Research article

Population size determines the type of nucleotide variations in humans

Sankar Subramanian

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

Running head: Patterns of nucleotide changes in human populations

Keywords: Type of nucleotide change, biased gene conversion, recombination, common and rare variants, deleterious variants, population bottleneck and human genomes

Address for correspondence:
Dr. Sankar Subramanian
GeneCology Research Centre
University of the Sunshine Coast
90 Sippy Downs Drive
Sippy Downs QLD 4556
Australia
Phone: +61-7-3735 7495
Fax: +61-7-3735 7459
E-mail: ssankara@usc.edu.au

Abstract

It is well known that the effective size of a population (N_e) is one of the major determinants of the amount of genetic variation within the population. Here, we examined whether the types of genetic variations are dictated by the effective population size. Our results revealed that for low frequency variants, the ratio of AT→GC to GC→AT variants (β) was similar across populations, suggesting the similarity of the pattern of mutation in various populations. However, for high frequency variants, β showed a positive correlation with the effective population size of the populations. This suggests a much higher proportion of high frequency AT→GC variants in large populations (e.g. Africans) compared to those with small population sizes (e.g. Asians). These results imply that the substitution patterns vary significantly between populations. These findings could be explained by the effect of GC-biased gene conversion (gBGC), which favors the fixation of G/C over A/T variants in populations. In large population, gBGC causes high β. However, in small populations, genetic drift reduces the effect of gBGC resulting in reduced β. This was further confirmed by a positive relationship between N_e and β for homozygous variants. Our results highlight the huge variation in the types of homozygous and high frequency polymorphisms between world populations. We observed the same pattern for deleterious variants, implying that the homozygous polymorphisms associated with recessive genetic diseases will be more enriched with G or C in populations with large N_e (e.g. Africans) than in populations with small N_e (e.g. Europeans).

Introduction

The out of Africa hypothesis predicts that the ancestors of the human populations around the world originated in Africa, migrated out of the continent and eventually colonized different parts of the world (Stringer 2003). During this process, the ancestral populations underwent a series of population bottlenecks along the migratory routes. Due to this founder effect, the ancestral population size is expected to decline with increasing distance from Africa. Previous empirical studies confirmed this prediction and showed that populations in Africa are the most genetically diverse and that the diversity declined with increasing geographic distance from Africa particularly along the colonization routes (Prugnolle, et al. 2005; Ramachandran, et al. 2005; Handley, et al. 2007; Li, et al. 2008; DeGiorgio, et al. 2009). These observations clearly suggest significant variation in the nucleotide diversity among global populations.

Nucleotide diversity ($\pi=4N_e\mu$) is a measure of genetic variation, which is determined by mutation rate (μ) and effective population size (N_e). Since mutation rate is similar across human populations, the observed difference in the diversity of world populations is largely due to the variations in effective population sizes. Although a recent study suggested a higher rate of mutation in non-Africans, the magnitude of this effect was very small (~5%) (Mallick, et al. 2016). Recent population genomic studies showed large variation in the number of polymorphisms observed between world populations (Genomes Project, et al. 2015). Populations in Africa have ~5 million Single Nucleotide Variations (SNVs) whereas those in East Asia have ~4.1 million, which is ~20% less. Although the variation in the number of polymorphisms is well known, it is unclear if there are differences in the types of polymorphisms between world populations. Are the frequencies of different types of

nucleotide changes (eg. A→G or T→C) similar across populations? This question arises from our understanding of the phenomenon of GC-biased gene conversion (gBGC).

gBGC is a recombination-associated process that favors G/C over A/T nucleotides during the repair of mismatches that occur in heteroduplex DNA during meiosis (Duret, et al. 2002; Marais 2003; Duret and Galtier 2009; Galtier, et al. 2009; Glemin, et al. 2015). Although this process is not associated with natural selection, the efficiency of gBGC could also be reduced by genetic drift. Therefore, the effect of gBGC is expected to be much weaker in small populations than in large populations (Galtier, et al. 2009; Glemin, et al. 2015) and consequently, the frequencies of AT→GC polymorphisms are expected to differ among world populations. It is important to characterize the population specific patterns of genetic variants, as these patterns may have immense implications for human health.

For instance, previous studies have shown a positive correlation between the number of deleterious homozygous SNVs present in human populations and their distance from East Africa (Henn, et al. 2016). Furthermore, non-Africans were found to have much higher proportion of high frequency or homozygous deleterious variants than Africans (Do, et al. 2015; Subramanian 2016). These observations suggest that due to the effects of genetic drift, small populations (typically located away from Africa) have a higher fraction of high frequency (and homozygous) deleterious mutations than large populations. On the other hand, a gBGC mediated skews in the frequencies of deleterious AT→GC (relative to GC→AT) polymorphisms were also reported (Lachance and Tishkoff 2014; Xue, et al. 2016). However, it is unclear whether the extent of such skews is influenced by the effective population sizes of various global populations. Therefore, using data from the 1000 Genomes Project we investigated the pattern of nucleotide changes observed in the SNVs segregating in different allele frequencies (Genomes Project, et al. 2015). Furthermore, we

also analyzed homozygous and heterozygous variant data for 126 distinct populations from around the world, obtained from the Simons Genome Diversity Project (Mallick, et al. 2016).

Results

Allele frequencies and types of nucleotide changes in human populations

To quantify the difference in the patterns of observed AT→GC and GC→AT changes we derived a measure β, based on the Waterston estimator (θ_W) as described in the methods (Equation 1). The measure β is the ratio of AT→GC ($\mu_{AT \to GC}$) and GC→AT ($\mu_{GC \to AT}$) mutation rates, which captures the mutational equilibrium between AT and GC nucleotides. The ratio β is expected to be 1 if the observed AT→GC and GC→AT changes are due to the forward and reverse mutation rates alone. Any deviation from this ratio (β=1) suggests a bias in the substitution of one type over the other. To examine the variation in the patterns of nucleotide changes we obtained the 1000 Genomes phase II data for 26 distinct populations of the world. Since most of the genomes from Latin America were admixed with Europeans/Africans, we separated 20 Peruvian genomes that had <0.5% admixture. We used these to represent an un-admixed Native American population and hence the total number of populations analyzed was 27. The SNVs of these populations were grouped into eleven categories based on their Derived Allele Frequencies (DAF). The ratio β was estimated for each category of SNVs belonging to each population. We estimated the effective size (N_e) of each population based on the mutation rate (μ) and nucleotide diversity (π) (see methods).

Figure 1 shows the relationship between N_e and β for SNVs belonging to two extreme allele frequencies. For DAF < 0.025, the estimates of β were almost equal to 1 for all populations. In contrast, we observed a significant positive correlation ($P < 10^{-6}$) for SNVs with very high DAF (> 0.9). Another major pattern was that the β estimates were similar among populations

belonging to the same geographical locations and very different among populations from distinct locations. However, this was not true for admixed Americans (black dots). We then examined this relationship for SNVs with different derived allele frequencies. We did not find any significant relationship between N_e and β for low frequency SNVs until the DAF was ≤ 0.2 ($P > 0.10$) (Figure 2A). However, significant positive correlations (at least $P < 10^{-2}$) were observed between N_e and β for SNVs with DAF > 0.2. The magnitude of the correlation increased with the increase in DAF, which is evident from the rise in the slopes of the regression lines. This increase is clearer in Figure 2B, which shows the positive relationship between DAF and the slopes of the regression lines shown in Figure 2A. The slope observed for SNVs with DAF > 0.9 was 6×10^{-5}, which is almost 15 times higher than for SNVs with DAF = 0.2-0.3 (4×10^{-6}). For high frequency SNVs (DAF > 0.9), the difference between the mean β estimated for African genomes (1.72) was 79% higher than that observed for Peruvian genomes (1.35).

We also plotted β against DAF to show how this estimate changes with increasing DAF. For this purpose, we selected five representative populations with significantly different N_e. As shown in Figure 3, positive relationships between DAF and β against were observed for all populations and β increased with increasing DAF. However, based on the slopes of the regression lines (see Figure 2 legend-inset) the rate of increase correlates with N_e. The slope was the highest for Africans (0.57), intermediate for Eurasians (0.44-0.32) and the lowest for Peruvians (0.27).

Patterns of homozygous and heterozygous variations

To further examine the patterns of substitutions in a much larger number of populations we obtained the genotype data of the Simons Genome Diversity Project for genomes

representing 126 distinct ethnic populations. We examined the patterns of nucleotide changes for homozygous and heterozygous SNVs. For each genome, we estimated β for homozygous and heterozygous SNVs. The nucleotide diversity was estimated by comparing the pairs of chromosomes of each genome and N_e was calculated using the mutation rate obtained from previous studies (see methods). We then plotted β against N_e. For homozygous SNVs, our regression analysis produced a significant positive correlation ($P < 10^{-6}$) between the two variables (Figure 4A). Although the effective population sizes vary widely between world populations, they are roughly similar for the populations in a geographical location. Interestingly, the β values for homozygous SNVs are also varied considerably but were similar for populations within a geographical region. This is clear from the mean estimates shown in the inset of Figure 4A. On the other hand, N_e for heterozygous SNVs did not show any significant relationship ($P = 0.36$) with β (Figure 4B).

Proportion of deleterious AT→GC and GC→AT changes in global populations

To understand the implications of the observed patterns on human health, we examined the patterns of nucleotide changes for deleterious SNVs. To determine the deleteriousness of the SNVs we used the robust method, Combined Annotation-Dependent Depletion (CADD). This method integrates over 60 diverse annotations (to measure the extent of deleteriousness of a variant) into a single measure (C score) (Kircher, et al. 2014). We designated SNVs with a C score of >15 as deleterious. To distinguish the frequency of different of types of nucleotide changes we estimated the proportion of deleterious AT→GC (P_{AT}) and GC→AT (P_{GC}) changes using equations 2 and 3 (methods). These estimates were plotted against N_e. For deleterious SNVs with DAF > 0.9 we found highly significant relationships between N_e and P_{GC} ($P < 10^{-6}$) and between N_e and P_{AT} ($P < 10^{-6}$) (Figure 5A). Importantly for large populations (Africans) the proportion of deleterious AT→GC mutations was 54% higher than

the proportion of GC→AT mutations. For small populations (Peruvians) this difference was only 26%. We performed a similar analysis using the homozygous SNVs from 126 world populations and obtained similar results (Figure 5B). The difference between P_{GC} and P_{AT} was highest for Africans (51%) and lowest (22%) for Native Americans.

Discussions

In this study, we showed that the types of nucleotide changes in world populations are shaped by their effective population sizes. Our results revealed a much higher proportion of AT→GC variations in populations with large effective sizes (eg. Africans) compared to those with small sizes (eg. Native Americans). These observations could be explained based on the well-known recombination associated GC-biased gene conversion (gBGC) (Duret, et al. 2002; Marais 2003; Duret and Galtier 2009; Galtier, et al. 2009; Glemin, et al. 2015). The two strands of DNA are connected by double hydrogen bonds between A and T bases (weak) as well as triple hydrogen bonds between G and C bases (strong). It has been shown that gBGC favors the changes involving AT→GC (or weak → strong) compared to GC→AT (or strong → weak) during the process of fixation. We developed a measure β to capture this bias in substitutions. For rare SNVs, β estimates were close to 1 ($\beta \approx 1$) for all populations (Figure 1). This suggests that the rate of AT→GC mutations per (A or T) base is equal to GC→AT mutations per (G or C) base. Therefore, the observed number of rare SNVs reflect the mutation patterns alone.

In contrast, for high frequency variations β estimates were significantly higher than 1 ($\beta > 1$). Importantly β increased with the increase in DAF as shown clearly in Figure 3. These results suggest a preferential fixation of AT→GC over GC→AT mutations over time. However, based on the slopes of the regression lines the results also suggest that the rate of preferential

of fixation of AT→GC mutations in small populations is low compared to that in large populations because of genetic drift that reduces the efficiency of gBGC. This result is supported by a previous study using human population genetic data that suggested that on an average gBGC is stronger in African than non-African populations (Glemin, et al. 2015).

To further support our claim, we examined the changes between A↔T (weak bond) and between G↔C (strong bond) nucleotides using high frequency SNVs (DAF > 0.9). Our results showed that β estimated for A→T/T→A or C→G/G→C were close to 1 for all populations and there was no significant relationship with N_e (Supplementary Figure S1). We also examined this using homozygous and heterozygous SNVs from the 126 populations and observed no significant relationship between β and N_e (Supplementary Figure S2A and S2B). Since the changes are between the same types of nucleotides (with respect to weak or strong hydrogen bonds) there was no effect of gBGC on the fixation of one type of nucleotide over other. This provides further evidence that the results of our study are not due to methodological artifacts.

Since gBGC does not affect the changes between A↔T (weak bond) and between G↔C (strong bond) previous studies have used the rate of these changes as a normalizing factor to assess the magnitude of gBGC on GC↔AT changes (Lachance and Tishkoff 2014; Glemin, et.al 2015; Xue and Chen 2016). Following this, we normalized AT→GC with A↔T changes and GC→AT with G↔C changes respectively and developed a normalized ratio, β' (equation 2 - methods). However, the relationship between N_e and β' was also highly significant and comparable to previous results obtained for high frequency (Supplementary Figure S3A) and homozygous SNVs (Supplementary Figure S3B). This further supports our

results as the normalization eliminates any variation in the mutation rates between populations.

The results based on homozygous and heterozygous SNVs shown in Figure 4 further support to the results based on DAF presented in Figures 1 and 2. For instance, Figures 1 and 2A showed there was no significant relationship between β and N_e for low frequency SNVs. Since low frequency SNVs predominantly exist as heterozygous SNVs in a genome, results based on the former are expected to be similar to those based on the latter (Figures 1 and Figure 4B). Similarly, as high frequency SNVs are more likely to be present as homozygous SNVs in genomes, the results for these two types of SNVs are alike. This is evident from the results shown in Figures 1 and 4A.

Since gBGC is mediated by recombination it effects were found to be strong for highly recombining regions. To examine this, we obtained SNVs from regions with low (<2 cM), medium (2-20 cM) and high (>20 cM) rates of recombination. The results showed a much higher β for the variants in high recombination regions (Supplementary Figure S4). However, the magnitude of the relationship between β and N_e was similar for all three regions; the slopes of the regression lines were 0.000039 (low), 0.000042 (medium) and 0.000041 (high). This suggests that the influence of population size on gBGC is relatively similar across chromosomal regions with varying degrees of recombination.

Previous studies have shown a significant negative correlation between heterozygosity and the distance (of the location of the populations) from Africa. This correlation is expected based on the prediction that during migration out of Africa human populations underwent a series of population bottlenecks or founder effects along the migratory route (Prugnolle, et al.

2005; Ramachandran, et al. 2005; Handley, et al. 2007; DeGiorgio, et al. 2009). This is because only a subset of people migrated from the original location to new sites and hence the size of the populations reduced with distance from Africa. From previous studies, we obtained the geographic distance of 41 non-African populations from Eastern Africa (Addis Ababa) (Li, et al. 2008; DeGiorgio, et al. 2009) and we plotted the estimates of β against them. We obtained a highly significant negative correlation for homozygous variants ($P < 10^{-6}$) (Figure 6A) but not for heterozygous variants ($P = 0.2$) (Figure 6B). This is very similar to the results shown in Figure 4. In this analysis, the geographic distance from Africa was used as a proxy for N_e. Hence this result independently confirms our findings and also justifies the method used in this study to estimate N_e from nucleotide diversity.

We estimated N_e under the assumption that mutation rate (μ) and the rate of accumulation of mutations are both similar between populations. However, a recent study suggested a slightly (~5%) higher rate of mutation accumulation in non-Africans compared to Africans (Mallick, et al. 2016). To accommodate the elevated diversity, we subtracted 5% of the observed divergence for non-Africans while estimating N_e and re-analyzed the data. This produced almost identical patterns and similar strengths of correlation to that reported in Figure 4 (Supplementary Figure S5).

We have shown that the types of SNVs observed in different human populations are very likely to be modulated by their effective population sizes. Since this pattern was universal for genome-wide variations we showed that deleterious SNVs also follow this pattern. Our results showed that populations with large effective sizes (e.g. Africans) displayed the greatest difference between the proportions of high frequency deleterious AT→GC and GC→AT SNVs. This difference was much lower in populations with small effective sizes

(e.g. Native Americans). This has significant implications in human health as it implies that high frequency diseases-associated mutations in Africans will be more enriched with AT→GC SNVs than in Native Americans. Furthermore, we showed that deleterious homozygous SNVs are also predominantly AT→GC in Africans, and to a greater extent than in non-Africans. This suggests the possibility that recessive genetic disorders in Africans are more likely to be caused by AT→GC variants than in non-Africans. Therefore, our study recommends that genome-wide association studies should consider the frequency of population specific nucleotide changes.

Materials and Methods

Genome data

We obtained genotype data from the 1000 genome project – Phase II (Genomes Project, et al. 2015). The genome-wide variations from the 26 populations including Africans (seven populations: 661 individuals), South Asians (five: 489), European (five: 503), East Asian (five: 504) and South Americans (four: 347). Although there were 85 Peruvian genomes were available, most of these were admixed with Europeans and Africans. Hence, we used the likelihood based clustering algorithm *Admixture* (Alexander, et al. 2009) and examined the proportion of admixture in each Peruvian genome. Our results showed that only 20 genomes (40 chromosomes) had <0.5% admixture from other populations and we included these un-admixed Peruvians as the 27th population in our analyses. To identify derived alleles, orientations of SNVs were determined using the ancestral state of the nucleotides, which was inferred from six primate EPO alignments (Genomes Project, et al. 2010). The SNVs were divided into eleven categories based on their derived allele frequencies (see Figure 2-legend). For the SNVs in each category we computed the counts of six types of changes: A/T→G/C, G/C→A/T, C→G, G→C, A→T and T→A (see below).

We also obtained the genotype data from the Simon Genome Diversity Project (Mallick, et al. 2016). To examine the patterns of nucleotide changes we used the homozygous and heterozygous SNVs present in a single representative genome from each of the 126 populations. We excluded four African hunter-gatherer populations from our analysis as it was difficult to ascertain the correct orientation of the nucleotide changes in these genomes. For each genome, we estimated the number of homozygous and heterozygous changes belonging to the six types described above.

Deleterious mutations

To determine the deleterious nature of a SNV we used a robust method, Combined Annotation-Dependent Depletion (CADD) that integrates diverse annotations into a single measure (C score) (Kircher, et al. 2014). The extent of deleteriousness was further determined by estimating the corresponding selective coefficients for these scores (Racimo and Schraiber 2014). The C scores for each SNV in the 1000 Genomes Project data were publically available (http://cadd.gs.washington.edu/download). Using an in-house Perl script, we combined this score with the genome data by using the chromosomal co-ordinates of the SNVs. For the deleterious variant analysis, we included only the SNVs for which the C score was available. We used a C score of ≥15 to determine a mutation to be deleterious in nature following previous studies (Kircher, et al. 2014; Subramanian 2015). However, using a different threshold produced almost identical results.

Estimating the ratio of mutation rates

The ratio of AT→GC and GC→AT mutation rates could be estimated based on the Waterson estimator (Watterson 1875), $\theta = 4N_e\mu = S/a_n$, where N_e is the effective population size and S

is the number of segregating sites per site and $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$. We can use this estimator considering only one type of mutation as:

$$\theta_{A \to G} = 4 N_e \mu_{A \to G} = \frac{S_{A \to G}}{a_n}$$

$$\theta_{G \to A} = 4 N_e \mu_{G \to A} = \frac{S_{G \to A}}{a_n}$$

The ratio of forward and reverse nucleotide changes (β) could be obtained as:

$$\beta = \frac{\mu_{A \to G}}{\mu_{G \to A}} = \frac{\theta_{A \to G}}{\theta_{G \to A}} = \frac{S_{A \to G}}{S_{G \to A}}$$

The number of segregating sites per site or SNVs per site of a genome can be estimated as:

$$S_{A \to G} = \frac{M_{A \to G}}{N_A} \; and \; S_{G \to A} = \frac{M_{G \to A}}{N_G}$$

where $M_{A \to G}$ and $M_{G \to A}$ are the number of observed A→G and G→A mutations in a genome respectively and N_A and N_G are the number of ancestral A and G nucleotides. This formula can be extended for the combined AT→GC and GC→AT mutation rates because each pattern is mutually exclusive and hence the ratio of nucleotide changes (β) is:

$$\beta = \frac{\mu_{AT \to GC}}{\mu_{GC \to AT}} = \frac{\theta_{AT \to GC}}{\theta_{GC \to AT}} = \frac{S_{AT \to GC}}{S_{GC \to AT}}$$

The number of segregating sites or SNVs in a genome can be calculated as:

$$S_{AT \to GC} = \frac{M_{A \to G} + M_{A \to C}}{N_A} + \frac{M_{T \to C} + M_{T \to G}}{N_T}$$

$$S_{GC \to AT} = \frac{M_{G \to A} + M_{G \to T}}{N_G} + \frac{M_{C \to T} + M_{C \to A}}{N_C}$$

Since A and T as well as C and G are complementary to each other in a double-stranded DNA they are equal in number. Therefore, β can be expressed as,

$$\beta = \frac{M_{A \to G} + M_{A \to C} + M_{T \to C} + M_{T \to G}}{M_{G \to A} + M_{G \to T} + M_{C \to T} + M_{C \to A}} \times \frac{N_{GC}}{N_{AT}}$$

$$\beta = \frac{M_{AT \to GC}}{M_{GC \to AT}} \times \frac{N_{GC}}{N_{AT}} \quad \to \quad (1)$$

This derivation proves that the ratio of forward and reverse rates of changes can be calculated by simply taking the ratio of the observed counts of AT→GC ($M_{AT \to GC}$) and GC→AT ($M_{GC \to AT}$) changes and multiplying with the ratio of the number of GC (N_{GC}) to AT (N_{AT}) nucleotides in a genome. Since equation 1 represents the ratio of mutation rates ($\mu_{AT \to GC}$ and $\mu_{GC \to AT}$) this ratio is expected to be 1 ($\beta = 1$) if the observed nucleotide changes are solely due to the result of these mutation rates. Any deviation from this suggests a bias in the substitution process. While $\beta > 1$ indicate an excess of AT→GC substitutions $\beta < 1$ imply an excess of GC→AT substitutions.

GC-biased gene conversion is known to affect only the changes involving weak (A or T) to strong (G or C) nucleotides but not the changes within weak (A↔T) or within strong (G↔C) nucleotides. Hence the latter is not expected to vary between populations with different N_e. Therefore, we used this as a normalization factor and developed a normalized ratio of AT→GC to GC→AT (β').

The A→G mutation rate can be normalized using A→T rate and the normalized rate ($\tau_{A \to G}$) can be expressed as:

$$\tau_{A \to G} = \frac{\mu_{A \to G}}{\mu_{A \to T}} = \frac{\theta_{A \to G}}{\theta_{A \to T}} = \frac{S_{A \to G}}{S_{A \to T}} = \frac{(M_{A \to G}/N_A)}{(M_{A \to T}/N_A)} = \frac{M_{A \to G}}{M_{A \to T}}$$

Similarly, we can obtain this expression for AT→GC and GC→AT rates as:

$$\tau_{AT \to GC} = \frac{\mu_{A \to G}}{\mu_{A \to T}} + \frac{\mu_{A \to C}}{\mu_{A \to T}} + \frac{\mu_{T \to C}}{\mu_{T \to A}} + \frac{\mu_{T \to G}}{\mu_{T \to A}}$$

$$\tau_{AT \to GC} = \frac{M_{A \to G} + M_{A \to C}}{M_{A \to T}} + \frac{M_{T \to C} + M_{T \to G}}{M_{T \to A}}$$

$$\tau_{GC \to AT} = \frac{\mu_{G \to A}}{\mu_{G \to C}} + \frac{\mu_{G \to T}}{\mu_{G \to C}} + \frac{\mu_{C \to T}}{\mu_{C \to G}} + \frac{\mu_{C \to A}}{\mu_{C \to G}}$$

$$\tau_{GC\to AT} = \frac{M_{G\to A} + M_{G\to T}}{M_{G\to C}} + \frac{M_{C\to T} + M_{C\to A}}{M_{C\to G}}$$

Therefore, the normalized ratio of AT→GC to GC→AT (β') is:

$$\beta' = \frac{\tau_{AT\to GC}}{\tau_{GC\to AT}} \qquad \rightarrow \qquad (2)$$

The relationship between nucleotide diversity (π), mutation rate (μ) and effective population size (N_e) for diploid organisms is $\pi = 4N_e\mu$. Using this relationship, we calculated the effective population size as $N_e = \pi/4\mu$. We used the observed nucleotide diversity of a population or of a diploid genome and used a mutation rate of 1.2×10^{-8} substitutions per site per generation following many studies based on human pedigree genome data (Roach, et al. 2010; Conrad, et al. 2011). A recent suggested that the rate of mutation accumulation in non-European genomes could be slightly (5%) higher than that of Africans (Mallick, et al. 2016). To accommodate this difference, we subtracted 5% of the nucleotide diversity for non-African populations only while calculating the effective population sizes and obtained almost identical results (Supplementary figure S4).

Estimation of the proportion of AT→GC counts

We also estimated the proportion of AT→GC counts (P_{GC}) and GC→AT ($N_{GC\to AT}$) counts, which were calculated as:

$$P_{GC} = \frac{N_{A\to G} + N_{A\to C} + N_{T\to C} + N_{T\to G}}{N} \qquad \rightarrow \qquad (3)$$

$$P_{AT} = \frac{N_{G\to A} + N_{G\to T} + N_{C\to A} + N_{C\to T}}{N} \qquad \rightarrow \qquad (4)$$

N is the number of all types of nucleotide changes. The standard error of $P_{AT\to GC}$ and $P_{GC\to AT}$ were calculated using the binomial variance.

Author Contributions

Conceptualization: SS

Data Analysis: SS

Writing: SS

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664.

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al. 2011. Variation in genome-wide mutation rates within and between human families. Nat Genet 43:712-714.

DeGiorgio M, Jakobsson M, Rosenberg NA. 2009. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proc Natl Acad Sci U S A 106:16057-16062.

Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet 47:126-131.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285-311.

Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162:1837-1847.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25:1-5.

Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. Nature 526:68-74.

Glemin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res 25:1215-1228.

Handley LJ, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. Trends Genet 23:432-439.

Henn BM, Botigue LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A 113:E440-449.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310-315.

Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. Am J Hum Genet 95:408-420.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538:201-206.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet 19:330-338.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159-160.

Racimo F, Schraiber JG. 2014. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. PLoS Genet 10:e1004697.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102:15942-15947.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636-639.

Stringer C. 2003. Human evolution: Out of Ethiopia. Nature 423:692-693, 695.

Subramanian S. 2016. Europeans have a higher proportion of highfrequency deleterious variants than Africans. Hum Genet 135:1-7.

Subramanian S. 2015. Using the plurality of codon positions to identify deleterious variants in human exomes. Bioinformatics 31:301-305.

Watterson GA. 1875. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256-276.

Xue C, Chen H, Yu F. 2016. Base-Biased Evolution of Disease-Associated Mutations in the Human Genome. Hum Mutat 37:1209-1214.

Figure legends

Figure 1. Relationship between the effective population size (N_e) and the ratio of AT→GC to GC→AT (β) changes. The genotype data was obtained for 27 populations of the world. The colors indicate different geographical locations of the populations. The relationship was not significant for SNVs with DAF <0.025 (P = 0.11, using the Spearman rank correlation) but highly significant for those with DAF > 0.9 ($P < 10^{-6}$).

Figure 2. (A) The correlation between N_e and β observed for SNVs belonging to eleven allele frequency categories with DAF: <0.025, 0.025-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9 and > 0.9. Only the upper values of the range are shown in X-axis. The correlations observed for SNVs with DAF < 0.2 were not statistically significant (P > 0.10) and remaining relationships were highly significant (at least $P < 10^{-2}$). **(B)** Scatter plot showing the magnitude of the relationship shown in figure 2A for difference allele frequency categories. X-axis shows the mid-values of the derived allele frequency categories and the slopes of the regression lines of figure 2A are shown on the Y-axis. The slopes of the lines for SNVs with DAF ≤ 0.2 are not included as those relationships were not statistically significant.

Figure 3. The magnitude of β with respect to the derived allele frequencies. The correlation between DAF and β were highly significant (at least P < 0.0017). Although all populations show an increasing trend, the rate of increase varies between them, which are manifested in the slopes of the regression lines (shown in the figure legend).

Figure 4. The relationship between effective population size (N_e) and the normalized ratio AT→GC to GC→AT (β) estimated for homozygous **(A)** and heterozygous **(B)** SNVs present

in individual genomes belonging to 126 distinct populations of the world. The data was obtained from the Simons Genome Variation Project. The correlation for homozygous SNVs was highly significant ($P < 10^{-6}$) but not statistically significant for heterozygous SNVs ($P = 0.36$). The colors represent populations from distinct geographical locations. The relationship shown in the inset was based on the estimates of N_e and β averaged for the populations belonging to a specific geographical location. The relationship shown in the inset was also significant ($P < 0.026$).

Figure 5. Variations in the proportions of deleterious AT→GC and GC→AT SNVs among world populations. These proportions correlate significantly with the effective population sizes of the populations. The trends were positive for AT→GC SNVs and negative for GC→AT SNVs. The relationships were based on **(A)** high frequency SNVs with DAF > 0.9 and **(A)** homozygous SNVs. All relationships were highly significant (at least, $P < 10^{-6}$).

Figure 6. Correlation between the geographical distance of non-African populations from East Africa (Addis Ababa) and the normalized ratio AT→GC/GC→AT (β) using **(A)** Homozygous **(B)** Heterozygous SNVs. The relationship was highly significant for homozygous SNVs ($P < 10^{-6}$) and not for heterozygous SNVs ($P = 0.2$). The distances between the locations of the non-African populations and Addis Ababa were obtained from previous studies (Li, et al. 2008; DeGiorgio, et al. 2009).
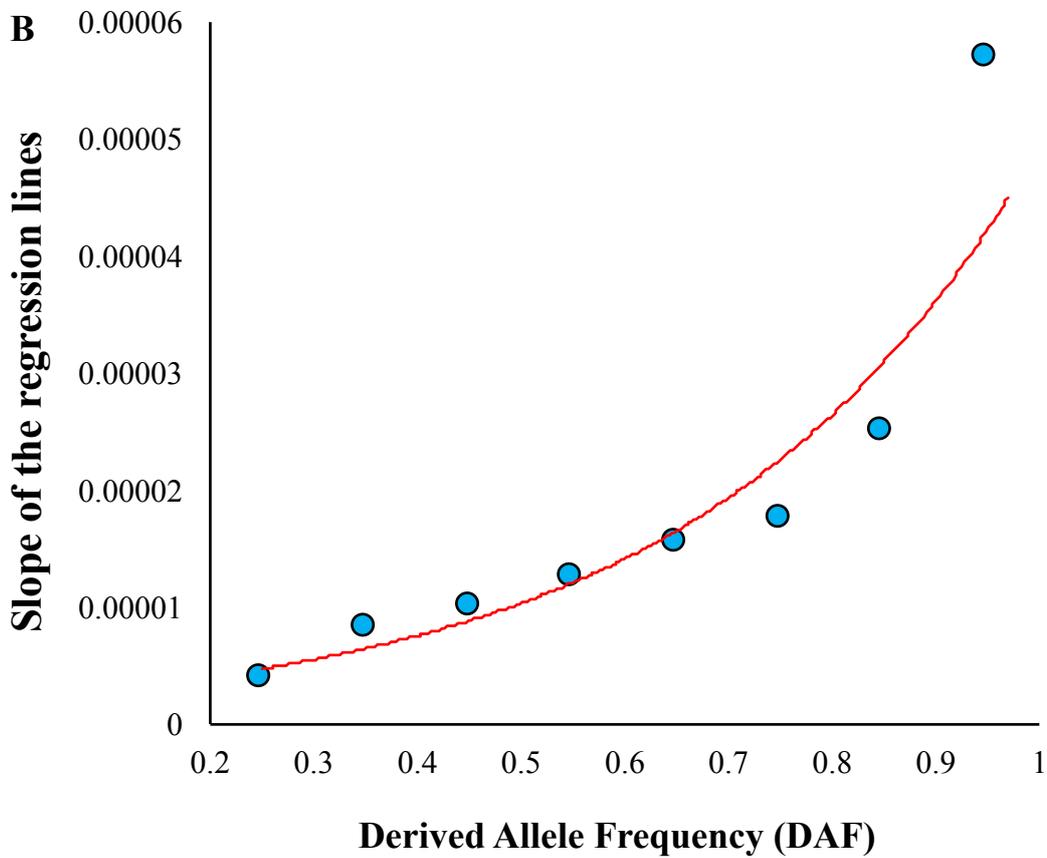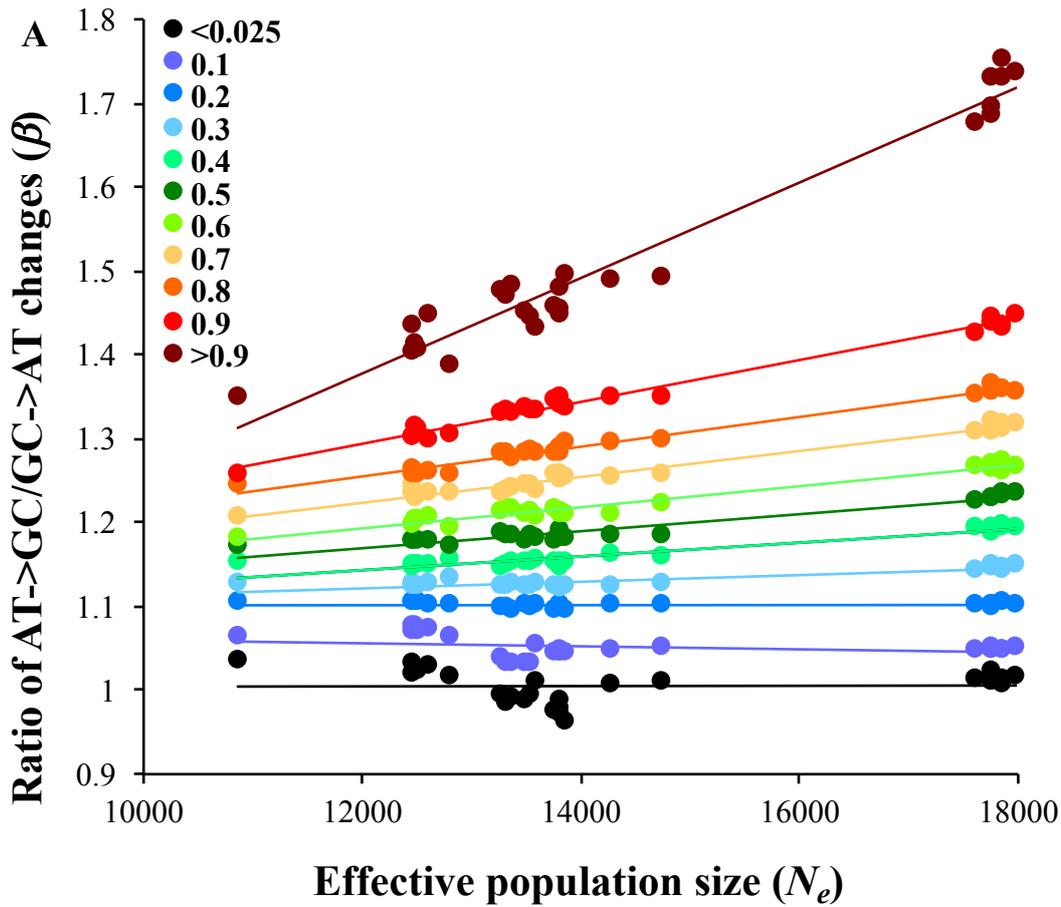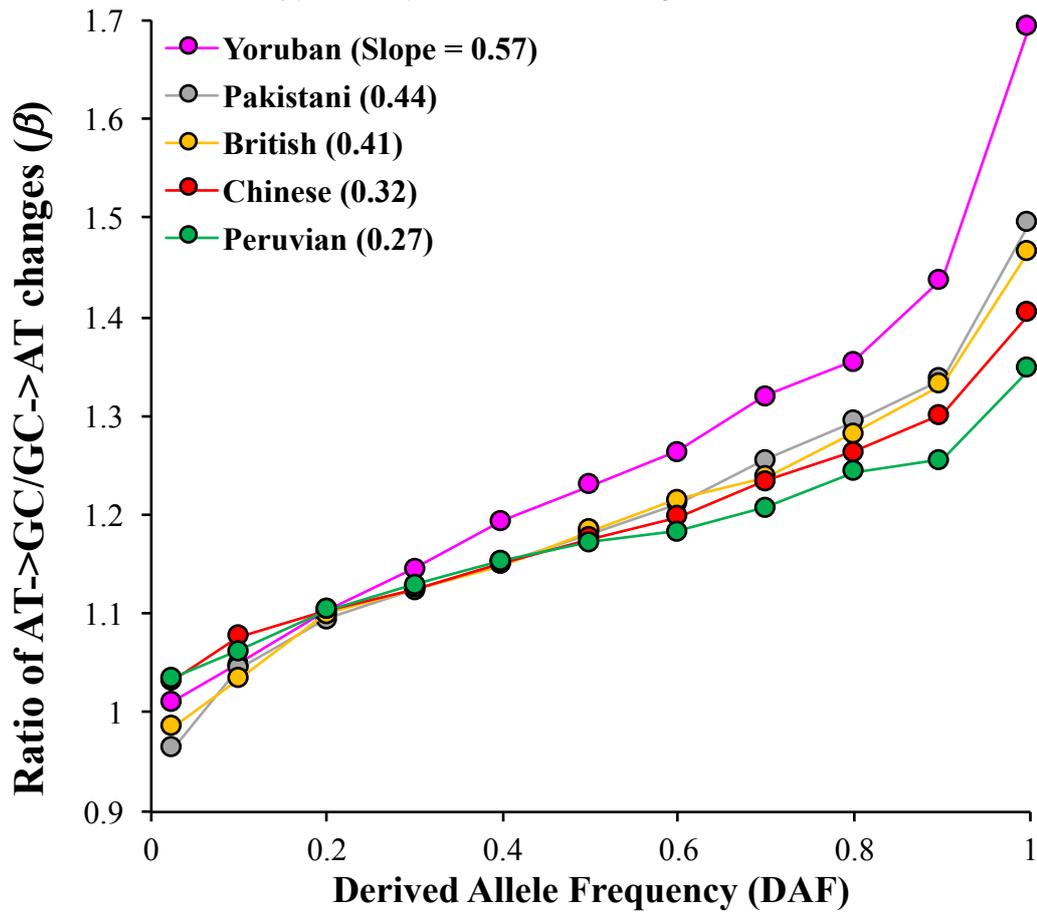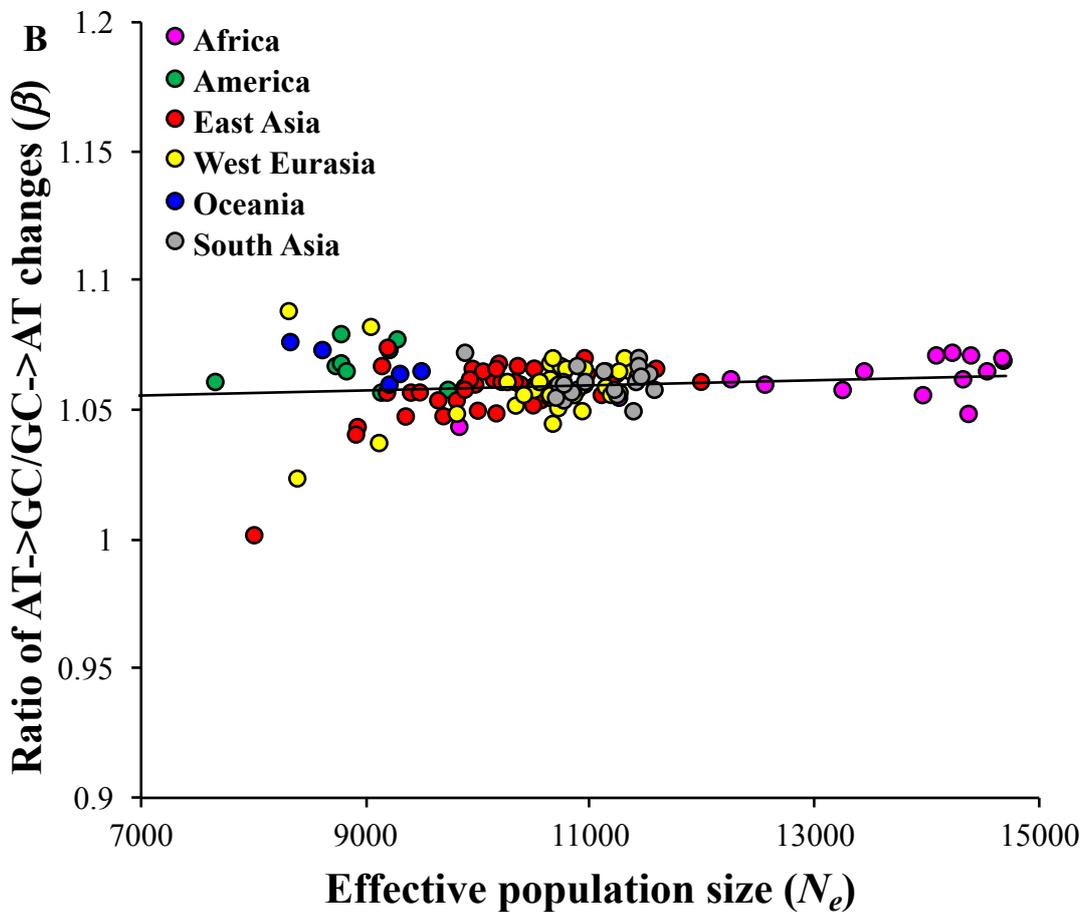
Figure 1
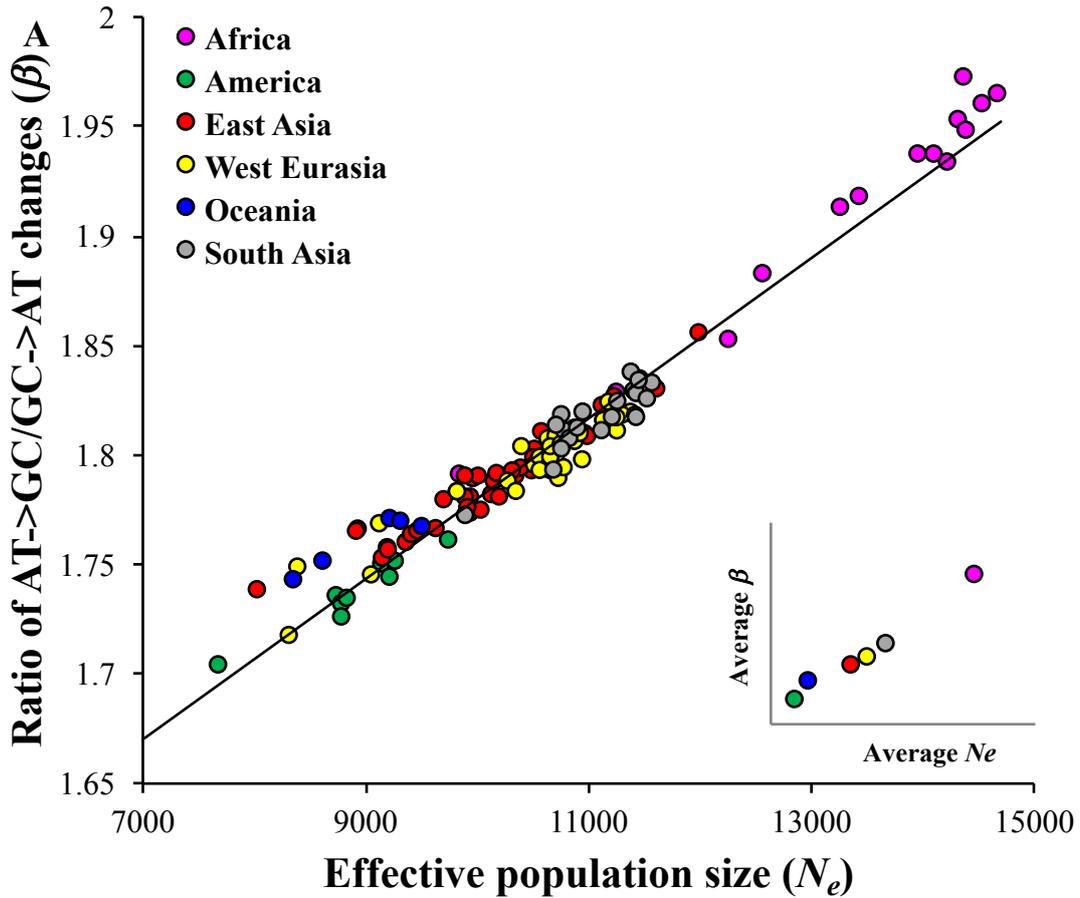
Figure 2

Figure 3

Figure 4

Figure 5

Figure 6