Rate, molecular spectrum, and consequences of human mutation

Michael Lynch[1]

Department of Biology, Indiana University, Bloomington, IN 47405

Although mutation provides the fuel for phenotypic evolution, it also imposes a substantial burden on fitness through the production of predominantly deleterious alleles, a matter of concern from a human-health perspective. Here, recently established databases on de novo mutations for monogenic disorders are used to estimate the rate and molecular spectrum of spontaneously arising mutations and to derive a number of inferences with respect to eukaryotic genome evolution. Although the human per-generation mutation rate is exceptionally high, on a per-cell division basis, the human germline mutation rate is lower than that recorded for any other species. Comparison with data from other species demonstrates a universal mutational bias toward A/T composition, and leads to the hypothesis that genome-wide nucleotide composition generally evolves to the point at which the power of selection in favor of G/C is approximately balanced by the power of random genetic drift, such that variation in equilibrium genome-wide nucleotide composition is largely defined by variation in mutation biases. Quantification of the hazards associated with introns reveals that mutations at key splice-site residues are a major source of human mortality. Finally, a consideration of the long-term consequences of current human behavior for deleterious-mutation accumulation leads to the conclusion that a substantial reduction in human fitness can be expected over the next few centuries in industrialized societies unless novel means of genetic intervention are developed.

base substitutions | human genetic disorders | introns | mutation rate | mutational spectrum

Despite its central significance to matters of health and phenotypic evolution, many uncertainties still remain about the rate and spectrum of mutations spontaneously arising in the human genome (1–3). How frequently do germline and somatic mutations arise, and to what extent does this vary between the sexes? What is the relative incidence of various forms of mutations, e.g., missense and nonsense base substitutions, insertions, duplications, and deletions, especially among alterations having major phenotypic effects? How does the mutational spectrum in humans compare with that in other species? And most importantly, what are the consequences of mutation for the long-term genetic well-being of our species?

In the near future, it should be possible to provide refined answers to these and many more questions by sequencing the complete genomes of well defined pedigrees and somatic tissues (4, 5). However, it is already possible to achieve a relatively complete picture of the point-mutation process from databases on mutations at loci known to underlie monogenic disorders with major phenotypic effects. In the case of autosomal-dominant and X-linked disorders, affected individuals can generally be identified as de novo mutants by comparison with the parental phenotypes, thereby providing a nearly unbiased view of the rate and spectrum of locus-specific mutations, similar to what has been achieved with reporter construct studies in microbes (6). Estimation of the human mutation rate from the incidence of monogenic disorders has a long history, dating back to Haldane (7), but Kondrashov (8) pioneered the use of allelic-sequence survey data to gain a broad perspective on the molecular mechanisms of mutation across multiple loci. Since the study of Kondrashov (8), there has been a substantial influx of new data on more alleles and more disorders, allowing for a refinement of prior estimates, and the very recent emergence of substantial data from other species provides a basis for comparative analysis.

The material covered herein touches upon a number of issues central to our understanding of human genetics and evolution: (*i*) The analyses provide a revised estimate of the human mutation rate per nucleotide site. (*ii*) Comparison with data for a diversity of species leads to a hypothesis regarding genome nucleotide composition that appears to be general across much of cellular life. (*iii*) An evaluation of the mutational cost of introns demonstrates that the vulnerability of human genes to degenerative mutations is a function not only of a high mutation rate per nucleotide site but also of aspects of gene structure. (*iv*) Taking into consideration the rates of mutation at both the germline and somatic-cell levels and their likely effects, the consequences of mutation for long-term human genetic well-being are explored, and some significant prior concerns are given credence.

Results

Base-Substitution Mutation Rate. Although an earlier attempt to estimate the human mutation rate using disease-gene data focused only on nonsense mutations (8), more accurate estimates of both the mutational spectrum and per-site rates may be obtained by including missense mutations, as done in the present study. For example, one limitation of a focus on nonsense mutations is that no mutations to nucleotide C can be detected on the coding strand, as the three termination codons (TAA, TAG, and TGA) are devoid of C. In addition, because termination codons are A+T rich, and there is a substantial mutational bias in the direction of A+T (where "+" implies total composition across both strands) production (9), a focus on these three codons may yield an overestimate of the overall mutation rate. Finally, the inclusion of missense mutations substantially increases sample sizes.

For the genes involved in this study, the average rates of base-substitutional mutation are 11.63 (1.80) and 11.22 (3.23) $\times 10^{-9}$ per site per generation for autosomal and X-linked loci (SDs in parentheses), respectively (Dataset S1). Modifying the latter estimate to scale to a 1:1 incidence of exposure across the sexes yields an autosomal equivalent estimate of 14.81 (4.26) $\times 10^{-9}$, which is not significantly different from the direct autosomal estimate. An average pooled estimate for genes that spend equal time in males and females is then 12.85 (1.95) $\times 10^{-9}$ per site per generation. Many sources of error contribute to the locus-specific estimates in Dataset S1, but these will all be subsumed into the SE of the overall mean estimate.

Universal Mutation Pressure in Direction of A+T. As in all well characterized species except *Caenorhabditis elegans* (10), the transition-to-transversion ratio for de novo mutations in humans is significantly greater than the value of 0.5 expected under a random mutation model (in which every nucleotide has an equal probability of mutating to each other state, one of which is always a transition; Table 1) (9–13). This is primarily a consequence of a high incidence of G:C → A:T transitions (where the colon denotes a Watson–Crick bond between strands). The raw human mutation spectrum is also consistent with observations in all other well characterized eukaryotic species in exhibiting a preponderance of mutations in the direction of A+T versus G+C. This disparity is contrary to the expectation for a genome in mutation equilibrium, which would necessarily exhibit equal numbers of mutations in both directions.

From the conditional mutation rates, i.e., the mutation rates weighted by the incidence of the starting base, it is possible to estimate the equilibrium A+T composition expected under mutation pressure alone (ref. 9, p. 130), and in all species with a well defined mutational spectrum this exceeds the actual A+T composition, even at silent sites (Table 2). As there is no evidence that all genomes are evolving toward new nucleotide-composition equilibria, the only explanation for this pattern is that directional mutational pressure toward A+T is countered by some form of selection in favor of C+G. Following the approach of Bulmer (14), under the assumption of drift-mutation-selection balance, it can be shown that the deviation of the observed base composition from the mutational expectation is entirely a function of $4N_es$ ($2N_es$ for haploids). This composite parameter is equivalent to twice the ratio of the power of selection (s) to the power of drift $1/(2N_e)$, where N_e is the effective population size, and s is the average selective advantage of C+G nucleotides.

The resultant estimates of $4N_es$ fall in the narrow range of 0.35 to 1.61 across all species, implying that the average magnitude of selection operating on base composition at the nucleotide level is of the same order of magnitude as the power of drift in a wide variety of species. A similar level of constancy of $4N_es$ across microbes and multicellular eukaryotes has been noted previously with respect to nucleotide composition in the third positions of codons (9). Remarkably, using phylogenetic data, Kondrashov et al. (15) obtained an estimate of $4N_es$ of 1.00 for codon usage in the great ape lineage, which is virtually identical to the estimate provided here (0.99).

As there is little question that effective population sizes decline by several orders of magnitude from microbes to multicellular species (16), the approximate constancy of $4N_es$ for nucleotide usage across disparate taxa requires a substantial increase in s operating on nucleotide composition in multicellular lineages. Although translation-associated selection (e.g., codon bias) is likely to be a factor in the evolution of nucleotide usage in coding DNA, biased gene conversion in the direction of

G+C composition appears to be an equally, if not more, powerful force that applies to all genomic regions in sexually reproducing species (9). However, because the power of gene conversion appears to be considerably greater in yeast than in animals (which have substantially lower rates of recombination per physical distance along chromosomes), this factor alone appears to be incapable of explaining the pattern noted previously. Still another factor favoring G/C composition is the enhanced stability of G:C relative to A:T Watson–Crick bonds.

Regardless of the causal mechanism(s), the existing data clearly indicate that the absolute intensity of selection favoring G+C composition is substantially magnified in species with reduced effective population sizes. This pattern is expected if the disadvantage of suboptimal nucleotides cumulatively increases as the genome-wide nucleotide composition deviates further from the selectively optimal state [a form of synergistic epistasis (17)]. As selection does not become effective until the selection coefficient of a mutation (s) approaches the power of drift ($1/2N_e$), under this hypothesis, one would expect the nucleotide composition to be driven from the selective optimum by mutation until s is approximately equal to $1/(4N_e)$, or equivalently $4N_es$ is ~1.00. Once this critical value of s is reached, mutation pressure would be incapable of pushing nucleotide composition to a more extreme value, with the equilibrium being defined by the ratio of mutation pressures and a scaled selection parameter near $4N_es$ of approximately 1.00. Consistent with this view, the results in Table 2 imply average values of $4N_es$ across species equal to 1.02 (0.18), 1.04 (0.08), and 0.78 (0.18) for coding DNA, silent sites, and total genomic DNA respectively, none of which are significantly different from 1.0.

If this interpretation is correct, then genomes that are in mutation-selection-drift equilibrium will have an A+T composition defined by the bias in mutation pressure alone, the approximate expectation being:

$$p_{A+T} = m/(m + e^{-1}), \quad \textbf{[1]}$$

from equation 6.2 in Lynch (9), where m is the ratio of the summed mutation rates involving G+C → A+T changes to the summed rates involving G+C ← A+T changes. Although this invariant pattern is "universal" only in the context of the current collection of species with well defined mutational features, because the phylogenetic diversity of this group is very substantial, it is likely to have broad applicability.

The predominance of A:T → G:C and G:C → A:T changes among base-substitutional mutations in the human genome has previously been inferred by other methods, some potentially influenced by selection biases (18–20). In primates, C → T transitions arise at CpG dinucleotide sites approximately 15 times the mutation rate observed at other sites (8, 21), ostensibly because of the spontaneous oxidative deamination of methylated cytosines at CpG sites. Consistent with this view, CpG sites in *Arabidopsis* that are specifically methylated do have elevated mutation rates,

Table 1. Summary of the raw base-substitution mutational spectrum in humans, and comparison with the de novo spectrum for other model species

Species	Transitions		Transversions				N	Ts:Tv	AT:GC
	AT > GC	GC > AT	AT > TA	GC > TA	AT > CG	GC > CG			
H. sapiens	0.221	0.408	0.067	0.110	0.078	0.125	3,336	1.697	1.732
D. melanogaster (12)	0.190	0.305	0.121	0.155	0.115	0.115	174	0.977	1.509
C. elegans (10)	0.105	0.205	0.220	0.281	0.113	0.077	391	0.448	2.235
A. thaliana (11)	0.118	0.588	0.059	0.094	0.071	0.071	85	2.400	3.625
S. cerevisiae (13)	0.103	0.286	0.060	0.312	0.057	0.182	1,250	0.638	3.740
E. coli (9)	0.162	0.317	0.195	0.115	0.155	0.056	1,037	0.920	1.362

N denotes the number of observed mutations, and the final 2 columns denote the ratios of transitions (Ts) and transversions (Tv) and of (GC→AT + GC→TA) to (AT→GC + AT→CG) mutations.

Table 2. Conditional mutation rates, equilibrium A/T composition expected under mutation pressure alone, and magnitude of average scaled selective disadvantage of A/T

	Conditional mutation rates						A/T composition				Efficiency of selection		
Species	AT > GC	GC > AT	AT > TA	GC > TA	AT > CG	GC > CG	Coding	Silent	Total	Equilibrium	Coding	Silent	Total
H. sapiens	4.29	9.61	1.29	2.58	1.52	2.95	0.48	0.44	0.55	0.68	0.82	0.99	0.55
D. melanogaster (12)	1.18	2.43	0.75	1.24	0.72	0.92	0.46	0.36	0.56	0.66	0.81	1.24	0.41
C. elegans (10)	1.48	5.26	3.10	7.23	1.59	1.97	0.57	0.65	0.65	0.80	1.14	0.79	0.80
A. thaliana (11)	1.32	11.38	0.66	1.82	0.79	1.37	0.56	0.66	0.63	0.86	1.61	1.19	1.29
S. cerevisiae (13)	0.14	0.62	0.08	0.68	0.08	0.39	0.60	0.65	0.62	0.86	1.38	1.19	1.32
E. coli (9)	0.08	0.16	0.10	0.06	0.08	0.03	0.48	0.37	0.49	0.57	0.35	0.83	0.31

Table shows conditional mutation rates ($\times 10^{-9}$/site/generation); genome-wide A/T composition (silent refers to synonymous sites in codons); equilibrium A/T composition expected under mutation pressure alone; and the magnitude of the average scaled selective disadvantage of A/T bases necessary to account for the observed A/T composition under the assumption of drift-selection-mutation balance. The latter quantity, which is equal to $4N_e s$ in the case of diploids and $2N_e s$ in the case of haploids, is equivalent to twice the ratio of the power of selection to the power of random genetic drift. Total refers to genome-wide composition.

although methylation alone does not completely explain the high rate of G:C → A:T transition in this species, even at CpG sites (11). It is common for the CpGs within human somatic cells to be 20% to 80% methylated (22), although the degree of methylation in germline cells is less clear, and in *Drosophila* and *Caenorhabditis*, which do not have detectable levels of methylation, other mechanisms must be responsible for the elevated rate of G:C → A:T mutation.

Insertions and Deletions. In the human genome, small (1–50-bp) deletions are approximately three times as common as insertions of the same size, with both types of changes exhibiting very similar scaling with the size of the fragment involved in the mutational event (Fig. 1). In both cases, the mutation frequency declines with the 1.82 power of fragment size. Among segregating mutations in the human population, deletions are also 2.3 to 4.1 times more common than insertions (23, 24), consistent with the threefold bias at the mutational level assuming that both types of alterations are equally deleterious.

From the fitted functions in Fig. 1, which ignore $3n$-sized mutations (some of which apparently go undetected because they leave the reading frame intact), the total numbers of expected 1- to 50-bp mutations in the data set (corrected for detectability) are 2,585 and 903 for deletions and insertions. Comparing these numbers with the expected number of base-substitutional mutations (after correcting for the number of undetected mutations of this sort; see *Methods*), the extrapolated estimates of the deletion and insertion rates become 0.58 and 0.20 × 10^{-9} per site per generation. Thus, taken together, insertion and deletion mutations are only approximately 6% as common as those involving single-base substitutions.

Cost of Introns. The mutational cost of introns can be estimated in units of numbers of coding-site equivalents by considering the ratio of targets, R_T (number of introns per gene divided by number of coding nucleotides per gene), and R_M (number of observed mutations to defective alleles resulting from altered splicing divided by number resulting from coding-region alterations; here defined by the total of all base-substitution mutations and all insertion/deletions smaller than 50 bp in length not known to affect splicing). The ratio of these ratios, R_M/R_T, provides an estimate of the average cost of an intron at a locus in units of numbers of coding nucleotides.

The average cost of an intron is approximately 30.8 (2.1) base pair equivalents (Fig. 2). In other words, in terms of the mutational target size to defective alleles, the addition of an intron to a human gene is on average equivalent to adding approximately 31 nucleotides to the coding region. This estimate is almost certainly downwardly biased because some coding-region mutations prob-

ably alter splicing (and go undetected in studies without cDNA sequencing) and some mutations undetected by target-locus sequencing may actually be deep within intron sequence (which, in almost all studies, has not been assayed).

It should also be realized that the relative intron costs presented here are functions of the selective constraints on the coding DNA at a locus. All other things being equal, genes with very strong constraints on coding sequence will yield lower relative measures of intron costs at the locus, despite the fact that the absolute cost of introns (in units of numbers of nucleotides reserved for proper splice-site recognition) is likely to be relatively constant across loci.

A more direct estimate of the mutational cost of introns can be obtained as follows. Conservatively assuming that all mutations affecting splicing are detectable, then after correcting for the detectability of base substitutions, there appear to be approximately 0.036 splicing mutations per base-substitution in an average human gene, which for the loci analyzed here, implies a mutation rate to defective alleles associated with splice-site mutations of 69.6×10^{-9} per intron per generation. Thus, taking into consideration that this estimate is likely to be downwardly biased, with an average of eight introns per protein region, a
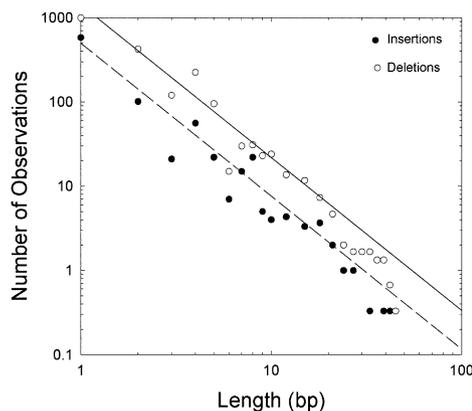

Fig. 1. The size-frequency spectrum of insertion and deletion mutations in human genes, summed over autosomal-dominant and X-linked genes. The scaling of frequency (f) with length of the mutation (L) is $f = 0.56 L^{-1.82}$ ($r^2 = 0.957$) and $f = 0.67 L^{-1.82}$ ($r^2 = 0.973$) for insertions and deletions, respectively. These regressions exclude the plotted data points for mutations with size changes that are multiples of 3, which leave the reading frame intact and in some cases have minimal phenotypic effects (and therefore go undetected), and also only employ mutations in size classes up to $L = 20$, beyond which sample sizes are very small and sporadic. For the latter reason, the data beyond $L = 10$ are also pooled into windows of three base sizes and divided by 3 to retain the appropriate scale.
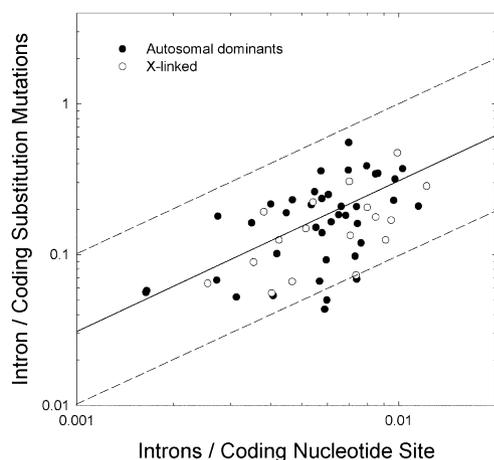
Fig. 2. The cost of introns in human genes in units of the cost of coding nucleotides. The solid diagonal line denotes the average ratio of the values on the vertical and horizontal axes, 30.8. The dashed lines, for reference, denote ratios of 10 (lower) and 100 (upper). Thus, the mutational cost of an intron in a typical human gene is equivalent to adding 30.8 coding nucleotides, and for the vast majority of loci this cost falls within the range of 10 to 100.

typical human gene experiences an elevation in the mutation rate to defective alleles of approximately 10^{-6} per generation that an intron-free allele would otherwise avoid.

Discussion

Despite the current status as the dominant organism on earth, the human species is confronted with substantial mutational challenges imposed by at least three baseline genetic features: (*i*) a relatively high per-generation germline mutation rate at the nucleotide level; (*ii*) a further inflation in the mutational rate of production of defective alleles associated with aspects of gene structure; and (*iii*) a large cumulative burden of somatic mutations imposed by a relatively late onset at maturity. The preceding results will now be used to obtain a quantitative perspective on these three issues, as well as an evaluation of the longer-term consequences of the mutational and environmental landscape for the future genetic well-being of our species.

Elevated Per-Generation Mutation Rate. Although the estimated base-substitutional mutation rate derived in this study, 12.8 (2.0) $\times$ 10^{-9} per site per generation, is approximately 25% lower than an earlier estimate of 17.0 (0.2) $\times 10^{-9}$ derived by Kondrashov (8), it is nevertheless higher than the rate for any other well studied species. A recent pedigree-based estimate derived from high-throughput sequencing of Y chromosomes separated by 13 generations (4) yields a base-substitutional mutation rate estimate of 17.3 (8.6) $\times 10^{-9}$ when scaled across the sexes under the assumption of a 6.5-fold inflation in males (see *Methods*), which is compatible with both previous estimates.

There are several potential explanations for the deviation of the current results from Kondrashov (8). First, the set of genes employed here extends beyond that used by Kondrashov (8), as a number of relevant data sets became available in the intervening period. Second, with new information on the incidences of disorders and the fractions of affected individuals as a result of de novo mutations, the estimated per-locus mutation rates to defective alleles are in some cases substantially different between our two studies. Third, because of data judged to be inadequate, three loci employed by Kondrashov (8) were discarded (ABCD1, AR, and EMD), and when these are excluded, Kondrashov's estimate declines to 16.0 (0.2) $\times 10^{-9}$, which is statistically compatible with the estimate provided here. Fourth, whereas the Kondrashov study (8) was confined to mutations to nonsense codons, the current

investigation integrated over a wider range of contextual sequence space by incorporating both nonsense and missense mutations.

Kondrashov's (8) decision to focus on termination-codon production was based on the fact that nonsense mutations are much more likely to have phenotypic effects than missense (and silent-site) mutations. However, Drake (6) showed that, by using a correction factor for the overall detectability of mutations, reasonable gene-wide estimates of the mutation rate can be obtained, and a similar approach was adhered to in this study. One reason to think that the base-substitutional mutation rate reported herein is reasonable concerns the estimated rates for insertion/deletion mutations. The latter were obtained by simply scaling the former by the ratio of the expected number of insertion/deletions to the expected number of base-substitutional changes, the former corrected for imperfect detectability of length classes that maintain the reading frame and the latter for undetectable missense and silent-site alterations. The resultant insertion- and deletion-rate estimates, 0.20 and 0.58 $\times 10^{-9}$ per site per generation, respectively, are not significantly different from those obtained by Kondrashov (8), 0.18 and 0.53 $\times 10^{-9}$ per site per generation, respectively. If the base-substitutional rates derived in this study were downwardly biased, the derived insertion and deletion rates would be expected to be as well.

The per-generation base-substitutional mutation rate for humans is approximately double the average rate in *Drosophila melanogaster* (approximately 4.65 $\times 10^{-9}$) (12, 25), *C. elegans* (approximately 5.60 $\times 10^{-9}$) (10, 26), and *Arabidopsis thaliana* ($\sim$6.50 $\times 10^{-9}$) (11), and substantially greater than that in *Saccharomyces cerevisiae* (approximately 0.33 $\times 10^{-9}$) (13) and *Escherichia coli* (approximately 0.26 $\times 10^{-9}$) (16). This pattern is consistent with the hypothesis that the ability of selection to minimize the mutation rate is compromised as the power of random genetic drift increases in response to reductions in effective population sizes that occur in the transitions from unicellular to multicellular species (9, 27).

Nevertheless, despite the high per-generation rate in humans, on a per–cell division basis, the human germline mutation rate is lower than that in any other organism for which reliable data are available. Assuming an average of approximately 216 germline cell divisions per generation across the sexes in humans (3), the rate of base-substitutional mutation is just 0.06 $\times 10^{-9}$ per site per germline cell division, which is approximately one fifth that in unicellular organisms. With approximately 36, 8.5, and 40 cell divisions separating gametic generations in fly (28), worm (29), and *Arabidopsis* (30), the germline per-cell division rates for these species are approximately 0.13, 0.65, and 0.16 $\times 10^{-9}$ per site, respectively. Thus, although the greatly magnified generation length in humans is not entirely compensated by the reduced mutation rate per cell division, the human per-generation rate would be 2 to 11 times higher if the rate of accumulation of mutations in replicating germline DNA were as high as that in invertebrates and *Arabidopsis*.

Inflation in the Mutational Target Size Associated with Gene Structural Complexity. In addition to having an inflated per-generation vulnerability to the accumulation of coding-region mutations, human genes also incur a considerable mutational burden associated with their inclusion of spliceosomal introns. Of the 10,930 small-scale disease-causing mutations recorded in the present study, 55% involve coding-sequence altering base substitutions (18% nonsense- and 37% missense-producing), 31% involve small coding-region insertions and deletions, and 14% involve alterations in splicing (13% base substitutions and 1% insertion/deletions). With an average intron density of approximately 1/180 coding bases, the genes in this study are not particularly intron-dense relative to other mammalian genes, whose average exon size is approximately 170 bp (9). Moreover, averaging over all loci in this study, approximately 25% of attempts to reveal a causative

mutation by sequencing coding regions in affected individuals failed to locate any mutation, and it is likely that a fraction of these harbored mutations in regions significant to gene expression (e.g., deep within introns, or in UTRs). Thus, given the likely downward bias in detecting splicing mutations, it can be concluded that at least 15% of human disease-causing mutations involve splicing.

This observation suggests that a substantial fraction of human mortality is an indirect consequence of gene structure (as opposed to amino acid sequence). From the statistics of the World Health Organization (31), approximately 49.7% of global human mortality is associated with various forms of cardiovascular disease, 23.2% with cancer, 9.7% with cerebrovascular disease, and 2.9% with other childhood diseases including congenital abnormalities, with the remainder resulting from pathogens, poisoning, accidents, war, and other sources of mainly environmental origin. Assuming all cancers arise as a consequence of mutation, and conservatively assuming that just half of these other disorders have genetic causes, then it appears that at least 8% of human deaths are a consequence of mutations incurred by the splice-site signatures of the approximately 150,000 introns carried within the nuclear genome.

Cumulative Burden of Somatic Mutations. Although only germline mutations have long-term effects on population fitness, the cumulative burden of somatic mutations compromises the demographic potential of multicellular individuals, acting as an "environmental" source of increased mortality and/or decreased fecundity. A number of studies have recorded the accumulation of various marker mutations in human tissues, and using the types of approaches outlined earlier, all results support the idea that the mutation rate per cell division is generally much greater in somatic cells than in the germline.

Based on observations on the development of retinoblastoma in heterozygous carriers of null alleles at the RB1 locus (32–35) and the parameters in Dataset S1, the average base-substitutional mutation rate in the cell lineage leading to retinal development is 0.99×10^{-9} per site per cell division. Similarly, studies of the APC gene in intestinal epithelial cells (36–38) imply a base-substitutional mutation rate of 0.27×10^{-9} per site per cell division. The rates for B and T lymphocytes derived from assays for defective alleles at the HPRT (39) and Pig-A (40) loci in cell cultures average 1.47×10^{-9} per site per cell division, and studies of the HPRT gene in human fibroblast cultures (41–43) imply a mutation rate of 0.34×10^{-9} per site per cell division. This diverse set of observations suggests that human somatic cells commonly accumulate mutations 4 to 25 times more rapidly than germline cells do, averaging approximately twice the mutation rate per cell division observed in microbial species. The male germline in the mouse is also known to have a 2- to 8-fold reduction in the mutation rate relative to that in somatic tissues (44–46). Despite these rate differences, the molecular spectrum of heritable germline mutations does not appear to be greatly different from that in somatic tissues. After correcting for detectability, the ratio of base substitutions to small insertion/deletions to splicing mutations in this study is 0.877 : 0.084 : 0.039, whereas the average over genome-wide surveys of human pancreatic, brain, colorectal, and breast cancers is 0.890 : 0.069 : 0.041 (47).

The potential consequences of somatic mutation accumulation for a species as long-lived and multicellular as humans are substantial. Even after reaching their terminal divisions, somatic cells still gradually acquire mutations in their nonreplicating DNA. Most notably, mutations arising at CpG sites appear to accumulate in an essentially replication-independent manner, i.e., proportional to absolute time (21), presumably because the deamination of methylated cytosines proceeds without regard to cell division. Thus, because approximately 9% of sites in human exons are of a CpG nature (based on the genes in this study), and because such sites experience a 15-fold inflation in the mutation rate, at least 60% of base-substitutional mutations are expected to arise in a replication-independent manner in human exons. Using the average somatic-cell mutation rate from the previously cited data, 0.77×10^{-9} per site per cell division, we can expect the mutation rate on the time scale of an average cell division to be no less than 0.46×10^{-9} per coding site. Human cell division times are typically on the order of 2 to 15 d depending on the tissue, so nonreplicating cells are expected to accumulate approximately 11 to 84×10^{-9} base-substitutional mutations per site per year in coding DNA, and replicating cells approximately 1.6 times more. Noncoding DNA harbors only approximately 1% CpG sites (48), so the corresponding values for nonreplicating cells are 1 to 9×10^{-9} base-substitutional mutations per site per year, and approximately 7.6 times higher in replicating cells.

Thus, even at the relatively young age of sexual maturity, approximately 15 y, the average somatic cell will have acquired at least 10^{-7} to 10^{-6} base substitutions per coding nucleotide, one to two orders of magnitude greater than the expectation for germline cells (approximately 10^{-8} per site, based on the present study). Over the same period, noncoding sites are expected to incur 10^{-8} to 10^{-7} base substitutions per site in nonreplicating cells, and nearly 10 times that in replicating cells. Assuming 1.5% coding DNA, these somatic rates roughly translate into approximately 100 to 1,000 mutations per nonreplicating diploid cell at the age of maturity and approximately 10^3 to 10^4 per replicating cell. These estimates, which do not include contributions from insertions, deletions, microsatellite instabilities, and segmental duplications, will be increased a further fourfold by the age of 60 y, assuming no age-specific inflation in the somatic mutation rate and no selection against mutant cells. Thus, proliferative cells (such as those in intestinal epithelium and the epidermis) at this age could very well contain 4,000 to 40,000 mutations.

Because the cell lineages of a complex organism are derived through developmental pathways, the total number of unique mutations per human soma is less than the preceding numbers times the number of cells per soma (approximately 10^{13} in humans). Nevertheless, the total somatic mutational load must be enormous. For example, the intestinal epithelium contains approximately 10^6 independent stem cells, each of which generates transient daughter cells every week or two. Thus, the intestinal epithelium of a 60-y-old is expected to harbor $>10^9$ independent mutations. This implies that, not far beyond the age of 60 y, nearly every genomic site is likely to have acquired a mutation in at least one cell in this single organ.

Although crude, these computations provide a compelling view of the enormous mutational challenges imposed by a large, long-lived soma. Much of medical research is focused on the development of strategies for thwarting the onset of cancer and other age-related disorders, but short of the acquisition of pharmaceuticals or dietary modifications for substantially reducing the somatic-cell mutation rate or the rejuvenation of tissues with relatively mutation-free germline stem cells, there appears to be little potential for substantially increasing the upper limit to the human lifespan.

Long-Term Consequences of Germline Mutations. Dating back to Muller (49), considerable thought has been given to the potential for a cumulative buildup of the deleterious-mutation load in the human population (2, 3, 50, 51). The motivation for this concern is the enormous change in the selective environment that human behavior has induced during approximately the past century. Innovations spawned by agriculture, architecture, industrialization, and most notably a sophisticated health care industry have led to a dramatic relaxation in selection against mildly deleterious mutations, and modern medical intervention is increasingly successful in ensuring a productive lifespan even in individuals carrying mutations with major morphological, metabolic, and behavioral defects. The statistics are impressive. For example, fetal mortality has declined by approximately 99% in England since the 1500s (52), and just since 1975, the mortality rate per diagnosed cancer has declined by approximately 20% in the

United States population (53). Because most complex traits in humans have very high heritabilities (54), the concern then is that unique aspects of human culture, religion, and other social interactions with well intentioned short-term benefits will eventually lead to the long-term genetic deterioration of the human gene pool. Of course, a substantial fraction of the human population still has never visited a doctor of any sort, never eaten processed food, and never used an automobile, computer, or cell phone, so natural selection on unconditionally deleterious mutations certainly has not been completely relaxed in humans. But it is hard to escape the conclusion that we are progressively moving in this direction.

The fundamental requirement for the maintenance of a species' genetic integrity and long-term viability is that the loss of mean fitness by the recurrent input of deleterious mutations each generation must be balanced by the removal of such mutations by natural selection. If the effectiveness of the latter is eliminated, normal viability and fertility can be maintained to a certain extent by modifying the environment to ameliorate the immediate effects of mutations, but this is ultimately an unsustainable situation, as buffering the effects of degenerative mutations would require a matching cumulative level of investment in pharmaceuticals, behavioral therapies, and other forms of medical intervention. Given the relatively high human mutation rate and the fact that a relaxation of natural selection typically leads to 0.1% to 1.5% decline in fitness per generation in other animal species with lower mutation rates (51), this type of scenario has now gained a level of quantitative credence that was absent when Muller (49) first raised the issue.

What is the likely magnitude of the per-generation loss in human fitness caused by recurrent introduction of deleterious mutations? From the present results, we infer that an average human gamete acquires approximately 38 de novo base-substitution mutations, approximately three small insertion/deletions in complex sequence, and approximately one splicing mutation. Transposable-element insertions, microsatellite instabilities, and segmental duplications and deletions of total or partial gene sequences will almost certainly sum to several additional events per gamete (9), so it is likely that an average newborn acquires a total of 50 to 100 new mutations at the diploid level, a small subset of which must be deleterious.

The net fitness consequences of human mutations remain unclear and will likely continue to be a major challenge, but some general arguments allow an order-of-magnitude assessment of the situation. Using rather different approaches, Yampolsky et al. (55) and Eyre-Walker et al. (56) have derived similar estimates of the distribution of fitness effects of new amino acid altering base-substitution mutations: approximately 11% cause fractional reductions in heterozygote fitness with $s < 10^{-5}$, 12% with $10^{-5} < s < 10^{-4}$, 50% with $10^{-4} < s < 10^{-2}$, and 27% with $s > 10^{-2}$, with an overall average selective disadvantage of approximately 0.04. Only approximately 1.5% of the human genome consists of coding DNA and approximately 25% of coding sites are silent, so we expect approximately 0.86 novel amino acid altering mutations per newborn. Approximately 5% of such mutations will lead to nonsense mutations, many of which will likely be in the category of $s > 10^{-2}$, but the remaining 95% will be missense in nature, with deleterious fitness effects averaging approximately 4% or less according to these results. Thus, with a complete relaxation of natural selection, the expected decline in fitness associated with mutations in coding DNA alone appears to be on the order of 1% to 3% per generation.

Less clear is the added contribution from other forms of mutations. The vast majority of point mutations reside outside of coding regions (on the order of 40 per gamete), and it is likely that most of these will have very minor fitness effects, with average s almost certainly $<< 10^{-2}$. Nevertheless, Eöry et al. (57) make a compelling case that approximately 4% of intergenic, 15% of UTR, and 22% of silent sites are under weak purifying

selection in humans, which is consistent with the arguments presented above for base-composition selection. Most major deletions and splicing mutations are probably highly deleterious, as they will generally render their host genes nonfunctional. Most transposable-element insertions and gene duplications appear to be at least weakly deleterious (9, 58); the average deleterious effects of such mutations are likely to be at least 1% per event, and as noted earlier, at least one such event is likely to arise per zygote per generation. Thus, although there is considerable uncertainty in the preceding numbers, it is difficult to escape the conclusion that the per-generation reduction in fitness due to recurrent mutation is at least 1% in humans and quite possibly as high as 5%. Although such a mutational buildup would be unnoticeable on a generation timescale, over the course of a couple of centuries (approximately six generations), the consequences are likely to become serious, particularly if human activities cause an increase in the mutation rate itself (by increasing levels of environmental mutagens). A doubling in the mutation rate would imply a 2% to 10% decline in fitness per generation, and by extension, a 12% to 60% decline in 200 years.

Because the genetic effective population size of humans is now very large, the alleles contributing to our growing mutation load will remain at very low frequencies for many centuries, so the damage incurred under the scenario outlined earlier need not be permanent. However, once a particular chromosomal region has acquired a mutation load in all chromosomes, even if distributed over different loci in different chromosomes, a full reassertion of the power of natural selection would be incapable of returning the population to a state better than that represented by the least-loaded chromosomal segment unless the region experienced sufficient recombination to reestablish less-loaded chromosomal segments. This issue is nontrivial in that the amount of recombination per physical distance in human chromosomes is exceptionally low, averaging approximately 10^{-5} crossover events per kb per meiosis (9), which is approximately 30% lower than the rate of origin of base-substitutional mutations alone in a 1-kb region.

In summary, it is remarkable that, 60 y ago, before the elucidation of the structure of DNA and drawing on very limited data under the assumption of just 5,000 genes in the human genome, Muller (49) estimated that the average newborn human acquires 0.2 to 1.0 de novo deleterious mutations. Using today's estimate of the human genome content (approximately 22,500 genes), these numbers scale up to 0.9 to 4.5 deleterious mutations per diploid genome per generation, a range that is quite consistent with the predictions from modern molecular data. Muller (49) was well aware of the enormous social barriers to solving the mutation-accumulation problem, but he held out hope that "a rationally directed guidance of reproduction" would eventually stabilize the situation. Unfortunately, it has become increasingly clear that most of the mutation load is associated with mutations with very small effects distributed at unpredictable locations over the entire genome, rendering the prospects for long-term management of the human gene pool by genetic counseling highly unlikely for all but perhaps a few hundred key loci underlying debilitating monogenic genetic disorders (such as those focused on in the present study).

Thus, the preceding observations paint a rather stark picture. At least in highly industrialized societies, the impact of deleterious mutations is accumulating on a time scale that is approximately the same as that for scenarios associated with global warming—perhaps not of great concern over a span of one or two generations, but with very considerable consequences on time scales of tens of generations. Without a reduction in the germline transmission of deleterious mutations, the mean phenotypes of the residents of industrialized nations are likely to be rather different in just two or three centuries, with significant incapacitation at the morphological, physiological, and neurobiological levels. Ironically, the genetic future of mankind may

reside predominantly in the gene pools of the least industrialized segments of society. Possible solutions to this problem, including multigenerational cryogenic storage and utilization of gametes and/or embryos, will raise significant ethical conflicts between short-term and long-term considerations.

Methods

Database. This study began with the Online Mendelian Inheritance in Man database maintained at the National Center for Biotechnology Information Web site. The total pool of listings for human genetic disorders with a known or suspected autosomal dominant or X-linked genetic basis yielded several thousand candidate disorders for analysis, many of which were redundantly annotated. For each disorder, the existing literature through 2008 was scrutinized for the data necessary for estimating mutation rates and spectra, resulting in sufficient data for the complete analysis of 21 autosomal-dominant and 13 X-linked disorders. An additional 23 autosomal-dominant and five X-linked loci yielded data sufficient for the analysis of the spectrum of mutational effects, while not harboring the data essential for mutation-rate estimation. Citations for the literature from which the informational details were derived for each disorder can be found in the *SI Text*.

Mutation Rate/Disorder. The total germline mutation rate to defective alleles at each locus (u_L) was estimated in two ways. First, a direct estimate of u_L was possible if the fraction of newborns in the population acquiring de novo mutations (f_N) was known. Letting p denote the penetrance of de novo mutations (the fraction of newborns carrying a newly arisen disease allele that exhibit the disorder) yields the following equation:

$$u_{LD} = f_N/(2p) \qquad [2]$$

in the case of a dominant autosomal disorder (the two accounting for mutational origin through either parent). Often, p is unknown, in which case $p = 1$ must be assumed and the previous expression would underestimate the mutation rate were this assumption violated. Second, if the incidence of the disease in the general population was known, an indirect estimate of u_L could be acquired under the assumption that the frequency of disease alleles has achieved selection-mutation balance. This further requires information on the fractional selective disadvantage of heterozygous carriers of disease alleles (s). Letting f_A denote the incidence of affected individuals, so that $f_A/(2p)$ estimates the frequency of the disease alleles, the indirect estimate of the per-locus mutation rate is as follows:

$$u_{LI} = f_A \cdot s/(2p) \qquad [3]$$

When both types of estimates were possible, they were averaged to yield a final estimate of the locus-specific mutation rate to defective alleles.

For X-linked diseases, where possible, direct estimates of the mutation rate were calculated via both mothers and fathers. The direct estimate of the female rate (u_{LFD}) was obtained as the fraction of male newborns exhibiting the trait, conditional on the mother not being a carrier, further divided by the penetrance. A direct estimate of the sum of male and female mutation rates ($u_{LMD} + u_{LFD}$) was obtainable for cases in which the fraction of female newborns who are new carriers was known. When both estimates were available, an estimate of the male rate was obtained by difference of these two measures. The average per-locus rate was then estimated as ($2u_{LFD} + u_{LMD}$)/3, as two thirds of X chromosomes are carried by females.

These indirect estimates of the male mutation rate are not necessarily very reliable, as the second X estimate is usually based on a small number of mothers observed to be carrying a new mutation, which further requires that the parents of carrier mothers be assayed. An alternative approach to obtaining the average rate from direct observational data on the X is to scale up the more easily obtained female rate by the observed inflation in the rates of paternally versus maternally derived mutations across all disorders for which such data are available. For the studies of autosomal mutations sur-

veyed here that discriminated between parental origins of mutations, 136 were male-derived and 18 female-derived. For X-linked disorders, the respective numbers were 99 and 18, so the overall ratio is 235:36, implying an average male mutation rate that is approximately 6.5 times that in females. This is not greatly different from a prior ratio of 5.2 obtained by phylogenetic comparisons of primate sequences (59). Using this scaling factor, the average mutation rate for an X-linked gene can then be estimated as $8.5u_{LFD}/3$. For comparative purposes, these sex-averaged estimates for X-linked genes can be converted to an autosomal equivalent (where the male:female weighting is 1:1) using ($u_{LFD} + 6.5u_{LFD}$)/2 = $3.75u_{LFD}$.

Indirect estimates of the average mutation rate for X-linked genes were acquired as follows:

$$u_{LI} = f_A \cdot (s/3)/p. \qquad [4]$$

where it is assumed that selection operates only within males, and the incidence is determined from the total fraction of affected male newborns.

Base-Substitutional Mutation Rate/Nucleotide Site. As the per-locus mutation rates to defective alleles are functions of many aspects of gene size, structure, and sensitivity, their primary utility here is in acquiring appropriately scaled estimates at the nucleotide level. The mutation rate to single-site base pair substitutions was computed as follows:

$$u_{BS} = u_L \cdot f_T \cdot f_{BS}/(L \cdot f_L \cdot f_D) \qquad [5]$$

where f_T is the fraction of mutations found after sequencing the candidate locus, f_{BS} is the fraction of detected mutations resulting from base substitutions, L is the length of the coding region (in bases), f_L is the fraction of the length wherein mutations have detectable phenotypic effects, and f_D is the fraction of base-substitutional mutations that are detectable. The latter is calculated as

$$f_D = x(n_m + n_n)/n_n \qquad [6]$$

where n_m and n_n denote the numbers of observed missense and nonsense mutations, and the factor x denotes the fraction of base substitutions at sense codons that are expected to lead to nonsense mutations. In introducing this general approach, Drake (6) assumed $x = 3/64$, based on the idea that three of 64 codons are chain terminating (ignoring bias in the mutational spectrum) and the fact that only a fraction of mutations at sense codons have functional effects. Here x was obtained more directly in a gene-specific manner by applying the full base-substitutional mutation spectrum (derived from all loci surveyed in this study) to the specific codons within each gene. The resultant values of x range from 0.027 to 0.057, with a mean of 0.040 (0.001). Using this approach, f_D takes on extreme values of x and 1.0 when the fractions of observable base-substitution mutations creating nonsense mutations at a locus are 1.0 and x, respectively. [In principle, n_m should also include synonymous (i.e., silent) mutations, but these are almost never recorded as being causal, and so with approximately 25% of sites within sense codons being silent, the upper limit to f_D is actually approximately 0.65 using the average value of x.] Use of the scaling factor f_D assumes that all nonsense mutations within the region of the gene susceptible to phenotypic effects are detectable, and to the extent that this is not the case, the mutation rate will be underestimated. In addition, f_L is often unknown and must then be assumed to equal 1.0, which will cause an underestimate of the mutation rate if the true value of f_L is <1.0.

1. Crow JF (1993) How much do we know about spontaneous human mutation rates? *Environ Mol Mutagen* 21:122–129.
2. Crow JF (1997) The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci USA* 94:8380–8386.
3. Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* 1:40–47.
4. Xue Y, et al.; Asan (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol* 19:1453–1457.
5. Mardis ER, Wilson RK (2009) Cancer genome sequencing: a review. *Hum Mol Genet* 18 (R2):R163–R168.

6. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88:7160–7164.
7. Haldane JBS (1935) The rate of spontaneous mutation of a human gene. *J Genet* 31:317–326.
8. Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 21:12–27.
9. Lynch M (2007) *The Origins of Genome Architecture* (Sinauer Assoc, Sunderland, MA).
10. Denver DR, et al. (2009) A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc Natl Acad Sci USA* 106:16310–16314.
11. Ossowski S, et al. (2009) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, in press.

12. Keightley PD, et al. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* 19:1195–1201.

13. Lynch M, et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 105:9272–9277.

14. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

15. Kondrashov FA, Ogurtsov AY, Kondrashov AS (2006) Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J Theor Biol* 240:616–626.

16. Lynch M (2006) The origins of eukaryotic gene structure. *Mol Biol Evol* 23:450–468.

17. Li W-H (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol* 24:337–345.

18. Cooper DN, Krawczak M (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum Genet* 85:55–74.

19. Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* 4:e1000071.

20. Albrecht-Buehler G (2009) The spectra of point mutations in vertebrate genomes. *Bioessays* 31:98–106.

21. Kim SH, Elango N, Warden C, Vigoda E, Yi SV (2006) Heterogeneous genomic molecular clocks in primates. *PLoS Genet* 2:e163.

22. Eckhardt F, et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38:1378–1385.

23. Weber JL, et al. (2002) Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet* 71:854–862.

24. Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet* 14:59–69.

25. Haag-Liautard C, et al. (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445:82–85.

26. Denver DR, Morris K, Lynch M, Thomas WK (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* 430:679–682.

27. Lynch M (2008) The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics* 180:933–943.

28. Drost JB, Lee WR (1995) Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among Drosophila, mouse, and human. *Environ Mol Mut* 26(suppl):48–64.

29. Wilkins AS (1992) *Genetic Analysis of Animal Development* (Wiley-Liss, New York), 2nd Ed.

30. Hoffman PD, Leonard JM, Lindberg GE, Bollmann SR, Hays JB (2004) Rapid accumulation of mutations during seed-to-seed propagation of mismatch-repair-defective *Arabidopsis*. *Genes Dev* 18:2676–2685.

31. World Health Organization (2004) Annex Table 2: Deaths by cause, sex and mortality stratum in WHO regions, estimates for 2002. *The World Health Report 2004 – changing history* (WHO, Geneva).

32. Hethcote HW, Knudson AG, Jr (1978) Model for the incidence of embryonal cancers: application to retinoblastoma. *Proc Natl Acad Sci USA* 75:2453–2457.

33. Fitzgerald PH, Stewart J, Suckling RD (1983) Retinoblastoma mutation rate in New Zealand and support for the two-hit model. *Hum Genet* 64:128–130.

34. Morris JA (1990) Spontaneous mutation rate in retinoblastoma. *J Clin Pathol* 43:496–498.

35. Lloyd RA, Papworth DG (1995) Retinoblastoma: a model for deriving the mutation rate without using any estimate of the size of the population at risk. *Mutat Res* 326:117–124.

36. Iwama T (2001) Somatic mutation rate of the APC gene. *Jpn J Clin Oncol* 31:185–187.

37. Luebeck EG, Moolgavkar SH (2002) Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci USA* 99:15095–15100.

38. Hornsby C, Page KM, Tomlinson I (2008) The *in vivo* rate of somatic adenomatous polyposis coli mutation. *Am J Pathol* 172:1062–1068.

39. Lichtenauer-Kaligis EG, et al. (1996) Comparison of spontaneous hprt mutation spectra at the nucleotide sequence level in the endogenous hprt gene and five other genomic positions. *Mutat Res* 351:147–155.

40. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.

41. Bhattacharyya NP, et al. (1995) Molecular analysis of mutations in mutator colorectal carcinoma cell lines. *Hum Mol Genet* 4:2057–2064.

42. Glaab WE, Tindall KR (1997) Mutation rate at the hprt locus in human cancer cell lines with specific mismatch repair-gene defects. *Carcinogenesis* 18:1–8.

43. Umar A, et al. (1998) Functional overlap in mismatch repair by human MSH3 and MSH6. *Genetics* 148:1637–1646.

44. Dycaico MJ, et al. (1994) The use of shuttle vectors for mutation analysis in transgenic mice and rats. *Mutat Res* 307:461–478.

45. Walter CA, Intano GW, McCarrey JR, McMahan CA, Walter RB (1998) Mutation frequency declines during spermatogenesis in young mice but increases in old mice. *Proc Natl Acad Sci USA* 95:10015–10019.

46. Hill KA, et al. (2005) Tissue-specific time courses of spontaneous mutation frequency and deviations in mutation pattern are observed in middle to late adulthood in Big Blue mice. *Environ Mol Mutagen* 45:442–454.

47. Jones S, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321:1801–1806.

48. Jabbari K, Bernardi G (2004) Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 333:143–149.

49. Muller HJ (1950) Our load of mutations. *Am J Hum Genet* 2:111–176.

50. Kondrashov AS (1995) Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J Theor Biol* 175:583–594.

51. Lynch M, et al. (1999) Spontaneous deleterious mutation. *Evolution* 53:645–663.

52. Woods R (2005) The measurement of historical trends in fetal mortality in England and Wales. *Popul Stud* 59:147–162.

53. Horner MJ, et al., eds (2009) *SEER Cancer Statistics Review, 1975-2006* (Natl Cancer Inst, Bethesda, MD).

54. Lynch M, Walsh JB (1998) *Genetics and Analysis of Quantitative Traits* (Sinauer Assoc, Sunderland, MA).

55. Yampolsky LY, Kondrashov FA, Kondrashov AS (2005) Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet* 14:3191–3201.

56. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900.

57. Eöry L, Halligan DL, Keightley PD (2009) Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol*, in press.

58. Nguyen DQ, et al. (2008) Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res* 18:1711–1723.

59. Makova KD, Li W-H (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416:624–626.