these analyses were extended to NAT1 flanking regions, the same tests lost significance in Bakola Pygmies, whereas they turned out to be even more significant in French, Gujarati, and Thai because of an excess of singletons when mutations are not orientated for their ancestral state (TD and F*) or because of an excess of highly frequent derived variants when their ancestral state is considered (H). These results are mainly due to the presence of the low-frequency and highly divergent NAT1*11A haplotype in the three Eurasian populations (fig. 2). As for the NAT2 coding region, both Tajima’s D and Fu and Li’s D* in the Ashkenazi, Sardinians, and French and Fu and Li’s F* in the Saami were significantly positive (table 7). However, all tests were not significant anymore when both flanking and exonic NAT2 variation was considered. As for NATP, although it was found to be in LD with NAT2, all tests yielded nonsignificant values.

To test significant differences in diversity levels among the three NAT loci, we performed the HKA test. The comparison between NAT1 and NATP was significant only among western and eastern Eurasians (P = .046 and .043, respectively), resulting from an excess of polymorphisms in NAT1, compared with fixed mutations. Again, these results are the consequence of the binary haplotype pattern observed at this locus (fig. 2). By contrast, the HKA tests comparing NAT1 versus NAT2 and NAT2 versus NATP yielded nonsignificant results. At the protein level, we compared the number of synonymous and nonsynonymous mutations in the two homologous genes. NAT1 exhibited a deficit in nonsynonymous mutations (Ks/Ka = 0.242). By contrast, NAT2 presented a Ks/Ka value closer to 1 (Ks/Ka = 0.802).

LRH Test for Recent Selection

We next performed the LRH test, which is designed to identify mutations/haplotypes under recent positive selection by comparing the frequency of a given allele with the breakdown of LD around it (Sabeti et al. 2002). Our results for both NAT1 and NAT2 are reported separately for African and Eurasian populations (fig. 4A and 4B). P values were estimated for all core haplotypes in all populations against both simulations and the empirical distributions of the HapMap in Yoruban and European-descent populations. A single haplotype of NAT2 (NAT2*5B) appeared to deviate from neutrality in western and central Eurasian populations: Psim = .0001 (Pemp = .0006) in Ashkenazi Jews, .0062 (.0085) in Saami, .0363 (.0063) in Turkmen, .0124 (.0067) in Moroccans, and .0464 (.1836) in Swedish, with the same haplotype in Sardinians being close to significance (.0576 [.2178]). Interestingly, the NAT2*5B haplotype, which exhibited the highest frequencies (>50%) in western Eurasians (table 6), bears the nonsynonymous mutation 341T→C (I114T), which has been shown to lead to a slow-acetylation status (Zang et al. 2004). In addition, a single NAT1 haplotype (NAT1*4) appeared to deviate from the simulated distribution in the same populations that showed signals of positive selection for NAT2*5B: Psim = .0008 (Pemp = .0319) in Ashkenazi Jews, .0115 (.1346) in Saami, .0445 (.2498) in Turkmen, .0293 (.2694) in Moroccans, and .0090 (.1987) in Swedish. In contrast to NAT2*5B, the protein encoded by the NAT1*4 haplotype does not differ in enzyme activity with those encoded by the other NAT1 haplotypes observed in this study (e.g., NAT1*10, NAT1*3, and NAT1*11) (Hughes et al. 1998; de Leon et al. 2000).

Growth Rate and Age of the NAT2 341T→C Mutation

The LD breakdown at the surrounding sites of a mutation is very informative for inferring allelic age estimates, through consideration of the recombination rate as a “genetic clock” (Labuda et al. 1997). The significant signal of selection detected for the NAT2 341T→C mutation, together with the functional consequences asso-

Table 4

<table>
<thead>
<tr>
<th>Allelic Composition and Frequency of NAT2 Haplotypes in Our Resequencing Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>The table is available in its entirety in the online edition of The American Journal of Human Genetics.</td>
</tr>
</tbody>
</table>

Table 5

<table>
<thead>
<tr>
<th>Allelic Composition and Frequency of NAT1 Haplotypes in Our Genotyping Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>The table is available in its entirety in the online edition of The American Journal of Human Genetics.</td>
</tr>
</tbody>
</table>
Table 6
Allelic Composition and Frequency of NAT2 Haplotypes in Our Genotyping Panel

The table is available in its entirety in the online edition of The American Journal of Human Genetics.

ciated with this variant (i.e., reduced acetylation activity), prompted us to estimate the age of this mutation by use of both maximum-likelihood and coalescent-based methods, both of which gave similar results. To provide a comparison, we performed the same analyses for the 590G→A mutation, which is located 250 bp from 341T→C and is never observed on the same haplotype. Our results indicated that these two mutations started to increase in frequency at similar times and with comparable growth rates in all populations (table 8). The only exception to this pattern was the 341T→C mutation in western/central Eurasians. Our estimations showed that this mutation started to increase in frequency 6,315 years ago (95% CI 5,797–7,005 years) at a growth rate (0.062) twice as big as the values observed for the same mutation in eastern Eurasians (0.031) and for 590G→A in the entire Eurasian sample. Indeed, the growth rate of 341T→C in western/central Eurasians was significantly different from all the others, since their 95% CIs did not overlap (table 8).

Acetylation Phenotype Inference

In view of the 95% NAT2 genotype-phenotype concordance (Cascorbi et al. 1995), we inferred the distribution of fast/slow-acetylation phenotypes across populations from our NAT2 genotyping panel and compared it with the phenotyping results of 23 healthy populations worldwide, reviewed for this occasion (fig. 5). To make both data sets comparable, we considered fast/slow heterozygotes to be fast acetylators because, even if they present a mean intermediate activity significantly different from that of fast homozygotes, they are mostly observed in the “fast acetylator activity peak” (Cascorbi et al. 1995). Phenotype frequencies showed strong variation among ethnic groups (fig. 5). The slow-acetylator phenotype is present at the lowest frequencies in eastern Asian and Native American populations as well as in the Pygmy groups studied here for the first time, whereas it exhibits the highest frequencies in Middle Eastern, European, central/south Eurasian, and African populations. The highest frequencies worldwide of the slow-acetylator phenotype are observed in the Ashkenazi population, in which it reaches 80%.

Discussion

The direct interaction of NAT1 and NAT2 gene products with the human chemical environment makes them potential targets of natural selection, provided that the exposure to xenobiotics has significantly influenced population fitness over time. Our population-based genetic study revealed distinctive patterns of variability for the two NAT genes, reflecting two very different evolutionary histories.

Reduced Variation and Deep-Rooting Genealogy in the NAT1 Gene

The NAT1 coding region is characterized by a reduced genetic diversity (π = 0.021%), with four populations of eight showing no variation at all (table 2). In addition, the nearby 3′ UTR (TAA) microsatellite of NAT1, which was also typed in the genotyping panel, presented low levels of heterozygosity (Hz = 0.073), with the allele (TAA)p accounting for 96.3% of the overall diversity (table 5). This deficit in heterozygosity was significant under both the stepwise mutation model and the two-phased mutation model (P < .05). Also, in strong contrast to NAT2, the functional mutations identified in NAT1 are present at very low frequencies (Upton et al. 2001). These observations, together with the Ks/Ka value of 0.242, are compatible with the action of purifying selection.
selection in shaping NAT1 diversity, in agreement with a previous study stating that the majority of human genes may be under weak negative selection (Bustamante et al. 2005). In this view, and considering that NAT1 is expressed in many tissues early in development and may play an additional role in the metabolism of folate (Sim et al. 2000), our genetic results suggest that its involvement in endogenous metabolic pathways might be more important than previously thought.

Purifying selection may not be the only evolutionary force that has influenced NAT1 diversity. Indeed, one of the most salient observations of this study is the highly divergent tree topology and high $T_{MRCA}$ (2.01 ± 0.29 MYA) of this locus (fig. 2). This binary pattern is translated into significant departures from neutrality in populations presenting the divergent haplotype NAT1*11A (see table 7 and results of the HKA test). The probability of finding such a high $T_{MRCA}$ under a Wright-Fisher model was found to be low ($P = .029$). Different hypotheses can be proposed to explain such long basal branches in the NAT1 gene tree. First, long-term balancing selection can result in divergent haplotype clusters, by maintaining two or more alleles over time, provided that they result in functional differences. Nevertheless, our data do not support this hypothesis, since the two nonsynonymous mutations separating the two clusters (fig. 2) have been shown to have no significant effects on the in vivo protein activity in human cells (Hein 2002) or on the stability and activity of the recombinant protein in yeast (Hughes et al. 1998). Any kind of selection due to a hitchhiking effect with neighboring genes is equally unlikely, because the two closest genes (ASAHI located 5′ and NAT2 located 3′) behave as independent haplotype blocks (this study and the HapMap database). Furthermore, our sequence data from the NAT1 coding region are consistent with the action of purifying selection rather than balancing selection, with the first selective regime having a minor influence on tree topologies (Williamson and Orive 2002). Second, gene conversion could also lead to such divergent haplotype patterns by the replacement of a segment of NAT1 with a tract from its nearby paralogs (NAT2 and/or NATP). This alternative is unlikely, however, since the 17 SNPs separating the two divergent NAT1 lineages are not physically clustered (fig. 2) as one would expect after gene conversion between duplicated loci (Innan 2003). Thus, if gene conversion formed the basis of such a haplotype pattern, multiple conversion events must be invoked, with some tracts of lengths <5 bp. Yet, the conversion-tract lengths have been estimated to range from 55 bp to 290 bp, through sperm-typing analyses (Jeffreys and May 2004).

In this view, an alternative and most likely scenario to explain our data is a demographic event such as ancient population structure. A number of studies have recently reported gene genealogies that present not only

Figure 3  Proportion of SNP pairs in significant LD against physical distance in the NATs 200-kb region. In each population, genotyped SNPs were selected to have a MAF >10%. The Fisher’s exact test was used to assess LD significance, followed by Bonferroni corrections. SNP pairs were grouped into 20-kb bins.
unexpectedly old coalescent times (~2 MYA) but also long basal branches (Harris and Hey 1999; Webster et al. 2003; Barreiro et al. 2005; Garrigan et al. 2005; Hayakawa et al. 2005). Our observations at NAT1, together with these studies, further support the view that some diversity in the genome of modern humans may have persisted from a structured ancestral population (Harding and McVean 2004). In addition, NAT1*11A appears to be absent in sub-Saharan Africa, since it was not detected in either our genotyping panel of 144 sub-Saharan Africans from distinct geographic locations or 600 African American individuals reported elsewhere (Upton et al. 2001). Therefore, the observation that the NAT1 gene tree is rooted in Eurasia questions the geographic location of such a structured ancestral population (Takahata et al. 2001). The origins of NAT1*11A could thus be placed either in sub-Saharan Africa, from where it must have subsequently disappeared, or in Eurasia. Should the latter be the case, the NAT1 gene tree is at odds with the commonly accepted replacement hypothesis (Lewin 1987) and is more parsimoniously explained by the occurrence of partial hybridization between modern humans expanding from Africa and pre-existing hominids in Eurasia, as recently sustained by the RRM2P4 locus (Garrigan et al. 2005). However, such inferences require further support from the analyses of multiple independent loci in increased numbers of samples and human populations.

The NAT2 Gene: An Advantage to Be a Slow Acetylator?

The significantly positive values observed for most sequence-based neutrality tests of NAT2 in western Eurasians (table 7) suggest the action of natural selection. Likewise, population size reductions—such as that probably experienced by non-African populations during the out-of-Africa exit (Marth et al. 2003) or, more recently, by Ashkenazi Jews (Behar et al. 2004)—could have also inflated TD and $F^*$ values (Przeworski et al. 2000). However, these demographic events should have equally influenced neutrality statistics for other non-African populations (i.e., eastern Eurasians), which is not the case. These observations argue thus in favor of the action of natural selection. Conversely, interspecific tests (i.e., $K_s/K_a$ and MK tests) do not rule out the neutral evolution of NAT2 diversity, which does not show any clear excess or lack of nonsynonymous mutations. In view of these apparently contradictory results, it is plausible that the frequencies, rather than the number of these mutations, have been influenced by selection, suggesting more subtle and recent fluctuations in the selective pressures operating on NAT2.

Such changes can be detected by the LRH test, which aims to identify haplotypes under recent positive selection. Moreover, this approach should be robust to the confounding effects of demography, since it corrects the extended haplotype homozygosity (EHH) of a given core haplotype by the EHH of all other haplotypes at the same core region (Sabeti et al. 2002). Overall, the LRH tests were more significant when NAT1 and NAT2 core haplotypes were compared with the simulated than with the empirical distribution. Manifestly, the empirical distribution also includes genes under selection that bias REHH toward higher values, and it allows the detection of selected haplotypes in a conservative way. Independently of which background distribution was used (e.g., simulated or empirical), the same NAT2 haplotype, NAT2*5B, was detected to depart from neutrality in western and central Eurasians (fig. 4B). These populations also showed significant $P$ values for a single NAT1 haplotype, NAT1*4. It is worth mentioning that ~80% of NAT2*5B haplotypes are associated with NAT1*4 in western/central Eurasians (vs. 43% and 60% in sub-Saharan Africans and eastern Eurasians, respectively).

### Table 7

Sequenced-Based Neutrality Tests in NAT1, NAT2, and NATP

<table>
<thead>
<tr>
<th>Population</th>
<th>NAT1 Coding Region*</th>
<th>NAT1 Coding and Flanking Regions (870 bp)</th>
<th>NAT2 Coding Region (870 bp)</th>
<th>NAT2 Coding and Flanking Regions (2,820 bp)</th>
<th>NATP Pseudogene (2,415 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TD</td>
<td>$F^*$</td>
<td>$H$</td>
<td>TD $F^*$ $H$</td>
<td>TD $F^*$ $H$</td>
</tr>
<tr>
<td>Bakola</td>
<td>$-1.513^a$</td>
<td>$-2.189^a$</td>
<td>0.19</td>
<td>0.007 $0.456$ $0.653$</td>
<td>0.090 $0.852$ $1.726$</td>
</tr>
<tr>
<td>Bantu</td>
<td>NA NA NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Ashkenazi</td>
<td>$-0.044$</td>
<td>$-1.311$</td>
<td>$-1.674$</td>
<td>2.516 $1.797$ $0.505$</td>
<td>0.374 $0.080$ $1.557$</td>
</tr>
<tr>
<td>Sardinian</td>
<td>NA NA NA</td>
<td>$0.723$</td>
<td>$-0.364$</td>
<td>3.143 $1.548$ $-0.091$</td>
<td>1.674 $2.516$ $0.505$</td>
</tr>
<tr>
<td>French</td>
<td>$-1.723$</td>
<td>$-2.535$</td>
<td>$-3.505$</td>
<td>2.048 $1.646$ $1.021$</td>
<td>1.017 $1.281$ $3.136$</td>
</tr>
<tr>
<td>Saami</td>
<td>NA NA NA</td>
<td>1.443</td>
<td>$-0.305$</td>
<td>1.355 $0.895$ $1.242$</td>
<td>0.113 $0.653$ $2.673$</td>
</tr>
<tr>
<td>Gujarati</td>
<td>$-1.723$</td>
<td>$-2.535$</td>
<td>$-3.505$</td>
<td>1.355 $0.895$ $1.242$</td>
<td>0.113 $0.653$ $2.673$</td>
</tr>
<tr>
<td>Thai</td>
<td>$-1.384$</td>
<td>$-0.410$</td>
<td>$-3.106^a$</td>
<td>0.818 $1.281$ $0.767$</td>
<td>0.441 $0.828$ $2.386$</td>
</tr>
</tbody>
</table>

* NA = not applicable (no variation observed).

$^b$ $0.05 < P < 0.01$

$P < 0.01$.  

Table 7.

430 The American Journal of Human Genetics Volume 78 March 2006 www.ajhg.org
Figure 4  REHH plotted against core haplotype frequencies. Circles represent NAT1 and NAT2 core haplotypes. The 95th and 99th percentiles were calculated from both simulated data (gray lines) and an empirical distribution obtained from the screening of the entire chromosome 8 (black lines). Circles above the 95th percentile of simulated and/or empirical distributions are blackened. A, NAT1 and NAT2 sub-Saharan African core haplotypes are plotted against both simulated data and the empirical distribution of ∼40,000 Yoruban core haplotypes from the HapMap. B, NAT1 and NAT2 core haplotypes in western/central and eastern Eurasian populations are plotted against both the simulated data and the empirical distribution of ∼40,000 European-descent core haplotypes from the HapMap. Numbers affiliated with significant core haplotypes refer to: (1) Moroccan NAT2*5B, (2) Swedish NAT2*5B, (3) Turkmen NAT1*4, (4) Moroccan NAT1*4, (5) Saami NAT1*4, and (6) Swedish NAT1*4.
Thus, the signals detected in both NAT1 and NAT2 may be the result of a single event of selection targeting a long-range haplotype composed of both NAT1*4 and NAT2*5B. Several lines of evidence, however, strongly support that the NAT1 haplotype does not harbor the functional cause of such a selective event: P values are globally less significant for NAT1*4 than for NAT2*5B, and, most importantly, all the NAT1 haplotypes observed in this study encode proteins with identical enzymatic activity (Hughes et al. 1998; de Leon et al. 2000). In sharp contrast, NAT2*5B bears the 341T →C mutation that is well known to encode an altered slow-acetylator protein (Zang et al. 2004).

Further support for the action of selection on the slow-encoding 341T →C NAT2 mutation comes from its growth-rate estimate, which showed that this mutation has increased in frequency twice as quickly as expected (r = 0.062) (table 8) only among western and central Eurasians. Previous estimations (Slatkin and Bertorelle 2001), together with our own, indicate that Eurasian populations have grown at a rate of ~0.030. Because population growth and selection have additive effects on the growth rate of a mutation (Slatkin and Rannala 2003), we estimated the 95% CI of this growth-rate estimate, which showed that this mutation has increased in frequency twice as quickly as expected (w = 0.032 – 0.030 = 0.032). Using a more accurate approach (Wright 1969), we estimated the 95% CI of this selective coefficient as 0.0124–0.0913. These figures reinforce the idea of a selective advantage of the 341C allele, even if weaker than other examples of recent but strong positive selection, such as lactase persistence (s ~0.09–0.19) (Bersaglieri et al. 2004) or the G6PD A– alleles (s > 0.1) (Saunders et al. 2005).

Altogether, both the LRH test and the growth rate of 341T →C argue in favor of positive selection acting on this slow-encoding mutation, which would imply a general selective advantage for the slow-acetylator phenotype. However, three other NAT2 variants (191G →A, 590G →A, and 857G →A) also encode slow proteins but do not show any significant departure from neutrality. Actually, the 341T →C mutation involves the greatest reduction in NAT2 enzymatic activity (Hein et al. 2000). In this context, it is very plausible that all NAT2 slow-acetylator variants have been subject to weak selective pressures, the signal of selection being detectable only in the mutation 341T →C causing the “slowest-acetylation” phenotype. Thus, the predicted increase in frequency of all slow mutations, in response to directional selection, would explain the observed excess of intermediate-frequency alleles in western Eurasian populations, as depicted by the significant positive values of TD (table 7).

The footprints of natural selection identified in western/central Eurasians raise the question of which event(s) may have provoked fluctuations in the spectrum of xenobiotics inactivated/activated by NAT2 (e.g., NAT2 activates heterocyclic carcinogens found in well-cooked meat [Hein et al. 2000; Hein 2002]) in these populations. In this context, given the geographic distribution of the slow-acetylator phenotype and the estimated expansion time of the slowest-encoding 341T →C mutation (5,797–7,005 years ago in western/central Eurasians), it is tempting to hypothesize that the emergence of agriculture in western Eurasia could be at the basis of such environmental changes. Indeed, there is accumulating evidence that this major transition resulted in a profound modification of human diets and lifestyles (Cordain et al. 2005) and, consequently, in the exposure of humans to chemical environments (Ferguson 2002). Moreover, the highest frequencies of slow acetylators are observed in the Middle East (fig. 5), one of the first regions where agriculture originated ~10,000 years ago, and these frequencies decrease toward western Europe, North Africa, and India, three regions where agriculture was subsequently diffused from the Fertile Crescent (Harris 1996). However, the hypothesis that the transition to agriculture influenced both the human exposure to xenobiotic environments and, consequently, the selective pressures

<table>
<thead>
<tr>
<th>Mutation and Parameter</th>
<th>Sub-Saharan Africans</th>
<th>Western/ Central Eurasians</th>
<th>Eastern Eurasians</th>
</tr>
</thead>
<tbody>
<tr>
<td>341T →C:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.270</td>
<td>.477</td>
<td>.181</td>
</tr>
<tr>
<td>r</td>
<td>.019 (.016–.024)</td>
<td>.062 (.055–.075)</td>
<td>.031 (.028–.037)</td>
</tr>
<tr>
<td>g</td>
<td>15,652 (13,797–18,435)</td>
<td>6,315 (5,797–7,005)</td>
<td>12,627 (11,427–14,685)</td>
</tr>
<tr>
<td>590G →A:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.185</td>
<td>.262</td>
<td>.362</td>
</tr>
<tr>
<td>r</td>
<td>.023 (.020–.030)</td>
<td>.031 (.028–.037)</td>
<td>.034 (.031–.041)</td>
</tr>
<tr>
<td>g</td>
<td>12,497 (10,932–14,958)</td>
<td>12,762 (11,657–14,385)</td>
<td>11,987 (10,940–13,643)</td>
</tr>
</tbody>
</table>

Note.—p = relative frequency of the mutation; r = growth rate; g = age of the mutation in years.

Table 8

Estimated Growth Rate and Age of the Mutations 341T →C and 590G →A (95% CI) of the NAT2 Gene
at NAT2 remains tentative and requires a better characterization of the naturally occurring substrates of the NAT2 enzyme.

Conclusion

The diversity patterns observed at the NATs region clearly illustrate our current vision of the human genome as a “mosaic of discrete segments,” each with its own individual evolutionary history (Paabo 2003). Whereas NAT1 could belong to a small proportion of nuclear loci that kept traces of an ancient population structure, the NAT2 gene gives some insights into the evolutionary processes that could make some present-day detrimental mutations frequent. The theory of the “thrifty genotype” proposes that common diseases, such as obesity and diabetes, are the result of a past advantage to efficiently metabolize rare food sources that are no longer restricted (Neel 1962). By analogy, the slow NAT2 acetylator haplotype (NAT2*5B), which is found at high frequencies worldwide and which we propose conferred some selective advantage, at least in western/central Eurasian populations, exhibits today the strongest association with susceptibility to bladder cancer and adverse drug reactions (Hein et al. 2000; Hein 2002). This “evolutionary conflict” could be widespread among genes involved in carcinogen metabolism, since two major events in human history—the Neolithic transition from foraging to agriculture and the more recent Industrial Revolution—may have dramatically changed our exposure to the damaging effects of environmental carcinogens. Consequently, dissecting the evolutionary processes that have shaped patterns of diversity of the genes involved in drug metabolism may represent a major analytical tool not only to identify those having played a crucial role in the past adaptation of Homo sapiens but also to better understand our present-day susceptibility to disease.

Acknowledgments

We warmly acknowledge M. Slatkin for supplying source codes for growth-rate estimations, R. R. Hudson for the modifications of the MS program, E. Sim for invaluable advice in the beginning of this project, and two anonymous reviewers for constructive criticisms on the early version of the manu-
script. We also thank A. Novellotto, A. Froment, J. M.Hombert, H. Roubia, and S. Santachiara-Benerecetti, for kindly providing us with DNA samples. This work was supported by CNRS and Institut Pasteur research funding. E.P. was supported by the French Ministry of Research Ph.D. program.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

Arlequin v.2.001, http://lgb.unige.ch/arlequin/
BEST v.1.0, http://genomethods.org/best/
DmLE v.2.2, http://www.dml.org/
DnaSP v.4.0, http://www.uab.es/dnasp/
GENALYS software, http://software.ceng.fr/
Genome-wide association study resource (GWAS), http://home.uchicago.edu/ 
GENETREE software, http://www.stats.ox.ac.uk/ 
Genome-wide association study resource (GWAS), http://home.uchicago.edu/ 

References

Hein DW (2002) Molecular genetics and function of NAT1 and NAT2: