9-3-2014

A homogenizing process of selection has maintained an 'ultra-slow' acetylation NAT2 variant in humans

Blandine Patillon
Université Paris Descartes, Université Paris Sud, France

Pierre Luisi
Institute of Evolutionary Biology, Catalonia, Spain

Estella S. Poloni
University of Geneva, Switzerland

Sotiria Boukouvala
Democritus University of Thrace, Alexandroupolis, Greece

Pierre Darlu
UMR7206, CNRS, Muséum National d'Histoire Naturelle, Université Paris, France

See next page for additional authors

Authors
Blandine Patillon, Pierre Luisi, Estella S. Poloni, Sotiria Boukouvala, Pierre Darlu, E. Genin, and Audrey Sabbagh

A homogenizing process of selection has maintained an 'ultra-slow' acetylation *NAT2* variant in humans

Running title: *NAT2**6 as a target of homogenizing selection

Patillon B[1,2,3], Luisi P[4], Poloni ES[5], Boukouvala S[6], Darlu P[7], Genin E[†,8] and Sabbagh A[*,†,1,2]

Keywords: *NAT2*; acetylation polymorphism; population differentiation; natural selection; linkage disequilibrium; rs1799930.

[1] IRD UMR216, Mère et enfant face aux infections tropicales, Paris, France

[2] PRES Sorbonne Paris Cité, Université Paris Descartes, Faculté de Pharmacie, Paris, France

[3] Université Paris Sud, Kremlin-Bicêtre, France

[4] Institute of Evolutionary Biology, CEXS-UPF-PRBB, Catalonia, Barcelona, Spain

[5] Laboratory of Anthropology, Genetics and Peopling History, Anthropology Unit, Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland

[6] Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis, Greece

[7] CNRS UMR7206, Muséum National d'Histoire Naturelle, Université Paris Diderot, Paris, France

[8] INSERM U1078, UBO EFS Bretagne, Brest, France

*Correspondence to: Audrey Sabbagh, UMR 216 IRD - Université Paris Descartes, Faculté de Pharmacie, 4 avenue de l'Observatoire, 75270 Paris Cedex 06, France. E-mail: audrey.sabbagh@ird.fr

†These authors contributed equally to the work and share the senior authorship.

Acknowledgements

This work was financially supported by the Institut Medicament-Toxicologie-Chimie. Environnement (IMTCE). BP is supported by a PhD fellowship from the doctoral program in Public Health from Paris Sud University. PL is supported by a PhD fellowship from 'Acción Estratrégica de Salud, en el Marco del Plan Nacional de Investigacio´n Científica, Desarrollo e Innovación Tecnológica 2008–2011' from Instituto de Salud Carlos III.

Abstract

N-acetyltransferase 2 (NAT2) is an important enzyme involved in the metabolism of a wide spectrum of naturally occurring xenobiotics, including therapeutic drugs and common environmental carcinogens. Extensive polymorphism in NAT2 gives rise to a wide interindividual variation in acetylation capacity which influences individual susceptibility to various drug-induced adverse reactions and cancers. Striking patterns of geographic differentiation have been described for the main slow acetylation variants of the *NAT2* gene, suggesting the action of natural selection at this locus. In the present study, we took advantage of the whole-genome sequence data available from the 1000 Genomes project to investigate the global patterns of population genetic differentiation at *NAT2* and determine whether they are atypical compared to the remaining variation of the genome. The non-synonymous substitution c.590G>A (rs1799930) defining the slow *NAT2*6* haplotype cluster exhibited an unusually low F_{ST} value when compared to the genome average (F_{ST} = 0.006, *P*-value = 0.016). It was pointed out as the most likely target of a homogenizing process of selection promoting the same allelic variant in globally distributed populations. The rs1799930 A allele has been associated with the slowest acetylation capacity *in vivo* and its substantial correlation with the subsistence strategy adopted by past human populations suggests that it may have conferred a selective advantage in populations shifting from foraging to agricultural and pastoral activities in the Neolithic period. Results of neutrality tests further supported an adaptive evolution of the *NAT2* gene through either balancing selection or directional selection acting on multiple standing slow-causing variants.

Introduction

The human acetylation polymorphism is one of the oldest and best-characterized pharmacogenetic traits that underlie interindividual and interethnic differences in response to xenobiotics. Acetylation catalised by NAT2 is a major biotransformation pathway for aromatic and heterocyclic amines present in the environment and diet, which can either be detoxified or bioactivated into metabolites that have the potential to cause toxicity and cancer (Butcher et al., 2002; Hein, 2002). From the clinical point of view, NAT2 acetylation is increasingly recognized as associated with significant health problems. Many clinically useful drugs are excreted by acetylation, some of them being crucial in the treatment of diseases representing a worldwide concern, such as tuberculosis, AIDS-related complex diseases, and hypertension. The individual acetylation status has proven to be an important determinant of both the effectiveness of prescribed medications and the development of adverse drug reactions and toxicity during drug treatment (Ladero, 2008; Meisel, 2002). Moreover, epidemiological studies have associated the acetylation phenotype with an increased susceptibility to various cancers following exposure to aromatic amine carcinogens (Agúndez, 2008; Hein, 2006; Sanderson et al., 2007; Selinski et al., 2013).

N-acetylation activity has been investigated in a wide range of populations, leading to a phenotype classification of humans in two main categories: fast acetylators, who exhibit the so-called ‘wild-type’ or normal acetylation activity, and slow acetylators, characterized by a decreased enzyme activity. The proportions of rapid and slow acetylators vary remarkably between populations of different ethnic and/or geographic origin (Sabbagh et al., 2011; Walker et al., 2009; Weber and Hein, 1985). Depending on the test substrate administered, a trimodal, rather than bimodal, distribution can be observed, revealing an additional, intermediate phenotype (Cascorbi et al., 1995; Kilbane et al., 1990; Parkin et al., 1997). Moreover, recent results suggest that the slow acetylator phenotype is not homogeneous and

that several slow acetylator phenotypes may rather exist, resulting from an allelic heterogeneity and differential functional effects of the slow acetylation alleles (Ruiz et al., 2012; Selinski et al., 2013). A refinement in phenotype inference, notably by the consideration of an 'ultra-slow' acetylator category, is advocated to help identifying new clinically relevant associations with one or more of these phenotype subcategories.

Acetylation polymorphism arises from allelic variations in the *NAT2* gene, which result in the production of arylamine *N*-acetyltransferase 2 (NAT2) proteins with variable enzyme activity or stability. The *NAT2* gene contains two exons with a relatively long intronic region of about 8.6 kb. Exon 1 is very short (100 bp) and the entire protein-coding region is contained within the 870-bp exon 2. Extensive polymorphism has been described in exon 2, with 38 nucleotide variations registered to date (http://nat.mbg.duth.gr/). Of these, four common non-synonymous substitutions at positions 191, 341, 590 and 857 are the most studied and characterize the major *NAT2* slow haplotype clusters (*NAT2*14*, *NAT2*5*, *NAT2*6* and *NAT2*7*, respectively). Individuals who are homozygous or compound heterozygous for two of these low-activity haplotypes are classified as slow acetylators.

NAT2 acetylation has attracted much research interest in evolutionary biology and several population genetic studies have attempted to clarify the role that slow acetylation could have played in the adaptation of our species (Fuselli et al., 2007; Luca et al., 2008; Magalon et al., 2008; Mortensen et al., 2011; Patin et al., 2006; Sabbagh and Darlu, 2006; Sabbagh et al., 2011). The high prevalence of slow acetylators in humans (well above 50% worldwide) is thought to be a consequence of the shift in modes of subsistence and lifestyle in the last 10,000 years, which triggered significant changes in diet and human exposure to xenobiotic compounds. Several surveys of *NAT2* sequence variation have indeed provided compelling evidence that at least some of the slow-causing variants of *NAT2* have been driven to present-day frequencies through the action of natural selection (Luca et al., 2008; Magalon et al.,

2008; Mortensen et al., 2011; Patin et al., 2006; Sabbagh et al., 2011). The slow acetylation phenotype may thus have been a key adaptation to increase our species fitness in response to the transition from foraging to farming.

Striking patterns of geographic differentiation have been described for the major *NAT2* slow-causing variants (García-Martín, 2008; Sabbagh et al., 2011). The function of NAT2 in mediating the interactions between humans and their chemical environment, which varies depending on diet and lifestyle, makes it an excellent candidate for population-specific selection pressures. Notably, an unusually high level of population differentiation between East Asians and other populations (F_{ST} values around 0.40) has been described for the c.341T>C slow-causing variant (rs1801280), as well as the two linked c.481C>T (rs1799929) and c.803A>G (rs1208) nonfunctional SNPs, when compared to an empirical distribution of F_{ST} computed across a 400-kb region encompassing the whole human *NAT* gene family (Sabbagh et al., 2008). In contrast, the slow 590A variant (rs1799930) was found to occur at roughly similar frequencies among widely dispersed populations (Luca et al., 2008; Sabbagh et al., 2011). Such a low level of geographic differentiation may rather suggest a homogenizing process of natural selection, promoting the same allelic variant in otherwise disparate populations (through either directional or balancing selection). Although many polymorphisms have been described in other regions of the *NAT2* gene (Mortensen et al., 2011; Patin et al., 2006), limited data exist on the geographic distribution of these variants in worldwide populations.

In this study, we took advantage of the whole-genome sequence data available from the 1000 Genomes (1KG) project to explore the global patterns of population genetic differentiation for the whole set of variants occurring in the entire *NAT2* gene sequence (~10 kb). An outlier approach was used to determine whether the patterns of geographic differentiation at this locus were atypical compared to those observed for the remaining

variation of the genome. Selection tests based on the site frequency spectrum and extended haplotype homozygosity were further applied to determine the possible role of natural selection in shaping the atypical patterns observed.

Materials and Methods

Data retrieval

Whole-genome variation data generated by the 1KG project in 1,089 unrelated individuals was directly downloaded from the 1000 Genomes ftp site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/), using the phase 1 integrated release version 3 of April 2012 (1000 Genomes Project Consortium et al., 2012). The 1,089 individuals are drawn from 14 different populations in sub-Saharan Africa, Europe, East Asia, and the Americas: Yoruba in Ibadan, Nigeria (YRI); Luhya in Webuye, Kenya (LWK); people with African ancestry in the Southwest United States (ASW); Utah residents with Northern and Western European ancestry (CEU); Tuscans in Italy (TSI); British in England and Scotland (GBR); Finnish in Finland (FIN); Iberians in Spain (IBS); Han Chinese in Beijing, China (CHB); Southern Han Chinese in China (CHS); Japanese in Tokyo, Japan (JPT); people with Mexican ancestry in Los Angeles, California (MXL); Colombians in Medellin, Colombia (CLM); and Puerto Ricans in Puerto Rico (PUR). From the obtained vcf (variant call format) files, we extracted exclusively the low-coverage VQSR (Variant Quality Score Recalibrator method) SNV calls in order to avoid any bias that might result from differences between low-coverage whole-genome calls and high-coverage exome SNV calls. Indels were not used. Functional annotation of the 36,382,866 SNVs retrieved was performed using classification from the dbSNP database (build 137) (http://genome.ucsc.edu/cgi-bin/hgTables:SNV137.txt). SNVs were assigned to two main classes: genic and nongenic

SNVs. Genic SNVs were further classified as intronic, 5'-UTR, 3'-UTR, coding synonymous, coding non-synonymous or splice-site.

Population genetic differentiation

Global levels of population genetic differentiation at *NAT2* (chr8:18248755-18258723 in the human GRCh37/hg19 assembly) were evaluated by using the fixation index F_{ST} (Wright, 1951) which quantifies the proportion of genetic variance explained by allele frequency differences among populations. F_{ST} ranges from 0 (for genetically identical populations) to 1 (for completely differentiated populations). F_{ST} scores were computed for all *NAT2* SNVs occurring with a minor allele frequency (MAF) $\geq$ 0.05 in at least one of the 14 1KG populations, using the BioPerl module PopGen (Stajich et al., 2002). Extreme values of F_{ST} can result from natural selection but also from nonselective events linked to the demography of populations, such as genetic drift. Because such nonselective processes randomly act on the genome, they are expected to have the same average effect across the genome, in contrast to natural selection, which impacts population differentiation in a locus-specific manner. The genome-wide variation data provided by the 1KG project can thus be used to infer the action of natural selection by adopting an outlier approach (Kelley et al., 2006). For that purpose, we built eight empirical distributions of the F_{ST} statistic by considering different subsets of SNVs defined according to their physical location and/or functional impact. To obtain distributions of likely independent observations, a LD-based pruning procedure was applied to each of these eight subsets using Plink (Purcell et al., 2007) with default parameters (pruning based on a variance inflation factor of at least 2 within each sliding window of 50 SNVs with a step of five SNVs). This resulted in a total of 25,532,386 independent autosomal SNVs included in the genome-wide empirical distribution. These numbers are respectively 15,141,160, 11,282,100, 10,477,050, 24,395, 198,718, 107,644, 146,572, and 1,912 in the nongenic, genic,

intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice-site distributions. These eight distributions were then used as reference to assess whether the patterns of genetic differentiation observed at *NAT2* are atypical. Empirical *P*-values were estimated as the proportion of F_{ST} scores in the empirical distribution that are either higher (diversifying selection) or lower (homogenizing selection) than the value observed at the locus of interest. Since F_{ST} strongly correlates with heterozygosity (Barreiro et al., 2008; Beaumont and Nichols, 1996; Elhaik, 2012), empirical *P*-values were calculated within bins of SNVs grouped according to their global MAF. A total of 27 bins were considered for the whole MAF range: 10 bins of size 0.001 for MAF between 0 and 0.01, 9 bins of size 0.01 for MAF between 0.01 and 0.10, and 8 bins of size 0.05 for MAF between 0.10 and 0.50.

Selection tests

To determine whether natural selection has played a role in the unusual patterns of geographic differentiation disclosed, we used two complementary approaches based on the allele frequency spectrum of segregating sites and on the local haplotype structure. Tajima's D (Tajima, 1989) is a classical neutrality test that compares estimates of the number of segregating sites and the average number of pairwise differences between nucleotide sequences (π). A zero value of the test statistic *D* is expected under the null hypothesis of selective neutrality, whereas a positive *D* is taken as indicative of balancing selection and a negative one of directional selection. Tajima's *D* scores were computed across the whole *NAT2* coding region by using a sliding window approach with a window size of 1 kb and a step size of 100 bp. Statistical significance of the test statistic was assessed using an empirical approach. From the genome-wide data available from the 1KG project, we selected a set of unlinked noncoding regions expected to be mostly neutrally evolving. A total of 100 autosomal regions of size 1kb were selected that met the following criteria: (i) to be at least

100 kb away from any known or predicted genes or expressed sequence tag or region transcribed into mRNA; (ii) to be outside any segmental duplication or region transcribed into a long noncoding RNA or conserved noncoding element, as defined in Woolfe *et al.* (Woolfe et al., 2007); (iii) to be distant from each other by at least 100 kb and not in linkage disequilibrium (LD) with each other; (iv) to contain a number of SNVs equal to the mean number of SNVs included in the 1-kb sliding windows spanning the *NAT2* coding region. Tajima's *D* scores were computed for these 100 regions so as to obtain the null (neutral) distribution of the test statistic in each population sample. An empirical *P*-value was estimated at each sliding window position within *NAT2* by considering the proportion of regions showing a test statistic greater (excess of intermediate-frequency variants) or lower (excess of low-frequency variants) than the value observed at that specific position.

We next used methods based on the extended haplotype homozygosity (EHH) measure, *i.e.* the sharing of identical alleles across relatively long distances by most haplotypes in a population sample (Sabeti et al., 2002). We calculated the integrated haplotype score (iHS) (Voight et al., 2006) that compares the rate of EHH decay observed for both the derived and ancestral allele at each core SNV. An extremely positive or negative value at the core SNV provides evidence of positive selection with unusually long haplotypes carrying the ancestral or the derived allele, respectively. The raw iHS scores were computed for all *NAT2* SNVs using the iHS option implemented in the WHAMM! Software (Voight et al., 2006), which we slightly modified in order to speed up computation times: thresholds for EHH decay were modified from 0.25 to 0.15 and the size of the analyzed region was set to 0.2 Mb instead of 2.5 Mb. Information on ancestral allele state was obtained from a four-way alignment of human, chimpanzee, orangutan and rhesus macaque species, provided by the 1KG

consortium (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/).

We also applied a cross-population test by computing the XP-EHH statistic (Sabeti et al., 2007) that compares the integrated EHH computed in a test population versus that of a reference population. XP-EHH scores were computed using the same EHH decay parameters and window size as for iHS. The Yoruba (YRI) sample was used as a reference for samples outside Africa, and the Utah residents of European ancestry (CEU) as a reference for African samples.

iHS and XP-EHH scores were also computed for all available SNVs in the 100 neutral regions described above, thus providing a reference distribution for each test statistic to estimate empirical *P*-values. Raw scores of iHS and XP-EHH in *NAT2* were standardized in bins of derived allele frequency (step size of 0.05) using the corresponding distribution of each statistic. For both iHS and XP-EHH, we used the phased data provided by the integrated Phase 1 release version 3 of the 1KG project (April 2012) and genetic distances were obtained from the high density genetic combined map based on 1KG pilot 1 data.

Estimation of the age of the derived allele at rs1799930

The age of the derived A allele at SNP rs1799930 was estimated using the maximum likelihood method implemented in the Estiage software (Génin et al., 2004). This method was originally developed to estimate the age of the most recent common ancestor of a rare mutation involved in rare disease using microsatellite data. It is based on the length of ancestral haplotype segments around the mutation shared by mutation carriers. Haplotypes in a 20 kb region centered around rs1799930 were obtained from the phased 1000 Genome data. Only SNV were considered and their genetic positions were obtained from the genetic maps provided on the Beagle website

(http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/). Only one SNV every 0.01 cM was kept in the final analysis with Estiage. Age estimates were derived in each population separately and a mutation rate of 10^{-3} per marker per individual per generation was assumed. The mutation rate is a key parameter in Estiage that accounts for the fact that the most proximal marker where a different allele from the ancestral haplotype is observed in an individual might in fact not indicate a recombination but rather a mutation or a genotyping error. For SNVs, mutation rates are assumed to be very low in the order of 10^{-6} but genotyping errors can be relatively high. A sensibility analysis was performed to choose the value of the mutation rate in a range between 10^{-6} and 10^{-2} that gave coherent estimates between the different populations and 10^{-3} was found to be the most plausible value. Population allele frequencies at the different markers were estimated on the haplotypes not carrying the A alleles at rs1799930.

In-silico prediction of SNV's functional effects

The F-SNP method (Lee and Shatkay, 2008) (http://compbio.cs.queensu.ca/F-SNP/) was applied to assess the potential functional effect of SNVs. This integrative scoring method combines assessments from 16 independent computational tools and databases, using a probabilistic framework that takes into account both the certainty of each prediction and the reliability of the different tools depending on the physical and functional annotation of the specific variant tested. It provides a functional significance (FS) score that quantitatively measures the possible deleterious effect of the tested SNV at the splicing, transcriptional, translational and post-translational levels. A FS score of 0.5 is considered as the cutoff point for predicting a deleterious effect (Lee and Shatkay, 2009).

Results and Discussion

Global patterns of population genetic differentiation were examined in the genomic region spanning the entire *NAT2* gene (~10 kb), using the F_{ST} statistic (Wright, 1951) and the sequence variation data provided by the 1KG project (1000 Genomes Project Consortium et al., 2012). *P*-values were estimated from empirical distributions built from the background genomic variation (see Materials and Methods). We assigned to each genetic variant of the *NAT2* gene a 'main *P*-value' derived from the genome-wide empirical distribution and a 'subset *P*-value' derived from the distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, nongenic, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice site).

No SNVs in *NAT2* exhibited significantly high F_{ST} values compared to the genomic background, when considering the global differentiation among the 14 worldwide populations from 1KG (all *P*-values > 0.05). No significant *P*-values were observed when contrasting East Asia to the rest of the world either (data not shown). Although high F_{ST} scores were observed for the three SNVs c.341T>C (rs1801280), c.481C>T (rs1799929) and c.803A>G (rs1208) in this specific pairwise comparison ($F_{ST} \approx 0.30$), they could not be considered as atypical when compared to the rest of the genome (*P*-values ranging from 0.06 to 0.09). This contrasts with previous findings pointing out an atypical pattern of differentiation for these three variants when considering HapMap data and only the set of variants located within a 400-kb region surrounding the *NAT2* gene as a reference distribution (Sabbagh et al., 2008). This difference may be due to the different set of populations surveyed or to the more accurate empirical distribution used to represent the background genomic variation in the present study.

In contrast, five *NAT2* SNVs exhibited unusually low F_{ST} values when compared to the genome average, with both main and subset *P*-values below 0.05 (Figure 1). Four of them (rs6984200, rs2087852, rs11996129 and rs1112005) are located in the intronic region of

NAT2, while the fifth one (rs1799930) is a non-synonymous substitution defining the *NAT2*6* slow haplotype cluster (c.590G>A resulting in R197Q). Note that the four intronic SNVs are in high LD with the rs1799930 variant (r^2 ranging from 0.80 to 0.88) (Table 1). They all occur at high frequencies in the global human population (within the 0.23-0.26 MAF range). Such low levels of population genetic differentiation suggest that at least one of these polymorphisms may be subject to balancing or species-wide directional selection, the rs1799930 being the most likely target given its gene location and functional impact.

To determine whether another putative candidate in the genomic region surrounding *NAT2* might explain the patterns observed, through a significant LD with these variants, we extended the analysis to a 600-kb region centered on the human *NAT* multigene family on chromosome 8 (Figure 2). All the variants exhibiting an r^2 value above the 0.10 threshold with the rs1799930 SNV are located within a 56-kb region (chr8:18229877-18285763 in hg19) in the direct vicinity of the *NAT2* gene (Figure 2A), making unlikely the involvement of another gene in the region. Table 1 provides the list of variants showing a significantly low interpopulation F_{ST} value for either the main or subset *P*-value and being in moderate to strong LD with the rs1799930 variant ($r^2 > 0.50$). They are all either intergenic (located up to ~3 kb upstream or 22 kb downstream of the *NAT2* gene) or intronic to *NAT2*. The prediction of their functional impact with the F-SNP method revealed high FS scores for some of them (FS = 0.50), denoting a potentially deleterious effect by affecting either the transcriptional or splicing regulation. However, the highest FS score was observed for the rs1799930 polymorphism (0.87), which also displayed the lowest subset *P*-value ($P = 0.016$; Table 1). Altogether, these results point to the rs1799930 polymorphism as the most likely target of homogenizing selection in the genomic region surveyed.

Interestingly, recent evidence suggest that the *NAT2*6* haplotype cluster (characterized by the rs1799930 A slow-causing allele) is related with the slowest acetylation capacity *in*

vivo, and that the homozygous genotype *NAT2*6/*6* thus defines a new category of 'ultra-slow' acetylators (Ruiz et al., 2012; Selinski et al., 2013). Ultra-slow acetylators have about 30% lower activities of caffeine metabolism compared with other slow acetylators. This is of the same order of magnitude than the reduction in enzyme activity between rapid and intermediate acetylators (Ruiz et al., 2012). These findings are consistent with a previous study by Cascorbi *et al.* (Cascorbi et al., 1995) that demonstrated a markedly decreased NAT2 activity *in vivo* in *NAT2*6*/**6* compared to *NAT2*5*/**5* genotypes. Indirect evidence is also provided by clinical association studies related to both drug toxicity and cancer risk. Anti-tuberculosis drug-induced hepatotoxicity risk has been shown to be particularly high in carriers of the *NAT2*6/*6* genotype (An et al., 2012; Huang et al., 2002; Lee et al., 2010; Leiro-Fernandez et al., 2011; Selinski et al., 2014; Teixeira et al., 2011). Similarly, the ultra-slow genotype, and not the common slow *NAT2* genotype, has been significantly associated with an increased risk of urinary bladder cancer (Selinski et al., 2013). Altogether, both *in vivo* phenotyping studies and clinical reports suggest that the *NAT2*6* variant is likely to be associated with a specific acetylation phenotype. This could explain why this particular *NAT2* slow-causing variant, and not another one, may have been a specific target of natural selection.

Although a highly homogenous distribution of the rs1799930 A allele is observed across worldwide populations (with a global frequency of 0.246), resulting in an unusually low F_{ST} value (interpopulation F_{ST} = 0.006), a three-fold lower frequency of this allele has been reported in hunter-gatherers (~ 0.08) as compared to agriculturalists and pastoralists (~ 0.25) in a comprehensive survey of human *NAT2* variation including 128 population samples classified according to their major subsistence strategy (Sabbagh et al., 2011). This significant difference in the frequency of *NAT2*6* alleles ($P < 0.0001$) was identified as the main genetic cause of the higher prevalence of the slow acetylation phenotype in populations practicing

farming and herding as compared to those mostly relying on hunting and gathering (46% *vs* 22%, respectively) (Sabbagh et al., 2011). Given this marked correlation between the rs1799930 A allele and the subsistence strategy adopted by past populations in the last 10,000 years, it has been suggested that this slow-causing allele may have conferred a selective advantage in populations shifting from foraging to agricultural and pastoral activities in the Neolithic period. New or more concentrated NAT2 substrates introduced in the chemical environment of food-producing communities have likely promoted a slower acetylation rate in these populations. This hypothesis is further supported by the age estimation of the rs1799930 A allele in the 1KG populations provided by a maximum-likelihood method implemented in the Estiage software (Génin et al., 2004) (Table 2). Our estimations showed that this allele started to increase in frequency at similar and recent times in all populations, roughly between ~1,500 and 6,000 years ago, thereby supporting a global expansion since the emergence of agriculture in the Neolithic. Consequently, the markedly low level of population differentiation observed at the rs1799930 locus may result from the convergent selection of the rs1799930 A allele in agriculturalist and pastoralist populations which are now present in most parts of the world.

Several lines of evidence support the hypothesis that the rs1799930 G>A non-synonymous substitution (R197Q) has specifically occurred in the human lineage. First, the NAT2 197R residue appears to be highly conserved throughout primate evolution, with 100% of the orthologous *NAT2* sequences generated in 19 distinct simian species harboring an arginine (R) at this position (Sabbagh et al., 2013). Second, the 197R position was found to be monomorphic in 103 individuals from six great ape species (*Pan troglodytes*, *Pan paniscus*, *Gorilla beringei*, *Gorilla gorilla*, *Pongo abellii*, *Pongo pygmaeus*) (Prado-Martinez *et al.* (Prado-Martinez et al., 2013), E.S. Poloni, personal communication), as well as in 28 Rhesus monkeys (*Macaca mulatta*) fully sequenced for the *NAT2* gene (A. Sabbagh, personal

communication), making the R197Q polymorphism a specific feature of the human lineage. The hypothesis of a trans-species polymorphism maintained for several million years, through shared balancing selection pressures, seems therefore unlikely.

Assuming that the rs1799930 A allele has conferred a selective advantage to populations shifting from food collection to farming and animal breeding in the Holocene, this could have happened either through directional selection or balancing selection. A gene-dose effect has been indeed described for this variant, with a significant trend toward a slower acetylation capacity in individuals carrying an increasing number of *NAT2*6* haplotypes (0, 1 or 2) (Ruiz et al., 2012; Selinski et al., 2013). Therefore, heterozygous individuals for this allele display an intermediate metabolic phenotype that may have been advantageous if one considers the competing needs of both maintaining an efficient detoxification of harmful xenobiotics and avoiding the damaging effects of the putative carcinogens that can be activated through NAT2 acetylation. In an attempt to provide further insights into the evolutionary mechanisms that might have driven and maintained the rs1799930 A allele at high frequencies in most human populations worldwide, we carried out several tests of selective neutrality based on the allele frequency spectrum: Tajima's D (Tajima, 1989) and haplotype structure: iHS (Voight et al., 2006) and XP-EHH (Sabeti et al., 2007). An empirical approach using sequence variation data from 100 unlinked noncoding regions was adopted to assess statistical significance.

All iHS and XP-EHH scores computed for the rs1799930 SNV in all individual populations from 1KG were not significant at the 0.05 threshold (Table 3). This precludes a clear signal of positive selection for this variant as the one expected under a 'hard sweep model', which assumes the rapid fixation of a single newly arisen advantageous mutation (Pritchard et al., 2010). In contrast, we found significant Tajima's *D* scores in the 1-kb regions encompassing the rs1799930 variant in five population samples: British, Finnish, Tuscans, Utah residents of European ancestry and Puerto Ricans ($P < 0.05$) (Tables 3 and 4). We

acknowledge that these results become non-significant when a correction for multiple testing is applied, but we also note that the ratio of five significant tests out of 14 is higher than the expected 5% proportion of false positives. Non-significant scores, but with *P*-values getting closer to the 5% threshold, were observed in two additional samples ($P = 0.07$ and $P = 0.08$ in Colombians and Luhya, respectively). Furthermore, although non-significant scores prevent rejection of the null hypothesis of selective neutrality, it is noteworthy that all populations tested but one (Japanese) gave positive Tajima's *D* values, suggesting a trend toward an excess of intermediate-frequency variants compatible with the action of balancing selection. Such consistent results for populations with different demographic pasts make it unlikely that they are due to demography rather than balancing selection. This is also in agreement with previous findings demonstrating globally positive and significant Tajima's *D* values in different continental populations and the absence of any signature of positive selection, as detected by EHH-based tests (Fuselli et al., 2007; Luca et al., 2008; Magalon et al., 2008; Mortensen et al., 2011). A notable exception concerns the c.341C>T slow-causing variant for which a selective sweep was detected in Western and Central Eurasian populations (Patin et al., 2006) with the long-range haplotype test (Sabeti et al., 2002). We did not confirm such signature of positive selection at this locus in any of the 14 populations from 1KG (both iHS and XP-EHH scores not significant at the 0.05 level). No significant scores were observed for any of the other slow-causing variants either (c.191G>A, c.341T>C and c.857G>A; data not shown). Therefore, patterns of diversity at *NAT2* seem compatible with either balancing selection or a more complex model of 'multiallelic' directional selection where different slow variants of *NAT2* may have simultaneously become targets of directional selection, thereby generating an excess of intermediate-frequency alleles. This would explain why our conventional tests of selection based on EHH, more suited to detect classical selective sweeps, failed to detect a signature of positive selection at the rs1799930 locus. The signature of

selection at this individual position could have been weakened by the global increase in frequency of other *NAT2* altering mutations. Note, however, that contrary to c.191G>A (*NAT2*14*), c.341T>C (*NAT2*5*) and c.857G>A (*NAT2*7*), which mainly cluster in specific continental regions (sub-Saharan Africa, Europe and Asia, respectively (Sabbagh et al., 2011), the cosmopolitan distribution of the c.590G>A variant (*NAT2*6*) suggests that it may have been positively selected in globally distributed food-producing communities. Finally, the hypotheses of balancing and directional selection are not mutually exclusive and multiple modes of selection may have operated at the *NAT2* locus on a population-specific basis, as previously suggested (Mortensen et al., 2011).

In conclusion, we have described an atypical pattern of geographic differentiation for five genetic variants of the *NAT2* gene, including the functional rs1799930 SNP defining the slow *NAT2*6* haplotype series, and four intronic SNPs in high LD with it. An extended analysis of a 600-kb region surrounding *NAT2* pointed to the rs1799930 polymorphism as the most likely target of a homogenizing process of natural selection promoting the same allelic variant in most human populations, resulting in an unusually low F_{ST} value (F_{ST} = 0.006). The rs1799930 A allele has been associated with the slowest acetylation capacity *in vivo* and is much more frequent in agriculturalists and pastoralists as compared to hunter-gatherers, suggesting it may have been positively selected in food-producing communities which are now present in most parts of the world. Neutrality tests based on the allele frequency spectrum revealed a trend toward an excess of intermediate-frequency variants at *NAT2*, compatible with either balancing selection or a more complex model of multiallelic directional selection. Our findings provide further insights into the functional importance of the rs1799930 polymorphism and the role it may have played in human adaptation to fluctuating xenobiotic environments.

Acknowledgements

This work was financially supported by the Institut Medicament-Toxicologie-Chimie. Environnement (IMTCE). BP is supported by a PhD fellowship from the doctoral program in Public Health from Paris Sud University. PL is supported by a PhD fellowship from ‘Acción Estratrégica de Salud, en el Marco del Plan Nacional de Investigacio´n Científica, Desarrollo e Innovación Tecnológica 2008–2011’ from Instituto de Salud Carlos III.

Literature Cited

1000 Genomes Project Consortium, G.R. Abecasis, A. Auton, L.D. Brooks et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature. 491, 56–65.

Agúndez, J.A.G. 2008. Polymorphisms of human N-acetyltransferases and cancer risk. Curr. Drug Metab. 9, 520–531.

An, H.-R., X.-Q. Wu, Z.-Y. Wang et al. 2012. NAT2 and CYP2E1 polymorphisms associated with antituberculosis drug-induced hepatotoxicity in Chinese patients. Clin. Exp. Pharmacol. Physiol. 39, 535–543.

Barreiro, L.B., G. Laval, H. Quach et al. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet. 40, 340–5.

Beaumont, M.A., and R.A. Nichols. 1996. Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society B. 263, 1619–1626.

Butcher, N.J., S. Boukouvala, E. Sim, and R.F. Minchin. 2002. Pharmacogenetics of the arylamine N-acetyltransferases. Pharmacogenomics J. 2, 30–42.

Cascorbi, I., N. Drakoulis, J. Brockmöller et al. 1995. Arylamine N-acetyltransferase (NAT2) mutations and their allelic linkage in unrelated Caucasian individuals: correlation with phenotypic activity. Am. J. Hum. Genet. 57, 581–592.

Elhaik, E. 2012. Empirical distributions of F(ST) from large-scale human polymorphism data. PloS One. 7, e49837.

Fuselli, S., R.H. Gilman, S.J. Chanock et al. 2007. Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. Pharmacogenomics. J. 7, 144–152.

García-Martín, E. 2008. Interethnic and intraethnic variability of NAT2 single nucleotide polymorphisms. Curr. Drug Metab. 9, 487–497.

Génin, E., A. Tullio-Pelet, S. Lyonnet, L. Abel. 2004. Estimating the age of rare disease mutations: the example of Triple A syndrome. J. Med. Genet. 41, 445-449.

Hein, D.W. 2002. Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. Mutat. Res. 506-507, 65–77.

Hein, D.W. 2006. N-acetyltransferase 2 genetic polymorphism: effects of carcinogen and haplotype on urinary bladder cancer risk. Oncogene. 25, 1649–1658.

Hill, W.G. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 226–231.

Huang, Y.-S., H.-D. Chern, W.-J. Su et al. 2002. Polymorphism of the N-acetyltransferase 2 gene as a susceptibility risk factor for antituberculosis drug-induced hepatitis. Hepatol. Baltim. Md 35, 883–889.

Kelley, J.L., J. Madeoy, J.C. Calhoun et al. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 16, 980–989.

Kilbane, A.J., L.K. Silbart, M. Manis et al. 1990. Human N-acetylation genotype determination with urinary caffeine metabolites. Clin. Pharmacol. Ther. 47, 470–477.

Ladero, J.M. 2008. Influence of polymorphic N-acetyltransferases on non-malignant spontaneous disorders and on response to drugs. Curr. Drug Metab. 9, 532–537.

Lee, P.H., and H. Shatkay. 2008. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res. 36, D820–D824.

Lee, P.H., and H. Shatkay. 2009. An integrative scoring system for ranking SNPs by their potential deleterious effects. Bioinformatics. 25, 1048–1055.

Lee, S.-W., L.S.-C. Chung, H.-H. Huang et al. 2010. NAT2 and CYP2E1 polymorphisms and susceptibility to first-line anti-tuberculosis drug-induced hepatitis. Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis. 14, 622–626.

Leiro-Fernandez, V., D. Valverde, R. Vázquez-Gallardo et al. 2011. N-acetyltransferase 2 polymorphisms and risk of anti-tuberculosis drug-induced hepatotoxicity in

Caucasians. Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis. 15, 1403–1408.

Luca, F., G. Bubba, M. Basile et al. 2008. Multiple advantageous amino acid variants in the NAT2 gene in human populations. PloS One. 3, e3136.

Magalon, H., E. Patin, F. Austerlitz et al. 2008. Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. Eur. J. Hum. Genet. 16, 243–251.

Meisel, P. 2002. Arylamine N-acetyltransferases and drug response. Pharmacogenomics. 3, 349–366.

Mortensen, H.M., A. Froment, G. Lema et al. 2011. Characterization of genetic variation and natural selection at the arylamine N-acetyltransferase genes in global human populations. Pharmacogenomics. 12, 1545–1558.

Parkin, D.P., S. Vandenplas, F.J. Botha et al. 1997. Trimodality of isoniazid elimination: phenotype and genotype in patients with tuberculosis. Am. J. Respir. Crit. Care Med. 155, 1717–1722.

Patin, E., L.B. Barreiro, P.C. Sabeti et al. 2006. Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. Am. J. Hum. Genet. 78, 423–436. doi:10.1086/500614

Prado-Martinez, J., P.H. Sudmant, J.M. Kidd et al. 2013. Great ape genetic diversity and population history. Nature. 499, 471–475.

Pritchard, J.K., J.K. Pickrell, and G. Coop. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 20, R208–15.

Purcell, S., B. Neale, K. Todd-Brown et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Ruiz, J.D., C. Martínez, K. Anderson et al. 2012. The differential effect of NAT2 variant alleles permits refinement in phenotype inference and identifies a very slow acetylation genotype. PloS One. 7, e44629.

Sabbagh, A., and P. Darlu. 2006. SNP selection at the NAT2 locus for an accurate prediction of the acetylation phenotype. Genet. Med. Off. J. Am. Coll. Med. Genet. 8, 76–85.

Sabbagh, A., P. Darlu, B. Crouau-Roy, and E.S. Poloni. 2011. Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. PloS One. 6, e18507.

Sabbagh, A., A. Langaney, P. Darlu et al. 2008. Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet. 9, 21.

Sabbagh, A., J. Marin, C. Veyssière et al. 2013. Rapid birth-and-death evolution of the xenobiotic metabolizing NAT gene family in vertebrates with evidence of adaptive selection. BMC Evol. Biol. 13, 62.

Sabeti, P.C., D.E. Reich, J.M. Higgins et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419, 832–7.

Sabeti, P.C., P. Varilly, B. Fry et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature. 449, 913–8.

Sanderson, S., G. Salanti, and J. Higgins. 2007. Joint effects of the N-acetyltransferase 1 and 2 (NAT1 and NAT2) genes and smoking on bladder carcinogenesis: a literature-based systematic HuGE review and evidence synthesis. Am. J. Epidemiol. 166, 741–751.

Selinski, S., M. Blaszkewicz, K. Ickstadt et al. 2013. Refinement of the prediction of N-acetyltransferase 2 (NAT2) phenotypes with respect to enzyme activity and urinary bladder cancer risk. Arch. Toxicol. 87, 2129–2139.

Selinski, S., M. Blaszkewicz, K. Ickstadt et al. 2014. Improvements in algorithms for phenotype inference: the NAT2 example. Curr. Drug Metab. 15, 233–249.

Stajich, J.E., D. Block, K. Boulez et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12, 1611–8.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123, 585–95.

Teixeira, R.L. de F., R.G. Morato, P.H. Cabello et al. 2011. Genetic polymorphisms of NAT2, CYP2E1 and GST enzymes and the occurrence of antituberculosis drug-induced hepatitis in Brazilian TB patients. Mem. Inst. Oswaldo Cruz. 106, 716–724.

Voight, B.F., S. Kudaravalli, X. Wen, and J.K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4, e72.

Walker, K., G. Ginsberg, D. Hattis et al. 2009. Genetic polymorphism in N-Acetyltransferase (NAT): Population distribution of NAT1 and NAT2 activity. J. Toxicol. Environ. Health B Crit. Rev. 12, 440–472.

Weber, W.W., and D.W. Hein. 1985. N-acetylation pharmacogenetics. Pharmacol. Rev. 37, 25–79.

Woolfe, A., D.K. Goode, J. Cooke et al. 2007. CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev. Biol. 7, 100.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15, 323–354.

Table 1. List of variants in the 600-kb region surrounding the *NAT2* gene with significantly low F_{ST} values and in high linkage disequilibrium with rs1799930

SNV	Physical position on chr 8 (hg19)	Functional annotation	Distance from *NAT2* (kb)	Global MAF	Inter-population F_{ST}	Main *P*-value	Subset *P*-value	r^2 with rs1799930	Functional significance score (F-SNP)
rs11992530	18246133	Intergenic	2.6	0.254	0.006	0.024	0.024	0.80	no known function
rs4646241	18246696	Intergenic	2.1	0.247	0.007	0.032	0.034	0.83	no known function
rs4646242	18247449	Intergenic	1.3	0.246	0.006	0.025	0.024	0.84	0.50
rs4646244	18247718	Intergenic	1.0	0.246	0.006	0.026	0.025	0.84	no known function
rs6984200	18250317	Intronic	0	0.255	0.007	0.028	0.028	0.80	0.50
rs2087852	18251926	Intronic	0	0.235	0.007	0.032	0.031	0.86	0.50
rs11996129	18254575	Intronic	0	0.236	0.005	0.021	0.020	0.85	0.37
rs1112005	18255876	Intronic	0	0.227	0.005	0.019	0.019	0.88	0.50
rs1799930	18258103	Coding non-synonymous	0	0.246	0.006	0.026	0.016	-	0.87
rs4646247	18258908	Intergenic	0.2	0.247	0.006	0.025	0.024	0.99	0.50
rs721398	18259305	Intergenic	0.6	0.244	0.005	0.019	0.020	0.97	no known function
rs45605031	18260239	Intergenic	1.5	0.248	0.006	0.024	0.026	0.98	no known function
rs872233	18280686	Intergenic	22.0	0.274	0.007	0.030	0.032	0.76	no known function

SNV, single nucleotide variant; MAF, minor allele frequency. Only variants with $r^2 > 0.50$ with rs1799930 are shown. SNVs located in the *NAT2* gene are shaded in grey. A functional significance score of 0.5 is considered as the cutoff point for predicting a deleterious effect (Lee and Shatkay, 2009).

Table 2. Estimation of the age of the derived allele at rs1799930

Population	Number of carrier haplotypes	Age in generations [95%CI]	Age in years [95%CI] (1 generation = 20 years)
YRI	40	294 [231-376]	5880 [4620-7520]
LWK	56	139 [111-174]	2780 [2220-3480]
ASW	34	202 [154-267]	4040 [3080-5340]
IBS	4	162 [76-356]	3240 [1520-7120]
TSI	54	129 [104-160]	2580 [2080-3200]
CEU	50	310 [250-386]	6200 [5000-7720]
GBR	51	74 [59-93]	1480 [1180-1860]
FIN	52	77 [61-188]	1540 [1220-3760]
JPT	43	101 [79-130]	2020 [1580-2600]
CHB	37	237 [184-308]	4740 [3680-6160]
CHS	49	203 [164-253]	4060 [3280-5060]
CLM	27	136 [100-188]	2720 [2000-3760]
MXL	16	138 [92-223]	2760 [1840-4460]
PUR	22	173 [124-248]	3460 [2480-4960]

The age of the derived A allele at SNP rs1799930 was estimated using the maximum likelihood method implemented in the Estiage software (Génin et al. 2004). YRI: Yoruba from Ibadan, Nigeria; LWK: Luhya fromWebuye, Kenya; ASW: people of African ancestry from the southwestern United States; IBS: Iberian populations from Spain; TSI: Tuscans from Italy; CEU: Utah residents with Northern and Westhern European ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; JPT: Japanese from Tokyo, Japan; CHB: Han Chinese from Beijing; CHS: Han Chinese from South China; CLM: Colombians from Medellín, Colombia; MXL: people of Mexican ancestry from Los Angeles, California; PUR: Puerto Ricans from Puerto Rico.

Table 3. Results of selection tests for the *NAT2* rs1799930 polymorphism

	Africa			Europe					Asia			America		
	YRI	LWK	ASW	IBS	TSI	CEU	GBR	FIN	JPT	CHB	CHS	CLM	MXL	PUR
iHS	0.18	-0.27	-0.05	-0.82	-0.08	0.17	0.02	-0.19	0.18	-0.11	0.06	0.33	0.09	-0.29
XP-EHH	-0.80	0.11	-0.50	1.38	0.54	0.80	0.68	0.42	0.20	0.42	0.02	0.19	0.07	0.13
Tajima's *D* [a]	0.33	1.00*	0.60	1.07	**2.58****	**2.49****	**3.28*****	**2.67****	-0.03	0.91	0.88	1.65*	0.99	**1.55****

Significant scores at the 0.05 threshold are shown in bold. YRI: Yoruba from Ibadan, Nigeria; LWK: Luhya fromWebuye, Kenya; ASW: people of African ancestry from the southwestern United States; IBS: Iberian populations from Spain; TSI: Tuscans from Italy; CEU: Utah residents with Northern and Westhern European ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; JPT: Japanese from Tokyo, Japan; CHB: Han Chinese from Beijing; CHS: Han Chinese from South China; CLM: Colombians from Medellín, Colombia; MXL: people of Mexican ancestry from Los Angeles, California; PUR: Puerto Ricans from Puerto Rico.

$^{*}\ P \leq 0.10,\ ^{**}\ P \leq 0.05,\ ^{***}\ P \leq 0.01$

[a] Highest Tajima's *D* score observed in the 1-kb sliding windows spanning the rs1799930 nucleotide position in each individual population. Note that, except in Asians (Japanese and both Han Chinese samples), Luhya and African-Americans, these scores also correspond to the highest values observed across the whole *NAT2* coding region (Table 3).

Table 4. Tajima's *D* scores in the 1-kb sliding windows spanning the whole *NAT2* coding region

Start[a]	End[a]	Slow-causing variant(s) included in the window[b]	Africa			Europe					Asia			America		
			YRI	LWK	ASW	IBS	TSI	CEU	GBR	FIN	JPT	CHB	CHS	CLM	MXL	PUR
18256555	18257555	none	0.12	-0.37	-0.40	-0.10	0.38	0.21	0.52	-0.45	-0.96	-1.17	-1.41	-0.27	-0.91	-0.23
18256655	18257655	none	0.16	-0.31	-0.32	-0.10	0.38	0.58	0.88	-0.26	-0.96	-1.17	-1.41	0.03	-0.78	0.32
18256755	18257755	c.191G>A	0.55	-0.28	0.04	0.46	0.38	0.78	0.92	0.02	-1.32	-1.05	-1.09	0.08	-0.67	0.23
18256855	18257855	c.191G>A, c.341T>C	1.22*	0.55	0.84	0.67	1.32	1.71	1.85	0.96	-0.61	-0.24	-0.26	0.93	0.09	1.10
18256955	18257955	c.191G>A, c.341T>C	0.94	0.41	0.55	0.67	1.32	1.71	1.85	0.96	-0.61	-0.24	-0.26	0.93	0.09	0.73
18257055	18258055	c.191G>A, c.341T>C	0.69	0.36	0.43	0.67	1.32	1.72	2.46**	1.35	-0.86	-0.24	-0.26	0.91	0.09	0.71
18257155	18258155	c.191G>A, c.341T>C, c.590G>A	0.33	0.82	0.60	0.58	1.57	1.95	3.28***	2.01*	-0.42	0.11	0.18	1.50*	0.13	1.14
18257255	18258255	c.191G>A, c.341T>C, c.590G>A	0.29	0.83	0.60	0.58	2.02	2.49**	3.28***	2.01*	-0.42	0.11	0.18	1.50*	0.13	1.14
18257355	18258355	c.191G>A, c.341T>C, c.590G>A	0.31	0.73	0.29	0.58	2.58**	2.49**	3.28***	2.01*	-0.42	0.11	0.19	1.50*	0.67	1.55**
18257455	18258455	c.191G>A, c.341T>C, c.590G>A, c.857G>A	0.17	1.00*	0.19	1.07	2.06	1.95	2.64***	2.67**	-0.37	0.35	0.78	1.65*	0.96	1.37*
18257555	18258555	c.191G>A, c.341T>C, c.590G>A, c.857G>A	-0.007	0.53	0.008	0.70	1.71	1.61	2.23*	2.18*	-0.12	0.35	0.78	1.28	0.99	1.13
18257655	18258655	c.191G>A, c.341T>C, c.590G>A, c.857G>A	-0.007	0.53	0.008	0.70	1.71	1.61	2.23*	2.18*	-0.12	0.35	0.78	1.28	0.99	1.13
18257755	18258755	c.341T>C, c.590G>A, c.857G>A	-0.04	0.58	0.06	0.70	2.16*	1.61	2.23*	2.18*	-0.12	0.35	0.41	1.28	0.99	1.51*
18257855	18258855	c.590G>A, c.857G>A	-0.70	-0.18	-0.62	0.54	0.96	0.87	1.51	1.44	-0.49	0.01	-0.007	0.61	0.03	0.83
18257955	18258955	c.590G>A, c.857G>A	-0.42	0.05	-0.47	0.45	1.27	1.21	1.81	1.73	-0.03	0.32	0.39	0.83	0.36	0.97
18258055	18259055	c.590G>A, c.857G>A	-0.82	-0.78	-0.88	0.10	0.22	-0.11	0.68	0.62	0.64	0.91	0.88	0.05	-0.22	0.21
18258155	18259155	c.857G>A	-0.67	-1.12	-1.05	0.14	-0.17	-0.51	0.29	0.23	0.23	0.70	0.54	-0.23	-0.03	0.004
18258255	18259255	c.857G>A	-0.47	-1.01	-0.92	0.14	-0.17	-0.51	0.29	0.23	-0.16	0.70	0.54	-0.23	-0.03	0.004
18258355	18259355	c.857G>A	-0.54	-0.90	-0.77	-0.79	-0.35	-0.61	0.12	0.07	0.58	1.42	1.20	-0.13	-0.51	-0.27

YRI: Yoruba from Ibadan, Nigeria; LWK: Luhya fromWebuye, Kenya; ASW: people of African ancestry from the southwestern United States; IBS: Iberian populations from Spain; TSI: Tuscans from Italy; CEU: Utah residents with Northern and Westhern European ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; JPT: Japanese from Tokyo, Japan; CHB: Han Chinese from Beijing; CHS: Han

Chinese from South China; CLM: Colombians from Medellín, Colombia; MXL: people of Mexican ancestry from Los Angeles, California; PUR: Puerto Ricans from Puerto Rico.

[*] $P \leq 0.10$, [**] $P \leq 0.05$, [***] $P \leq 0.01$

[a] Genomic position of each 1-kb window on chromosome 8 in the human GRCh37/hg19 assembly.

[b] The four slow causing variants c.191G>A, c.341T>C, c.590G>A and c.857G>A correspond to the rs1801279, rs1801280, rs1799930, rs1799931 SNPs, respectively.

Figure 1.

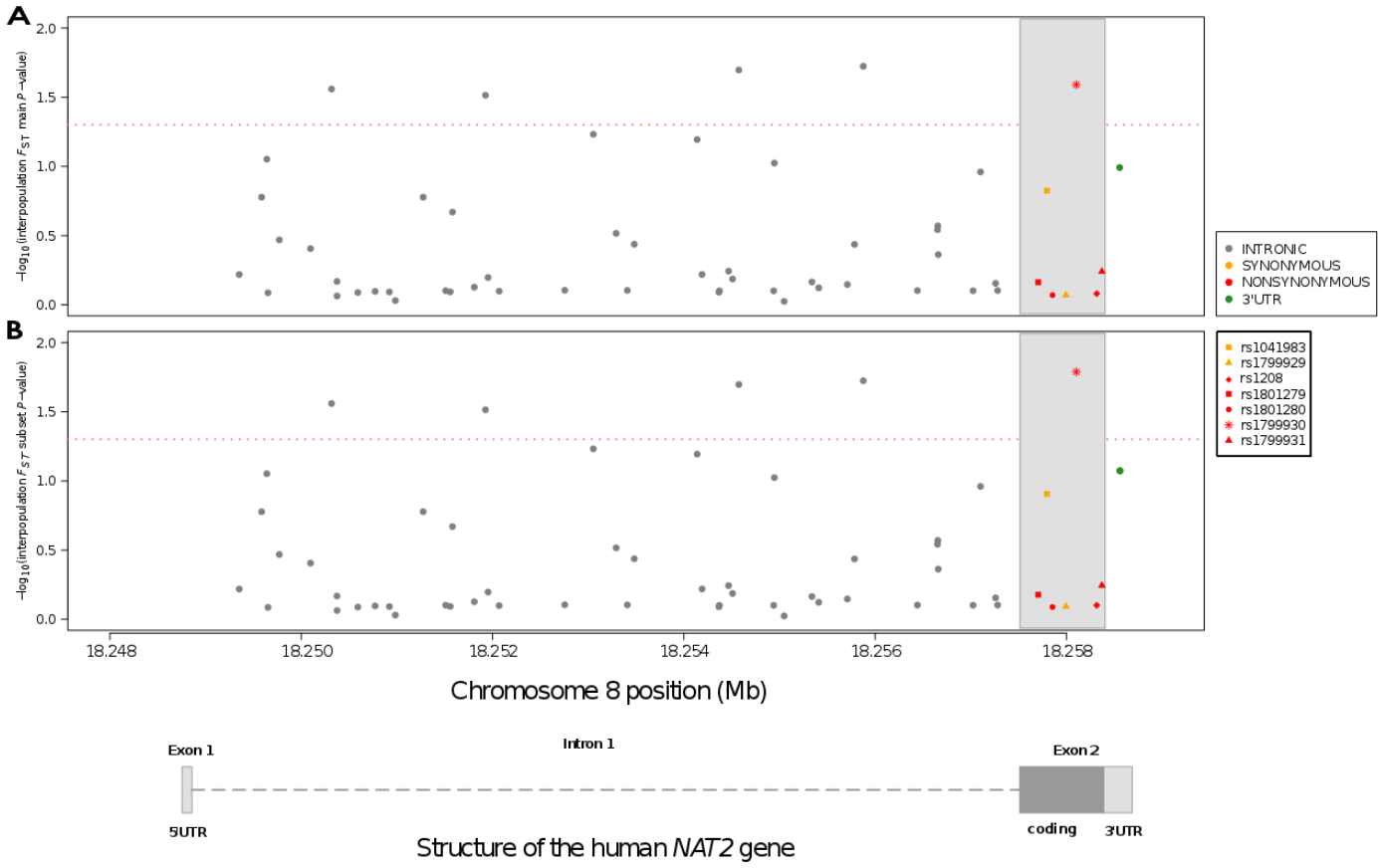
A
B
2.0
1.5
1.0
0.5
0.0
INTRONIC
SYNONYMOUS
NONSYNONYMOUS
3'UTR
rs1041983
rs1799929
rs1208
rs1801279
rs1801280
rs1799930
rs1799931
18.248
18.250
18.252
18.254
18.256
18.258
Chromosome 8 position (Mb)
Exon 1
Intron 1
Exon 2
5'UTR
coding
3'UTR
Structure of the human NAT2 gene

Figure 2.

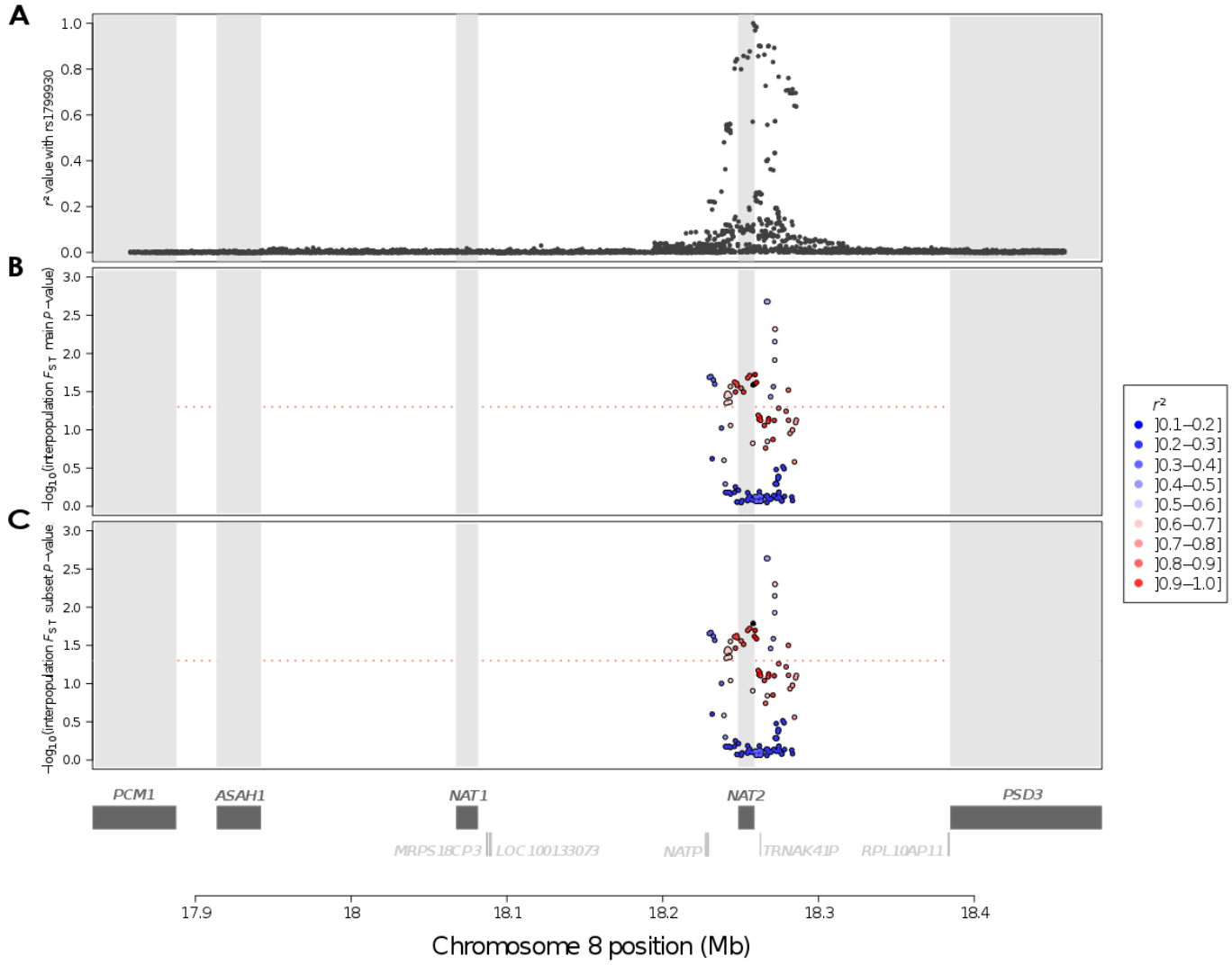
A
B
C
r² value with rs1799930
−log10(interpopulation FST main P–value)
−log10(interpopulation FST subset P–value
r²
]0.1–0.2]
]0.2–0.3]
]0.3–0.4]
]0.4–0.5]
]0.5–0.6]
]0.6–0.7]
]0.7–0.8]
]0.8–0.9]
]0.9–1.0]
PCM1
ASAH1
NAT1
NAT2
PSD3
MRPS18CP3
LOC100133073
NATP
TRNAK41P
RPL10AP11
17.9
18
18.1
18.2
18.3
18.4
Chromosome 8 position (Mb)

Figure 3.

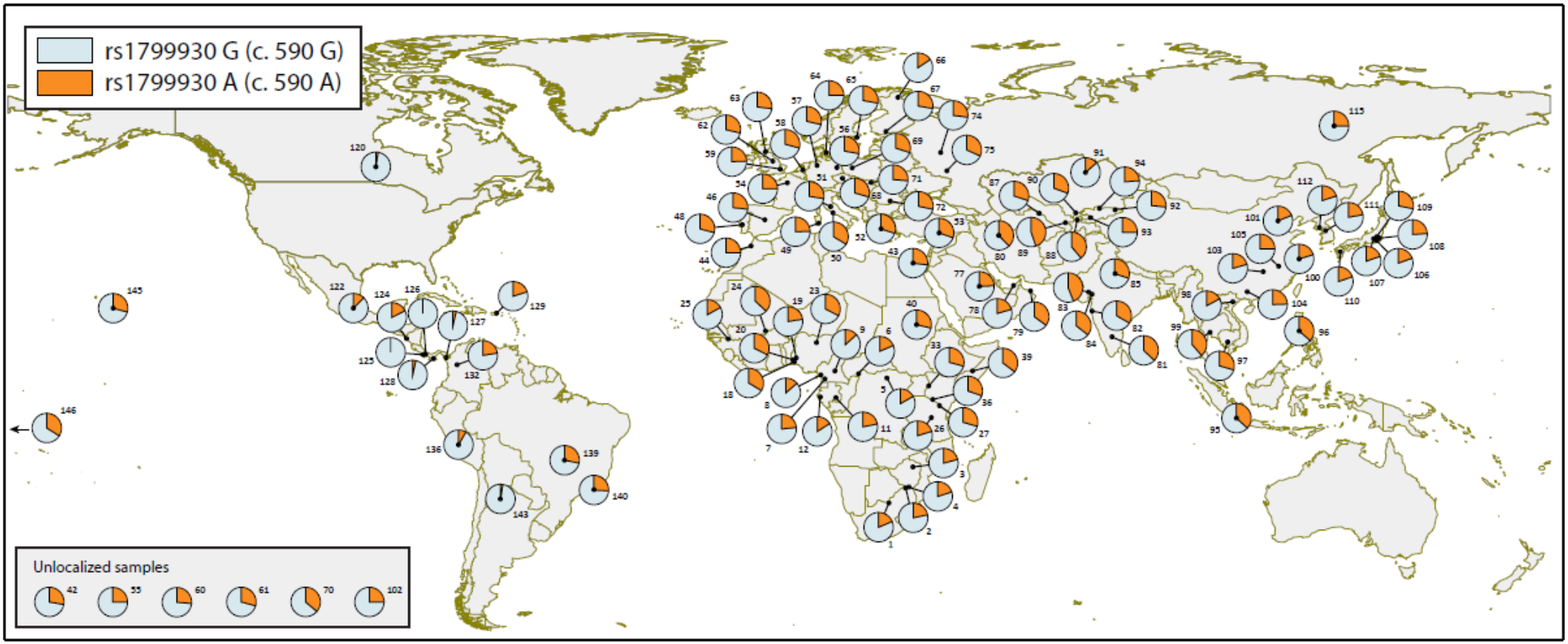
rs1799930 G (c. 590 G)
rs1799930 A (c. 590 A)
Unlocalized samples
42
55
60
61
70
102

Figure Captions

Figure 1. Distribution of -$\log_{10}$ (*P*-values) of interpopulation F_{ST} scores across the human *NAT2* gene. F_{ST} scores were computed among the 14 worldwide populations from the 1000 Genomes project. *P*-values were estimated from the lower tail of the empirical distributions. (A) Main *P*-value derived from the genome-wide empirical distribution (including 25,532,386 SNVs). (B) Subset *P*-value derived from the empirical distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, intronic, 3'UTR, coding synonymous, and coding non-synonymous). The red dotted line indicates the 0.05 significance threshold.

Figure 2. Distribution of -$\log_{10}$ (*P*-values) of interpopulation F_{ST} scores across a 600-kb region centered on the human *NAT* gene family for those variants in linkage disequilibrium (r^2 > 0.10) with the rs1799930 polymorphism. (A) Level of linkage disequilibrium, as measured by the r^2 statistic (Hill, 1968), with the rs1799930 genetic variant. **(B) Main *P*-value** estimated from the lower tail of the genome-wide empirical distribution (including 25,532,386 SNVs). **(C) Subset *P*-value** estimated from the lower tail of the empirical distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, nongenic, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice site). The red dotted line indicates the 0.05 significance threshold. SNVs are displayed in different colors according to their r^2 value with rs1799930, ranging from dark blue (r^2 = 0.10) to dark red (r^2 = 1.0). The rs1799930 polymorphism is represented as a black triangle. Coding genes and pseudogenes in the 600-kb region are represented below as dark grey and light grey boxes, respectively. The genomic position (in megabases) on chromosome 8 is indicated on the horizontal axis (human GRCh37/hg19 assembly).

Figure 3. Distribution of rs1799930 (c.590G>A) allele frequencies in human populations. Allele frequency data for the rs1799930 SNP were collected for 146 worldwide samples by performing an extensive survey of the literature. Description of samples and retrieved allele frequency data are provided in Table S1. Only those samples composed of at least 20 individuals (N = 105) were represented on the map. Samples are numbered as reported in Table S1 ('sample ID'). Six samples could not be localized on the map because of unspecified sampling location (sample 70) or because of divergence between sampling location and region of origin (samples 42, 55, 60, 61, and 102); these samples are displayed in a box beneath the caption.