HAL
open science

Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia.

Hélène Magalon, Etienne Patin, Frédéric Austerlitz, Tatyana Hegay, Almaz Aldashev, Lluis Quintana-Murci, Evelyne Heyer

To cite this version:

Hélène Magalon, Etienne Patin, Frédéric Austerlitz, Tatyana Hegay, Almaz Aldashev, et al.. Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia.. European Journal of Human Genetics, 2008, 16 (2), pp.243-51. 10.1038/sj.ejhg.5201963 . hal-00271644

HAL Id: hal-00271644

https://hal.science/hal-00271644

Submitted on 6 May 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Population genetic diversity of the *NAT2* gene supports a role of acetylation in human adaptation to farming in Central Asia

Hélène Magalon[1,6], Etienne Patin*[,1,2,6], Frédéric Austerlitz[3], Tatyana Hegay[4], Almaz Aldashev[5], Lluís Quintana-Murci[2] and Evelyne Heyer[1]

[1]*Unité d'Eco-Anthropologie et Ethnobiologie, CNRS UMR 5145, Musée de l'Homme, Paris, France;* [2]*Human Evolutionary Genetics, CNRS URA 3012, Institut Pasteur, Paris, France;* [3]*Laboratoire Ecologie Systématique et Evolution, CNRS UMR 8079, Orsay, France;* [4]*Institute of Immunology, Academy of Sciences, Tashkent, Uzbekistan;* [5]*National Academy of Sciences, Bishkek, Kyrgyz Republic*

The arylamine *N*-acetyltransferase 2 (NAT2) enzyme detoxifies a wide spectrum of naturally occurring xenobiotics including carcinogens and drugs. Variation at the *NAT2* gene has been linked to the human acetylation capacity, either 'slow' or 'fast', which modifies susceptibility to cancer and adverse drug reactions. We investigated the possible influence of natural selection in shaping the acetylation phenotype and the *NAT2* gene variability in six Central Asian populations, who are either long-term sedentary agriculturalists (two Tajik populations), recent sedentary agriculturalists (Kazakhs, Uzbeks) or nomad pastoralists (two Kirghiz populations). To this end, we sequenced the entire *NAT2* coding exon, as well as genotyping nine intergenic SNPs covering a 200-kb region. Our results revealed that the two Tajik populations exhibited significantly higher proportions of slow acetylators than the nomadic populations. In addition, sequence-based neutrality tests yielded significantly positive values in Central Asian populations following an agriculturalist lifestyle, due to an excess of haplotypes at intermediate frequencies. Taken together, our data suggest that balancing selection, and/or directional selection on standing low-frequency alleles, have shaped *NAT2* genetic diversity and the human acetylation phenotype in Central Asian agriculturalists. These results further support the hypothesis that a major transition in human lifestyle, such as the emergence of farming has dramatically changed human chemical environments and the selective pressures they imposed.

Introduction

The dispersals of human populations from Africa into new environments in the last 75 000 years, and the introduction of farming starting ~10 000 years ago, involved profound changes in human lifestyle, including diet and exposure to pathogens and xenobiotics.[1] During the Palaeolithic period, the dietary choices of modern humans must have been limited to minimally processed wild plants and animal foods. With the advent of agriculture, novel foods were introduced, of which the human genome had little evolutionary experience. Thus, farming communities were challenged by new selective pressures that led to the emergence of

*Correspondence: Dr E Patin, Human Evolutionary Genetics, Metchnikoff Building, 5th floor, 25 rue du Dr Roux Paris, Institut Pasteur, Cedex 15, Paris 75724, France. Tel: +33 1 44 38 94 41; Fax: +33 1 45 68 86 39; E-mail: epatin@pasteur.fr
[6]Joint first authors.

genetically transmitted phenotypes increasing human survival. Besides digesting new energy-rich nutrients, humans also had to detoxify a wide range of novel xenobiotics, including toxins and carcinogens.[2] For example, changes in the temperature at which meat and fish are cooked modify human exposure to exogenous carcinogens (heterocyclic amines and polycyclic aromatic hydrocarbons), and therefore, increases the risk of developing colon cancer.[2] This example suggests that the genes involved in the detoxification of exogenous molecules, generally speaking, might have played an important role in human adaptation during the transition from foraging to farming.

Among the enzymes involved in xenobiotic detoxification, arylamine *N*-acetyltransferase 2 (NAT2) is a phase II drug-metabolizing enzyme (DME), which catalyses the *N* acetylation of aromatic amines.[3] This DME reached prominence initially as one of the first enzymes to be recognized as a cause of interindividual variation in drug metabolism. Functional polymorphisms at *NAT2* gene segregate in humans and in other mammals into rapid and slow acetylation phenotypes.[4] The acetylation status has been previously shown to modify the frequency and/or the severity of drug and xenobiotic toxicity in human populations.[5,6] For example, slow acetylators are at increased risk of hepatotoxicity to isoniazid, a major antitubercular drug, because the high and extended circulating concentration of this drug gives rise to toxic effects in humans.[7] Slow acetylators are also at increased risk for prostate and urinary bladder cancers following exposure to aromatic amine carcinogens.[8,9] Conversely, for example, fast acetylators are at increased risk for colon cancer (also lung, breast and laryngeal cancers) when individuals are highly exposed to heterocyclic amines found in well-cooked meat.[9] In this context, the involvement of acetylation phenotypes in the metabolism of several xenobiotics suggests that they do change the fitness of human populations. Consequently, the changes in human exposure to toxic molecules and carcinogens introduced by the transition to agriculture might have implied a number of modifications in the selective pressures acting on DMEs, as recently shown for the *NAT2* gene.[10]

To test in depth whether past demography and lifestyles have influenced the selective pressures acting on *NAT2*, we characterized genetic diversity and the patterns of linkage disequilibrium (LD) of the *NAT2* gene region in six different human populations from Central Asia. Populations inhabiting Central Asia exhibit marked differences in terms of lifestyle (sedentary agriculturalists *vs* nomad pastoralists), dietary choices and past demographic histories, and therefore represent an excellent model system to test the extent to which changes in human lifestyle might have influenced human adaptation to xenobiotic environments.

Materials and methods

DNA samples

In our study, we analysed a total of 138 unrelated individuals (276 chromosomes) from six populations of Central Asia, including: Kirghizs who were sampled in the south (lowland) (population named KRA, $2n = 94$) and in the north of Kirghizstan (highland) (KRM, $2n = 46$), Kazakhs from Kazakhstan (KZ, $2n = 52$), Tajiks from the Uzbek cities of Agalic (TJA, $2n = 64$) and Urgut (TJU, $2n = 58$) and Uzbeks from Uzbekistan (UZ, $2n = 62$) (see Figure 1). The Kirghizs sampled in this study are thought to have always been nomad pastoralists, because of their origins from southern Siberia where the Neolithic revolution appeared late and marginally.[11,12] Their diet is dominated by meat. It is widely accepted that the Uzbeks and the Kazakhs were nomads who became sedentarized in the 14–15th century and the 20th century, respectively.[13] They are now agriculturalists and have a diet based mainly on meat and dairy.[14] By contrast, the Tajiks are long-term agriculturalists and may represent the descendants of people who made the Neolithic transition in this area (~4000–3000 years ago[15]). They are sedentary agriculturalists rearing cattle. Their food diet is less dominated by meat consumption, contrary to the other populations cited above.[16]

Genotyping and sequencing

All the individuals were typed for a total of 15 polymorphic SNPs. These SNPs correspond to the nine intergenic SNPs

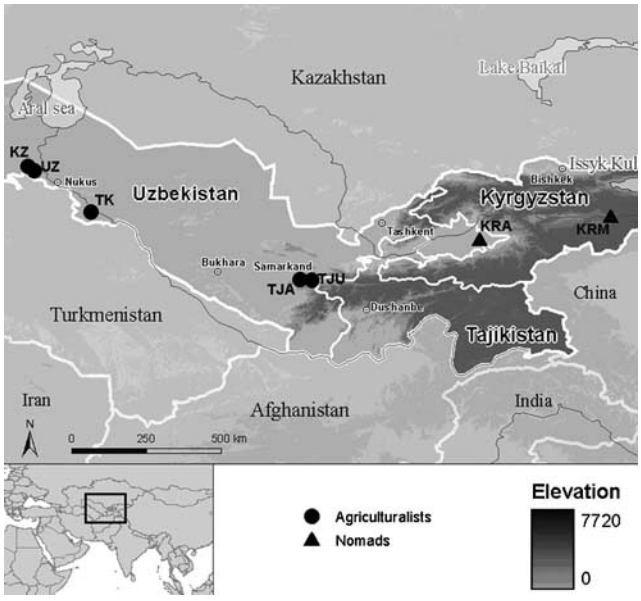

Figure 1 Localities of the seven Central Asian samples. Population names are labelled as follows: KRA, KRM: Kirghizs; KZ: Kazakhs; TJA, TJU: Tajiks; UZ: Uzbeks; TK: Turkmen.

selected in a previous study[10] and to the six SNPs found in *NAT2* coding region, except SNP 191G>A as this is monomorphic in non-African samples (282C>T, 341T>C, 481C>T, 590G>A, 803A>G, 857G>A; for more details, see Patin *et al*[10]). The nine intergenic SNPs, covering a region of 200 kb, were genotyped by *Taq*Man (ABI Prism® 7000 Sequence Detection System, Applied Biosystems, Foster City, CA, USA) assays. As to the six *NAT2* SNPs, a region of 1500 bp including *NAT2* coding exon (870 bp) and its flanking regions were PCR amplified and sequenced with the 'ABI Prism® 3100 Genetic Analyser' (Applied Biosystems). Protocols were reported in a previous study.[10] Sequences were analysed using GENALYS.[17]

Data analysis

Allele frequencies were calculated by gene counting. Using Arlequin v.2.000,[18] F_{ST} statistics[19] from all pairwise population comparisons were calculated and deviations from the Hardy–Weinberg equilibrium were tested. Haplotype reconstruction from unphased genotype was performed using the Bayesian method implemented in PHASE v.2.1.1.[20] The phenotype (fast/slow acetylator) proportions were calculated for each population considering that the 'slow acetylator' allele is codominant.

Using DnaSP v.4.10,[21] we calculated the nucleotide diversity π and Watterson's estimator θ_W of θ and we performed the neutrality tests Tajima's D_T,[22] Fu and Li's F* and D^*.[23] These statistics are functions of the number of segregating sites in a sample of sequences, the mean pairwise difference between the sequences and the number of derived variants that are observed only once in the sample, all of which are affected by natural selection.[22,23] *P*-values were estimated by the use of coalescent simulations under a neutral model of evolution ($n = 10\,000$).

Pairwise LD between the 15 genotyped SNPs was estimated in each population. We calculated the statistics D'[24] and r^2 Hill and Robertson[25] using DnaSP v.4.10.[26] Their statistical significance was tested using a Fisher's exact test followed by Bonferroni corrections for multiple testing.[27] To represent the decay of LD over distance, we plotted the proportion of SNP pairs showing significant Fisher's exact tests for D' against the physical distance between SNP pairs.

We also performed the long-range haplotype (LRH) test using SWEEP v.1.0[28] on a core region corresponding to *NAT2* coding exon.[10] Then we assessed for each core haplotype its relative extended haplotype homozygosity (REHH) 200 kb apart. To test the significance of potentially selected core haplotypes, our data set was compared with both empirical (obtained from the screening of the entire chromosome 8 in European-descent populations; HapMap database) and simulated distributions[10] of 'core haplotype frequencies *vs* REHH'.

We inferred the population growth rate *r* and the age *g* of *NAT2* non-synonymous mutations by the use of a joint maximum-likelihood estimation of these parameters, as described in a previous study.[29] Since demographic events leave the same signature on the whole genome, we also estimated the growth rate and age of all the nine intergenic SNPs and considered them as references to discern the respective influences of population expansion and selection on *NAT2* non-synonymous mutations. The recombination parameter required for these analyses was estimated by comparing deCODE and Marshfield genetic and physical distances in the *NAT2* region (UCSC Genome Bioinformatics). We performed our estimations with the effective population size N_e fixed to 20 000 individuals. We also used the data set from Patin *et al*[10] to assess the mutations' ages in populations from Western Eurasia (Ashkenazi Jews, Sardinians, Swedes, Saami populations) and Eastern Eurasia (Gujarati from India, Chinese and Thai). We also added the agriculturalist Turkmen (TK) population from Uzbekistan.[10]

Results

Sequence variation at the *NAT2* gene region

To evaluate the extent to which natural selection has shaped the patterns of variability of the *NAT2* gene in Central Asian populations, we re-sequenced *NAT2* in six populations presenting different lifestyles (see Material and methods). We identified only six SNPs in the *NAT2* coding region (282C>T, 341T>C, 481C>T, 590G>A, 803A>G and 857G>A), which have all been previously reported.[8,10] All populations were at the Hardy–Weinberg equilibrium at all SNPs. Haplotype reconstruction and frequency estimation were performed in each population separately (Table 1). As to haplotypes encoding an altered NAT2 protein (noted 'slow' in Table 1), the Tajiks and the Kazakhs exhibited the highest frequency of haplotype *NAT2*5B* (TJA: 26%; TJU: 22% and KZ: 23%), which carries the non-synonymous mutation 341T>C encoding the NAT2 enzyme with the most altered activity as compared to the other NAT2 slow proteins.[8,9] The *NAT2*6A* haplotype, which carries the non-synonymous mutation 590G>A, was present at high frequencies, especially in Tajik populations (TJA: 39% and TJU: 45%), while it was found at the lowest frequency in Kazakhs (KZ: 13%). Finally, the Kazakhs exhibited the highest frequency (23%) of *NAT2*7B*, which carries the non-synonymous slow-acetylation mutation 857G>A, which is mainly restricted to East Eurasian populations.[10] As to 'fast' haplotypes, the haplotype *NAT2*4*, which is the ancestral state of the gene and is defined as the reference *NAT2* haplotype,[30] was found at high frequency in the Kirghizs (KRA: 48% and KRM: 46%) and in the Uzbeks (UZ: 42%), at nearly two-fold lower frequencies in the Tajiks (TJA: 23% and TJU: 26%) and at intermediate frequency in the Kazakhs (KZ: 38%).

Table 1 Allelic composition and frequency of *NAT2* haplotypes in Central Asia

		Mutations						Populations							
Haplotypes	*Acetylation status*	*282*	*341*	*481*	*590*	*803*	*857*	*KRA*	*KRM*	*KZ*	*TJA*	*TJU*	*UZ*	*TK*[a]	*Total*
*NAT2*4*	Fast	.	.	.	.	.	.	0.48	0.46	0.38	0.23	0.26	0.42	0.31	0.38
*NAT2*12A*	Slow	.	.	.	.	G	.							0.01	0.002
*NAT2*5A*	Slow	.	C	T	.	.	.				0.01		0.01		0.005
*NAT2*5B*	Slow	.	C	T	.	G	.	0.12	0.15	0.23	0.26	0.22	0.14	0.23	0.18
*NAT2*5C*	Slow	.	C	.	.	G	.		0.02	0.02	0.06	0.02	0.01	0.03	0.02
*NAT2*6A*	Slow	T	.	.	A	.	.	0.26	0.24	0.13	0.39	0.45	0.31	0.30	0.30
*NAT2*7B*	Slow	T	.	.	.	.	A	0.15	0.13	0.23	0.03	0.05	0.10	0.12	0.11
2N								94	46	52	64	58	62	100	376

Notes: dots represent the ancestral state of each mutation, deduced from the chimpanzee *NAT2* sequence.
[a]Turkmen population, data from Patin *et al.*[10]

Table 2 Acetylation phenotype frequencies of Central Asian populations

	Populations						
Phenotypes	*KRA*	*KRM*	*KZ*	*TJA*	*TJU*	*UZ*	*TK*[a]
Fast	0.21	0.22	0.11	0.09	0.07	0.16	0.08
Intermediate	0.53	0.48	0.54	0.28	0.38	0.52	0.48
Slow	0.26	0.30	0.35	0.63	0.55	0.32	0.44

Notes: heterozygous individuals bearing a fast and a slow haplotype are considered intermediate acetylators.
[a]From Patin *et al.*[10]

Acetylation phenotype inference

We inferred from *NAT2* genotypes the distribution of fast/slow acetylation phenotypes across populations (Table 2). Heterozygotes bearing a 'fast' and a 'slow' haplotype were considered 'intermediate acetylators', because they present a mean intermediate NAT2 activity significantly different from that of fast homozygotes.[31] Phenotype frequencies showed strong variation among the different studied populations. Indeed, the Tajiks exhibited significantly higher proportions of slow acetylators (TJA: 63% and TJU: 55%; χ^2 test, $P<0.05$) as compared to the Uzbeks, the Kirghizs and the Kazakhs, who presented proportions of slow acetylators ranging from 26% (KRA) to 35% (KZ).

Population differentiation

We analysed the pattern of population differentiation by using the F_{ST} statistics, estimated from haplotype frequencies (Table 3a). When considering the *NAT2* coding region, the tests of exact differentiation for all pairs of populations first indicated that both Tajik populations (TJA and TJU) were significantly differentiated from the Kirghiz and Kazakh populations (KRA, KRM and KZ) ($F_{ST}=0.046$–0.078; $P<0.05$). Conversely, Tajiks were not statistically different from Uzbek ($F_{ST}=0.027$; $P=0.058$) and Turkmen populations. In addition, the two Tajik populations were not differentiated between each other. Finally, the south Kirghiz population (KRA) was differentiated from Turkmens. When all the SNPs were considered in the analysis (Table 3b), including the genotyped intergenic SNPs, the patterns of differentiation became slightly different: the TJA and TJU populations were no longer differentiated from the KZ; the TJA population was no longer differentiated from the KRM population; significant F_{ST} values were weaker between the Tajiks and the other populations ($F_{ST}=0.021$–0.038), while the Turkmen population (TK) remained differentiated from KRA population ($F_{ST}=0.011$; $P=0.002$). In both cases, the UZ population was undifferentiated from all the other populations.

Statistical tests of neutrality

Estimates of diversity (π and θ_W) at the *NAT2* gene for all populations are shown in Table 4. To test whether patterns of DNA sequence variation at *NAT2* fit expectations under the neutral model of evolution, we analysed the sequences by use of D_T, D^* and F^* (Table 4). For all populations, D_T was positive and ranged from 1.074 (KRM) to 2.08 (TJA). It reached significance in the Kazakhs ($D_T=1.657$, $P=0.044$) and in both Tajik populations TJU and TJA ($D_T=1.745$ and 2.08; $P=0.044$ and 0.02, respectively). D^* ranged from 1.125 (KRA) to 1.181 (KRM) and no value was significant. In contrast, F^* exhibited the same trend as D_T: all values were positive, and ranged from 1.327 (KRA) to 1.709 (TJA) and reached significance for the same three populations KZ, TJU and TJA ($F^*=1.554$, 1.584 and 1.709; $P=0.039$, 0.035 and 0.028, respectively). The significance of these three statistics was assessed without any assumption about populations' demography (ie neither population growth neither bottleneck events were taken into account).

Long-range haplotype test for recent directional selection

We next investigated whether recent directional selection has influenced the *NAT2* region by the use of the LRH test. This test is made to detect recent directional selection by

Table 3 Pairwise estimates of F_{ST} between Central Asian populations using (a) the nine intergenic SNPs (below diagonal) and the six *NAT2* SNPs (above diagonal) and (b) all 15 genotyped SNPs (below diagonal) and phenotype frequencies (above diagonal)

	KRA	*KRM*	*KZ*	*TJA*	*TJU*	*UZ*	*TK*
(a)							
KRA	—	−0.014	0.016	**0.067*** **	**0.061** **	−0.006	**0.021* **
KRM	0.005	—	0.002	**0.046* **	**0.046* **	−0.014	0.005
KZ	−0.001	−0.003	—	**0.067** **	**0.078*** **	0.021	0.015
TJA	**0.011* **	**0.008**	**0.008**	—	−0.011	0.027	0.004
TJU	**0.010* **	**0.007**	−0.0004	**0.005**	—	0.022	0.006
UZ	0.003	−0.002	−0.0006	0.002	**0.005**	—	0.001
TK	**0.010** **	−0.004	−0.001	**0.005**	**0.008* **	−0.0004	—
(b)							
KRA	—	−0.015	0.003	**0.107*** **	**0.083** **	−0.006	**0.041* **
KRM	0.004	—	−0.010	0.089*	0.064*	−0.016	0.023
KZ	−0.001	−0.003	—	0.035	0.018	−0.015	−0.005
TJA	**0.011** **	**0.006**	0.007	—	−0.015	0.060*	0.005
TJU	**0.016** **	**0.012* **	**0.008**	**0.005**	—	0.040	−0.005
UZ	0.005	−0.003	0.001	0.003	0.005	—	0.008
TK	0.011**	−0.004	0.001	0.004	**0.009* **	0.001	—

Notes: data in bold indicate that Fisher's exact test of differentiation is significant at the 0.05 level. Asterisks are reported for F_{ST} values significantly different from zero (*$P<0.05$; **$P<0.01$; ***$P<0.001$).

Table 4 Summary statistics and tests of neutrality based on the sequenced *NAT2* coding and flanking regions in Central Asian populations

	π (10^{-3})	θ_W	D_T	D*	*F**
KRA	1.44	1.173	1.124	1.125	1.327
KRM	1.59	1.365	1.074	1.181	1.344
KZ	1.81	1.327	1.657*	1.172	1.554*
TJA	1.94	1.268	2.08*	1.155	1.709*
TJU	1.82	1.296	1.745*	1.163	1.584*
UZ	1.60	1.277	1.276	1.158	1.405
TK	1.87	1.158	2.255*	1.120	1.732*

Notes: *$P=0.05$ probability level.

comparing two allele age estimates: the allele frequency and the breakdown of LD around the allele.[28] An allele under recent directional selection will be typically too frequent given its age. *P*-values were estimated for all core haplotypes in all populations from the simulated distribution and from empirical data. None of the haplotypes in any of the sampled populations appeared to deviate from simulated neutral and empirical genomic expectations.

Growth rate and mutation age

LD breakdown at the surrounding sites of a mutation is very informative for inferring allele age estimates by considering the recombination rate as a 'genetic clock'. Taking into account the results of neutrality tests indicating a possible effect of natural selection, we investigated growth rates and ages of both functional *NAT2* mutations and intergenic mutations as a reference (Table 5). Our estimations showed that the non-functional mutations exhibited a mean growth rate of 0.012 (CI = 0.008; 0.020). Conversely, we observed that all functional mutations did not experience the same history among the different populations. The mutation 341T > C, together with mutations 481C > T and 803A > G, which are strongly associated with 341T > C (Table 1), showed a high growth rate in KRM and TK populations (mean $r=0.029$, CI = 0.018; 0.049), while both Tajik populations TJA and TJU presented a weak growth rate (mean $r=0.0115$, CI = 0.006; 0.021) in the range of growth rate estimates for intergenic mutations. The remaining KRA, KZ and UZ populations showed intermediate growth rates (with overlapped confidence intervals). The non-synonymous mutation, 590G > A, exhibited a high growth rate in the KRA, KRM and KZ (mean rate = 0.023), while this value was weaker (in the same order as intergenic mutations) in the four remaining populations TJA, TJU, TK and UZ (mean rate = 0.013). The mutation 857G > A exhibited a different pattern with no clear tendency, KZ showing the highest growth rate ($r=0.021$) and KRA the weakest ($r=0.012$). No growth rate estimate was obtained for 857G > A in the Tajik populations because the mutation frequency in these populations was too low.

Linkage disequilibrium levels

We determined pairwise LD between SNPs using D' and r^2, which are useful for modelling recombination rates and association power, respectively.[32] LD generally decreased with physical distance, although large variation in the LD–distance relationship was observed between populations (Figure 2). The Tajik population, TJA, exhibited a more

Table 5 Mutation frequency *P*, growth rate *r* and mutation age *g* with 95% confidence intervals for all *NAT2* functional mutations in Central Asian populations

Population	*KRA*	*KRM*	*KZ*	*TJA*	*TJU*	*UZ*	*TK*[a]	*CHI*[a]	*GUJ*[a]	*THAI*[a]
Intergenic mutations										
Mean *r*	0.013 (0.008; 0.021)	0.013 (0.008; 0.022)	0.015 (0.009; 0.024)	0.010 (0.006; 0.017)	0.011 (0.007; 0.019)	0.011 (0.007; 0.018)	0.012 (0.008; 0.020)	0.010 (0.007; 0.017)	0.011 (0.006; 0.017)	0.010 (0.006; 0.017)
Mean age (years)	11 596 (9425; 14 662)	11 635 (9380; 16 112)	11 317 (10 662; 17 905)	13 886 (12 537; 19 840)	13 300 (11 455; 18 462)	13 354 (11 740; 19 405)	12 533 (10 060; 14 922)	13 660 (10 847; 17 573)	13 851 (11 233; 17 358)	14 353 (11 495; 18 270)
Mutation 341T>C										
P	0.12	0.17	0.25	0.34	0.24	0.18	0.26	0.07	0.34	0.11
r	0.015 (0.008; 0.028)	0.029 (0.018; 0.049)	0.019 (0.012; 0.032)	0.012 (0.007; 0.021)	0.011 (0.006; 0.020)	0.013 (0.007; 0.023)	0.029 (0.019; 0.047)	0.138 (0.006; 0.028)	0.009 (0.005; 0.016)	0.009 (0.004; 0.018)
Age (years)	7775 (5495; 11 480)	4962 (3620; 7337)	7332 (5452; 10 382)	11 280 (8585; 15 295)	11 327 (8262; 16 187)	9562 (6835; 14 012)	5292 (4107; 7075)	7497 (4925; 12 075)	14 440 (11 177; 19 065)	11 502 (7845; 17 540)
Mutation 590G>A										
P	0.26	0.24	0.13	0.39	0.45	0.31	0.30	0.25	0.44	0.39
r	0.022 (0.014; 0.037)	0.024 (0.015; 0.041)	0.022 (0.013; 0.039)	0.013 (0.008; 0.021)	0.013 (0.008; 0.021)	0.013 (0.008; 0.023)	0.013 (0.008; 0.022)	0.010 (0.005; 0.017)	0.014 (0.009; 0.023)	0.013 (0.009; 0.023)
Age (years)	6490 (4997; 8717)	5977 (4445; 8565)	5917 (4187; 8977)	10 967 (8457; 14 647)	11 217 (8697; 14 905)	10 217 (7712; 14 027)	10 432 (8095; 13 737)	12 840 (9602; 17 655)	10 327 (8287; 13 082)	10 470 (8240; 13 585)
Mutation 857G>A										
P	0.15	0.13	0.23	0.03	0.05	0.10	0.12	0.14	0.06	0.19
r	0.012 (0.006; 0.023)	0.017 (0.009; 0.031)	0.021 (0.013; 0.035)	NA	NA	0.015 (0.008; 0.029)	0.018 (0.010; 0.032)	0.010 (0.005; 0.019)	0.013 (0.005; 0.027)	0.013 (0.007; 0.020)
Age (years)	9610 (6915; 13 820)	7285 (5037; 11 287)	6705 (4967; 9570)	NA	NA	7377 (4982; 11 635)	6902 (4942; 10 075)	10 922 (7755; 15 942)	7542 (4905; 12 247)	9515 (7025; 13 315)

Abbreviations: CHI, Chinese; GUJ, Gujarati; THAI, Thai; TK, Turkmen.
Notes: the intergenic mutations are considered as a reference.
NA stands for not applicable.
[a]From Patin *et al.*[10]

important decrease in LD compared to the other populations. Conversely, the population KZ presented the slowest decrease in LD with distance.

Discussion

The diversity survey of the *NAT2* gene in Central Asians revealed a clear contrast between the Tajik populations as compared to the remaining groups. Both Tajik groups exhibited a frequency of slow acetylators (0.55 and 0.63) twice higher than in the other populations (KZ, UZ, KRA and KRM: 0.26–0.35). Accordingly, the frequencies of the fast haplotype *NAT2*4* (0.23 and 0.26) were approximately half of *NAT2*4* frequency in the remaining four populations (0.38–0.48). Interestingly, the lifestyle of both Tajik populations strongly differs from that of the remaining populations. Indeed, Tajiks shifted to a sedentary lifestyle the earliest and were the first to practice agriculture in Central Asia.[15] Accordingly, this early transition might have implied differences in the dietary conditions of Tajiks, as compared to nomadic populations. The harsh climatic conditions of Central Asia might have discouraged the spread of agriculture but it is likely that some fertile spots, like the Fergana Valley (between Kirghizstan and Uzbekistan) and other areas where water was available, could have sustained early farming communities, such as the Tajiks. A similar dichotomous pattern at the *NAT2* gene has been observed among sub-Saharan Africans, where the Bantu-speaking agriculturalists presented a higher frequency of slow acetylators, as compared to the Pygmy hunter-gatherers (0.46 and 0.10, respectively).[33] These significant differences in the frequency distribution of slow/fast acetylation phenotypes depending on lifestyle, attested by significant F_{ST} estimates, strongly suggest that being slow acetylator has been an advantage in long-term agriculturalist populations in Central Asia.

The results of sequence-based neutrality tests give further support to this observation. Indeed, the agriculturalist Tajik populations exhibited significantly positive Tajima's D_T and Fu and Li's F* values, whereas *NAT2* variation in nomadic populations was compatible with neutrality. The significantly positive values observed for these neutrality tests in the two Tajik populations (TJA and TJU), as well as in the Kazakh population (KZ), are due to an excess of haplotypes at intermediate frequency. This observation can be explained by three different, non-mutually exclusive hypotheses: (1) demographic processes,[34] (2) balancing selection or (3) directional selection on standing variation.[35] In the context of a demographic explanation, population subdivision may lead to an excess of intermediate-frequency haplotypes[36] by wrongly considering two isolated populations as a unique virtual population. Recent bottlenecks can also generate the same diversity patterns, because they involve the loss of low-frequency haplotypes.[37] Because both population subdivision and

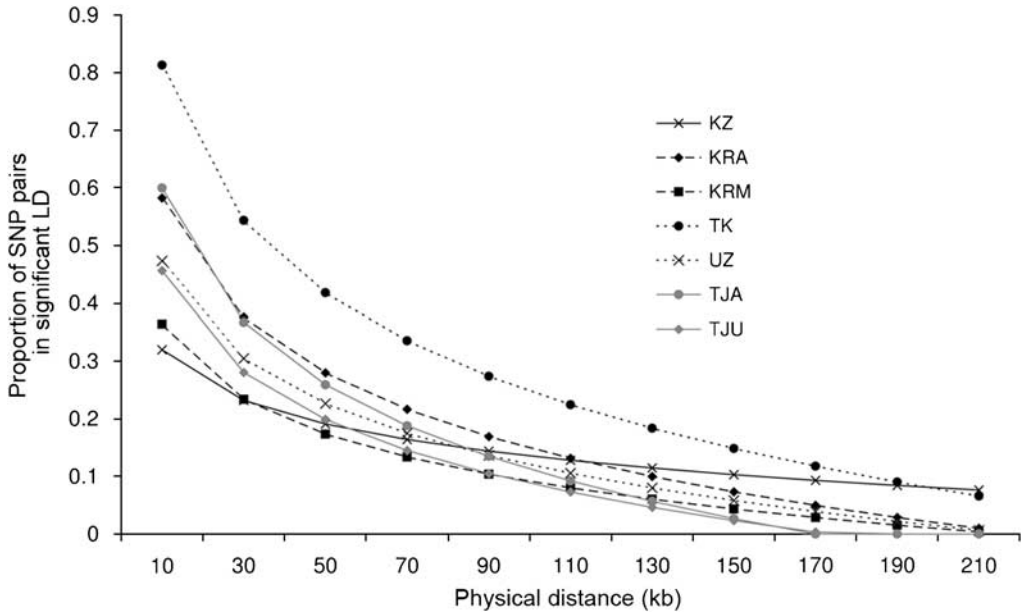

Figure 2 Levels of linkage disequilibrium (as measured by the significance of *D′*) against physical distance at *NAT2* region in Central Asian populations.

bottlenecks tend to increase allelic association over distance, we would expect LD levels in agriculturalist populations to be higher than in other populations.[38] Actually, we observed the inverse pattern in the *NAT2* region, the TJA and TJU populations showing the strongest decreases of LD with distance (Figure 2). In addition, if the patterns of genetic diversity at *NAT2* are the sole result of demography and genetic drift, its diversity would be comparable to that inferred from neutral markers. Mitochondrial DNA and Y-chromosome uniparentally inherited markers show no signs of population subdivision or reduction in Tajik populations (E Heyer, unpublished data). In this view, the hypothesis that demographic forces yielded to the significantly positive Tajima's D_T and Fu and Li's F* values at *NAT2* seems unlikely.

The second hypothesis invokes the action of balancing selection, corresponding to the heterozygote advantage. This selective regime can maintain different alleles at intermediate frequencies over time, yielding to an excess of intermediate-frequency alleles. The significantly positive values of D_T observed at *NAT2* gene support the action of balancing selection favouring heterozygous individuals having both a 'fast' and a 'slow' *NAT2* haplotype. Indeed, we observed a frequency distribution of *NAT2* slow and fast haplotypes in Kazakhs (0.38 *vs* 0.62) and Tajiks (0.25 *vs* 0.75), which could be consistent with balancing selection. However, balancing selection favours heterozygotes only if the heterozygote status represents a distinct phenotype compared to homozygotes. Depending on the substrate used to test NAT2 acetylation activity, fast/slow heterozygotes have, or have not, a different phenotype than both fast and slow homozygotes. In fact, when phenotypes are determined with caffeine as a test substrate, a distinct dose effect could be observed: fast/slow heterozygous individuals present an intermediate phenotype between the phenotypes of fast and slow homozygotes.[31] Given that this intermediate acetylator status is observed in humans, we can easily imagine why being fast/slow heterozygotes would be advantageous: fast acetylators exposed to xenobiotics becoming harmful when acetylated would be disadvantageous, whereas slow acetylators exposed to xenobiotics becoming harmless when acetylated would, in turn, be disadvantageous. Because the human chemical environment is probably made of the two types of xenobiotic molecules, the best strategy to avoid the harmful effects of numerous toxic products would be to have an intermediate acetylator status. It is also interesting to note that the fast/fast homozygotes are more susceptible to colon cancer whereas the slow/slow homozygotes are more susceptible to bladder cancer. In this view, being fast/slow heterozygous could be an advantage as compared to the two individual homozygous states.

Alternatively, the observed patterns of genetic diversity at the *NAT2* gene can result from the action of directional selection acting on standing low-frequency alleles. Indeed, human adaptation to new environments might have involved more likely alleles already being present at appreciable frequency, resulting eventually in an excess of intermediate-frequency alleles. For example, directional selection exerting on standing variation with $P = 0.05$, the frequency at which a given allele starts to be favoured, can lead to positive values of D_T.[35] In our case, the allele

carrying the slow mutation 341T>C might have been selectively neutral and present at low frequency in Central Asia prior to the emergence of agriculture. Subsequently, it might have been selected with ensuing environmental changes, such as the appearance of new toxic or antibiotic molecules linked with agricultural practices (for instance, cereals storage[34,39]). However, the LRH test, which is conceived to directly detect the action of recent directional selection, was not significant for any of the *NAT2* haplotypes, in contrast with the previous results observed in the Turkmen population.[10] Several reasons can explain why this test failed to detect any signature of directional selection. First, the selective advantage conferred by *NAT2* mutations might have been too weak to be detected by the LRH test. This hypothesis is likely if we assume a global advantage of being a slow acetylator. In this case, each of the three mutations altering the NAT2 activity (341T>C, 590G>A and 857G>A) would have been advantageous, which would globally weaken the individual signature of selection at each altering mutation. Second, demographic events can affect the extent of LD around a given allele. The LRH test corrects for the spurious effects of demography by dividing the extent of LD around the tested core haplotype by the extent of LD around the remaining core haplotypes, which are all equally influenced by demography and expected to evolve under neutrality. However, this assumption can be misleading in the *NAT2* case, because a global advantage of being a slow acetylator would have favoured several *NAT2* core haplotypes bearing an altering mutation, a situation that would short circuit the correction of the test for demography.

Several other aspects of *NAT2* diversity are consistent with a scenario where several mutations are simultaneously targeted by directional selection. For example, growth rates and F_{ST} values at sites under directional selection are expected to be higher than at neutral sites. We observed such trends at NAT2-altering mutations (Tables 3a and 5), even if they did not reach statistical significance. In addition, a selective advantage of being a slow acetylator would make altering mutations increase slowly in frequency, which could also explain the significant excess of intermediate-frequency alleles detected by Tajima's D_T and Fu and Li's F*. This scenario is further supported by the fact that the populations that show the highest and most significant values of Tajima's D_T (TJA, TJU, KZ and TK) correspond to the populations with the highest proportions of slow haplotypes (correlation coefficient $r^2 = 0.95$, $P = 0.0038$). Theoretical predictions are nevertheless required to conclude if the genetic diversity at *NAT2* is best explained by this particular case of 'multiallelic' directional selection.

In conclusion, the patterns of genetic diversity at the *NAT2* gene in six Central Asian populations revealed striking differences between long-term agriculturalists and nomadic populations, in terms of both genetic variability and acetylation phenotypes. As previously reported in other geographic regions,[10] these observations further support the observation that the acetylation status of agriculturalist populations has been the target of natural selection, in the form of balancing or directional selection. The *NAT2* example, together with an increasing number of genes showing signs of recent directional selection, suggests that the selective pressures acting on our species have frequently changed due to rapid evolution of human-driven environments. In this context, detailed evolutionary studies of populations with well-defined lifestyles will provide valuable clues to better understand the consequences that major human cultural transitions had on the population genome diversity, both neutral and disease-related.

Acknowledgements

We thank all the participants in the study. This work was supported by the CNRS grant project ATIP, the CNRS grant project OHLL and the European Science Foundation project OMLL. We also thank Institut Français d'Etudes sur l'Asie Centrale (IFEAC) for some technical support.

References

1 Cordain L, Eaton SB, Sebastian A *et al*: Origins and evolution of the Western diet: health implications for the 21st century. *American J Clin Nutr* 2005; **81**: 341–354.
2 Ferguson LR: Natural and human-made mutagens and carcinogens in the human diet. *Toxicology* 2002; **181–182**: 79–82.
3 Hein DW, Doll MA, Rustan TD *et al*: Metabolic activation and deactivation of arylamine carcinogens by recombinant human NAT1 and polymorphic NAT2 acetyltransferases. *Carcinogenesis* 1993; **14**: 1633–1638.
4 Hein DW, Doll MA, Fretland AJ *et al*: Rodent models of the human acetylation polymorphism: comparisons of recombinant acetyltransferases. *Mutat Res* 1997; **376**: 101–106.
5 Jiang W, Feng Y, Hein DW: Higher DNA adduct levels in urinary bladder and prostate of slow acetylator inbred rats administered 3,2′-dimethyl-4-aminobiphenyl. *Toxicol Appl Pharmacol* 1999; **156**: 187–194.
6 Feng Y, Fretland AJ, Rustan TD, Jiang W, Becker WK, Hein DW: Higher frequency of abberant crypt foci in rapid than slow acetylator inbred rats administered the colon carcinogen 3,2′-dimethyl-4-aminobiphenyl. *Toxicol Appl Pharmacol* 1997; **147**: 56–62.
7 Huang Y-S, Chern H-D, Su W-J *et al*: *N*-acetyltransferase 2 gene as a susceptibility risk factor for antituberculosis drug-induced hepatitis. *Hepatology* 2002; **35**: 883–889.
8 Hein DW, Doll MA, Fretland AJ *et al*: Molecular genetics and epidemiology of the *NAT1* and *NAT2* acetylation polymorphisms. *Cancer Epidemiol Biomark Prev* 2000; **9**: 29–42.
9 Hein DW: Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. *Mutat Res* 2002; **506–507**: 65–77.
10 Patin E, Barreiro LB, Sabeti PC *et al*: Deciphering the ancient and complex evolutionary history of human arylamine *N*-acetyltransferase genes. *Am J Hum Genet* 2006; **78**: 423–436.
11 Oshibkina SV: *Neolit Severnoj Evrazii*. Moscow: Nauka (eds) 1996.
12 Troickaya TN, Novikov AV: *Arkheologiya Zapadno-Sibirskoy ravniny*. Novossibirsk, Russia, 2004.
13 Sellier A, Sellier J: *Atlas Des Peuples d'Orient*, Paris: La Découverte (eds) 2004.

14 Mishra V, Arnold F, Semenov G, Hong R, Mukuria A: Epidemiology of obesity and hypertension and related risk factors in Uzbekistan. *Eur J Clin Nutr* 2006; **60**: 1355–1366.
15 Brunet F: La néolithisation en Asie centrale: un état de la question. *Paléorient* 1998; **24**: 27–48.
16 FAO: *FAO Statistical Yearbook 2004, Vol. 1/1* 2004.
17 Takahashi M, Matsuda F, Margetic N, Lathrop M: Automated identification of single nucleotide polymorphisms from sequencing data. *J Bioinform Comput Biol* 2003; **1**: 253–265.
18 Schneider S, Roessli D, Excoffier L: *Arlequin ver. 2.000: A Software for Population Genetics Data Analysis*. Switzerland: Genetics and Biometry Laboratory, University of Geneva, 2000.
19 Weir B, Cockerham C: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
20 Stephens M, Donelly P: A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**: 1162–1169.
21 Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003; **19**: 2496–2497.
22 Tajima F: Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
23 Fu YX, Li WH: Statistical tests of neutrality of mutations. *Genetics* 1993; **133**: 693–709.
24 Lewontin RC: The interaction of selection and linkage. II. Optimum models. *Genetics* 1964; **50**: 757–782.
25 Hill WG, Robertson A: The effects of inbreeding at loci with heterozygote advantage. *Genetics* 1968; **60**: 615–628.
26 Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003; **19**: 2496–2497.
27 Rice WR: Analyzing tables of statistical tests. *Evolution* 1989; **43**: 223–225.
28 Sabeti PC, Reich DE, Higgins JM, *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
29 Austerlitz F, Kalaydlieva L, Heyer E: Detecting population growth, selection and inherited fertility from haplotypic data in humans. *Genetics* 2003; **165**: 1579–1586.
30 Hein DW, Grant DJ, Sim E: *N*-acetyltransferase gene nomenclature. *Pharmacogenetics* 2000; **10**: 291–292.
31 Cascorbi I, Drakoulis N, Brockmoller J, Maurer A, Sperling K, Roots I: Arylamine *N*-acetyltransferase (NAT2) mutations and their allelic linkage in unrelated Caucasian individuals: correlation with phenotypic activity. *Am J Hum Genet* 1995; **57**: 581–592.
32 Delvin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.
33 Patin E, Harmant C, Kidd KK *et al*: Sub-Saharan African coding sequence variation and haplotype diversity at the *NAT2* gene. *Hum Mutat* 2006; **27**: 720.
34 Akey JM, Eberle MA, Rieder MJ *et al*: Population history and natural selection shape patterns of genetic variation in 132 genes. *Plos Biology* 2004; **2**: 1591–1599.
35 Przeworski M, Coop G, Wall JD: The signature of positive selection on standing genetic variation. *Evolution* 2005; **59**: 2312–2323.
36 Tajima F: The effect of change in population size on DNA polymorphism. *Genetics* 1989; **123**: 597–601.
37 Laporte V, Charlesworth B: Effective population size and population subdivision in demographically structured populations. *Genetics* 2002; **162**: 501–519.
38 Ardlie KG, Kruglyak L, Seielstad M: Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2002; **3**: 299–309.
39 Bassett EJ, Keith MS, Armelagos GJ, Martin DL, Villanueva AR: Tetracycline-labeled human bone from ancient Sudanese Nubia (A.D. 350). *Science* 1980; **209**: 1532–1534.