Genetic Adaptation to Levels of Dietary Selenium in Recent Human History

Louise White,*1 Frédéric Romagné,1 Elias Müller,1 Eva Erlebach,1 Antje Weihmann,1 Genís Parra,1 Aida M. Andrés,1 and Sergi Castellano*1

1Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

*Corresponding author: E-mail: louise.white@eva.mpg.de; sergi.castellano@eva.mpg.de.

Abstract

As humans migrated around the world, they came to inhabit environments that differ widely in the soil levels of certain micronutrients, including selenium (Se). Coupled with cultural variation in dietary practices, these migrations have led to a wide range of Se intake levels in populations around the world. Both excess and deficiency of Se in the diet can have adverse health consequences in humans, with severe Se deficiency resulting in diseases of the bone and heart. Se is required by humans mainly due to its function in selenoproteins, which contain the amino acid selenocysteine as one of their constituent residues. To understand the evolution of the use of this micronutrient in humans, we surveyed the patterns of polymorphism in all selenoprotein genes and genes involved in their regulation in 50 human populations. We find that single nucleotide polymorphisms from populations in Asia, particularly in populations living in the extreme Se-deficient regions of China, have experienced concerted shifts in their allele frequencies. Such differentiation in allele frequencies across genes is not observed in other regions of the world and is not expected under neutral evolution, being better explained by the action of recent positive selection. Thus, recent changes in the use and regulation of Se may harbor the genetic adaptations that helped humans inhabit environments that do not provide adequate levels of Se in the diet.

Key words: micronutrients, polygenic adaptation, local adaptation.

Introduction

As humans spread out of Africa about 60,000–100,000 years ago (Stringer and Andrews 1988; Fu et al. 2013) and migrated around the world, they came to inhabit a vast range of environments. These environments differ widely in their geology, and the abundance of the chemical elements required by humans varies in their rocks and soils (Selinus and Alloway 2005). Elements such as zinc, iron, manganese, copper, iodine, molybdenum, and selenium (Se), which are essential to the human diet but needed only in trace amounts (Mertz 1981), accumulate in varying quantities in the plants and animals in these environments. The intake of these micronutrients, which may have changed with the domestication of plants and animals in the past 10,000 years (Diamond 2002), can exceed or fail the dietary needs of human populations today (Selinus and Alloway 2005).

Both excess and deficiency of micronutrients in the diet have adverse health consequences in humans, but deficiency is more common with iron and iodine deficiencies today affecting almost one-quarter of the world’s population (Caballero 2002; Bhutta and Salam 2012). Deficiency of other micronutrients, such as Se, appears less widespread but severe in some regions of the world. Studies on the level and bioavailability of Se in plant and animal foods across the world suggest that it varies widely at both global and local scales (Oldfield 2002; Selinus and Alloway 2005), with relatively little Se in countries such as Finland (Koivistoinen and Huttunen 1986), New Zealand (Thomson 2004), possibly Malawi (Hurst et al. 2013) and, most notably, some parts of China (Oldfield 2002; Xia et al. 2005). Heart (Keshan) and bone (Kashin–Beck) diseases, which are treated with Se supplementation, are endemic to large areas of China (Keshan Disease Research Group of the Chinese Academy of Medical Sciences 1979; Moreno-Reyes et al. 2001). More generally, Se deficiency is associated with low immune response, infertility, cognitive decline, and increased risk of mortality (Hatfield et al. 2012; Rayman 2012).

Se is required by humans mainly due to its function in 25 selenoproteins (Kryukov et al. 2003), which contain the amino acid selenocysteine (Sec) as one of their constituent residues. This amino acid is encoded by a codon (UGA) that in all other genes represents a stop codon. Sec is analogous to the amino acid cysteine (Cys) in its molecular structure with an atom of Se replacing that of sulfur (Hatfield et al. 2012). These elements confer different properties when incorporated into proteins (Gromer et al. 2003; Steinmann et al. 2010; Snider et al. 2013) and although substitutions between Sec and Cys in orthologous proteins are uncommon in vertebrates (Castellano et al. 2009), paralogous proteins with Cys instead of Sec exist in most vertebrate species, with humans having six of them. Selenoproteins and their Cys-containing paralogs are generally involved in preventing oxidative damage (Steinbrenner and Sies 2009).

The homeostasis and metabolism of Se and Sec involves the transport of dietary Se from the liver to other organs in the body, its receptor-mediated uptake into cells, the
synthesis of Sec, and the readthrough of a UGA stop codon with a tRNA charged with Sec. Se deficiency tightly regulates these processes to favor the supply of Se to some tissues and selenoproteins at the expense of others (Schomburg and Schweizer 2009). The molecular basis of such hierarchy in the distribution of Se is not fully understood but some 19 proteins may be involved (Allmang et al. 2009; Burk and Hill 2009; Hatfield et al. 2012), including different receptors for the uptake of Se in cells (Burk and Hill 2009) and proteins directing the recoding of the UGA codon in selenoproteins (Dumitrescu et al. 2005; Allmang et al. 2009; Bifano et al. 2013) through their differential binding of an RNA secondary structure in each selenoprotein mRNA (Latreche et al. 2009).

To explore whether changes in dietary Se have shaped the evolution of genes that use or regulate Se, we have conducted a survey of single nucleotide polymorphisms (SNPs) in selenoprotein genes and genes involved in the regulation of Se and Sec in 50 human populations (Cann et al. 2002). We investigated whether human populations have adapted to the different levels of Se in their diets by assessing the signatures of positive selection in these genes. We investigate both classical signatures of selection and signatures of polygenic adaptation, including population differentiation and the concerted change in allele frequencies in different regions of the world. We find evidence for polygenic adaptation in the groups of selenoprotein genes and genes involved in the regulation of Se and Sec, with the strongest signatures in populations from China living in regions characterized by Se deficiency (Oldfield 2002; Xia et al. 2005).

Results

Candidate Genes and Their Genetic Variation in Humans

We used a hybridization approach (see Materials and Methods) to enrich and sequence the protein-coding and regulatory regions of 25 selenoprotein genes and 19 genes involved in the regulation of Se and Sec (table 1) in 855 unrelated individuals from the Human Genome Diversity Panel (HGDP-CEPH) (supplementary tables S1 and S2, Supplementary Material online) (Cann et al. 2002). These individuals belong to 50 populations and provide an unbiased sample of Se nutritional histories in humans across Africa, Middle East, Europe, Central South Asia, East Asia, Oceania, and America. We also sequenced the six Cys-containing paralogs of selenoproteins (table 1) in the same populations. We called SNPs in the HGDP-CEPH individuals, with an average coverage of 20-fold across the 50 candidate genes, and identified 7,989 SNPs (see Materials and Methods). To derive a neutral expectation for these SNPs and genes, we sequenced 7,989 SNPs (see Materials and Methods). To derive a neutral expectation for these SNPs and genes, we sequenced 1,035 SNPs in them (see Materials and Methods).

Signatures of Positive Selection in the Candidate Genes

We investigated whether certain genes have differentiated, in certain populations, beyond what is expected under neutrality. Specifically, we tested whether the groups of selenoprotein genes, their Cys-containing paralogs, or the genes involved in the regulation of Se and Sec (table 1) show concerted shifts in their SNP allele frequencies in certain regions of the world. We first computed the FST statistic

<table>
<thead>
<tr>
<th>Table 1. Candidate Genes Grouped According to Their Type and the Biological Process They Participate in.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genes</strong></td>
</tr>
<tr>
<td><strong>Selenoproteins</strong></td>
</tr>
<tr>
<td>Glutathione peroxidase (GPs) 1, 2, 3, 4, and 6</td>
</tr>
<tr>
<td>Iodothyronine deiodinase (DI) 1, 2, and 3</td>
</tr>
<tr>
<td>15 kDa selenoprotein (Sel15)</td>
</tr>
<tr>
<td>Selenoprotein H (SelH)</td>
</tr>
<tr>
<td>Selenoprotein I (SelI)</td>
</tr>
<tr>
<td>Selenoprotein K (SelK)</td>
</tr>
<tr>
<td>Selenoprotein M (SelM)</td>
</tr>
<tr>
<td>Selenoprotein N (SelN)</td>
</tr>
<tr>
<td>Selenoprotein O (SelO)</td>
</tr>
<tr>
<td>Selenoprotein R 1 (SelR1)</td>
</tr>
<tr>
<td>Selenoprotein S (SelS)</td>
</tr>
<tr>
<td>Selenoprotein T (SelT)</td>
</tr>
<tr>
<td>Selenoprotein V (SelV)</td>
</tr>
<tr>
<td>Selenoprotein W 1 (SelW1)</td>
</tr>
<tr>
<td>Thioredoxin reductase (TR) 1, 2, and 3</td>
</tr>
<tr>
<td>Cys-containing paralogs</td>
</tr>
<tr>
<td>Glutathione peroxidase (GPs) 5, 7, and 8</td>
</tr>
<tr>
<td>Selenoprotein R 2 and 3</td>
</tr>
<tr>
<td>Selenoprotein W 2 (SelW2)</td>
</tr>
<tr>
<td>Regulation of Se and Sec</td>
</tr>
<tr>
<td>Transport and uptake of Se into cells</td>
</tr>
<tr>
<td>Selenoprotein P (SelP)</td>
</tr>
<tr>
<td>Apolipoprotein E receptor 2 (ApoER2)</td>
</tr>
<tr>
<td>Megalin</td>
</tr>
<tr>
<td>Metabolism of Se</td>
</tr>
<tr>
<td>Selenocysteine lyase (SCLY)</td>
</tr>
<tr>
<td>Se binding protein 1 (SELENBP1)</td>
</tr>
<tr>
<td>Biosynthesis of Sec</td>
</tr>
<tr>
<td>O-phosphoryl tRNasec Kinase (PSTK)</td>
</tr>
<tr>
<td>Selenophosphate synthetase 2 (SPS2)</td>
</tr>
<tr>
<td>O-phosphoryltRNAsec Ser transferase (SEPS2C)</td>
</tr>
<tr>
<td>Seryl-tRNA synthetase (SARS2)</td>
</tr>
<tr>
<td>tRNA</td>
</tr>
<tr>
<td>Ser-tRNAsec, copy in chromosome 17</td>
</tr>
<tr>
<td>Ser-tRNAsec, copy in chromosome 19</td>
</tr>
<tr>
<td>Ser-tRNAsec, copy in chromosome 22</td>
</tr>
<tr>
<td>Incorporation of Sec into proteins</td>
</tr>
<tr>
<td>CUGBP, Elav-like family member 1 (CELF1)</td>
</tr>
<tr>
<td>Elongation factor for Sec (EFSec)</td>
</tr>
<tr>
<td>Eukaryotic translation initiation factor 4A3 (eIF4A3)</td>
</tr>
<tr>
<td>ELAV like RNA binding protein 1 (ELAVL1)</td>
</tr>
<tr>
<td>Ribosomal protein L30 (RPL30)</td>
</tr>
<tr>
<td>SECIS binding protein 2 (SBP2)</td>
</tr>
<tr>
<td>Selenophosphate synthetase 1 (SPS1)</td>
</tr>
<tr>
<td>tRNAsec 1 associated protein 1 (TRNAU1AP)</td>
</tr>
<tr>
<td>Exportin 1 (XPO1)</td>
</tr>
</tbody>
</table>

**Note:** SelP and SPS2 are selenoproteins with a regulatory role.
(Weir and Cockerham 1984) for each of these candidate genes by combining the $F_{ST}$ values of all SNPs in each gene (see Materials and Methods). This $F_{ST}$ value takes into account the differentiation of functional alleles, if any, and other linked alleles and, thus, measures the overall genetic differentiation of each gene between each pair of populations in our study. We then compared the $F_{ST}$ values of each gene with the $F_{ST}$ values of our set of neutral loci, for the same population pair (fig. 1). This provides a rank for each $F_{ST}$ value with respect to its neutral expectation, which measures how unusual the differentiation for a gene is in a given population pair.

We divided the genes into the three functional groups above and focused only on genes whose $F_{ST}$ values rank in the top 5% of the distribution of neutral $F_{ST}$ values, as these extreme genes bear the strongest signature of local adaptation. We then grouped populations by world region (supplementary table S2, Supplementary Material online) to consider together the extreme $F_{ST}$ ranks from populations in each world region (excluding pairs of populations in the same world region). We then compared the $F_{ST}$ ranks across world regions with a two-sided Mann–Whitney $U$ test (fig. 1A). The skewness of the $F_{ST}$ ranks measures the deviation of these genes from the neutral expectation, with stronger skewness (higher $F_{ST}$ ranks) indicating higher allele frequency differentiation in a given world region. The $P$ values from the Mann–Whitney $U$ pairwise comparisons are represented as colors on a heat map in figure 2, with red indicating a significantly stronger departure from the neutral expectation in the left-hand world region than in the world region at the top. Such unexpected and region-specific allele frequency differentiation is a signature of local adaptation in that region of the world, through positive selection.

**Fig. 1.** Computation of $F_{ST}$ ranks for each set of candidate genes (selenoproteins, Cys-containing paralogs, or genes involved in the regulation of Se and Sec) between pairs of populations. The $F_{ST}$ ranks are based on the distribution of $F_{ST}$ values from neutral loci (pseudogenes) between the same pair of populations. (A) Comparison of $F_{ST}$ ranks between pairs of world regions: 1) Comparison of the ranks of $F_{ST}$ values in the top 5% of the distribution of neutral $F_{ST}$ with a two-sided Mann–Whitney $U$ test between a pair of world regions; and 2) $P$ values from these pairwise comparisons are shown as colors on a heat map. (B) Contribution of individual genes to the $F_{ST}$ ranks in a world region: 1) The first (lowest) $F_{ST}$ rank a value of 1, the next rank a value of 2, and so on. This produces a combined measure of the number of times the $F_{ST}$ of a gene (X or Y) significantly departs from the neutral expectation between region A and the rest of the world; and 2) Gene X has the highest sum of rank values and contributes more to the signatures of positive selection in region A than gene Y.

---

**Mann–Whitney U-test**

Compute $F_{ST}$ ranks per population pair

**A – World regions**

1 – Test shift in $F_{ST}$ ranks

2 – P-value heat map

**B – Genes**

1 – $F_{ST}$ ranks by gene

2 – Sum rank values

Gene X: $20 + 17 + 16 + 13 = 66$ (70.2%)

Gene Y: $14 + 8 + 5 + 1 = 28$ (29.8%)
Cys-Containing Paralogs

We observe no signatures of local positive selection in the group of Cys-containing paralogs, as no comparison in figure 2 shows a significant difference between pairs of world regions (lowest $P = 0.292$, for Middle East vs. Africa). These genes have functions that, while related to those of selenoproteins, are independent of Se and its abundance throughout the world. It is perhaps then not surprising that as a group Cys-containing paralogs have not experienced local adaptation in recent human history. If changes in dietary Se pose weak selective pressures in humans, we expect a similar pattern in selenoprotein genes.

Selenoprotein Genes

The pattern in selenoprotein genes (fig. 2A) is quite different from their Cys-containing homologs (fig. 2A). We observe significant signatures of local positive selection in selenoprotein genes, with several world region comparisons below or close to the 5% significance level. These include comparisons that involve Central South Asia ($P = 0.035$ vs. Oceania), East Asia ($P = 0.11$ vs. Africa, $P = 0.180$ vs. Middle East, $P = 0.133$ vs. Europe, $P = 0.027$ vs. Oceania), and America ($P = 0.010$ vs. Africa, $P = 0.025$ vs. Middle East, $P = 0.011$ vs. Europe, $P = 0.054$ vs. Central South Asia, $P = 0.053$ vs. East Asia, $P = 0.004$ vs. Oceania). Thus, local allele frequency differentiation in selenoprotein genes departs from neutral expectations significantly more in Asia and America than in other regions of the world. Furthermore, in these regions they depart from neutral expectations significantly more than their Cys-containing paralogs when all $F_{ST}$ ranks are considered ($P = 2.53 \times 10^{-4}$ in Central South Asia, $P < 2.20 \times 10^{-16}$ in East Asia, $P = 4.87 \times 10^{-11}$ in America). This is consistent with selenoprotein genes, unlike genes that use Cys in their proteins, having experienced local adaptation in some regions of the world.

Genes Involved in the Regulation of Se and Sec

Interestingly, the pattern in genes that regulate Se and Sec (fig. 2A) resembles that of selenoprotein genes rather than that of their Cys-containing paralogs, with the strongest differences in allele frequencies found in Central South Asia ($P = 0.207$ vs. Africa, $P = 0.057$ vs. America) and East Asia ($P = 0.147$ vs. Africa, $P = 0.173$ vs. Europe, $P = 0.028$ vs. America). Overall, the differences among pairs of world regions are somewhat weaker in these genes than in selenoprotein genes. Still, they also depart from neutral expectations significantly more than the Cys-containing paralogs when all $F_{ST}$ ranks are considered ($P < 2.20 \times 10^{-16}$ in Central South Asia, $P < 2.20 \times 10^{-16}$ in East Asia). This is consistent with regulatory changes in the homeostasis of Se and Sec having experienced local adaptation in human populations in Asia. These regulatory adaptations are likely to have a broader physiological impact than adaptations in selenoprotein genes.

Se Deficiency in China

The majority of populations from East Asia, a region with signatures of positive selection (fig. 2A), are from China (supplementary table S1, Supplementary Material online). Because China has large Se deficient areas (Oldfield 2002; Xia et al. 2005), we further investigated whether these signatures are due to local differences in dietary Se intake. To do this, we compared the $F_{ST}$ ranks of Se-related genes from six
Chinese populations living in known Se deficient areas to eight Chinese populations living in areas that are not deficient (supplementary table S7, Supplementary Material online) (see Materials and Methods). The group of populations living in Se deficient areas has a significantly stronger enrichment in highly differentiated genes than populations living in nondeficient areas ($P = 1.42 \times 10^{-10}$ for selenoproteins, $P = 2.00 \times 10^{-8}$ for regulation of Se and Sec). Furthermore, as a group, populations living in Se deficient areas show significant signatures of positive selection compared with other world regions (fig. 2B) in selenoprotein genes ($P = 3.22 \times 10^{-4}$ vs. Africa, $P = 3.16 \times 10^{-3}$ vs. Middle East, $P = 1.02 \times 10^{-4}$ vs. Europe, $P = 1.51 \times 10^{-3}$ vs. Central South Asia, $P = 1.72 \times 10^{-3}$ vs. Oceania) and genes involved in the regulation of Se and Sec ($P = 6.62 \times 10^{-3}$ vs. Africa, $P = 0.04$ vs. Middle East, $P = 7.49 \times 10^{-3}$ vs. Europe, $P = 0.05$ vs. Central South Asia, $P = 5.84 \times 10^{-4}$ vs. America). This signature is absent in the set of Chinese populations living in Se adequate areas (fig. 2B). This is consistent with Se-related genes having experienced local adaptation in populations from geographic regions whose soil does not provide adequate levels of dietary Se.

### Polygenic Adaptation

We next investigated which particular genes, among the selenoproteins and genes involved in the regulation of Se and Sec, contribute most to the signatures of local adaptation in Asia. We aimed to determine whether the signatures of positive selection in these regions are spread across genes. For each gene we counted the number of times its $F_{ST}$ values appear in the upper 5% tail of the neutral $F_{ST}$ distribution, and measured the contribution of each gene as its fraction in the total number of extreme $F_{ST}$ values (fig. 1B and Materials and Methods). Signatures of positive selection in Asia are the result of allele frequency shifts in many genes, each of them having a contribution to the extreme $F_{ST}$ values in this region (supplementary table S4, Supplementary Material online). In fact, only three and six genes do not contribute to the signatures in selenoproteins from East Asia and Central South Asia, respectively (supplementary tables S4 and S5, Supplementary Material online). For genes involved in the regulation of Se and Sec, the number of genes not contributing is 2 and 5, respectively (supplementary tables S4 and S5, Supplementary Material online). This suggests that local adaptation was polygenic in nature in Asia, with SNPs in several genes changing in allele frequency in response to differences in environmental Se.

Still, some genes have a particularly large contribution to these signatures of local adaptation. The five selenoprotein genes that show the strongest evidence of local adaptation in East Asia already have 75.6% of the extreme $F_{ST}$ values in this world region. These genes are, in order: DI2, SelS, GPx1, SelM, and SelF (table 2 and supplementary table S4, Supplementary Material online). With the addition of DI3, the same genes contribute the most to the signatures of positive selection in Central South Asia (supplementary table S5, Supplementary Material online). In addition, multiple populations contribute to these signatures of positive selection in Asia, with the Hezhen, Naxi and Oroqen populations in East Asia and the Balochi, Brahui, Kalash, Sindhi and Uygur in Central South Asia contributing the most. The five regulatory genes that show the strongest evidence of local adaptation in East Asia have 64.4% of the extreme $F_{ST}$ values in this world region. These genes are, in order: CELF1, SP52, SEPCES, ELAVL1 and Ser-tRNA$^{Sec}$ (table 2, fig. 3, and supplementary table S4, Supplementary Material online). With the addition of SelP, the same genes contribute the most to the signatures in Central South Asia (table 2 and supplementary table S5, Supplementary Material online). We note that the picture in America is very different, as the signatures from the allele frequency differentiation in selenoprotein genes (fig. 2A) are almost exclusive due to the DI2 gene (supplementary table S6, Supplementary Material online), particularly in the Colombian population.

In any case, we do not find a significant excess of low frequency or high frequency derived alleles in any of the 50 candidate genes using classic neutrality tests, such as Tajima's $D$ and Fay & Wu's $H$ (see Materials and Methods). The neutral expectation for these tests was based on coalescent simulations for the genes and populations in our study using a modified demography from Gutenkunst et al. (2009), which accounts for the low admixture of Europeans with the American populations analyzed (see Materials and Methods). We conclude that polygenic adaptation was important in those genes that use or regulate Se and Sec in Asia.

### Discussion

An interesting question in biology is whether shifts in micronutrients intake as humans migrated around the world or changed their dietary practices are an important selective pressure in humans. This is the case for certain macronutrients, such as starch (Perry et al. 2007) and lactose (Tishkoff et al. 2007), but the effect of micronutrients such as Se remains largely unexplored. We have addressed this question by surveying the patterns of polymorphism in selenoprotein genes and genes involved in the regulation of Se and Sec in 50 human populations, which have a range of Se nutritional histories.

Although a few regions in Africa, middle East, Europe or Oceania have abnormal levels of Se (Koivisto and Huttunen 1986; Thomson 2004; Hurst et al. 2013), Se deficiency or toxicity has not been broadly reported (Oldfield 2002; Safaralizadeh et al. 2005; Selinus and Alloway 2005; Novakovic et al. 2014), with no Se-related diseases being common in these regions (Oldfield 2002; Selinus and Alloway 2005; Novakovic et al. 2013). This agrees well with the absence of signatures of local positive selection in these world regions (fig. 2A).

In contrast, levels of Se in the diet vary widely across some parts of Asia (Oldfield 2002), with Pakistan, where most Central South Asian populations were sampled (supplementary tables S1 and S2, Supplementary Material online), having levels of Se in the soil that are borderline deficient across large areas (Khan et al. 2006, 2008; Ahmad et al. 2009). This agrees with the signatures of positive selection identified in...
Table 2. Genes and SNPs Contributing Most to the Signatures of Positive Selection in East Asia (figs. 2 and 3).

<table>
<thead>
<tr>
<th>Gene Rank</th>
<th>Gene</th>
<th>SNP Rank</th>
<th>SNPs ID in SelenoDB 2.0, Chromosome and Coordinate (hg19)</th>
<th>Gene Region, SNP ID in dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DI2</td>
<td>1</td>
<td>SNP00000190_2.0, 14, 80,677,639 Coding/Intron</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SelS</td>
<td>1</td>
<td>SNP00005709_2.0, 15, 101,815,150 Intron</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GPx1</td>
<td>1</td>
<td>SNP00000352_2.0, 3, 49,395,766 5'-UTR/Promoter region</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SelM</td>
<td>1</td>
<td>SNP00005262_2.0, 22, 31,498,917 Downstream region</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Sel15</td>
<td>1</td>
<td>SNP00001985_2.0, 1, 87,381,198 Promoter region</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CELF1</td>
<td>1</td>
<td>SNP0005905_2.0, 11, 47,500,661 Intron</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>SPS2</td>
<td>1</td>
<td>SNP00001698_2.0, 16, 30,457,350 5'-UTR</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SEPSECS</td>
<td>1</td>
<td>SNP00001861_2.0, 4, 25,161,929 Coding</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ELAVL1</td>
<td>1</td>
<td>SNP00006324_2.0, 19, 8,027,942 3'-UTR/Downstream region</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Ser-tRNA Sec</td>
<td>1</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

Note.—The SNP ID in SelenoDB 2.0 is given for all SNPs in protein-coding genes. The SNP ID in dbSNP is given for those SNPs with known functional consequences discussed in the text.

**Fig. 3.** Genes involved in the biosynthesis of Sec and its incorporation into proteins (in orange) that contribute most to the signatures of positive selection in regulatory genes from East Asian populations.
populations from these regions in both selenoprotein genes and genes involved in the regulation of Se and Sec (fig. 2A). Furthermore, the most Se deficient areas worldwide, where Se-related diseases are endemic (Keshan Disease Research Group of the Chinese Academy of Medical Sciences 1979; Moreno-Reyes et al. 2001; Oldfield 2002; Xia et al. 2005), are found in China in East Asia. This agrees with the signatures of local adaptation identified in those populations in China living in Se deficient areas (fig. 2B). Thus, local adaptation influenced not only selenoprotein genes but also the tight regulation in the use of Se that has been observed in humans under deficiency conditions (Allmang et al. 2009; Burk and Hill 2009; Hatfield et al. 2012). The absence of signatures in Cys-containing genes (fig. 2A) shows that the differentiation observed in Se-related genes is not due to the genetic drift experienced by populations in East Asia (Keinan et al. 2007).

We are more cautious in the interpretation of the signatures of positive selection in America (fig. 2A). First, they are observed only in selenoprotein genes and are largely due to a single gene, the DI2 gene. So, it is unclear whether Se availability (rather than other selective pressures in this gene) is responsible for them. Second, the signatures are not widespread in this world region but strongest in one particular population, the Colombian. In any case, little is known about the levels of dietary Se in humans in America. Although white muscle disease, a symptom of Se deficiency, is quite prevalent in the levels of dietary Se in humans in America. Although white muscle disease, a symptom of Se deficiency, is quite prevalent in

The selenoprotein genes contributing to the signatures of positive selection in Asia (table 2) contain SNPs with known functional consequences, some also with high levels of population differentiation. First, DI2, an oxidoreductase that catalyzes the conversion of the iodine-dependent hormone T4 to its active form T3 in the thyroid, has a Thr to Ala substitution (rs225014) in exon 2 with high $F_{ST}$ values for many East Asian populations. Tissue samples from individuals homozygous for the Ala allele at this locus have been shown to exhibit significantly lower DI2 enzyme velocity than samples with Ala/Thr or Thr/Thr genotypes (Canani et al. 2005). The activity of this enzyme thus depends on two micronutrients, Se and iodine, whose deficiency has been linked to Kashin–Beck disease in China (Yao et al. 2011). Second, SeiS, a gene involved in stress response in the endoplasmic reticulum and genes involved in the regulation of Se and Sec (fig. 2A). Furthermore, the most Se deficient areas worldwide, where Se-related diseases are endemic (Keshan Disease Research Group of the Chinese Academy of Medical Sciences 1979; Moreno-Reyes et al. 2001; Oldfield 2002; Xia et al. 2005), are found in China in East Asia. This agrees with the signatures of local adaptation identified in those populations in China living in Se deficient areas (fig. 2B). Thus, local adaptation influenced not only selenoprotein genes but also the tight regulation in the use of Se that has been observed in humans under deficiency conditions (Allmang et al. 2009; Burk and Hill 2009; Hatfield et al. 2012). The absence of signatures in Cys-containing genes (fig. 2A) shows that the differentiation observed in Se-related genes is not due to the genetic drift experienced by populations in East Asia (Keinan et al. 2007).

We are more cautious in the interpretation of the signatures of positive selection in America (fig. 2A). First, they are observed only in selenoprotein genes and are largely due to a single gene, the DI2 gene. So, it is unclear whether Se availability (rather than other selective pressures in this gene) is responsible for them. Second, the signatures are not widespread in this world region but strongest in one particular population, the Colombian. In any case, little is known about the levels of dietary Se in humans in America. Although white muscle disease, a symptom of Se deficiency, is quite prevalent in

The selenoprotein genes contributing to the signatures of positive selection in Asia (table 2) contain SNPs with known functional consequences, some also with high levels of population differentiation. First, DI2, an oxidoreductase that catalyzes the conversion of the iodine-dependent hormone T4 to its active form T3 in the thyroid, has a Thr to Ala substitution (rs225014) in exon 2 with high $F_{ST}$ values for many East Asian populations. Tissue samples from individuals homozygous for the Ala allele at this locus have been shown to exhibit significantly lower DI2 enzyme velocity than samples with Ala/Thr or Thr/Thr genotypes (Canani et al. 2005). The activity of this enzyme thus depends on two micronutrients, Se and iodine, whose deficiency has been linked to Kashin–Beck disease in China (Yao et al. 2011). Second, SeiS, a gene involved in stress response in the endoplasmic reticulum and genes involved in the regulation of Se and Sec (fig. 2A). Furthermore, the most Se deficient areas worldwide, where Se-related diseases are endemic (Keshan Disease Research Group of the Chinese Academy of Medical Sciences 1979; Moreno-Reyes et al. 2001; Oldfield 2002; Xia et al. 2005), are found in China in East Asia. This agrees with the signatures of local adaptation identified in those populations in China living in Se deficient areas (fig. 2B). Thus, local adaptation influenced not only selenoprotein genes but also the tight regulation in the use of Se that has been observed in humans under deficiency conditions (Allmang et al. 2009; Burk and Hill 2009; Hatfield et al. 2012). The absence of signatures in Cys-containing genes (fig. 2A) shows that the differentiation observed in Se-related genes is not due to the genetic drift experienced by populations in East Asia (Keinan et al. 2007).

We are more cautious in the interpretation of the signatures of positive selection in America (fig. 2A). First, they are observed only in selenoprotein genes and are largely due to a single gene, the DI2 gene. So, it is unclear whether Se availability (rather than other selective pressures in this gene) is responsible for them. Second, the signatures are not widespread in this world region but strongest in one particular population, the Colombian. In any case, little is known about the levels of dietary Se in humans in America. Although white muscle disease, a symptom of Se deficiency, is quite prevalent in
magnitudes of their functional impact is in line with the magnitude of their differentiation, which is spread across genes, making such future assessment challenging. In any case, our understanding of local adaptations to environmental Se will only improve with better knowledge of the levels of Se in the diet of human populations worldwide. Today, the lack of comparable quantitative data on soil, plant or dietary Se levels across the world hampers the correlation of human Se status with genetic data. Progress on the biology of micronutrients and the evolution of their use by humans is thus likely to come from linking alleles at functional sites with precise measurement of these essential trace elements across the world.

Materials and Methods

Candidate Genes
We compiled a list of 50 human genes (table 1). This list includes the 25 selenoprotein genes in humans and the 6 genes that encode proteins with sulfur (in the form of cysteine, Cys) instead of Se (in the form of selenocysteine, Sec) (Kryukov et al. 2003). These are paralogs of selenoprotein genes and their functions may be related to their Se-dependent counterparts. We have also included 19 genes involved in the homeostasis and metabolism of Se and Sec (Allmang et al. 2009). These are genes whose function and regulation may depend on levels of Se in the diet.

We used the human selenoprotein gene annotation from SelenoDB 1.0 (Castellano et al. 2008), which includes the annotation of the SECS element of each selenoprotein gene. The annotation of the Cys-containing homologs was also obtained from SelenoDB 1.0. The annotation of genes involved in the homeostasis and metabolism of Se and Sec, as well as the alternative transcript forms for all genes, was obtained from RefSeq (Pruitt et al. 2009). For analysis, the 50 genes were split into three groups as defined in table 1: 1) Selenoprotein genes, 2) Cys-containing paralogous genes, and 3) genes involved in the regulation of Se and Sec, which includes the SelP and SP52 selenoprotein genes.

Pseudogenes
Following Andrés et al. (2010), we used a group of 48 neutrally evolving loci to derive the neutral expectation for some of the analyses (supplementary table S3, Supplementary Material online). These loci are unlinked (physically dispersed) known pseudogenes that are greater than 400 nt in length and located over 2,000 nt from any known gene. To ensure that these pseudogenes are neutral with respect to natural selection, they 1) do not overlap evolutionary conserved regions of the human genome; 2) do not include simple repeats or low complexity DNA; 3) do not differ from the rest of the human genome in terms of recombination rate; 4) are present as pseudogenes in chimpanzee, orangutan, and rhesus; and 5) are present in a single copy in the human and chimpanzee genomes or, when a member of a gene family, only one member is included. In addition, processed ribosomal pseudogenes and olfactory receptor pseudogenes are not included. The length of these pseudogenes ranges from 422 to 2,832 nt, and their combined length is 43,700 nt.

Human Samples
We used samples from the Foundation Jean Dausset-Centre d’Etude du Polymorphisme Humain; Human Genome Diversity Panel (HGDP-CEPH) (supplementary table S1, Supplementary Material online). We restricted our analysis to the HGDP-CEPH set detailed in Rosenberg (2006), which excludes individuals with second-degree relationships or closer. Of this set 940 individuals were available for sequencing. Additionally, we removed one Naxi individual (HGDP01339) prior to variant calling as a result of likely contamination during the sequencing process. The set used for variant calling consisted of 928 individuals (supplementary table S2, Supplementary Material online). These individuals from 53 populations were grouped in one of the following geographical regions for analysis: Africa, Middle East, Europe, Central South Asia, East Asia, Oceania, and America. Previous analyses have shown the Middle East, Central South Asia, and America to contain inbred populations (Pemberton et al. 2012), which may bias analyses based on allele frequency differences among populations. We therefore excluded from our analysis the population from each of these regions having the most inbred individuals (Pemberton et al. 2012) (supplementary table S2, Supplementary Material online). Our final analysis is thus based on 855 individuals from 50 worldwide populations.

Array Capture Design
We designed an Agilent custom array (Agilent Technologies) to target all exons (coding and noncoding untranslated regions, UTRs) plus 400 nt of the surrounding introns (200 nt at each side of each exon) and 2,000 nt upstream (to include promoter regions) of all candidate genes (table 1). Additionally, for genes where the target region was shorter than 4,000 nt we added 2,000 nt of extra downstream region and for tRNA genes 3,000 nt were added upstream and downstream. The complete coding region of each pseudogene was targeted (supplementary table S3, Supplementary Material online).

Using the human reference sequence (hg18) as a template, we designed a total of 469,695 probes (419,468 and 50,227 for the candidate and the putatively neutrally evolving genes, respectively). The probes were of 60 nt in length, started 50 nt before and after each targeted region, and were set every 1 nt to maximize capture efficiency. The probes target a total of 524,875 nt (470,800 and 54,083 nt for the candidate and the putatively neutrally evolving genes, respectively) of the human genome.

Library Preparation, Capture, and Sequencing
We prepared 940 libraries from the HGDP-CEPH samples following a double index protocol (Meyer and Kircher 2010). The genomic DNA (100µl, 10 ng/µl) of each sample was sonicated (Bioruptor) to fragments of 150–300 nt in length. Target capture was performed in batches of pooled
libraries with around 90 samples per pool (Hodges et al. 2009). The enriched libraries were amplified and sequenced using the Illumina GAIIx platform yielding 76-bp paired-end reads.

**Variant Calling**

Base-calling was performed with Ibis (Kircher et al. 2009). Reads were mapped to the human reference genome (hg19) using BWA (Li and Durbin 2009) yielding an average on-target coverage of 20-fold per individual and per gene after filtering out duplicate reads and those with a mapping quality less than 25. The GATK IndelRealigner was used to improve read alignment in indel regions (McKenna et al. 2010). We called genotypes using the GATK UnifiedGenotyper version 2.2 (DePristo et al. 2011), and filtered them to remove genotypes in sites that 1) had a coverage below 8 x in more than 50% of samples, 2) had an average coverage above 100 x, 3) were indels or SNPs within 5 bp of an indel, 4) were multi-allelic sites, 5) had an SNP quality less than 20, and 6) had a strand bias greater than 10. We additionally filtered out sites that did not have 1:1 human–chimp correspondence in the Ensembl EPO 6 primate alignments (Paten et al. 2008) or were identified as being prone to systematic error (Castellano et al. 2014). Genotypes with a quality (GQ) less than 20 were masked out but the positions retained unless all genotypes at that position had a GQ less than 20. This resulted in 7,989 and 1,035 SNPs in the candidate genes and pseudogenes, respectively, across populations.

**Population Differentiation**

We measured population differentiation with the $F_{ST}$ statistic according to the formula of Weir and Cockerham (1984). This formula calculates $F_{ST}$ per SNP in an unbiased way (that accounts for both genetic and statistical sampling) using an analysis of variance approach. The formula represents the proportion of the overall genetic diversity that is accounted for by allele frequency differences between pairs of populations. As originally defined by Wright (1951), $F_{ST}$ values range from 0 (no population differentiation) to 1 (no allele sharing among populations); however, the Weir and Cockerham (1984) method can result in negative $F_{ST}$ values which were considered as 0 in our analyses. We did not calculate $F_{ST}$ values for sites where fewer than 50% of individuals in a population had a genotype call. To calculate an $F_{ST}$ value per gene (based on all SNPs across that gene), we used a weighted average by first summing the numerator over all SNPs then summing the denominator over all SNPs before finding the quotient (Weir and Cockerham 1984). Thus, the resulting value is not the same as the average of single SNP $F_{ST}$ values. We calculated $F_{ST}$ between each pair of populations in supplementary table S2, Supplementary Material online, and for each pair of populations, an $F_{ST}$ value was obtained for each gene in table 1.

**Genes and SNPs Contributing to the Signatures of Positive Selection**

An overview of this analysis can be found in figure 1. To identify the genes that contribute most to the patterns in figure 2, we used the rank of their $F_{ST}$ values in the 5% upper tail of the neutral distribution. We then sorted the ranks of these extreme $F_{ST}$ values in ascending order and gave the first rank a value of 1, the next rank a value of 2, and so on (fig. 1B). Any identical ranks were given the same value. All values for each gene were then summed (fig. 1B). This produces a combined measure of the number of times the $F_{ST}$ of a gene departs from the neutral expectation between populations in two geographical regions and the strength of such departure. Genes can achieve a high summed value by having many $F_{ST}$ values within the 5% tail or by having a few very extreme $F_{ST}$ values, or more usually by a combination of both. The genes with the highest summed values contribute most to the signatures of positive selection shown in figure 2A.

We present these genes in table 2 and supplementary tables S4–S6, Supplementary Material online. We then took the five highest ranking genes in East Asia from each group (table 2 and supplementary table S4, Supplementary Material online) and performed the same ranking process using the per-SNP $F_{ST}$ values from each of these genes, this time summing values for each SNP across pairs of populations (supplementary tables S8 and S9, Supplementary Material online). This indicates those SNPs within each gene with the highest contribution to the signatures of positive selection in East Asia in figure 2A.

We further classified the populations from China into two groups: Those populations sampled from within the most Se deficient regions (pink areas of map in Oldfield, page 45; less than 0.02 ppm of Se based on local animal feeds and forages) (Oldfield 2002) and those populations that were sampled from regions likely to be Se adequate (more than 0.02 ppm of Se) or where it was difficult to tell (supplementary table S7, Supplementary Material online), making our test conservative. A two-sided Mann–Whitney U test was then performed using vectors of $F_{ST}$ ranks from these two groups to ascertain whether the signals of positive selection in Se-related genes are stronger in populations that are likely to be Se deficient (fig. 2B).

**Classic Neutrality Tests**

We computed Tajima’s $D$ and Fay & Wu’s $H$ statistics (Tajima 1989; Fay and Wu 2000) using the Perl program “Neutrality Test Pipeline” (NTP) (Andres et al. 2010). Tajima’s $D$ compares the number of segregating sites ($S$) against the average nucleotide diversity ($\pi$) in a gene. $S$ and $\pi$ are both independent estimators of $\theta$ ($4N_e\mu$) and are expected to be equal under neutrality, but behave differently when a population evolves nonneutrally. Significantly negative values of Tajima’s $D$ are consistent with positive or purifying selection (excess of low frequency alleles). Fay & Wu’s $H$ can help to distinguish positive selection from purifying selection and population size expansion in cases where Tajima’s $D$ is negative. A negative Fay & Wu’s $H$ detects the excess of high-frequency-derived alleles produced when an advantageous variant has rapidly increased to high frequencies pulling linked variation with it. $H$ has the greatest power to detect selection when the
selected allele is very close to being fixed or has just fixed in the population.

NTP runs a modified version of msstats (Thornton 2003) on genotype input data and computes a P value for each output statistic based on supplied coalescent simulations. Each gene and each control region were run individually. Fay & Wu’s H test requires an outgroup in order to distinguish the derived allele from the ancestral one. The inferred sequence of the human–chimp ancestor for each gene and control region was obtained from Ensembl (Flicek et al. 2013) and added to the NTP input. There were a number of positions at which the inferred ancestral allele did not match either the present human reference or alternative allele. These positions were removed prior to the NTP analysis because in this situation neither allele can be assigned as being derived. We used the Benjamini and Hochberg method (Benjamini and Hochberg 1995) for multiple test correction.

Coalescent Simulations
We simulated a neutral expectation using a slightly modified four population demography (Africa, Europe, Asia, and America) from Gutenkunst et al. (2009). This demographic scenario accounts for both population expansion and admixture between populations and includes a 48% admixture between the European and American populations after the split between the Asian and European populations. This model was based on a population of Mexican ancestry living in the United States. The American populations in our data set show very little European admixture (Li et al. 2008) and so this European and American admixture was removed from our demography. The use of this demographic scenario required the assignment of our 50 populations to the four populations included in the demography. This was done with reference to the k = 4 plot in Li et al. (2008). This grouped the populations as shown in supplementary table S2, Supplementary Material online, for NTP.

We performed 10,000 coalescent simulations with ms (Hudson 2002) for each gene and each pseudogene in each HGDP-CEPH population. Population and gene/pseudogene-specific details, for example, the number of chromosomes and segregating sites were used in each simulation. NTP accounts for missing genotype data by marking those positions as also missing in the ms simulations.

Data Deposition
Human gene annotations and their SNPs in the HGDP-CEPH populations are available in SelenoDB 2.0 (Romagne et al. 2014) at http://www.selenodb.org.

Supplementary Material
Supplementary tables S1–S9 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
The authors thank Raymond Burk, Kristina Hill, and Alain Krol for help with the list of genes involved in the homeostasis and metabolism of Se and Sec. They also thank Mark Stoneking for access to the HGDP-CEPH samples and Ryan Gutenkunst for advice about population admixture in these samples. They also thank Felix M. Key for help with the NTP and ms analyses. This work was supported by the Max Planck Society.

References


