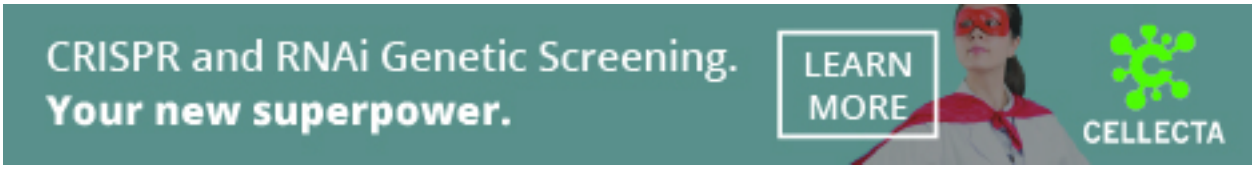
Genome-wide evidence for selection acting on single amino acid repeats

Wilfried Haerty and G. Brian Golding

Genome Res. 2010 20: 755-760 originally published online January 7, 2010
Access the most recent version at doi:10.1101/gr.101246.109



Genome-wide evidence for selection acting on single amino acid repeats

Wilfried Haerty and G. Brian Golding[1]

Biology Department, McMaster University, Hamilton, Ontario L8S4L8, Canada

Low complexity and homopolymer sequences within coding regions are known to evolve rapidly. While their expansion may be deleterious, there is increasing evidence for a functional role associated with these amino acid sequences. Homopolymer sequences are thought to evolve mostly through replication slippage and, therefore, they may be expected to be longer in regions with relaxed selective constraint. Within the coding sequences of eukaryotes, alternatively spliced exons are known to evolve under relaxed constraints in comparison to those exons that are constitutively spliced because they are not included in all of the mature mRNA of a gene. This relaxed exposure to selection leads to faster rates of evolution for alternatively spliced exons in comparison to constitutively spliced exons. Here, we have tested the effect of splicing on the structure (composition, length) of homopolymer sequences in relation to the splicing pattern in which they are found. We observed a significant relationship between alternative splicing and homopolymer sequences with alternatively spliced genes being enriched in number and length of homopolymer sequences. We also observed lower codon diversity and longer homocodons, suggesting a balance between slippage and point mutations linked to the constraints imposed by selection.

[Supplemental material is available online at http://www.genome.org.]

One of the most commonly shared features between proteins in eukaryotic genomes is the abundance of simple sequences, characterized by low information content due to amino acid compositional bias (Golding 1999; Huntley and Golding 2000). Simple sequence composition varies from stretches of a single amino acid (hereafter homopolymer sequences) to repeats of a few residues. According to the paradigm associating protein function with three-dimensional structure, these simple sequences have long been considered the protein counterpart of "junk" DNA, as they often have undefined three-dimensional structure as revealed from X-ray crystallography (Bannen et al. 2007).

Several studies have shown these sequences to be highly polymorphic and to evolve rapidly between species (Brown et al. 2002; Huntley and Golding 2002; Huntley and Clark 2007). In accord with the idea that homopolymers and low complexity sequences are considered nonfunctional, these sequences are thought to evolve nearly neutrally (Lovell 2003). The widely observed polymorphisms in repeated motifs is mainly the consequence of DNA slippage (Levinson and Gutman 1987; Dieringer and Schlotterer 2003), which generates long runs of the same codon (hereafter termed homocodons) that become interrupted by point mutations over time.

Although the concept of neutral evolution seems valid for noncoding, nonfunctional repeated sequences, there is, however, increasing evidence that suggests another evolutionary scenario for repeated motifs found in coding sequences. The expansion of some homopolymers is known to be directly responsible for human genetic disorders, such as Huntington disease and spinobulbar muscular atrophy (Karlin et al. 2002; Usdin 2008). In dogs, variation of homopolymer size in transcription factors has also been linked to morphological differences among breeds (Fondon and Garner 2004). Furthermore, proteins involved in transcription, DNA/RNA binding, cellular signal transduction, reproduction, and gametogenesis are enriched in homopolymer sequences (Karlin et al. 2002; Alba and Guigo 2004; Faux et al. 2005; Huntley and Clark 2007). In addition, there is a relationship between homopolymer composition and protein function (Alba et al. 1999a; Alba and Guigo 2004; Faux et al. 2005). All these observations are suggestive of the action of selection acting on these sequences. If simple sequences are selectively deleterious, then purifying selection is expected to either remove them altogether or to reduce their instability by favoring point mutations that reduce the opportunity for DNA slippage (Kruglyak et al. 1998; Alba et al. 1999b; Rolfsmeier and Lahue 2000; Hancock and Simon 2005). Although the observation of homopolymer sequences encoded by a heterogeneous set of codons may be suggestive of the action of selection, because most of the amino acids are encoded by codons that differ only at a single position, such an observation may also be the consequence of slippage followed by the accumulation of mutations over time. Nonetheless, the action of selection was detected in poly-serine runs by Huntley and Golding (2006) , who showed evidence for the action of slippage in combination with selection for 12 loci and selection alone driving the evolution for two others out of 31 loci analyzed (Huntley and Golding 2006).

If single amino acid repeats are under selection, differences in size and/or codon composition of homopolymer sequences should be expected, depending on their degree of exposure to selection. In a recent analysis of five *Drosophila* genomes, we reported a rapid evolution of alternatively spliced exons (ASE) in comparison to constitutively spliced exons (CSE) (Haerty and Golding 2009). This observation is likely the consequence of lower levels of inclusion of alternatively spliced exons in the mature transcripts of genes leading to a lower degree of exposure to selective constraints (Haerty and Golding 2009). A similar result has also been reported in mammalian genomes (Kriventseva et al. 2003; Modrek and Lee 2003; Xing and Lee 2005; Chen et al. 2006; Ermakova et al. 2006). Therefore, one way to test for selection on the structure (size, codon composition) of homopolymer sequences is to determine their variation as exon splicing patterns change. We tested this

[1]Corresponding author.
E-mail golding@mcmaster.ca; fax (905) 522-6066.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.101246.109.

hypothesis, using the genome of *Drosophila melanogaster*, whose recent reannotation (Stark et al. 2007) includes alternative splicing events identified through EST, gene prediction, cDNA sequencing, and manual curation.

We observed a significant enrichment in homopolymer sequences in genes undergoing alternative splicing in comparison to genes with a single annotated protein isoform. Furthermore, the comparison of alternatively and constitutively spliced exons revealed a significant difference in homopolymers abundance, size, and codon composition depending on the splicing pattern of the exon in which they are found. Although the extension of this analysis to other eukaryotic genomes with different proportions of genes undergoing alternative splicing confirm some of the observations made in *D. melanogaster*, we also noted some differences between organisms, suggesting different selective pressures acting on low complexity sequences depending on the taxa.

Results

Simple sequences in the *D. melanogaster* genome

We found a total of 17,463 low complexity regions (LCR) and 16,073 homopolymer sequences among *D. melanogaster* proteins, corresponding to 9809 LCR (including 4965 without homopolymer sequences) and 9123 single amino acid repeats in 3671 and 4678 exons from 2991 and 3907 genes, respectively (Table 1). Exons containing homopolymers, are longer than exons without such sequences regardless of the exon splicing pattern (Kruskal–Wallis rank sum test, $P < 0.001$ in all the comparisons; see Supplemental material 1). Given that we previously showed that genes undergoing alternative splicing and genes with a single protein isoform evolve differently (Haerty and Golding 2009), we compared their homopolymer content and found alternatively spliced genes to be enriched with homopolymer sequences (1451/3612 vs. 2456/10,446; χ^2 test, $P < 0.001$; Fig. 1). We also observed alternatively spliced genes to be enriched in poly-A, N, Q , whereas genes with a single annotated protein isoform are enriched in poly-E, K, L, T ($P < 0.01$; Fig. 1). This bias in amino acid composition reflects the Gene Ontology functional differences between the two gene categories (see Supplemental material 2). Furthermore, in agreement with previous studies (Alba and Guigo 2004; Faux et al. 2005; Huntley and Clark 2007), we found that genes involved in development or transcription are enriched with homopolymers regardless of their splicing patterns (see Supplemental material 2).

We investigated the potential effects of relaxed selective constraints on the frequency, length, and codon composition of homopolymer sequences found in ASE, in comparison to CSEs. We

Table 1. Number of low complexity (without homopolymer sequences) and homopolymer sequences per splicing categories in *D. melanogaster*

	No. of exons	LCR	Homopolymer
ASE	8920	1169 (828)	2456 (1190)
Single	3591	314 (227)	693 (354)
Multiple	2757	250 (190)	579 (291)
Complex	2572	605 (411)	1163 (545)
CSE	13,783	1139 (824)	2226 (1145)
ONE	33,900	2657 (2027)	4437 (2343)

The number of exons with LCR or homopolymers sequence is given in parentheses. ASE, Alternatively spliced exon; CSE, constitutively spliced exon; ONE, exon found in gene with a single annotated protein isoform.

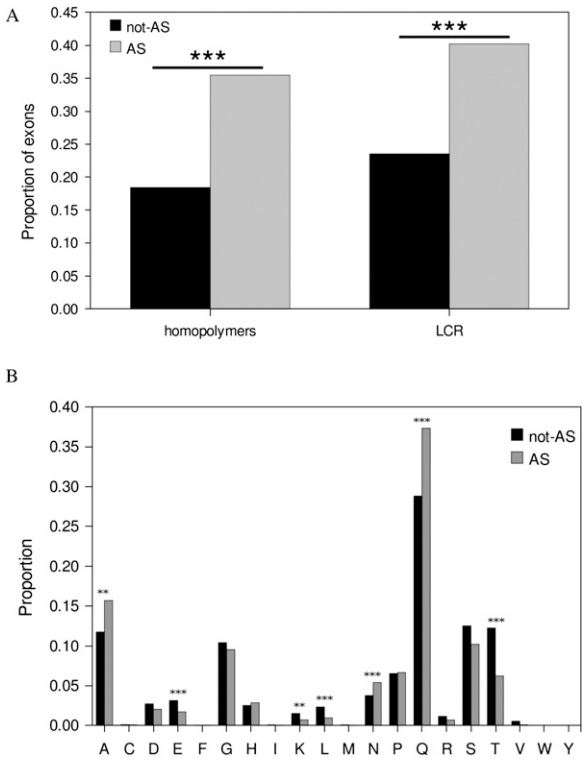


Figure 1. (*A*) Proportion of exons found in genes with a single annotated protein isoform (black bars) and in alternatively spliced genes (gray bars) with homopolymer sequences and low complexity protein regions *D. melanogaster*. (*B*) Comparison of homopolymer composition between genes undergoing alternative splicing and genes with a single protein isoform. $^{**}P < 0.01$; $^{***}P < 0.001$.

observed a greater proportion of homopolymer sequences in all ASE categories, in comparison to either CSE or to exons found in genes with a single annotated protein isoform (χ^2 test, $P < 0.01$ in all comparisons). Furthermore, the proportion of the exonic sequence occupied by a homopolymer is larger in both ASE found in a single and multiple transcripts in comparison to CSE (Kruskal–Wallis rank sum test, $P < 0.001$ in all comparisons; Fig. 2). In contrast, no difference in size is found between homopolymers from ASE using alternative 5′ and/or 3′ splice sites ("complex") and CSE ($P > 0.05$). To rule out any possible effect of greater divergence in these ASE leading to interrupted shorter homopolymers stretches, we analyzed low complexity sequences or homopolymer sequences interrupted by a single different amino acid and reached the same conclusions as before in both analyses ($P < 0.001$ in all comparisons). We also compared the sizes of the homopolymer sequences found in ASE and CSE within a gene to rule out any possible bias due to the use of genes with different evolutionary histories. The sizes of homopolymer sequences relative to the exon size is larger in ASE than in CSE (Kolmogorov–Smirnov test, $P < 0.001$).

If homopolymers sequences are under relaxed constraints when located in ASE, we should expect a larger absolute variation of their size across species in comparison to single amino acid repeats found in constitutively spliced exons. Using homopolymer sequences found in orthologous exons between five *Drosophila* species (Haerty and Golding 2009), we observed that homopolymers from ASE found in a single transcript are more variable than similar sequences found in CSE (Kruskal–Wallis rank sum test, $P < 0.001$; see Supplemental material 3).

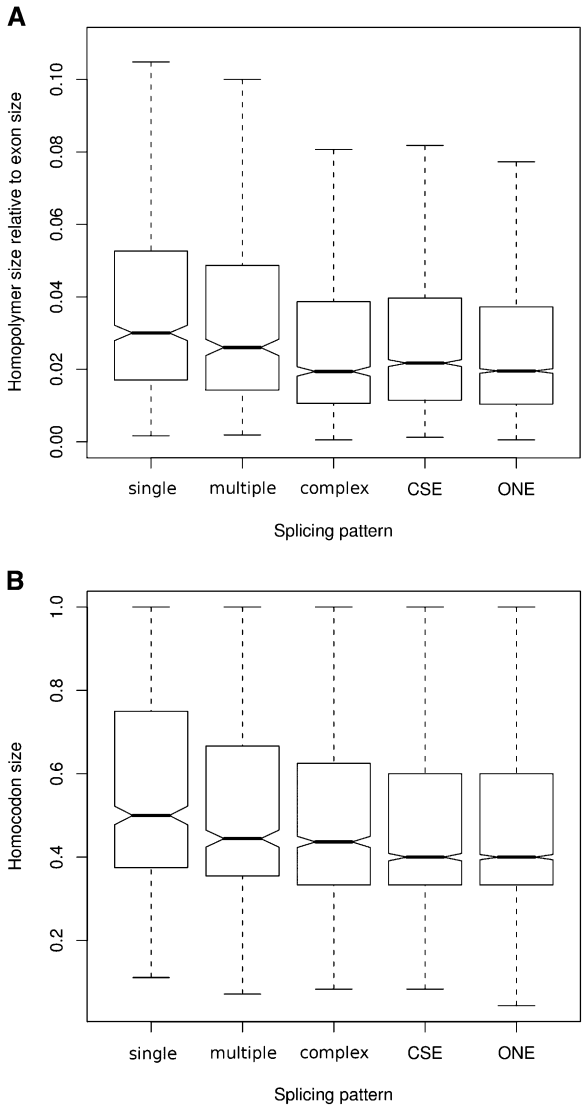


Figure 2. (*A*) Comparison of homopolymer sequence size relative to exon size between exon splicing patterns. (*B*) Comparison of homocodon size difference between exon splicing patterns. (CSE) Constitutively spliced exon; (ONE) exon found in gene with a single annotated protein isoform. The box shows the first and third quartiles; the dotted lines extend to the fifth and 95th percentiles; the size of the notch indicates level of uncertainty associated with the median.

At the nucleotide level, there is a higher proportion of homopolymers composed of a single codon in ASE in comparison to CSE (χ^2 test, $P < 0.01$). Furthermore, homopolymers within ASE are also characterized by longer homocodon stretches, as well as a lower codon diversity estimated through a Shannon-Weaver index than in CSE (Kruskal–Wallis rank sum test, $P < 0.001$ in both comparisons; Fig. 2). No difference is observed between homopolymers found in CSEs and exons found in genes with a single protein isoform ($P > 0.05$; Fig. 2). Similar conclusions are reached when using homopolymers that are composed of residues encoded by at least four codons. Interestingly, when dividing homopolymer sequences found in ASE with a complex splicing pattern into those falling in the constitutive region of the exon and those falling in the spliced/expanded region of the exons, we found similar results as in the comparison between ASE and CSE. Homopolymers found in the spliced/expanded region have a lower codon diversity ($P < 0.001$) and longer homocodon runs than homopolymers found in the constitutive part of the exon ($P < 0.001$). The conclusions remain the same when using amino acids encoded by at least four codons ($P < 0.01$ and $P < 0.001$ for the Shannon-Weaver index and the homocodon length, respectively).

Simple sequences in other eukaryotic genomes

Because the occurrence of alternative splicing is known to vary widely between eukaryotic genomes (Kim et al. 2007), we investigated whether the results found in *D. melanogaster* could be generalized to other eukaryotic genomes with different proportions of alternatively spliced genes. Exons with homopolymers are longer than exons without such sequences in each species ($P < 0.001$; see Supplemental material 1). As for *D. melanogaster*, we observed an enrichment of low complexity sequences or homopolymers in genes undergoing alternative splicing in comparison to genes with a single annotated protein isoform in the genomes of *C. elegans*, *D. rerio*, *M. musculus*, and *H. sapiens* (χ^2 test, $P < 0.001$ in all comparisons). However, as previously reported (Alba et al. 2001; Huntley and Clark 2007), the comparison between genomes of the amount of low complexity and homopolymer sequences revealed a striking difference between the species. *D. melanogaster* which has the smallest genome (bp, 14,141 genes) has the largest number of low complexity sequences, in comparison to worm, zebrafish, mouse and human genomes ($P < 0.001$; Fig. 3).

In the five species, ASE are enriched in homopolymer sequences in comparison to CSE (χ^2 test, $P < 0.001$ in all comparisons, Table 2). This effect appears to be associated with complex ASE. Furthermore, in *D. rerio*, longer homopolymer sequences in ASE, in comparison to CSE, are observed (Kruskal–Wallis rank sum test, $P < 0.01$; Supplemental material 4), while no difference between splicing patterns is found for the other genomes. With the exception of *C. elegans*, low complexity and homopolymer sequences found in complex ASE or in genes with a single protein isoform are significantly smaller than similar sequences found in other ASE categories or CSE ($P < 0.001$; see Supplemental material 4). In all genomes, exons with homopolymer sequences are longer than exons without such sequences ($P < 0.001$ in all comparisons). The size difference between ASE and CSE is not conserved in human

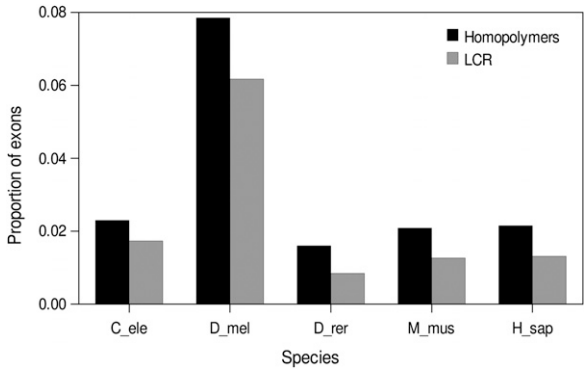


Figure 3. Proportion of exons with homopolymer sequences (black bars) or low complexity sequences without single amino acid repeats (gray bars) in the *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *H. sapiens* genomes.

Table 2. Number of homopolymer sequences in exons with different splicing patterns in *C. elegans*, *D. rerio*, *M. musculus*, and *H. sapiens*

Splicing pattern	*C. elegans*	*D. rerio*	*M. musculus*	*H. sapiens*
LCR				
Single	108 (1.76)	218 (0.52)	236 (0.93)	307 (0.95)
Multiple	61 (2.12)	65 (0.44)	248 (0.88)	387 (0.82)
Complex	144 (1.97)	595 (12.06)	530 (1.67)	698 (1.54)
CSE	322 (1.99)	245 (0.94)	738 (0.94)	1011 (1.09)
ONE	1965 (1.66)	1323 (0.57)	1697 (1.54)	1522 (1.79)
Homopolymers				
Single	245 (3.18)	451 (1.30)	496 (1.85)	588 (1.88)
Multiple	139 (2.14)	113 (2.12)	445 (1.72)	671 (1.65)
Complex	170 (4.09)	1106 (2.01)	817 (2.65)	1073 (2.32)
CSE	498 (2.82)	441 (1.98)	1314 (1.85)	1524 (1.87)
ONE	2710 (2.13)	2368 (1.66)	2594 (2.24)	2414 (2.69)

The percentages of exons with LCR and homopolymers sequences are given in parentheses. (CSE) Constitutively spliced exon; (ONE) exon found in gene with a single annotated protein isoform.

and mouse for exons with/without homopolymer sequences ($P > 0.05$). In each of the other genomes ASE are shorter than CSE.

At the nucleotide level, we observed a significantly greater proportion of homopolymers composed of a single codon in ASE found in a single transcript, in comparison to CSE, in the *D. rerio*, *M. musculus*, and *H. sapiens* genomes (χ^2 test, $P < 0.01$ in all comparisons), while no difference was observed in *C. elegans*. This result remained significant only in *M. musculus* when using homopolymers with amino acid encoded by at least four codons ($P < 0.001$). The absence of significant effects in zebrafish and human is likely attributable to low statistical power. Homopolymers found in ASE have a lower codon diversity in *D. rerio*, *M. musculus*, and *H. sapiens* (Kruskal–Wallis rank sum test, $P < 0.001$, $P < 0.05$, and $P < 0.001$, respectively), however no significant difference is found in *C. elegans* ($P > 0.05$ after Bonferroni correction). Similar conclusions are reached when limiting the analysis to amino acids encoded by at least four codons for both *M. musculus* and *H. sapiens*, while the effect vanishes in *D. rerio* ($P > 0.05$). Homopolymers are also encoded by longer homocodons in ASE of *C. elegans*, *M. musculus*, and *H. sapiens* ($P < 0.01$, $P < 0.001$, and $P < 0.001$, respectively; Supplemental material 5). The conclusions remain the same when using amino acids encoded by at least four codons ($P < 0.05$ in all comparisons). In *D. rerio*, although the length of homocodon runs tends to be longer in ASE, in comparison to CSE, the difference is not significant after Bonferroni correction ($P = 0.179$).

Discussion

The observation of rapid rates of evolution in intrinsically disordered regions, homopolymers, and low complexity sequences has led some authors to propose that such sequences may be hotspots for mutations and could generate material upon which selection may operate (Brown et al. 2002; Kashi and King 2006; Romero et al. 2006). In a similar fashion, alternatively spliced exons have also been proposed to be part of the raw material upon which selection acts, as relaxed selection allows the accumulation of mutations in these exons (Modrek and Lee 2003; Xing and Lee 2006). Here, we report that the difference in exposure to selection between alternatively and constitutively spliced exons leads to a significant difference in the size and codon composition of the homopolymer sequences found in these exons.

Although the genomes analyzed differ in the proportion of genes undergoing alternative splicing, we consistently observed alternatively spliced genes to be significantly enriched in both low complexity and homopolymer sequences in comparison to genes with a single protein isoform. This observation, as well as the enrichment of alternatively spliced regions in intrinsically disordered regions (Romero et al. 2006), may reflect the observed functional bias of genes including low complexity or homopolymer sequences toward DNA/RNA binding and transcription (Alba and Guigo 2004; Faux et al. 2005; Huntley and Clark 2007). Furthermore, while homopolymers found in genes with a single annotated protein isoform are directly exposed to selection, those found in alternatively spliced genes benefit from lower levels of inclusion, which may explain the observed difference in homopolymer density between gene categories.

In *D. melanogaster*, the density and structure of homopolymer sequences in ASE differ from similar sequences located in CSE, both when pooling exons across genes or when exons within the same gene are compared. Previous studies showed a faster evolution of alternatively spliced exons in comparison to constitutively spliced exons in *D. melanogaster* as a consequence of relaxation of selective constraints acting on alternatively spliced exons (Modrek and Lee 2003; Chen et al. 2006; Ermakova et al. 2006; Malko et al. 2006; Haerty and Golding 2009) or positive selection (Ramensky et al. 2008). The enrichment of ASE with potentially deleterious homopolymers, as well as the lower codon diversity and longer homocodon runs for homopolymers in ASE, support the hypothesis of a rapid evolution of single amino acid repeats, due to lowered selective constraints as a consequence of lower inclusion levels.

The higher codon diversity in homopolymers, both within constitutively spliced exons or exons found in genes with a single annotated protein isoform, could be explained by the action of replication slippage followed by the accumulation of synonymous substitutions as CSE are relatively older than ASE and have a higher rate of synonymous substitutions (Modrek and Lee 2003; Chen et al. 2006; Malko et al. 2006; Xing and Lee 2006; Haerty and Golding 2009). However, the action of selection can also explain this observation. In comparison to ASE, both CSE and exons found in genes with a single annotated protein isoform are found in all the mature transcripts of a gene, therefore, the expansion of homopolymer sequences is expected to have stronger deleterious effects, and selection may favor point mutations that disrupt homocodon runs limiting the action of slippage in these coding sequences (Kruglyak et al. 1998; Alba et al. 1999b; Rolfsmeier and Lahue 2000; Hancock and Simon 2005). In agreement with this last hypothesis, evidence of purifying selection acting on CSE have previously been reported (Haerty and Golding 2009) and conserved proteins between human and mice were found to be enriched in homopolymers with an heterogeneous codon composition (Mularoni et al. 2007). Whether the observed variation in homopolymer structures are linked to the direct effect of selection on single amino acid repeats or due to the indirect effect of selection acting on the exon in which the homopolymer is found remain to be formally tested. Hence these results should be interpreted with caution.

Our observations agree with a proposed scenario explaining the evolution of homopolymers sequences, according to which homopolymers arise in genomic regions under low selective constraints, such as ASE, and with time selection either removes the repeated sequence altogether or reduces homopolymer instability through the accumulation of point mutations, reducing the

opportunity for slippage to occur (Hancock et al. 2001; Hancock and Simon 2005; Simon and Hancock 2009).

Among homopolymer sequences found in alternatively spliced exons, those within exons spliced through the use of alternative and/or splice sites strongly differ from sequences observed in alternatively spliced exons found in single or in multiple transcripts. This may be explained by a different origin of these exons. While alternatively spliced exons (cassette exons) originate mostly through exonization of intronic sequences, exon duplication, or exon shuffling (Ast 2004; Blencowe 2006; Lev-Maor et al. 2007), exons spliced through the use of 5′ and/or 3′ splice sites have been suggested to evolve from existing constitutively spliced exons (Zhang and Chasin 2006). Therefore, the observed differences in the length and composition of homopolymer sequences between cassette exons and complex exons may be the consequence of different evolutionary histories among exons. However, the higher proportion of homopolymer sequences in these exons in all the genomes analyzed is puzzling. This could be explained by a higher rate of nonsynonymous substitutions in these exons leading to both more numerous and shorter interrupted homopolymer sequences. However, the analyses of both homopolymer sequences and sequences composed of a few different amino acids revealed the same effect as observed on homopolymer sequences.

The analyses of these five genomes reveal large differences in the effect of splicing on the number, length, and composition of homopolymer sequences. Although alternatively spliced genes are enriched in single amino acid repeats in all the species and we found an enrichment of homopolymer sequences in *C. elegans*, *D. melanogaster*, and *D. rerio* in alternatively spliced exons relative to constitutively spliced exons, we do not observe these results in *M. musculus* and *H. sapiens*, even though these species have a greater proportion of alternatively spliced genes (Kim et al. 2007). One possibility explaining this phenomenon may be the lower proportion of genes with alternative splicing in the Ensembl database, only 37.89% and 51.89% genes are annotated as alternatively spliced in mouse and human, respectively, whereas estimates of the proportion of alternatively spliced genes reach up to 56% in mouse (Kim et al. 2007) and 94% in human (Wang et al. 2008). In such a case, the observed difference between alternatively spliced genes and genes with a single protein isoform should have been less clear. However, their homopolymer sequences have similar codon composition properties, suggesting that the conclusions previously drawn from the observation of the fruit fly genome may also be valid for the zebrafish, mouse, and human genomes. In these genomes, a balance between slippage and point mutations linked to the variation of selective constraints associated with alternative splicing is likely the cause of the difference in the structure of homopolymer sequences.

Methods

The genomes of *Danio rerio* (assembly version 7), *Mus musculus* (NCBI 37 assembly), and *Homo sapiens* (NCBI 36 assembly) were downloaded from the Ensembl database (http://www.ensembl.org/). The genomes of *Caenorhabditis elegans* and *D. melanogaster* were downloaded from WormBase (release WS198, http://www.wormbase.org/) and FlyBase (release 5.12, http://flybase.org/), respectively. Using the genome annotations, we classified exons according to their splicing patterns. Exons found in a single alternatively spliced transcript, in more than one transcript (cassette exons) or if alternatively spliced via the use of alternative and/or splice sites were respectively classified as "single," "multiple," or "complex." The sequences of complex exons were also divided into a constitutive part and a spliced/extended part. Exons found in all the transcripts of genes undergoing alternative splicing were classified as constitutively spliced. We also consider the exons found in genes with a single annotated protein isoform (labeled as "ONE" in Figs. 1–3).

Low complexity protein sequences were identified using SEG (Wootton and Federhen 1993) and the parameters previously established by Huntley and Golding (2002) in order to select for longer and more repetitive low complexity sequences (windows size 15, complexity threshold 1.9). We separately collected all the homopolymer sequences defined as runs of at least five identical amino acids and runs of at least five amino acids interrupted by a single different residue. We use the term "low complexity sequences" for amino acid sequences characterized by a low information content composed of more than one amino acid and use the term "homopolymer" or "single amino acid repeats" to indicate stretches of a single amino acid. The size of exons is known to vary depending upon their splicing pattern (Zavolan et al. 2003; Sorek et al. 2004; Zheng et al. 2005), therefore, to allow a comparison between exon splicing patterns, the size of the homopolymers is expressed as the proportion of the exon length. For the homopolymers found in exons spliced through the use of alternative and/or splice sites, we used the size of the smallest exon in which a homopolymer is found in order to avoid underestimating the relative size of the homopolymer. The codon diversity of homopolymer sequences was assessed using a Shannon-Weaver index:

$$S = \frac{p_i \log_2 (p_i)}{L}, \quad (1)$$

with the frequency of codon i and the length of the homopolymer L. The size of the longest homocodon run within a homopolymer was retrieved. We tested for an association between gene splicing patterns, homopolymer enrichment or homopolymer composition and molecular function using FATIGO (http://www.babelomics.org/; Al-Shahrour et al. 2006).

If homopolymers are submitted to different selective pressures depending upon the splicing pattern of the exon in which they are found, different evolutionary patterns are expected. We assessed the size variation of homopolymers found in a set of 38,762 exons with orthologs in *Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, and *Drosophila erecta* used in a previous analysis (Haerty and Golding 2009). This set includes exons greater than 50 bp with best reciprocal blast hits in all species and covering at least 80% of the *D. melanogaster* query. We implemented a similar approach as the one used by Huntley and Clark (2007) to compute the variation of size of a homopolymer across the five species. We computed the distance between species as the absolute difference in size of homopolymers. We used the total length of the neighbor-joining tree build using the NEIGHBOR package from PHYLIP (Felsenstein 1989) as a measure of the homopolymer variability across species.

The comparisons of low complexity and homopolymers sequence sizes between exon splicing categories were performed using a Kruskal–Wallis rank sum test with 10,000 permutations. Bonferroni corrections for multiple tests were applied when necessary. The comparison of low complexity and homopolymers sequence distributions between splicing patterns were performed using a test with one degree of freedom. Differences in homopolymer sizes and compositions between ASE and CSE within the same gene were compared to the difference of homopolymer sizes and compositions from independent CSE from the same gene, a total of 1000 permutations were preformed. The distributions were compared using a Kolmogorov–Smirnov test.

Acknowledgments

We thank Carlo Artieri, Ben Evans, and Richard Morton for their comments on the early versions of this manuscript. This work was supported by a Canada Research Chair and a Natural Science and Engineering Research Council grant to G.B.G.

References

Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J, Dopazo J, et al. 2006. BABELOMICS: A systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* **34:** W472–W476.

Alba MM, Guigo R. 2004. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res* **14:** 549–554.

Alba MM, Santibanez-Koref MF, Hancock JM. 1999a. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. *J Mol Evol* **49:** 789–797.

Alba MM, Santibanez-Koref MF, Hancock JM. 1999b. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Mol Biol Evol* **16:** 1641–1644.

Alba MM, Santibanez-Koref MF, Hancock JM. 2001. The comparative genomics of polyglutamine repeats: Extreme differences in the codon organization of repeat-encoding regions between mammals and *Drosophila*. *J Mol Evol* **52:** 249–259.

Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* **5:** 773–782.

Bannen RM, Bingman CA, Phillips GN Jr. 2007. Effect of low-complexity regions on protein structure determination. *J Struct Funct Genomics* **8:** 217–226.

Blencowe BJ. 2006. Alternative splicing: New insights from global analyses. *Cell* **126:** 37–47.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55:** 104–110.

Chen FC, Wang SS, Chen CJ, Li WH, Chuang TJ. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol* **23:** 675–682.

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.

Ermakova EO, Nurtdinov RN, Gelfand MS. 2006. Fast rate of evolution in alternatively spliced coding regions of mammalian genes. *BMC Genomics* **7:** 84. doi: 10.1186/1471-2164-7-84.

Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res* **15:** 537–551.

Felsenstein J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* **5:** 164–166.

Fondon JW III, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci* **101:** 18058–18063.

Golding GB. 1999. Simple sequence is abundant in eukaryotic proteins. *Protein Sci* **8:** 1358–1361.

Haerty W, Golding B. 2009. Similar selective factors affect both between-gene and between-exon divergence in *Drosophila*. *Mol Biol Evol* **26:** 859–866.

Hancock JM, Simon M. 2005. Simple sequence repeats in proteins and their significance for network evolution. *Gene* **345:** 113–118.

Hancock JM, Worthey EA, Santibanez-Koref MF. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol Biol Evol* **18:** 1014–1023.

Huntley MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol* **24:** 2598–2609.

Huntley M, Golding GB. 2000. Evolution of simple sequence in proteins. *J Mol Evol* **51:** 131–140.

Huntley MA, Golding GB. 2002. Simple sequences are rare in the Protein Data Bank. *Proteins* **48:** 134–140.

Huntley MA, Golding GB. 2006. Selection and slippage creating serine homopolymers. *Mol Biol Evol* **23:** 2017–2025.

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci* **99:** 333–338.

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

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res* **35:** 125–131.

Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S. 2003. Increase of functional diversity by alternative splicing. *Trends Genet* **19:** 124–128.

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* **95:** 10774–10778.

Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast G. 2007. The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet* **3:** e203. doi: 10.1371/journal.pgen.0030203.

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

Lovell SC. 2003. Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Lett* **554:** 237–239.

Malko DB, Makeev VJ, Mironov AA, Gelfand MS. 2006. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res* **16:** 505–509.

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* **34:** 177–180.

Mularoni L, Veitia RA, Alba MM. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89:** 316–325.

Ramensky VE, Nurtdinov RN, Neverov AD, Mironov AA, Gelfand MS. 2008. Positive selection in alternatively spliced exons of human genes. *Am J Hum Genet* **83:** 94–98.

Rolfsmeier ML, Lahue RS. 2000. Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae*. *Mol Cell Biol* **20:** 173–180.

Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, et al. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci* **103:** 8390–8395.

Simon M, Hancock JM. 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* **10:** R59. doi: 10.1186/gb-2009-10-6-r59.

Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, Shamir R. 2004. A non-EST-based method for exon-skipping prediction. *Genome Res* **14:** 1617–1623.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232.

Usdin K. 2008. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Res* **18:** 1011–1019.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* **17:** 149–163.

Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci* **102:** 13526–13531.

Xing Y, Lee C. 2006. Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* **7:** 499–509.

Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* **13:** 1290–1300.

Zhang XH, Chasin LA. 2006. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci* **103:** 13427–13432.

Zheng CL, Fu XD, Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11:** 1777–1787.

Received September 28, 2009; accepted in revised form December 30, 2009.