Genome Biology and Evolution

SMBE

Selection for Translation Efficiency on Synonymous Polymorphisms in Recent Human Evolution

Yedael Y. Waldman[1,*], Tamir Tuller[2,3,6,*,†], Alon Keinan[4,*,†], and Eytan Ruppin[1,5,*,†]

[1]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

[2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

[3]Faculty of Mathematics and Computer science, Weizmann Institute of Science, Rehovot, Israel

[4]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York

[5]School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

[6]Present address: Department of Biomedical Engineering, Faculty of Engineering, Tel Aviv University, Tel Aviv, Israel

†These authors contributed equally to this work.

*Corresponding author: E-mail: tamirtul@post.tau.ac.il; ak735@cornell.edu; ruppin@post.tau.ac.il.

Accepted: 19 July 2011

Abstract

Synonymous mutations are considered to be "silent" as they do not affect protein sequence. However, different silent codons have different translation efficiency (TE), which raises the question to what extent such mutations are really neutral. We perform the first genome-wide study of natural selection operating on TE in recent human evolution, surveying 13,798 synonymous single nucleotide polymorphisms (SNPs) in 1,198 unrelated individuals from 11 populations. We find evidence for both negative and positive selection on TE, as measured based on differentiation in allele frequencies between populations. Notably, the likelihood of an SNP to be targeted by positive or negative selection is correlated with the magnitude of its effect on the TE of the corresponding protein. Furthermore, negative selection acting against changes in TE is more marked in highly expressed genes, highly interacting proteins, complex members, and regulatory genes. It is also more common in functional regions and in the initial segments of highly expressed genes. Positive selection targeting sites with a large effect on TE is stronger in lowly interacting proteins and in regulatory genes. Similarly, essential genes are enriched for negative TE selection while underrepresented for positive TE selection. Taken together, these results point to the significant role of TE as a selective force operating in humans and hence underscore the importance of considering silent SNPs in interpreting associations with complex human diseases. Testifying to this potential, we describe two synonymous SNPs that may have clinical implications in phenylketonuria and in Best's macular dystrophy due to TE differences between alleles.

Key words: translation efficiency, SNP, synonymous mutations, population genetics, causal variants, allele frequency differentiation.

Introduction

Synonymous mutations are traditionally considered to be "silent," as they do not affect protein sequence and are often taken as a measure for neutral evolution rate (King and Jukes 1969; Nei and Gojobori 1986; Bustamante et al. 2005; Yang 2007). However, as more and more genomic data accumulated, it became evident that synonymous mutations may have functional outcome and hence can be targeted by natural selection: they can affect splicing events, messenger RNA (mRNA) stability, microRNA binding, and nucleosome formation, sometimes even causing disorders (Chamary et al. 2006). Synonymous mutations can also influence a gene's translational efficiency (TE)—the speed or accuracy of translation—because different codons exhibit different TE, mainly due to the abundance of the corresponding transfer RNAs (tRNAs); higher tRNA abundance leads to faster and/or more accurate ribosomal translation (Bulmer 1991; Gustafsson et al. 2004; Kramer and Farabaugh 2007; Stoletzky and Eyre-Walker 2007; Hershberg

and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011). Changes in translation rate do not only result in different protein levels—thereby having a regulatory effect—but they can also affect protein function via folding (Kimchi-Sarfaty et al. 2007)—as in many cases folding is performed during translation (Komar 2009)—or via differential arginylation (Zhang et al. 2010). Studies have shown that the observed bias in codon usage in various organisms is toward codons with abundant tRNAs, marking the importance of TE (Ikemura 1981, 1982, 1985; Moriyama and Powell 1997; Percudani et al. 1997; Kanaya et al. 1999; Duret 2000; Man and Pilpel 2007; Tuller, Waldman, et al. 2010). Although the role of codon bias in relation to TE in humans is still under debate (Chamary et al. 2006), recent studies support the hypothesis that codon bias plays a significant role in TE in humans, both in normal conditions (Urrutia and Hurst 2003; Lavner and Kotlar 2005; Parmley and Huynen 2009; Waldman et al. 2010) and in cancerous mutations, where it has been shown to be targeted by natural selection (Waldman et al. 2009).

Previous studies of TE, in humans as well as in other organisms, often considered a single "reference" genome and measured codon bias in this genome with respect to gene expression (Urrutia and Hurst 2003; dos Reis et al. 2004), tRNA pool (dos Reis et al. 2004), and other parameters (Akashi 1994; Stoletzky and Eyre-Walker 2007). These studies did not seek patterns of natural selection on TE or its evolution, but a few recent studies analyzed interspecies selection on TE between different yeast species (Man and Pilpel 2007; Zhou et al. 2010), different worm species (Zhou et al. 2010), between eubacterial and archaeal organisms (Chen et al. 2004), and between human and chimpanzee (Comeron 2006). The availability of data on intraspecies variation can help uncover evidence for TE selection by considering a more refined timescale. Specifically, human population genetic data can hold evidence of selection in the last tens of thousands of years of human evolution (Nielsen et al. 2007; Novembre and Di Rienzo 2009; Keinan and Reich 2010). A recent analysis of differences in allele frequencies between human populations showed that nonsynonymous single nucleotide polymorphism (SNPs), which alter protein sequences, are under stronger selection, both positive and negative, as compared with other SNPs that do not change protein sequences (Barreiro et al. 2008). A similar methodology can be employed to analyze TE selection in recent human evolution, and Comeron (2006) applied a similar approach on two populations, though his results concerning selection on TE have been limited by the small sample size of $<$500 SNPs in 90 chromosomes that was available at that time.

Here we perform the first genome-wide analysis of natural selection related to TE in recent human evolution. The usage of genome-wide data enables us not only to determine that there is selection for TE in recent human evolution but also to address related questions that were never addressed before in this scope: what are the factors influencing TE selection? Does this force vary between different genes or between different parts of genes? And, do the targets of TE selection correlate with those of natural selection in general? Our results show that natural selection has been operating on TE in recent human evolution. Moreover, we find marked differences in TE selection between different classes of genes and within different locations along gene sequences that are related to both translation rate and accuracy.

Materials and Methods

SNPs Data Sets

We obtained SNP data from the HapMap3 project (The International HapMap 3 Consortium 2010), release 3. Using annotation from dbSNP (Sherry et al. 2001) build 130, we focused on 30,080 coding SNPs, 13,798 of which are synonymous. Genotype data are obtained from HapMap3 for 1,198 unrelated individuals from 11 populations (supplementary table S1, Supplementary Material online). As a further validation, we repeated some of our analyses using allele frequency information from exon sequencing (34,983 coding SNPs, of which 18,608 are synonymous) of individuals from two population samples, African Americans and European Americans (Lohmueller et al. 2008).

TE Measure

We calculated a codon's TE following dos Reis et al. (2004). Briefly, Let n_i be the number of tRNA isoacceptors recognizing codon i. Let tCGN_{ij} be the genomic copy number of the jth tRNA recognizing the ith codon, and let S_{ij} be the selective constraint on the efficiency of the codon–anticodon coupling. We define the absolute adaptiveness, W_i, for each codon i as follows:

$$Wi = \sum_{j=1}^{n_i} (1 - Sij)\mathrm{tCGN}_{ij}$$

W_i is a measure for the abundance of the tRNAs that translate codon i. As both translation rate and accuracy are influenced by tRNA abundance (Kramer and Farabaugh 2007; Stoletzky and Eyre-Walker 2007; Gingold and Pilpel 2011), W_i is a measure for TE, both for translation elongation rate as well as accuracy. Let i and j be the two codons defined by the two alleles of a given SNP. Following Waldman et al. (2009), we define the ratio W_i/W_j to be the measure for the effect on TE (ΔTE) of the two codons i and j. As the order of alleles is arbitrary, we considered the bigger ratio (i.e., $\Delta\mathrm{TE} \geq 1$ for all SNPs). We decided to use this approach rather than define preferred (i.e., with larger TE) and unpreferred alleles because efficiency is context

dependent, and in some cases (e.g., near the beginning of transcripts), the codon that is translated more slowly is more optimal (or improves the fitness) to the organism (Fredrick and Ibba 2010; Tuller et al. 2010; Gingold and Pilpel 2011). As our study is genomic (SNPs) and not tissue specific, we followed dos Reis et al. (2004) and used genomic copy number (Chan and Lowe 2009) as a global measure for tRNA levels. Nevertheless, this measure was previously shown in detail to be adequate in TE analysis in humans, also in tissue-specific context (Waldman et al. 2010). S_{ij} values were taken from table 2 in dos Reis et al. (2004).

TE Evolution since Human/Mouse Divergence

In our analysis of TE evolution, we calculated the TE of an entire gene based on its codon composition (dos Reis et al. 2004). Let W_i be the measure for TE for codon i, as defined above. By normalizing W_i's values (dividing them by the maximal W_i), we obtain w_i, the relative adaptiveness value of codon i. The tRNA adaptation index (tAI) of a gene g is the geometric mean of its codons:

$$\mathrm{tAI}(g) = \left(\prod_{k=1}^{lg} w_{ikg}\right)^{1/lg},$$

where i_{kg} is the codon defined by the k'th triplet on gene g and lg is the length of the gene (excluding stop codons). tAI of human and mouse orthologues was calculated for each species using its own genomic tRNA pool (Chan and Lowe 2009). Orthology information was retrieved from the Mouse Genome Informatics (Bult et al. 2008).

F_{ST} Calculation

To estimate allele frequency differentiation across populations, we used the F_{ST} statistic as formulated by Keinan et al. (2007). F_{ST} captures the fraction out of the variation in allele frequencies that is attributed to between-population variation rather than within-population diversity. As a consequence, SNPs with similar allele frequencies between populations are assigned lower F_{ST} estimates, whereas differences in allele frequencies will yield higher F_{ST} estimates (Weir and Cockerham 1984). As F_{ST} was originally defined for two populations, its values are between 0 (no difference between populations) and 1 (complete difference, i.e., the SNP is fixed for one allele in one population and fixed for the other allele in another population). However, our estimator is unbiased and therefore can produce slightly negative estimates of F_{ST}, which should not affect our calculations because the effect will not depend on ΔTE.

Formally, let p_i be the frequency of a variant in biallelic SNP in each of two populations ($i = 1,2$). Set $q_i = 1 - p_i$ to be the frequency of the other variant in population i. We define F_{ST} as N/D where

$$N = p_1(q_2 - q_1) + p_2(q_1 - q_2)$$

and

$$D = p_1q_2 + q_1p_2.$$

We use the following estimators for N and D:

$$\widehat{N} = (a_1/n_1 - a_2/n_2)^2 - \frac{h_1}{n_1} - \frac{h_2}{n_2}$$

and

$$\widehat{D} = \widehat{N} + h_1 + h_2$$

where a_i and n_i are allele count and total number of alleles for population i and h_i is the heterozygosity estimate for population i:

$$h_i = \frac{a_i(n_i - a_i)}{n_i(n_i - 1)}.$$

We generalize this definition to more than two populations as follows: Let k be the number of populations examined. For each pair of populations i and j, N_{ij} and D_{ij} are defined as above. A global F_{ST} can then be defined as $\frac{N_\Sigma}{D_\Sigma}$, where

$$N_\Sigma = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} N_{ij}, \quad D_\Sigma = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} D_{ij}.$$

Further description and analysis of this definition of F_{ST} can be found in Keinan et al. (2007) and in Reich et al. (2009).

Gene Sets

Protein complexes data were downloaded from the CORUM database (Ruepp et al. 2010) for 2,527 genes (3,957 [2,037 synonymous] SNPs). For each gene, we determined complex size as the number of proteins forming it and averaged across complexes in cases where the gene participates in more than one complex.

We defined essential genes as described in Waldman et al. (2010) (1,765 essential genes and 1,836 [891 synonymous] SNPs).

Protein–protein interaction data were taken from Bossi and Lehner (2009) (10,024 proteins and 78,799 interactions; 16,386 [8,053 synonymous] SNPs). Expression data for 12,726 genes with 20,133 (9,720 synonymous) SNPs were taken from Su et al. (2004). Based on 30 adult human tissues, we defined expression rate and breadth as the mean expression level and the number of tissues in which the gene is expressed, respectively (Waldman et al. 2010). dN and dS values (human/mouse, human/chimpanzee) were downloaded from BioMart (Durinck et al. 2005). Gene functional classification was downloaded from Gene Ontology (GO) (Ashburner et al. 2000).

Functional Annotation

Division of gene regions into functional and nonfunctional regions was based on InterPro (Hunter et al. 2009) classification and was done for genes with at least one functional region (i.e., annotated genes). According to this definition, we had 17,163 (8,466 synonymous) SNPs within functional regions and 8,939 (3,919 synonymous) SNPs outside these regions in 10,026 genes.

Significance Assessment

Significance test was performed similar to the scheme used by Barreiro et al. (2008) when comparing between different classes of SNPs. For each group of SNPs of interest, we ranked the SNPs according to their ΔTE measure. Next, we considered two groups: those SNPs with ΔTE above the median ΔTE ("large" ΔTE) as SNPs with relatively large effect on TE (mean ΔTE = 1.93) and those below the 10th percentiles ("small" ΔTE) as SNPs with almost no effect on TE (mean ΔTE = 1.07) as a baseline. Under the null hypothesis of no TE selection, we expect to see no difference in F_{ST} measures between the two SNP groups in negative (low F_{ST} values) or positive (high F_{ST} values) selection. To test for the significance of the difference between the two groups in negative (positive) selection, we defined the second (98.5th) percentile of F_{ST} values as a threshold and compared the number of SNPs in each group below (above) that threshold using a χ^2 test with one degree of freedom (taking the small ΔTE group as our expected distribution). For each group, we measure the fraction of SNPs passing the cutoff, and the enrichment reported in this work is the ratio between the two fractions. Results were robust across other ΔTE and F_{ST} percentiles (supplementary note 2, Supplementary Material online).

When analyzing groups with smaller number of SNPs (GO terms, complex and essential genes, and in the first 100 codons of genes), we used the fifth (95th) percentile as a threshold for negative (positive) selection. When focusing on the first 100 codons of highly/lowly expressed genes, we took the 10th and 90th percentiles as thresholds for selection.

In addition, we also measured the significance of enrichment of the group (e.g., a specific GO term) as compared with the genome-wide enrichment: we compared the number of SNPs that passed the threshold in the specific group of interest with that expected to pass it according to the F_{ST} and ΔTE distributions of all the SNPs outside this group. Again, this was done using a χ^2 test with one degree of freedom.

For GO analysis, we focused on terms with at least 700 synonymous SNPs (to have enough detection power) and with at most 4,000 SNPs (i.e., ignoring too general terms). When two terms were almost identical (Jaccard index between gene groups being above 0.95), we removed the term with smaller number of synonymous SNPs, resulting in 130 GO terms. We considered only results that were significantly enriched as compared with the global enrichment after false discovery rate correction, performing this for each ontology ("molecular function," "biological process," and "cellular component") separately. When we compared between selection within GO term with selection outside this term, we focused only on SNPs within genes in some GO term to avoid bias between annotated and unannotated genes.

Evidence for general selection was done in a similar way. For general selection in recent human evolution (human lineage), we used the F_{ST} (dN/dS and dN for human/chimpanzee and human/mouse) percentiles as in TE selection. Next, we measured how many SNPs (genes) passed the threshold in the group being analyzed and compared it (χ^2 test with one degree of freedom) to the expected number of SNPs (genes) outside this group. In difference from the TE analysis, we considered all coding SNPs for this analysis.

All correlations reported in this work are nonparametric (Spearman correlation).

Results

Analysis Overview

We examined the differentiation in allele frequencies of individual SNPs between different human populations via F_{ST} (Weir and Cockerham 1984; Holsinger and Weir 2009) using an unbiased estimator that is insensitive to differential sample sizes (Keinan et al. 2007). SNPs with very low values of F_{ST} are those with almost no differentiation between populations, suggestive of negative selection. On the other hand, SNPs with very high levels of differentiation between populations exhibit high F_{ST} values, which is suggestive of geographically localized positive selection that drives allele to high frequency in some but not all the populations between which F_{ST} is measured (Weir and Cockerham 1984; Barreiro et al. 2008; Holsinger and Weir 2009; Keinan and Reich 2010). Under the assumption of neutrality, F_{ST} is determined by genetic drift alone and should exhibit a consistent distribution across all SNPs throughout the genome. Deviation from this distribution for a specific set of SNPs suggests that natural selection has targeted SNPs from this set more often than the genome-wide background (Barreiro et al. 2008). To test for natural selection on TE in recent human evolution, we examined whether SNPs with different levels of effect on TE (ΔTE) also exhibit a different distribution of F_{ST} values. This approach is in the same spirit as that used in a recent study, which showed that coding SNPs are under stronger negative selection compared with noncoding SNPs and that nonsynonymous SNPs and SNPs in the 5′ untranslated regions are under stronger positive selection (Barreiro et al. 2008). To estimate ΔTE, we first calculated

the TE of each codon (Waldman et al. 2009). This TE measure is based on the abundance of the corresponding tRNAs that translate the codon and quantifies both translation rate and accuracy as tRNA abundance is a major factor for both features (Kramer and Farabaugh 2007; Stoletzky and Eyre-Walker 2007; Gingold and Pilpel 2011). We then estimated ΔTE for each SNP as the ratio between the TE of the two codons resulting from the SNP's two alleles (Materials and Methods).

Using the HapMap3 data set (The International HapMap 3 Consortium 2010), we obtained genome-wide allele frequency data for 13,798 synonymous SNPs in 7,957 autosomal human genes in 1,198 unrelated individuals from 11 populations (Materials and Methods). We considered F_{ST} between this set of 11 populations, which is sensitive to the impact of natural selection after these populations have split.

The Genome-wide Pattern of TE Selection

Our analysis showed evidence for genome-wide selection for TE. Thus, focusing on negative selection, SNPs with larger ΔTE show enrichment for extremely low F_{ST} values (below second percentile) as compared with SNPs with lower values (fig. 1*A*). Notably, TE manifests a 2.45-fold change (χ^2 test, $P = 1.44 \times 10^{-10}$) in enrichment among synonymous SNPs, which is comparable to the 2.54-fold enrichment ($P < 10^{-16}$) found between synonymous and noncoding SNPs (supplementary note 1, Supplementary Material online). Turning to positive selection, SNPs with larger ΔTE show enrichment for extremely high F_{ST} values (above 98.5th percentile) as compared with SNPs with lower values (fig. 1*B*). Again, TE manifests a 1.57-fold change (χ^2 test, $P = 1.01 \times 10^{-2}$) among synonymous SNPs that is even larger than the 1.15-fold ($P = 2.49 \times 10^{-2}$) change between synonymous and noncoding SNPs.

Similarly, we contrasted the distribution of F_{ST} between two categories of synonymous SNPs: those with almost no effect on TE (below the 10th percentile; mean ΔTE = 1.07), that were taken as a baseline, and those with large ΔTE (above the median; mean ΔTE = 1.93). Comparison between the two classes can allow exploring the selection on TE. In agreement with the above results, we found a 1.78-fold ($P = 9.77 \times 10^{-13}$) and a 1.48-fold ($P = 2.01 \times 10^{-5}$) enrichment in low and high F_{ST} values, respectively, in large ΔTE SNPs as compared with small ΔTE SNPs (supplementary fig. S1, Supplementary Material online). Similar results were obtained for other definitions as well, reflecting the robustness of the results (supplementary note 2, Supplementary Material online).

Even within SNPs of already extreme F_{ST} values, F_{ST} correlates with their ΔTE: focusing on the 500 SNPs with lowest F_{ST} values, we partitioned the data based on ΔTE values and found a significant negative correlation between ΔTE and F_{ST} ($R = -0.534$, $P = 7.80 \times 10^{-3}$; fig. 1*C*), as suggested if negative selection explained the lower F_{ST} values. Similarly, focusing on the 350 SNPs with highest F_{ST} values, we found a significant positive correlation between ΔTE and F_{ST} ($R = 0.554$, $P = 2.30 \times 10^{-2}$; fig. 1*D*) as expected by more extensive positive selection on larger ΔTE SNPs.

These results remain significant after controlling for various factors such as GC content, recombination rate, exon splicing signals, and mutational biases such as biased gene conversion (supplementary notes 3–4, fig. S1, and table S2, Supplementary Material online). We next turned to investigate whether TE selection has had more of an impact on certain groups of genes and in certain positions within genes. Similar to the genome-wide analysis, we contrasted F_{ST} distribution between two categories of synonymous SNPs: those with small ΔTE (below the 10th percentile) and those with large ΔTE (above the median) and looked for enrichment in extreme F_{ST} values in the last group.

Gene TE Selection Is Influenced by Connectivity, Expression, and Essentiality

Previous studies have shown that several measures, conventionally considered to be related to the functional importance of genes, affect selection in general: connectivity (Fraser et al. 2002), expression (Drummond et al. 2005), and essentiality (Liao et al. 2006). Moreover, some studies support the hypothesis that highly expressed genes evolve slowly due to TE constraints (Akashi 2001). We found that TE selection is also affected by these measures. Negative TE selection was stronger in both highly connected proteins in the human interactome (4.81-fold enrichment, $P \ll 10^{-16}$) and in highly expressed genes (1.67-fold enrichment, $P = 4.69 \times 10^{-5}$; expression rate). Both lowly expressed genes and lowly interacting proteins showed no evidence for TE selection, and the difference between the groups was significant (supplementary table S3, Supplementary Material online). Positive TE selection was found in lowly (4.15-fold enrichment, $P \ll 10^{-16}$) but not highly ($P = 0.89$) interacting proteins (fig. 2). Highly expressed genes were slightly more affected by positive TE selection but not significantly different than lowly expressed proteins (supplementary fig. S2, Supplementary Material online). Similar results were obtained for expression breadth (supplementary fig. S2, Supplementary Material online). TE may be more preserved in highly expressed genes as there are more mRNA copies that potentially consume more ribosomes and also more protein copies that may have translation errors, resulting in stronger TE selection. Turning to essential genes (where we used slightly different threshold due to smaller number of SNPs; Materials and Methods), there was a 2.16-fold enrichment ($P = 3.02 \times 10^{-5}$) for negative TE selection, but the enrichment was not significantly higher than nonessential genes ($P = 0.09$). Interestingly, we observed a decrease in positive

Genome Biology and Evolution

SMBE

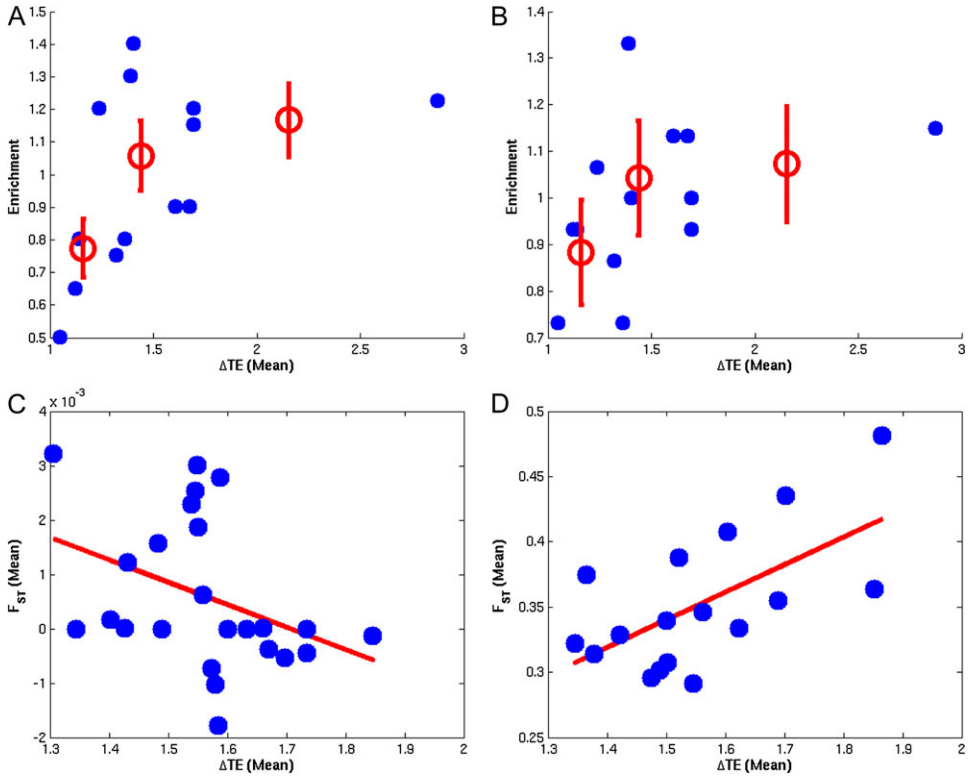

Fig. 1.—Genome-wide selection for TE. Using nonoverlapping bins (1,000 SNPs in each bin, 1,798 in the last bin), we found a correlation between mean ΔTE and the enrichment in (*A*) low F_{ST} (below the second percentile; negative selection) and (*B*) high F_{ST} (above the 98.5th percentile; positive selection) values in each bin ($R = 0.64$, $P = 9.00 \times 10^{-3}$ and $R = 0.55$, $P = 2.55 \times 10^{-2}$ for negative and positive selection, respectively; Spearman correlation). Similarly, we divided the data into three equal bins and measured the enrichment in each bin for extreme F_{ST} values. To assess the significance of the difference between the three bins, we performed 10,000 bootstrapping sampling and compared the distributions ($P \ll 10^{-16}$ between all bins, Wilcoxon test). Mean and standard deviation of enrichment of each bin are shown in red. Second, we focused on SNPs with extreme F_{ST} values (500 and 350 SNPs with lowest F_{ST} [$F_{ST} < 0.0036$] and highest F_{ST} [$F_{ST} > 0.289$] values for negative and positive selection, respectively). We divided these SNPs into nonoverlapping equally sized bins (20 SNPs in each bin) according to their F_{ST} measure and found significant correlation between F_{ST} and ΔTE for both (*C*) negative ($R = -0.534$, $P = 4.07 \times 10^{-3}$) and (*D*) positive ($R = 0.554$, $P = 1.10 \times 10^{-2}$) selection. Results are also significant without binning ($R = -0.141$, P value $= 1.55 \times 10^{-3}$ and $R = 0.113$, P value $= 0.034$ for negative and positive selection, respectively).

TE selection in essential genes (0.40-fold enrichment, $P = 7.18 \times 10^{-6}$). This decrease is particularly pronounced when comparing with the enrichment in nonessential genes ($P = 2.94 \times 10^{-10}$).

Groups of Genes Enriched for TE Selection

Complex members tend to have similar protein abundances, presumably for efficient production of the complex (Carmi et al. 2006, 2009; Tuller et al. 2007). As TE changes may affect protein levels and hence disrupt complex formation, we hypothesized that negative TE selection will be more prevalent within complex members. Indeed, we found a 3.08-fold enrichment ($P \ll 10^{-16}$) for negative TE selection within complexes. This enrichment is significantly higher as compared with the enrichment found in SNPs in noncomplex genes ($P = 3.84 \times 10^{-5}$). Furthermore, TE selection varies between complexes according to their size: genes in large complexes (mean complex size of 25.9 genes) exhibited a 5.37-fold enrichment ($P \ll 10^{-16}$) for negative TE selection as compared with no significant enrichment in genes in small complexes (mean complex size of 3.11 genes). Positive TE selection analysis did not reveal significant differences between complex and noncomplex members ($P = 0.82$).

In addition, we used the GO classification (Ashburner et al. 2000) and found that several GO terms show TE selection significantly different than the genome-wide TE pattern (fig. 3, supplementary tables S5 and S6, Supplementary Material online; Materials and Methods). Interestingly, several of these groups (for both positive and negative TE selection) contain regulatory genes—keynote genes in both inter- and intraspecies variation (Levine and Tijan 2003; Chen and Rajewsky 2007).

TE Selection along the Gene

TE selection may vary along the gene's sequence. For example, codon bias is known to be stronger in conserved sites, which are presumably more important for function,

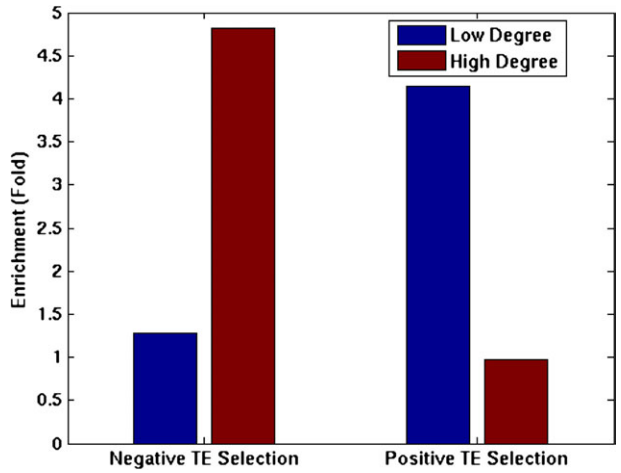

Fig. 2.—TE selection and connectivity in the human protein–protein interaction network. Negative TE selection is stronger in lowly interacting proteins (i.e., below the median of five interactions; 2,384 genes, 4,156 SNPs), whereas positive TE selection is stronger in highly interacting proteins (i.e., with more than five interactions; 2,016 genes, 3,897 SNPs). See also supplementary table S2 (Supplementary Material online).

apparently to reduce translation errors of amino acid substitution in these regions (Akashi 1994; Stoletzky and Eyre-Walker 2007). Using InterPro (Hunter et al. 2009) classification, we found a 2.23-fold enrichment ($P = 1.11 \times 10^{-16}$) for negative TE selection within functional regions while no evidence for selection outside these regions (supplementary table S3, Supplementary Material online). There was no difference between the two regions in respect to positive TE selection ($P = 0.43$).

In contrast, less efficient codons are favored near the start sites of genes, presumably to reduce ribosomal collisions and minimize protein production cost (Tuller et al. 2010). Such a ramp may also prevent various types of translation errors (such as truncated proteins) that may cause toxic misfolded proteins (Drummond and Wilke 2008, 2009). Hence, we hypothesized that negative TE selection will be stronger in the initial segments of genes to maintain this "ramp." Moreover, we expected stronger selection on highly expressed genes where efficient translation is more critical. Indeed, we found a 1.53-fold enrichment ($P = 5.27 \times 10^{-3}$) for negative TE selection in the first 100 codons of highly expressed genes (748 SNPs). This enrichment was higher as compared outside this region, but the difference was only borderline significant ($P = 0.08$). Interestingly, we found evidence for positive selection in the first 100 codons, mainly in lowly expressed genes: there was a 3-fold enrichment ($P = 1.15 \times 10^{-14}$) for positive TE selection. This enrichment was significant as compared with SNPs outside this region ($P = 1.17 \times 10^{-6}$; see supplementary note 5, Supplementary Material online, for additional results).

TE Selection and General Selection

An interesting question concerns the interplay between TE selection in recent human evolution and other general forces of selection that are unrelated to TE. Do these forces show similar or different selection patterns? To address this intriguing question, we used several approaches.

First, we directly checked whether genes under selection show stronger TE selection. For that purpose, we used the ratio between nonsynonymous (dN) and synonymous substitutions (dS) between human and chimpanzee as a measure for selection pressure on protein sequence since the last common ancestor. SNPs within genes below the median dN/dS, suggestive of stronger constraints imposed on protein sequence, exhibited 2.55-fold enrichment ($P \ll 10^{-16}$) for negative TE selection that was significantly higher than the group above the median ($P = 0.01$). Similar results were obtained for dN/dS values between human and mouse (supplementary table S3, Supplementary Material online).

Second, we examined whether TE selection pattern in various gene sets resembles that of other selective pressures that are not related to TE. These general selective pressures were measured using dN/dS ratio between human and chimpanzee and with F_{ST} within human populations (Materials and Methods). Remarkably, we found that for many of the gene groups analyzed above, TE selection resembled general selection in the two timescales (supplementary tables S3–S6, Supplementary Material online). Similar results were also obtained by taking dN values (human/chimpanzee) and dN/dS values (human/mouse) (supplementary tables S4–S6, Supplementary Material online).

Nevertheless, there were some differences between TE and general selection, mainly in respect to positive selection. Specifically, regulatory genes (defined by GO classification) showed decrease in positive selection on protein sequence (dN/dS measure) as opposed to enrichment in TE selection (supplementary table S6, Supplementary Material online). Regulatory genes can evolve through changes in abundance or sequence and function (Wittkopp 2005). These results suggest that the former (via TE changes) are more common than functional (nonsynonymous) changes. Indeed, transcription factors show enrichment for positive selection in their expression levels in the human lineage while showing lower dN/dS values (Blekhman et al. 2008). Similarly, metabolic genes (GO:0008152) showed enrichment for both positive and negative TE selection, whereas only negative and even decrease in positive general selection (supplementary table S3, Supplementary Material online), in agreement with studies showing that metabolic genes underwent positive selection in the human lineage in respect to expression levels (Khaitovich et al. 2006; Blekhman et al. 2008).

TE Selection on Codons and Genes

In our analysis thus far, we focused on TE selection as reflected by SNPs. Hence, we study TE selection on single

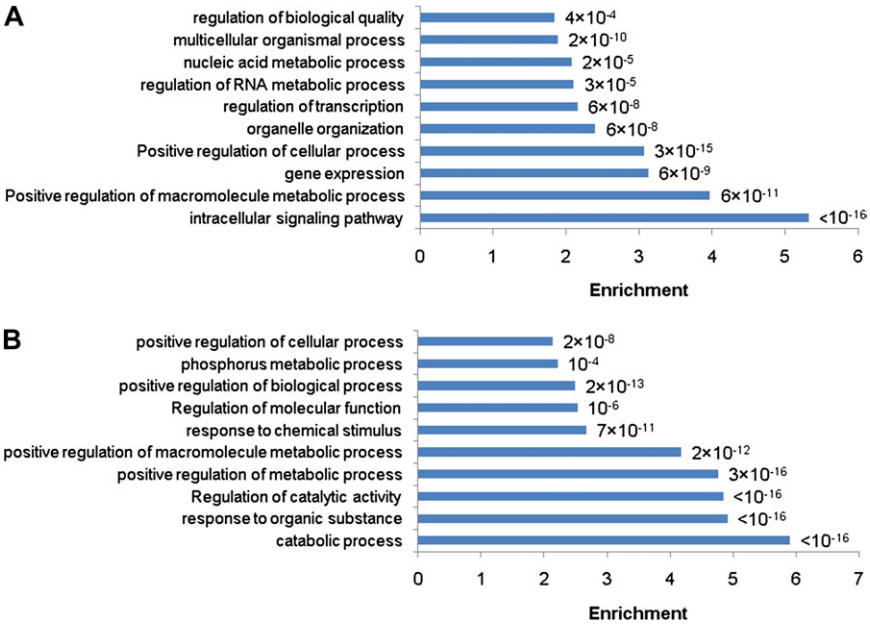

Fig. 3.—GO terms with significant TE selection. Ten most significant terms under (*A*) negative and (*B*) positive TE selection are shown. *P* values to the right of each bar indicate the significance of the enrichment. For additional results and details see supplementary tables S3–S4 (Supplementary Material online).

nucleotides in recent human evolution. Similarly, we can also look on TE selection in larger timescales, from a whole gene's perspective: during evolution, changes in codon composition can also affect the TE of the entire gene. It is interesting to examine what is the relationship between changes in TE of genes during evolution and selection on TE in specific codons in recent human evolution within these genes. For that purpose, we calculated for each gene its TE, both in human and in mouse orthologues and compared between the TE rankings of each gene in the two species (Materials and Methods). This allowed us to analyze different set of genes, according to their TE evolution since human/mouse divergence. Interestingly, we found that genes that changed their TE during human evolution (either increase or decrease as compared with mouse) showed stronger TE selection than those with little change in TE. This trend was observed for both negative (fig. 4*A*) and positive (fig. 4*B*) TE selection. However, negative TE selection was stronger in genes showing increase in TE in human evolution as compared with genes with decreased TE in human: there was a 1.94-fold enrichment ($P = 2.64 \times 10^{-10}$) in genes with increased TE in human but only a 1.54-fold enrichment ($P = 5.95 \times 10^{-4}$) in genes with decreased TE in human. The difference between the groups was highly significant ($P = 8.20 \times 10^{-4}$; see also fig. 4*A*).

These results demonstrate that genes under stronger TE selection in human evolution, with respect to mouse, are also under stronger TE selection in recent human evolution.

TE Variants Potentially Involved in Diseases

The results presented above imply that ΔTE synonymous variants are not silent and hence may have significant and even clinical implications. As preliminary results, we present several examples demonstrating the potential value of TE in SNP analysis. The first example is rs1042503, a synonymous SNP (c.735G>A, p.V245V) located in the phenylalanine hydroxylase gene, which encodes a rate-limiting enzyme in phenylalanine catabolism. Mutations in this gene lead to phenylketonuria (PKU) disease, a disease that if not treated properly causes impaired cognitive development and neurological function (Scriver 2007). This SNP exhibits a very high F_{ST} (0.42, 99.6th percentile) and a relatively high value of ΔTE (3.52, 95th percentile). Using each of these measures alone would probably not underscore its possible importance. However, there are only four other synonymous SNPs (out of 13,798 SNPs) with both higher F_{ST} and ΔTE (supplementary table S7, Supplementary Material online). Notably, although most mutations associated with PKU in this gene are nonsynonymous (Scriver et al. 2003), this synonymous variant was also associated with PKU and is among the most associated mutations in this gene (Dworniczak et al. 1990; Scriver et al. 2003). The PKU-associated allele (GTA) is translated less efficiently than the other allele (GTG) and thus can cause ribosomal stalling and leading to aberrant folding (Kimchi-Sarfaty et al. 2007; Komar 2009) or higher probability for translation error (Kramer and Farabaugh 2007; Stoletzky and Eyre-Walker 2007). Interestingly, the associated allele is almost absent in African

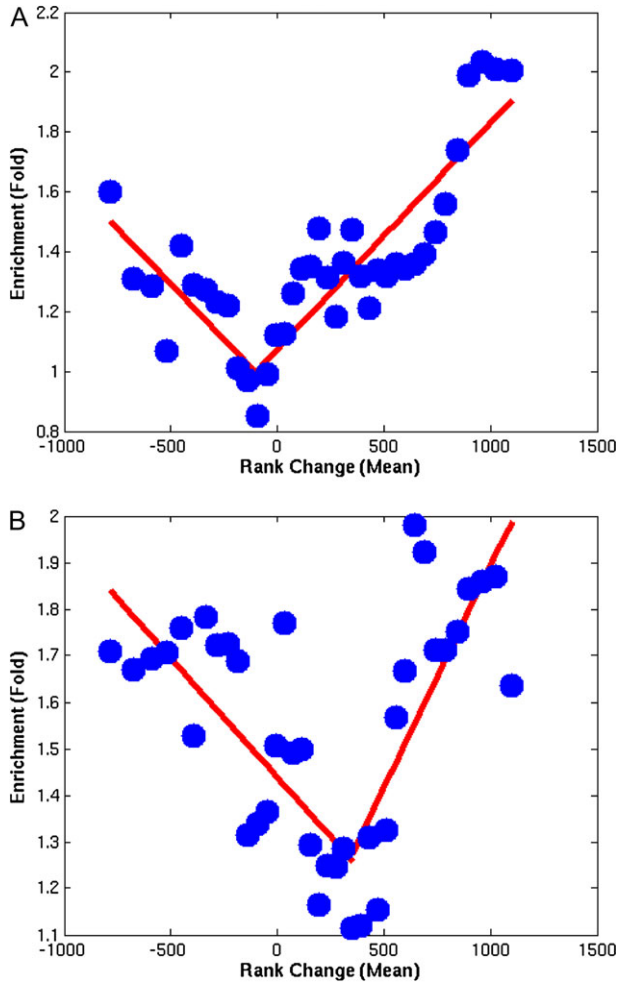

Fig. 4.—TE selection in genes varies in respect to their TE evolution. For each gene, we computed its TE in human and mouse and compared between their rankings. Next, we sorted all genes according to the change in their ranking in the two species and used sliding windows to compute enrichment for (*A*) negative and (*B*) positive TE selection (3,500 genes in each window, 100 genes difference between consecutive bins). Positive rank change reflects increase in TE in human as compared with mouse.

populations (fig. 5*A*), in concordance with PKU prevalence which is at least one order of magnitude lower in sub-Saharan populations as compared with European and Asians (Scriver 2007; Hardelid et al. 2008). Positive selection to increase the associated allele in non-African population may be a result from an overdominant selection where the heterozygote allele has an advantage (Krawczak and Zschocke 2003).

Another interesting example is the rs1109748, also among the top five SNPs with both high F_{ST} (0.44) and high of ΔTE (3.61) (supplementary table S7, Supplementary Material online). This SNP (c.219C>A, p.I73I) is located in a transmembrane region of BEST1 (also known as VMD2) and has the highest F_{ST} among all HapMap3 SNPs (coding and noncoding) within 500 kb. Mutations in BEST1 cause Best's macular dystrophy, a rare retina disorder (Petrukhin et al. 1998). Interestingly, a nonsynonymous mutation in the same residue of the SNP (c.218T>A, p.I73N) was found in a patient with Best's macular dystrophy (Marchant et al. 2001) and was shown to have a measurable effect on membrane insertion (Milenkovic et al. 2007), marking the importance of this residue. Changes in TE can increase translation error and amino acid substitutions of the residue or affect protein folding. The frequency of the more efficient allele (C) is much higher in both European and African populations as compared with East Asian populations (fig. 5*B* and *C*), suggesting differential prevalence rates of Best's macular dystrophy between these populations. However, Best's macular dystrophy is a rare disease, and therefore, prevalence rates in different populations are currently unavailable to the best of our knowledge.

One can also use other measures to detect selection and combine them with TE. For example, a recent study set out to find causal variants within regions under recent positive selection using a composite of multiple signals (CMS) (Grossman et al. 2010). In their analyses, they mainly focused on nonsynonymous variants or variants within regulatory regions. We reanalyzed their results, focusing on silent variants. We found that for the region 43,400,000–43,600,000 in chromosome 19, the SNP rs3178327 had both high genome-wide CMS value (12.904, $P = 4.5 \times 10^{-5}$, placing it second in this region and first among coding variants in this region) and high ΔTE (3.52), suggesting that its ΔTE may explain its relatively high probability to be a casual variant. This polymorphism is between two valine alleles (GTG and GTA) and is located within a transmembrane region of the YIF1B protein. Further analysis is needed to verify the possibility of this SNP to be a casual variant. Supplementary table S8 (Supplementary Material online) contains CMS and ΔTE values for other synonymous SNPs, some of them may be potentially casual variants.

Discussion

In this work, we showed that TE has been targeted by natural selection, both positive and negative, during relatively recent human evolution. In addition to the observed genome-wide TE selection, there are marked differences between and within genes. Negative TE selection is stronger in complex members and essential genes, as well as in highly expressed and highly interacting proteins. Stronger negative TE selection was also observed in initial segments of highly expressed genes and in functional regions and positive TE selection in the same region for lowly expressed genes. Interestingly, we found that regulatory genes are under stronger than average TE selection, both positive and negative, in agreement with previous studies that have highlighted the importance of regulation in inter- and intraspecies variation

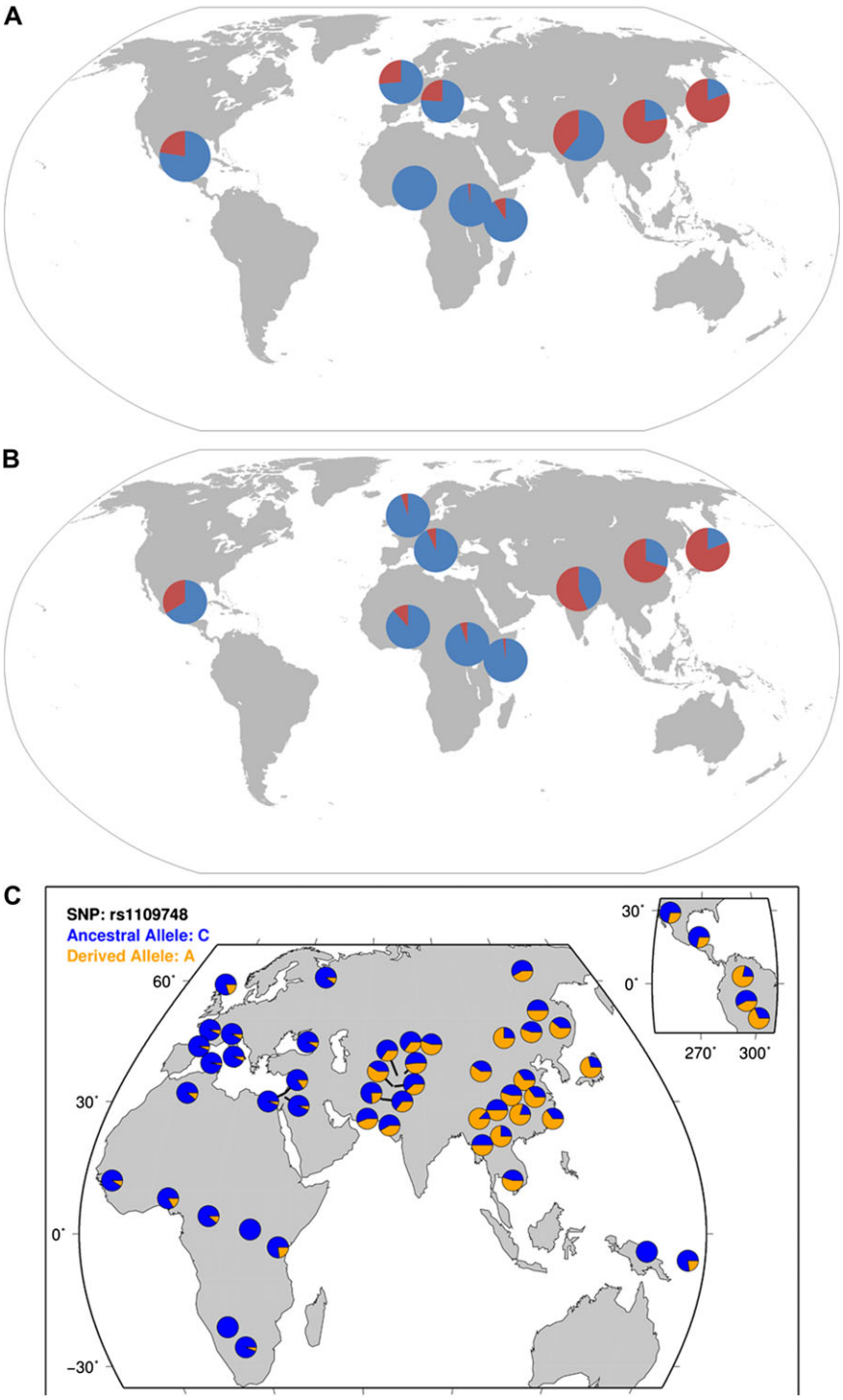

Fig. 5.—Worldwide allele distributions for variants putatively involved in diseases. Allele frequencies are shown (*A*) for rs1042503 (phenylalanine hydroxylase gene) in HapMap3 populations and for rs1109748 (BEST1 gene) in (*B*) HapMap3 and in (*C*) HGDP populations. These two synonymous SNPs show both very high F_{ST} value (>0.4) and very large ΔTE (>3.5). The former, showing a pattern consistent with positive selection outside Africa, has been clinically associated with PKU, whereas the latter, showing a pattern consistent with selection in Asian populations, is located in a residue associated with Best's macular dystrophy (Results). For HapMap3 data, we pooled together the CHB and CHD samples and ignored the ASW population (see supplementary table S1, Supplementary Material online). Panel (*C*) was downloaded from the HGDP selection browser (Pickrell et al. 2009).

(Levine and Tijan 2003; Chen and Rajewsky 2007). Furthermore, TE selection shows similar pattern across gene classes to that of general selection in recent human evolution as well as general selection on a much deeper timescale (supplementary table S3, Supplementary Material online). Nevertheless, we also found differences, suggesting that positive selection in regulatory and metabolic genes is mainly obtained via changes in protein levels (TE changes) and less by nonsynonymous changes, in accordance with earlier studies (Khaitovich et al. 2006; Blekhman et al. 2008).

Ascertainment bias in SNP detection between different groups of genes (Clark et al. 2005; Keinan et al. 2007; Nielsen et al. 2007; Barreiro et al. 2008) can potentially influence some of the results concerning general selection in recent human evolution, and therefore they should be treated with caution. General selection in recent human evolution was the subject of many previous studies (Voight et al. 2006; Nielsen et al. 2007; Sabeti et al. 2007; Barreiro et al. 2008; Akey 2009; Tennessen et al. 2010) and is not the central subject of the current study. Ascertainment biases should not affect the main focus of this study—TE selection—because the groups we compared (small and large ΔTE SNPs) are both synonymous and were taken from the same set of genes. Nevertheless, we repeated some of our analyses on a set of SNPs discovered by homogenous sequencing (Lohmueller et al. 2008). Despite its smaller size (35 individuals from 2 populations as compared with 1,198 individuals from 11 populations), we found evidence for TE selection also in this data set: a 1.32-fold enrichment (χ^2 test, $P = 1.34 \times 10^{-4}$) for low F_{ST} values (below second percentile) in large ΔTE SNPs as compared with small ΔTE SNPs, indicating negative TE selection. Taking a slightly different threshold (below fourth percentile), the results were even more significant: a 1.43-fold enrichment (χ^2 test, $P = 2.59 \times 10^{-13}$). Similarly, we found a 1.31-fold enrichment ($P = 8.14 \times 10^{-4}$) for high F_{ST} values (above 98.5th percentile), testifying for positive selection. As more and more data that is based on full genome sequencing will become available, it will be interesting to repeat the analyses reported here.

The importance of this work lies in several aspects. First, it further highlights the importance of TE in humans (Urrutia and Hurst 2003; Lavner and Kotlar 2005; Parmley and Huynen 2009; Waldman et al. 2010). However, in contrast to previous studies, we focused on recent human evolution. Thus, we provide for the first time a genome-wide evidence for TE selection in humans in a relatively recent epoch. Similarly, other studies used polymorphism data to study TE in Drosophila (Akashi and Schaeffer 1997; Akashi 1999), but their analysis was based on relatively small number of genes. In addition, recent study analyzed genome-wide data on yeast SNPs, finding evidence for selection for TE (Vishnoi et al. 2011). To the best of our knowledge, this is the first comprehensive study that uses genome-wide population genetics data in any multicellular organism to address questions on TE selection on a relatively short time period.

In conclusion, this study underscores the importance of synonymous variants, which are often neglected and considered as silent and nonfunctional. Specifically, we showed two examples where TE may have significant clinical implications in human diseases. With the rapid advancement in sequencing techniques, there is vast increase in whole-genome sequencing data. We hope that this study will not only encourage the usage of these data when studying TE but also further mark the importance of silent SNPs and TE when looking for causal variants in evolution, disease states, and other related studies.

Supplementary Material

Supplementary notes 1–5, tables S1–S8, and figures S1–S2 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Edmond J. Safra Bioinformatics program (Tel Aviv University), Eshkol (Israeli Ministry of Science and Technology) and Dan David fellowships (to Y.Y.W.); Alfred P. Sloan Research Fellowship and NIH grant U01-HG005715 (to A.K.); Regular and Converging Technologies research grants from the Israel Science Foundation and Tauber Fund (to E.R.). T.T. is a Koshland Scholar at Weizmann Institute of Science.

Literature Cited

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136:927–935.
Akashi H. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221–238.
Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660–666.
Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146:295–307.
Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 19:711–722.
Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. Nat Genet. 25:25–29.
Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet. 40:340–345.
Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet. 4:e1000271.
Bossi A, Lehner B. 2009. Tissue specificity and the human protein interaction network. Mol Syst Biol. 5:260.
Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.

Bult CJ, et al. 2008. The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res. 36:D724–D728.
Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.
Carmi S, Levanon EY, Eisenberg E. 2009. Efficiency of complex production in changing environment. BMC Syst Biol. 3:3.
Carmi S, Levanon EY, Havlin S, Eisenberg E. 2006. Connectivity and expression in protein networks: proteins in a complex are uniformly expressed. Phys Rev E. 73:031909.
Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 7:98–108.
Chan PP, Lowe LM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. 37:D93–D97.
Chen K, Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 8:93–103.
Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A. 101:3480–3485.
Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15:1496–1502.
Comeron JM. 2006. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. Proc Natl Acad Sci U S A. 103:6940–6945.
dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32:5036–5044.
Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.
Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715–724.
Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16:287–289.
Durinck S, et al. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439–4340.
Dworniczak B, Aulehla-Scholz C, Horst J. 1990. Phenylalanine hydroxylase gene: silent mutation uncovers evolutionary origin of different alleles. Clin Genet. 38:270–273.
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.
Fredrick K, Ibba M. 2010. How the sequence of a gene can tune its translation. Cell 141:227–229.
Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. Mol Syst Biol. 7:481.
Grossman SR, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886.
Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends Biotechnol. 22:346–353.
Hardelid P, et al. 2008. The birth prevalence of PKU in populations of European, South Asian and sub-Saharan African ancestry living in South East England. Ann Hum Genet. 72:65–71.
Hershberg R, Petrov D. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.
Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet. 10:639–650.
Hunter S. et al. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D224–D228.
Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 151:389–409.
Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol. 158:573–597.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.
International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.
Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143–155.
Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet. 39:1251–1255.
Keinan A, Reich D. 2010. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 6:e1000886.
Khaitovich P, et al. 2006. Positive selection on gene expression in the human brain. Curr Biol. 16:R356–R358.
Kimchi-Sarfaty C, et al. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315:525–528.
King JL, Jukes TH. 1969. Non-Darwinian evolution. Science 16:788–798.
Komar AA. 2009. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 34:16–24.
Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA 13:87–96.
Krawczak M, Zschocke J. 2003. A role for overdominant selection in phenylketonuria? Evidence from molecular data. Hum Mutat. 2:394–397.
Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene. 345:127–138.
Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature 424:147–151.
Liao BY, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 23:2072–2080.
Lohmueller KE, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994–997.
Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 39:415–421.
Marchant D, et al. 2001. Identification of novel VMD2 gene mutations in patients with best vitelliform macular dystrophy. Hum Mutat. 17:235.

Milenkovic VM, Rivera A, Horling F, Weber BH. 2007. Insertion and topology of normal and mutant bestrophin-1 in the endoplasmic reticulum membrane. J Biol Chem. 282:1313–1321.
Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol. 45:514–523.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet. 8:857–868.
Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet. 10:745–755.
Parmley JL, Huynen MA. 2009. Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. PloS Genet. 5:e1000548.
Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol. 268:322–330.
Petrukhin K, et al. 1998. Identification of the gene responsible for Best macular dystrophy. Nat Genet. 19:241–247.
Pickrell JK, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19:826–837.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.
Ruepp A, et al. 2010. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. 38:D497–D501.
Sabeti PC, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.
Scriver CR. 2007. The PAH gene, phenylketonuria, and a paradigm shift. Hum Mutat. 28:831–845.
Scriver CR, et al. 2003. PAHdb 2003: what a locus-specific knowledge-base can do. Hum Mutat. 21:333–344.
Sherry ST, et al. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29:308–311.
Stoletzky N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol. 24:374–381.
Su AI, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101:6062–6067.
Tennessen JA, Madeoy J, Akey JM. 2010. Signatures of positive selection apparent in a small sample of human exomes. Genome Res. 20:1327–1334.
Tuller T, Kupiec M, Ruppin E. 2007. Determinants of protein abundance and translation efficiency in S. cerevisiae. PLoS Comput Biol. 3:e248.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 107:3645–3650.
Tuller T, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res. 13:2260–2264.
Vishnoi A, Sethupathy P, Simola D, Plotkin JB, Hannenhalli S. 2011. Genome-wide survey of natural selection on functional, structural, and network properties of polymorphic sites in Saccharomyces paradoxus. Mol Biol Evol. doi:10.1093/molbev/msr085
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.
Waldman YY, Tuller T, Sharan R, Ruppin E. 2009. TP53 cancerous mutations exhibit selection for translation efficiency. Cancer Res. 69:8807–8813.
Waldman YY, Tuller T, Shlomi T, Sharan R, Ruppin E. 2010. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. Nucleic Acids Res. 38:2964–2974.
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.
Wittkopp PJ. 2005. Genomic sources of regulatory variation in cis and in trans. Cell Mol Life Sci. 62:1779–1783.
Yang Z. 2007. PAML4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zhang F, Saha S, Shabalina SA, Kashina A. 2010. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. Science 329:1534–1537.
Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol. 27:1912–1922.

Associate editor: Eugene Koonin

Genome Biology and Evolution SMBE