

# Unstable DNA Repair Genes Shaped by Their Own Sequence Modifying Phenotypes

Daniel S. Falster · Sigve Nakken · Marie Bergem-Ohr ·  
Einar Andreas Rødland · Jarle Breivik

Received: 21 January 2010 / Accepted: 10 February 2010 / Published online: 6 March 2010  
© The Author(s) 2010. This article is published with open access at Springerlink.com

**Abstract** The question of whether natural selection favors genetic stability or genetic variability is a fundamental problem in evolutionary biology. Bioinformatic analyses demonstrate that selection favors genetic stability by avoiding unstable nucleotide sequences in protein encoding DNA. Yet, such unstable sequences are maintained in several DNA repair genes, thereby promoting

breakdown of repair and destabilizing the genome. Several studies have therefore argued that selection favors genetic variability at the expense of stability. Here we propose a new evolutionary mechanism, with supporting bioinformatic evidence, that resolves this paradox. Combining the concepts of gene-dependent mutation biases and meiotic recombination, we argue that unstable sequences in the DNA mismatch repair (MMR) genes are maintained by their own phenotype. In particular, we predict that human MMR maintains an overrepresentation of mononucleotide repeats (monorepeats) within and around the MMR genes. In support of this hypothesis, we report a 31% excess in monorepeats in 250 kb regions surrounding the seven MMR genes compared to all other RefSeq genes (1.75 vs. 1.34%,  $P = 0.0047$ ), with a particularly high content in PMS2 (2.41%,  $P = 0.0047$ ) and MSH6 (2.07%,  $P = 0.043$ ). Based on a mathematical model of monorepeat frequency, we argue that the proposed mechanism may suffice to explain the observed excess of repeats around MMR genes. Our findings thus indicate that unstable sequences in MMR genes are maintained through evolution by the MMR mechanism. The evolutionary paradox of genetically unstable DNA repair genes may thus be explained by an equilibrium in which the phenotype acts back on its own genotype.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-010-9328-0) contains supplementary material, which is available to authorized users.

D. S. Falster · M. Bergem-Ohr · J. Breivik (✉)  
Institute of Basic Medical Science, University of Oslo,  
P.O. Box 1018 Blindern, 0315 Oslo, Norway  
e-mail: jbreivik@medisin.uio.no

*Present Address:*

D. S. Falster  
Department of Biological Sciences, Macquarie University,  
Sydney, Australia

S. Nakken  
Centre for Molecular Biology and Neuroscience,  
Institute of Medical Microbiology, Rikshospitalet  
University Hospital, 0027 Oslo, Norway

*Present Address:*

S. Nakken  
Bioinformatics Core Facility, Institute of Medical Informatics,  
Rikshospitalet, 0310 Oslo, Norway

E. A. Rødland  
Department of Informatics and Center for Cancer Biomedicine,  
University of Oslo, 0316 Oslo, Norway

E. A. Rødland  
Norwegian Computing Center, 0314 Oslo, Norway

**Keywords** DNA repair · Microsatellites ·  
Genetic instability · Cancer · DNA mismatch repair ·  
Recombination · Mutation bias

## Introduction

DNA mismatch repair (MMR) is an enzymatic mechanism that recognizes and corrects single nucleotide and

insertion–deletion mismatches in DNA (Lyer et al. 2006; Marti et al. 2002). It thereby maintains the overall stability of the genome and is central to the prevention of cancer (Lynch et al. 2006; Peltomaki 2005). MMR is particularly important in stabilizing the length of microsatellites (also known as short tandem repeats or simple sequence repeats), and MMR deficiency is recognized as microsatellite instability throughout the genome (Ellegren 2004). Concurrently, several of the MMR genes, in human and other eukaryotes, contain microsatellites within their own coding sequence (Chang et al. 2001). These monorepeats make MMR genes particularly susceptible to deactivation by frame-shift mutation and a mutational target in cancer development (Venkatesan et al. 2006; Ohmiya et al. 2001; Perucho 1996). Thus, the very genes that protect against genetic instability and cancer are themselves unstable. In this article, we provide a mechanistic explanation for this seeming evolutionary paradox.

Chang et al. (2001) previously proposed that the unstable sequences in the MMR genes have been selected because they provide genetic variability. This idea of selection for variability has been proposed to explain a number of biological phenomena (Kashi and King 2006; Li et al. 2004), but evidence for this interpretation is limited. Other authors have therefore argued that although instability is not selected per se, unstable sequences may spread when linked to other favorable properties (Sniegowski et al. 2000; Baer et al. 2007). In general, however, full genome analyses demonstrate that selection favors stability by avoiding nucleotide repeats in coding sequences (Ackermann and Chao 2006; Wanner et al. 2008). The question thus remains: Why are unstable microsatellites overrepresented in the very MMR genes responsible for maintaining microsatellite stability?

Another relationship between MMR and microsatellites gives hint of a possible solution. Numerous studies show that MMR not only stabilizes microsatellites, but can also induce different types of mutation biases in such sequences (Burt and Trivers 2006; Sleckman 2005; Ellegren 2002; Pearson et al. 2005; Shah et al. 2010). As a primary example, wild-type MSH2 promotes expansion of trinucleotide repeats related to inheritance and progression of neurodegenerative disorders in mouse models (Subramanian et al. 2003; Manley et al. 1999), whereas the homologous gene in *Drosophila melanogaster* (Spell) causes genome-wide contraction of dinucleotide repeats (Harr et al. 2002).

In humans, mutation of MSH2 and other MMR genes is related to the Lynch syndrome (Lynch et al. 2006; Felton et al. 2007). This condition, with an incidence of approximately 1:1000 in the general population (de la Chapelle 2005), is characterized by early development of tumors with microsatellite instability. The affected individual is generally heterozygous, and MMR deficiency arise as a

consequence of somatic inactivation of the normal allele. The instability is particularly evident in monorepeats (Lynch et al. 2006; Peltomaki 2001), and the mutated repeats show a strong overrepresentation (89%) of contractions (Sammalkorpi et al. 2007; Zhou et al. 1997), implying that MMR proficiency maintains the length and stability of monorepeats.

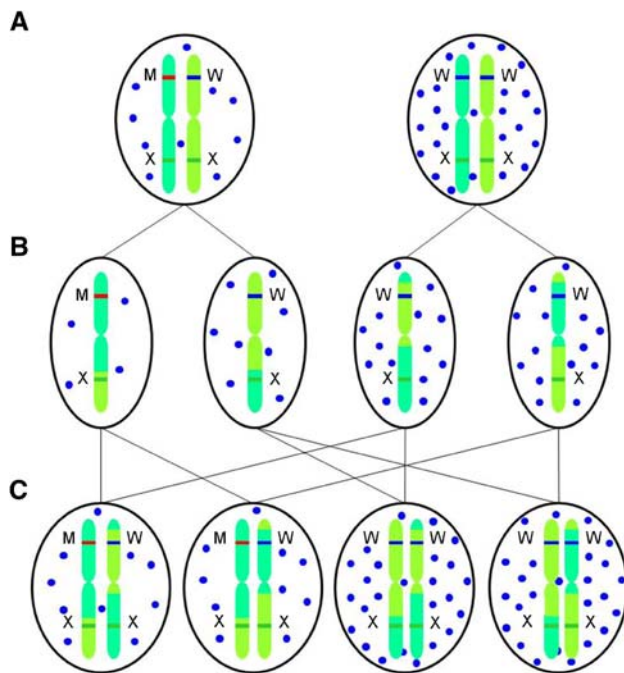
Microsatellite instability in Lynch syndrome is generally confined to the tumor cells, and little is known about the effect of MMR mutations through the human germline. Still, evidence from animal studies and cell lines, show that even heterozygous MMR mutations may produce an increase in mutation rate (Zhang et al. 2002; Alazzouzi et al. 2005; Bouffler et al. 2000), and such haploinsufficiency has also been detected in the germline (Larson et al. 2004; Gurtu et al. 2002; Baida et al. 2003).

Summing up, there are two different connections between MMR and monorepeats. First, several of the MMR genes are destabilized by monorepeats within their own coding regions (Chang et al. 2001). Second, MMR activity introduces a mutation bias that maintains the length and stability of monorepeats in somatic cells, and probably also through the germline. These observations led us to propose a mechanism that links these two phenomena. More specifically, we predict that the paradoxical occurrence of unstable monorepeats within the MMR genotypes is maintained by the mutation bias of the MMR phenotype.

### Proposed Evolutionary Mechanism

The evidence summarized above indicates that the length of monorepeats is determined by a dynamic balance between expansion and contraction of repeat sequences, and that this equilibrium is influenced by different MMR phenotypes. Specifically, it suggests that the homozygous wild-type maintains the length and stability of long monorepeats, whereas the heterozygous mutant show a tendency for contraction due to haploinsufficiency.

For a random region of the genome, rearranged with new MMR alleles every generation, the state of equilibrium will be determined by the relative strength and frequency of the different MMR phenotypes in the population. For a wild-type MMR allele itself, however, this point of equilibrium will be shifted toward expansion. The reason may be illustrated by a Mendelian crossing scheme (Fig. 1). In brief, due to meiotic recombination through the course of evolution, an MMR allele will be more exposed to its own phenotype than to the phenotypes of the alternative alleles. Accordingly, an allele whose phenotype promotes a particular composition of nucleotides should in general contain more of such sequence elements than other sequences of the genome.



**Fig. 1** Proposed mechanism by which an MMR protein (*blue dots*) selectively affects its own genotype. To illustrate the evolutionary dynamics we regard the crossing between a homozygous wild-type, W/W, and a heterozygous mutant, W/M (A). The W/W phenotype maintains the length and stability of monorepeats, whereas the insufficient phenotype (W/M) leads to contraction of these sequences. Regarding possible offspring (C), a random allele in the genome, X, is exposed to the insufficient phenotype in 4 of 8 cases (50%), whereas the W allele is exposed to this phenotype in 2 of 6 cases (33%). Regarding the haploid gametes (B), the W allele is physically separated from the M allele and may involve a differentiated mutagenic effect in the early stages of development. Combined, these effects of meiotic recombination suggest that an allele should be more influenced by its own phenotype than by the phenotype of alternative alleles. Or more specifically, a wild-type MMR allele should maintain longer monorepeats than other regions of the genome (Color figure online)

From this deduction we thus made the following predictions: (1) Wild-type MMR alleles, which maintain the stability of monorepeats, should have more monorepeats than other regions of the genome; (2) This effect should be seen throughout the haplotype block (McVean et al. 2004), not just as individual repeats in coding sequences (Chang et al. 2001); and (3) The amount of repeats in an MMR allele should correlate to the strength and frequency of its mutator phenotype (Marti et al. 2002).

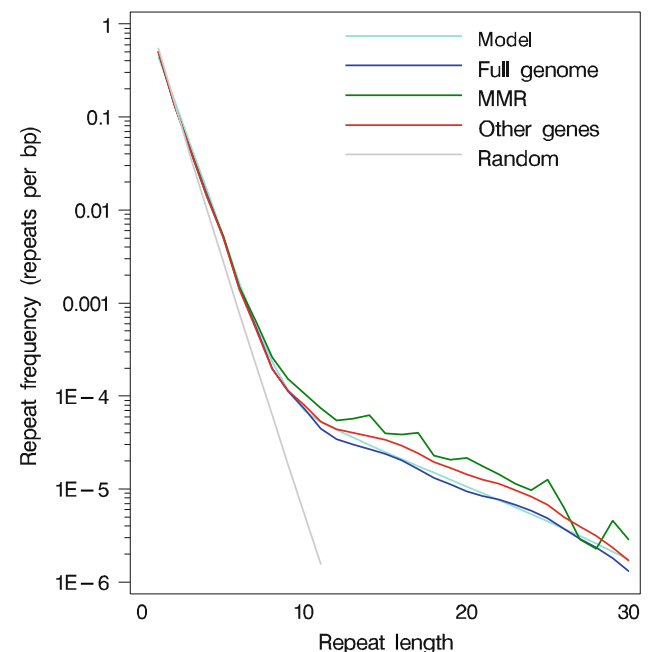
### Sequence Analysis

To test the hypotheses outlined above we performed a complete mapping of monorepeats in the human genome. Sequence data comprising 21,958 defined RefSeq gene sequences (hg19, NCBI Build 37.1) were analyzed for

monorepeats. The MMR system was defined by the seven genes *MSH2*, *MSH3*, *MSH6*, *PMS1*, *PMS2*, *MLH1*, and *MLH3* (Marti et al. 2002). Comparisons were made between standardized genomic regions of 250 kb centered to the defined gene sequences, thus spanning the average length of haplotype blocks in the human genome, which is approximately 200 kb (McVean et al. 2004).

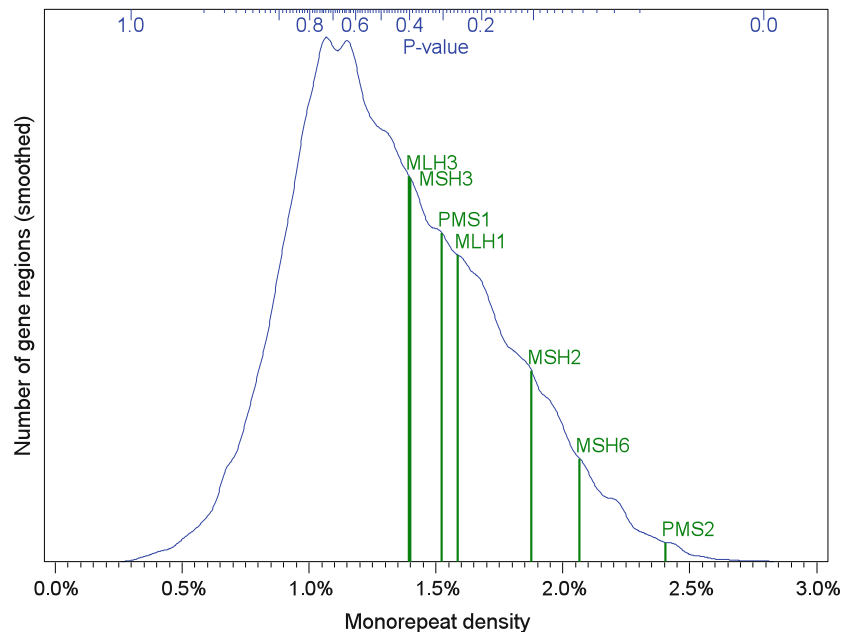
The dataset confirmed previous reports that monorepeats are overrepresented in the human genome compared to expectations based on random nucleotide sequences with similar base compositions (Subramanian et al. 2003; Borstnik and Pumpernik 2002). In particular, there was a marked deviation for long repeat lengths, starting from about 7 bp (Fig. 2). This pattern of deviation was matched by the 250 kb regions for all genes and for those comprising the MMR genes. The observed pattern is also consistent with experimental studies showing that there exists a threshold length about which monorepeats become intrinsically unstable and subject to the stabilizing effect of MMR (Lai and Sun 2003). Therefore, we considered only repeats of length 7 bp or longer in subsequent analyses.

To test for differences in the cumulative number of repeats among sequences, we calculated the proportion of



**Fig. 2** Frequency of monorepeats in MMR genes and the genome. The frequency of monorepeats of increasing length was predicted based on the assumption of random distribution of nucleotides (*gray line*) (Borstnik and Pumpernik 2002), as well as, the presented mathematical model (*light blue line*). These predictions were then plotted against the observed frequency in the full genome (*blue line*), MMR gene regions (*green line*), and all other 250 kb gene regions (*red line*). MMR gene regions show a general excess of repeat lengths of 7 bp and longer compared to all other gene regions and the genome in general (Color figure online)

**Fig. 3** Distribution of repeat content. The graph illustrates the distribution of all 250 kb gene regions relative to their content of monorepeat (7 bp and longer). Positions of the seven MMR regions are indicated. Top scale represents *P*-values for the distribution. The *PMS2* and *MSH6* regions each had a significant overrepresentation of repeats. All seven regions had above median repeat content and scored significantly as a group (Table 1)



the 250 kb gene regions made up of monorepeats (hereafter called repeat content, %) and compared the repeat content of MMR regions to the remaining gene regions. Repeat content varied greatly with respect to chromosome position (supporting information, Fig. S1) and showed a non-normal distribution (Fig. 3). Accordingly, statistical comparison of monorepeats between MMR and other gene regions were performed using Wilcoxon rank-sum test (one-sided,  $\alpha = 0.05$ ).

The primary results are summarized in Table 1. Combined, the MMR regions had a 31% higher content of monorepeats than other gene regions (1.75 vs. 1.34%,  $P = 0.0047$ ), with the excess of repeats distributed evenly across repeat lengths (Fig. 2). The seven MMR regions varied in repeat content from 1.39 to 2.41%. Two of the MMR regions differed significantly from the other gene regions when analyzed individually, *PMS2* (2.41%,  $P = 0.0047$ ) and *MSH6* (2.07%,  $P = 0.043$ ). All MMR regions scored above median repeat content (Fig. 3).

An excess of monorepeats in MMR coding sequences has previously been reported (Chang et al. 2001). Our results confirmed these findings, with a repeat content of 0.26% in protein coding parts of the 250 kb in MMR regions compared to 0.13% for other genes. Still, coding sequences had a lower repeat content than the non-coding sequences (0.26 vs. 1.79% for MMR regions, 0.13 vs. 1.38% in other gene regions) and contributed only 0.39% of the monorepeats in the 250 kb regions around the MMR genes. The contribution of the protein coding repeats, known prior to our analysis (Chang et al. 2001), was thus negligible for the overall repeat content of the MMR regions.

### Analyses of Potential Confounding Factors

We found that monorepeat density varied between chromosomes ( $P < 0.0001$ , Kruskal–Wallis test). Moreover, we found that it was correlated (using Spearman correlation) with the GC content of the region (corr = 0.13), the fraction of region that was protein coding (corr = 0.26) and the level of gene expression (only available for 71% of genes; corr = 0.20), all highly significant ( $P < 0.0001$ ). There was also a weak correlation to codon bias (corr =  $-0.012$ ,  $P = 0.068$ ).

In order to check if these factors could explain the observed density of monorepeats within and around the MMR genes, we applied a general linear model. Because repeat density had a slightly skewed distribution, we ran these analyses on the square root of the repeat density, which was less skewed. We then fitted a linear model using the above listed factors, with log-transformed gene expression values. Since we only had gene expression data for 71% of the genes, we first did the analyses without accounting for gene expression level, then an additional analysis including this factor.

The residuals from these analyses, i.e., the difference between the actual value and the value predicted by the linear model, were used as a measure of over- or underrepresentation of monorepeats corrected for chromosome differences and correlations. Wilcoxon analyses were then performed on these residuals comparing the MMR regions against the remaining.

The GLM model, with all factors included except gene expression level, explained 11.0% of the variance in repeat density, strengthening the difference between MMR

**Table 1** Characteristics of MMR and other genes

Gene regions	Genomic location	Repeat content <sup>a</sup>	GC content <sup>b</sup>	Coding content <sup>c</sup>	Codon bias <sup>d</sup>	Expression <sup>e</sup>
<i>MSH2</i>	2p22-p21	1.88 ( $P = 0.10$ )	45.16	1.86	0.56	290.30
<i>MSH3</i>	5q11-q12	1.40 ( $P = 0.40$ )	38.36	1.37	0.56	401.65
<i>MSH6</i>	2p16	2.07 ( $P = 0.043$ )	42.39	2.46	0.57	2016.15
<i>PMS1</i>	2q31.1	1.52 ( $P = 0.30$ )	36.92	2.58	0.54	2517.20
<i>PMS2</i>	7p22.2	2.41 ( $P = 0.0047$ )	46.52	4.11	0.57	62.35
<i>MLH1</i>	3p21.3	1.59 ( $P = 0.26$ )	40.96	2.47	0.57	1883.90
<i>MLH3</i>	14q.24	1.39 ( $P = 0.40$ )	45.37	4.90	0.55	133.90
All MMR		1.75 ( $P = 0.0047$ )	42.24	2.82	0.56	1043.64
All other genes		1.34	44.75	2.96	0.60	1259.61

<sup>a,b,c</sup> Repeat content, GC content, and coding content are given as percentages.  $P$ -values were computed using a one-sided Wilcoxon rank-sum test

<sup>d</sup> Codon bias was computed using the *B measure* (Karlin et al. 1998)

<sup>e</sup> Gene expression data from testis germ cells were collected from Gene Atlas v2 (Su et al. 2004), and are given as gcRMA-condensed intensities

regions and control regions slightly (to  $P = 0.0046$ ). When gene expression levels were included, all seven MMR genes, but only 71% of the other genes could be included in the analyses. This increased the explained variance to 14.0% and weakened the difference between MMR regions and control regions somewhat (to  $P = 0.0102$ ). However, even when controlling for the effects of confounding factors, the differences between MMR genes and the remainder of the genome remained statistically significant. Thus, we may conclude that these factors, although contributing somewhat to observed differences, cannot explain the differences in repeat content between MMR genes and the rest of the genome. Further details are available as Supplementary Information.

### Mathematical Model of Monorepeat Frequency

Our bioinformatic analyses support the hypothesis that differential exposure of MMR and other genes to MMR activity has led to differences in repeat content. In this section, we consider what size difference in expansion and contraction mutation rates are needed to explain these differences.

To assess the impact of varying mutation rate on repeat content, we modelled a stochastic process describing the evolution of repeat content due to slippage and point mutations. Our approach is based on the model presented by Lai and Sun (2003), which describes the effects of slippage mutation (contractions and expansions) on equilibrium repeat frequency. However, their model only treats the evolution of repeats after they have arisen, not the processes by which short repeats are created by point mutations. We therefore extended their model to include the processes by which point mutations maintain a

background frequency of short monorepeats such as that expected in a purely random sequence.

The model is described in brief here; a full mathematical description is given in Supplementary Information. The genome was considered as a sequence of monorepeats and repeat evolution modeled as a stochastic process. The ordering of monorepeats was not modeled explicitly, only the frequency of repeats of different length. Repeat frequencies are influenced by point and slippage mutations, which extend, contract, join, or split existing repeats. Slippage mutations were assumed to expand or contract existing repeats by a single nucleotide, with mutation rates for expansion and contraction mutation increasing exponentially with repeat length. The effect of point mutations depends on their location within a repeat: point mutations can split an existing repeat, extend an existing repeat by a single base pair, or by join nearby repeats of similar type. The effects of slippage and point mutations combine to give transition probabilities for each repeat length. To simplify the dynamics, we assumed that sizes of neighboring repeats were independent. We then solved for the equilibrium length distribution (see Supporting Information for more details).

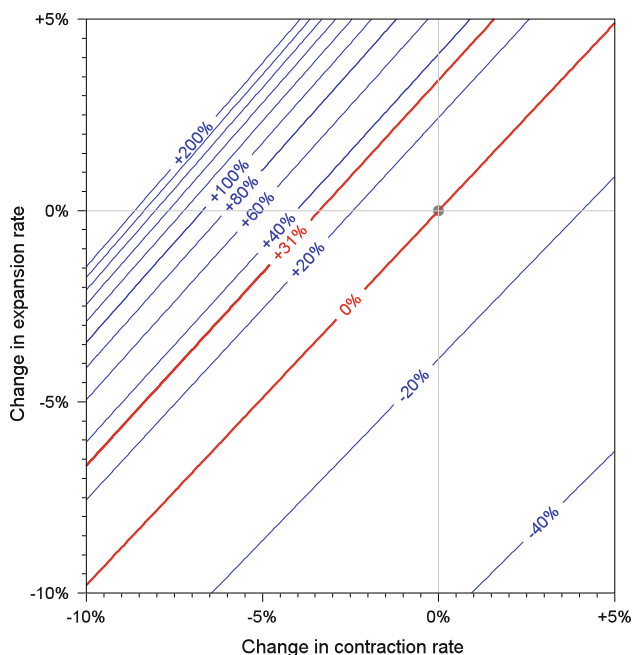
With relatively few parameters, the model described gave a good fit to the observed repeat distribution in the whole genome for repeats of length 2–30 bp (Fig. 2). To achieve this fit, we used a combination of observed mutation spectra and empirical fitting. The frequency of short repeats (2–5 bp) was influenced primarily by the probability that a point mutation extends a neighboring repeat sequence. This parameter was empirically fitted to match the observed repeat distribution. Based on data from Kelkar et al. (2008), the slippage mutation rate was set to increase exponentially with repeat length, starting at approximately 1000 times the point mutation ratio for



11-repeats and increasing by a factor 10 for every 15 nucleotides of length added. The ratio of expansion to contraction was adjusted to fit the observed repeat distribution. In order to get a reasonable fit for repeats of intermediate length, a correction term was needed to reduce the slippage mutation rate for repeats of less than 11 bp.

To explore influence of different levels of MMR activity on repeat content, we varied expansion and contraction rates across a range of values around the fitted values and assessed the effect on repeat content of the genome (Fig. 4). These adjustments represent possible effects of going from the general mutation rates experienced by the genome, to the mutation rates experienced by proficient MMR alleles. The results from the model indicate that small changes in rate of contraction mutation can alter mean repeat content in line with observed data. In particular, a 31% increase in repeat content, as observed in the MMR regions, might be explained by as little as a 3.4% reduction in the contraction frequency. An 81% increase in repeat content, as observed in the PMS2 region, requires only a 6.1% reduction in contraction frequency.

If MMR activity reduces expansion as well as contraction mutations, then a proportionately larger effect on contractions is needed to generate the observed repeat content. For example, if 89% of the slippage mutations



**Fig. 4** Influence of expansion and contraction mutation rates on equilibrium repeat content predicted from stochastic model of repeat evolution. The contours show the change in repeat content (7 bp and longer) when contraction rates (*X* axis) and expansion rates (*Y* axis) are modified. The 31% change contour corresponds to the difference between MMR genes and other genes

caused by a defective MMR allele are contractions (Sammalkorpi et al. 2007), a 3.8% reduction in contraction rates and 0.5% reduction in expansion rate will again give 31% increase in repeat content. Similarly, increasing the rate of contraction mutation (as occurs in MMR deficient cells) caused a decrease in repeat content, as occurs in genetically unstable tumors and cell lineages.

Dunlop et al. (2000) have estimated the carrier frequency of MLH1 and MSH2 mutations to approximately 1:3139. Based on the approximate 1:1000 incidence of Lynch syndrome (de la Chapelle 2005), of which 40% are related to MSH2 (Peltomaki 2005) with a penetrance of 54% (Choi et al. 2009), we estimate the carrier frequency of mutated MSH2 to 1:1350 and the allele frequency to 1:2700. In order to get an overall increase of 3.4%, the mutated alleles must then increase the contraction rates  $\sim 100$ -fold ( $2700 \times 0.034 = 91.8$ ) to explain the observed differences in repeat content. Note that these numbers are very approximate, and merely serve to indicate the order of magnitude.

## Discussion

Combining gene-dependent mutation biases with Mendelian inheritance (Fig. 1), we have deduced that an allele should be more affected by its own mutation bias than should other sequences of the genome. In particular, we predicted that the stabilizing effect of MMR on monorepeats has promoted an excess of such repeats within the MMR haplotype blocks. Confirming this prediction, we found a general expansion of monorepeats in 250 kb regions surrounding the MMR genes. This finding was based on a conservative statistical assessment controlling for the overrepresentation and uneven distribution of monorepeats in the genome. Furthermore, controlling for covariation of repeat density with protein coding content, GC content, codon bias or level of expression did not have significant influence on the results. The evolutionary dynamic proposed thus provides a novel explanation for the prevalence of unstable sequences in several MMR genes.

In accordance with previous analyses (Subramanian et al. 2003), we found a general overrepresentation of monorepeats longer than 7 bp in the human genome (Fig. 2), indicating a mechanism that promotes such sequences through the course of evolution. The same pattern was mirrored in the MMR regions, suggesting that the 31% excess of monorepeats is caused by the same mechanism that promotes such sequences throughout the genome. The statistical analysis and the pattern of repeat lengths thus support our hypothesis that the MMR proteins promote expansion of monorepeats in the human germline, and that this effect is particularly strong within and around their own nucleotide sequence.

Looking at the individual MMR regions, the highest content of monorepeats was found for *PMS2* and *MSH6*, followed by *MSH2* and *MLH*. These four genes cooperate in the recognition of small DNA loops that frequently arise in monorepeats during DNA replication (Lyer et al. 2006; Marti et al. 2002). Correspondingly, loss of function of any of these genes has been related to a particularly high degree of instability in monorepeats, whereas the other MMR genes have a limited effect (Lyer et al. 2006; Marti et al. 2002). *MLH1*, *MSH2*, *MSH6*, and *PMS2* are also the genes of which mutated alleles are related to the Lynch syndrome (Lynch et al. 2006), with an incidence of 1:1000 in the general population. Moreover, all four genes are expressed in oocytes and embryos of rhesus monkeys (Zheng et al. 2005), indicating a key function also in the human germline (Jaroudi and SenGupta 2007). In line with our predictions, we thus found that the MMR genes, which reportedly have the strongest effect on monorepeat stability, also contain the largest amount of such sequences. These findings contrast the conclusion of Chang et al. that monorepeats are particularly related to the “minor” components of MMR (Chang et al. 2001).

Our hypothesis also predicts that mutated MMR alleles should experience their own contraction bias more often than other regions of the genome. This effect of MMR deficiency has been extensively demonstrated in cancer cells (Sammalkorpi et al. 2007). In particular, MMR deficiencies have been directly related to contractions of the *BAT-26* microsatellite marker (also a monorepeat) located within *MSH2* (Boyer et al. 2002; de Leeuw et al. 2001; Zhou et al. 1997; Hoang et al. 1997). However, as homozygous and heterozygous germline mutations in MMR involve strong risk for early cancer, such alleles are probably short-lived in the population (Desai et al. 2000; Sun et al. 2005; Felton et al. 2007). A germline effect of the contraction bias on deficient MMR alleles may thus be hard to detect and has not been tested for in this study, as full genomic sequences of mutated MMR alleles are presently unavailable.

Chang et al. (2001) have argued that “the exceptional density of microsatellites in the minor MMR genes represents a genetic switch that allows the adaptive mutation rate to be modulated over evolutionary time.” This hypothesis cannot explain the excess in monorepeats in non-coding regions within and around MMR genes, several of which have a major role in the prevention of genetic instability and cancer. Nor can it explain the striking association between the mutation bias of the MMR phenotype and repeat content in the MMR genotype. Based on the proposed evolutionary mechanism, we therefore argue that the overrepresentation of monorepeats within and around the MMR genes is maintained by the MMR mechanism.

The population frequency of MMR deficient alleles, including complete as well as partial loss of function, is unknown as we generally only recognize the polymorphisms that cause disease. Nor do we know the effect of human MMR on the germline mutation rate. However, based on the presented model, we argue that the high repeat content in MMR regions may be explained by less than 100-fold difference in microsatellite mutation rate between the MMR wild-type and the heterozygous mutant. This level of instability is in the lower range of that observed in MMR deficient tumors (Lynch et al. 2006; Sammalkorpi et al. 2007) and in the germline of MMR deficient and insufficient mice (Larson et al. 2004; Gurtu et al. 2002).

Most interestingly, the study by Larson et al. (2004) suggests that embryos formed from *PMS2*-deficient eggs have a strong increase in monorepeat mutation rate limited to the earliest stages of development. Heterozygous MMR mutations may thus have significant effect on germline mutation rate, even though the resulting offspring is phenotypically normal. It is therefore interesting to speculate that a similar maternal effect occurs in the human germline.

Moreover, the proposed evolutionary mechanism might be related to the phenomenon of genetic anticipation in Lynch syndrome, i.e., the observation that the disease occurs at an earlier age in successive generations (Nilbert et al. 2009). As the MMR proteins maintain the length of monorepeats within their own nucleotide sequences, they establish a network of self-sustaining loops propagating through the generations. Although the high content of monorepeats makes the MMR genes vulnerable to MMR deficiency, the interdependency of gene and protein may be understood as a stable evolutionary strategy. When a loop is broken, however, it triggers a cascade of events leading to accumulated breakdown of the regulatory network and increasing cancer risk through the generations.

In conclusion, we demonstrate an overrepresentation of monorepeats within and around the MMR genes, and provide an evolutionary and mechanistic explanation to this paradox. In brief, we argue that the MMR proteins have shaped the sequence composition of their own alleles. This concept challenges the dogma that flow of information is unidirectional from DNA to protein (Thieffry and Sarkar 1998; Crick 1970), but is based on simple deduction from well-established molecular mechanisms. In theory, the concept is applicable to any protein that either directly or indirectly affects the nucleotide composition. Other DNA repair genes may also induce mutation biases leading to accumulation of particular sequences within the genome (Pearson et al. 2005; Burt and Trivers 2006). Further testing of the hypothesis will thus require a systematic mapping of sequence-modifying phenotypes and their respective genotypes.

**Acknowledgments** J.B. conceived and developed the theoretical model, initiated the project, interpreted results, and wrote the manuscript. D.S.F. and E.A.R. developed the theoretical model and the methodology for testing it, performed bioinformatics analysis, interpreted results and wrote the manuscript. E.A.R. developed the mathematical model for monorepeat evolution. M.B.-O. contributed to developing the theoretical model and interpreted results. S.N. developed methodology and performed bioinformatic analysis. All authors discussed the results and commented on the manuscript. We thank Andrés Ögmundsson for technical assistance and Eivind Hovig for insightful comments on the manuscript.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Ackermann M, Chao L (2006) DNA sequences shaped by selection for stability. *PLoS Genet* 2:224–230
- Alazzouzi H, Domingo E, Gonzalez S, Blanco I, Armengol M, Espin E, Plaja A, Schwartz S, Capella G, Schwartz SJ (2005) Low levels of microsatellite instability characterize MLH1 and MSH2 HNPCC carriers before tumor diagnosis. *Hum Mol Genet* 14:235–239
- Baer CF, Miyamoto MM, Denver DR (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet* 8:619–631
- Baida A, Lopez A, Marcos R, Velazquez A (2003) Germline mutations at microsatellite loci in homozygous and heterozygous mutants for mismatch repair and PCNA genes in *Drosophila*. *DNA Repair* 2:827–833
- Borstnik B, Pumpernik D (2002) Tandem repeats in protein coding regions of primate genes. *Genome Res* 12:909–915
- Bouffler SD, Hofland N, Cox R, Fodde R (2000) Evidence for Msh2 haploinsufficiency in mice revealed by MNU-induced sister-chromatid exchange analysis. *Br J Cancer* 83:1291–1294
- Boyer JC, Yamada NA, Roques CN, Hatch SB, Riess K, Farber RA (2002) Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. *Hum Mol Genet* 11:707–713
- Burt A, Trivers R (2006) *Genes in conflict: the biology of selfish genetic elements*. Belknap Press, Cambridge
- 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
- Choi YH, Cotterchio M, McKeown-Eyssen G, Neerav M, Bapat B, Boyd K, Gallinger S, McLaughlin J, Aronson M, Briollais L (2009) Penetrance of colorectal cancer among MLH1/MSH2 carriers participating in the colorectal cancer familial registry in Ontario. *Hered Cancer Clin Pract* 7
- Crick F (1970) Central dogma of molecular biology. *Nature* 227:561–563
- de la Chapelle A (2005) The incidence of Lynch syndrome. *Fam Cancer* 4:233–237
- de Leeuw WJF, van Puijenbroek M, Merx R, Wijnen JT, Brocker-Vriends AHJT, Tops C, Vasen H, Cornelisse CJ, Morreau H (2001) Bias in detection of instability of the (C)8 mononucleotide repeat of MSH6 in tumours from HNPCC patients. *Oncogene* 20:6241–6244
- Desai DC, Lockman JC, Chadwick RB, Gao X, Percesepe A, Evans DGR, Miyaki M, Yuen ST, Radice P, Maher ER, Wright FA, de la Chapelle A (2000) Recurrent germline mutation in MSH2 arises frequently de novo. *J Med Genet* 37:646–652
- Dunlop MG, Farrington SM, Nicholl I, Aaltonen L, Petersen G, Porteous M, Carothers A (2000) Population carrier frequency of hMSH2 and hMLH1 mutations. *Br J Cancer* 83:1643–1645
- Ellegren H (2002) Mismatch repair and mutational bias in microsatellite DNA. *Trends Genet* 18:552
- Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5:435–445
- Felton KEA, Gilchrist DM, Andrew SE (2007) Constitutive deficiency in DNA mismatch repair. *Clin Genet* 71:483–498
- Gurtu VE, Verma S, Grossmann AH, Liskay RM, Skarnes WC, Baker SM (2002) Maternal effect for DNA mismatch repair in the mouse. *Genetics* 160:271–277
- Harr B, Todorova J, Schlotterer C (2002) Mismatch repair-driven mutational bias in *D. melanogaster*. *Mol Cell* 10:199–205
- Hoang JM, Cottu PH, Thuille B, Salmon RJ, Thomas G, Hamelin R (1997) *BAT-26*, an indicator of the replication error phenotype in colorectal cancers and cell lines. *Cancer Res* 57:300–303
- Jaroudi S, SenGupta S (2007) DNA repair in mammalian embryos. *Mutat Res* 635:53–77
- Karlin S, Mrazek J, Campbell AM (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol Microbiol* 29:1341–1355
- 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
- Lai Y, Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol Biol Evol* 20:2123–2131
- Larson JS, Stringer SL, Stringer JR (2004) Impact of mismatch repair deficiency on genomic stability in the maternal germline and during early embryonic development. *Mutat Res* 556:45–53
- Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 21:991–1007
- Lyer RR, Pluciennik A, Burdett V, Modrich PL (2006) DNA mismatch repair: functions and mechanisms. *Chem Rev* 106:302–323
- Lynch HT, Boland CR, Gong G, Shaw TG, Lynch PM, Fodde R, Lynch JF, de la CA (2006) Phenotypic and genotypic heterogeneity in the Lynch syndrome: diagnostic, surveillance and management implications. *Eur J Hum Genet* 14:390–402
- Manley K, Shirley TL, Flaherty L, Messer A (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet* 23:471–473
- Marti TM, Kunz C, Fleck O (2002) DNA mismatch repair and mutation avoidance pathways. *J Cell Physiol* 191:28–41
- McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584
- Nilbert M, Timshel S, Bernstein I, Larsen K (2009) Role for genetic anticipation in Lynch syndrome. *J Clin Oncol* 27:360–364
- Ohmiya N, Matsumoto S, Yamamoto H, Baranovskaya S, Malkhosyan SR, Perucho M (2001) Germline and somatic mutations in *hMSH6* and *hMSH3* in gastrointestinal cancers of the microsatellite mutator phenotype. *Gene* 272:301–313
- Pearson CE, Edamura KN, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* 6:729–742
- Peltomaki P (2001) DNA mismatch repair and cancer. *Mutat Res* 488:77–85
- Peltomaki P (2005) Lynch syndrome genes. *Fam Cancer* 4:227–232



- Perucho M (1996) Microsatellite instability: the mutator that mutates the other mutator. *Nat Med* 2:630–631
- Sammalkorpi H, Alhopuro P, Lehtonen R, Tuimala J, Mecklin JP, Jarvinen HJ, Jiricny J, Karhu A, Aaltonen LA (2007) Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer Res* 67:5691–5698
- Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res* 70:431–435
- Sleckman BR (2005) Lymphocyte antigen receptor gene assembly—multiple layers of regulation. *Immunol Res* 32:253–258
- Sniegowski PD, Gerrish PJ, Johnson T, Shaver A (2000) The evolution of mutation rates: separating causes from consequences. *Bioessays* 22:1057–1066
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101:6062–6067
- Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* 4:R13
- Sun S, Greenwood CMT, Thiffault I, Hamel N, Chong G, Foulkes WD (2005) The HNPCC associated MSH2\*1906G>C founder mutation probably originated between 1440 CE and 1715 CE in the Ashkenazi Jewish population. *J Med Genet* 42:766–768
- Thieffry D, Sarkar S (1998) Forty years under the central dogma. *Trends Biochem Sci* 23:312–316
- Venkatesan RN, Bielas JH, Loeb LA (2006) Generation of mutator mutants during carcinogenesis. *DNA Repair* 5:294–302
- Wanner RM, Guthlein C, Springer B, Bottger EC, Ackermann M (2008) Stabilization of the genome of the mismatch repair deficient *Mycobacterium tuberculosis* by context-dependent codon choice. *BMC Genom* 9:294
- Zhang SL, Lloyd R, Bowden G, Glickman BW, de Boer JG (2002) Msh2 deficiency increases the mutation frequency in all parts of the mouse colon. *Environ Mol Mutagen* 40:243–250
- Zheng P, Schramm RD, Latham KE (2005) Developmental regulation and in vitro culture effects on expression of DNA repair and cell cycle checkpoint control genes in rhesus monkey oocytes and embryos. *Biol Reprod* 72:1359–1369
- Zhou XP, Hoang JM, Cottu P, Thomas G, Hamelin R (1997) Allelic profiles of mononucleotide repeat microsatellites in control individuals and in colorectal tumors with and without replication errors. *Oncogene* 15:1713–1718