Title: Positive selection of *AS3MT* to arsenic water in Andean populations

Author: Christina A. Eichstaedt Tiago Antao Alexia Cardona Luca Pagani Toomas Kivisild Maru Mormina

PII: S0027-5107(15)30026-9
DOI: http://dx.doi.org/doi:10.1016/j.mrfmmm.2015.07.007
Reference: MUT 11496

To appear in: *Mutation Research*

Received date: 14-4-2015
Revised date: 8-7-2015
Accepted date: 15-7-2015

Please cite this article as: Christina A.Eichstaedt, Tiago Antao, Alexia Cardona, Luca Pagani, Toomas Kivisild, Maru Mormina, Positive selection of AS3MT to arsenic water in Andean populations, Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis http://dx.doi.org/10.1016/j.mrfmmm.2015.07.007

Positive selection of *AS3MT* to arsenic water in Andean populations

Christina A. Eichstaedt[a, 1]*, Tiago Antao[b], Alexia Cardona[a], Luca Pagani[a, c], Toomas Kivisild[a], Maru Mormina[a, 2]*

[a]Division of Biological Anthropology, University of Cambridge, Cambridge CB2 1QH, Cambridgeshire, UK; ac812@cam.ac.uk, tk331@cam.ac.uk

[1]Centre for Pulmonary Hypertension, Thoraxclinic at the University Hospital Heidelberg, 69126 Heidelberg, Baden-Württemberg, Germany; Christina.Eichstaedt@med.uni-heidelberg.de

[b]Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK; tiagoantao@gmail.com

[c]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, UK; lp8@sanger.ac.uk

[2]Faculty of Humanities and Social Sciences, University of Winchester, Winchester SO22 4NR, Hampshire, UK; Maru.Mormina@winchester.ac.uk

*Corresponding authors:

E-mail: Christina.Eichstaedt@med.uni-heidelberg.de
Postal address: Centre for Pulmonary Hypertension
Thoraxclinic at the University Hospital Heidelberg
Amalienstr. 5
69126 Heidelberg
Germany
Phone: 0049/6221-3961221

E-mail: Maru.Mormina@winchester.ac.uk
Postal address: School of Humanities and Social Sciences
University of Winchester
Sparkford Rd
Winchester, SO22 4NR
United Kingdom
Phone: 0044/1962-827589

Abstract

Arsenic is a carcinogen associated with skin lesions and cardiovascular diseases. The Colla population from the Puna region in Northwest Argentinean is exposed to levels of arsenic in drinking water

exceeding the recommended maximum by a factor of 20. Yet, they thrive in this challenging environment since thousands of years and therefore we hypothesise strong selection signatures in genes involved in arsenic metabolism. We analysed genome-wide genotype data for 730,000 loci in 25 Collas, considering 24 individuals of the neighbouring Calchaquíes and 24 Wichí from the Gran Chaco region in the Argentine province of Salta as control groups. We identified a strong signal of positive selection in the main arsenic methyltransferase *AS3MT* gene, which has been previously associated with lower concentrations of the most toxic product of arsenic metabolism monomethylarsonic acid. This study confirms recent studies reporting selection signals in the *AS3MT* gene albeit using different samples, tests and control populations.

Highlights

- Argentine Collas have adapted to minimise consequences of arsenic water
- The selection pressure of arsenic can be identified at genome-wide level
- Beneficial allele frequencies are reduced with distance from arsenic exposure

Key words: arsenic drinking water; Collas; Puna; methyltransferase; Calchaquíes

Abbreviations

<xps:span class=deft>DMA </xps:span> <xps:span class=defd>dimethylarsinic acid</xps:span>
<xps:span class=deft>F_{ST} </xps:span> <xps:span class=defd>fixation index</xps:span>
<xps:span class=deft>iHS</xps:span> <xps:span class=defd>integrated haplotype score</xps:span>
<xps:span class=deft>LSBL </xps:span> <xps:span class=defd>locus specific branch length</xps:span>
<xps:span class=deft>MMA </xps:span> <xps:span class=defd>monomethylarsonic acid </xps:span>
<xps:span class=deft>PBS </xps:span> <xps:span class=defd>population branch statistic </xps:span>
<xps:span class=deft>XP-EHH </xps:span> <xps:span class=defd>cross population extended haplotype homozygosity</xps:span>

1. Introduction

High levels of arsenic in drinking water are found in countries all over the world [1]. Arsenic mainly originates from minerals in the ground and enters the food chain through drinking water and food sources such as crop plants [2]. Anthropogenic actions like mining and pesticide use contribute to elevated levels of arsenic [3].

Long-term exposure to arsenic can result in cancer, skin lesions, as well as cardiovascular and pulmonary diseases [4]. However, not only at a later stage in life but already at an early age arsenic exposure can have drastic consequences. Arsenic can cross the placental barrier and thus affect the

foetal development. Arsenic alters immune response modulator concentrations measured in breast milk [5] as well as in newborn cord blood [6]. Subsequently, high arsenic intake by drinking water in early childhood increases the risk of respiratory infections and diarrhoea in infants [7] as well as liver cancer associated mortality [8]. This suggests that populations exposed to high levels of arsenic over long periods of time may possess some kind of protection against arsenic toxicity.

In the body, inorganic arsenic is modified to monomethylarsonic acid (MMA) and subsequently to dimethylarsinic acid (DMA) [9] by methyltransferases. The second reaction occurs much faster due to an increased substrate affinity of the enzyme for MMA and therefore DMA is the predominant end product of arsenic metabolism [10]. Inorganic arsenic, MMA and DMA are excreted in the urine and can be used to measure arsenic metabolism. The most toxic arsenic product is MMA; thus, the first step in the arsenic metabolism is considered to be rather an activation than a detoxification of arsenic [11]. Hence, low levels of MMA in comparison to DMA in urine are beneficial to reduce its toxicity [12].

In the highlands of Northwest Argentina, the Puna, high levels of arsenic in water have been present since many thousands of years [13]. In some locations levels exceed the maximum safe level set by the WHO of 10 µg/l by a factor 20 [9]. San Antonio de los Cobres, in the heart of the Puna region, is one of such localities [9, 14]. Yet, its inhabitants show unusually low levels of excreted MMA metabolite relative to DMA and inorganic arsenic [9]. In agreement with this observation, Puna highlanders show increased frequencies of arsenic methyltransferase (*AS3MT*) alleles that have been associated with low MMA urine concentrations [15-17]. Allele differences in Collas were associated with enzyme expression levels [16] and resulting concentrations of arsenic metabolites in Collas. Lower levels of MMA were found in Collas compared to Bangladeshi [15], Chinese or Tibetans [18] exposed to permanently elevated arsenic levels in drinking water. Genes responsible for the metabolism of arsenic, therefore, may have been targets of strong positive selection among these populations. Levels of MMA and DMA have been recently associated with various SNPs near *AS3MT* in women from the Colla population of San Antonio de los Cobres in the Argentinean Puna region [19]. Moreover, an allele frequency based selection test applied on genome-wide genotype data in the same study suggested *AS3MT* as one of the main candidates of selection in this population.

In this study, we investigate the strength of the selection pressure exerted by elevated arsenic levels on the genome of a different subset of men and women from the Colla population from San Antonio de los Cobres and surrounding villages. We use two neighbouring groups, the Calchaquí and the Wichí as control populations. We also assessed genome-wide genotype data using distinct allele frequency based selection test and were able to confirm strong signatures within and near the *AS3MT* gene, thus underlining the key role of this gene in the adaptation to environmental arsenic.

2. Materials and Methods

2.1 Subjects and ethical approval

Individuals with indigenous ancestry from three regions of the Northwestern Argentinean province of Salta were recruited to participate in this study in April 2011: (1) Collas from the Andean Plateau or Puna (>3500 m), (2) Calchaquíes from Cachi in the Calchaquí valleys at 2300 m and (3) Wichí from the plains of the Gran Chaco region near Embarcación (Figure 1). We used our previously published data [20, 21] for 730,525 single nucleotide polymorphisms (SNPs) genotyped in 25 Collas (11 men, 14 women), 24 Calchaquíes (10 men, 14 women) and 24 Wichí (12 men, 12 women). In the Colla sample, 16 individuals were from San Antonio de los Cobres, where arsenic levels reach 214 µg/l [14]; 7 were from Tolar Grande with an arsenic level of 4 µg/l and one individual was from Olacapato, where arsenic levels are 12 µg/l. Arsenic concentrations for the exact sampling locations in the Gran Chaco regions were not available, however in surrounding locations arsenic concentrations were measured to be: Las Varas 0 µg/l, Pinchanal 19.5 µg/l, General Ballivián 4 µg/l, Tartagal 2.3 µg/l [22]. Concentrations in Cachi (Río Las Trancas) were 3.1 µg/l [22].

Only healthy unrelated adults who gave written informed consent were included in the study. The study was approved by the University of East Anglia Research Ethics Committee, the Ministry of Health of the Province of Salta (Ministerio de Salud Pública, Salta, Argentina) and the University of Cambridge's Human Biology Research Ethics Committee (HBREC.2011.01).

Figure 1: Sampling locations and arsenic levels in the province of Salta, Argentina

Stars denote sampling locations, circles levels of arsenic. Sampling locations of the Wichí population in the Gran Chaco region are (left to right): Embarcación, Carboncito, Misión Chacheña, Dragones (purple stars). Arsenic concentrations in surrounding locations were measured to be: Las Varas 0 µg/l, Pinchanal 19.5 µg/l, General Ballivián 4 µg/l, Tartagal 2.3 µg/l [22]. Calchaquíes originated from Cachi (turquoise star) with an arsenic level of 3.1 µg/l [22]. Collas (pink stars) were sampled in: San Antonio

de los Cobres (arsenic level: 214 μg/l), Tolar Grande (arsenic level: 4 μg/l) and Olacapato (arsenic level of 12 μg/l) [14].

2.2 Genotype data analysis

In total, 726,090 SNPs passed a genotype call rate of >98% and were included in downstream analyses [20, 21]. Two tests for positive selection were employed to analyse genome-wide signatures of arsenic adaptation. The pairwise fixation index (F_{ST}) was used as a measure of population differentiation [23] between Collas and Wichí, and between Calchaquí and Wichí using the programme GENEPOP [24]. We defined genomic windows of 200 kb and used maximal F_{ST} values to rank them. Only the top 1% was considered for analyses. Because the direction of the pairwise F_{ST} signatures cannot be determined (i.e. the signal can be due to extreme allele frequencies in either of the two populations), we also used the population branch statistic (PBS) to pinpoint allele differentiation to the population of interest [25, 26]. PBS is based on pairwise F_{ST} of three populations. Collas and Calchaquíes were each compared to Wichí and Eskimos [27]. Eskimos were chosen as the closest non-American outgroup genotyped on the same genotyping platform as Collas, Calchaquíes and Wichí. They originated from Novoe Chaplino, Chukotka Autonomous Okrug in Northeast Siberia [27]. PBS was calculated following Yi et al. [25] using a modified approach from Pickrell et al. [28] for 100 kb windows ranked by maximum PBS values [21]. Regional analysis of linkage disequilibrium was carried out with HaploView 4.2 [29].

As the first step in functional interpretation of the results of selection scanning, we compiled a list of genes known to be involved in arsenic metabolism. We included genes from three different sources: a) from the Gene Ontology (GO) database AmiGO we extracted genes that matched the search keyword 'arsen' to include metabolites of arsenic such as arsenate and arsenite [30]; b) from the gene information database GeneCards [31] we extracted genes associated with any compound containing the keyword 'arsen'; c) additional methyltransferases were extracted from literature [15, 32]. The final candidate gene list consisted of 35 unique genes (Table A.1). The selection test results were subsequently screened for these 35 candidate genes of arsenic metabolism.

Allele frequency differences between the three populations were assessed with One Way Analysis of Variance (ANOVA) implemented in the Statistical Package for Social Sciences (SPSS), Version 20.

3. Results

We conducted whole genome scans in Collas and Calchaquíes to identify genetic loci that showed higher than genome-wide average allelic differences between populations (F_{ST} and PBS tests). These scans highlighted the arsenic methyltransferase (*AS3MT*) gene as being highly differentiated in the Colla population. The gene was among the top 15 windows in PBS of Colla highlanders (Figure 2) and among the top 40 windows of pairwise F_{ST} between Collas and Wichí.

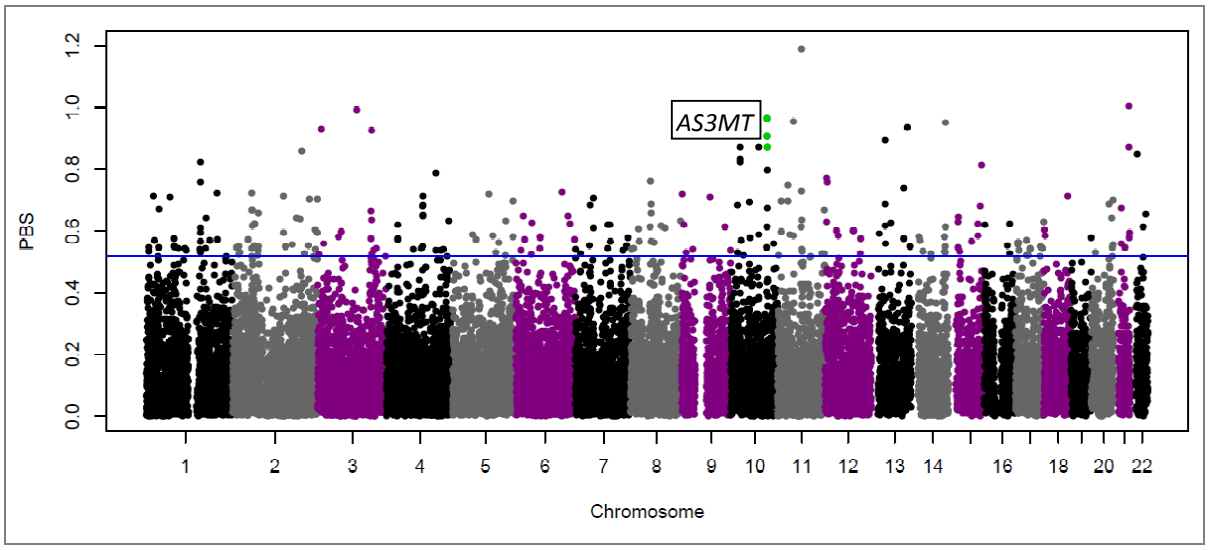

Figure 2: Window PBS scores across the genome in Collas

The blue line indicates the top 1% of hits. The fourth highest cluster overall is found on chromosome 10. The green circles indicate the PBS hits ± 1 Mb of *AS3MT*. The highest scoring SNP overall lies on chromosome 11 falls within a gene free region. The hit on chromosome 21 is located within *CBS*, which regulates cerebral blood flow velocity.

The pairwise F_{ST} signal was exclusively driven by two SNPs, one within the *AS3MT* gene (rs1046778, F_{ST}=0.606) and another one 1 kb upstream of *AS3MT* (rs7085104, F_{ST}=0.564; genome-wide mean F_{ST}=0.041). Specific variants of these alleles have been associated with beneficial arsenic metabolism [15]. The C allele of the T/C SNP rs1046778 was more frequent in Collas than in Wichí (74% and 8% respectively). The G allele of the G/A SNP rs7085104 was also prevalent in Collas (78% compared to 15% in Wichí). This is consistent with previously reported frequencies for these alleles [15], which have been associated with overall decreased expression of *AS3MT* and lower excreted MMA levels [15]. Engström and colleagues showed a 175% increase of *AS3MT* expression in homozygous carriers of the T allele at the rs1046778 locus compared to homozygotes of the C allele. Overall, 92% of Collas were at least heterozygous for the C and G allele on the same chromosomal strand corresponding to both functionally advantageous alleles (Figure 3). The percentage of homozygotes for both beneficial alleles is decreased in individuals from the Calchaquí valley, but not significantly. However, allele frequencies in both Collas and Calchaquíes differed significantly from Wichí ($p< 0.001$, ANOVA). A recent study using a dataset with greater SNP density, however, could not identify these two previously highlighted SNPs among the top 20 SNPs associated with MMA or DMA concentrations in 124 women from San Antonio de los Cobres [19].

In agreement with our F_{ST} results, PBS comparisons of Collas, Wichí and Eskimos highlighted a window containing *AS3MT* and two neighbouring genes, *CNNM2* and *WBP1L* (Table A.2). However, this test identified a different set of SNPs than F_{ST} in the surrounding region of *AS3MT*. The SNP (rs12221064) nearest to the gene region identified by PBS was located 15 kb downstream of *AS3MT* within *CNNM2* (Table A.2) and ranked 11th. Other high-ranking SNPs included rs17115100, within *CYP17A1*, 38 kb upstream of *AS3MT* (ranking 4th), and rs11191514 within *CNNM2,* 112 kb

downstream (ranking 10th) (Table A.2). The recent study by Schlebusch and colleagues associated rs17115100 and rs11191514 with percentage of MMA in urine and rs17115100 also with percentage of DMA in urine [19]. The allele frequency F_{ST} based selection test used by these authors (LSBL, locus specific branch length test) also highlighted *AS3MT* as top candidate of selection in the Colla population with Peruvians and Colombians as control populations. We previously reported haplotype based selection tests in Collas [21] but *AS3MT* was not among the top 1% haplotypes. However, a regional haplotype analysis 1 Mb up and downstream of *AS3MT* identified a haplotype block of 499 kb containing *AS3MT* (Figure A.1).

We repeated the PBS test using a neighbouring population to the Collas, the Calchaquíes, comparing it to Wichí and Eskimos. This test identified the same upstream SNP (rs17115100), albeit the SNP containing window ranked much lower (50th). While the top 1% F_{ST} results from Calchaquíes lacked *AS3MT*, it contained another gene from the candidate gene list, the cyclin-dependent kinase inhibitor 1A (*CDKN1A*) gene (rank 60). This kinase inhibitor is a modulator of the cell cycle and was inferred by orthologs to respond to an arsenic-containing substance (GO:0046685, evidence: inferred through electronic annotation).

AS3MT was the only of the 35 arsenic candidate genes (Table A.1) showing a signature of selection with two selection tests in the same population.

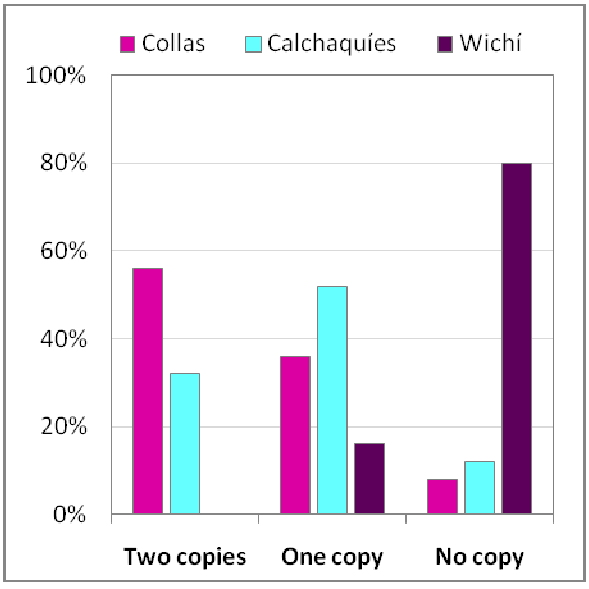

Figure 3: Distribution of beneficial CG alleles in Argentinean populations

The majority of Collas has two copies of the beneficial CG alleles (rs1046778, rs7085104) within and near *AS3MT*, while Calchaquíes mainly carry one copy of the specific alleles. In Wichí most individuals have no copy of the beneficial alleles. Allele frequencies differ significantly between Collas and Wichí and Calchaquíes and Wichí ($p<0.001$).

4. Discussion

High concentrations of arsenic in drinking water represent a strong environmental stressor, driving significant adaptive change in the highland populations of the Argentinean Puna. In this study, *AS3MT* was identified by our genome-wide scans as the main outcome of positive selection. Alleles

within or nearby this gene are highly differentiated and appear within the top 1% of ca. 13,000 windows across the genome. *AS3MT* had not been identified previously the top 1% of two haplotype based tests (integrated haplotype score, iHS and cross population extended haplotype homozygosity, XP-EHH) in Collas [21]. However, the minimum SNP density for iHS in a 200 kb window was not reached in the respective window containing the gene; therefore, no iHS test statistic could be calculated. XP-EHH neither highlighted the respective window as a particular long high frequency haplotype [21].

Thus, the selection signature of *AS3MT* was not identified by our previous haplotype based tests [21] but only by allele frequency based tests. Though a similar study also failed to identify a strong selection signal with iHS, it reported the average iHS values in a 1 Mb window around *AS3MT t*o be among the top 3%. In both studies, allele frequency based tests lead to more conclusive results, suggesting selection from standing variation in the ancestral population prior to the exposure to high arsenic concentrations. The alleles identified by our present study have been functionally evaluated and associated with reduced MMA concentrations in the Colla population of San Antonio de los Cobres [15, 19]. High concentrations of MMA are associated with arsenic related diseases [12]; thus, the metabolism of Argentine Puna inhabitants seems fine-tuned to reduce toxic MMA [9]. Only *AS3MT* could be highlighted from the arsenic candidate gene list by two selection tests using a genome-wide genotype approach. An alternative arsenic methyltransferase *N6AMT1*, which was also associated with lower MMA in Collas [33], did not reach genome-wide significance (data not shown).

The findings of our study are therefore well in agreement with a previous recent report [19] suggesting selection pressure from arsenic water in the Colla population, albeit analysing different individuals, using distinct control populations and different F_{ST} based selection tests (PBS and F_{ST} instead of LSBL). Alleles both within and around *AS3MT* appear to be the target of strong positive selection. The SNPs around *AS3MT* could be in linkage with a regulatory or functional variant or could itself influence *AS3MT* expression. An analysis of the region revealed a haplotype block of approximately 499 kb around the gene region (Figure A.1), thus suggesting selection of surrounding SNPs. Schlebusch et al. [19] also highlighted selection signatures outside the coding region of the *AS3MT* gene. Whole genome scans have the potential to reveal more distantly located loci with functional relevance, which may be overlooked by targeted resequencing of specific gene regions. Besides reporting strong signatures around *AS3MT*, we also highlighted adjacent genes*,* such as *CMMN2* or *CYP17A1,* and cannot unequivocally exclude that these may also contribute in particular to the PBS selection signal. However, considering the high F_{ST} scores within *AS3MT*, the functional relevance of this gene in the arsenic metabolism and association of overrepresented alleles in Collas with its expression [15], *AS3MT* is a likely candidate of selection. Nevertheless functional *in vitro* and *in vivo* studies of alleles are necessary for a more conclusive interpretation. In this regard, it is worth noting that the

neighbouring genes *CNNM2* and *WPB1L* (Table A.2), have been shown to be differentially methylated in the Colla population [16]. Since methylation reduces gene expression, a decreased level of the arsenic methyltransferase in peripheral blood was observed [16]. The reduced expression of this enzyme is associated with lower levels of MMA [15] and thus most likely beneficial in an environment with elevated arsenic concentrations.

It is interesting to note, that the F_{ST} values for the two highlighted alleles within 1 kb upstream of the *AS3MT* gene were 10 fold higher (0.606 and 0.564) than the gene's average F_{ST} of 0.053 calculated in another study, which compared Collas to indigenous Peruvians [17]. This underlines the extreme allele differentiation of two functionally associated SNPs compared to the complete gene region.

Significant differences in the allele frequency of *AS3MT* were also observed between Calchaquíes, Wichí and Eskimos, even though arsenic levels in ground water in the Calchaquí region are lower than those in the Puna [22]. The selection signature of *AS3MT* ranks lower in Calchaquíes than in Collas albeit still among the top 1%, thus implying either a reduced selection pressure in the Calchaquí population or gene flow from Collas [20]. Calchaquíes also show a selection signature around *CDKN1A*, as indicated by pairwise F_{ST}, although this signature is less strong than that of *AS3MT* in Collas. The functional significance of this cell cycle regulator for arsenic metabolism remains to be clarified.

In summary, our study confirms previous claims that positive selection has shaped allele frequencies of *AS3MT* to allow adaptation to the extremely toxic element arsenic [19]. We show signatures of positive selection driving allele frequencies in Collas and, to a smaller degree, in the neighbouring Calchaquí population. Selected alleles have enabled these populations to thrive for thousands of years despite their constant exposure to high levels of arsenic in drinking water.

5. Conclusion

The toxicant arsenic was shown to shape allele frequencies of the main arsenic methyltransferase in Argentinean Collas and Calchaquíes. This study confirms recent findings highlighting the strong selection pressure of the environmental carcinogen arsenic at a genome-wide level. This suggests that natural selection has given carriers of beneficial alleles higher reproductive success to thrive despite the daily consumption of high levels of arsenic.

Conflict of interest statement

The authors declare that there are no conflicts of interests.

Funding sources

This work was supported by European Research Council Starting Investigator (FP7-261213, TK), a starting investigator grant from the University of East Anglia (RC-158, MM), a Young Explorers Grant from the National Geographic Society (8900-11, CE) and a Sir Henry Wellcome Postdoctoral Fellowship (WT100066MA, TA). The funding bodies had no influence on the study design or analysis, data interpretation or article preparation.

Acknowledgements

We greatly appreciate the support of the Ministry of Health of the Province of Salta, Argentina and local hospital authorities for facilitating the data collection. We are particularly indebted to the people of Cachi, the Puna and the Gran Chaco region for their generous participation in this study.

References

[1] B. van Halem, S.A., G.L. Amy, J.C. van Dijk, Arsenic in drinking water: a worldwide water quality concern for water supply companies, Drink. Water Eng. Sci., 2 (2009) 29-34.
[2] M. Azizur Rahman, H. Hasegawa, M. Mahfuzur Rahman, M.A. Mazid Miah, A. Tasmin, Arsenic accumulation in rice (Oryza sativa L.): human exposure through food chain, Ecotoxicol Environ Saf, 69 (2008) 317-324.
[3] D.K. Nordstrom, Public health. Worldwide occurrences of arsenic in ground water, Science, 296 (2002) 2143-2145.
[4] WHO, Researchers warn of impending disaster from mass arsenic poisoning, Press Release2000.
[5] R. Raqib, S. Ahmed, R. Sultana, Y. Wagatsuma, D. Mondal, A.M. Hoque, B. Nermell, M. Yunus, S. Roy, L.A. Persson, S.E. Arifeen, S. Moore, M. Vahter, Effects of in utero arsenic exposure on child immunity and morbidity in rural Bangladesh, Toxicology letters, 185 (2009) 197-202.
[6] R.C. Fry, P. Navasumrit, C. Valiathan, J.P. Svensson, B.J. Hogan, M. Luo, S. Bhattacharya, K. Kandjanapa, S. Soontararuks, S. Nookabkaew, C. Mahidol, M. Ruchirawat, L.D. Samson, Activation of inflammation/NF-kappaB signaling in infants born to arsenic-exposed mothers, PLoS Genet, 3 (2007) e207.
[7] S.F. Farzan, S. Korrick, Z. Li, R. Enelow, A.J. Gandolfi, J. Madan, K. Nadeau, M.R. Karagas, In utero arsenic exposure and infant infection in a United States cohort: a prospective study, Environmental research, 126 (2013) 24-30.
[8] J. Liaw, G. Marshall, Y. Yuan, C. Ferreccio, C. Steinmaus, A.H. Smith, Increased childhood liver cancer mortality and arsenic in drinking water in northern Chile, Cancer Epidemiol Biomarkers Prev, 17 (2008) 1982-1987.
[9] M. Vahter, G. Concha, B. Nermell, R. Nilsson, F. Dulout, A.T. Natarajan, A unique metabolism of inorganic arsenic in native Andean women, Eur J Pharmacol, 293 (1995) 455-462.
[10] S. Lin, Q. Shi, F.B. Nix, M. Styblo, M.A. Beck, K.M. Herbin-Davis, L.L. Hall, J.B. Simeonsson, D.J. Thomas, A novel S-adenosyl-L-methionine:arsenic(III) methyltransferase from rat liver cytosol, J Biol Chem, 277 (2002) 10795-10803.
[11] M.N. Hall, M.V. Gamble, Nutritional manipulation of one-carbon metabolism: effects on arsenic methylation and toxicity, Journal of toxicology, 2012 (2012) 595307.

[12] A.H. Smith, C.M. Steinmaus, Health effects of arsenic and chromium in drinking water: recent human findings, Annual review of public health, 30 (2009) 107-122.
[13] G. Concha, B. Nermell, M. Vahter, Spatial and temporal variations in arsenic exposure via drinking-water in northern Argentina, Journal of health, population, and nutrition, 24 (2006) 317-326.
[14] G. Concha, K. Broberg, M. Grander, A. Cardozo, B. Palm, M. Vahter, High-level exposure to lithium, boron, cesium, and arsenic via drinking water in the Andes of northern Argentina, Environ Sci Technol, 44 (2010) 6875-6880.
[15] K. Engström, M. Vahter, S.J. Mlakar, G. Concha, B. Nermell, R. Raqib, A. Cardozo, K. Broberg, Polymorphisms in arsenic(+III oxidation state) methyltransferase (*AS3MT*) predict gene expression of *AS3MT* as well as arsenic metabolism, Environ Health Perspect, 119 (2012) 182-188.
[16] K.S. Engström, M.B. Hossain, M. Lauss, S. Ahmed, R. Raqib, M. Vahter, K. Broberg, Efficient arsenic metabolism-the *AS3MT* haplotype is associated with DNA methylation and expression of multiple genes around AS3MT, PLoS One, 8 (2013) e53732.
[17] C.M. Schlebusch, C.M. Lewis, Jr., M. Vahter, K. Engström, R.Y. Tito, A.J. Obregón-Tito, D. Huerta, S.I. Polo, A.C. Medina, T.D. Brutsaert, G. Concha, M. Jakobsson, K. Broberg, Possible positive selection for an arsenic-protective haplotype in humans, Environ Health Perspect, 121 (2013) 53-58.
[18] S. Fu, J. Wu, Y. Li, Y. Liu, Y. Gao, F. Yao, C. Qiu, L. Song, Y. Wu, Y. Liao, D. Sun, Urinary arsenic metabolism in a Western Chinese population exposed to high-dose inorganic arsenic in drinking water: influence of ethnicity and genetic polymorphisms, Toxicol Appl Pharmacol, 274 (2014) 117-123.
[19] C.M. Schlebusch, L.M. Gattepaille, K. Engström, M. Vahter, M. Jakobsson, K. Broberg, Human Adaptation to Arsenic-Rich Environments, Mol Biol Evol, (2015).
[20] C.A. Eichstaedt, T. Antao, A. Cardona, L. Pagani, T. Kivisild, M. Mormina, Genetic and Phenotypic Differentiation of an Andean Intermediate Altitude Population, Phys Reports, in press (2015).
[21] C.A. Eichstaedt, T. Antao, L. Pagani, A. Cardona, T. Kivisild, M. Mormina, The Andean Adaptive Toolkit to Counteract High Altitude Maladaptation: Genome-wide and Phenotypic Analysis of the Collas, PLoS One, 9 (2014) e93314.
[22] Centro de Ingeniería en Medio Ambiente del Instituto Tecnológico de Buenos Aires, Map of arsenic levels in Argentina [Mapa de arsenico en Argentina - Resultados recopilados de análisis de arsénico], La Comisión de Aguas, NutiRed.org, ITBA, TECHO 2015.
[23] B.S. Weir, C.C. Cockerham, Estimating F-statistics for the analysis of population structure, Evolution, 38 (1984) 1358-1370.
[24] F. Rousset, Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux, Mol Ecol Resour, (2008) 103-106.
[25] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z.X. Cuo, J.E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T.S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, H. Yang, R. Nielsen, J. Wang, Sequencing of 50 human exomes reveals adaptation to high altitude, Science, 329 (2010) 75-78.
[26] M.D. Shriver, R. Mei, A. Bigham, X. Mao, T.D. Brutsaert, E.J. Parra, L.G. Moore, Finding the genes underlying adaptation to hypoxia using genomic scans for genetic adaptation and admixture mapping, Adv Exp Med Biol, 588 (2006) 89-100.
[27] A. Cardona, L. Pagani, T. Antao, D.J. Lawson, C.A. Eichstaedt, B. Yngvadottir, M.T. Shwe, J. Wee, I.G. Romero, S. Raj, M. Metspalu, R. Villems, E. Willerslev, C. Tyler-Smith, B.A. Malyarchuk, M.V. Derenko, T. Kivisild, Genome-wide analysis of cold adaptation in indigenous siberian populations, PLoS One, 9 (2014) e98076.
[28] J.K. Pickrell, G. Coop, J. Novembre, S. Kudaravalli, J.Z. Li, D. Absher, B.S. Srinivasan, G.S. Barsh, R.M. Myers, M.W. Feldman, J.K. Pritchard, Signals of recent positive selection in a worldwide sample of human populations, Genome Res, 19 (2009) 826-837.

[29] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics, 21 (2005) 263-265.
[30] S. Carbon, A. Ireland, C.J. Mungall, S. Shu, B. Marshall, S. Lewis, AmiGO: online access to ontology and annotation data, Bioinformatics, 25 (2009) 288-289.
[31] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel, D. Lancet, GeneCards Version 3: the human gene integrator, Database : the journal of biological databases and curation, 2010 (2010) baq020.
[32] X. Ren, M. Aleshin, W.J. Jo, R. Dills, D.A. Kalman, C.D. Vulpe, M.T. Smith, L. Zhang, Involvement of N-6 adenine-specific DNA methyltransferase 1 (*N6AMT1*) in arsenic biomethylation and its role in arsenic-induced toxicity, Environ Health Perspect, 119 (2012) 771-777.
[33] F. Harari, K. Engström, G. Concha, G. Colque, M. Vahter, K. Broberg, N-6-adenine-specific DNA methyltransferase 1 (N6AMT1) polymorphisms and arsenic methylation in Andean women, Environ Health Perspect, 121 (2013) 797-803.

Appendix

Table A.1: Arsenic detoxification associated candidate genes

Source	Gene	Name
AmiGO: "arsen" associated ontology	*ABCC2*	ATP-Binding Cassette Sub-Family C Member 2
	AS3MT	Methylarsonite Methyltransferase (Arsenite Methyltransferase)
	ASNA1	Arsa (Bacterial) Arsenite Transporter, Atp-Binding, Homolog 1
	CDKN1A	Cyclin-Dependent Kinase Inhibitor 1
	CPEB2	Cytoplasmic Polyadenylation Element-Binding Protein 2
	CPOX	Coproporphyrinogen Oxidase
	CYP1A1	Cytochrome P1-450, Dioxin-Inducible
	DDX3X	DEAD (Asp-Glu-Ala-Asp) Box Helicase 3, X-Linked
	FECH	Errochelatase
	GCLC	Glutamate-Cysteine Ligase, Catalytic Subunit
	GLRX2	Glutaredoxin 2
	GSTO1	Glutathione S-Transferase Omega-1 (Monomethylarsonic Acid Reductase)
	GSTO2	Glutathione S-Transferase Omega-2 (Monomethylarsonic Acid Reductase)
	HMOX1	Heme Oxygenase 1
	MKNK2	MAP Kinase Interacting Serine/Threonine Kinase 2
	PPIF	Peptidylprolyl Isomerase F (Cyclophilin F)
	PTEN	Phosphatidylinositol-3,4,5-Trisphosphate 3-Phosphatase
	RBM4	RNA-Binding Motif Protein 4
	RNF4	RING Finger Protein 4
	SLC34A1	Solute Carrier Family 34 Member 1
	SRRT	Serrate RNA Effector Molecule Homolog (Arsenite-Resistance Protein 2)
	TNFRSF11B	Tumor Necrosis Factor Receptor Superfamily Member 11B
	UROS	Uroporphyrinogen-III Synthase
	ZFAND1	Zinc Finger, AN1-Type Domain 1
	ZFAND2A	Zinc Finger, AN1-Type Domain 2A (Arsenite Inducible RNA Associated Protein)
	ZFAND2B	Zinc Finger, AN1-Type 2B (Arsenite-Inducible RNA-Associated Protein-Like Protein)
Gene cards: "arsen"	*METTL18*	Methyltransferase-Like Protein 18 (Arsenic-Transactivated Protein 2)
	POLE3	Polymerase (DNA Directed), Epsilon 3 (Arsenic-Transactivated Protein)

associated gene name	*SERPINH1*	Serine (Or Cysteine) Proteinase Inhibitor, Clade H (Arsenic-Transactivated Protein 3)
	SPDL1	Spindly Homolog (Drosophila) (Arsenite-Related Gene 1 Protein)
Literature: arsenic associated methyl-transferases	*BHMT*	Betaine-Homocysteine S-Methyltransferase [15]
	DNMT1	DNA (Cytosine-5-)-Methyltransferase 1 [15]
	DNMT3B	DNA (Cytosine-5)-Methyltransferase 3B [15]
	N6AMT1	N-6 Adenine-Specific DNA Methyltransferase 1 [32]
	PEMT	Phosphatidylethanolamine N-Methyltransferase [15]

Table A.2: Top 15 windows of PBS in Collas

Rank	Genes in window	Window location	PBS	Max. score position
1	*ACY3, ALDH3B2, TBX10*	11: 67400000 - 67500000	1.188	23 kb downstream *ALDH3B2*
2	*CBS, MX2, PKNOX1*	21: 44400000 - 44500000	1.005	Within *CBS* and *MX2*
3	*HRH1, RP11-572M11.3, SNORD112*	3: 112800000 - 112900000	0.993	Within *HRH1*
4	*CYP17A1, CYP17A1-AS1, WBP1L*	10: 104500000 - 104600000	0.965	Within *CYP17A1*, 38 kb upstream of *AS3MT*
5	*TSPAN18*	11: 44700000 - 44800000	0.955	27 kb upstream of *TSPAN18*
6	*RN7SL710P*	14: 97900000 - 98000000	0.950	36 kb upstream of *RN7SL710P*
7	No gene	13: 103800000 - 103900000	0.934	n.a.
8	*ATP2B2, ATP2B2-IT1, ATP2B2-IT2*	3: 10600000 - 10700000	0.929	Within *ATP2B2*
9	*PLCH1, PLCH1-AS1*	3: 155100000 - 155200000	0.925	Within *PLCH1*
10	*CNNM2*	10: 104700000 - 104800000	0.906	Within *CNNM2*
11	*CNNM2, C10orf32,* ***AS3MT***	10: 104600000 - 104700000	0.906	Within *CNNM2*, 15 kb downstream of *AS3MT*
12	*LHFP*	13: 40100000 - 40200000	0.893	Within *LHFP*
13	*CRYAA, U2AF1*	21: 44500000 - 44600000	0.872	Within *MX2*
14	*LINC00856*	10: 80300000 - 80400000	0.871	Within *LINC00856*
15	*PDSS1*	10: 26900000 - 27000000	0.870	617 bp upstream of *PDSS1*

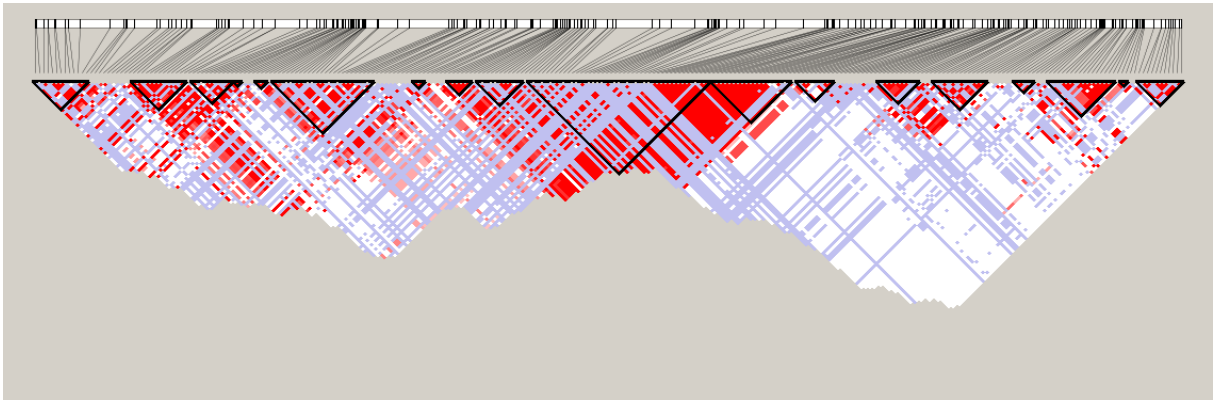

Figure A.1: Haplotype analysis of SNPs ± 1 Mb around *AS3MT*

The region includes SNPs located 1 Mb up and downstream of *AS3MT* on chromosome 10 between positions 104,629,210 - 104,661,656. The central largest triangle includes *AS3MT* and spans 499 kb.

SNPs are represented on the top and the linkage disequilibrium (LD) degree is displayed below by colour. Red squares represent strong LD (D'= 1) with a high logarithm of odds (LOD) score, i.e. the probability of linkage between 2 loci is very high; purple: high D', low LOD score; white: low D' and low LOD score. Black triangles indicate haplotype blocks calculated by the programme Haploview [29]. Only heterozygous SNPs are displayed.