BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Variation in NAT2 acetylation phenotypes is associated with differences in food-producing subsistence modes and ecoregions in Africa

Eliška Podgorná[1,2], Issa Diallo[3], Christelle Vangenot[2], Alicia Sanchez-Mazas[2], Audrey Sabbagh[4], Viktor Černý[1] and Estella S. Poloni[2*]

Abstract

Background: Dietary changes associated to shifts in subsistence strategies during human evolution may have induced new selective pressures on phenotypes, as currently held for lactase persistence. Similar hypotheses exist for arylamine N-acetyltransferase 2 (NAT2) mediated acetylation capacity, a well-known pharmacogenetic trait with wide inter-individual variation explained by polymorphisms in the *NAT2* gene. The environmental causative factor (if any) driving its evolution is as yet unknown, but significant differences in prevalence of acetylation phenotypes are found between hunter-gatherer and food-producing populations, both in sub-Saharan Africa and worldwide, and between agriculturalists and pastoralists in Central Asia. These two subsistence strategies also prevail among sympatric populations of the African Sahel, but knowledge on *NAT2* variation among African pastoral nomads was up to now very scarce. Here we addressed the hypothesis of different selective pressures associated to the agriculturalist or pastoralist lifestyles having acted on the evolution of *NAT2* by sequencing the gene in 287 individuals from five pastoralist and one agriculturalist Sahelian populations.

Results: We show that the significant *NAT2* genetic structure of African populations is mainly due to frequency differences of three major haplotypes, two of which are categorized as decreased function alleles (*NAT2*5B* and *NAT2*6A*), particularly common in populations living in arid environments, and one fast allele (*NAT2*12A*), more frequently detected in populations living in tropical humid environments. This genetic structure does associate more strongly with a classification of populations according to ecoregions than to subsistence strategies, mainly because most Sahelian and East African populations display little to no genetic differentiation between them, although both regions hold nomadic or semi-nomadic pastoralist and sedentary agriculturalist communities. Furthermore, we found significantly higher predicted proportions of slow acetylators in pastoralists than in agriculturalists, but also among food-producing populations living in the Sahelian and dry savanna zones than in those living in humid environments, irrespective of their mode of subsistence.

Conclusion: Our results suggest a possible independent influence of both the dietary habits associated with subsistence modes and the chemical environment associated with climatic zones and biomes on the evolution of *NAT2* diversity in sub-Saharan African populations.

Keywords: NAT2, Acetylation polymorphism, African Sahel, Pastoral nomads, Subsistence mode, Ecoregion, Natural selection

* Correspondence: estella.poloni@unige.ch
[2]Department of Genetics and Evolution, Anthropology Unit, Laboratory of Anthropology, Genetics and Peopling History, University of Geneva, 12 Rue Gustave-Revilliod, 1211 Geneva 4, Switzerland
Full list of author information is available at the end of the article

Background

All along human evolution and continuing in present times, populations have developed a wide variety of cultural innovations. Many of these may have affected the interactions between humans and their environment in such a way as to leave gene-culture coevolution imprints in the genome, due to both demographic and selective processes [1–5]. Among these innovations, new subsistence strategies represent major shifts with notable consequences on the pathogenic and dietary environments in which populations have been living, thus possibly also inducing new selective pressures on phenotypes [6–9]. Subsistence strategies are broadly classified in two opposite categories, i.e., foragers (or hunter-gatherers), relying on the collection of nutriments naturally occurring in the wild, and food-producers whose appearance in the prehistoric record is generally associated with the emergence of the Neolithic [10–12]. Among the latter, two major modes of production are distinguished, namely agriculture based on the cultivation of domesticated plants, and pastoralism based on the herding of domesticated animals. These two modes of subsistence also imply differential usage of space, in that most agricultural societies developed sedentarism whereas communities relying on animal husbandry developed nomadism or transhumance to exploit seasonal variations in the location of pasture areas.

It is currently held that these differential developments in subsistence strategies also influenced the selective regimes to which past populations were subjected, including through the action of new population-specific (i.e., culturally-related) selective pressures induced by dietary changes [13–15]. Lactase persistence, a heritable condition in which the physiological ability to digest lactose (the sugar contained in fresh milk) is maintained throughout adulthood, probably represents the best-known example of an adaptation related to diet [16]. Indeed, convergent evolution of the trait has been demonstrated [17], and is explained by the emergence of similar selective pressures resulting from adopting a diet heavily relying on milk. However, demographic processes (as opposed to selective pressure) playing a significant role in the spread of the trait constitutes a valid, mutually non-exclusive alternative explanation [18–22]. Hence, disentangling the respective contributions of selective and demographic forces to extant levels and patterns of genetic differentiation between populations represents a challenging task [23].

Reduced arylamine *N*-acetyltransferase 2 (NAT2) activity is another trait whose evolution was probably also shaped by differential, population-specific selective pressures [24–26]. Due to its early discovery linked to its major role in the treatment of tuberculosis with isoniazid, inherited variation in *N*-acetylation activity is currently one of the best known pharmacogenetic traits [27]. The phenotype is driven by the existence of a polymorphic *N*-acetyltransferase 2 (NAT2) enzyme. This cytosolic enzyme, encoded by the gene *NAT2* of chromosome 8, is mainly expressed in the liver, small intestine and colon [28] where it catalyzes a Phase II acetylation reaction, i.e., the transfer of an acetyl functional group to the terminal nitrogen of aromatic amines, heterocyclic amines, and hydrazines [29]. Hydrazines are used in the synthesis of numerous organic molecules, like the anti-tubercular agent isoniazid, while aromatic and heterocyclic amines are known to be produced, for instance, in meat and fish cooked at high temperatures, as well as in combustion smokes such as those from tobacco, grasses and wood chips [30, 31]. Thus, the NAT2 enzyme plays a crucial role in the detoxification of numerous xenobiotic compounds, including common therapeutic drugs and exogenous chemicals present in the diet and the environment [32]. Acetylation capacity (also coined acetylation status) can be measured by an individual's response to drugs metabolized by the enzyme, and this phenotype is now known to depend on the individual's genotype at the single coding exon of the *NAT2* gene. Mutations in *NAT2* result indeed in diverse acetylation phenotypes that explain inter-individual variation in response to standard drug dose administration (which can vary from lack of therapeutic efficacy to adverse drug reactions) [27, 33, 34]. Moreover, mutations in *NAT2* may also act as risk factors for different types of cancers [35, 36]. The gene displays a high degree of polymorphism in humans, with 88 alleles (i.e., haplotypes) listed to date by the official consensus gene nomenclature of human *NAT2* alleles (Arylamine *N*-acetyltransferase Gene Nomenclature Committee, nat.mbg.duth.gr, see [37]). Inter-population variation in *NAT2* allele frequencies has been intensively documented, particularly so for those alleles defined by the combination of nucleotides at four major functional single nucleotide polymorphisms (rs1801279, rs1801280, rs1799930, and rs1799931) of the coding exon [38] (and references therein).

Although response to drug intake is a quantitative variable, the distribution of the phenotype in tested human groups was shown to be at least bi- (or tri-) modal [39]. A simplified model of genotype-phenotype relationships was therefore adopted from the early studies onwards, in which those alleles considered as fully functional are responsible for the fast acetylator status, whereas decreased function alleles are responsible for the slow acetylator status [39–41]. Hence, acetylation status of an individual that is carrier of two fully functional alleles is classified as fast, that of a carrier of one fully functional allele and one allele with decreased function is classified as intermediate, and that of a carrier of two decreased function alleles as slow. Also, many studies do not distinguish between fast and intermediate acetylators, categorizing both types of subjects as fast (or rapid) acetylators. Similarly to lactase persistence, *NAT2* phenotypes prevalence varies substantially between populations [24–26, 38, 42–44], and hypotheses of gene-culture coevolution of *NAT2* polymorphisms have been put forwards, notably the idea that the slow acetylator phenotype became selectively advantageous in those populations adopting a food-producing mode of subsistence in the Neolithic. Indeed, given its role in the detoxification of numerous exogenic compounds, the NAT2 enzyme acts at the interface between the organism and its chemical environment, and hence is a likely target for natural selection. However, in contrast to lactose for lactase persistence, a dietary or environmental causal factor driving *NAT2* evolution is not identified, but hypotheses involving meat consumption and availability in food supplied folates related to nutritional shifts during the Neolithic have been proposed [25].

These hypotheses have been fuelled by the finding of significantly different frequency distributions of acetylation phenotypes between agriculturalists and hunter-gatherers, both in sub-Saharan Africa [44] and globally at the worldwide scale [25, 38, 45]. In addition, significant differential acetylation prevalence between sedentary agriculturalists and nomadic pastoralists has been reported in Central Asia [43]. Here, a higher proportion of slow acetylators was found among Tajik agriculturalists than among Kirghiz nomadic pastoralists. These two distinct food-producing strategies also prevail among sympatric populations of the African Sahel, but knowledge on *NAT2* genetic variation among African pastoral nomads was up to now very scarce, thus hindering a meaningful statistical analysis of the likely differences between food-producing subsistence modes in sub-Saharan Africa (see Supporting Information in [38]).

Current research holds that the earliest food-producing communities in sub-Saharan Africa relied on nomadic cattle herding, while sedentary farming would be a more recent phenomenon [46–50]. As the southern part of the African continent was occupied rather recently by Bantu herders, the most diversified pastoral societies live nowadays in the Sahelian belt, and these are, from west to east (and mentioning only the most important ones), the Moors, Fulani, Tuareg, Tubu, Zaghawa, nomadic Arabs (Kababish and Baggara) and Beja. This region forms a unique ecosystem bordered by the Saharan desert and the tropical rainforests, extending from east to west all across the continent. Its climate is characterized by annual cycles of wet and dry seasons allowing the co-existence of sedentary farmers and nomadic pastoralists who move with their livestock between wet- and dry-season pastures [51, 52]. In this study, we carried out an investigation of *NAT2* sequence variation among six Sahelian populations relying either on nomadic pastoralism or on sedentary agriculture sampled in an area extending from western Burkina Faso to northeastern Chad, in order to explicitly address the hypothesis that different selective pressures associated to these lifestyles acted on the evolution of *NAT2*. Moreover, by combining our dataset with published African samples of *NAT2* sequences (Table 1 and Additional file 1: Figure S1), we also address the hypothesis that selective pressures, stemming from an environmental factor linked to the ecoregion in which populations have been living (i.e., climatic zone and biome), might also have shaped the evolution of this gene.

Results

NAT2 diversity in the Sahel

We analyzed a total number of 287 samples from six well-defined Sahelian populations relying on two different modes of subsistence, namely pastoralism (Fulani from Banfora and from Tindangou in Burkina Faso, Fulani from Ader in Niger, Fulani from Bongor in Chad, and Daza in Chad) and agriculture (Kanembou in Chad). We completed this new dataset of *NAT2* sequences with those from other African population published samples (Table 1 and Additional file 1: Figure S1). For each of the 287 Sahelian samples, 1,396 base pairs (bp) encompassing the 870 bp *NAT2* coding exon were successfully sequenced (Additional file 2). In total, 15 polymorphic positions were observed, 11 of which are located in the coding exon, and 9 being non-synonymous (Fig. 1).

From the 15 single nucleotide polymorphisms (SNPs) observed, a total of 21 haplotypes were inferred with the expectation-maximisation (EM) algorithm of Arlequin for the six Sahelian populations (Fig. 1, Table 2 and Additional file 2). Among these, four haplotypes differed from known haplotypes at positions outside the coding exon, and were named with alphabetical suffixes (*NAT2*4a*, *NAT2*5Ba*, *NAT2*6Aa*, and *NAT2*14Ba*), while their predicted effect on the enzyme's activity was considered unchanged from the defining haplotype (respectively, *NAT2*4*, *NAT2*5B*, *NAT2*6A*, and *NAT2*14B*). The EM algorithm also led to the inference of three new haplotypes, and the following names were attributed by the official *NAT2* nomenclature committee (nat.mbg.duth.gr): *NAT2*12N*, *NAT2*13D* and *NAT2*14K*. All three are defined by a new combination of recognized signature SNPs from the official *NAT2* gene nomenclature with other SNPs, so that enzymatic activity could not be predicted for any of these three haplotypes.

Frequency distributions of the 21 *NAT2* haplotypes are reported in Table 2. Low-activity haplotype *NAT2*5B* is

Table 1 African population samples studied for *NAT2* sequence variation

Population	Code	Sample size	Geographical region (Country)	Living in dry savanna biome[1]	Linguistic affiliation	Subsistence mode	Reference
Egyptians	EGY	10	Northern (Egypt)	no	Afro-Asiatic	Agricultural	[25]
Mandenka	MAN	97	Western (Senegal)	yes	Niger-Congo	Agricultural	[24]
Fulani Banfora	FBAN	49	Western (Burkina Faso)	yes	Niger-Congo	Pastoralist	this study
Fulani Tindangou	FTIN	50	Western (Burkina Faso)	yes	Niger-Congo	Pastoralist	this study
Fulani Ader	FADE	48	Western (Niger)	yes	Niger-Congo	Pastoralist	this study
Yoruba in Ibadan	YRI	88	Western (Nigeria)	no	Niger-Congo	Agricultural	[53]
Yoruba Bantus	YOR	31	Western (Nigeria)	no	Niger-Congo	Agricultural	[44]
Yoruba	YRB	18	Western (Nigeria)	no	Niger-Congo	Agricultural	[97]
Yoruba CEPH	YO	12	Western (Nigeria)	no	Niger-Congo	Agricultural	[42]
Ibo	IBO	19	Western (Nigeria)	no	Niger-Congo	Agricultural	[97]
Hausa	HAU	17	Western (Nigeria)	yes	Afro-Asiatic	Agricultural	[97]
Kanembou	KANE	49	Central (Chad)	yes	Nilo-Saharan	Agricultural	this study
Daza	DAZ	41	Central (Chad)	yes	Nilo-Saharan	Pastoralist	this study
Fulani Bongor	FBON	50	Central (Chad)	yes	Niger-Congo	Pastoralist	this study
Fulani	FU	13	Central (Cameroon)	yes	Niger-Congo	Pastoralist	[42]
Kanuri	KN	12	Central (Cameroon)	yes	Nilo-Saharan	Agricultural	[42]
Mada	MD	14	Central (Cameroon)	yes	Afro-Asiatic	Agricultural	[42]
Ngumba Bantus	NGU	16	Central (Cameroon)	no	Niger-Congo	Agricultural	[44]
Lemande	LM	14	Central (Cameroon)	no	Niger-Congo	Agricultural	[42]
Bakola Pygmy	PYG	26	Central (Cameroon)	no	Niger-Congo	Hunter-gatherer	[44]
Bedzan Pygmy	BEZ	32	Central (Cameroon)	no	Niger-Congo	Hunter-gatherer	[44]
Baka Pygmy Cameroon	BAKC	31	Central (Cameroon)	no	Niger-Congo	Hunter-gatherer	[44]
Baka Pygmy Gabon	BAKG	16	Central (Gabon)	no	Niger-Congo	Hunter-gatherer	[44]
Akele Bantus Gabon	GAB	26	Central (Gabon)	no	Niger-Congo	Agricultural	[44]
Biaka Pygmy	BIA	24	Central (C. A. R.)	no	Niger-Congo	Hunter-gatherer	[44]
Mbuti Pygmy	MBU	24	Central (D. R. C.)	no	Nilo-Saharan	Hunter-gatherer	[44]
Dinka	DN	13	Eastern (South Sudan)	yes	Nilo-Saharan	Pastoralist	[42]
Luhya in Webuye	LWK	97	Eastern (Kenya)	yes	Niger-Congo	Agricultural	[53]
Maasai	MAS	12	Eastern (Kenya)	yes	Nilo-Saharan	Pastoralist	[97]
Luo	LUO	14	Eastern (Kenya)	yes	Nilo-Saharan	Pastoralist	[97]
Somali	SOM	20	Eastern (Somalia)	yes	Afro-Asiatic	Pastoralist	[44]
Turu	TR	15	Eastern (Tanzania)	yes	Niger-Congo	Agro-pastoralist	[42]
Hadza	HZ	14	Eastern (Tanzania)	yes	Khoisan	Hunter-gatherer	[42]
Sandawe	SW	18	Eastern (Tanzania)	yes	Khoisan	Hunter-gatherer	[42]
Burunge	BG	17	Eastern (Tanzania)	yes	Afro-Asiatic	Agro-pastoralist	[42]
Chagga Bantus	CHA	32	Eastern (Tanzania)	yes	Niger-Congo	Agricultural	[44]
Maasai	MS	14	Eastern (Tanzania)	yes	Nilo-Saharan	Pastoralist	[42]
San	SAN	38	Eastern (Zimbabwe)	yes	Khoisan	Hunter-gatherer	[97]
African Americans[2]	ASW	61	ND[3]	ND[3]	ND[3]	ND[3]	[53]

[1] Classification according to climatic zone and biome (ecoregion), based on [99]

[2] African Americans: Americans of African ancestry in Southwestern US

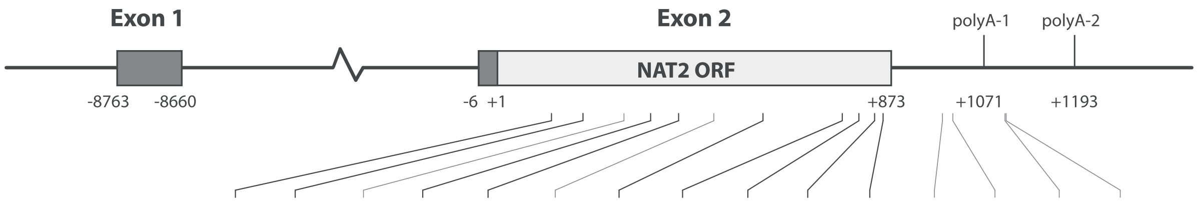
[3] ND: not defined

the most common haplotype in all six populations (41.9 % on average, Fig. 1). It is followed by *NAT2*6A* (24.5 % on average), also a low-activity haplotype. Thus, these two slow haplotypes together account for 67.4 % of the gene copies in the total sample of our study. Further low-activity haplotypes detected in most of our samples include *NAT2*5C*, **14A*, and **14B*, as well as **5A*, **7B* and *6C*. In turn, only three different fast haplotypes were detected in the Sahelian samples (*NAT2*4*, *NAT2*12A* and *NAT2*13A*), and these represent on average less than 18 % of the gene copies. Finally, six haplotypes could not be classified according to their effect on acetylation status. Their total frequency is of 3 % or less in the Fulani samples, whereas it is of 8.2 % and 6.1 % in the Kanembou and Daza, respectively.

Thus, when considering the whole 1,396 bp sequenced segment, 6 to 11 segregating sites, and 8 to 13 haplotypes were observed in each of the six Sahelian samples, and similar levels of diversity were estimated, as indicated by gene diversity (h) values ranging from 0.72 to 0.78, and nucleotide diversity (π) values ranging from 0.0018 to 0.0019 (Table 2). No departure from Hardy-Weinberg equilibrium, nor from selective neutrality and demographic equilibrium (with any of the Ewens-Watterson homozygosity or Fu's F_s tests) was found (all *P*-values > 5 %; results not shown). On another hand, a significant departure from the neutral equilibrium model in favor of overdominant selection was found with Tajima's *D* test for the Fulani nomads from the Banfora area ($D = 2.647$, $P = 0.006$, Table 2). This value remained significant after Bonferroni correction for multiple testing. We note that in the five other tested populations, estimates of *D* were all positive although not significant.

NAT2 diversity in African populations

The complete list of sixty-one *NAT2* haplotypes detected in the 39 African samples of coding-exon sequences is provided in Additional file 3: Figure S2 and Additional file 4: Table S1. Among these, seventeen haplotypes are newly described (notably due to the inclusion of samples from the 1000 Genomes Project [53]), and were assigned new names by the official *NAT2* nomenclature committee (Additional file 4: Table S1).



SNP		A121T	G191A	C282T	T341C	C403G	C481T	G590A	A766G	A803G	G838A	G857A	C989G	T1021C	C1134T	A1138G	
Amino acid change		N41Y	R64Q	-	I114T	L135V	-	R197Q	K256E	K268R	V280M	G286E	-	-	-	-	
Haplotype	Acetylation activity																Average frequency
NAT2*4	fast																5.1
NAT2*4a	fast														T		0.2
NAT2*12A	fast									G							5.8
NAT2*13A	fast			T													6.8
NAT2*5A	slow				C		T										0.8
NAT2*5B	slow				C		T			G							41.9
NAT2*5Ba	slow				C		T			G				C			0.7
NAT2*5C	slow				C					G							3.3
NAT2*6A	slow			T				A									24.5
NAT2*6Aa	slow			T				A								G	0.2
NAT2*6C	slow			T				A		G							0.2
NAT2*7B	slow			T								A					1.2
NAT2*14A	slow		A														2.4
NAT2*14B	slow		A	T													3.1
NAT2*14Ba	slow		A	T									G				0.2
NAT2*6F	unknown							A		G							0.2
NAT2*6O	unknown			T				A			A						0.4
NAT2*12H	unknown					G				G							2.5
NAT2*12N	unknown	T								G							0.2
NAT2*13D	unknown			T					G								0.2
NAT2*14K	unknown		A	T							A						0.2

Fig. 1 Schematic diagram of the *NAT2* locus on 8p22, including the non-coding 100 bp-long Exon 1, the 8.6 Kb-long intronic region, and the 870 bp-long single protein-coding Exon 2, adapted from [102] and [77]. The first (+1) and last positions (+873) of the open reading frame (ORF) of Exon 2 are indicated at the bottom of it, as well the relative positions of Exon 1 and of the polyadenylation signals. The positions of the 15 polymorphic sites (SNPs) observed among the 287 individuals from the six Sahelian samples sequenced in this study are shown as heavy-black (non-synonymous mutations) or light-gray (synonymous mutations) vertical bars below the diagram. Segments link the SNPs positions to the list of 21 haplotypes inferred from the combination of the 15 SNPs. Haplotypes' associated acetylation activity (taken from the official NAT2 gene nomenclature, http://nat.mbg.duth.gr/) and average frequency among the six Sahelian samples are shown on the left and right sides of the list, respectively. A diagram displaying the position of the 30 SNPs observed in protein-coding Exon 2 among the 39 African population samples (1,192 individuals) analyzed in this study is shown in Additional file 3: Figure S2

Table 2 Haplotype frequencies and molecular diversity of the six Sahelian samples in a 1,396 bp sequence encompassing the *NAT2* coding exon

		Population[1]						
Haplotype	Acetylation activity[2]	FBAN	FTIN	FADE	FBON	DAZ	KANE	Total
*NAT2*4*	fast	2	4	6	8	7	2	29
*NAT2*4a*	fast						1	1
*NAT2*12A*	fast	6	5		4	11	6	32
*NAT2*13A*	fast	11	10	8	7	2	2	40
*NAT2*5A*	slow		1		1		3	5
*NAT2*5B*	slow	39	46	42	47	28	40	242
*NAT2*5Ba*[3]	slow			2	1		1	4
*NAT2*5C*	slow	3	4	6	2	1	3	19
*NAT2*6A*	slow	32	17	20	23	23	25	140
*NAT2*6Aa*[3]	slow					1		1
*NAT2*6C*	slow				1			1
*NAT2*7B*	slow		1		2	2	2	7
*NAT2*14A*	slow	2	4	7		1		14
*NAT2*14B*	slow	3	5	2	2	1	5	18
*NAT2*14Ba*[3]	slow			1				1
*NAT2*6F*	unknown		1					1
*NAT2*6O*	unknown					1	1	2
*NAT2*12H*	unknown			2	1	4	7	14
*NAT2*12N*[4]	unknown		1					1
*NAT2*13D*[4]	unknown				1			1
*NAT2*14K*[4]	unknown		1					1
Total (2N chromosomes)		98	100	96	100	82	98	574
Number of haplotypes (k)		8	13	10	13	12	13	
Number of segregating sites (S)		6	9	9	10	10	11	
Gene diversity (expected heterozygosity, h)		0.72	0.75	0.75	0.72	0.78	0.76	
Nucleotide diversity (π) x 10^{-3}		1.81	1.79	1.85	1.78	1.83	1.93	
Tajima's *D* (*P*-value)[5]		**2.65 (0.994)**	1.09 (0.875)	1.18 (0.891)	0.73 (0.801)	0.72 (0.799)	0.68 (0.875)	

[1] Population codes as in Table 1

[2] Reported activity in the official *NAT2* gene nomenclature (nat.mbg.duth.gr)

[3] Small caps alphabetical suffixes were added to the names of haplotypes that differ from known haplotypes in the flanking region of the *NAT2* coding exon (see text)

[4] New haplotypes submitted to the official *NAT2* gene nomenclature and included in it (see text)

[5] *P*-value associated with Tajima's *D* test for departure from selective neutrality: it is given as the proportion of random *D* values generated under the neutral equilibrium model that are smaller than, or equal to the observed value. The sole significant result is shown in bold; it corresponds to a type I error rate of 0.006, and it remains significant after Bonferroni correction for multiple testing

Analysis of estimated diversity levels for the 870 bp *NAT2* coding exon evidenced substantial variation between populations, ranging from 0.65 to 0.91 for expected heterozygosity (h), and from 0.0022 to 0.0033 for nucleotide diversity (π) (Additional file 5: Table S2). Heterozygosity levels estimated for the six Sahelian populations are located in the lower half of the distribution (h varying from 0.71 to 0.78). For nucleotide diversity instead, four of our Sahelian samples display intermediate values (around 0.0029), while the Fulani from Bongor display a slightly lower value (0.0028), and the Kanembou a slightly higher one (0.0030). However, these estimates are associated with large standard deviations thus precluding finding significant differences in diversity levels between populations. This is especially the case of many of the published datasets, which include samples of less than 20 individuals. In support of this, we found a high and significant correlation between sample size and number of haplotypes detected ($r = 0.738$, $P < 0.00001$, Additional file 6: Figure S3), and it remained significant even after removal of the four samples with sizes larger than 50 individuals ($r = 0.398$, $P = 0.018$).

Substantial variation in frequency distributions of *NAT2* haplotypes was observed among African populations, particularly so for slow haplotypes *NAT2*5B* and **6A*, and the fast one *NAT2*12A* (Additional file 7: Figure S4). The variance in haplotype frequency among African populations is highest for *NAT2*5B* (0.018, Additional file 8: Figure S5). This low-activity haplotype is indeed generally more frequent in populations living in the Sahel or in the dry savannas of East Africa (more than 35 % on average) than in populations living to the south of it (less than 19 % on average), such as the Yoruba or the Pygmy populations (Additional file 7: Figure S4). *NAT2*6A*, also a slow haplotype, displays a similar trend. Conversely, the fast haplotype *NAT2*12A* displays a somewhat opposite trend, in that it is frequent in the Pygmy populations and in the San as well (from 22 % to 48 %), whereas its frequency is lower than 14 % in most Sahelian and East African populations. It is also rather frequent in those two Yoruba samples of small size (28 % and 21 % in the YRB and YO, respectively, both samples with less than 20 individuals), but not in the larger sized ones (9 % and 6 % in the YRI and YOR, respectively).

When considering only the 870 bp of the *NAT2* coding exon, a positive and significant Tajima's *D* value was found again for the Fulani from Banfora, as well as for the Fulani from Ader, the Fulani tested by Mortensen et al. [42], the Mbuti Pygmies, and the Egyptians (Additional file 5: Table S2). These five rejection cases result from an excess of intermediate frequency variants (i.e., positive *D* values ranging from 1.92 to 2.65). For the Mbuti Pygmies, a significant excess of observed heterozygosity was detected also with the Ewens-Watterson test ($P = 0.0463$), thus leading to the rejection of selective neutrality and demographic equilibrium. However, none of these results remained significant after Benjamini and Hochberg false discovery rate (FDR) adjustment for multiple testing.

NAT2 population structure in Africa

To investigate for the presence of a genetic structure differentiating populations with distinct food-producing subsistence modes (and for the influence of samples of small size on the results as well), analyses of population structure were performed on three subsets of the population data (Table 1), namely: (1) on the African populations dataset (AFR, i.e., 38 samples, excluding African Americans), (2) on the African food-producing populations dataset (FP, i.e., 29 samples, excluding hunter-gatherers), and (3) on an African food-producing populations dataset comprising only those samples with size ≥ 20 individuals (FPLS, i.e., 13 samples, excluding both hunter-gatherers and all other samples with sizes smaller than 40 chromosomes). Taking into account the molecular diversity of *NAT2* sequences (in terms of the number of pairwise differences between *NAT2* haplotypes), the global level of population structure estimated for the African continent is of 3.3% (AFR dataset, $\Phi_{ST} = 0.033$, $P < 0.0001$), which means that 3.3% of the total genetic variation is attributed to differentiation among populations, while 96.7% is due to differences among individuals within populations. In the AFR dataset, 33.4 % of the pairwise genetic distances between populations were found to be significant at the 5 % level. This proportion drops to 19.5 % in the FP dataset, thus suggesting substantial *NAT2* genetic differentiation both among hunter-gatherer populations and between hunter-gatherer and food-producing societies. Consistently, the proportion of genetic variance explained by differences among populations drops to 2.2% in the FP dataset, but it is still significant (FP dataset, $\Phi_{ST} = 0.022$, $P < 0.0001$). Thus, while differentiation of hunter-gatherer populations contributes to *NAT2* population structure in the African continent, population structure among food-producing communities is also significant. Among the non-significant pairwise genetic distances, a rather large proportion could actually be due to lack of power in significance testing because of small sample sizes. Indeed, the proportion of significant distances increases again to 34.6% in the FPLS dataset, i.e., in the dataset that excludes food-producing population samples of less than 20 individuals. In line with this, the proportion of genetic variance explained by differences among populations also raises to 2.6% (for the FPLS dataset, $\Phi_{ST} = 0.026$, $P < 0.0001$), thereby suggesting that the inclusion of samples of small size in the AFR and FP analyses leads to a lowering of the power to differentiate populations.

Most of the significant pairwise genetic distances in the FPLS dataset differentiate the Yoruba and Akele populations from the others (Additional file 9: Figure S6). Conversely, none of the six Sahelian populations was found significantly differentiated from the others, irrespective of language affiliation or lifestyle, and most of the genetic distances between our samples and the East African samples were also found statistically not significant.

Accordingly, no correlation of genetic distances with geographic distances was found for any of the AFR and FP datasets ($r = 0.054$, $P = 0.223$, and $r = -0.012$, $P = 0.531$, respectively), whereas the Mantel test with the FPLS dataset led to a significant negative correlation ($r = -0.218$, $P = 0.034$). This somewhat surprising result stems from both a set of large genetic distances between populations located in close geographic proximity and a set of small genetic distances between populations geographically far apart (Additional file 10: Figure S7), and does not fit the expectation of an isolation-by-distance model of evolution. In line with this, no significant correlogram of Moran's *I* spatial autocorrelation indices was found with any of the *NAT2* haplotypes observed in Africa, except for haplotype *NAT2*5C*, a rather infrequent haplotype.

Association of *NAT2* genetic structure with geography, culture, or climatic zone and biome

Hierarchical analyses of molecular variance (AMOVA) were carried out to gain further insight into a possible association of the genetic structure of populations with factors related to their demographic and cultural history. Four categorization criteria were tested (Table 1): geography, language, subsistence mode and ecoregion (biome). The results highlight a marked genetic structure associated with the fourth categorization criterion that considers whether populations live within the dry savanna biome or outside of it (Table 3). Indeed, with a classification of populations in ecoregions, Φ_{CT} indices were found high and significant for all three data subsets (Φ_{CT} of 2.3 %, 3.6 % and 5.4 %, for AFR, FP, and FPLS, respectively, all *P*-values < 0.05 and remaining significant after Bonferroni correction for multiple testing) and

Table 3 Analysis of molecular variance (AMOVA) under four criteria of classification of populations, based on the 870 bp long *NAT2* coding exon

				Percentage of variation		Fixation indexes			
Dataset[1]	Categories grouping[2]	Number of population samples	Number of groups	Between groups	Between populations within groups				
						Φ_{CT}	*P*-value[3]	Φ_{SC}	*P*-value[4]
AFR	Geo[5]	37	3	−0.01[6]	3.23	−0.00012[6]	0.3722	0.03230	**<0.0001** (**0.0004**)
	Lang	38	4	0.43	3.04	0.00425	0.1323	0.03053	**<0.0001** (**0.0004**)
	Subsist	38	4	1.82	2.02	0.01819	**0.0003** (**0.0012**)	0.02055	**<0.0001** (**0.0004**)
	Clim	38	2	2.35	2.11	0.02345	**0.0001** (**0.0004**)	0.02161	**<0.0001** (**0.0004**)
FP	Geo[5]	28	3	−0.04[6]	2.21	−0.00037[6]	0.4572	0.02207	**<0.0001** (**0.0004**)
	Lang	29	3	−0.21[6]	2.30	−0.00207[6]	0.6201	0.02297	**<0.0001** (**0.0004**)
	Subsist	29	3	1.16	1.58	0.01156	**0.0251** (0.1004)	0.01600	**0.0001** (**0.0004**)
	Clim	29	2	3.57	0.71	0.03572	**<0.0001** (**0.0004**)	0.00739	**0.0170** (0.0680)
FPLS	Geo	13	3	−0.53[6]	2.99	−0.00527[6]	0.7238	0.02974	**<0.0001** (**0.0004**)
	Lang[7]	12	2	−0.45[6]	2.84	−0.00447[6]	0.5612	0.02832	**<0.0001** (**0.0004**)
	Subsist	13	2	1.20	1.99	0.01202	0.0530	0.02016	**<0.0001** (**0.0004**)
	Clim	13	2	5.40	0.54	0.05403	**0.0035** (**0.0140**)	0.00568	**0.0462** (0.1848)

[1] As described in Methods, the three population data subsets are AFR: 38 samples, excluding the Americans of African ancestry (ASW of [53], see Table 1); FP: 29 samples of African food-producing populations; FPLS: 13 samples of African food-producing populations with sample size ≥ 20 individuals. Thus, the ASW sample was not considered in any of the AMOVA analyses

[2] Categories as in Table 1 : classification according to geographical region (Geo), subsistence mode (Subsist), linguistic affiliation (Lang), and ecoregion (Clim), namely climatic zone and biome, which defines the fourth categorization criterion that considers whether populations live within the dry savanna biome or outside of it

[3] Significance of the Φ_{CT} index and of the corresponding percentage of variation due to differences between groups. Significant *P*-values (i.e., <5 %) are shown in bold, and adjusted *P*-values after Bonferroni correction for multiple testing (here, four tests) are provided in brackets

[4] Significance of the Φ_{SC} index and of the corresponding percentage of variation due to differences between populations within groups. Significant *P*-values (i.e., <5%) are shown in bold, and adjusted *P*-values after Bonferroni correction for multiple testing (here, four tests) are provided in brackets

[5] Only three geographical regions are considered here (Western, Central and Eastern, Table 1) because the fourth region (Northern) is represented by one population sample only (EGY)

[6] Because variance components in AMOVA are actually defined as covariances, negative values can occur [95]. A negative Φ_{CT} value would be expected if gene copies were more correlated between groups than between populations within groups. However, none of the negative Φ_{CT} values in the table are statistically significant, thus indicating that they are equal to zero

[7] Only two linguistic families are considered here (Niger-Congo and Nilo-Saharan, Table 1) because the third family (Afro-Asiatic) is represented by one population sample only (SOM)

greater than Φ_{SC} indices in all cases (corresponding Φ_{SC} for AFR, FP and FPLS of 2.2 %, 0.7 % and 0.6 %, respectively, all *P*-values < 0.05). Significant genetic structure was also detected with a classification according to subsistence strategy, but under this categorization criterion and for each of the three datasets, the Φ_{CT} index is lower than the Φ_{SC} index, thereby meaning that more differentiation is found among populations in groups with matching subsistence mode than between those groups, even if only slightly more so. Finally, the results indicate that neither differentiation among geographic groups nor among linguistic groups does associate with the genetic structure of populations displayed by *NAT2* sequences, since Φ_{CT} indices were never found significant for any of the AFR, FP, or FPLS datasets (Table 3 and Additional file 2).

The two-dimensional plots resulting from the nonmetric multidimensional scaling (MDS) analyses of genetic distances among populations are shown in Fig. 2 for the AFR dataset, and in Additional file 11: Figure S8 and Additional file 12: Figure S9 for the FP and FPLS datasets, respectively. In each of these figures, the MDS plot is displayed four times, with populations being highlighted according to our four categorization criteria, namely geography, language, subsistence mode and biome. Consistent with the AMOVA results, categorization according to the environment seems to better fit with the location of populations in the plots than any of the three other criteria (see also Additional file 2).

NAT2 Phenotypes

The frequency of slow acetylators in each population sample was predicted from individuals' *NAT2* genotypes (i.e., diplotypes). Predicted proportions of slow acetylators in African populations (Additional file 13: Table S3) are displayed on a map in Fig. 3. Boxplots of predicted prevalence of slow acetylators (Fig. 4) indicate higher median proportion of slow acetylators for pastoralists than for the other subsistence modes (pastoralists > agro-pastoralists > agriculturalists > hunter-gatherers, Fig. 4a), although the variances of these proportions

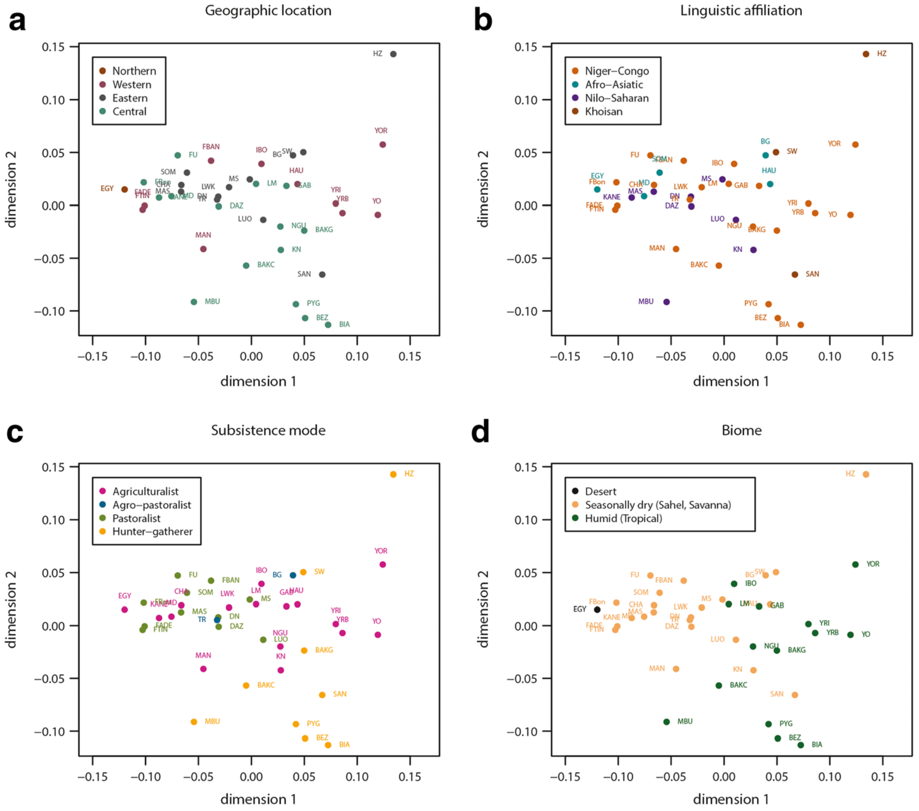


Fig. 2 MDS plot of pairwise Reynolds genetic distances between the 38 populations of the AFR dataset. The Stress value is 0.071. The same plot is reproduced 4 times, with populations color-coded according to: (**a**) geographical region, (**b**) linguistic affiliation, (**c**) subsistence mode, and (**d**) biome (see text)

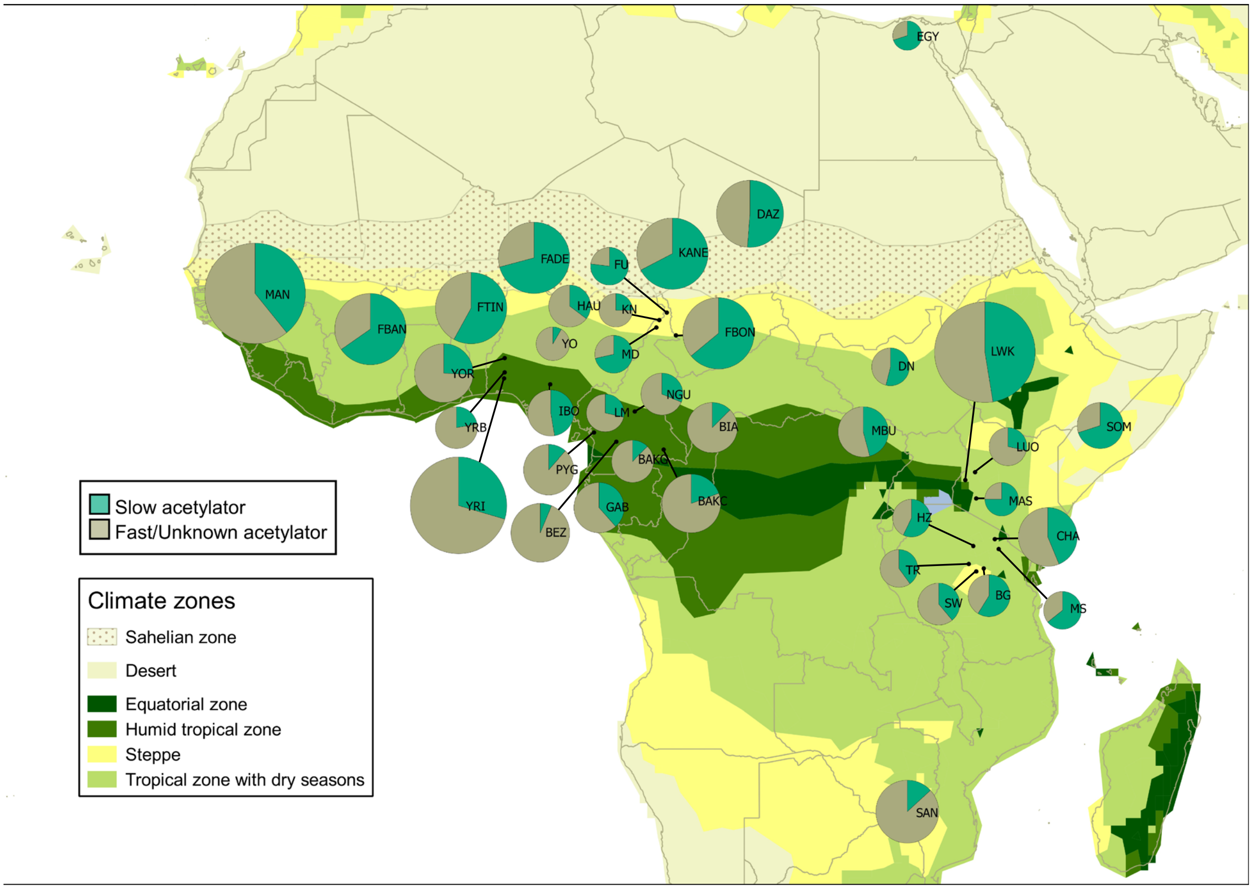


Fig. 3 Map showing the frequency distributions of predicted *NAT2* phenotypes in African populations screened for sequence variation in the coding-exon (pie charts are proportional to sample size). Map created with the QGis open source software [98], with climatic zones defined according to [99]

among populations are very large. On another hand, slow acetylation is more frequent in populations living/dwelling in the Sahelian belt or in the dry savanna surrounding it (including in the East African dry savanna zones) than in populations living to the South of it (Fig. 3), apparently irrespective of their subsistence strategy (Fig. 4b). Actually, crossing the information on subsistence mode with that on biome suggests higher slow acetylation in agricultural and hunter-gatherer populations living in the seasonally dry zones (Sahel, Savanna) than in those living in the humid tropical and equatorial zones. This trend seems even more marked for agriculturalists when only population samples of at least 20 individuals are considered (Fig. 4c).

Kruskal-Wallis tests for homogeneity of slow acetylation frequency among groups corroborated these observations (Table 4). Neither groups defined by geographic locations (Geo) nor groups defined by linguistic families (Lang) associate with differences in slow acetylation prevalence among populations. In turn, a significant difference in the frequency of slow acetylators was found among populations relying on different subsistence strategies (Subsist, $P = 0.0014$). This difference remains both when hunter-gatherers and when small samples are excluded from the analysis (FP and FPLS datasets, $P = 0.0161$, and $P = 0.0152$, respectively). When hunter-gatherers are excluded (FP dataset), only three subsistence modes are represented in the data, namely agriculturalists, agro-pastoralists, and pastoralists (Table 1). When small samples are excluded (FPLS dataset), only agriculturalists and pastoralists are represented, thus implying that slow acetylation prevalence is significantly higher among pastoralist than among agriculturalist populations. These results were confirmed by pairwise Wilcoxon tests of equality in average slow acetylation frequencies among subsistence strategies (Additional file 14: Table S4). A significant difference in slow acetylation prevalence was consistently found between pastoralists and agriculturalists (and between pastoralists and hunter-gatherers as well). Nevertheless, high prevalence of slow acetylation was also predicted for several agriculturalist populations from dry savanna regions, such as the Kanembou (67.3 %, 95 % confidence interval = [55.1%; 79.6 %]) and Luhya (47.4 %, 95 % confidence

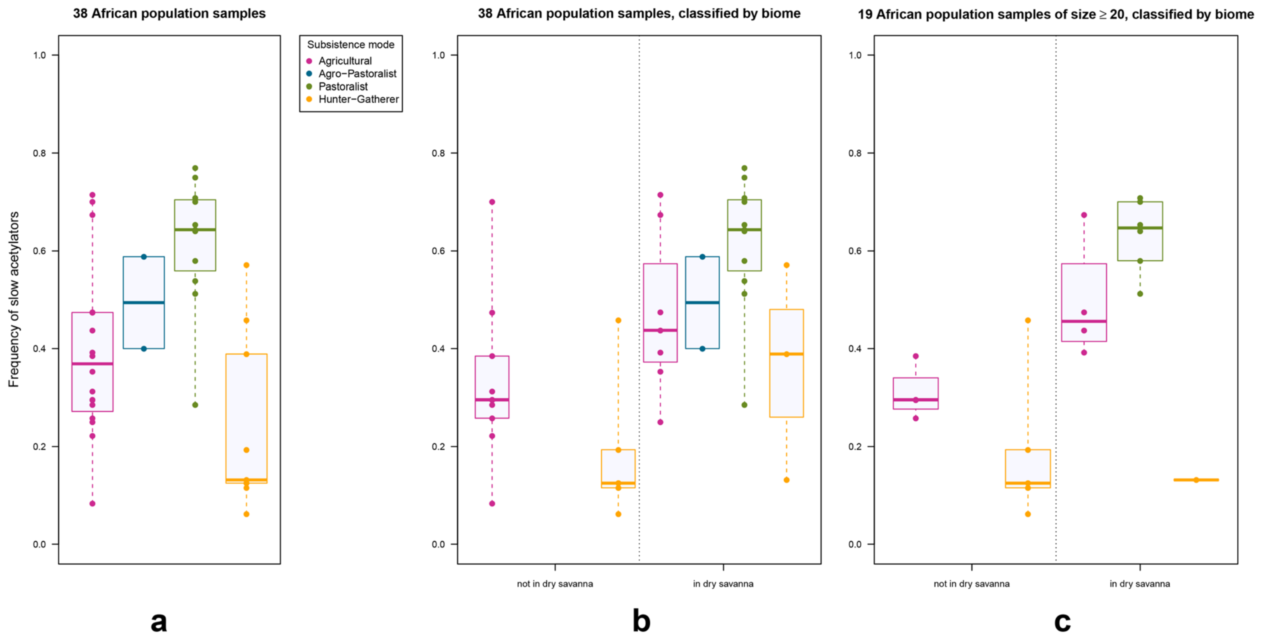


Fig. 4 Boxplots of predicted prevalence of slow acetylators in populations classified according to subsistence mode and further to ecoregion. Each box extends from the first to the third quartile and displays the median (thick line), and the dispersion of populations' individual values (filled circles) is shown by dashed lines. (**a**) Thirty-eight African populations (AFR dataset); (**b**) same dataset crossed with biome information, i.e., separating populations living in the seasonally dry zones (Sahel, Savanna) from those living in the humid tropical and equatorial zones; (**c**) same cross-analysis as in (**b**) but excluding population samples of less than 20 individuals

Table 4 Kruskal-Wallis test for equality of frequency of the slow acetylation phenotype across geographical regions, linguistic families, subsistence modes, and climatic zones

Dataset[1]	Categories grouping[2]	Number of populations	Number of groups	Kruskal-Wallis H statistic	*P*-value[3]
AFR	Geo[4]	37	3	2.38	0.3039
	Lang	38	4	7.13	0.0680
	Subsist	38	4	15.62	**0.0014** (**0.0056**)
	Clim	38	2	13.17	**0.0003** (**0.0012**)
FP[5]	Geo[4]	28	3	2.52	0.2843
	Lang	29	3	3.61	0.1649
	Subsist	29	3	8.25	**0.0161** (0.0644)
	Clim	29	2	7.74	**0.0054** (**0.0216**)
FPLS	Geo	13	3	0.33	0.8480
	Lang[5]	12	2	1.15	0.2827
	Subsist	13	2	5.90	**0.0152** (0.0608)
	Clim	13	2	6.43	**0.0112** (**0.0448**)

[1] As described in Methods, the three population data subsets are AFR: 38 samples, excluding the 1KG Americans of African ancestry (ASW, see Table 1); FP: 29 samples of African food-producing populations; FPLS: 13 samples of African food-producing populations with sample size ≥ 20 individuals. Thus, the ASW sample was not considered in any of the Kruskal-Wallis tests

[2] Categories as in Table 1 : classification according to geographical region (Geo), subsistence mode (Subsist), linguistic affiliation (Lang), and ecoregion (Clim), namely climatic zone and biome, which defines the fourth categorization criterion that considers whether populations live within the dry savanna biome or outside of it

[3] Significant *P*-values (i.e., <5 %) are shown in bold, and adjusted *P*-values after Bonferroni correction for multiple testing (here, four tests) are provided in brackets

[4] Only three geographical regions are considered here (Western, Central and Eastern, Table 1) because the fourth region (Northern) is represented by one population sample only (EGY)

[5] Only two linguistic families are considered here (Niger-Congo and Nilo-Saharan, Table 1) because the third family (Afro-Asiatic) is represented by one population sample only (SOM)

interval = [38.1 %; 56.7 %]), thus suggesting a possible association of acetylation status with ecoregion, namely with the climatic zone and biome populations live in (Fig. 4c). In line with this observation, a highly significant difference in average slow acetylation frequency was consistently found between populations living within the dry savanna biome or outside of it (Clim, P = 0.0003, P = 0.0054, and P = 0.0112, for the AFR, FP, and FPLS datasets, respectively, Table 4), irrespective of subsistence strategy.

Single factor analyses of variance (ANOVA) corroborated the results of the Kruskal-Wallis rank sum tests in that predicted slow acetylation frequency is observed to differ significantly between populations classified both by ecoregion (Clim) and subsistence mode (Subsist), but not by geography (Geo) or linguistic classification (Lang) (Additional file 15: Table S5). Moreover, ecoregion and subsistence mode were shown to be independently associated with differences in mean predicted slow acetylation frequency by multi-factor ANOVAs, as no significant interaction between these two factors (categorization groups) was found. Finally, Tukey's honest significant difference (HSD) tests for multiple comparisons of means confirmed the significant difference in predicted frequency of slow acetylators between populations living/dwelling in seasonally dry zones (Sahel, savanna biome) and those living in humid tropical and equatorial zones (i.e., "Living outside"), as well as between populations relying on different subsistence modes, notably between those relying on pastoralism and those relying on agriculture (Additional file 15: Table S5).

Taken together, our results thus indicate that factors associated both to the ecoregion in which populations have been living/dwelling and the mode of subsistence on which they have been relying independently explain a significant proportion of NAT2 phenotypic diversity in Africa.

Discussion

In this study, we have generated 1,396 bp of sequence encompassing the *NAT2* coding exon in 574 chromosomes sampled in six Sahelian populations. The activity of the enzyme encoded by this highly polymorphic gene is of clinical importance as it influences individuals' physiological response to the absorption of numerous chemical compounds present in the diet and the environment, including several nowadays useful medications, and it is also involved in cancer development risk [37]. Five of the populations studied rely on pastoralism as a mode of subsistence, four of which are the Fulani nomads from Niger, Burkina Faso and Chad, all speaking dialects of a Niger-Congo language (Fulfulde) [54, 55], whilst in the fifth one, the Daza from Chad, a Nilo-Saharan language (Dazaga) is spoken [54, 56]. The sixth population sample studied represents the Kanembou, a sedentary population, also from Chad, but relying on agriculture, and speaking a Nilo-Saharan language [54, 57].

Intermediate to low *NAT2* molecular diversity in Sahelian populations

Twenty-one haplotypes were parsimoniously inferred for the six Sahelian samples, of which three haplotypes not yet described to the best of our knowledge (*NAT2*12N*, *NAT2*13D* and *NAT2*14K*, Fig. 1). In order to compare our newly generated dataset with published *NAT2* sequences from other African samples, we restricted the analysis to the 870 bp of the *NAT2* single coding exon. We have thus shown that levels of genetic diversity in the six Sahelian populations are amidst the low to intermediate range of values in the scale of *NAT2* polymorphism known for the African continent, especially so when measured by expected heterozygosity (Additional file 5: Table S2). This result could be due to faster genetic drift in these populations compared to others, which is congruent with the idea, based on analyses of polymorphisms in uniparentally transmitted DNA (mitochondrial DNA and Y chromosome) showing that some Sahelian populations experienced a demographic contraction in their history [54, 55]. Actually, in line with mitochondrial DNA (mtDNA), *NAT2* diversity was found somewhat lower in the Fulani pastoral groups than in the Kanembou agriculturalists (both when measured by expected heterozygosity and by nucleotide diversity), consistent with recent genomic results that inferred smaller expansion rates for pastoral populations than for those having adopted agriculture [58, 59]. Moreover, while nearly none of the Ewens-Watterson homozygosity or Fu's F_S tests on *NAT2* variation rejected the null hypothesis of selective neutrality and demographic equilibrium, Tajima's D tests were found significantly positive for three Fulani groups, although not anymore after correction for multiple testing. Nevertheless, for one of those Fulani groups (FBAN), Tajima's D was positive when tested on the entire sequenced segment of 1,396 bp, and it remained significant after correction (Table 2). Thus, this trend towards Tajima's D positive values could also point to a reduction in population size [60], particularly so in the pastoral nomads. However, irrespective of subsistence mode, most African populations analyzed here displayed the same trend towards positive D values, which are also expected if the gene is submitted to balancing selection. The signature left by balancing selection should be detected by the Ewens-Watterson homozygosity test, since this selective regime should lead to an excess of heterozygotes, as has been shown for the Human Leukocyte Antigen (HLA) loci [61]. As already stated, among the 39 Ewens-Watterson tests performed on the complete population dataset, a single case of significant rejection of neutrality was found for the Mbuti Pygmies (Additional file 5: Table S2), because of heterozygosity in excess (which would fit the balancing selection model), but not anymore after adjustment for type I error rate. While all other tests were not significant, the proportion of cases in which the observed heterozygosity was lower than the expected was of nearly 40% (i.e., 15 out of 38). However, positive *D* values can also result from the presence, in the populations' gene pools, of several molecularly rather distant haplotypes at frequencies higher than expected under neutrality. Such frequency distributions can develop if several standing variants become submitted to positive directional selection [62]. This interpretation was favored in a recent study that screened variation of the entire *NAT2* locus sequence (ca. 10 kb) in 14 world populations from the 1000 Genomes Project [63], but no pastoral population was represented in that collection.

The genetic structure of *NAT2* in Africa is neither associated to geography nor to linguistics

Consistent with previous studies [24], we found a significant genetic structure of *NAT2* in Africa (Φ_{ST} = 3.3 % for the AFR dataset of 38 populations). Here we further showed that this structure is not only attributable to the genetic differentiation of hunter-gatherer populations but is also found among African food-producing communities, as attested by the significant Φ_{ST} value of 2.6 % for the FPLS dataset of 13 populations. This significant genetic structure is mainly due to the variation in frequencies of three major haplotypes, two of which are categorized as decreased function alleles (*NAT2*5B* and *NAT2*6A*), which are particularly frequent in populations living in the dry savanna biome, and one fast allele (*NAT2*12A*), more frequently detected outside this region (i.e., in populations living in tropical humid environments) than within it. Accordingly, both the AMOVA (Table 3) and MDS analyses (Fig. 2 and Additional file 11: Figure S8 and Additional file 12: Figure S9) indicate that neither differentiation among geographic groups nor among linguistic groups does associate with the genetic structure of populations displayed by *NAT2* sequences, contrarily to what is known for polymorphisms at other loci of the genome, such as mtDNA [64], Y chromosome and classical polymorphisms [55, 65, 66], or even for genome-wide variation [67, 68]. Furthermore, the genetic structure of *NAT2* in Africa does not strongly associate with subsistence strategy either, mainly because most Sahelian and East African populations display little to no genetic differentiation between them (Additional file 9: Figure S6), although both regions are populated by nomadic or semi-nomadic pastoralist and sedentary agriculturalist communities. Actually, this low genetic differentiation between most Sahelian and East African populations suggests a possible climate and biome link that is supported by the significant association of *NAT2* genetic structure with a classification of populations according to ecoregions. However, in contrast to biological functions for which the link with environmental pressures is obvious, such as immunity and pathogens [69–72], a climate and biome related factor that would affect *NAT2* evolution, if any, remains to be determined [25].

Differences in slow acetylation prevalence across the southern Sahelian limit

The frequency of slow acetylators in each population sample was predicted from individuals' *NAT2* genotypes (i.e., diplotypes). Kruskal-Wallis tests of homogeneity in proportions of slow acetylators among groups of populations did not evidence significant variation between geographic groups, nor between linguistic groups. In turn, significant difference of slow acetylators were found between subsistence strategies, and specifically between pastoralist and agriculturalist populations (Table 4, P = 0.016 and P = 0.015, for the FP and FPLS datasets, respectively; see also Additional file 14: Table S4 and Additional file 15: Table S5). Considering only those populations for which sample sizes included at least 20 individuals, average frequency of slow acetylators among pastoralists is of 63.2 % (±7.5), while it is of 41.6 % (±13.6) among agriculturalists (Additional file 13: Table S3 and Fig. 4). The higher standard deviation estimated for the latter group reflects a high variance in slow acetylator frequencies among agriculturalist populations such as the Yoruba (less than 30 %) and the Kanembou (more than 67 %). By contrast, among pastoralists, these frequencies vary from more than 50 % in the Daza to around 70 % in the Somali and the Fulani from Ader. For comparison, slow acetylators average to 18.1 % (±14.2) among hunter-gatherers.

Similarly to the finding of a significant difference in proportions of acetylation phenotypes between sedentary farmers and nomadic pastoralists in Central Asia [43], these results are compatible with the hypothesis of a differential evolution of *NAT2* acetylation capacity according to lifestyle among food-producing societies in Africa. It is thus tempting to suppose that the driving selective force behind this differential evolution is to be found in the development of dietary differences due to lifestyle. Polymorphisms in the *NAT2* gene or downstream from it and tagging its phenotypic expression have been recently identified by genome-wide association studies for their role in the metabolism of sugars and carbohydrates [73, 74], or in that of lipids [75, 76], thus making a strong case for dietary influences. However, we have also found a significant difference in frequencies of the slow acetylator phenotype across the southern Sahelian limit, thus showing that such difference is actually enclosed in the zones defined by climate and biome in which the populations are living. Irrespective of lifestyle (whether pastoralist, agriculturalist, or even agro-pastoralist), we found significantly higher proportions of slow acetylators in populations living in the Sahelian and tropical with dry seasons zones than in those living in the humid tropical and equatorial zones (Table 4, P = 0.005 and P = 0.011, for the FP and FPLS datasets, respectively). Multi-factor ANOVAs further showed that no significant interaction between subsistence mode and ecoregion explained the observed differences in predicted mean slow acetylation frequency (Additional file 15: Table S5). Altogether, these results point to a possible independent influence of dietary habits on one hand (as reflected by subsistence modes), and of the chemical environment on the other hand (as reflected by climatic zones), in the evolution of *NAT2* diversity. Thus, this hypothesis does not necessarily call for the existence of specific dietary components associated with subsistence modes, and hence has the advantage of being compatible with the opposite pattern of variation observed in sub-Saharan Africa with respect to Central Asia, where higher slow acetylation prevalence is observed in sedentary agriculturalists than in nomadic pastoralists [43]. Further population sampling will thus be needed to substantiate it. At present, we can only acknowledge that the difference in slow acetylation frequency between African sedentary agriculturalist populations living within the dry savanna biome or outside of it was not found significant (Wilcoxon rank sum test P-values of 0.11 and 0.057, respectively, for the comparison of the sixteen agriculturalist samples, and that of the seven agriculturalist samples of size ≥ 20 individuals).

Causative factors driving *NAT2* evolution

Consistent with previous investigations on *NAT2* molecular variation [42] (and references therein), no statistically significant signal of selection was found in our datasets after correction type I error rate. This suggests that if positive directional selection for slow-causing *NAT2* haplotypes in specific dietary and environmental conditions has been acting on the genetic evolution of populations, such selective pressures might have acted on multiple standing genetic variation rather than on specific new *NAT2* mutations [25, 26, 38, 43, 77]. We acknowledge indeed that classical selective neutrality tests on population molecular data (such as the Ewens-Watterson homozygosity, Tajima's D and Fu's F_S tests used here) are known to have little power to detect a signal of selection on standing genetic variation [62, 78]. Thus, the possibility exists that the genetic structure of African food-producing populations inferred through the analysis of *NAT2* molecular diversity, which differentiates populations from distinct climatic zones, could result from differentiated selective pressures on standing variation related to the xenobiotic environment of distinct climate and biome zones, and explain the close genetic proximity of Sahelian populations with East African populations independently of their lifestyle.

Alternatively, archaeological evidence suggests that the Sahel could have been first inhabited only by pastoralists, and adoption of farming by some populations would represent a recent, secondary change [47, 48]. Thus the lack (or nearly so) of significant genetic structure of *NAT2* molecular variation among Sahelian populations would be compatible with the hypothesis of differentiated selective pressures on standing variation related to xenobiotic intake of distinct dietary habits. This hypothesis has the advantage of explaining the significant difference in frequencies of the slow acetylator phenotype observed between African pastoralists and agriculturalists. Consequently, either secondary adoption of farming by formerly pastoralist populations should also be postulated for some East African populations, such as the Luhya from Kenya (of the 1000 Genomes Project [53]), or a rather common occurrence of gene flow between pastoralists and agriculturalists populations in East Africa should be envisioned. This latter case is compatible with the complex history of East African populations evidenced for mitochondrial DNA [64], as well as at the genome-wide level [68, 79–81].

Finally, as discussed in [24], it is also possible to envision that no selective constraint has acted on the evolution of *NAT2*, notably if the function of the enzyme encoded by this gene is redundant with other enzymes such as NAT1. In this case, the low genetic differentiation observed between Sahelian and East African populations could be due to a recent common origin. Under this last hypothesis, we would expect a consistent pattern of variation of *NAT2* with the genetic structure inferred from other independent genetic markers in the genome, thus pointing to the demographic history of these populations rather than to specific selective pressures having acted on *NAT2* evolution. This last alternative is not favored at present, given the patterns of variation known for other genomic segments, as discussed above, but it calls for further investigation because it would allow disentangling the effects of demographic and selective processes in the evolution of *NAT2*.

Our analysis suffers of course of some drawbacks. Firstly, as already stated, because it makes the simplistic assumption of a discrete categorization model of acetylation phenotypes (i.e., slow versus intermediate/fast acetylators) on the basis of genotypes, although heterogeneity in these categories has been recently demonstrated [33, 34]. For instance, acetylation status was found to vary significantly between the three genotypes formed by the two slow alleles *NAT2*5* and *NAT2*6*, with *NAT2*5* homozygotes being faster acetylators than *NAT2*5*/*NAT2*6* heterozygotes, and the latter displaying faster acetylation than *NAT2*6* homozygotes. In addition, the existence of variation in acetylation capacity due to polymorphisms outside the coding-exon of the *NAT2* gene is not considered here, although we know that regulatory polymorphisms could alter levels of protein expression and thereby acetylation activity. Besides, the existence of epistatic relationships between the expression of *NAT2* and its ortholog *NAT1* has been described [82], thus further emphasizing the complexity of the acetylation function in the organism and its evolution.

Secondly, the high level of polymorphism displayed by *NAT2* in human populations is probably very often underestimated with the samples sizes currently in use, as shown by the high and significant correlation of the latter with the number of alleles detected, at least for African populations (Additional file 6: Figure S3). Interestingly, we found that this correlation holds both for slow-causing variants (r = 0.655, P < 0.00001) and variants with unknown effect on phenotype as well (r = 0.605, P < 0.00001), but not for fast acetylation alleles (Additional file 16: Figure S10). This last result suggests that many more slow-causing variants could exist, probably at low frequencies, which is compatible with the hypothesis of a relaxation of functional constraints on *NAT2* in the course of human evolution, notably in those populations that adopted a food-producing mode of subsistence, where these variants are more frequent.

Thirdly, we should also consider that the discrete lifestyle categories in which we have classified our population samples (pastoralists, farmers or agro-pastoralists) are rather gross approximations of much more complex subsistence strategies. Extant cultures are not mono-subsistent as their diets also rely on various complementary modes of food supply, such as fishing and gathering, and complex networks of food exchange are concurrently used [83, 84]. For instance, some groups among the Kanembou agriculturalists from Chad are specialized in hunting and gathering [85]. Notably, the collection of a wild grass species known as *kreb* and of Spirulina algae known as *dihe'* makes a substantial contribution to the diets of these peoples [86]. At present unfortunately, to the best of our knowledge no accurate estimation of the relative dependence on various forms of subsistence is available for our samples.

Finally, the ecological zones considered in our analyses have been shifting in the past, so that populations living in one zone today might have coped with a different environment several generations ago. Extreme environmental changes are documented specifically in the Lake Chad basin; the current extension of the lake is ~20,000 km^2 but it is estimated that it was ~350,000 km^2 some ten thousand years ago [87]. In fact, with the onset of Holocene, the southern Sahara changed into a landscape rich in water resources, as evidenced by archaeological findings of items related to an aquatic life in places where we can barely imagine them today: in addition to direct evidence, such as hippopotamus and crocodile bones in dry river systems, there are findings of harpoons, which people used to hunt these animals [88].

Conclusion

The pattern of variation of *NAT2* in sub-Saharan African food-producing populations differs from that expected under a model of isolation-by-distance or from those observed with other genetic systems, but it is compatible with the hypothesis that it was shaped by selective pressures linked to the chemical environment in which populations evolved. We have shown that it is possible that differences in xenobiotic environments associate with climates and biomes that oppose arid, seasonally dry regions, such as the Sahel to tropical humid regions, such as around the Gulf of Guinea, hence explaining *NAT2* genetic and phenotypic differentiation across the southern Sahelian limit. However, the possibility exists that differences in xenobiotic environments interacting with NAT2 could also result from differential dietary habits linked to subsistence modes. Under this hypothesis, the genetic similarity between eastern and central-western African populations from the Sahel would then be explained by two non-mutually exclusive processes, namely significant gene flow across the Sahel, or secondary shift from pastoralism to agriculturalism in those Sahelian and East African populations that practice agriculture nowadays. Future studies including measurements of phenotypes and more sampling of populations (with large sample sizes) will certainly shed a new light on these conjectures.

Methods

Samples

Biological samples (buccal swabs or saliva samples) were collected during several missions in 2003, 2004, 2005 and 2010 in Niger, Burkina Faso and Chad, from unrelated anonymous volunteers, including only individuals with four grand-parents born to the population. Variability of DNA extracted from these samples was analysed in numerous former studies, such as in [54–57, 89–91].

Fulani nomads, representing the nomadic pastoral lifestyle, were sampled at different places within their geographic range across the Sahelian zone [54, 89]. The westernmost part of the Sahelian belt is represented by Fulani from the western and eastern parts of Burkina Faso, respectively (Banfora area, n=49, and Tindangou area, n=50).

More eastern Fulani groups were sampled in southern Niger (Ader area, *n*=48) and western Chad (Bongor area, *n*=50). As Fulani groups move frequently, the samples were secured in the temporary camps during the dry seasons, when the nomads rested at the southern extremity of their migration routes or transhumance corridors. The semi-nomadic Daza people (*n*=41) were collected in the northernmost regions of Lake Chad basin [56]. The Daza also rely mainly on animal husbandry, although some groups are semi-sedentary, cultivating palms in small oases. The sedentary Kanembou (*n*=49) from the northern fringes of Lake Chad represent agriculturalists in our dataset [54]. Informed verbal consent was obtained from all donors together with sampling authorization from the appropriate state institutions of the respective countries involved, namely from the Ministère des Enseignements Secondaire et Supérieur de la Recherche et de la Technologie (ref. 1534), and the Ministère de la Santé Publique et de la Lutte contre les Endémies, both in Niamey, Niger (ref. 13/2004/CCNE) for sampling in Niger, from the Ministère des Enseignements Secondaire et Supérieur de la Recherche Scientifique, in Ouagadougou, Burkina Faso (ref. 1029) for sampling in Burkina Faso, and from the Centre National d'Appui à la Recherche, in N'Djamena, Chad (ref. 01/CNAR/2010) for sampling in Chad. Ethics approval was required in Niger, and obtained on September 23, 2004, from the Comité Consultatif National d'Ethique (Ministère de la Santé Publique et de la Lutte contre les Endémies, Niamey, Niger). In Burkina Faso and Chad, no ethics approval was required by National Authorities, possibly because ethics commissions were not operating, but the same study protocol was followed in all three countries. Only the geographic location of sampling sites was recorded (i.e., no name or or any other identifiable information about volunteers was asked). An interpreter assisted the sampling collection, since the majority of participants expressed themselves in local languages/dialects, and most of them could not read or write in official language(s). Each participant provided informed consent orally and we repeatedly ensured, by means of the interpreter, that all participants fully understood the purpose of providing their saliva or buccal swabs: investigate the genetic variability of their population and its evolutionary causes. Participants were unrelated healthy adults, age range was 18–55 years, and both males and females were included. As many participants as technical limitations permitted (i.e., materials and costs) were included for sampling.

DNA samples preparation and *NAT2* sequencing

DNA was extracted from buccal swabs using the method described in [92]. Daza saliva samples were collected with Oragene™ DNA (OG-500) collection kit and extractions were performed according to the manufacturer's protocol. A segment of ca. 1.5 kb in the region encompassing the 870 bp single coding exon of the *NAT2* gene was amplified and subsequently sequenced in both forward and reverse directions. Two partially overlapping pairs of primers were used (number-named according to the *NAT2* exon): −128F/+914R [24] and +702F/+1373R [42]. Sequencing service was provided by Macrogen, Seoul, Korea. To confirm the detection of mutation 121A>T in a single Fulani individual in heterozygous state, another pair of primers was used (i.e., -15F and +779R of [42]).

Inference of *NAT2* haplotypes

The sequence haplotypes and their associated maximum likelihood (ML) frequencies were inferred separately for each population sample, using the Bayesian approach based on an approximate coalescent model implemented in the software PHASE v.2.1 [93, 94], and the maximum likelihood (ML) approach based on the expectation-maximisation (EM) algorithm implemented in Arlequin ver. 3.5 [95]. To this end, PHASE and Arlequin were run independently, and the resulting inferences were compared. Because we observed that PHASE results were less stable among runs (slightly different haplotype calls output with different seed numbers but with the same settings) than Arlequin results (identical haplotype calls with the same settings), we chose the latter software to obtain the haplotype calls and their associated ML frequencies (see also Additional file 2).

Following the standards of the official nomenclature of human *NAT2* alleles [37], haplotype *NAT2*4* (GenBank accession X14672) was used as a reference for the coding exon. However, because of known variation in the flanking region of the coding exon of *NAT2*4* alleles [26, 42, 44], we chose the human genome reference sequence (GRCh37/hg19, http://genome.ucsc.edu/) as a reference for the upstream and downstream flanking regions, and created a *NAT2*4*/hg19 construct. The sequences generated in this study were thus aligned with this construct, using ClustalW [96]. Two in-house programs were designed so as to (1) identify and map variant positions with respect to the *NAT2*4*/hg19 construct, and (2) classify inferred phased haplotypes according to the official *NAT2* gene nomenclature.

Collection of published *NAT2* sequences from African populations

Our new dataset was completed with published *NAT2* sequences from sampled African populations obtained through a comprehensive search of publications. Only populations represented by samples including at least 10 individuals (i.e., 20 chromosomes) were considered. Sequence genotypes from [97], obtained through personal communication with the authors, had to be phased in order to infer sequence haplotypes. The new dataset also comprises *NAT2* sequences reconstructed from the phased variant call file extracted from the 1000 Genomes Project (phase 1, release version 3 of April 2012, [53]), which were obtained with a Python program developed for this purpose. Together with the new 574 *NAT2* sequences generated in this study (i.e., 287 individuals), the complete dataset assembled includes 38 African population samples and one sample of African Americans, totalizing 1,192 individuals (Table 1 and Additional file 1: Figure S1). Geographic maps reporting the samples (Fig. 3, and Additional file 1: Figure S1 and Additional file 7: Figure S4) were created with the QGis open source software [98], and climatic zones were defined according to [99].

Estimates of diversity within populations

The ML frequency distributions of *NAT2* haplotypes in the six Sahelian population samples and published frequency distributions from other African samples were used to estimate molecular diversity indices (gene and nucleotide diversity), and to test for possible deviation from Hardy Weinberg equilibrium and selective neutrality (the latter through the use of three tests, Ewens-Watterson homozygosity test, Tajima's D test and Fu's F_s test) with the Arlequin software.

For all tests that revealed at least one significant departure from the null hypothesis in one population, we used R [100] to apply a correction method to control for type I error rate (either Bonferroni, or the less conservative Benjamini and Hochberg FDR when the number of tests exceeded ten), so as to obtain adjusted *P*-values. The R environment was also used to calculate Pearson's product–moment correlation coefficient between samples sizes and observed levels of diversity (e.g., number of haplotypes).

Estimates of differentiation between populations

The Arlequin software was also used to estimate population pairwise Φ_{ST} values (using the observed number of pairwise differences between haplotypes as a measure of molecular distance, with a 4:1 transition:transversion ratio), test their statistical significance (10,000 permutations) and compute Reynold's genetic distances, as well as to perform AMOVA analyses to test the significance of population groups' structures (here with100,000 permutations) according to geographic location, linguistic affiliation, subsistence mode or climatic zone and biome (i.e., ecoregion, following [99]). Under this last classification criterion, all populations living in the Sahel or in areas defined by a tropical with dry seasons climate were grouped into the "dry savanna" biome, whereas those living to the south and west of it, namely in the tropical humid and equatorial zones, were grouped into the "humid" or "desert" biome (Table 1). Matrices of pairwise Reynold's genetic distances were submitted to nonmetric multidimensional scaling analyses, using the function metaMDS of the vegan package in R. The mantel function of vegan was used to test the significance of correlation coefficients between matrices of pairwise geographic and genetic distances with Mantel tests. Spatial autocorrelation analyses were performed with the PASSaGE software [101].

Prediction of acetylation capacity and phenotype diversity among populations

Following previous reports [38, 63] (and references therein), the predicted acetylation phenotype associated with a given diploid haplotype combination was categorized as "slow" only if the latter is made of two haplotypes with reported low-activity in the official *NAT2* gene nomenclature. All other combinations (including those made of a low-activity haplotype and an unknown-activity haplotype or a newly described haplotype) were parsimoniously pooled together in a "fast/unknown" phenotype category, thus ensuring conservative estimation of slow acetylation prevalence in each population sample. Ninety-five percent non-parametric confidence intervals of slow-acetylation frequencies were generated by bootstrapping over individuals in each sample with the boot package of R (using 10,000 bootstrap samples). Homogeneity of inferred phenotype frequencies (i.e., "slow" versus "fast/unknown" acetylators) among geographic regions, linguistic groups, subsistence modes, and climatic zones was tested with R by performing Kruskal-Wallis rank sum tests and single/multiple factors ANOVAs, and differences in average slow acetylation frequencies among subsistence modes or between ecoregions were tested with pairwise Wilcoxon rank sum and Tukey's HSD tests.

Availability of supporting data

The data sets supporting the results of this article are available within the article and its additional files. The 574 *NAT2* sequences generated in this study are available under accessions KR231046 to KR231619 in the GenBank database at http://www.ncbi.nlm.nih.gov/genbank/.

Additional files

Additional file 1: Figure S1. Map showing the location of African populations screened for sequence variation in *NAT2*, including the six Sahelian populations of this study. The ASW sample of African Americans from the 1000 Genomes Project is not located on this map. Map created with the QGis open source software [98]. (PDF 2240 kb)

Additional file 2. Results. Supplementary text for Results section. (PDF 157 kb)

Additional file 3: Figure S2. Schematic diagram of the NAT2 870 bp-long single protein-coding Exon (Exon 2) on 8p22. The first (+1) and last positions (+873) of the ORF are indicated on top of it. The positions of the 30 polymorphic sites (SNPs) observed among the 1192 individuals from the 39

African samples analyzed in this study are shown as heavy-black (non-synonymous mutations) or light-gray (synonymous mutations) vertical bars below the diagram. Segments link the SNPs positions to the list of 61 haplotypes inferred from the combination of the 30 SNPs. Haplotypes' associated acetylation activity (taken from the official NAT2 gene nomenclature, http://nat.mbg.duth.gr/) and average frequency among the 39 population samples are shown on the left and right sides of the list, respectively. (PDF 21 kb)

Additional file 4: Table S1. *NAT2* coding-exon SNPs and haplotypes observed in all sequenced African population samples. (XLSX 32 kb)

Additional file 5: Table S2. Diversity of the *NAT2* 870 bp coding-exon in the 39 African populations included in this study. (XLSX 16 kb)

Additional file 6: Figure S3. Plot of the number of alleles (distinct *NAT2* sequence haplotypes) observed in samples as a function of sample size. The dashed line shows the linear regression of number of haplotypes on sample size, and Pearson's product–moment correlation coefficient is shown in the bottom-right caption. (PDF 2 kb)

Additional file 7: Figure S4. Frequency distributions of *NAT2* haplotypes in African populations screened for sequence variation in the coding-exon: (a) map showing frequency distributions as pie charts at geographic locations (with size of pie proportional to sample size; map created with the QGis open source software [98]); (b) frequency distributions shown as an area chart, and including the ASW sample of African Americans from the 1000 Genomes Project. Low-activity haplotypes are shown in shades of blue and green, fast haplotypes in shades of red, and those of unknown consequence in shades of gray. (PDF 238 kb)

Additional file 8: Figure S5. Variance in *NAT2* haplotype frequencies among African populations. The frequency variance in the complete 39 populations dataset is shown by light-gray bars, that of the AFR dataset (excluding ASW) by dark-gray bars. Haplotypes are ordered by decreasing variance, and only haplotypes displaying variance > 4.83e-4 are shown. (PDF 1 kb)

Additional file 9: Figure S6. Graphical representation of the matrix of pairwise Reynolds genetic distances among the 13 populations of the FPLS dataset (left-pane) and of the associated significance (right-pane). (PDF 41 kb)

Additional file 10: Figure S7. Plot of Reynolds pairwise genetic distances among the 13 populations of the FPLS dataset as a function of geographic distance separating them (great-circle distances, in km). (PDF 2 kb)

Additional file 11: Figure S8. MDS plot of pairwise Reynolds genetic distances between the 29 populations of the FP dataset. The Stress value is 0.045. The same plot is reproduced 4 times, with populations color-coded according to: (a) geographical region, (b) linguistic affiliation, (c) subsistence mode, and (d) biome (see text). (PDF 11 kb)

Additional file 12: Figure S9. MDS plot of pairwise Reynolds genetic distances between the 13 populations of the FPLS dataset. The Stress value is 0.014. The same plot is reproduced 4 times, with populations color-coded according to: (a) geographical region, (b) linguistic affiliation, (c) subsistence mode, and (d) biome (see text). (PDF 10 kb)

Additional file 13: Table S3. Predicted frequency of slow acetylators per population, with 95% non-parametric (bootstrap) confidence interval. (XLSX 12 kb)

Additional file 14: Table S4. Pairwise Wilcoxon rank sum tests of differences in predicted frequency of NAT2 slow acetylation among subsistence modes. (XLSX 11 kb)

Additional file 15: Table S5. Single factor and multiple factors ANOVAs of differences in predicted frequency of NAT2 slow acetylation, and Tukey's HSD tests for multiple comparisons of means. (XLSX 16 kb)

Additional file 16: Figure S10 Plot of the number of alleles (distinct *NAT2* sequence haplotypes) of each functional category (green for slow, brown for fast, orange for unknown) observed in samples as a function of sample size. The dashed lines show the linear regression of number of haplotypes on sample size, and Pearson's product–moment correlation coefficients are provided in the top-left caption. (PDF 2 kb)

Abbreviations

NAT2: Arylamine N-acetyltransferase 2; ORF: Open reading frame; SNPs: Single nucleotide polymorphisms; bp: Base pairs; AFR: African populations dataset; FP: African food-producing populations dataset; FPLS: African food-producing populations dataset comprising only those samples with size ≥ 20 individuals; AMOVA: Analysis of molecular variance; ANOVA: Analysis of variance; MDS: Nonmetric multidimensional scaling; Geo: Categorization according to geographic location; Lang: Categorization according to linguistic affiliation; Subsist: Categorization according to subsistence mode; Clim: Categorization according to climatic zone and biome; ND: Not defined; mtDNA: Mitochondrial DNA; HLA: Human Leukocyte Antigen; ML: Maximum likelihood; EM: Expectation-maximisation; FDR: False discovery rate; HSD: Honest significant difference; GRCh37/hg19: Human genome reference sequence (Genome Reference Consortium GRCh37, hg19 assembly).

Competing interests

The authors declare there are no competing interests.

Authors' contributions

ESP and VC conceived and designed the research. VC and ID designed and implemented samples collection. EP and ESP performed research and analyzed data, and CV, AS and ASM participated in some analyses. ESP, EP and VC wrote the manuscript. CV, AS and ASM participated in drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The project was supported by the Sciex-NMS.ch programme of the Rector's Conference of the Swiss Universities (Project 12.377 to ESP) and by the Grant Agency of the Czech Republic (grant 13-37998S-P505 to VC). We wish to express our gratitude to all donors for providing the samples for this study, to Pavel Podgorný for invaluable programing support in data analyses, and to Sotiria Boukouvala for her dedicated assistance in data submission to the official *NAT2* nomenclature committee. We would also like to thank two anonymous reviewers for their valuable comments and suggestions that helped to improve an earlier version of this manuscript.

Author details

[1]Department of the Archaeology of Landscape and Archaeobiology, Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, Czech Republic. [2]Department of Genetics and Evolution, Anthropology Unit, Laboratory of Anthropology, Genetics and Peopling History, University of Geneva, 12 Rue Gustave-Revilliod, 1211 Geneva 4, Switzerland. [3]Département de Linguistique et Langues Nationales, Institut des Sciences des Sociétés, CNRST, Ouagadougou, Burkina Faso. [4]IRD, UMR216, Mère et enfant face aux infections tropicales, Université Paris Descartes, Sorbonne Paris Cité, Faculté des Sciences Pharmaceutiques et Biologiques, Paris, France.

Received: 17 April 2015 Accepted: 13 November 2015
Published online: 01 December 2015

References

1. Laland KN, Odling-Smee J, Myles S. How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet. 2010;11(2):137–48.
2. Richerson PJ, Boyd R, Henrich J. Colloquium paper: gene-culture coevolution in the age of genomics. Proc Natl Acad Sci U S A. 2010;107 Suppl 2:8985–92.
3. Balaresque PL, Ballereau SJ, Jobling MA. Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet. 2007; 16(Spec No. 2):R134–139.
4. Harris EE, Meyer D. The molecular signature of selection underlying human adaptations. Am J Phys Anthropol. 2006;Suppl 43:89–130.
5. Feldman MW, Laland KN. Gene-culture coevolutionary theory. Trends Ecol Evol. 1996;11(11):453–7.
6. Lachance J, Tishkoff SA. Population Genomics of Human Adaptation. Annu Rev Ecol Evol Syst. 2013;44:123–43.

7. Vasseur E, Quintana-Murci L. The impact of natural selection on health and disease: uses of the population genetics approach in humans. Evol Appl. 2013;6(4):596–607.
8. Barbujani G, Colonna V. Human genome diversity: frequently asked questions. Trends Genet. 2010;26(7):285–95.
9. Novembre J, Pritchard JK, Coop G. Adaptive drool in the gene pool. Nat Genet. 2007;39(10):1188–90.
10. Bellwood P. First Farmers: The Origins of Agricultural Societies. Malden (MA): Blackwell; 2005.
11. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002;418(6898):700–7.
12. Cavalli-Sforza LL, Menozzi P, Piazza A. The History and Geography of Human Genes. Princeton, N.J.: Princeton University Press; 1994.
13. Jobling MA, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. Human Evolutionary Genetics. 2nd ed. New York and London: Garland Science; 2014.
14. Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, et al. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc Natl Acad Sci U S A. 2010;107 Suppl 2:8924–30.
15. Patin E, Quintana-Murci L. Demeter's legacy: rapid changes to our genome imposed by diet. Trends Ecol Evol. 2008;23(2):56–9.
16. Swallow DM. Genetics of lactase persistence and lactose intolerance. Annu Rev Genet. 2003;37:197–219.
17. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007;39(1):31–40.
18. Macholdt E, Slatkin M, Pakendorf B, Stoneking M. New insights into the history of the C-14010 lactase persistence variant in Eastern and Southern Africa. Am J Phys Anthropol. 2015;156:661–4.
19. Ranciaro A, Campbell MC, Hirbo JB, Ko WY, Froment A, Anagnostou P, et al. Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet. 2014;94(4):496–510.
20. Priehodova E, Abdelsawy A, Heyer E, Cerny V. Lactase persistence variants in Arabia and in the African Arabs. Hum Biol. 2014;86(1):7–18.
21. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, et al. Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond Ser B Biol Sci. 2011;366(1566):863–77.
22. Gerbault P, Moret C, Currat M, Sanchez-Mazas A. Impact of selection and demography on the diffusion of lactase persistence. PLoS ONE. 2009;4(7):e6369.
23. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The role of geography in human adaptation. PLoS Genet. 2009;5(6):e1000500.
24. Sabbagh A, Langaney A, Darlu P, Gerard N, Krishnamoorthy R, Poloni ES. Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet. 2008;9:21.
25. Luca F, Bubba G, Basile M, Brdicka R, Michalodimitrakis E, Rickards O, et al. Multiple advantageous amino acid variants in the NAT2 gene in human populations. PLoS ONE. 2008;3(9):e3136.
26. Patin E, Barreiro LB, Sabeti PC, Austerlitz F, Luca F, Sajantila A, et al. Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. Am J Hum Genet. 2006;78(3): 423–36.
27. Meyer UA. Pharmacogenetics - five decades of therapeutic lessons from genetic diversity. Nat Rev Genet. 2004;5(9):669–76.
28. Husain A, Zhang X, Doll MA, States JC, Barker DF, Hein DW. Identification of N-acetyltransferase 2 (NAT2) transcription start sites and quantitation of NAT2-specific mRNA in human tissues. Drug Metab Dispos. 2007;35(5):721–7.
29. Sim E, Abuhammad A, Ryan A. Arylamine N-acetyltransferases: from drug metabolism and pharmacogenetics to drug discovery. Br J Pharmacol. 2014;171(11):2705–25.
30. Sugimura T, Wakabayashi K, Nakagama H, Nagao M. Heterocyclic amines: Mutagens/carcinogens produced during cooking of meat and fish. Cancer Sci. 2004;95(4):290–9.
31. Kataoka H, Kijima K, Maruo G. Determination of mutagenic heterocyclic amines in combustion smoke samples. Bull Environ Contam Toxicol. 1998;60(1):60–7.
32. Hein DW. Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. Mutat Res. 2002;506–507:65–77.
33. Ruiz JD, Martinez C, Anderson K, Gross M, Lang NP, Garcia-Martin E, et al. The differential effect of NAT2 variant alleles permits refinement in phenotype inference and identifies a very slow acetylation genotype. PLoS ONE. 2012;7(9):e44629.
34. Selinski S, Blaszkewicz M, Ickstadt K, Hengstler JG, Golka K. Improvements in algorithms for phenotype inference: the NAT2 example. Curr Drug Metab. 2014;15(2):233–49.
35. Meisel P. Arylamine N-acetyltransferases and drug response. Pharmacogenomics. 2002;3(3):349–66.
36. Butcher NJ, Boukouvala S, Sim E, Minchin RF. Pharmacogenetics of the arylamine N-acetyltransferases. Pharmacogenomics J. 2002;2(1):30–42.
37. McDonagh EM, Boukouvala S, Aklillu E, Hein DW, Altman RB, Klein TE. PharmGKB summary: very important pharmacogene information for N-acetyltransferase 2. Pharmacogenet Genomics. 2014;24(8):409–25.
38. Sabbagh A, Darlu P, Crouau-Roy B, Poloni ES. Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. PLoS ONE. 2011;6(4):e18507.
39. Evans DA, Manley KA, Mc KV. Genetic control of isoniazid metabolism in man. Br Med J. 1960;2(5197):485–91.
40. Weber WW. The acetylator genes and drug response. 1987.
41. Weber WW, Hein DW. N-acetylation pharmacogenetics. Pharmacol Rev. 1985;37(1):25–79.
42. Mortensen HM, Froment A, Lema G, Bodo JM, Ibrahim M, Nyambo TB, et al. Characterization of genetic variation and natural selection at the arylamine N-acetyltransferase genes in global human populations. Pharmacogenomics. 2011;12(11):1545–58.
43. Magalon H, Patin E, Austerlitz F, Hegay T, Aldashev A, Quintana-Murci L, et al. Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. Eur J Hum Genet. 2008;16(2):243–51.
44. Patin E, Harmant C, Kidd KK, Kidd J, Froment A, Mehdi SQ, et al. Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene. Hum Mutat. 2006;27(7):720.
45. Valente C, Alvarez L, Marks SJ, Lopez-Parra AM, Parson W, Oosthuizen O, et al. Exploring the relationship between lifestyles, diets and genetic adaptations in humans. BMC Genet. 2015;16:55.
46. Smith AB. African herders : emergence of pastoral traditions. Walnut Creek: AltaMira Press; 2005.
47. Neumann K. The late emergence of agriculture in sub-Saharan Africa: archaeobotanical evidence and ecological considerations. In: Neumann K, Butler A, Kahlheber S, editors. Food, fuel and fields Progress in African archaeobotany. Koln: Heinrich-Barth-Institute; 2003. p. 71–92.
48. Marshall F, Hildebrand E. Cattle Before Crops: The Beginnings of Food Production in Africa. J World Prehistory. 2002;16(2):99–143.
49. Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Rege JE. African pastoralism: genetic imprints of origins and migrations. Science. 2002; 296(5566):336–9.
50. Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. Evidence for two independent domestications of cattle. Proc Natl Acad Sci U S A. 1994; 91(7):2757–61.
51. Homewood K. Ecology of African Pastoralist Societies. Oxford and Athens: James Currey and Ohio University Press; 2008.
52. Pedersen J, Benjaminsen TA. One Leg or Two? Food Security and Pastoralism in the Northern Sahel. Hum Ecol. 2008;36(1):43–57.
53. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
54. Cerny V, Pereira L, Musilova E, Kujanova M, Vasikova A, Blasi P, et al. Genetic structure of pastoral and farmer populations in the African Sahel. Mol Biol Evol. 2011;28(9):2491–500.
55. Buckova J, Cerny V, Novelletto A. Multiple and differentiated contributions to the male gene pool of pastoral and farmer populations of the African Sahel. Am J Phys Anthropol. 2013;151(1):10–21.
56. Podgorna E, Soares P, Pereira L, Cerny V. The genetic impact of the lake chad basin population in North Africa as documented by mitochondrial diversity and internal variation of the L3e5 haplogroup. Ann Hum Genet. 2013;77(6):513–23.
57. Cerny V, Salas A, Hajek M, Zaloudkova M, Brdicka R. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet. 2007;71(Pt 4):433–52.

58. Aime C, Verdu P, Segurel L, Martinez-Cruz B, Hegay T, Heyer E, et al. Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. Eur J Hum Genet. 2014;22(10):1201–7.
59. Aime C, Laval G, Patin E, Verdu P, Segurel L, Chaix R, et al. Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. Mol Biol Evol. 2013;30(12):2629–44.
60. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–95.
61. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PLoS ONE. 2011;6(2):e14643.
62. Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution. 2005;59(11):2312–23.
63. Patillon B, Luisi P, Poloni ES, Boukouvala S, Darlu P, Genin E, et al. A Homogenizing Process of Selection Has Maintained an "Ultra-Slow" Acetylation NAT2 Variant in Humans. Hum Biol. 2014;86(3):185–214.
64. Poloni ES, Naciri Y, Bucho R, Niba R, Kervaire B, Excoffier L, et al. Genetic evidence for complexity in ethnic differentiation and history in East Africa. Ann Hum Genet. 2009;73(Pt 6):582–600.
65. Sanchez-Mazas A, Poloni ES. Genetic Diversity in Africa. In: Encyclopedia of Life Sciences. Chichester: John Wiley & Sons, Ltd.; 2008.
66. Excoffier L, Pellegrini B, Sanchez-Mazas A, Simon C, Langaney A. Genetics and history of sub-Saharan Africa. Yearb Phys Anthropol. 1987;30:151–94.
67. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 2010;107(2):786–91.
68. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009;324(5930):1035–44.
69. Sanchez-Mazas A, Lemaître JF, Currat M. Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. Philos Trans R Soc Lond Ser B Biol Sci. 2012;367(1590):830–9.
70. Novembre J, Han E. Human population structure and the adaptive response to pathogen-induced selection pressures. Philos Trans R Soc Lond Ser B Biol Sci. 2012;367(1590):878–86.
71. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7(11):e1002355.
72. Dos Santos FR, Buhler S, Nunes JM, Bitarello BD, Franca GS, Meyer D, et al. HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. Immunogenetics. 2015;67:651–63.
73. Eny KM, Lutgers HL, Maynard J, Klein BEK, Lee KE, Atzmon G, et al. GWAS identifies an NAT2 acetylator status tag single nucleotide polymorphism to be a major locus for skin fluorescence. Diabetologia. 2014;57(8):1623–34.
74. Knowles JW, Xie W, Zhang Z, Chennemsetty I, Assimes TL, Paananen J, et al. Identification and validation of N-acetyltransferase 2 as an insulin sensitivity gene. J Clin Invest. 2015;125(4):1739–51.
75. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707–13.
76. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–+.
77. Fuselli S, Gilman RH, Chanock SJ, Bonatto SL, De Stefano G, Evans CA, et al. Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. Pharmacogenomics J. 2007;7(2):144–52.
78. Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 2012;8(10):e1003011.
79. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327–32.
80. Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 2014;111(7):2632–7.
81. Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, et al. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet. 2012;91(1):83–96.
82. Wang D, Para MF, Koletar SL, Sadee W. Human N-acetyltransferase 1 *10 and *11 alleles increase protein expression through distinct mechanisms and associate with sulfamethoxazole-induced hypersensitivity. Pharmacogenet Genomics. 2011;21(10):652–64.
83. Linseele V. From first stock keepers to specialised pastoralists in the West African savannah. In: Bollig M, Schnegg M, Wotzka H-P, editors. Pastoralism in Africa: Past, Present and Future. New York and Oxford: Berghahn Books; 2013. p. 145–70.
84. Sellen DW, Mace R. Fertility and mode of subsistence: a phylogenetic analysis. Curr Anthropol. 1997;38(5):878–89.
85. Nicolaisen I. Elusive hunters: the Haddad of Kanem and the Bahr el Ghazal. Copenhagen: Aarhus University Press; 2010.
86. Batello C, Marzot M, Touré AH. The future is an ancient lake : traditional knowledge, biodiversity and genetic resources for food and agriculture in Lake Chad Basin ecosystems. Rome: FAO; 2004.
87. Bouchette F, Schuster M, Ghienne J-F, Denamiel C, Roquin C, Moussa A, et al. Hydrodynamics in Holocene Lake Mega-Chad. Quat Res. 2010;73(2):226–36.
88. Drake NA, Blench RM, Armitage SJ, Bristow CS, White KH. Ancient watercourses and biogeography of the Sahara explain the peopling of the desert. Proc Natl Acad Sci U S A. 2011;108(2):458–62.
89. Cerny V, Hajek M, Bromova M, Cmejla R, Diallo I, Brdicka R. MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. Hum Biol. 2006;78(1):9–27.
90. Hajek M, Cerny V, Bruzek J. Mitochondrial DNA and craniofacial covariability of Chad Basin females indicate past population events. Am J Hum Biol. 2008;20(4):465–74.
91. Cerezo M, Cerny V, Carracedo A, Salas A. New insights into the Lake Chad Basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs. PLoS ONE. 2011;6(4):e18682.
92. Cerny V, Hajek M, Cmejla R, Bruzek J, Brdicka R. mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa. Ann Hum Biol. 2004;31(5):554–69.
93. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 2005;76(3):449–62.
94. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001;68(4):978–89.
95. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10(3):564–7.
96. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673–80.
97. Matimba A, Del-Favero J, Van Broeckhoven C, Masimirembwa C. Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects. Hum Genomics. 2009;3(2):169–90.
98. QGIS Development Team. QGIS Geographic Information System. In: Open Source Geospatial Foundation Project. 2014.
99. UNEP. Africa: Atlas of Our Changing Environment. In: Division of Early Warning and Assessment (DEWA). Nairobi, Kenya: UNEPU; 2008.
100. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
101. Rosenberg MS, Anderson CD. PASSaGE: Pattern Analysis, Spatial Statistics and Geographic Exegesis. Version 2. Methods Ecol Evol. 2011;2(3):229–32.
102. Boukouvala S, Sim E. Structural analysis of the genes for human arylamine N-acetyltransferases and characterisation of alternative transcripts. Basic Clin Pharmacol Toxicol. 2005;96(5):343–51.