Relative mutation rates of each nucleotide for another estimated from allele frequency spectra at human gene loci

LEEYOUNG PARK
Natural Science Research Institute, Yonsei University, 134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-749, Korea

(*Received 28 January 2009 and in revised form 12 June 2009*)

Summary

This study aims to comprehensively examine the mutation rates of one base for another in human gene loci. In contrast to most previous efforts based on divergence data from untranscribed regions, the present study employs the basic theory of the reversible recurrent mutation model using large-scale, high-quality re-sequencing data from public databases of gene loci. Population mutation parameters ($4N\nu$ and $4N\mu$) are obtained for each pair of base substitutions. The estimated parameters show good strand reversal symmetry, supporting the existence of mutation-drift equilibrium. Analysis of specific gene regions including mRNA, coding sequence (CDS), 5′-untranslated region (5′-UTRs), 3′-UTR and intron shows that there are clear differences in the mutation rates of each base for another depending on the location of the base in question. Results from analyses that take the adjacent bases into account exhibit excellent strand reversal symmetry, confirming that the identity of an adjacent base influences mutation rates. The CpG to TpG (or CpG to CpA) substitution is found at a rate approximately seven-fold higher than the reverse transition in intron regions due to cytosine deamination, but the effect is strongly reduced in mRNA regions and almost entirely lost in 5′-UTRs. However, from the overall increased transitions in sites other than CpGs and the proportion of CpGs in the total sequence, CpG methylation is not the main factor responsible for the increased rate of transitions as compared with transversions. In this report, after adjusting average mutation rates to the sequence compositions, no substitution bias is found between A+T and C+G, indicating base composition equilibrium in human gene loci. Population differences are also identified between groups of people of African and European descent, presumably due to past population histories. By applying the basic theory of population genetics to re-sequenced data, this study contributes new, detailed information regarding mutations in human gene regions.

1. Introduction

Mutation rate is one of the most important parameters in genetics (Crow, 2000). It can provide crucial information about the evolution of genome sequence composition. Unfortunately, direct estimation of mutation rates in higher eukaryotes is not as easy as in microbes (Drake *et al.*, 1998). An attempt has been made to measure bidirectional somatic mutation rates using hyper-mutating chicken B cells (Jolly *et al.*, 2007), but the method is not widely applicable. Human mutations are much more complicated than expected (Duret, 2009). It is well known that human mutations are influenced by factors such as age and gender (Drake *et al.*, 1998; Crow, 2000, 2006). Moreover, transcription-induced mutation biases are observed in gene loci (Green *et al.*, 2003; Polak & Arndt, 2008). Direct estimation of mutation rates found differential mutation rates across loci (Drake *et al.*, 1998; Reich & Lander, 2001; Kondrashov, 2002). However, this direct estimation, using the incidence of dominant Mendelian diseases in humans, would only be appropriate for providing approximate information.

Estimates of human mutation rates are usually based on theoretical population genetics. One earlier estimate using a phylogenetic approach resulted in an average mutation rate of $\sim 2{\cdot}5 \times 10^{-8}$ per nucleotide, assuming the effective population size to be 10^4 or 10^5

Tel: (82)2 2123 7615. Fax: (82)2 313 8892. e-mail: lypark@yonsei.ac.kr

(Nachman & Crowell, 2000). This mutation rate is very similar to that found by direct estimation (Drake *et al.*, 1998; Kondrashov, 2002). There are many ways to estimate mutation rates based on theoretical population genetics, including coalescent-based approaches (Kimura & Ohta, 1971; Clark *et al.*, 2005; Hartl & Clark, 2007). However, most previous estimates are based on phylogeny, probably due to the availability of appropriate data. All of the approaches based on population genetics should be applied with caution, since past population histories and selective pressure can influence the results (Hartl & Clark, 2007).

It is well known that transition mutations are much more frequent than transversions and that the biochemical properties of CpG methylation help facilitate C-to-T transition mutations (Krawczak *et al.*, 1998; Strachan & Read, 2004). Therefore, it is highly plausible that mutation rates will vary depending on the adjacent nucleotide. Moreover, recent bovine genome data strongly support neighbouring nucleotide effects on mutations (Jiang *et al.*, 2008). Phylogeny-based methods that account for nearest-neighbour interactions (Hwang & Green, 2004; Lunter & Hein, 2004; Arndt & Hwa, 2005) also confirm the high C-to-T transition mutation rate, especially in CpG sites. More recently, a similar method based on phylogeny was developed based on the general time-reversible model, but the authors did not apply the method to real data (Catanzaro *et al.*, 2006). However, the sequences of untranscribed regions were used for most of the estimates based on phylogeny. Moreover, rate estimates generated using the Human Genome Mutation Database (HGMD) are limited because the data rely solely upon reported mutations, which are usually associated with diseases (Krawczak *et al.*, 1998). The analysis of re-sequencing data of human gene loci can provide more detailed general information regarding these issues in human gene loci.

Previous studies have found that the evolution of genome sequence composition is quite complex (Duret *et al.*, 2002; Lercher *et al.*, 2002; Piganeau *et al.*, 2002; Webster *et al.*, 2003; Belle *et al.*, 2004; Duret, 2009) and is even influenced by recombination (Duret & Arndt, 2008). In previous comparative genomics studies, more fixed nucleotide substitutions of G/C to A/T were found, but polymorphism data suggested more nucleotide substitutions of A/T to G/C. To address such discrepancies, more extensive estimates using public re-sequencing data based on the basic theories of population genetics would be helpful. The number of polymorphic sites from re-sequencing data provides the overall mutation rate, and the allele frequency spectra are used to estimate the mutation rate of one base for another (Wright, 1931; Kimura & Ohta, 1971). Mutation rates in this study will be estimated using allele frequency spectra from large-scale, high-quality, public re-sequencing data generated by the SeattleSNPs Program for Genome Application (PGA). Like other estimates based on theoretical population genetics, these methodologies require knowledge of the effective population size in order to estimate mutation rates. Since the primary purpose of this study is to enhance our knowledge of mutations and the evolution of the human genome, the current study focuses on determining relative rates of each base being mutated for another, without discussing the actual effective population sizes.

2. Methods

(i) *Data*

Data from the SeattleSNPs PGA were used for the analyses. A detailed description of the data and the patterns of variations has been reported previously (Crawford *et al.*, 2005). There were 24 samples of African descent (AD) and 23 of European descent (ED), consisted of the first and second study populations in the SeattleSNPs PGA. Excluding X-linked genes and re-sequenced genes using the third study population, 292 genes were selected to estimate the population mutation parameters for each base substitution (Supplementary Table 1). Because the evolutionary processes governing mutations in repeat regions such as Alu, Mer and others, may be different from those of the rest of the genome, such regions were also excluded. Finally, tri-allelic sites were excluded, bringing the total number of base pairs (bp) examined to 4 437 493.

(ii) *Mutation rates for each base to another*

Let μ represent the rate per generation of the mutation of 'A' to 'a' and ν represent the rate of the opposite mutation. Let the frequency of 'A' be denoted as q. If mutations are subject to uniform rates of recurrence of mutation and reverse mutation over the whole genome, the distribution of allele frequencies for reversible recurrent mutation at equilibrium can be expressed as the following beta distribution, in which N is the effective population size (Wright, 1931; Kimura & Ohta, 1971):

$$\phi(q)=\frac{\Gamma(4N\mu+4N\nu)}{\Gamma(4N\mu)\Gamma(4N\nu)}q^{4N\nu-1}(1-q)^{4N\mu-1}. \tag{1}$$

In contrast to a scenario in which there is no mutation, where the derived allele frequency distribution would ignore the fixed alleles, the frequency, q, in this study involves fixed alleles. From eqn (1), the population mutation parameter of A-to-a is $4N\mu$, and that of a-to-A is $4N\nu$. These parameters can be estimated

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press

from the mean and variance of the beta distribution. The estimates of $4N\mu$ and $4N\nu$ are expressed as follows, where m is the mean and 'var' is the variance:

$$4N\nu = m \times \left(\frac{m \times (1-m)}{\text{var}} - 1\right),$$
$$4N\mu = (1-m) \times \left(\frac{m \times (1-m)}{\text{var}} - 1\right). \tag{2}$$

Data extraction and handling were performed using Perl programming, and the actual estimations were conducted using the R statistical package. First, the total sequence was analysed and sub-sequences were then analysed based on whether or not they were transcribed and/or translated. If there were transcript isoforms, the longer isoform was selected for the region classification. In total, analyses were conducted on the following eight regions: complete sequence (indicated as 'Total' in tables), mRNA, coding sequence (CDS), 5′-untranslated region (5′-UTR), 3′-UTR, all UTRs, intron and untranscribed region. Here, the mRNA region included the fully processed, mature mRNA without any introns. For estimations of population mutation parameters while taking into account the adjacent nucleotide effect, the data were divided into groups based on the identity of the next and previous nucleotides. Correlations between pairs of mutation rates expected to be equal due to strand reversal symmetry, assuming mutation-drift equilibrium were analysed by non-parametric rank-based measures for association, Kendall's *tau* and Spearman's *rho* statistics in the R package. Additionally, correlations between rates of AD and ED samples were tested to examine whether there was a population difference.

Another approach used to estimate the population mutation parameters was based on the number of polymorphic sites. If θ is the population mutation parameter $4N\mu$ and P is the number of polymorphic sites, then θ can be expressed as eqn (3) (Kimura & Ohta, 1971; Hartl & Clark, 2007). Here, μ is the overall mutation rate. In this equation, q is an arbitrary frequency that defines whether or not a variant is polymorphic. If the frequency of a variant is higher than q, the site is considered to be polymorphic. In this study, q is equal to one divided by the number of sample chromosomes (1/48 for the AD samples and 1/46 for the ED samples):

$$\theta = \ln(1-P)/\ln(q) \tag{3}$$

(iii) *Simulations*

To test whether the differences in mutation rates in the AD and ED samples come from the past population histories, samples were generated based on expected past population history using a program, ms, which generates samples under the Wright–Fisher neutral model (Hudson, 1983, 2002). The same parameters of past population histories as a previous study were used in this study. The parameters were derived by extensive assessment of the possible past population histories of African–Americans and European–Americans (Boyko *et al.*, 2008). However, the number of samples was changed to 48 for the AD samples and 46 for the ED samples to reflect the data set used in this study. The African expansion model was applied for the AD samples, and the European simple bottleneck and European complex bottleneck models were applied for the ED samples. Simulations were repeated ten times.

3. Results

Mutation rates of one base for another were estimated with re-sequencing data, using the reversible recurrent mutation model at equilibrium. As described in the Methods section, repeat regions, insertions and tri-allelic sites were excluded, and a total of 4 437 493 bp in 292 genes were examined, including 21 472 polymorphic sites. Sixteen polymorphic sites found within inserted regions were also excluded from the data. Among the polymorphic sites, 8444 sites were polymorphic in both populations, 10 015 sites were polymorphic only in the AD population and 3013 sites were polymorphic only in the ED population. Bases analysed, excluding the polymorphic sites, included: A, 1 182 263; C, 986 450; G, 1 007 768; and T, 1 239 540. The data set contained 1·21 times more A+T than C+G (45·2% GC content).

(i) *Overall estimates*

The population mutation parameters ($4N\mu$) were estimated to examine relative mutation rates, which are the product of effective population size (N) and mutation rate per base per generation (μ). Estimates of the population mutation parameter based on the number of polymorphic sites (Kimura & Ohta, 1971; Hartl & Clark, 2007) were 0·00108 for the AD samples and 0·000675 for the ED samples. These estimates are similar to the previous estimates using coalescent- or phylogeny-based methods (Nachman & Crowell, 2000; Bhangale *et al.*, 2005; Carlson *et al.*, 2005). The estimated mutation rates based on the reversible recurrent mutation model were also similar to values from these earlier studies. Taking sequence constitution into consideration, the average population mutation parameters were 0·000810 for the AD samples and 0·000613 for the ED samples. If the effective population number is assumed to be approximately 10 000, the same as in the previous study (Nachman & Crowell, 2000), the actual mutation rates would be $2{\cdot}02 \times 10^{-8}$ in the AD samples and $1{\cdot}53 \times 10^{-8}$ in the ED samples, which is similar to the

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press

Table 1. *Estimates of* θ ($4N\nu$ *and* $4N\mu$) *for the AD samples, which are scaled up by* 10^3

Population	Direction	AC	AG	AT	CG	CT	GT
Total	$X\rightarrow Y$	0·118	0·530	0·087	0·173	0·606	0·149
	$X\leftarrow Y$	0·141	0·621	0·083	0·170	0·483	0·121
	[a]	1433	6429	1072	1809	6188	1528
mRNA	$X\rightarrow Y$	0·094	0·483	0·043	0·107	0·463	0·089
	$X\leftarrow Y$	0·091	0·480	0·044	0·110	0·498	0·093
	[a]	167	821	77	211	815	152
CDS	$X\rightarrow Y$	0·076	0·439	0·019	0·096	0·419	0·068
	$X\leftarrow Y$	0·068	0·404	0·021	0·099	0·518	0·081
	[a]	95	461	24	121	484	74
5′-UTR	$X\rightarrow Y$	0·164	0·730	0·039	0·098	0·567	0·054
	$X\leftarrow \mathrm{Y}$	0·108	0·515	0·043	0·105	0·959	0·085
	[a]	15	53	8	21	72	12
3′-UTR	$X\rightarrow Y$	0·108	0·521	0·076	0·125	0·532	0·137
	$X\leftarrow Y$	0·130	0·638	0·070	0·127	0·412	0·104
	[a]	57	307	45	69	259	66
UTRs	$X\rightarrow Y$	0·118	0·549	0·075	0·124	0·543	0·126
	$X\leftarrow Y$	0·130	0·619	0·071	0·127	0·469	0·106
	[a]	72	360	53	90	331	78
Intron	$X\rightarrow Y$	0·122	0·545	0·096	0·193	0·658	0·169
	$X\leftarrow Y$	0·159	0·681	0·089	0·185	0·471	0·125
	[a]	1009	4518	821	1243	4249	1106
Untranscribed region	$X\rightarrow Y$	0·116	0·506	0·085	0·166	0·563	0·129
	$X\leftarrow Y$	0·124	0·537	0·084	0·165	0·521	0·120
	[a]	257	1090	174	355	1124	270

[a] Number of polymorphic sites.

previous estimate of approximately $2{\cdot}5\times10^{-8}$. In this study, the estimates of population mutation parameters were considered as relative mutation rates, instead of assuming an effective population size. Therefore, rates described in this study refer to relative mutation rates.

(ii) *Mutation-drift equilibrium*

There are six possible combinations of two of the four nucleotides; therefore, there are 12 mutation rates of one base for another. Assuming that the effective population size is constant and that substitutions are in equilibrium, the following equalities should be observed:

- A→C = T→G Transversion
- A→G = T→C Transition
- C→A = G→T Transversion
- C→T = G→A Transition
- A→T = T→A Transversion
- C→G = G→C Transversion

As shown in Tables 1 and 2, estimates for the AD and ED samples correspond relatively well to the expected equalities of mutation rates between certain bases corresponding to the strand reversal symmetry. Table 3 shows the excellent overall correlation between the observed rates and expected equalities. Poor correlation between the observed rates and expected equalities was noted for 5′-UTR regions, which contained the least amount of data of all the regions analysed.

(iii) *Population differences*

Increased population mutation parameters were observed in the AD samples compared with those in the ED samples. Tables 1 and 2 reflect the presence of more polymorphic sites in the AD samples than in the ED samples, which is suspected as a main reason for the increased population mutation parameters in AD samples. The parameters in the AD and ED samples showed very tight correlations (Table 3), indicating that there is a very similar pattern of base substitution in each sample set. These results suggest that the same mechanisms affecting genomic composition may underlie both populations.

Differences in the magnitude of rates between the AD and ED samples may exist as a result of differential effective population sizes between AD and ED samples. Using the effective population sizes estimated from linkage disequilibrium (Tenesa *et al.*, 2007), the average mutation rates were obtained as $2{\cdot}70\times10^{-8}$ in the AD samples and $4{\cdot}94\times10^{-8}$ in the ED samples. For the calculation, the estimate of 7500 for the population of Yoruba in Ibadan, Nigeria, was

Table 2. *Estimates of* θ (*4*Nν *and 4*Nμ) *for the ED samples, which are scaled up by* 10^3

Population	Direction	AC	AG	AT	CG	CT	GT
Total	$X \rightarrow Y$	0·090	0·401	0·069	0·134	0·454	0·110
	$X \leftarrow Y$	0·108	0·470	0·065	0·131	0·362	0·089
	[a]	859	4047	689	1121	3815	926
mRNA	$X \rightarrow Y$	0·067	0·346	0·029	0·086	0·338	0·054
	$X \leftarrow Y$	0·064	0·344	0·030	0·089	0·363	0·057
	[a]	100	507	59	140	520	87
CDS	$X \rightarrow Y$	0·058	0·324	0·015	0·080	0·297	0·045
	$X \leftarrow Y$	0·052	0·299	0·016	0·083	0·366	0·054
	[a]	55	294	22	84	323	47
5′-UTR	$X \rightarrow Y$	0·086	0·411	0·008	0·033	0·462	0·020
	$X \leftarrow Y$	0·057	0·290	0·009	0·035	0·780	0·031
	[a]	9	29	3	10	42	7
3′-UTR	$X \rightarrow Y$	0·071	0·367	0·046	0·102	0·394	0·073
	$X \leftarrow Y$	0·086	0·450	0·043	0·104	0·305	0·055
	[a]	36	184	34	46	155	33
UTRs	$X \rightarrow Y$	0·077	0·376	0·046	0·093	0·411	0·067
	$X \leftarrow Y$	0·085	0·425	0·044	0·096	0·355	0·056
	[a]	45	213	37	56	197	40
Intron	$X \rightarrow Y$	0·095	0·414	0·077	0·146	0·495	0·127
	$X \leftarrow Y$	0·123	0·517	0·072	0·140	0·354	0·094
	[a]	603	2870	509	770	2604	675
Untranscribed region	$X \rightarrow Y$	0·088	0·392	0·064	0·134	0·423	0·096
	$X \leftarrow Y$	0·094	0·416	0·063	0·133	0·391	0·089
	[a]	156	670	121	211	691	164

[a] Number of polymorphic sites.

Table 3. *Test of mutation-drift equilibrium and the correlation between the rates in the AD and ED samples; Spearman rank correlation rho and Kendall's rank correlation tau are indicated with their* P*-values*

Target		All[a]	Total	mRNA	CDS	5′-UTR	3′-UTR	UTR	Intron	Untranscribed
AD	rho	0·952	1	1	1	0·771	1	0·943	1	1
	P-value	$<10^{-3}$	0·003	0·003	0·003	0·103	0·003	0·017	0·003	0·003
	tau	0·858	1	1	1	0·6	1	0·867	1	1
	P-value	$<10^{-3}$	0·003	0·003	0·003	0·136	0·003	0·017	0·003	0·003
ED	rho	0·950	1	0·943	1	0·714	1	0·943	1	1
	P-value	$<10^{-3}$	0·003	0·017	0·003	0·136	0·003	0·017	0·003	0·003
	tau	0·840	1	0·867	1	0·467	1	0·867	1	1
	P-value	$<10^{-3}$	0·003	0·017	0·003	0·272	0·003	0·017	0·003	0·003
AD+ED[b]	rho	0·967	1	0·993	1	0·846	0·986	0·965	0·979	1
	P-value	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	0·001	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
	tau	0·877	1	0·970	1	0·667	0·939	0·879	0·909	1
	P-value	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	0·002	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
AD versus ED	rho	0·969	0·993	0·993	0·979	0·993	0·937	0·944	0·993	0·993
	P-value	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
	tau	0·874	0·970	0·970	0·909	0·970	0·848	0·848	0·970	0·970
	P-value	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$

[a] Combined estimates of total, mRNA, CDS, 5′-UTR, 3′-UTR, UTRs, intron and untranscribed regions.
[b] Test for estimates of both AD and ED samples.

used for the AD samples as the effective population size; the estimate of 3100 for the population from Utah with European ancestry was used for the ED samples. Compared with the estimates with the same effective population size of 10 000 ($2·02 \times 10^{-8}$ in the AD samples and $1·53 \times 10^{-8}$ in the ED samples), the gap between the mutation rates of AD and ED samples became larger after using the estimated effective population sizes (Tenesa *et al*., 2007). The larger gap between samples was observed again for the estimates using the number of polymorphic sites. This phenomenon may result because the method for estimating effective population size (Tenesa *et al*., 2007) did not reflect the past population histories.

To test the effect of past population history directly, simulations were conducted using the most probable parameters of past population history for each sample set as described in the methods section (Boyko *et al*., 2008). The simulations showed that the number of polymorphic sites for the AD population is 1·75-fold higher than the number of polymorphic sites for the ED population. The same ratio was calculated when the complex bottleneck model was applied instead of the simple bottleneck model for the ED population (Boyko *et al*., 2008). The simulated ratio was very similar to the real ratio (1·61) found in the SeattleSNPs PGA data. Results from the simulations support that the higher population mutation parameters in the AD samples might come from differences in the past population histories of the AD and ED samples, which influence the effective population sizes of populations.

(iv) *Regional differences and transition versus transversion*

As shown in Tables 1 and 2, the population mutation parameters were estimated in several specific gene regions: mRNA, CDS, 5′-UTR, 3′-UTR, all UTRs, intron and untranscribed region. The relative mutation rates were different depending on the region. The mRNA regions showed lower mutation rates than intronic and untranscribed regions, primarily attributable to the lower rates in the CDS. In particular, introns showed higher mutation rates than overall sequences and even untranscribed regions. These lower rates in the CDS may result from the negative selection pressures in genes as indicated in a recent study (Schmidt *et al*., 2008). One interesting observation is that T-to-C transitions were more frequent than C-to-T transitions in the 5′-UTR and CDS regions. In 5′-UTRs, the increased rate of T-to-C transitions may indicate the importance of CpG maintenance in this region. However, it is not clear whether the difference in transition rates is due to selective pressure or variation in overall mutation rates in the region.

Although overall estimates show mutation-drift equilibrium with good strand reversal symmetry, the transcription-induced mutation biases were observed mainly in intronic regions and weakly in 3′-UTR regions, showing increased A-to-G over T-to-C mutation rates compared with G-to-A over C-to-T. Additionally, as observed in a previous study (Green *et al*., 2003; Polak & Arndt, 2008), the increased C-to-T over G-to-A is observed in the 5′-UTR regions. However, contrary to the previous results (Polak & Arndt, 2008), the global asymmetry in transcripts of exceeding A-to-G rates over T-to-C rates was not observed in the 5′-UTR region, which showed even opposite results. The asymmetry in the 5′-UTR was higher in ED samples than AD samples. Overall, these effects of mutation biases due to transcription were not strong in this study, and were primarily observed in intronic regions, possibly because the target genes were selected to examine the inflammatory pathway so that they may be less regularly transcribed.

The mutation biases can induce an excess of G+T over A+C on the coding strand (Green *et al*., 2003). As expected, intronic regions showed 51·4% GT content, which is slightly increased compared with 50·9% GT content in the total sequence. This 51·4% GT content in introns is the highest among all investigated regions. However, in this study, taking the sequence compositions into account, the average G+T-to-A+C mutation rate is equal to the reverse rate, as the population mutation parameters were 0·000355 for AD samples and 0·00027 for ED samples, if all possible relevant exchanges were considered (A-to-G, G-to-A, C-to-T, T-to-C, A-to-T, T-to-A, C-to-G and G-to-C). Not only A-to-G, T-to-C and their reverse mutations but also A-to-T, C-to-G and their reverse mutations can eventually contribute to the excess of G+T over A+C in sequence composition. Consideration of sequence compositions and inclusion of these rates counterbalance the increased A-to-G over T-to-C mutation rates, with an expectation of the equilibrium of the sequence composition in intronic regions.

Not surprisingly, there were more transition polymorphisms than transversion polymorphisms. The number of transitions in the AD samples was 2·16-fold higher than the number of transversions. This ratio was 2·19 in the ED samples. The average population mutation parameters for transversions, adjusting for the base composition of the sequence, were 0·000255 for the AD samples and 0·000195 for the ED samples. The parameters for transitions were 0·000555 for the AD samples and 0·000418 for the ED samples (a 2·18- and 2·14-fold increase in rate over transversions, respectively). For transitions, C-to-T and G-to-A mutations were slightly more frequent than the reverse (Tables 1 and 2), presumably due to CpG methylation. The transition to transversion ratios also differed slightly depending on the region (Table 4). Interestingly, much higher ratios were observed in the CDS and the 5′-UTR. However, the increase in transitions versus transversions was not due to a higher level of CpG-to-TpG transitions (or CpG-to-CpA), but rather to an increase in other transition substitutions, especially T-to-C (Supplementary Tables 2 and 3).

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press

Table 4. *Average transition to transversion ratios and average A/T-to-G/C to G/C-to-A/T ratios for given target regions (Ts, average rate of transitions; Tv, average rate of transversions). The estimates of average θ for transition, transversion, A/T-to-G/C and G/C-to-A/T are scaled up by 10*^{3}

Target regions	Samples	Ts	Tv	Ts/Tv	AT→GC	GC→AT	AT→GC/ GC→AT
Total	AD	0·55	0·25	2·18	0·343	0·343	1·001
	ED	0·42	0·19	2·14	0·258	0·258	1·001
mRNA	AD	0·48	0·17	2·86	0·286	0·286	1·001
	ED	0·35	0·12	2·91	0·204	0·204	1·001
CDS	AD	0·44	0·13	3·30	0·257	0·257	1·001
	ED	0·32	0·10	3·11	0·186	0·185	1·001
5′-UTR	AD	0·66	0·18	3·72	0·379	0·379	1·001
	ED	0·46	0·07	6·45	0·254	0·254	1·000
3′-UTR	AD	0·52	0·21	2·41	0·318	0·318	1·001
	ED	0·37	0·14	2·65	0·222	0·222	1·001
UTR	AD	0·54	0·22	2·50	0·331	0·330	1·001
	ED	0·39	0·14	2·81	0·230	0·230	1·001
Intron	AD	0·58	0·28	2·10	0·359	0·359	1·001
	ED	0·44	0·21	2·06	0·272	0·272	1·001
Untranscribed	AD	0·53	0·25	2·16	0·327	0·326	1·001
	ED	0·40	0·19	2·14	0·248	0·248	1·001

(v) *Genomic composition equilibrium*

To examine the nature of changes in genomic composition, the average rates of G/C-to-A/T and A/T-to-G/C mutations were estimated with consideration of current base composition. Without adjusting the sequence composition, the raw estimates showed that there were higher mutation rates of G/C-to-A/T than A/T-to-G/C. After considering the current sequence composition of each base, the average rates of G/C-to-A/T mutations were almost the same as A/T-to-G/C, which were 0·000343 for the AD samples and 0·000258 for the ED samples. These estimates indicate that there is the same number of changes between A/T and G/C bases. This result supports the notion that current sequence compositions due to the nucleotide substitutions between A/T and G/C are in equilibrium at these gene loci. As shown in Table 4, the average rate ratios of A/T-to-G/C versus G/C-to-A/T substitutions indicate that all the regions are in equilibrium.

(vi) *Adjacent nucleotide effect*

To examine the neighbouring nucleotide effects of mutation, rates dependent on the previous and subsequent nucleotide in the sequence were estimated. As shown in Tables 5 and 6, different mutation rates were observed depending on the adjacent sequence. Again, the correlation between estimates for the AD and ED samples was highly significant, indicating a similar mechanism of mutation in both populations. In principle, due to equilibrium in the mutation rates between paired base substitutions, estimates of the effect of the previous nucleotide should match the estimates of the effect of the next nucleotide, i.e. CpG-to-TpG = CpG-to-CpA. As shown in Table 7, overall association tests between the estimates considering the previous and the reversely corresponding next nucleotides were more significant than those between the estimates not taking the adjacent sequence into account (Table 3). This confirms that the adjacent sequence influences the mutation rates.

As expected, the highest rates were for CpG-to-TpG and CpG-to-CpA, indicating that CpG methylation is an important factor in the generation of specific mutations. To examine regional differences in these rates, the transition rates were estimated again after separating the data into sub-sequences (Supplementary Tables 2 and 3). Because transitions were more common than transversions and provided sufficient data for generating estimates, they were chosen for analysis. CpG methylation effects were reduced in mRNA regions and enhanced in intron regions. Like the estimates in Tables 1 and 2, the effects were lowest in the 5′-UTR and next lowest in the CDS. However, in this analysis, the rates of CpG-to-TpG rates were still slightly higher than those for TpG-to-CpG, although increased rates of TpG-to-CpG could certainly be seen in these regions compared with other regions. The increase in T-to-C transition rates as compared with the reverse substitution for 5′-UTR and CDS (Tables 1 and 2) comes from the overall effects of the adjacent sequence (Supplementary Tables 2 and 3). Combined with the results in Table 4, it may be suspected that a

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press

Table 5. *Estimates of* θ ($4N\nu$ *and* $4N\mu$) *from the distribution of allele frequencies in the AD samples based on either the next or previous nucleotide. The estimates of* θ *are scaled up by* 10^3

		Next nucleotide effect						Previous nucleotide effect					
[a]	Direction	AC	AG	AT	CG	CT	GT	AC	AG	AT	CG	CT	GT
A	$X \rightarrow Y$	0·110	0·329	0·065	0·146	0·407	0·110	0·096	0·396	0·081	0·207	0·882	0·112
	$X \leftarrow Y$	0·127	0·437	0·097	0·167	0·521	0·123	0·164	0·443	0·102	0·136	0·648	0·124
	[b]	417	1203	260	461	1301	294	355	1456	304	460	1944	356
C	$X \rightarrow Y$	0·172	0·593	0·091	0·133	0·399	0·142	0·135	0·777	0·080	0·123	0·561	0·468
	$X \leftarrow Y$	0·127	0·570	0·071	0·173	0·422	0·116	0·146	3·900	0·076	0·581	0·488	0·086
	[b]	394	1290	208	400	1174	334	420	2421	263	381	1690	284
G	$X \rightarrow Y$	0·090	0·511	0·075	0·575	3·581	0·137	0·118	0·417	0·089	0·199	0·573	0·131
	$X \leftarrow Y$	0·476	0·566	0·072	0·120	0·661	0·119	0·145	0·390	0·105	0·151	0·554	0·167
	[b]	264	1726	249	369	2266	403	343	1142	235	441	1271	412
T	$X \rightarrow Y$	0·119	0·765	0·120	0·162	0·428	0·216	0·127	0·552	0·099	0·180	0·471	0·143
	$X \leftarrow Y$	0·105	0·968	0·088	0·234	0·358	0·125	0·114	0·400	0·061	0·145	0·324	0·122
	[b]	355	2196	349	576	1439	491	313	1401	266	522	1265	473

[a] Adjacent sequence.
[b] Number of polymorphic sites.

Table 6. *Estimates of* θ ($4N\nu$ *and* $4N\mu$) *from the distribution of allele frequencies in the ED samples based on either the next or previous nucleotide. The estimates of* θ *are scaled up by* 10^3

		Next nucleotide effect						Previous nucleotide effect					
[a]	Direction	AC	AG	AT	CG	CT	GT	AC	AG	AT	CG	CT	GT
A	$X \rightarrow Y$	0·084	0·256	0·048	0·112	0·310	0·084	0·076	0·303	0·068	0·166	0·666	0·083
	$X \leftarrow Y$	0·097	0·340	0·071	0·128	0·397	0·094	0·129	0·340	0·085	0·109	0·489	0·092
	[b]	239	795	148	302	785	194	217	926	198	287	1193	212
C	$X \rightarrow Y$	0·131	0·438	0·075	0·100	0·294	0·113	0·106	0·590	0·059	0·093	0·424	0·336
	$X \leftarrow Y$	0·098	0·422	0·059	0·130	0·310	0·092	0·115	2·988	0·056	0·439	0·368	0·062
	[b]	230	787	134	246	749	196	267	1507	167	236	1021	179
G	$X \rightarrow Y$	0·065	0·387	0·059	0·471	2·705	0·094	0·083	0·313	0·076	0·147	0·431	0·103
	$X \leftarrow Y$	0·347	0·429	0·057	0·098	0·495	0·081	0·102	0·292	0·090	0·112	0·416	0·131
	[b]	170	1057	175	239	1410	249	185	758	164	278	818	245
T	$X \rightarrow Y$	0·097	0·575	0·098	0·122	0·320	0·157	0·100	0·409	0·073	0·142	0·345	0·101
	$X \leftarrow Y$	0·085	0·728	0·072	0·176	0·267	0·091	0·089	0·296	0·045	0·114	0·237	0·086
	[b]	220	1396	228	331	866	284	188	849	159	317	778	289

[a] Adjacent sequence.
[b] Number of polymorphic sites.

CpG maintenance system may exist in these two regions.

4. Discussion

This study is the first report of the relative mutation rates of one base for another by applying the recurrent reversible mutation model in equilibrium to resequencing data. The overall estimated population mutation parameters found were similar to previous studies using coalescent- and phylogeny-based approaches (Nachman & Crowell, 2000; Bhangale *et al.*, 2005; Carlson *et al.*, 2005). This study provides several important insights into mutation in the human genome. First, mutation-drift equilibrium was observed in human gene loci from the relative mutation rates. Second, the sequence composition of the human genome is in equilibrium using this method, at least in these gene loci. Third, mutation rates differ depending on the region of a gene locus in question. In particular, the lower rate in mRNA is observed primarily due to CDS regions, which may result from negative selection pressure, as described in a recent report (Schmidt *et al.*, 2008). This study also confirms several expectations with more detailed information: higher rates of transition substitutions, mutations due to CpG methylation, adjacent sequence effects and transcription-induced mutation biases.

Table 7. *Spearman rank correlation rho and Kendall's rank correlation tau of effects from the next and previous nucleotides; all tests show strong significance with* P-*values of less than* $1{\cdot}0\times10^{-5}$ *by each test*

Target	Test method	Total	mRNA	Intron	Untranscribed region	All transitions[a]	Transitions in total sequence[b]
AD	Spearman rho	0·982	0·921	0·977	0·936	0·868	0·979
	Kendall's tau	0·911	0·806	0·888	0·805	0·715	0·920
ED	Spearman rho	0·971	0·895	0·964	0·919	0·830	0·980
	Kendall's tau	0·880	0·770	0·861	0·776	0·694	0·915
AD + ED[c]	Spearman rho	0·982	0·918	0·976	0·945	0·864	0·977
	Kendall's tau	0·902	0·795	0·879	0·814	0·721	0·896

[a] Combined transition rates of total, mRNA, 5′-UTR, 3′-UTR, UTRs, intron and untranscribed regions.
[b] Transition rates in total sequence only.
[c] Test for estimates of both AD and ED samples.

The frequent C-to-T substitutions occurring due to CpG methylation are well known and have been observed in mutation databases (Krawczak *et al*., 1998; Olivier *et al*., 2002; Petitjean *et al*., 2007). This phenomenon is also common in other organisms (Morton *et al*., 2006). Phylogeny-based estimates with the nearest-neighbour interactions showed that the CG-to-TG substitution rate is more than 10-fold higher than other substitution rates (Hwang & Green, 2004; Lunter & Hein, 2004). However, in this study, the pattern was quite different from previously published results. The differences between CG-to-TG (and CG-to-CA) and other rates were not as large as in the previous estimates. Moreover, the rates of all other transitions except CpG-to-TpG (or CpG-to-CpA) lay in a relatively small range, in contrast to the previous study. Differences from the previous results may be a consequence of analysing human gene loci, in which negative selections are underway with observations of lower substitution rates at non-synonymous CpG sites (Schmidt *et al*., 2008). However, the possibility that differences in methodology may have led to these discrepancies cannot be ruled out.

The investigated sequence had a GC content of 45·2%, which is higher than the genome-wide average of 41% (Strachan & Read, 2004). The number of CpGs analysed was 60 035, which was only 27% of the expected amounts, taking the sequence compositions into account. However, this is a bit higher than the regular expectation of the human genome, which is one-fifth of the expected value given the overall contents of C and G (Strachan & Read, 2004). This discrepancy may result from the examination of gene loci in this study. Previous studies have indicated that the transition bias is due mainly to CpGs (Rosenberg *et al*., 2003; Keller *et al*., 2007). However, considering the small number of CpGs in the human genome, it is unlikely that CpG methylation completely explains the transition bias. The estimates in this study show that other transition rates are also much higher than transversion rates (Tables 5 and 6).

Previous measurements relying primarily on phylogeny-based methods suggested higher G/C to A/T substitution rates (Hwang & Green, 2004; Lunter & Hein, 2004; Lipatov *et al*., 2006; Duret & Arndt, 2008). However, several studies using polymorphism data found evidence for a fixation bias in favour of mutations that increase GC content (Eyre-Walker, 1999; Webster *et al*., 2003). A more recent study suggested that these observations could be caused by misidentification of the ancestral state of partial data (Hernandez *et al*., 2007). However, their results did not show the reverse observation, that is, a higher rate of G/C to A/T mutations. The estimates in the current study indicate that there is no mutational preference for G/C or A/T and that the overall sequence is in base compositional equilibrium (Table 4). Since the estimates in this study also indicate higher rates of G/C to A/T mutations than the reverse before adjusting for the current sequence composition, strict adjustment for the sequence composition may somewhat alter the previous estimates based on phylogeny. However, unfortunately, it is still true that the current results cannot explain the observation of higher rates of G/C to A/T substitutions than the reverse in previous studies using comparative methods.

Increased transition to transversion ratios in the 5′-UTR and CDS were observed in this study (Table 4). As shown in Supplementary Tables 2 and 3, these phenomena appear to result from increased T-to-C transition rates rather than increased CpG-to-TpG transition rates. Decreased CpG-to-TpG transition rates were observed in 5′-UTR and CDS regions, indicating that these rates did not contribute to the increased transition to transversion ratios in the 5′-UTR and CDS. Since the reversible recurrent mutation model assumes mutation-drift equilibrium and the estimates are dependent on the sequence composition, it cannot be ruled out that these observations might be a consequence of methodology. However, it should also be noted that these estimates may provide a reasonable explanation for the evolution of

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press

sequence composition and the maintenance of CpG islands in 5′-UTR and CDS regions.

The estimated population mutation parameters for the ED samples are lower than the estimates for the AD samples, which may arise from the different effective population sizes between samples. As shown in the results section, application of the effective population sizes estimated from linkage disequilibrium did increase the discrepancy between the estimates of AD and ED samples (Tenesa *et al.*, 2007). The increased discrepancy may arise because the estimation method of effective population sizes lacks the consideration of past population history. Previous studies found a demographic history of expansion in AD populations and a bottleneck in ED populations (Akey *et al.*, 2004; Marth *et al.*, 2004; Boyko *et al.*, 2008). Using the same parameters suggested in a previous study (Boyko *et al.*, 2008), simulation results provided evidence that the increase in polymorphic sites in the AD samples were due to the differences between the past demographic histories of AD and ED populations, which may influence effective population sizes. However, these simulations can only superficially explain the increased estimates calculated for the AD samples. Moreover, considering the overall effect of past population histories, higher asymmetry of transition rates of the 5′-UTR in ED samples than AD samples may come from other effects, such as selection pressures, rather than the effect of past population history. For a complete understanding of population differences, more studies would be necessary.

There are limitations to the interpretation of the findings of this study. Direct estimation from the incidence of dominant Mendelian diseases in humans indicates differential mutation rates across loci (Drake *et al.*, 1998; Reich & Lander, 2001; Kondrashov, 2002). The gene lengths in this study are varied; some of the genes have a very small number of polymorphic sites due to their short lengths. To address these concerns, further study of gene clusters containing a sufficient number of polymorphic sites may be useful. It should also be noted that mutation rates are influenced by gender and age (Crow, 2000, 2006). In addition, attention must be paid to population demography and selective pressure. Although they are not expected to influence comparison of the relative rates significantly, these other factors may nevertheless alter the mutation rate estimates. Moreover, all the regions examined in this study are gene loci. Mutation rates may be different for regions that do not harbour genes.

Using high-quality, large-scale, public re-sequencing data, this study explored the relative mutation rates of each base for each of the other bases. This study is the first report of population mutation parameters as relative mutation rates by applying the recurrent reversible mutation model to re-sequencing data. By providing more detailed information on the mutations in human gene loci, this study offers reasonable explanations for the evolution of the human genome by mutational events. Nevertheless, further studies using larger data sets and theoretical studies with consideration of population genetic factors would be helpful in providing better understanding for human mutation processes. In addition, as suggested (Duret, 2009), direct estimation using large-scale re-sequenced data of families would be extremely useful.

The author appreciates the reviewers' comments, which substantially improve this manuscript. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-532-C00017). Key calculations were performed using the supercomputing resource at the Korea Institute of Science and Technology Information (KISTI).

References

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A. & Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology* **2**, e286.

Arndt, P. F. & Hwa, T. (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**, 2322–2328.

Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. (2004). The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *Journal of Molecular Evolution* **58**, 653–660.

Bhangale, T. R., Rieder, M. J., Livingston, R. J. & Nickerson, D. A. (2005). Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Human Molecular Genetics* **14**, 59–69.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G. & Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics* **4**, e1000083.

Carlson, C. S., Reider, M. J., Nickerson, D. A., Eberle, M. A. & Kruglyak, L. (2005). Comment on 'Discrepancies in dbSNP confirmations rates and allele frequency distributions from varying genotyping error rates and patterns'. *Bioinformatics* **21**, 141–143.

Catanzaro, D., Pesenti, R. & Milinkovitch, M. C. (2006). A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model. *Bioinformatics* **22**, 708–715.

Clark, R. M., Tavare, S. & Doebley, J. (2005). Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Molecular Biology and Evolution* **22**, 2304–2312.

Crawford, D. C., Akey, D. T. & Nickerson, D. A. (2005). The patterns of natural variation in human genes. *Annual Review of Genomics and Human Genetics* **6**, 287–312.

Crow, J. F. (2000). The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics* **1**, 40–47.

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press

Crow, J. F. (2006). Age and sex effects on human mutation rates: an old problem with new complexities. *Journal of Radiation Research* (*Tokyo*) **47** (Suppl. B), B75–B82.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* **148**, 1667–1686.

Duret, L. (2009). Mutation patterns in the human genome: more variable than expected. *PLoS Biology* **7**, e1000028.

Duret, L. & Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics* **4**, e1000071.

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

Eyre-Walker, A. (1999). Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152**, 675–683.

Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics* **33**, 514–517.

Hartl, D. L. & Clark, A. G. (2007). *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates, Inc.

Hernandez, R. D., Williamson, S. H., Zhu, L. & Bustamante, C. D. (2007). Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Molecular Biology and Evolution* **24**, 2196–2202.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Hwang, D. G. & Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the USA* **101**, 13994–14001.

Jiang, Z., Wu, X. L., Zhang, M., Michal, J. J. & Wright, R. W. Jr (2008). The complementary neighborhood patterns and methylation-to-mutation likelihood structures of 15,110 single-nucleotide polymorphisms in the bovine genome. *Genetics* **180**, 639–647.

Jolly, C., Cook, A. J., Raftery, J. & Jones, M. E. (2007). Measuring bidirectional mutation. *Journal of Theoretical Biology* **246**, 269–277.

Keller, I., Bensasson, D. & Nichols, R. A. (2007). Transition–transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genetics* **3**, e22.

Kimura, M. & Ohta, T. (1971). *Theoretical Aspects of Population Genetics*. Princeton, NJ: Princeton University Press.

Kondrashov, A. S. (2002). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation* **21**, 12–27.

Krawczak, M., Ball, E. V. & Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *American Journal of Human Genetics* **63**, 474–488.

Lercher, M. J., Smith, N. G., Eyre-Walker, A. & Hurst, L. D. (2002). The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**, 1805–1810.

Lipatov, M., Arndt, P. F., Hwa, T. & Petrov, D. A. (2006). A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. *Journal of Molecular Evolution* **62**, 168–175.

Lunter, G. & Hein, J. (2004). A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* **20** (Suppl. 1), i216–i223.

Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372.

Morton, B. R., Bi, I. V., McMullen, M. D. & Gaut, B. S. (2006). Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **172**, 569–577.

Nachman, M. W. & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304.

Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C. & Hainaut, P. (2002). The IARC TP53 database: new online mutation analysis and recommendations to users. *Human Mutation* **19**, 607–614.

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S. V., Hainaut, P. and Olivier, M. (2007). Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Human Mutation* **28**, 622–629.

Piganeau, G., Mouchiroud, D., Duret, L. & Gautier, C. (2002). Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *Journal of Molecular Evolution* **54**, 129–133.

Polak, P. & Arndt, P. F. (2008). Transcription induces strand specific mutations at the 5′ end of human genes. *Genome Research* **18**, 1216–1223.

Reich, D. E. & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics* **17**, 502–510.

Rosenberg, M. S., Subramanian, S. & Kumar, S. (2003). Patterns of transitional mutation biases within and among mammalian genomes. *Molecular Biology and Evolution* **20**, 988–993.

Schmidt, S., Gerasimova, A., Kondrashov, F. A., Adzhubei, I. A., Kondrashov, A. S. & Sunyaev, S. (2008). Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genetics* **4**, e1000281.

SeattleSNPs. (2006). NHLBI Program for Genomic Applications. Seattle, WA: SeattleSNPs. Available from http://pga.gs.washington.edu (Accessed 3 June 2008).

Strachan, T. & Read, A. P. (2004). *Human Molecular Genetics 3*. New York: Garland Publisher.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. and Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**, 520–526.

Webster, M. T., Smith, N. G. & Ellegren, H. (2003). Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Molecular Biology and Evolution* **20**, 278–286.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

https://doi.org/10.1017/S0016672309990164 Published online by Cambridge University Press