



HAL
open science

Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in Central Asia

Laure Ségurel, Frédéric Austerlitz, Bruno Toupance, Mathieu M. Gautier, Joanna L Kelley, Patrick Pasquet, Christine Lonjou, Myriam Georges, Sarah Voisin, Corinne Cruaud, et al.

► To cite this version:

Laure Ségurel, Frédéric Austerlitz, Bruno Toupance, Mathieu M. Gautier, Joanna L Kelley, et al.. Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in Central Asia. *European Journal of Human Genetics*, 2013, 21 (10), pp.1146-1151. 10.1038/ejhg.2012.295 . hal-02120826

HAL Id: hal-02120826

<https://hal.science/hal-02120826>

Submitted on 6 May 2019

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

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

ARTICLE

Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in Central Asia

Laure Ségurel^{1,2}, Frederic Austerlitz¹, Bruno Toupance¹, Mathieu Gautier³, Joanna L Kelley², Patrick Pasquet¹, Christine Lonjou⁴, Myriam Georges¹, Sarah Voisin¹, Corinne Cruaud⁵, Arnaud Couloux⁵, Tatyana Hegay⁶, Almaz Aldashev⁷, Renaud Vitalis^{1,3,8,9} and Evelyne Heyer^{1,9}

The high prevalence of type 2 diabetes and its uneven distribution among human populations is both a major public health concern and a puzzle in evolutionary biology. Why is this deleterious disease so common, while the associated genetic variants should be removed by natural selection? The ‘thrifty genotype’ hypothesis proposed that the causal genetic variants were advantageous and selected for during the majority of human evolution. It remains, however, unclear whether genetic data support this scenario. In this study, we characterized patterns of selection at 10 variants associated with type 2 diabetes, contrasting one herder and one farmer population from Central Asia. We aimed at identifying which alleles (risk or protective) are under selection, dating the timing of selective events, and investigating the effect of lifestyle on selective patterns. We did not find any evidence of selection on risk variants, as predicted by the thrifty genotype hypothesis. Instead, we identified clear signatures of selection on protective variants, in both populations, dating from the beginning of the Neolithic, which suggests that this major transition was accompanied by a selective advantage for non-thrifty variants. Combining our results with worldwide data further suggests that East Asia was particularly prone to such recent selection of protective haplotypes. As much effort has been devoted so far to searching for thrifty variants, we argue that more attention should be paid to the evolution of non-thrifty variants.

European Journal of Human Genetics (2012) 0, 000–000. doi:10.1038/ejhg.2012.295

Keywords: type 2 diabetes; genetic adaptation; Central Asia; thrifty genotype; human evolution

INTRODUCTION

During the Neolithic Revolution (8500–2500 BC), many human populations shifted from a nomadic hunter-gathering to a farming or herding mode of subsistence.¹ This transition was accompanied by dramatic changes in diet and likely reshuffled selective pressures acting on metabolic genes, presumably with local adaptations in populations with contrasted lifestyles (for a review, see Brown²). More recently, in many populations, the Industrial Revolution yielded new nutritional and cultural changes, which were accompanied by an important rise in the prevalence of metabolic disorders like type 2 diabetes (T2D). T2D, a major public health problem,³ is characterized by an elevation of blood glucose levels due to prolonged impaired insulin secretion and/or insulin resistance, the low efficiency of insulin to store glucose into cells where it is later used for energy. To date, around 40 genetic variants have been associated with polygenic T2D.^{4,5} These causal variants are expected to reduce fitness even if the mean age of onset for T2D is relatively late. Indeed, as observed in other diseases,⁶ the variance in age of onset could lead to enough cases occurring during reproductive life to compromise the fitness of

individuals carrying a risk allele. Furthermore, causal variants of T2D may also affect fitness indirectly, through their association with other syndromes occurring earlier in life, notably insulin resistance and cardiovascular diseases⁷ or gestational diabetes.⁸ As a result, one would expect these variants to be selected against by natural selection, and therefore the high prevalence of T2D as well as its variability among populations has remained a puzzle in evolutionary biology.⁹

Several hypotheses have been put forward to explain this paradox. Most of them hypothesized a past evolutionary advantage to carry T2D risk variants. The ‘thrifty genotype’ hypothesis^{10–12} proposed that in populations facing strong food insecurity (hunter-gatherers, but also pre-industrialized farmers^{13,14}), insulin resistance was selected for because it ensured a more efficient use of available nutritional resources. On the other hand, the ‘carnivore connection’ hypothesis^{15,16} proposed that insulin resistance was selected for in herders and hunter-gatherers, but not in farmers, as an adaptive response to low-carbohydrate and protein-rich diet. Finally, the ‘variable disease selection’ hypothesis¹⁷ suggested that thrifty variants were selected for in response to infectious diseases,¹⁸ as the

¹Eco-Anthropologie et Ethnobiologie, Univ Paris Diderot, Sorbonne Paris Cité, Paris, France; ²Department of Human Genetics, University of Chicago, Chicago, IL, USA; ³INRA, UMR CBGP (INRA – IRD – CIRAD – Montpellier SupAgro), Campus International de Baillarguet, France; ⁴C2BiG (Centre de Bioinformatique/Biostatistique Génomique d’Île de France), Plateforme Post-génomique P3S, Hôpital Pitié Salpêtrière, Paris, France; ⁵Genoscope, Centre National de Séquençage, Evry, France; ⁶Uzbek Academy of Sciences, Institute of Immunology, Tashkent, Uzbekistan; ⁷Institute of Molecular Biology and Medicine, National Center of Cardiology and Internal Medicine, Bishkek, Kyrgyzstan; ⁸Institut des Sciences de l’Évolution, UMR 5554, CNRS, Université Montpellier 2, Place Eugène Bataillon, France

⁹These authors equally supervised this work.

*Correspondence: L. Ségurel, Department of Human Genetics, University of Chicago, CLSC 418, 920 East 58th Street, Chicago IL 60637, USA. Tel: (001)-773-834-8932; E-mail: lsegurel@uchicago.edu.

Received 8 May 2012; revised 30 October 2012; accepted 11 December 2012

Q1

Q2

immune response has energetic costs that, in turn, affect the metabolic syndrome. Although they rely on different mechanisms, these hypotheses all predict that the variants associated with the risk of T2D were favored in pre-Neolithic hunter-gatherer populations, as well as in some, but not all, post-Neolithic populations.

On the other hand, Allen and Cheer¹⁹ proposed that milk consumption in Europe was responsible for an increased uptake of glucose in the diet, providing an opportunity for the selection of protective (non-thrifty) variants. This mechanism would lead to the positive selection of protective variants in milk-consuming populations, which might explain the low prevalence of T2D in Europeans. Hancock *et al.*²⁰ proposed that genes associated with common metabolic diseases like T2D have been targeted by selection during adaptation to climate, leading to the increase in frequency of alleles with potentially opposing effects on disease susceptibility. Last, non-adaptationist theories have also been proposed, suggesting that drift alone might be responsible for the differences in prevalence observed between populations.^{21,22}

Which hypothesis do the genetic data support? A handful of studies have investigated the signatures of selection on candidate genes for T2D,^{23–29} but most of them did not assess whether the risk or the protective alleles were targeted by selection. On the contrary, Vander-Molen *et al.*³⁰ and Helgason *et al.*³¹ studies reported signatures of positive selection on protective haplotypes for T2D, the latter study providing evidence for an increase in frequency of these haplotypes ~8400 years ago. These results seemingly contradict the thrifty genotype hypothesis,³² but as each study was based on the analysis of a single gene (*CAPN10* and *TCF7L2*, respectively), they cannot be considered as definitive evidence. Hancock *et al.*²⁰ analyzed worldwide correlations between climate variables and allele frequencies at metabolic genes and found evidence of selection on both protective (in *LEPR* and *PON1*) and risk haplotypes (in *FABP2* and *EPHX2*). Klimentidis *et al.*³³ found that for three genes, selection targeted the risk haplotypes (in *IGF2BP2*, *WFS1* and *SLC30A8*) and that T2D-associated loci are highly differentiated on a worldwide scale, with Sub-Saharan Africans and East Asians being particularly prone to positive selection. Given the latter two studies did not provide estimates of the timing of the selective event, it remains unclear which evolutionary hypothesis is supported by the genetic data.

In this study, we investigated the selective pressures acting on genes associated with T2D in two neighboring populations with contrasted lifestyles and dietary habits. We concentrated on Central Asia, a region with high ethnic diversity, which has been the focus of a recent population genetics survey.³⁴ We studied a population of ancestrally nomadic herders (Kyrgyz) and a population of long-term agriculturalists (Tajiks), given that the main hypotheses to explain the paradox of the high prevalence of T2D proposed that some lifestyles and/or dietary habits could have provided a selective advantage to individuals carrying T2D risk variants. We analyzed re-sequencing data and genotyping data in genomic regions located around 10 T2D-associated mutations and in 20 presumably neutral regions in order to (i) investigate the selective patterns in one herder and one farmer population, controlling for the unknown demographic history; (ii) identify which allele (risk or protective) has been or still is targeted by selection; and (iii) infer the timing of onset of selection (before or after the Neolithic).

MATERIAL AND METHODS

DNA samples and re-sequencing

We collected DNA from 40 Kyrgyz individuals in Bishkek, Kyrgyzstan, as well as 39 Tajik individuals in Bukhara, Uzbekistan (see Supplementary Note 1). We

chose 10 mutations associated with T2D or related phenotypes, in *FABP2*, *PPARG*, *TCF7L2*, *LEPR*, *KCNJ11*, *SLC30A8*, *HHEX*, *CDKALI*, *KCNQ1* and *PON1* (see Supplementary Table S1 and Supplementary Note 2) and obtained re-sequencing data for ~1.1 kb regions around them. We also sequenced 20 presumably neutral regions, designed by Patin *et al.*³⁵ (average length: ~1.3 kb). PCR and sequencing reactions were performed as described in Supplementary Note 1 and Supplementary Table S2.

Sequence analyses and within-population neutrality tests

We used DnaSP version 5³⁶ to estimate several genetic diversity statistics and to compute neutrality tests based on the site-frequency spectrum (Tajima's D ,³⁷ Fu and Li's D and F ³⁸ and Fu's F_s ³⁹). Significance was tested by means of 10 000 coalescent simulations for each test statistics and each sequence, using DnaSP, assuming a large constant population size and a neutral infinite-sites model of mutation. The simulations were conditioned on the observed level of nucleotide diversity (θ) in each population, estimated as the average number of nucleotide differences between individuals per sequence and per population. We also computed Zeng *et al.*'s⁴⁰ E statistics and tested its significance in a similar manner as for the other tests using the java program kindly provided to us by K. Zeng. One-tailed P -values were computed as the probability of obtaining lower values than the ones observed, further transformed into $(1-2|P-0.5|)$, and then corrected for the false discovery rate (FDR) for multiple testing.⁴¹

Between-populations neutrality tests

Indices of population differentiation (F_{ST}) were computed with Genepop v.4.7.⁴² We used the software package Dfdist⁴³ to generate 1 000 000 coalescent simulations in a symmetrical 10-demes island model at migration-drift equilibrium,⁴⁴ conditional on the observed level of differentiation measured at the 147 presumably neutral SNPs ($F_{ST} = 0.006$). As Dfdist was originally designed for the analysis of bi-allelic, dominant markers (see, eg,⁴⁵), we modified it in order to simulate co-dominant, bi-allelic markers (see, eg,⁴⁶). The coalescent simulations were performed using $\theta = 2nN\mu = 0.2$ (where n is the number of demes of size N , and μ is the mutation rate), in order to match the observed overall gene diversity of the presumably neutral SNPs in the pooled sample ($H_e = 0.182$). We checked that the distribution of F_{ST} conditional on heterozygosity was robust to a range of alternative values (from $\theta = 0.02$ to $\theta = 2.0$; results not shown). One-tailed P -values were computed for the 10 candidate genic mutations as the probabilities of getting values as small or smaller than the ones observed, and corrected for the FDR in multiple testing.⁴¹

Haplotype-based tests of selection

We used the data from a companion paper, for which all individuals were genotyped using the microarray Human-660 W-Quad-V1.0 from Illumina (see Supplementary Note 3). SNPs were phased with the software fastPhase 1.4,⁴⁷ using population label information to estimate phased haplotypes. We computed the integrated haplotype scores (iHS) to compare the decay of homozygosity around the mutations of interest between the ancestral and the derived background, in each population, using the rehh package.⁴⁸ The iHS estimates were standardized per allelic frequency bins, using genome-wide data.⁴⁹ P -values for the 10 candidate genic mutations were corrected for the FDR in multiple testing.⁴¹

For mutations with a significant iHS value (corrected P -value < 0.05), we used Austerlitz *et al.*'s method⁵⁰ as implemented in their Mathematica⁵¹ notebook. This method provides a maximum-likelihood estimate of the time elapsed since the appearance of the mutation and its intrinsic growth rate, using the number of copies of the mutant allele in the population and the level of allelic association between this allele and one or several closely linked markers. Nine SNPs were chosen for that purpose on each side of each target SNP, at about 20, 35, 50, 75, 100, 125, 150, 200 and 250 kb, respectively, from the target mutation. We assumed a population size of 100 000 individuals for the computations and checked that alternative choices of population size (10 000 or 1 000 000) did not affect our results.

RESULTS

Genetic diversity

Genetic diversity statistics are provided in Table 1 for genic candidate sequences and in Supplementary Table S3 for presumably neutral sequences. Overall nucleotide diversity was not significantly different between presumably neutral and genic candidate sequences (in average 1×10^{-3} vs 1.5×10^{-3} , respectively, Wilcoxon's rank sum test, P -value = 0.29). The highest nucleotide diversity was observed in *FABP2*, in both populations: 4.1×10^{-3} in Kyrgyz and 4.3×10^{-3} in Tajiks. We tested for deviation from Hardy–Weinberg equilibrium for the 208 observed polymorphisms in each population. Four mutations, located in presumably neutral sequences, departed significantly from Hardy–Weinberg equilibrium (Chi-square test, P -value < 0.001). These polymorphisms were re-genotyped and confirmed by an independent PCR.

Neutrality tests within population

As shown in Supplementary Table S4, we found that only one genomic region departed from neutrality, namely the region around rs1799883 in *FABP2* (Tajima's D = 2.88, FDR-corrected P -value = 0.045) in Kyrgyz. No significantly positive values were found in the presumably neutral sequences. Given that the candidate mutations were associated to the phenotype of interest through genome-wide association studies (GWAS), they might be biased toward intermediate allelic frequencies, which might inflates Tajima's D values.⁵² This is indeed visible when we compare the allelic frequency spectrum of genic and presumably neutral sequences (compare Supplementary Figure S1a and 1b). In order to test whether such a bias could result in spurious signatures of selection, we ran the same tests on a subset of presumably neutral sequences with the same allele frequency spectrum as the target mutations (see Supplementary Figure S1c). None of the so-ascertained presumably neutral sequences departed from neutrality (see Supplementary Table S5), which suggests that the observed pattern in *FABP2* is unlikely to be caused by demographic history, but could rather be taken as evidence for selection.

Between-population tests

The 10 candidate mutations tended to be more differentiated on average than neutral regions (F_{ST} = 0.030 vs 0.006), although not significantly (Wilcoxon's rank sum test, P -value = 0.06). Using

Beaumont and Nichols⁴³ approach, we found two out of these 10 candidate mutations departing from neutral expectations (Figure 1, rs1137100 in *LEPR*; P -value = 0.02 and rs2237892 in *KCNQ1*; P -value = 0.01). This suggests that genetic variation at these mutations may have been affected by natural selection, most probably by differential local adaptation between Kyrgyz and Tajiks.

Haplotype-based tests

As shown in Table 2, we found a significant iHS score for *LEPR* in Kyrgyz (iHS = -2.7, P -value = 0.01). Using Austerlitz *et al.*'s⁵⁰ method, we found that this mutation started to increase in frequency 7500 ya (95% confidence interval (CI_{95%}): 6500–8900) in this population, with a growth rate of 1.027 (CI_{95%}: 1.020–1.040). The same trend was observed in Tajiks (iHS = -1.9, P -value = 0.04), even though the growth rate was much lower (1.010, CI_{95%}: 1.007–1.016, starting 14 400 ya, CI_{95%}: 11 900–17 600). We also observed a significantly positive iHS value for *HHEX* in both populations: iHS = 2.9 in Kyrgyz (P -value = 0.01) and 2.6 in Tajiks (P -value = 0.02). This selective event started around 10 500 ya (CI_{95%}: 8700–12 700) and 10 700 ya (CI_{95%}: 9000–13 100), respectively, in Kyrgyz and Tajiks, with respective growth rates of 1.027 (CI_{95%}: 1.020–1.040) and 1.021 (CI_{95%}: 1.018–1.032). We also found a signal of selection on *PON1* in Tajik: iHS = 2.0 (P -value = 0.04), which was not detected using Austerlitz *et al.*'s⁵⁰ method (very low growth rate of 1.005, CI_{95%}: 1.003–1.008, starting 43,000 ya, CI_{95%}: 35 500–52 200). In all the cases, the haplotypes targeted by positive selection carried the protective allele, which corresponded either to the derived (*LEPR*) or the ancestral allele (*HHEX*, *PON1*).

The mutation rs2237892 on *KCNQ1*, which was more differentiated between populations than expected under neutrality, did not present a significant iHS score in either population. However, we found two SNPs around rs2237892 with significant iHS values in Kyrgyz only (see Supplementary Note 5), suggesting that the differential selection inferred on *KCNQ1* with the F_{ST} -based tests likely results from recent selection acting in Kyrgyz only.

Table 1 Summary statistics of genetic diversity per genic region and per population

Gene	Kyrgyz (herders)					Tajiks (farmers)				
	S	Sin	π	H	Hdiv	S	Sin	π	H	Hdiv
<i>FABP2</i>	12	0	0.00413	5	0.703	14	1	0.00431	8	0.776
<i>PPARG</i>	1	0	0.00016	2	0.182	2	1	0.00018	3	0.172
<i>TCF7L2</i>	3	1	0.00031	4	0.188	2	0	0.00048	3	0.295
<i>LEPR</i>	4	0	0.00098	4	0.605	5	1	0.00099	5	0.723
<i>KCNJ11</i>	4	1	0.00119	4	0.679	6	3	0.00121	6	0.690
<i>SLC30A8</i>	8	4	0.00189	8	0.724	8	4	0.00191	7	0.697
<i>HHEX</i>	2	0	0.00076	2	0.478	2	0	0.00080	2	0.505
<i>CDKAL1</i>	10	1	0.00310	5	0.537	9	0	0.00244	4	0.411
<i>KCNQ1</i>	4	0	0.00062	5	0.522	4	0	0.00048	4	0.282
<i>PON1</i>	5	0	0.00153	4	0.685	6	1	0.00189	5	0.687
Average	5.3	0.7	0.00147	4.3	0.530	5.8	1.1	0.00147	4.7	0.524

Abbreviations: S, number of polymorphic sites; Sin, number of singletons; π , nucleotide diversity (number of pairwise differences between sequences per site); H, number of haplotypes; Hdiv, haplotype diversity.

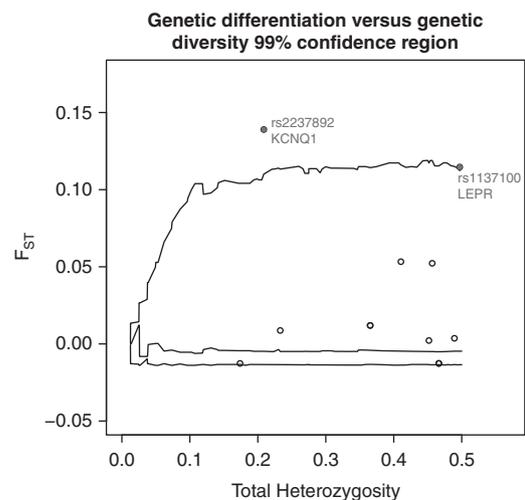


Figure 1 Genetic differentiation (F_{ST}) as a function of heterozygosity for 10 SNPs associated with the risk of type 2 diabetes. The upper and lower lines represent the 99% confidence region expected under neutrality (see Material and Methods), with the middle line showing the median genetic differentiation under neutrality. Each circle represents a SNP associated with the risk of type 2 diabetes. The names and corresponding genes of SNPs with a significant corrected P -value (≤ 0.05) are shown (excess of differentiation).

Table 2 Allelic frequencies, F_{ST} and iHS values for each variant in Kyrgyz and Tajiks from Central Asia, as well as in two populations from the HGDP-CEPH panel: ASN (individuals of Japanese and Chinese ancestries) and CEU (individuals of European ancestry)

Mutation	Gene	Anc/Der allele	Risk allele frequency in				F_{ST} Kyrgyz-Tajiks	P -val	F_{ST} ASN-CEU	iHS in Kyrgyz	P -val	iHS in Tajiks	P -val	iHS in ASN	iHS in CEU
			Kyrgyz	Tajiks	ASN	CEU									
rs1799883	<i>FABP2</i>	C/T ^R	0.39	0.31	0.31	0.38	0.002	0.44	0.005	NA		NA		-1.3	-1.2
rs1801282	<i>PPARG</i>	C ^R /G	0.90	0.91	0.96	0.93	0.009	0.91	0.006	NA		NA		NA	-0.6
rs7903146	<i>TCF7L2</i>	T ^R /C	0.10	0.17	0.02	0.25	-0.013	0.44	0.124	-1.3	0.09	-0.4	0.19	NA	0.1
rs1137100	<i>LEPR</i>	A ^R /G	0.41	0.66	0.17	0.66	0.115	0.02	0.244	-2.7	0.01	-1.9	0.04	-2.4	-1.4
rs5215	<i>KCNJ11</i>	T/C ^R	0.36	0.37	0.34	0.41	-0.013	0.91	0.005	-0.9	0.11	-1.2	0.08	0.1	-0.4
rs13266634	<i>SLC30A8</i>	C ^R /T	0.62	0.80	0.54	0.75	0.053	0.13	0.046	-1.2	0.09	-1.2	0.08	-1.2	-1.9
rs1111875	<i>HHEX</i>	T/C ^R	0.40	0.46	0.36	0.56	0.004	0.44	0.038	2.9	0.01	2.6	0.02	1.9	0.5
rs7754840	<i>CDKAL1</i>	G/C ^R	0.30	0.19	0.39	0.31	0.012	0.44	0.008	-1.4	0.09	1.1	0.09	-0.9	1.1
rs2237892	<i>KCNQ1</i>	C ^R /T	0.79	0.97	0.62	0.93	0.139	0.01	0.115	-0.2	0.22	-0.3	0.19	-0.6	0.0
rs3917498	<i>PON1</i>	G/T ^R	0.44	0.25	0.63	0.40	0.052	0.13	0.053	0.9	0.11	2.0	0.04	0.8	0.9

For each mutation, we determined the ancestral and derived alleles (given on the + strand) using the chimpanzee as outgroup. We further determined the risk allele (denoted ^R) for each candidate SNP, following the original studies that reported the association with T2D. The mutations of interest in *FABP2* and *PPARG* were not present on the genotyping array, so the iHS value could not be computed. Significant P -values are shown in bold (P -value ≤ 0.05); the P -values are shown corrected for multiple correction. Data for the ASN and CEU populations come from the Haplotter database based on HapMap Phase II data (<http://haplotter.uchicago.edu/>).

Intriguingly, we did not find signals of selection for the mutation rs7903146 in *TCF7L2*, although it represents the strongest signal of association with T2D to date in other populations and the most consistent signal of selection.^{20,31,33} However, the estimated growth rate of the protective allele in Kyrgyz was higher than that of the risk variant (1.027, $CI_{95\%}$: 1.019–1.040 vs 1.009, $CI_{95\%}$: 1.007–1.017) and pointed to a selective event starting 12 000 years ago. Tajiks did not show such a signal (growth rate of 1.018, $CI_{95\%}$: 1.013–1.026 for the protective vs 1.017, $CI_{95\%}$: 1.012–1.027 for the risk allele).

DISCUSSION

Characterizing the patterns of selection at genes associated with T2D in Central Asia

Using neutrality tests based on within-population diversity, haplotype structure, and between-population differentiation, we were able to identify complementary signals of selection among candidate genes for T2D. In particular, we found evidence for differential positive selection at rs1137100 in *LEPR*, at which (i) the F_{ST} between Kyrgyz and Tajiks was higher than expected under neutrality and (ii) the iHS statistic provided evidence for positive selection acting in Kyrgyz and to a lesser extent in Tajiks. We also found evidence for balancing selection (or a recent partial selective sweep or selection on standing variation⁵³) on *FABP2* in both populations, where we found a high nucleotide diversity and a significantly positive Tajima's D in Kyrgyz. Consistently, two mutations in *FABP2* presented the lowest F_{ST} estimates between Kyrgyz and Tajiks of the full data set ($F_{ST} = -0.014$).

Identifying the targeted alleles and the timing of selection

We did not find any evidence in Central Asia of *pre-Neolithic* selection favoring T2D risk variants, in contradiction with the thrifty genotype,¹⁰ the carnivore connection^{15,16} and the variable disease selection¹⁷ hypotheses. However, we cannot exclude that some of these variants were selected for in a distant past, as signatures of ancient selection might be difficult to detect by means of population genetic approaches. We did not find any evidence of *post-Neolithic* selection acting on T2D risk variants, which contradicts the thrifty genotype hypothesis (the more recent version of which considers that food insecurity is stronger in farming populations, where it should select for thrifty variants^{13,14}), as well the carnivore connection

hypothesis (which predicts that insulin resistance should be selected for in herders in response to their low-carbohydrate diet).

Contrastingly, we found signatures of positive selection of protective variants (either ancestral or derived) in both populations. Our analyses further showed that selection of protective variants occurred between 5500 and 12 000 years ago (depending on the gene considered), which corresponds to the earliest stages of the Neolithic Revolution. Our results therefore suggest that protective variants were selected for during and/or after this transition, which echoes previous studies that reported signals of positive selection of protective alleles for genes associated with risks of heart diseases⁵⁴ and hypertension.^{55,56}

Possible evolutionary scenarios

We identified footprints of selection that are likely to reflect a shift in the metabolic constraints accompanying the Neolithic transition, with protective alleles becoming advantageous. Interestingly, the *KCNQ1* protective haplotype (which frequency is higher in Kyrgyz than in Tajiks, see Table 2) has been shown to be at low frequency in populations where cereals are the main dietary component⁵⁷ and under recent selection in four out of seven pastoral populations from South Asia.³³ It seems therefore that this protective haplotype was recently targeted by selection in response to specific pastoral dietary habits. However, we have shown in this study that signals of selection toward protective variants have been found in both herders and farmers from Central Asia, suggesting that the same phenotype is favored in both the lifestyles. It could be that the input of cereals in the diet of farmers at the Neolithic is responsible for reshuffling the selective pressures on genes involved in glucose metabolism, while the consumption of milk in pastoral populations have led to similar major metabolic changes.¹⁹ Milk consumption is indeed widespread among Kyrgyz populations (even though the frequency of lactase persistency is low⁵⁸), as in the Hausa population, where a signal of selection has been detected in *CAPN10*.²³

On the other hand, evidence of recent selection for the protective haplotype at *LEPR* and *TCF7L2* (as we found in Kyrgyz) is also documented in other East Asian populations^{20,31,33,49} (see also the high F_{ST} and iHS values in ASN, Table 2), from which the Kyrgyz are genetically closer than the Tajiks.³⁴ These signatures of selection are, therefore, more likely to reflect a differential adaptation between East

Asians and other populations than between herders and farmers. Similarly, a strong differentiation at *HHEX* was found between East Asians and other groups³³ (see also the high *iHS* value in ASN, Table 2), a gene for which we showed that the protective mutation was under recent selection in both Kyrgyz and Tajiks. This suggests that selection might act for this gene at a broader geographical scale, encompassing both ethnic groups from Central Asia. These results point to a recent selection of protective haplotypes in Asia, which might be the result of a specific type of agriculture developed in this part of the world, or because of particular climatic and/or pathogenic conditions.

Perspectives

We acknowledge that our conclusions are based on a limited number of common variants, which do not necessarily represent the genetic architecture of T2D susceptibility as a whole. This complex disease is indeed only partially explained by the common variants identified so far, and further studies based on additional risk variants are now required to complete our knowledge of the evolution of T2D susceptibility.

Much effort has been devoted so far in the search for thrifty variants, and the thrifty genotype hypothesis has a deep impact into the therapeutic diet strategies adopted in modern societies to manage chronic diseases.^{59,60} Yet our results, along with those from other authors,^{20,29–31,33} support a radically different scenario, in which protective (non-thrifty) haplotypes have been and might still be under positive selection in many populations worldwide. This suggests that the biological constraints driving the evolution of genetic variants associated with T2D are still poorly understood. There is, therefore, a need to reconsider the selective pressures acting on these genes. In particular, it is crucial to analyze additional populations with contrasted lifestyles and modes of subsistence, as most populations studied so far are farmers. Furthermore, we believe that considerable progress could be made if forthcoming studies were based on the analysis of individual populations (rather than geographical groups of populations), provided that detailed investigations on their lifestyle and mode of subsistence are undertaken. It is also important that these studies infer the time of onset of selection. Only with this information will we be able to evaluate the extent to which positive selection of protective variants occurred since the Neolithic.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all the people who volunteered to participate in this study or who helped us in the field. We thank K. Zeng for providing us his program for computing the *E* statistics. This work was supported by the 'Service de Systématique Moléculaire' of the Muséum national d'Histoire naturelle (UMS 2700 CNRS) and the 'Consortium National de Recherche en Génomique': it is part of the agreement number 2005/67 between the Genoscope and the Muséum National d'Histoire Naturelle on the project 'Macrophylogeny of life' directed by Guillaume Lecointre. This work was funded by the ANR grant 'NUTGENEVOL' (07-BLAN-0064). RV also acknowledges support from the ANR grant 'EMILE' (09-BLAN-0145-01). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

- 3 Zimmet P, Alberti KG, Shaw J: Global and societal implications of the diabetes epidemic. *Nature* 2001; **414**: 782–787.
- 4 Herder C, Roden M: Genetics of type 2 diabetes: pathophysiologic and clinical relevance. *Eur J Clin Invest* 2011; **41**: 679–692.
- 5 Morris AP, Voight BF, Teslovich TM et al: Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012; **44**: 981–990.
- 6 Pavard S, Metcalf CJ: Negative selection on BRCA1 susceptibility alleles sheds light on the population genetics of late-onset diseases and aging theory. *PLoS One* 2007; **2**: e1206.
- 7 Stern MP: Diabetes and cardiovascular disease. The 'common soil' hypothesis. *Diabetes* 1995; **44**: 369–374.
- 8 Robitaille J, Grant AM: The genetics of gestational diabetes mellitus: evidence for relationship with type 2 diabetes mellitus. *Genet Med* 2008; **10**: 240–250.
- 9 Diamond J: The double puzzle of diabetes. *Nature* 2003; **423**: 599–602.
- 10 Neel JV: Diabetes mellitus: a 'thrifty' genotype rendered detrimental by 'progress'? *Am J Hum Genet* 1962; **14**: 353–362.
- 11 Neel JV: The thrifty genotype revisited; in: Kobberling J, Tattersall RB (eds) *The Genetics of Diabetes Mellitus*. London: Academic Press 1982 vol Serono Symposium No 47.
- 12 Neel JV, Weder AB, Julius S: Type II diabetes, essential hypertension, and obesity as 'syndromes of impaired genetic homeostasis': the 'thrifty genotype' hypothesis enters the 21st century. *Perspect Biol Med* 1998; **42**: 44–74.
- 13 Prentice AM: Starvation in humans: evolutionary background and contemporary implications. *Mech Ageing Dev* 2005; **126**: 976–981.
- 14 Benyshek DC, Watson JT: Exploring the thrifty genotype's food-shortage assumptions: a cross-cultural comparison of ethnographic accounts of food security among foraging and agricultural societies. *Am J Phys Anthropol* 2006; **131**: 120–126.
- 15 Brand Miller JC, Colagiuri S: The carnivore connection: dietary carbohydrate in the evolution of NIDDM. *Diabetologia* 1994; **37**: 1280–1286.
- 16 Colagiuri S, Brand Miller J: The 'carnivore connection'—evolutionary aspects of insulin resistance. *Eur J Clin Nutr* 2002; **56**: Suppl 1 S30–S35.
- 17 Wells JC: Ethnic variability in adiposity and cardiovascular risk: the variable disease selection hypothesis. *Int J Epidemiol* 2009; **38**: 63–71.
- 18 Roth J: Evolutionary speculation about tuberculosis and the metabolic and inflammatory processes of obesity. *JAMA* 2009; **301**: 2586–2588.
- 19 Allen JS, Cheer SM: 'Civilisation' and the thrifty genotype. *Asia Pacific J Clin Nutr* 1996; **4**: 341–342.
- 20 Hancock AM, Witonsky DB, Gordon AS et al: Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet* 2008; **4**: e32.
- 21 Speakman JR: Thrifty genes for obesity, an attractive but flawed idea, and an alternative perspective: the 'drifty gene' hypothesis. *Int J Obes (Lond)* 2008; **32**: 1611–1617.
- 22 Klopstein S, Currat M, Excoffier L: The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol* 2006; **23**: 482–490.
- 23 Fullerton SM, Bartoszewicz A, Ybaza G et al: Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am J Hum Genet* 2002; **70**: 1096–1106.
- 24 Ruiz-Narvaez E: Is the Ala12 variant of the PPARG gene an 'unthrifty allele'? *J Med Genet* 2005; **42**: 547–550.
- 25 Myles S, Hradetzky E, Engelken J et al: Identification of a candidate genetic variant for the high prevalence of type II diabetes in Polynesians. *Eur J Hum Genet* 2007; **15**: 584–589.
- 26 Myles S, Davison D, Barrett J, Stoneking M, Timpson N: Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics* 2008; **1**: 22.
- 27 Pickrell JK, Coop G, Novembre J et al: Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 2009; **19**: 826–837.
- 28 Southam L, Soranzo N, Montgomery SB et al: Is the thrifty genotype hypothesis supported by evidence based on confirmed type 2 diabetes- and obesity-susceptibility variants? *Diabetologia* 2009; **52**: 1846–1851.
- 29 Chen R, Corona E, Sikora M et al: Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. *PLoS Genet* 2012; **8**: e1002621.
- 30 Vander Molen J, Frisse LM, Fullerton SM et al: Population genetics of CAPN10 and GPR35: implications for the evolution of type 2 diabetes variants. *Am J Hum Genet* 2005; **76**: 548–560.
- 31 Helgason A, Palsson S, Thorgeirsson G et al: Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet* 2007; **39**: 218–225.
- 32 Gibson G: Human evolution: thrifty genes and the dairy queen. *Curr Biol* 2007; **17**: R295–R296.
- 33 Klimentidis YC, Abrams M, Wang J, Fernandez JR, Allison DB: Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Hum Genet* 2011; **129**: 407–418.
- 34 Martinez-Cruz B, Vitalis R, Segurel L et al: In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur J Hum Genet* 2011; **19**: 216–223.
- 35 Patin E, Laval G, Barreiro LB et al: Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* 2009; **5**: e1000448.
- 36 Librado P, Rozas J: DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 2009; **25**: 1451–1452.

1 Diamond J: Evolution, consequences and future of plant and animal domestication. *Nature* 2002; **418**: 700–707.

2 Brown EA: Genetic explorations of recent human metabolic adaptations: hypotheses and evidence. *Biol Rev Camb Philos Soc* 2012.

- 37 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
- 38 Fu YX, Li WH: Statistical tests of neutrality of mutations. *Genetics* 1993; **133**: 693–709.
- 39 Fu YX: Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 1997; **147**: 915–925.
- 40 Zeng K, Fu YX, Shi S, Wu Cl: Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 2006; **174**: 1431–1439.
- 41 Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statistic Soc B* 1995; **57**: 289–300.
- 42 Rousset F: GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Res* 2008; **8**: 103–106.
- 43 Beaumont M, Nichols RA: Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond* 1996; **263**: 1619–1626.
- 44 Wright S: The genetical structure of populations. *Ann Eugen* 1951; **15**: 323–354.
- 45 Bonin A, Taberlet P, Miaud C, Pompanon F: Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol Biol Evol* 2006; **23**: 773–783.
- 46 Segurel L, Lafosse S, Heyer E, Vitalis R: Frequency of the AGT Pro11Leu polymorphism in humans: Does diet matter? *Ann Hum Genet* 2010; **74**: 57–64.
- 47 Scheet P, Stephens M: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006; **78**: 629–644.
- 48 Gautier M, Vitalis R: rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 2012; **28**: 1176–1177.
- 49 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **4**: e72.
- 50 Austerlitz F, Kalaydjieva L, Heyer E: Detecting population growth, selection and inherited fertility from haplotypic data in humans. *Genetics* 2003; **165**: 1579–1586.
- 51 Wolfram Research I. Mathematica; in: 8.0V (ed). Wolfram Research, Inc.: Champaign, Illinois, 2010.
- 52 Casto AM, Feldman MW: Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? *PLoS Genet* 2011; **7**: e1001266.
- 53 Przeworski M, Coop G, Wall JD: The signature of positive selection on standing genetic variation. *Evolution* 2005; **59**: 2312–2323.
- 54 Fullerton SM, Clark AG, Weiss KM *et al*: Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* 2000; **67**: 881–900.
- 55 Nakajima T, Wooding S, Sakagami T *et al*: Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. *Am J Hum Genet* 2004; **74**: 898–916.
- 56 Thompson EE, Kuttub-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A: CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet* 2004; **75**: 1059–1069.
- 57 Hancock AM, Witonsky DB, Ehler E *et al*: Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci USA* 2010; **107**: Suppl 2 8924–8930.
- 58 Heyer E, Brazier L, Segurel L *et al*: Lactase persistence in central Asia: phenotype, genotype, and evolution. *Hum Biol* 2011; **83**: 379–392.
- 59 O'Keefe Jr. JH, Cordain L: Cardiovascular disease resulting from a diet and lifestyle at odds with our Paleolithic genome: how to become a 21st-century hunter-gatherer. *Mayo Clin Proc* 2004; **79**: 101–108.
- 60 Jew S, AbuMweis SS, Jones PJ: Evolution of the human diet: linking our ancestral diet to modern functional foods as a means of chronic disease prevention. *J Med Food* 2009; **12**: 925–934.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)