Hypermutability of Genes in *Homo sapiens* Due to the Hosting of Long Mono-SSR

Etienne Loire,†‡§||¶#** Françoise Praz,††‡‡ Dominique Higuet,§||¶# Pierre Netter,*†‡ and Guillaume Achaz§||¶#***

*Université Pierre et Marie Curie-Paris 6, Unité Mixte de recherche (UMR) 7592, Institut Jacques Monod, Paris, France; †Centre National de la Recherche Scientifique (CNRS), UMR 7592, Institut Jacques Monod, Paris, France; ‡Université Denis Diderot-Paris7, UMR 7592, Institut Jacques Monod, Paris, France; §Université Pierre et Marie Curie Paris 6, UMR 7138, Systématique, Adaptation, Evolution, Paris, France; ||CNRS, UMR 7138, Systématique, Adaptation Evolution, Paris, France; ¶Museum National d'Histoire Naturelle, UMR 7138, Systématique Adaptation Evolution, Paris, France; #Institut National de al Sauté et de la Recherche Médicale, UMR 7138, Systématique, Adaptation Evolution, Paris, France; **Université Pierre et Marie Curie-Paris 6, Atelier de Bioinformatique, Paris, France; ††Université Pierre et Marie Curie-Paris 6, UMR_S 893, CdR Saint-Antoine, Paris, France; and ‡‡INSERM, UMR_S 893, CdR Saint-Antoine, Paris, France

Simple sequence repeats (SSRs) are very common short repeats in eukaryotic genomes. "Long" SSRs are considered "hypermutable" sequences because they exhibit a high rate of expansion and contraction. Because they are potentially deleterious, long SSRs tend to be uncommon in coding sequences. However, several genes contain long SSRs in their exonic sequences. Here, we identify 1,291 human genes that host a mononucleotide SSR long enough to be prone to expansion or contraction, being called hypermutable hereafter. On the basis of Gene Ontology annotations, we show that only a restricted number of functions are overrepresented among those hypermutable genes including cell cycle and maintenance of DNA integrity. Using a probabilistic model, we show that genes involved in these functions are expected to host long SSRs because they tend to be long and/or are biased in nucleotide composition. Finally, we show that for almost all functions we observe fewer hypermutable sequences than expected under a neutral model. There are however interesting exceptions, for example, genes involved in protein and RNA transport, as well as meiosis and mismatch repair functions that have as many hypermutable genes as expected under neutrality. Conversely, there are functions (e.g., collagen-related genes) where hypermutable genes are more often avoided than in other functions. Our results show that, even though several functions harbor unusually long SSR in their exons, long SSRs are deleterious sequences in almost all functions and are removed by purifying selection. The strength of this purifying selection however greatly varies from function to function. We discuss possible explanations for this intriguing result.

Introduction

Microsatellites or simple sequence repeats (SSRs) are arrays of DNA with short motifs—1–6 nt—repeated in tandem (Tautz 1994). SSRs are ubiquitous in all genomes explored so far and are especially abundant in eukaryote genomes (Toth et al. 2000). Strikingly, the number and the sizes of SSRs in genomes are typically much larger than expected from simple substitution models (Pupko and Graur 1999). This overabundance of SSRs is, most likely, a consequence of their specific mutational properties; these repeats are prone to expansion and contraction through polymerase strand slippage (Levinson and Gutman 1987) and, to a lesser extent, to recombination (Li et al. 2002). For strand slippage, after the replication fork has run, template and neosynthesized strands can be reannealed with the slippage of one (or more) motifs. If the "mismatch repair" (MMR) complex does not correct the resulting loop, a subsequent round of replication changes the number of repeated units by a specific amount. This translates into an insertion or a deletion of one (or more) motifs in the SSR.

Various factors have been shown to modulate the rate of SSR expansion/contraction, although their relative strength varies from species to species (Toth et al. 2000). It appears that, for "long" SSRs, contraction prevails over expansion (Xu et al. 2000). This bias in favor of contraction, along with a higher chance of being interrupted by a substitution for even longer SSRs, prevents their infinite growth (Kruglyak et al. 1998; Ellegren 2000; Dieringer and Schlotterer 2003). The nature of the motif itself also greatly modulates the mutation rate. For example, GC-rich SSRs are more unstable than others (Sagher et al. 1999; Gragg et al. 2002), and long motifs are more stable than short ones (Rose and Falush 1998; Legendre et al. 2007). Overall, the relative role of all these factors makes difficult to predict a mutation rate for a given SSR. However, it remains true that the SSR mutation rate is typically several orders of magnitude higher than the average substitution rate (Drake et al. 1998).

A number of studies in a wide range of organisms have attempted to delimit characteristics of SSRs that are predictive of the variability of the repeats. They concluded that the number of repeats was among the strongest predictors of the slippage probability during replication (Rose and Falush 1998; Lai and Sun 2003b; Legendre et al. 2007; Kelkar et al. 2008). Repeat variability is not an all-or-nothing phenomenon but rather increases exponentially with increasing number of repeat units, as initially established in yeast (Sia et al. 1997).

At what length should SSRs be deemed "hypermutable"? Using a simple probabilistic model, Rose and Falush (1998) proposed a threshold size for slippage mutations around 8 bp for mono-, di-, and tetranucleotide SSRs. Based on a different model, similar thresholds were proposed: 9 units for mononucleotide SSRs (mono-SSRs) and 4 units for dinucleotide (8 bp) and tetranucleotide (16 bp) SSRs (Lai and Sun 2003a). Using a human/chimpanzee complete genome comparison, it appears that, in this lineage, a mononucleotide of 9 units exhibit a similar mutability than a dinucleotide of 6 units or a tetranucleotide SSR of 5 units (Kelkar et al. 2008). Alternatively, we can also infer that a mononucleotide of 8 units exhibit

Key words: microsatellites, SSR, evolution, mutability, *Homo sapiens*.

E-mail: loire@abi.snv.jussieu.fr.

Mol. Biol. Evol. 26(1):111–121. 2009
doi:10.1093/molbev/msn230
Advance Access publication October 8, 2008

a similar mutability than a dinucleotide of 5 units or a tetranucleotide SSR of 4 units.

Interestingly, similar threshold sizes for mononucleotide and dinucleotide SSRs instability were observed in vitro during polymerase chain reaction (Lai and Sun 2003a; Shinde et al. 2003). For mono-SSRs, the observation of human oncogenesis associated with microsatellite instability (MSI) also highlights 8 units as an instability threshold. MSI has been shown to underlie hereditary nonpolyposis colorectal cancer (HNPCC) (Aaltonen et al. 1993). HNPCC patients carry a germ-line mutation in one of the postreplicative MMR genes, mainly MLH1 or MSH2 (Jacob and Praz 2002; Woerner et al. 2006). Once the corresponding normal allele is lost through somatic inactivation, cells become totally devoid of MMR activity and are left with unrepaired polymerase errors that arise during replication. Rates of mutation arising in microsatellite repeats are drastically enhanced by mutations affecting postreplicative DNA MMR (Strand et al. 1993). In this context, only genes with an SSR of at least 8 nt have been reported to exhibit a significant instability (Duval and Hamelin 2003; Woerner et al. 2006; Miquel et al. 2007). Altogether, these results suggest common features of microsatellite mutation mechanisms both in vivo and in vitro with evidence of a slippage mutation threshold at around 8 or 9 units for mono-SSRs.

SSRs tend to be less common in coding sequences (Metzgar et al. 2000; Ackermann and Chao 2006) as a change in nucleotide number often has disastrous functional consequences. If the unit length is a multiple of three, there will be an expansion or a contraction of the particular amino acids encoded by the 3-mer (codon). It is well established that long expansions of such coding microsatellites are responsible for many neurodegenerative disorders (Everett and Wood 2004). When the unit length is not a multiple of three, a change in unit number produces a frameshift (Strauss 1999). If the slippage occurs during the replication process, it may create an allele that contains a premature stop codon either in somatic cells or in the germ line. Slippage can also occur during transcription (Fabre et al. 2002), leading to abnormal messenger RNA that is usually degraded by the nonsense-mediated mRNA decay system (Conti and Izaurralde 2005). Because SSRs in coding sequences are typically associated with deleterious effects, they tend to be subject to purifying selection. We want to emphasize that SSRs that have unit lengths that are not a multiple of three have a direct, harmful potential in coding sequences because no slippage can be tolerated; therefore, they should be even less common within exons. Intriguingly, it has been observed that many genes involved in DNA repair, including MMR, carry a long mono-SSR in their coding sequences (Mori et al. 2001; Miquel et al. 2007). If these particular SSR experience an expansion or a contraction, the MMR system will become deficient and will lead to a higher mutation rate (as observed in some HNPCC-associated tumors). It has been postulated that a deficient MMR system could be advantageous when the environment is stressful. In this case, organisms with a higher mutation rate could adapt more easily to environmental challenges. Consequently, mono-SSRs in these genes could have been positively selected for their mutational potential (Moxon and Wills 1999; Chang et al. 2001; Kashi and King 2006).

In the present study, we have detected all strict SSRs—that is perfect repeats without any "interruption" in the pattern—in all human genes. We used the presence of a long SSR (with at least 8 [or 9] units for mono-SSR, 5 [or 6] units for di-SSR, and 4 [or 5] units for tetra- and penta-SSRs) as a proxy for the hypermutability of genes (Rose and Falush 1998; Lai and Sun 2003b). Even though many other factors can influence the mutability of genes, the presence of a long SSR greatly increases the chances for a gene to be inactivated. Indeed, the probability of a nonsense substitution is several orders of magnitude lower than the rate of slippage of a long enough SSR. We found mono-SSRs to be the most abundant unstable SSR as well as the most biased in term of hosting genes' function. Consequently, we focused our study on mono-SSR and used the term "hypermutable genes" to refer to genes that carry a long (and therefore potentially unstable) mono-SSR in their coding sequence hereafter.

Using annotations from the Gene Ontology (GO) database (Ashburner et al. 2000), we performed an in silico functional analysis of all genes that are a priori hypermutable. We found a cohesive restricted subset of functions that are overrepresented among hypermutable genes. To take into account differences due to the length and the composition of genes, we computed for each gene the probability to host a long mono-SSR. In this statistical framework, we observe less hypermutable genes than expected in almost all functions, including the ones we found overrepresented. This shows that, typically, hypermutable genes are removed by purifying from the human genome because of their deleterious potential. Interestingly, we observe that the strength of the purifying selection, that removes long mono-SSR, varies from function to function.

Materials and Methods

Microsatellites in Human-Coding Sequences

We extracted all exons and introns from all transcripts of the 22,218 genes from the human genome of the database Ensembl v37. Each gene sequence was then reduced to its exonic sequences only. When exons of different transcripts were overlapping, we merged them into an artificial exonic-like sequence. For each gene, we then concatenated all its nonredundant exonic-like sequences into a single sequence and inserted an "*X*" at each junction. The *X* tag ensures that no microsatellite can be detected astride two different exons. The same procedure was applied to introns to build up a unique intronic sequence for each gene. We built two sets, each composed of 22,218 artificial exonic sequences and 18,384 artificial intronic sequences derived from all transcripts.

We detected all strict SSRs (no interruption in the pattern) of a motif whose length ranges from 1 to 5.

Statistics on Mono-SSR

The following model is very similar to previous models that were used to describe the probability of observing a given SSR in sequences (de Wachter 1981).

Interestingly, the functional bias of hypermutability is only driven by mono-SSR. Therefore, the statistical framework focused on mono-SSR exclusively. Extensions for longer motifs are given in Robin et al. (2005).

Probability of a Given Mono-SSR

We will give here an approximation of the probability to observe at least one occurrence of an X-SSR of size m^+ (m or more) in a random sequence of L independent letters of the {A, T, G, and C} alphabet. P_x will denote the probability to generate a nucleotide X in such random sequence.

Let us first note that the number of occurrences of an X-SSR of size m^+, denoted by N_x, is exactly the number of clumps of the m-mer $(X)_m$. A clump of a motif is defined here as the maximal set of overlapping occurrences of this motif in the sequence (Robin et al. 2005). The expectation of N_x is thus given by

$$E(N_X) = (1 - P_X) \times (P_X)^m \times (L - m + 1),$$

and N_x can be approximate by a Poisson random variable (Robin et al. 2005). Therefore, we have

$$P(N_X \geq 1) = 1 - P(N_X = 0),$$

where

$$P(N_X = 0) = e^{-E(N_X)}.$$

Expected Size of a Mono-SSR

We first computed, for each sequence and for each type of nucleotide, the m value that corresponds to $P(N_x \geq 1) \geq 0.5$. This value will be named $m_{1/2}$. If the model fits the data, a given gene has 50% chance of having its longest SSR larger than $m_{1/2}$. We can then affect all genes to either a "larger" or a "smaller" category depending whether its longest mono-SSR is larger or smaller than its $m_{1/2}$. Because these are independent Bernoulli trials, we expect for a set of genes that half of it should be in the larger category. We can then test if the genes tend to have a smaller/larger mono-SSR than expected using a χ^2 test.

Expected Fraction of Hypermutable Genes

We also computed the expected fraction of genes carrying a long mono-SSR (m fixed) in their coding sequences for a set of genes. To do so, we calculated, for each gene of this set, the probability of observing at least one mono-SSR of length m^+ of any type of nucleotide (with m = 8 or 9). We assume that the probability for each type of SSR is independent. Because m is not very small, this approximation is reasonable. In a given gene, the probability to find at least one mono-SSR of length m^+ is

$$P\left(N_{\mathrm{A,C,G,T}} \geq 1\right) = 1 - P(N_{\mathrm{A}} = 0) \times P(N_{\mathrm{C}} = 0) \times P(N_{\mathrm{G}} = 0) \times P(N_{\mathrm{T}} = 0).$$

The average of all these probabilities for a given function is an unbiased estimator of the expected fraction of hypermutable genes in this function.

Finally, using this model, we can compute the confidence interval (CI) associated with its expected fraction of hypermutable gene. To do so, one needs to compute the probability that, among N genes, each having a probability $P(N_{\mathrm{A,\ C,\ G,\ T}} \geq 1)$ to host a mono-SSR at size m^+, n genes have such an SSR. These are N independent Bernoulli trials with different probabilities of success. We estimate the probability to obtain at least n hypermutable genes for a given term by simulations. For each term, we randomly run N Bernoulli trials with respect to the individual probability of each gene. This procedure is repeated 10^5 times for a given term. The empirical distribution is then used to compute a 95% CI for a given set of genes.

Functional Group of Human Genes

We used GO (Ashburner et al. 2000) as well as Panther Ontology (Mi et al. 2005) to assign human genes to functional groups. Both databases are based on organized ontologies, a controlled vocabulary for the description of gene products. More precisely, there are constituted of terms (i.e., GO term or PantherID) that describe a "biological process" (BP), a "molecular function" (MF), or a "cellular component" (CC) (although this latter category does not exist in Panther Ontology). For all genes, we considered all available annotations. We retrieved GO terms from Ensembl (http://www.ensembl.org/biomart/martview/) and Panther IDs from the Panther database Web page (http://www.pantherdb.org/).

Here, we defined the level of a term as the number of nodes that exists between this term and the root of the graph (level 0). In the cases of multiple paths, we keep the shortest one. We decided to compare only terms lying at the same level. We used the annotated term of a gene to browse the ontologies and collect all its parental terms. For each level, we considered only genes that have at least one defined term.

Representation of Gene Functions among the Data Set

We wanted to test if any function were overrepresented among genes carrying a long SSR. For that purpose, we used a cumulative hypergeometric law (see e.g., Castillo-Davis and Hartl 2003 as suggested Rivals et al. 2007).

We perform our tests level by level to compare comparable terms. For each level of the ontologies, we performed one test per term. To correct for multiple tests, we considered that terms lying at the same level of the ontology were independent and therefore can be corrected using the Bonferroni correction. On the contrary, we considered that tests between levels were fully dependent because they use the same annotations but with different accuracies.

Results

In this study, we restricted ourselves to strict SSRs that contain no nucleotide interruption, which tend to stabilize microsatellites (Ellegren 2004) and thus lower their intrinsic mutability. For each of the 22,218 annotated genes in the human genome, we detected all strict SSRs in concatenated exonic and intronic sequences. Because we were interested in studying genes that are susceptible to direct inactivation by SSR contraction or expansion, we excluded SSR that had a unit length that was a multiple of three.

Long SSRs in Coding Sequences Are Mononucleotide and Dinucleotide SSRs

For each gene, we identified the largest mono-, di-, tetra-, and pentanucleotide SSR (frameshifting SSR) in exonic and, when available, in intronic sequence. Because the rate of insertion/deletion grows exponentially with the number of repeat units (Tran et al. 1997; Legendre et al. 2007), the longest SSR in the exons of a given gene provides a good approximation for gene hypermutability.

Results (fig. 1) show that all types of SSRs are smaller in exons than in introns. Indeed, intronic SSRs are four times longer than exonic SSR for mono-SSR (5.8 vs. 21.9) and 2.5 times longer for pentanucleotide SSR (1.3 vs. 3.4). One could relate this observation to the purifying selection that acts against the expansion of SSR in coding sequence; however, intronic sequences are much longer than exonic sequences (i.e., on average 30 times). Both factors contribute to this difference, as it will be shown below.

As mentioned above, mono-SSRs are estimated to be unstable when they reach a length of 8 units (Rose and Falush 1998) or 9 units (Lai and Sun 2003b). If we consider a threshold of 8 units, the corresponding mutabilities are reached for di-, tetra-, and penta-SSRs for 5, 4 and 4 units, respectively. In this case, the numbers of genes, in the human genome, having an SSR longer or equal than the threshold, are 1,291 for mono-SSR (5.8% of all genes), 678 for di-SSR (3.1%), 39 for tetra-SSRs (0.2%), and 11 for penta-SSRs (<0.1%) and a total of 1,935 (8.7%) genes. Using thresholds of 9, 6, 5, and 5 units for mono-, di-, tetra-, and penta-SSRs yields to 417 for mono-SSR (1.9%), 116 for di-SSR (0.52%), 8 for tetra-SSRs (<0.1%), and 1 for penta-SSRs (<<0.1%) and a total of 475 (2.1%) genes.

If we assume that those thresholds represent the minimum numbers of units to observe instability, the SSRs that mostly participate to gene hypermutability are clearly mono-SSR and di-SSR.

Hypermutable Genes Are Overrepresented in a Restricted Subset of Functions

Using either the lower (8 units for mono-SSRs) or the higher (9 units for mono-SSRs) threshold, we define a set of genes that have, a priori, a high probability to be disrupted by a nonsense mutation due to the expansion/contraction of the SSR they host. We then searched for overrepresented terms of GO (Ashburner et al. 2000) among the set of genes.

We worked on the subset of 15,385 genes (69% of total) that had at least one term in one of the three graphs. Note that 57% of all genes have one term in BP, 63% in MF, 54% in CC, and 48% in the three. The fraction of genes with a long SSR within each subset is identical (data not shown). It is however important to note that more specific levels are made up of fewer annotated genes.

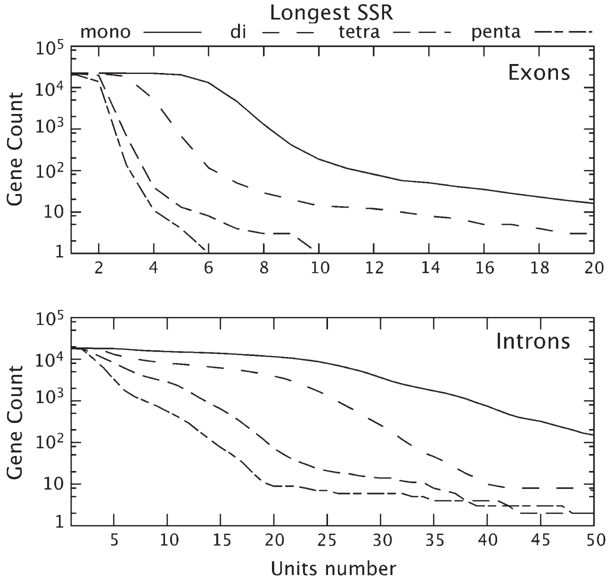

FIG. 1.—Distribution of SSR length in human genes. Counts of human genes that contain an SSR which size is equal or larger than the value given in *x* axis. The size is expressed in number of units. We only report the results for SSRs whose motif length is not a multiple of three. In the top panel, we report results for exonic sequences, whereas results for intronic sequences are displayed on the bottom panel. This figure illustrates that introns carry larger SSRs than exons do and that long SSRs in exons are mostly mono- or di-SSRs.

From all terms that were annotated at least once in the human genome (supplementary table S1, Supplementary Material online), only a few were found overrepresented. No function was overrepresented if only genes hosting a long tetra- or a penta-SSR were considered, and their removal has no impact on the results. More surprisingly, there is no function overrepresented among genes with long di-SSR, and their removal leaves the results almost unchanged (supplementary table S2, Supplementary Material online). Therefore, the only SSRs that are not uniformly distributed among functions are the mono-SSRs.

Figure 2 shows all terms we found overrepresented in hypermutable genes when mono-SSRs of 8 bp or more are considered. Results with mono-SSR of 9 bp are consistent with the former and are presented in supplementary table S2 (Supplementary Material online). Among the 3,122 BP terms, only 10 were statistically overrepresented (fig. 2*a*). Interestingly, genes with mono-SSRs are enriched for functions involved in either "cell cycle" or "response to DNA damage stimulus." Many of these hypermutable genes carry both types of annotations or related ones. The overrepresented terms are more or less precise descriptions of the same subset of functions. Following Alexa et al. (2006), if we remove the 12 genes that are annotated as functioning in meiosis (the most specific overrepresented term), no BP terms are found to be overrepresented. Therefore, genes with this function are responsible for the more general terms found to be overrepresented. Because there is no reason to believe that the most precise terms are most informative, we present results for all levels.

The same trend is observed for MF (fig. 2*b*) and CC (fig. 2*c*). In MF, out of the 2,600 terms, only 15 highly connected terms are found overrepresented. These terms all relate to "hydrolase" (especially "ATPase"), "helicase," "GTPase regulator," and "ATP binding." Removing ATPase and GTPase regulator genes from the data set suppresses other overrepresentations in MF. As for CC, only five terms (out of 583) are overrepresented and all are related to "nucleus." The "intracellular nonmembrane-bound organelle" term encompasses intracellular molecular components such as the kinetochores, the chromosomes, and the nucleosome. Ignoring genes from nucleus does not alter the overrepresentation in intracellular nonmembrane-bound organelle and vice versa. Obviously, removing genes annotated by the latter suppresses the overrepresentations of shallower related terms.

Among Overrepresented Functions, Genes Are Longer and/or More Biased in Composition

Only a restricted number of functions are overrepresented in hypermutable genes. In the three graphs, these functions all relate to cell cycle and "DNA maintenance." We wanted to test whether genes involved in these functions have a higher chance of hosting a long mono-SSR. In this respect, we computed, for each gene, the probability of finding a long mono-SSR (8 bp or more) given its length and composition. The probability model we used here assumes that mono-SSRs are only generated by several independent substitutions that keep the average nucleotide content of the gene unchanged. It is therefore used to check whether the presence of a given mono-SSR in a given gene can be explained by random point mutations only. This model does not include the possibility of slippage for modifying the size of coding mono-SSR. Indeed, insertion or deletion of 1 or 2 units in a coding SSR whose motif length is not a multiple of three leads to a frameshift mutation. Thus, fixation of such events must be extremely rare.

The average probability of having a mono-SSR of 8 units or more in genes involved in the function we find overrepresented is 0.184, that is higher than 0.142, the average for the other annotated genes ($P << 10^{-16}$, Wilcoxon U test). This shows that, on average, genes involved in the function we found overrepresented have a higher probability to host a long mono-SSR.

Mono-SSRs Are Typically Shorter than Expected in Exons

Because this model assumes that all substitutions can occur freely with respect to the gene nucleotide composition, this model can be used as a neutral model. Indeed, this model corrects for local composition and therefore for potential local mutation biases. Furthermore, it assumes that all substitutions occur freely within the sequence, which implies the neutrality of substitutions. From the comparison of what is expected under the model to what we observe, we are able to test for the neutrality of mono-SSR.

We first tested whether the length of mono-SSR, we observe in genes, is expected under the neutral model. To do so, we computed for each gene, $m_{1/2}$, the size of the SSR that corresponds to a probability $P = 0.5$. If an SSR originates from several independent selectively neutral substitutions, half of the genes will have a mono-SSR larger than $m_{1/2}$, the other half will have a mono-SSR smaller than $m_{1/2}$. We counted all genes that were hosting a smaller or a larger SSR than $m_{1/2}$. Results are given in table 1.

In exons, we find that all types of SSRs are smaller than expected (χ^2 test; $P < 10^{-16}$), which agrees with previous studies (Metzgar et al. 2000; Ackermann and Chao 2006). For introns, we find that G-SSR and C-SSR are smaller than expected (χ^2 test; $P < 10^{-16}$), whereas A-SSR and T-SSR are longer than expected (χ^2 test; $P < 10^{-16}$). Interestingly, introns where "Alu" were removed by RepeatMasker (Smit 1999) show the same pattern. Actually, masking Alu reduces the length of intronic sequence and increases the number of sequences that host larger than expected G-SSR or C-SSRs.

There Are Less Hypermutable Genes than Expected in Almost All Functions

If we find as many hypermutable genes (i.e., genes with a mono-SSR of 8 bp or more) as the neutral model predicts, we will have to acknowledge that long mono-SSRs are virtually neutral for these genes. If we find more hypermutable genes than expected, it suggests that mono-SSRs were positively selected in these genes. Indeed, in exons, mono-SSRs are created by the accumulation of substitutions and if they improve the fitness of their host genome, they will be selected for. If we find less long mono-SSR than expected, it suggests that mono-SSRs are removed by purifying selection from the coding sequences.

In all, 1,291 genes (5.8% of the total) contain a mono-SSR of 8 units or more. Using the model, we expect 14.2% of such genes (with a 95% conservative CI of [13.8%, 14.7%]). This again highlights that, on average, there are less long mono-SSR in genes than expected by chance, most likely due to their removal by purifying selection.

We further wanted to test if this trend is shared by all functions taken individually. Therefore, we compared for each term of GO, the expected fraction of genes with long mono-SSR to the expected one. Results are shown in figure 3. Overall, among the functions that have at least 20 genes, 734/1,238 functions (59.3%) exhibit a fraction of hypermutable genes outside the 95% CI that was computed under the neutral model—406/679 (59.8%) in BP, 233/404 (57.7%) in MF, and 95/155 (61.3%) in CC. These functions are colored in blue in figure 3. For all, except one, there are less hypermutable genes than predicted by the neutral model. Taking into account also the terms with less than 20 genes, we observe a lower, though significant, number of terms outside the 95% CI: 788/6,305 terms (12.5%). This demonstrates that for almost all functions, hypermutable genes are removed by purifying selection. Considering mono-SSR of 9 bp or more (instead of 8 bp or more) leads to identical results (supplementary fig. S1, Supplementary Material online).

The functions that we found overrepresented among hypermutable genes (colored in red)—the functions given

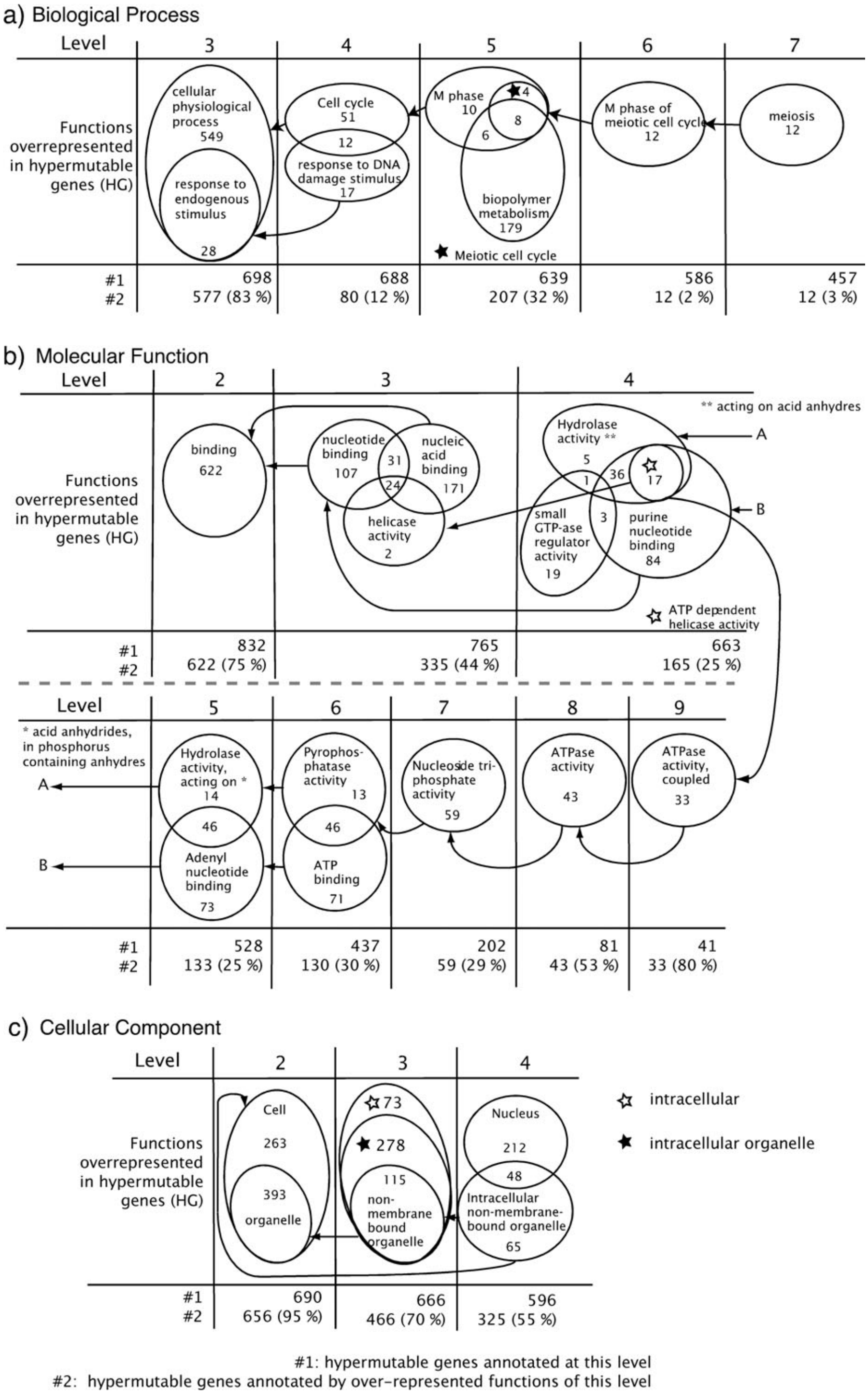
a) Biological Process
Level
3
4
5
6
7
Functions overrepresented in hypermutable genes (HG)
cellular physiological process 549
response to endogenous stimulus
28
Cell cycle 51
12
response to DNA damage stimulus 17
M phase 10
4
8
6
biopolymer metabolism 179
★ Meiotic cell cycle
M phase of meiotic cell cycle 12
meiosis 12
#1
#2
698
577 (83 %)
688
80 (12 %)
639
207 (32 %)
586
12 (2 %)
457
12 (3 %)
b) Molecular Function
Level
2
3
4
** acting on acid anhydres
binding 622
nucleotide binding 107
31
24
nucleic acid binding 171
helicase activity 2
Hydrolase activity **
5
1
36
17
A
B
3
small GTP-ase regulator activity 19
purine nucleotide binding 84
ATP dependent helicase activity
#1
#2
832
622 (75 %)
765
335 (44 %)
663
165 (25 %)
Level
5
6
7
8
9
* acid anhydrides, in phosphorus containing anhydres
A
B
Hydrolase activity, acting on * 14
46
Adenyl nucleotide binding 73
Pyrophos-phatase activity 13
46
ATP binding 71
Nucleoside tri-phosphate activity 59
ATPase activity 43
ATPase activity, coupled 33
#1
#2
528
133 (25 %)
437
130 (30 %)
202
59 (29 %)
81
43 (53 %)
41
33 (80 %)
c) Cellular Component
Level
2
3
4
Functions overrepresented in hypermutable genes (HG)
Cell 263
393 organelle
73
278
115 non-membrane bound organelle
Nucleus 212
48
Intracellular non-membrane-bound organelle 65
intracellular
intracellular organelle
#1
#2
690
656 (95 %)
666
466 (70 %)
596
325 (55 %)
#1: hypermutable genes annotated at this level
#2: hypermutable genes annotated by over–represented functions of this level

FIG. 2.—GO terms overrepresented among hypermutable genes. Here, we report for all three branches of the ontology, the functions we found overrepresented among hypermutable genes in the human genome. Results are given for (*a*) BP, (*b*) MF, and (*c*) CC. Each column is a level in the ontology (the higher the level, the more precise the annotation). It contains ellipses representing overrepresented functions lying at this level. The encapsulated numbers are the numbers of hypermutable genes in these functions. Genes shared by several functions are given in the intersection of ellipses. Arrows indicate a complete inclusion into another term at a shallower adjacent level. We also give, under the picture, the total number of hypermutable genes that is annotated at this level as well as the number among them that is embedded in the functions we found overrepresented.

in figure 2—have a larger observed fraction than the average. They, however, exhibit usually less hypermutable genes than expected from neutrality. This shows that even though we find them overrepresented, hypermutable genes are also avoided in these functions. It is noteworthy to mention that the hypergeometric statistics we used to estimate the overrepresentation among hypermutable genes depends on the number of genes within a term. Therefore, we

Table 1
Mono-SSR Probability in Human Exons and Introns

	Exons		Introns		Alu-masked Introns	
Mono-SSR	Smaller	Larger	Smaller	Larger	Smaller	Larger
A	20,271	1,947	3,441	1,4943	5,045	13,339
T	20,279	1,939	3,254	1,5130	4,458	13,926
G	21,660	558	16,059	2,325	13,651	4,733
C	21,342	976	15,139	3,245	12,484	5,900
Number expected	11,109	11,109	9,192	9,192	9,192	9,192

NOTE.—For each type of mono-SSRs (A, C, G, and T), we compute for each human gene an expected length value ($m_{1/2}$) beyond which there is a 50% chance of finding an SSR of size $m_{1/2}$ or longer. Each gene was then assigned to the larger or the smaller category depending on the comparison of the length of its longest mono-SSR to $m_{1/2}$. If the neutral model were fitting, we would expect half of the genes to host a mono-SSR larger than $m_{1/2}$. This table shows the results for exonic and intronic sequences and for each type of repeat nucleotide. We also examined intronic sequences masked for Alu sequences because their presence in an intron adds A/T repeats to these sequences. Deviation from the expectation (0.5 vs. 0.5) is significant for all types of sequences and mono-SSRs (χ^2 test, $P < 10^{-16}$ for all tests).

observed terms with a high fraction of hypermutable genes that are not significantly overrepresented (e.g., MMR with an observed fraction of 26.1%) and, conversely, terms we found significantly overrepresented even though they exhibit a moderate fraction of hypermutable genes (e.g., "biopolymer metabolism," which has an observed fraction of 7.3%). This latter case happens when the number of genes is very large for a given function, which improves the power of the statistical test we used.

Generally, the comparison between the observed and the expected fraction of hypermutable genes for all terms (fig. 3) reveals a weak though positive correlation between the observed and the expected values ($r = 0.35$ for BP, $r = 0.56$ for MF, and $r = 0.43$ for CC, $P << 10^{-4}$ for all regressions). This implies that typically the presence of long mono-SSR in genes can be partially explained by their length and their nucleotide composition.

The Strength of Purifying Selection Varies from Function to Function

Results highlight interesting functions that appear different from the others. First, we observed some functions with a particularly small observed/expected ratio. The most striking example is the "collagen" term (fig. 3*c*) for which the ratio is 0.075. Even though one would expect a large proportion (40.0%) of hypermutable genes within this term, we found only very few (3.1%). Conversely, there are 36 terms with a ratio observed/expected larger than 1 (e.g., "endoplasmic reticulum to Golgi transport" as well as meiosis and MMR in fig. 3*a*).

This variation could be solely due to the random sampling of genes within functions. Modeling the probability of having long mono-SSR under purifying selection may allow the test for this hypothesis. As a first approximation, we used the observed density of long mono-SSR in coding sequences to compute an average rate of SSR per base. If all genes were under the same selective constraints, the number of SSR per gene should be Poisson distributed with this average rate multiplied by their length. Accordingly, we computed the probability to host at least one long mono-SSR (i.e., to be an hypermutable gene) for all genes. We then computed, for each function, a 95% CI for the expected number of hypermutable genes. Among terms with more than 20 genes, we found 171/1,238 (13.8%) terms outside the CI; this is larger than the 5% we expected if coding mono-SSRs were under the same selective pressure in all functions.

Discussion

In this study, we assumed that all genes hosting a long enough mono-SSR can be considered as hypermutable genes. Whatever the chosen threshold for hypermutability, we show that only a cohesive restricted set of functions are overrepresented among hypermutable genes. Interestingly, we show that this is only due to the mono-SSR within genes, the other type of SSRs being uniformly distributed among functions. Using a probabilistic model, we were able to show that mono-SSRs are shorter than expected by a model of neutral substitution (which is coherent with previous studies, e.g., Metzgar et al. [2000]; Ackermann and Chao [2006]) and that hypermutable genes are avoided in almost all functions. Finally, our study shows that the strength of purifying selection, that removes hypermutable genes from the human genomes, varies greatly from function to function.

SSRs Are Kept Small by Purifying Selection in Exons

The comparison between introns and exons suggests that frameshifting SSRs are subject to a strong purifying selection in coding sequences. Indeed, if one considers that intron evolution is almost neutral, then the length of intronic SSRs must be solely the consequence of their mutation process. The differences observed between length of exonic and intronic SSRs reflect the existence of selection that acts against free expansion of those SSRs in coding sequence.

Indeed, using a model that predicts the size of the longest mono-SSR expected in a coding sequence of a given length and composition, we showed that, in exons, mono-SSR length is globally smaller than expected. In introns, G/C-SSRs are also shorter than expected but A/T-SSRs are usually longer than expected. This is consistent with the observation that G/C-SSRs are generally smaller than A/T-SSRs (Li et al. 2002). This suggests that A/T- and G/C-SSRs should be considered separately. Insertion of Alu sequences in introns contributes to the abundance of long A/T-SSRs but is not sufficient to explain their

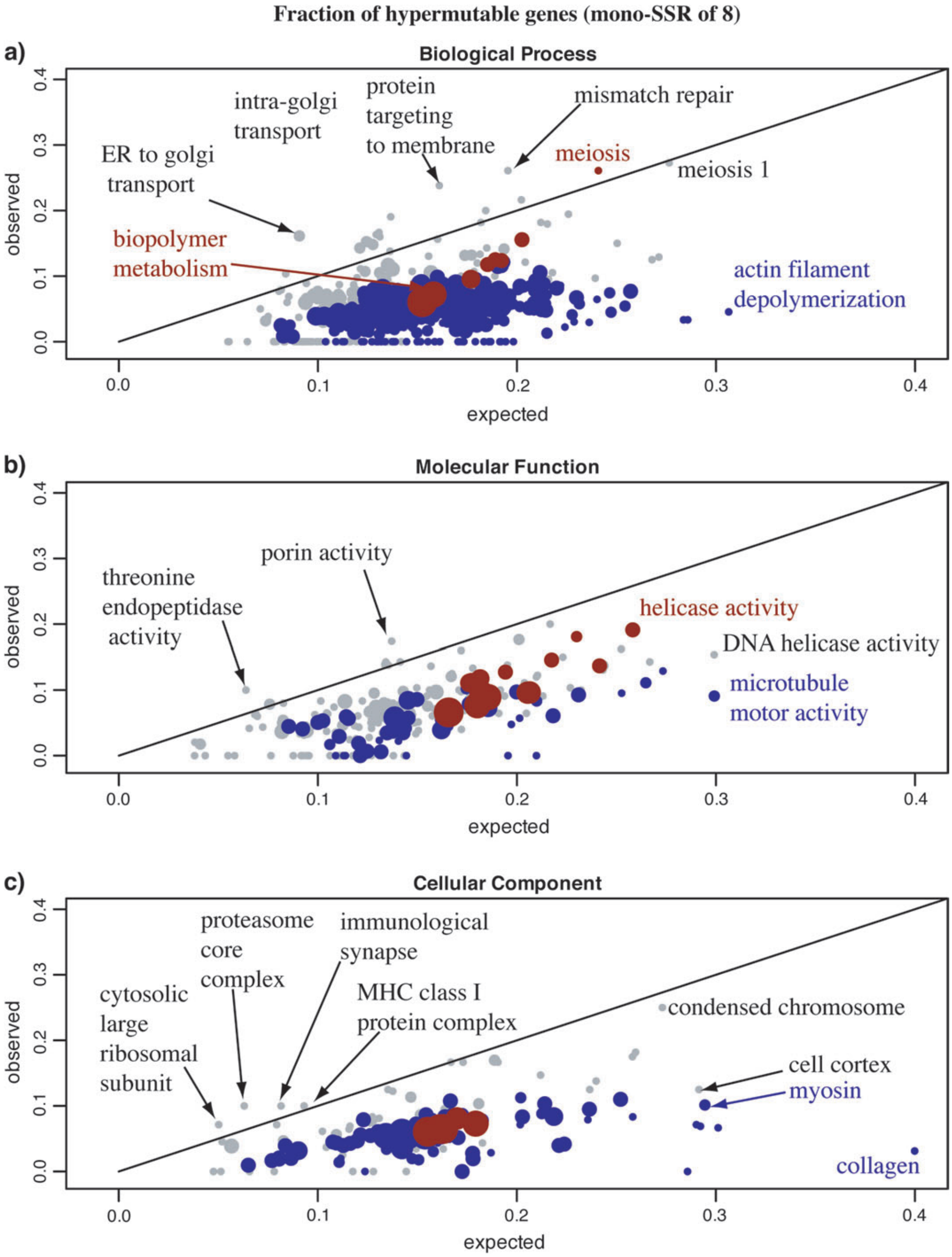

FIG. 3.—Expected and observed fractions of hypermutable genes for all GO terms. Here we represent for each term, the observed proportion of genes that contain a mono-SSR larger than eight as a function of its expected fraction. The terms are extracted from (*a*) BP, (*b*) MF, and (c) CC ontologies. The size of each dot is proportional to the total number of genes that term encompasses (taken as discrete intervals: [20, 50], [50, 100], [100, 500], [500, 10^3], and [10^3, infinity]). Terms with less than 20 genes were not represented. Terms we found statistically overrepresented among hypermutable genes (terms from fig. 2) are colored in red. The line represents the ratio observed/expected = 1. Terms that are significantly outside the 95% CI predicted under neutrality are colored in blue. This figure shows that almost all functions contain less genes carrying a long mono-SSR than expected. This again illustrates that most, if not all, long mono-SSR tends to be removed by purifying selection. Although, it also shows that some functions (e.g., meiosis, MMR, and "condensed chromosome") encompass many genes with long mono-SSR along with an observed/expected ratio close to 1. This suggests that genes involved in those functions are under relaxed purifying selection.

abundance. Because there is no reading frame, one could imagine that A/T-SSRs can undergo free expansion. The model we used as a reference assumes that all SSRs are created by an accumulation of substitutions. Beyond a threshold size, SSRs experience expansions through replication slippage (or recombination) and then become longer than expected. Obviously, there are additional factors that prevent G/C-SSRs to expand. As for coding sequences, we suspect that G/C-SSRs are kept short by purifying selection in introns. Two molecular evidences are compatible with this hypothesis. First, G-rich tracts are known to adopt unusual DNA structure (parallel quadruplex) involved in different biological functions (Sen and Gilbert 1988). Second, G-rich tracts are also prone to electron transfer that causes oxidative damage (Hall et al. 1996). For one or the other (or both) reasons, there is a good chance that G/C-SSRs have an impact on fitness even in introns. This effect should equally apply in exons. Those deleterious effects certainly add up with those previously highlighted (selection against frameshifts).

Functions of Hypermutable Genes

Despite this global underrepresentation of SSRs in exonic sequences, several genes still host a long SSR. Defining a threshold for long SSRs is not trivial. Thus, we used two sets of values that are relevant for the minimum size beyond which SSRs are subject to expansion and contraction. It is important to mention that both sets of thresholds lead to extremely similar results. This highlights the robustness of our results to the choice of a threshold for hypermutability. Among the very large number of terms that were annotated in the human genes, only a restricted number exhibits an overrepresentation of hypermutable genes.

Legendre et al. (2007) conducted a similar analysis on a data set that includes all genes that contain any type of SSR. BP overrepresented among this data set is different from the ones we report here. An analysis of the 1,266 genes hosting a long tri-SSR reveals a similar set of functions (data not shown), with the exception of neurogenesis and related terms. We suspect that the difference in metric for hypermutability explains this difference. Because many neurological disorders are caused by the presence of a coding tri-SSR, we conclude that the overrepresentation of the functions described by Legendre et al. is mainly driven by genes hosting a tri-SSR that we ignored in our study.

Importantly, one could argue that this is a consequence of large duplicate families that share often the same annotations. However, using Ensembl definition of gene family (Enright et al. 2002), we computed for each function the fraction of genes that contain a duplicate within the function. No differences were observed between the overrepresented functions and the others (0.35 vs. 0.41, $P = 0.18$ when considering all genes, 0.25 vs. 0.31, $P = 0.32$ when considering genes with mono-SSR, Mann–Whitney *U* test). Therefore, this overrepresentation is not an artifact of large duplicate families. Our analysis shows that those functions are generally devoted to cell cycle and maintenance of genome integrity (DNA repair, meiosis, cell cycle, helicase domain–containing genes, nuclear localized genes, etc.). It should be mentioned that a similar set of functions is overrepresented among genes that host at least two long mono-SSRs (data not shown). Furthermore, the same analysis with annotations from PantherDB (Mi et al. 2005) also leads to a similar set of functions (data not shown). Overall, we think that our results are robust to the most obvious artifacts and that the restricted cohesive set of functions we find overrepresented in hypermutable genes are meaningful.

The Strength of Purifying Selection against Hypermutable Genes Varies from Function to Function

We computed an expected fraction of hypermutable genes in all functional groups of genes and compared it with the observed fraction. We show that almost all functions clearly harbor less hypermutable genes than expected under neutrality. This strongly suggests that the vast majority of long mono-SSRs are kept out of coding sequences by purifying selection.

Functions overrepresented among the hypermutable genes (i.e., those dedicated to genomic stability and cell cycle) are expected to contain a large fraction of hypermutable genes. They are longer and/or more biased in composition than the average genes. Therefore, the overrepresentation of hypermutable genes in those functions can be explained by the length and the nucleotide composition of genes among those functions. This points out the importance of using a statistical framework that tests for the effect of length and composition of the genes.

An overestimation of the expected number of long mono-SSRs would diminish the strength of the purifying selection we observe. At least three properties of DNA-coding sequences were neglected in our model. First, slippage process was ignored, although almost none is expected in coding sequence. Slippage, however, leads to larger mono-SSR than what is observed in coding sequence. Therefore, ignoring slippage lowers the expected number and size of mono-SSR. Second, we also ignored the dependency of nucleotide context in coding sequences. We estimated the probabilities of mono-SSR in coding sequences using a simulated data set of random sequences modeled by a Markov model of size 2 (using the frequency of the 3-mers). Using these probabilities instead of the one given by the Poisson model does not qualitatively changes our results. Finally, we ignored the amino acid sequences of the genes. Ackermann and Chao (2006) fixed the amino acid sequences of genes and showed that mono-SSRs are underrepresented.

There are few functions for which we observed as many hypermutable genes as expected under a neutral model. For these functions, long mono-SSRs are virtually neutral. On another extreme, we shall consider functions that are expected to contain long mono-SSR but do not (e.g., cytoskeleton- and collagen-related genes). Overall, we have to acknowledge that the strength of the purifying selection that acts against long mono-SSR varies from function to function, from very strong (e.g., for collagen) up to its complete absence (e.g., for ER to Golgi transport). We can propose several hypotheses to explain this observation.

First, the rate of instability for SSR within the same genome may greatly vary from one locus to another. Therefore, we can imagine that the hypermutable genes are located in peculiar loci in the genome where SSRs are stabilized.

Alternatively, it is possible that the functions where mono-SSRs are apparently neutral could be composed of genes that are more "dispensable" than others. Here we used dispensable to refer to a low cost in fitness when the gene is not properly expressed. For the human genome, we however do not have a list of phenotype associated with the absence of all genes. The use of Online Mendelian

Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim/) seems inappropriate because although half of the genes carry an entry, the entries clearly do not have the same meaning in terms of individual fitness and the genes with no annotation cannot be considered as dispensable.

Finally, it is possible that the apparent neutrality of SSR could be the result of a balance between positive and negative selection. If the expression of a gene is associated to sometimes positive, sometimes negative fitness, one could imagine that the evolution of such a gene would look neutral even though it is always under selection. Here, we find that genes that host a long SSR are devoted to the maintenance of DNA integrity. Why did such genes retain hypermutable motifs in their coding sequences? Previous studies (Moxon and Wills 1999; Chang et al. 2001; Rocha et al. 2002; Kashi and King 2006) reported the presence of long mono-SSR in MMR genes and proposed that these genes tune the global mutation rate of the organism by switching on and off after a loss-of-frame mutation caused by replication slippage. Mutator phenotypes, generally caused by a mutated MMR gene (Rosenberg et al. 1998), have been shown to be evolutionary advantageous in bacteria facing an environmental challenge (Taddei et al. 1997). Among a population under stress, individuals with a new advantageous mutation (most likely individuals bearing the mutator allele) will improve in fitness. Thus, this advantageous mutation will increase in frequency along with the mutator (by hitchhiking). If genetic linkage is likely to be strong in bacteria, it is not in eukaryotes. Therefore, the possibility of mutators in the human lineage seems difficult. We can nonetheless intuitively suspect that selection could favor a premutator state (i.e., unstable mono-SSR hosted in coding sequence) in some function (e.g., genes devoted to genomic stability), although it would require more theoretical investigations that will not be conducted here.

It seems difficult at this stage to definitely support or reject one of the hypotheses. However, we would like to mention that the last hypothesis (hidden positive selection) should be regarded with caution. If long mono-SSR looks neutral in these genes, the most parsimonious explanation is that they are neutral.

As a consequence, we do not favor this “oscillating mode of selection” hypothesis and challenge the existence of mutator genes in human and more generally in eukaryotes.

Conclusion

The hypermutability of the human genes (when considering only potentially unstable SSR) is typically a consequence of their length and/or nucleotide composition. Most long SSRs are removed from coding sequence by purifying selection. However, a restricted set of functions seems to be insensitive to the presence of a priori deleterious long SSR. The mystery of this apparent relaxed purifying selection needs more thought and data. In that respect, we think that there is a need for more theory along with a phylogenetic perspective on the evolution of coding SSR to gather further insight in this unclosed debate.

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Authors would like to thank S. Schbath, E. Rocha, I. Gonçalves, S. Baulac, E. Leguern, and C. Castillo-Davis for comments on previous versions of the manuscript.

Literature Cited

Aaltonen LA, Peltomaki P, Leach FS, et al. (15 co-authors). 1993. Clues to the pathogenesis of familial colorectal cancer. Science. 260:812–816.

Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. PLoS Genet. 2:e22.

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 22:1600–1607.

Ashburner M, Ball CA, Blake JA, et al. (17 co-authors). 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 25:25–29.

Castillo-Davis CI, Hartl DL. 2003. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 19:891–892.

Chang DK, Metzgar D, Wills C, Boland CR. 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Res. 11: 1145–1146.

Conti E, Izaurralde E. 2005. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. Curr Opin Cell Biol. 17:316–325.

de Wachter R. 1981. The number of repeats expected in random nucleic acid sequences and found in genes. J Theor Biol. 91: 71–98.

Dieringer D, Schlotterer C. 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13: 2242–2251.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics. 148:1667–1686.

Duval A, Hamelin R. 2003. Replication error repair, microsatellites, and cancer. Med Sci (Paris). 19:55–62.

Ellegren H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet. 24:400–402.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 5:435–445.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Everett CM, Wood NW. 2004. Trinucleotide repeats and neurodegenerative disease. Brain. 127:2385–2405.

Fabre E, Dujon B, Richard GF. 2002. Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. Nucleic Acids Res. 30:3540–3547.

Gragg H, Harfe BD, Jinks-Robertson S. 2002. Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae*. Mol Cell Biol. 22:8756–8762.

Hall DB, Holmlin RE, Barton JK. 1996. Oxidative DNA damage through long-range electron transfer. Nature. 382:731–735.

Jacob S, Praz F. 2002. DNA mismatch repair defects: role in colorectal carcinogenesis. Biochimie. 84:27–47.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22:253–259.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18:30–38.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA. 95:10774–10778.

Lai Y, Sun F. 2003a. Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. J Theor Biol. 224:127–137.

Lai Y, Sun F. 2003b. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 20:2123–2131.

Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. 17:1787–1796.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 4:203–221.

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 11:2453–2465.

Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10:72–80.

Mi H, Lazareva-Ulitsky B, Loo R, et al. (12 co-authors). 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 33:284–288.

Miquel C, Jacob S, Grandjouan S, Aime A, Viguier J, Sabourin JC, Sarasin A, Duval A, Praz F. 2007. Frequent alteration of DNA damage signalling and repair pathways in human colorectal cancers with microsatellite instability. Oncogene. 26:5919–5926.

Mori Y, Yin J, Rashid A, Leggett BA, Young J, Simms L, Kuehl PM, Langenberg P, Meltzer SJ, Stine OC. 2001. Instabilotyping: comprehensive identification of frameshift mutations caused by coding region microsatellite instability. Cancer Res. 61:6046–6049.

Moxon ER, Wills C. 1999. DNA microsatellites: agents of evolution? Sci Am. 280:94–99.

Pupko T, Graur D. 1999. Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. J Mol Evol. 48:313–316.

Rivals I, Personnaz L, Taing L, Potier MC. 2007. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 23:401–407.

Robin S, Rodolphe F, Schbath S. 2005. DNA words and models. Cambridge: Cambridge University Press.

Rocha EPC, Matic I, Taddei F. 2002. Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? Nucleic Acids Res. 30: 1886–1894.

Rose O, Falush D. 1998. A threshold size for microsatellite expansion. Mol Biol Evol. 15:613–615.

Rosenberg SM, Thulin C, Harris RS. 1998. Transient and heritable mutators in adaptive evolution in the lab and in nature. Genetics. 148:1559–1566.

Sagher D, Hsu A, Strauss B. 1999. Stabilization of the intermediate in frameshift mutation. Mutat Res. 423:73–77.

Sen D, Gilbert W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature. 334:364–366.

Shinde D, Lai Y, Sun F, Arnheim N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. 31:974–980.

Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD. 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol. 17:2851–2858.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657–663.

Strand M, Prolla TA, Liskay RM, Petes TD. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature. 365:274–276.

Strauss BS. 1999. Frameshift mutation, microsatellites and mismatch repair. Mutat Res. 437:195–203.

Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. 1997. Role of mutator alleles in adaptive evolution. Nature. 387:700–702.

Tautz D. 1994. Simple sequences. Curr Opin Genet Dev. 4:832–837.

Toth G, Gaspari Z, Jurka J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10: 967–981.

Tran HT, Keen JD, Kricker M, Resnick MA, Gordenin DA. 1997. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Mol Cell Biol. 17:2859–2865.

Woerner SM, Kloor M, von Knebel Doeberitz M, Gebert JF. 2006. Microsatellite instability in the development of DNA mismatch repair deficient tumors. Cancer Biomark. 2:69–86.

Xu X, Peng M, Fang Z. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 24: 396–399.

Naruya Saitou, Associate Editor

Accepted October 5, 2008