Difference in the Proportions of Deleterious Variations Within and Between Populations Influences the Estimation of FST

Sankar Subramanian (ssankara@usc.edu.au)
University of the Sunshine Coast

Research Article

Keywords: Population differentiation, FST, deleterious mutations, temporal distributions and population genetics theory

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-153346/v1

Version of Record: A version of this preprint was published at Genes on January 22nd, 2022. See the published version at https://doi.org/10.3390/genes13020194.

Difference in the proportions of deleterious variations within and between populations influences the estimation of F_{ST}

Sankar Subramanian

GeneCology Research Centre, The University of the Sunshine Coast, 90 Sippy Downs Drive, Sippy Downs Qld 4556, Australia

Running head: Deleterious variations influence the estimation of F_{ST}

Keywords: Population differentiation, FST, deleterious mutations, temporal distributions and population genetics theory

Address for correspondence:

GeneCology Research Centre
University of the Sunshine Coast
90 Sippy Downs Drive
Sippy Downs QLD 4556
Australia
Phone: + 61-7-5430 2873
Fax: +61-7- 5430 2881
E-mail: ssankara@usc.edu.au

Abstract

Estimating the extent of genetic differentiation between populations is an important measure in population genetics, ecology and evolutionary biology. Fixation index or F_{ST} is an important measure, which is routinely used to quantify this. Previous studies have shown that F_{ST} estimated for selectively constrained regions was significantly lower than that estimated for neutral regions. By deriving the theoretical relationship between F_{ST} at neutral and constrained sites we show that an excess in the fraction of deleterious variations segregating within populations compared to that segregates between populations is the cause for the reduction in F_{ST} estimated at constrained sites. Using whole genome data, our results revealed that the magnitude of reduction in F_{ST} estimates obtained for selectively constrained regions was much higher for distantly related populations compared to those estimated for closely related pairs. For example, the reduction was 47% for comparison between Europeans-Africans, 30% for European-Asian comparison, 16% for the Northern-Southern European pair and only 4% for the comparison involving two Southern European (Italian and Spanish) populations. Since deleterious variants are purged over time due to purifying selection, their contribution to the among-population diversity at constrained sites decreases with the increase in the divergence between populations. However, within-population diversity remains the same for all pairs compared and therefore F_{ST} estimated at constrained sites for distantly related populations are much smaller than those estimated for closely related populations. Our results suggest that the level of population divergence should be considered when comparing constrained site F_{ST} estimates from different pairs of populations.

Keywords: Population differentiation, FST, deleterious mutations, temporal distributions and population genetics theory

Significance statement

FST is a fundamental parameter that is routinely used in population genetics to measure the level of population differentiation. Previous studies reported a much smaller FST estimates at selectively constrained regions compared to those at neutral regions. However, what is exactly causing the reduction in FST for constrained sites or SNPs is not known. Here we show that an excess fraction of deleterious variations is segregating within population compared to that segregates between populations and this is the cause for the reduction in F_{ST} estimates at constrained regions. We also show that the excess fraction is also depended upon the level of divergence between populations. A much higher fraction was observed for comparisons involving distantly related populations than those involving closely related populations.

Introduction

Since the introduction of F-statistics by Sewall Wright [1], fixation index or F_{ST} has been routinely used to measure the extent of differentiation between populations [2-12]. F_{ST} compares the heterozygosities within and between (or total) populations to measure the level of genetic structure among populations. Apart from being an integral part of the descriptive statistics to describe a population, F_{ST} has direct applications in conservation biology, ecology, evolutionary biology and clinical genetics. F_{ST} reveals the extent of genetic drift and the level of migrations between populations, which is useful to understand the population dynamics of an ecosystem [13]. The level of differentiation in populations helps conservation biologists to measure risk of extinction of a population or species [14]. Furthermore, F_{ST} is used to identify candidate genetic variants and genes associated with Mendelian and complex genetic diseases [2,3,9].

In evolutionary biology F_{ST} is used to detect the signature of positive selection [3,4,6,7,10-12,15]. Since adaptive mutations quickly spread through a population, their relative high frequency in one population (compared to the other) elevates the F_{ST} estimates. In whole genome analyses, F_{ST} is estimated using a sliding window to detect genomic regions showing high differentiation between populations [10]. Although a number of previous studies have investigated the relationship between F_{ST} and positive selection only a handful of studies examined the influence of negative selection on F_{ST}. A previous study reported lower F_{ST} for genic compared to nongenic SNPs [3]. The reduction in F_{ST} was more pronounced when only the amino acid changing nonsynonymous SNPs (nSNPs) were considered and a similar reduction was observed for mutations in disease-related genes. This suggests that purifying selection does not allow an increase in the frequency of SNPs, which could have led to the observed low F_{ST} [16]. Later a more systematic investigation was conducted to examine this issue using human

genome data [17]. This study grouped nSNPs based on the evolutionary rates of sites in which they were present and showed a positive correlation between the rates and F_{ST}. Hence F_{ST} estimated for the nSNPs present in selectively constrained sites (with low rate of evolution) was much smaller than that estimated for those present in neutral sites with high evolutionary rates. A similar observation was made by another study on the populations of fruit flies from France and Rwanda [18]. F_{ST} estimates obtained for long introns (known to be under high purifying selection) and conserved genes were typically lower than those estimated for short introns (under relaxed selective constraints) and less conserved genes.

Although the influence of purifying selection on F_{ST} estimates has been well documented, how exactly selective constraint affects F_{ST} estimations or the mechanism by which purifying selection influences these estimates is unclear. Furthermore, whether the magnitude of reduction in F_{ST} is depended on the divergence between populations or whether the magnitude of reduction is similar between closely related and distantly related populations is also unknown. To examine these, we first investigated the theoretical relationship between F_{ST} at neutral and constrained sites. Using the data from the 1000 genome project - Phase 3 [19] we then estimated F_{ST} for pairs of populations with different levels of divergence.

Materials and Methods

Theoretical relationship between F_{ST} and Neutrality Index (N_I)

F_{ST} at synonymous sites ($F_{ST(S)}$) can be expressed using Hudson et.al [20] as:

$$F_{ST(S)} = 1 - \frac{H_{ws}}{H_{bs}} \qquad (1)$$

where H_b and H_w are mean number of synonymous nucleotide differences between and within populations respectively. Similarly, F_{ST} at nonsynonymous sites ($F_{ST(N)}$) is given as:

$$F_{ST(N)} = 1 - \frac{H_{wn}}{H_{bn}} \qquad (2)$$

Multiplying $\left(\frac{H_{ws}}{H_{ws}}\right)$ and $\left(\frac{H_{bs}}{H_{bs}}\right)$ with the numerator and denominator respectively:

$$F_{ST(N)} = 1 - \frac{H_{wn}\left(\frac{H_{ws}}{H_{ws}}\right)}{H_{bn}\left(\frac{H_{bs}}{H_{bs}}\right)}$$

Rearranging the above gives

$$F_{ST(N)} = 1 - \frac{\left(\frac{H_{wn}}{H_{ws}}\right)H_{ws}}{\left(\frac{H_{bn}}{H_{bs}}\right)H_{bs}} \qquad (3)$$

Substituting $\omega_b = \frac{H_{bn}}{H_{bs}}$ and $\omega_w = \frac{H_{wn}}{H_{ws}}$

$$F_{ST(N)} = 1 - \frac{\omega_w H_{ws}}{\omega_b H_{bs}} \qquad (4)$$

The ratio of the mean number of nonsynonymous to synonymous differences $\left(\frac{H_{bn}}{H_{bs}}\right)$ between populations denotes the proportion of nonsynonymous polymorphisms ($_b$) segregating between populations and the ratio of nonsynonymous to synonymous differences $\left(\frac{H_{wn}}{H_{ws}}\right)$ denotes the proportion of nonsynonymous polymorphisms ($_w$) present within populations. For comparisons involving two populations $\omega_w = (\omega_1 + \omega_2)/2$ where $_1$ and $_2$ are proportions of nonsynonymous polymorphisms in populations 1 and 2 respectively.

Previous studies introduced a measure Neutrality Index (N_I), which is the ratio of the ratio of nonsynonymous to synonymous polymorphisms to the ratio of nonsynonymous to synonymous substitutions$\left[\left(\frac{P_n}{P_s}\right)/\left(\frac{D_n}{D_s}\right)\right]$ [21]. N_I was used to compare ***within species*** polymorphisms at neutral

and constrained sites with the substitutions ***between species***. In this study we compare mean number of polymorphisms ***within population*** at neutral and constrained sites with those of ***between populations*** $\left[\left(\frac{H_{wn}}{H_{ws}}\right)/\left(\frac{H_{bn}}{H_{bs}}\right)\right]$ *or* $\frac{\omega_w}{\omega_b}$, which is qualitatively similar to N_I. Therefore

$$F_{ST(N)} = 1 - N_I \frac{H_{ws}}{H_{bs}} \qquad (5)$$

$$F_{ST(N)} = 1 - \left[N_I\left(1 - F_{ST(S)}\right)\right] \qquad (6)$$

If $_b$ and $_w$ fractions of nonsynonymous mutations are equal (that result in $N_I = 1$), then equation 6 reduces to

$$F_{ST(N)} = F_{ST(S)} \qquad if\ \ \omega_w = \omega_b\ or\ N_I = 1 \quad (7)$$

This suggests that F_{ST} at synonymous sites is equal to that at nonsynonymous sites. However, it is well known that the fraction of slightly deleterious mutations present within population is higher than that observed between populations. This is because a much higher fraction of those present within population are young and yet to be purged from the population by natural selection. Therefore, $_w$ is expected to be higher than $_b$. Therefore,

$$F_{ST(N)} < F_{ST(S)} \qquad if\ \ \omega_w > \omega_b\ or\ N_I > 1 \quad (8)$$

The above theoretical relationships demonstrate that F_{ST} at nonsynonymous sites is expected to be smaller than that observed for synonymous sites if there is an excess in the proportion of nonsynonymous deleterious mutations present within populations compared to that of between populations.

Population genome data

Whole genome data for 26 world-wide populations was downloaded from the 1000 genome project – Phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) [19]. Only biallelic single nucleotide polymorphisms (SNPs) from the autosomes were included for the

analyses and the allele frequencies of each SNP in 26 populations were computed, which were used for estimating F_{ST} using the estimators described below. Pairwise FSTs were computed for the exomes of human populations including CHB (Northern Chinese), CHS (Southern Chinese), GBR (British), IBS (Spanish), JPT (Japanese), TSI (Italian) and YRI (Nigerian). To determine the magnitude of selective constraints at each site of the human genome we used a robust method, Combined Annotation-Dependent Depletion (CADD) that integrates many diverse annotations into a single measure (*C* score). [22]. The precomputed C scores for each genome position are available at: http://cadd.gs.washington.edu/download/1000G_phase3_inclAnno.tsv.gz and these scores were mapped to the genotype data from the 1000 genome project. To identify derived alleles, orientations of SNVs were determined using the ancestral state of the nucleotides, which was inferred from six primate EPO alignments [19].

F_{ST} estimation

For estimating F_{ST} from human exome data we used the method developed by Hudson et.al [20] and used the following estimator [23]:

$$\tilde{F}_{ST}^{Hudson} = \frac{H_B - H_S}{H_B} \quad (9)$$

and

$$H_S = \tilde{p}_1(1-\tilde{p}_1) + \tilde{p}_2(1-\tilde{p}_2)$$

$$H_B = \tilde{p}_1(1-\tilde{p}_2) + \tilde{p}_2(1-\tilde{p}_1)$$

where p_1, p_2 are frequencies of the bialleles. To combine F_{ST} estimated for different SNPs of the genome we used the ratio of averages approach suggested by Bhatia et.al. [23]. To estimate the variance, we used a bootstrap resampling procedure with 1000 replicates.

The magnitude of reduction in the F_{ST} of nonsynonymous sites was quantified as:

$$\rho = 1 - \frac{F_{ST(N)}}{F_{ST(S)}} \qquad (10)$$

Results

The effect of purifying selection on F_{ST}

To examine the influence of purifying selection on F_{ST} we used European and African exome data from the 1000 genome project - Phase 3 (see methods). In order to examine the magnitude of selection pressure the Combined Annotation Dependent Depletion (CADD) score or *C*-score was used [22]. Nonsynonymous SNPs were grouped into seven categories based on their *C*-scores. Figure 1A shows the relationship between selection pressure and F_{ST} estimated for synonymous (sSNPs) and nonsynonymous SNPs (nSNPs) using the exome data for the Italian (TSI) – Nigerian (YRI) pair. Clearly F_{ST} is the highest for the neutral sSNPs, which declines with increase in selection magnitude. F_{ST} estimate for highly constrained nSNPs with a *C*-score >30 was only 0.082, which is much smaller than that estimated for sSNPs (0.154). We introduced a measure, □ to capture the magnitude of reduction in F_{ST} estimates of nSNPs compared to that of sSNPs (Equation 10 - see methods). Figure 1B that shows the positive relationship between the extent of selection constraint and magnitude of reduction of F_{ST} (□). The reduction in the F_{ST} estimate was only 2% for nSNPs under relaxed selection pressure (*C*-Score <=5), which increases with the magnitude of selection pressure. For highly constrained nSNPs (*C*-score >30), the reduction in F_{ST} was 47%, which is 24 times higher than that observed for nSNPs under relaxed constraint.

Relationship between F_{ST} at neutral and constrained genomic regions

To understand the actual cause of the reduction in F_{ST} for constrained SNPs we derived the theoretical relationship between F_{ST} at neutral ($F_{ST(S)}$) and constrained ($F_{ST(N)}$) regions, which is shown in equation 6. From this it is clear that this relationship in modulated by Neutrality

Index (NI), which is the ratio of ratio of nonsynonymous-to-synonymous polymorphisms observed within populations ($_w$) to that segregates between ($_b$) populations ($N_I =$ $_w/$ $_b$). As shown in equation 7, $F_{ST(S)}$ and $F_{ST(N)}$ are equal when the proportions of nonsynonymous polymorphisms observed within and between populations are the same ($_w =$ $_b$). However, it is well known that a higher proportion of slightly deleterious nSNPs is expected to segregate within populations than that of between populations. This is because a significant fraction of them are purged by purifying selection over time and hence their fraction gets diminished for between population comparisons. Therefore, F_{ST} estimated for nSNPs is expected to be smaller than that of sSNPs as the fraction of nSNPs segregating within populations is higher than that segregates between populations ($_w >$ $_b$) (Equation 8).

Plotting the relationship of equation 6 shows that the difference between $F_{ST(S)}$ and $F_{ST(N)}$ is much higher when $F_{ST(S)}$ is small (Figure 2A). For instance, when $F_{ST(S)}$ was 0.02 the corresponding $F_{ST(N)}$ was 0.0004 (for N_I=1.02). In contrast, when $F_{ST(S)}$ when was 0.1 the corresponding $F_{ST(N)}$ was 0.082, which is only slight smaller than the former. The magnitude of reduction of $F_{ST(N)}$ compared to $F_{ST(S)}$ is very clear in figure 2B. For N_I=1.02, the magnitude of reduction of $F_{ST(N)}$ was 98% when $F_{ST(S)}$ = 0.02. Whereas this reduction was only 18% for $F_{ST(S)}$ = 0.1. Figure 2B also shows that the magnitude of reduction is also increases with increasing N_I. For example, when $F_{ST(S)}$ = 0.2 and N_I=1.02 the magnitude of reduction was only 8% but it was 48% for N_I=1.12.

F_{ST} estimates and population divergence

Next, we investigated the effects of purifying selection on F_{ST} with respect to population divergence. This is to compare the magnitude of reduction of F_{ST} estimated for closely and distantly related populations. For this purpose, we used four pairs of comparisons with

different levels of divergence, European (Italian/TSI) – African (Nigerian/YRI), European (Italian/TSI) – Asian (Chinese/CHB), Southern European (Italian/TSI) – Northern European (British/GBR) and two Southern Europeans (Italian/TSI – Spanish/IBS). Figure 3 shows F_{ST} estimates obtained for sSNPs and highly constrained nSNPs (C score > 30) for the four pairs of populations. By comparing of the pairs of columns from Figure 3A to 3D reveals a reduction in the difference between the F_{ST} estimated for synonymous ($F_{ST(S)}$) and highly constrained ($F_{ST(N)}$) nonsynonymous sites. The positive correlation between the population divergence and the extent of reduction of $F_{ST(N)}$ is clear in Figure 4. The $F_{ST(N)}$ observed for constrained nSNPs of the distantly related Italian-Nigerian pair was 47% smaller than that of sSNPs (Figure 4A). While this reduction was 30% for the Italian-Chinese pair and 16% for Italian-British comparison, it was only 4% for the closely related Italian-Spanish pair (Figure 4). We then estimated N_I for the four pairs, which were found to be 1.0861, 1.0400, 1.0015 and 1.0003 for the Italian-Nigerian, Italian-Chinese, Italian-British and Italian-Spanish pairs respectively. This translates to an excess 8.6%, 4.0%, 0.15% and 0.03% of nSNPs were present in within populations compared that segregating between the populations of the Italian-Nigerian, Italian-Chinese, Italian-British and Italian-Spanish pairs respectively. The presence of these excess fractions resulted in 47.1%, 30.0%, 15.9% and 4.2% reduction in the F_{ST} estimated for the highly constrained nSNPs (*C*-score > 30) of the corresponding pairs of populations respectively. A similar analysis was performed for the Chinese lineage using Chinese-Nigerian, Chinese-British, Chinese-Japanese and Northern-Southern Chinese population pairs. This analysis showed an excess 10.3%, 4.0%, 0.1% and 0,03% of nSNPs were present in within population comparisons compared to those in between these pairs of populations. These excess proportions resulted in 46.6%, 29.9%, 8.6% and 4.1% reduction in $F_{ST(N)}$ compared to $F_{ST(S)}$.

Discussion

Although previous studies have observed a reduction in F_{ST} estimates of selectively constrained sites [3,17,18]. Particularly Maruki et al. (2012) showed that genes and genomic regions under purifying selection have low F_{ST} values. Our results confirmed their finding as shown in Figure 1. However, the focus of this study is to find the true cause for that reduction in the F_{ST} of constrained sites. Using the theoretical relationship between F_{ST} at neutral and constrained sites we showed that an excess fraction of nSNPs segregating within population compared to that between populations is the reason for the reduction in F_{ST} at constrained sites. The reason for the excess fraction of nSNPs present in within populations is due to that fact that a high proportion of deleterious mutations segregating within populations are relatively young and hence were not removed by natural selection. Therefore, they contribute significantly to the constrained site heterozygosity within populations. In contrast, a much higher proportion of the harmful mutations have been purged due to the time elapsed and hence their contribution to the constrained site heterozygosity between populations is relatively less. Hence within population heterozygosity at constrained sites is much more inflated than that observed for inter-population comparison. This results in the reduction of F_{ST} estimates, as it is based on the normalized difference between the inter- and intra-population diversities.

The results of this study highlight two important patterns and provide theoretical and empirical explanations for them. First the reduction in the F_{ST} estimates positively correlates with the magnitude of selection suggesting a much higher underestimation for nSNPs at highly constrained regions of the genome. This is because high magnitude of purifying selection leads to segregation of more slightly deleterious mutations within populations (as more genomic sites are under selection) and hence the fraction of deleterious nSNPs segregating within populations will be much higher than that segregates between populations ($_w >> _b$ or $N_I >> 1$). Hence this leads to a much higher underestimation of F_{ST} of nSNPs at highly constrained regions

compared to that of sSNPs ($F_{ST(N)} << F_{ST(S)}$). In contrast, there are fewer deleterious nSNPs in less constrained regions and hence the fraction of harmful polymorphisms segregating within populations is expected to be only modestly higher than that segregates between populations ($_w >$ $_b$ or $N_I > 1$). This results in a much smaller reduction in the F_{ST} estimated for nSNPs present in regions under relaxed selective constraints ($F_{ST(N)} < F_{ST(S)}$).

Second, we have shown that the magnitude of reduction of F_{ST} at constrained sites for comparisons involving distantly related populations was much higher than that of those involving closely related pairs. For instance, this reduction for the Nigerian-Italian comparison (47%) is more than ten-fold higher than that of the Southern European (Spanish-Italian) pair (4.2%). It is well known that deleterious variants are removed over time and hence the only a small fraction ($_b << 1$) of them segregate and contribute to constrained site inter-population diversity for distantly related populations. However, a relatively modest fraction ($_b < 1$) of harmful nSNPs contribute to the inter-population diversity for closely related population as the elapsed time not enough to purge most of them. On the other hand, the fraction of deleterious nSNPs within population ($_w$) remain the same (eg. within Italians) for comparisons involving both distantly (eg. Nigerian-Italian) as well as closely (eg. Spanish-Italian) related populations. Hence the magnitude of reduction in constrained site F_{ST} (with respect to neutral site F_{ST}) for distantly related populations is much higher ($F_{ST(N)} << F_{ST(S)}$) than that observed for closely related populations ($F_{ST(N)} < F_{ST(S)}$).

The findings of this study suggest that the F_{ST} estimated for different genes or genomic regions of a genome are not comparable if the level of selective constrains are different between them. This is particularly important while using F_{ST} estimates to detect positive selection because such methods assume neutral evolution in genes and genomic regions and hence do not account

for purifying selection in the estimations [4,6,10-12,15]. Our results also strongly indicate that F_{ST} obtained from the constrained regions of different pairs of populations are not comparable if the population divergence times between the pairs are not the same. In such cases F_{ST} estimations should include only neutral sites to obtain unbiased estimates. However, this is only possible for large genomes such as vertebrate in which constrained regions constitute only a small fraction (~10%) of the genome [24,25]. This is an important issue for small genomes such as those of fruit flies with >50% of the genome is under selection [26]. Therefore, population divergence time need to be considered when comparing the genome-wide F_{ST} estimates from different populations.

Acknowledgments

The author acknowledges the support from the Australian Research Council (DP18010089) and the University of the Sunshine Coast.

Data Availability Statement

The whole genome data used in this study is available at:

https://ftp.ncbi.nlm.nih.gov/1000genomes/

References

1 Wright, S. The genetical structure of populations. *Ann Eugen* **15**, 323-354 (1951).

2 Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**, 1805-1814, doi:10.1101/gr.631202 (2002).

3 Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**, 340-345, doi:10.1038/ng.78 (2008).

4 Beaumont, M. A. & Balding, D. J. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**, 969-980 (2004).

5 Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**, 1111-1120, doi:10.1086/421051 (2004).

6 Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res* **20**, 393-402, doi:10.1101/gr.100545.109 (2010).

7 Excoffier, L., Hofer, T. & Foll, M. Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* **103**, 285-298, doi:10.1038/hdy.2009.74 (2009).

8 Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet* **41**, 66-70, doi:10.1038/ng.303 (2009).

9 Sams, A. & Hawks, J. Patterns of population differentiation and natural selection on the celiac disease background risk network. *PLoS One* **8**, e70564, doi:10.1371/journal.pone.0070564 (2013).

10 Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu Rev Genet* **47**, 97-120, doi:10.1146/annurev-genet-111212-133526 (2013).

11 Wu, D. D. & Zhang, Y. P. Positive selection drives population differentiation in the skeletal genes in modern humans. *Hum Mol Genet* **19**, 2341-2346, doi:10.1093/hmg/ddq107 (2010).

12 Xue, Y. *et al.* Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. *Genetics* **183**, 1065-1077, doi:10.1534/genetics.109.107722 (2009).

13 Whitlock, M. C. & McCauley, D. E. Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). *Heredity (Edinb)* **82 (Pt 2)**, 117-125, doi:10.1038/sj.hdy.6884960 (1999).

14 Frankham, R., Ballou, J. D. & Briscoe, D. A. *Introduction to conservation genetics*. (Cambridge University Press, 2002).

15 Bonhomme, M. *et al.* Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* **186**, 241-262, doi:10.1534/genetics.104.117275 (2010).

16 Nielsen, R. Molecular signatures of natural selection. *Annu Rev Genet* **39**, 197-218, doi:10.1146/annurev.genet.39.073003.112420 (2005).

17 Maruki, T., Kumar, S. & Kim, Y. Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. *Mol Biol Evol* **29**, 3617-3623, doi:10.1093/molbev/mss187 (2012).

18 Jackson, B. C., Campos, J. L. & Zeng, K. The effects of purifying selection on patterns of genetic differentiation between Drosophila melanogaster populations. *Heredity (Edinb)* **114**, 163-174, doi:10.1038/hdy.2014.80 (2015).

19 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

20 Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-589 (1992).

21 Rand, D. M. & Kann, L. M. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* **13**, 735-748, doi:10.1093/oxfordjournals.molbev.a025634 (1996).

22 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

23 Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res* **23**, 1514-1521, doi:10.1101/gr.154831.113 (2013).

24 Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res* **21**, 1769-1776, doi:10.1101/gr.116814.110 (2011).

25 Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* **10**, e1004525, doi:10.1371/journal.pgen.1004525 (2014).

26 Andolfatto, P. Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437**, 1149-1152, doi:10.1038/nature04107 (2005).

Figure Legends

Figure 1. **(A)** Relationship between selection intensity and F_{ST}. Whole exome data comprising synonymous SNPs (sSNPs) and nonsynonymous SNPs (nSNPs) for the Italian (TSI)-Nigerian (YRI) population pair was used to estimate F_{ST}. The magnitude of selection intensity on nSNPs is measured by the Combined Annotation-Dependent Depletion (CADD) method that integrates many diverse annotations into a single measure (*C* score) [22]. A bootstrap resampling procedure (1000 replicates) was used to estimate the standard error. **(B)** Magnitude of reduction of F_{ST} estimates and selection intensity. X-axis shows the reduction in F_{ST} estimates of nSNPs in comparison with that of sSNPs (□) using equation 10 (see methods) for the exome data described above. Error bars show standard error of the mean.

Figure 2. **(A)** Theoretical relationship between neutral $F_{ST(S)}$ and constrained $F_{ST(N)}$ sites shown in equation 6 (see methods). **(B)** Relationship between neutral $F_{ST(S)}$ and the magnitude of reduction ( ) in constrained site $F_{ST(N)}$ (Equation 10). The above relationships were shown for different N_I values.

Figure 3. F_{ST} estimates for synonymous and highly constrained nonsynonymous SNPs of the **(A)** Italian-Nigerian **(B)** Italian-Chinese **(C)** Italian-British and **(D)** Italian-Spanish population pairs. Error bars are the standard error of the mean and a bootstrap resampling procedure (1000 replicates) was used to estimate the variance. The difference between the FST estimates of synonymous and constrained sites are highly significant ($P < 0.01$, *Z* test).

Figure 4. The magnitude of reduction in F_{ST} estimates of nSNPs obtained for four population pairs. The population tree on top is drawn to highlight the correlation between the population divergence and the magnitude of reduction in F_{ST}. **(A)** Italian-Nigerian, Italian-Chinese,

Italian-British and Italian-Spanish population pairs. **(B)** Chinese-Nigerian, Chinese-British, Chinese-Japanese and Northern Chinese (Beijing)-Southern Chinese (Shanghai) population pairs. Error bars are the standard error of the mean and a bootstrap resampling procedure (1000 replicates) was used to estimate the variance.

Figures

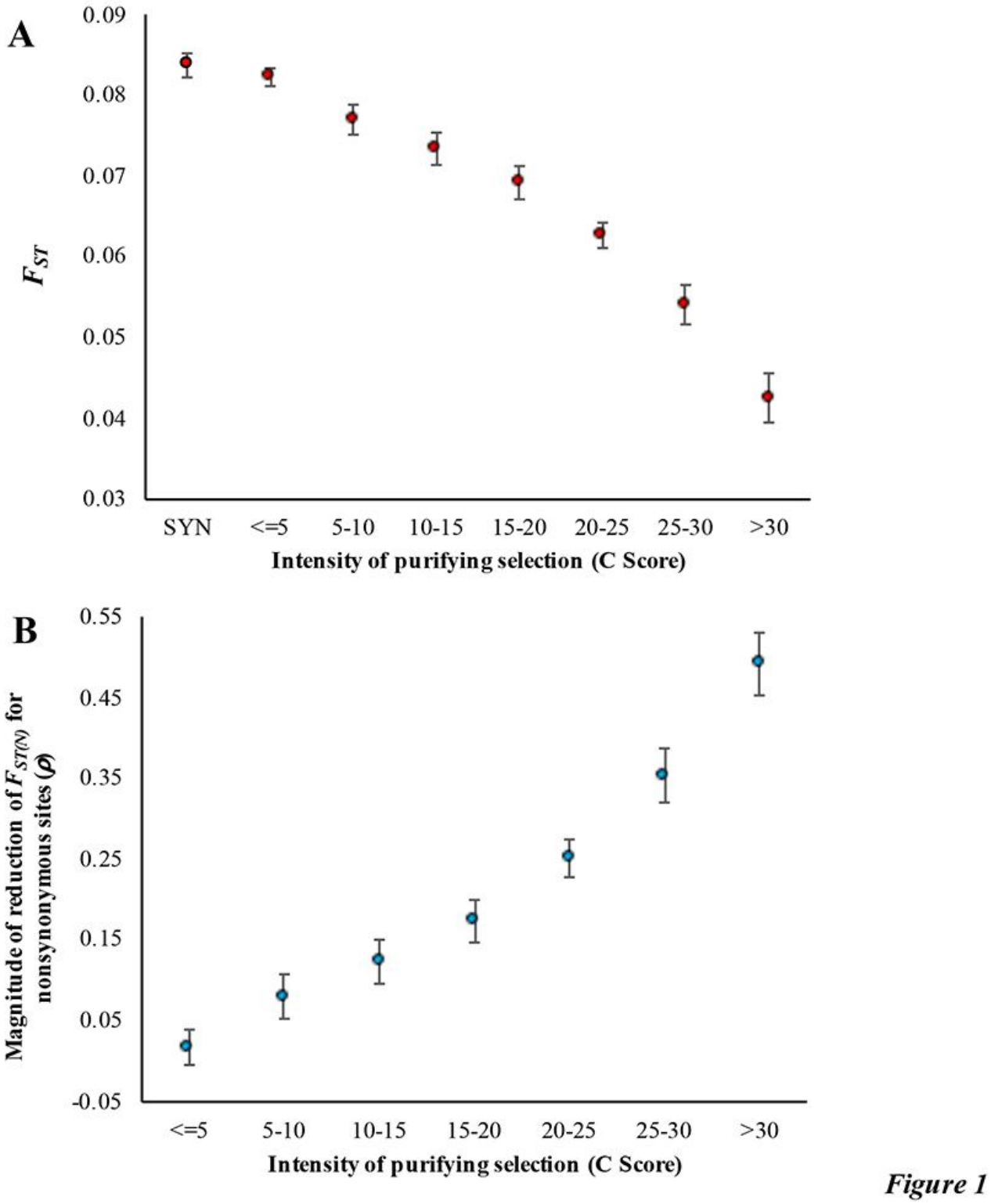

Figure 1

(A) Relationship between selection intensity and FST. Whole exome data comprising synonymous SNPs (sSNPs) and nonsynonymous SNPs (nSNPs) for the Italian (TSI)-Nigerian (YRI) population pair was used to estimate FST. The magnitude of selection intensity on nSNPs is measured by the Combined

Annotation-Dependent Depletion (CADD) method that integrates many diverse annotations into a single measure (C score) 22. A bootstrap resampling procedure (1000 replicates) was used to estimate the standard error. (B) Magnitude of reduction of FST estimates and selection intensity. X-axis shows the reduction in FST estimates of nSNPs in comparison with that of sSNPs (□) using equation 10 (see methods) for the exome data described above. Error bars show standard error of the mean.

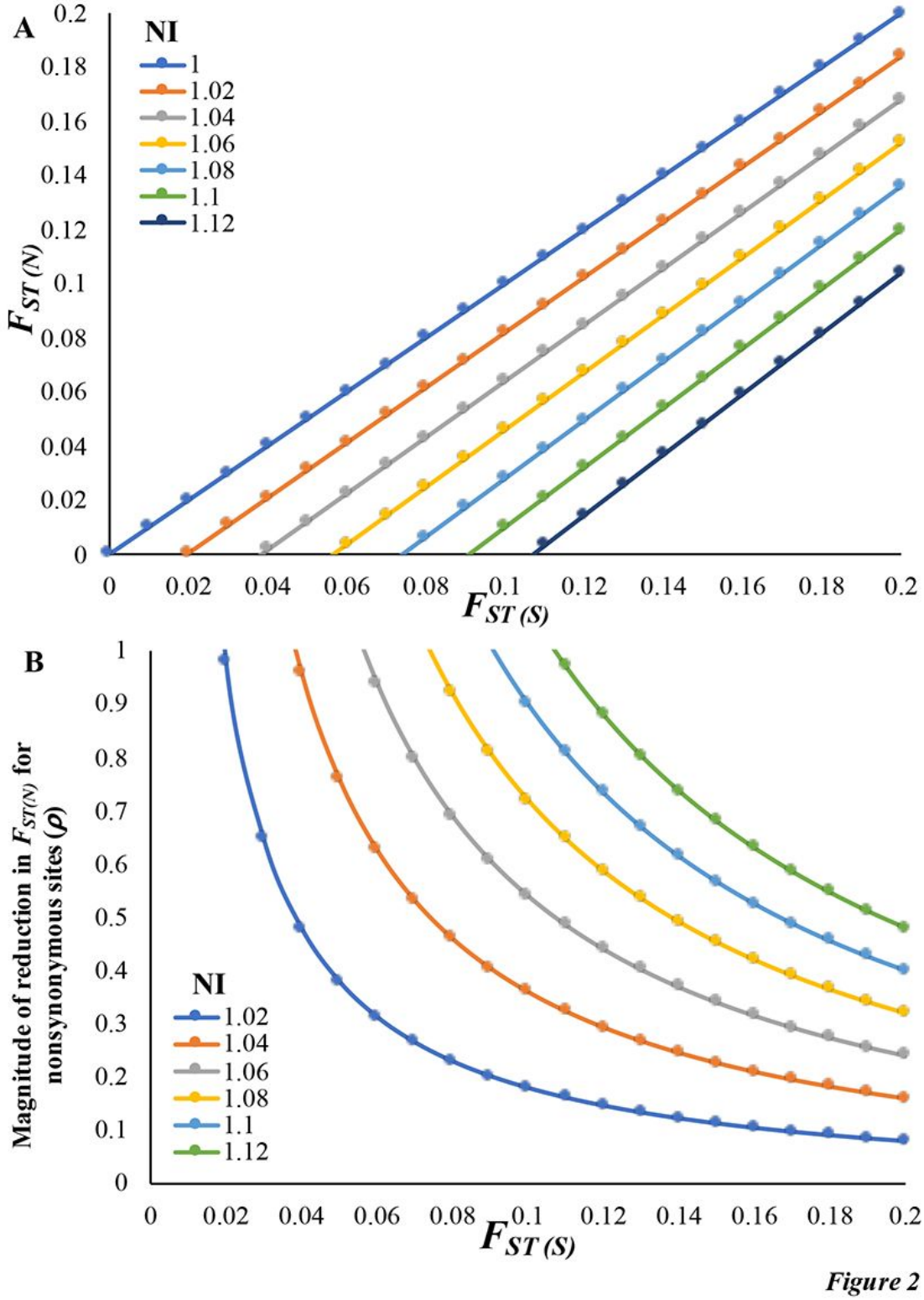

Figure 2

Figure 2

(A) Theoretical relationship between neutral FST(S) and constrained FST(N) sites shown in equation 6 (see methods). (B) Relationship between neutral FST(S) and the magnitude of reduction (□) in constrained site FST(N) (Equation 10). The above relationships were shown for different NI values.

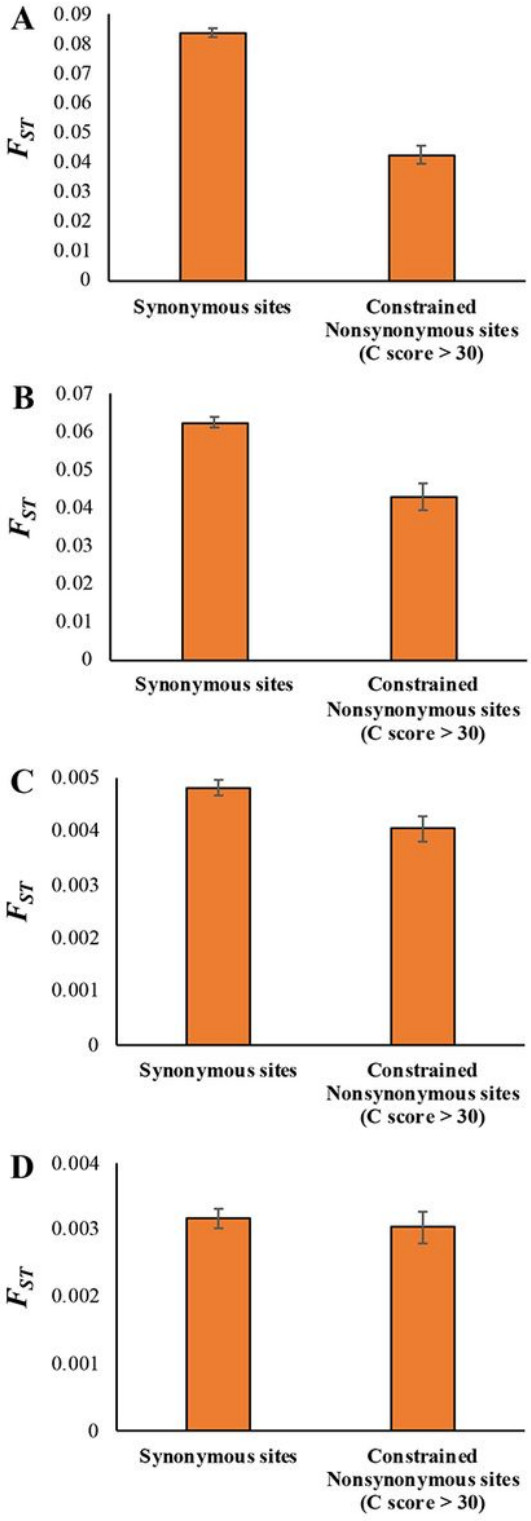

Figure 3

Figure 3

FST estimates for synonymous and highly constrained nonsynonymous SNPs of the (A) Italian-Nigerian (B) Italian-Chinese (C) Italian-British and (D) Italian-Spanish population pairs. Error bars are the standard

error of the mean and a bootstrap resampling procedure (1000 replicates) was used to estimate the variance. The difference between the FST estimates of synonymous and constrained sites are highly significant (P < 0.01, Z test).

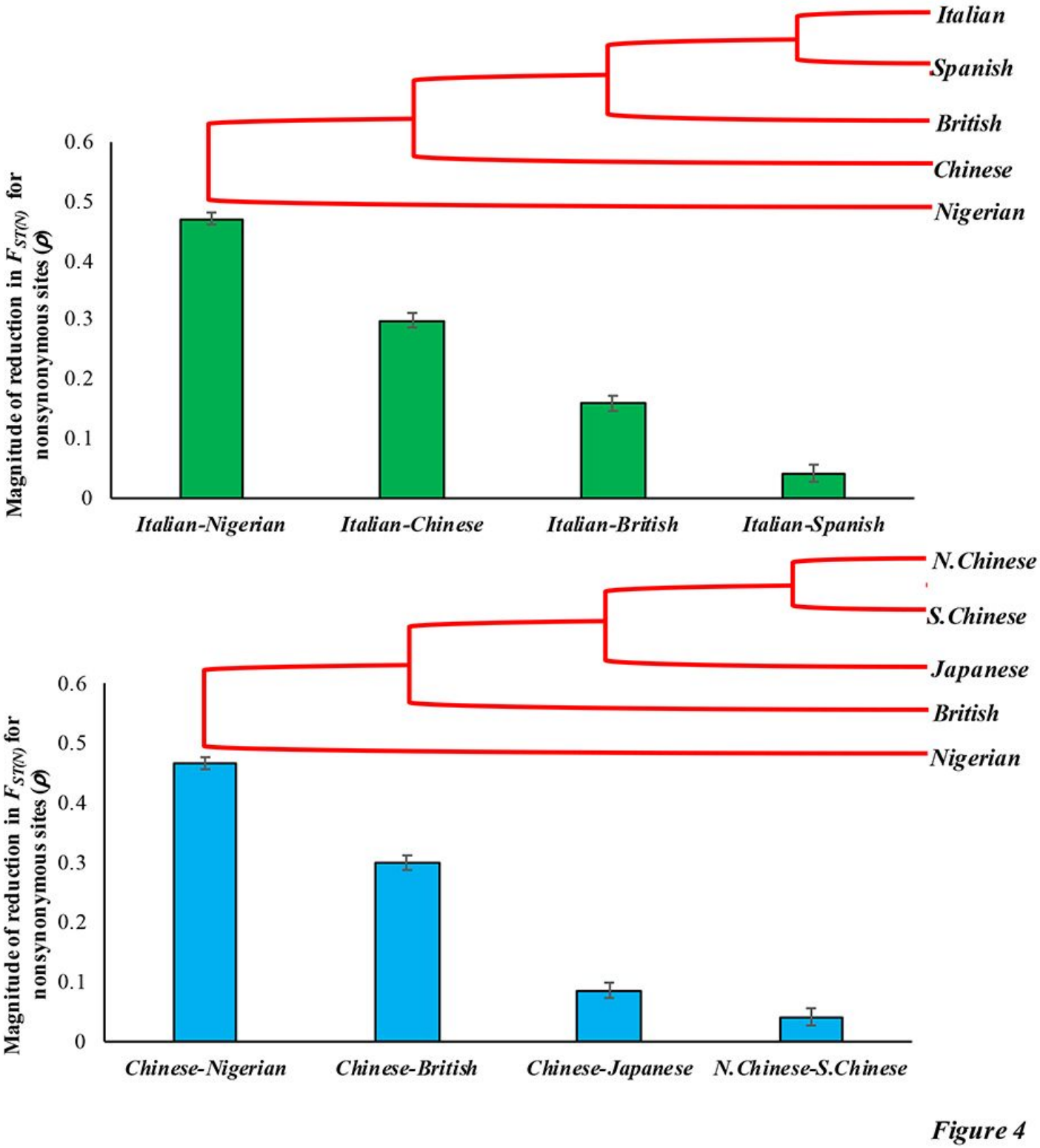

Figure 4

The magnitude of reduction in FST estimates of nSNPs obtained for four population pairs. The population tree on top is drawn to highlight the correlation between the population divergence and the

magnitude of reduction in FST. (A) Italian-Nigerian, Italian-Chinese, Italian-British and Italian-Spanish population pairs. (B) Chinese-Nigerian, Chinese-British, Chinese-Japanese and Northern Chinese (Beijing)-Southern Chinese (Shanghai) population pairs. Error bars are the standard error of the mean and a bootstrap resampling procedure (1000 replicates) was used to estimate the variance.