BMC Genomics

BioMed Central

Research article

The expansion of amino-acid repeats is not associated to adaptive evolution in mammalian genes

Fernando Cruz*[1,2], Julien Roux[1,2] and Marc Robinson-Rechavi[1,2]

Address: [1]Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland and [2]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

Email: Fernando Cruz* - Fernando.CruzRodriguez@unil.ch; Julien Roux - julien.roux@unil.ch; Marc Robinson-Rechavi - Marc.Robinson-Rechavi@unil.ch

* Corresponding author

Published: 18 December 2009

BMC Genomics 2009, **10**:619 doi:10.1186/1471-2164-10-619

Received: 1 September 2009
Accepted: 18 December 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/619

Abstract

Background: The expansion of amino acid repeats is determined by a high mutation rate and can be increased or limited by selection. It has been suggested that recent expansions could be associated with the potential of adaptation to new environments. In this work, we quantify the strength of this association, as well as the contribution of potential confounding factors.

Results: Mammalian positively selected genes have accumulated more recent amino acid repeats than other mammalian genes. However, we found little support for an accelerated evolutionary rate as the main driver for the expansion of amino acid repeats. The most significant predictors of amino acid repeats are gene function and GC content. There is no correlation with expression level.

Conclusions: Our analyses show that amino acid repeat expansions are causally independent from protein adaptive evolution in mammalian genomes. Relaxed purifying selection or positive selection do not associate with more or more recent amino acid repeats. Their occurrence is slightly favoured by the sequence context but mainly determined by the molecular function of the gene.

Background

Microsatellites or simple sequence repeats (SSRs) are DNA tracts composed of 1-6 bp long motifs repeated in tandem. A balance between slippage events, that increase the purity of the repeat, and point mutations, that tend to eliminate perfect repeats, determines their length distribution. However, as the slippage rate is higher than the point mutation rate, the purity of the repeated tract will be an inverse measure of the age of the SSR [1-3].

Triplet repeats are more common within coding regions [4], as they are less likely to alter the reading frame and can be translated into amino-acid repeats (AARs). AARs are frequently associated with disease [e.g. [5,6]]. Strong effects on morphology and phenotype have also been described in dog breeds [7]. Examples of AARs contributing to adaptive evolution [2,8] have been found in case studies in insects [9], plants [10,11] and mammals [12].

Genomic comparisons have shown that highly variable AARs have a higher purity in their coding sequence [13,14]. AAR expansion has been found to correlate with the non-synonymous rate of substitution [13,15,16] supporting a role of selection in their expansion. The correlation is consistent with either relaxed purifying selection, or with positive selection; the latter is suggested by case

studies of adaptive evolution [9-12]. Previous studies [13,15,16] have been restricted in their taxonomic scale, did not take into account exon boundaries, and did not integrate potential confounding parameters into their analyses. Here we perform a systematic study of mammalian genomes. We contrasted AARs in positively selected genes (PSGs) and non-PSGs [17] to examine their relationship with protein adaptive evolution. We also analyzed other factors correlating with AARs in 6 high coverage mammalian genomes. The results were confirmed on a dataset of orthologous exons with wider species diversity. Thus, the relative contribution of each parameter to the expansion of AARs has been determined.

Our results indicate that AAR expansion is not causally associated to protein adaptive evolution on a genome scale. However, there is a minor contribution of the GC context surrounding the AARs for an increased slippage rate. AARs are over-represented in genes involved in DNA binding and transcriptional activity.

Results

Recent expansions in mammalian Positively Selected Genes

Under the hypothesis of AARs as a resource for adaptation, genes that have experienced adaptive evolution are expected to show more and more recent (i.e. purer) AARs associated with a higher substitution rate. To test this prediction, we used the PSGs identified in a thorough study of mammalian genes [17]. First, we compared the amount of repeat containing genes (RCGs) and non-repeat containing genes (non-RCGs) between positively selected genes (PSGs) and non-positively selected genes (non-PSGs) (Table 1). A Fisher's Exact Test shows a weak but significant association between repeats and positive selection (p = 0.042). Repeats were then split in two classes, young repeats with high purity (>= 0.9) and old repeats with low purity (<0.9) (Table 1). The PSGs have significantly more young repeats (p = 0.0004), suggesting that adaptive evolution in mammals could be associated with recent expansion of repeats.

We also analyzed the physical properties of the AARs. The Lehninger classification describes four categories of amino acids: acidic, basic, polar uncharged and hydrophobic amino acids [6]. All simple amino acid repeats were classified into the corresponding category for PSGs and non-PSGs (Table 2). The distribution of amino acid repeats differed significantly between PSGs and non-PSGs in a chi-square test (p = 0.0003). The differences remain significant after *Yate's correction for continuity* [18] (Yates' p = 0.001) and are mainly due to an excess of repeats of acidic and hydrophobic amino acids in the PSGs. The excess of repeats of hydrophobic AARs explains 77.3% of the differences between PSGs and non-PSGs. However this excess is essentially due to an excess of Leucine repeats. Removing these, the Chi-square is not significant after *Yate's correction for continuity* (Yates' p = 0.067).

Table 1: Counts of AARs in Positively versus non-Positively Selected Genes in Mammals

	RCGs	non-RCGs	Pure	Impure
PSGs	19	381	26	8
non-PSGs	1207	14922	2021	2448

Counts of repeat containing genes (RCGs), repeat-free genes (non-RCGs), and of number of pure and impure amino-acid repeats (AARs), of the PSGs and non-PSGs classes. These numbers were used to perform two different Fisher's Exact Tests.

The correlation of amino acid repeats with positive selection and evolutionary rates is spurious

Previous studies in human and mouse have suggested that AAR expansion could be favoured by relaxed purifying selection, repeat length being associated with higher rates of non-synonymous substitutions [13,15]. While our analyses of 6 high-quality mammalian genomes confirm a positive correlation between d_N and repeat length (ρ = 0.043, p = 0.002), this is very weak. A stronger correlation is observed between the average purity of AARs and d_N (ρ = 0.111, $p = 1.54 \cdot 10^{-12}$), but there is a similar correlation with d_S (ρ = 0.112, $p = 7.8 \cdot 10^{-13}$), and the correlation with ω, which should be most indicative of selection, is the weakest (ρ=0.058, p = 0.00017). The similar values of correlation with d_N and d_S may be related to the correlations between these rates (d_N *vs.* d_S ρ = 0.485, $p < 2.16 \cdot 10^{-16}$), and with the GC context surrounding the repeats (d_N *vs.* $GC_{context}$ ρ = 0.115, $p < 2.16 \cdot 10^{-16}$; d_S *vs.* $GC_{context}$ ρ = 0.478, $p < 2.16 \cdot 10^{-16}$). Indeed the $GC_{context}$ also correlates with the purity (ρ = 0.09, $p = 4.272 \cdot 10^{-08}$) and the number of AARs (ρ = 0.06, $p < 2.16 \cdot 10^{-16}$).

In order to disentangle the effect of these features of gene evolution we fitted the observed variation to a linear model and performed an analysis of variance [e.g. [19]]. We performed this analysis on 3 different mammalian datasets: PSGs, the 6 high-coverage genomes, and orthologous exons (Material and Methods). We detail only the analyses of the PSG dataset (Tables 3 and 4). The other two datasets, with a majority of genes under purifying selection (mean ω = 0.161 ± 0.21), provide similar results and conclusions with slight variations in the percentage of explained variance (Additional file 1, Tables S1-S4). Adaptive AAR expansions should result in high average purities (i.e., recent or frequent slippage events) and many AARs per positively selected gene. Although the contribution of evolutionary parameters is statistically significant, it is minimal and unlikely to be biologically relevant. For the average purity of the repeats on a gene, ω explains only 0.4% of the variance, while the fact of detecting adaptive

Table 2: Physicochemical Properties of the AARs in Positively Selected versus Non-Positively Selected Mammalian Genes

	Acidic	Basic	Polar	Hydrophobic
PSGs	10 (0.95)	0 (-1.08)	7 (-2.51)	17 (3.23)
non-PSGs	970 (-0.083)	154 (0.094)	2314 (0.22)	1031 (-0.28)

Counts of amino acid categories using the Lenhinger classification for each AAR in PSGs and non-PSGs. Values shown in brackets correspond to the residuals for each cell obtained in a Pearson's χ^2 test.

evolution on any branch of the tree (i.e. significant Likelihood Ratio Test) explains <0.1% of the variance observed for the number of repeats. This shows that the enrichment for recent repeats observed using Fisher's Exact Test was a spurious association. Protein length explains 2% of the variance for AARs, which is not surprising as longer proteins have a greater potential to host repeats. Of note, it has been shown that positive selection tests are also more significant on longer proteins [e.g. [19]], which may contribute to the association between PSGs and AARs.

The excess of leucine repeats also appears spurious, as there is no significant correlation between the ω values of each branch in the tree and the length of the leucine repeats ($\rho = 0.36$, $p = 0.25$) or their purity ($\rho = -0.17$, $p = 0.59$).

GC rich contexts can favour the expansion of amino acid repeats

The $GC_{context}$ is the only parameter highly significant in both analyses of variance (on AAR purity and on AAR number). It explains only 1.6% and 0.7% of the variance, but this is 3-fold more than the percentage explained by ω or by significant evidence of positive selection. Thus GC-rich sequences appear more prone to the expansion of repeats. To explore this question, we analyzed 16 exons showing accelerated evolution in primates due to GC-biased gene conversion (gBGC) [20]. Two out of these 16 exons have AARs, or 12% of this small dataset. Interestingly, the purity of these repeats highly correlates with the $GC_{context}$ ($\rho = 0.85$, $p = 0.002$, in 10 mammalian sequences), indicating that a GC increase due to gBGC might sometimes favour the expansion of AARs.

Previous studies have also shown that nucleotide compositional constraints increasing the GC content at 3rd codon positions (GC3) influence the expansion of homopolymeric AARs in mammalian and reptilian transcription factors [21]. Analyses of mammalian exons and of complete protein coding genes (Figure 1) shows that there is a weak, but highly significant, positive correlation between purity and GC3 in the DNA sequence surrounding the repeats ($\rho = 0.28$, $p < 2.2 \cdot 10^{-16}$ and $\rho = 0.126$, $p < 2.2 \cdot 10^{-16}$, for exons and whole genes, respectively). A Welch's t-test comparing the GC3 context of exons containing pure and impure repeats indicates that genes hosting pure repeats have on average a higher GC3 than impure repeats (0.75 and 0.66 respectively, $p < 2.2 \cdot 10^{-16}$). In summary, these results consistently indicated that in mammals there is a small but significant increase of AAR expansion in regions with high GC.

Aminoacid repeats and gene expression

The main reasons that led us to study the relationship between repeat expansion and expression levels are: 1) The observed excess of hydrophobic repeats is likely to lead to aggregation and misfolding in PSGs [22]. 2) The correlation between substitution rates and $GC_{context}$, that also correlates with the average purity of AARs, has been shown to be limited by expression-related purifying selection [23]. 3) In *E. coli* it has been observed that the stability of the structure around the translation start is directly related with the expression level [24].

Table 3: ANOVA of Linear Model to Explain the Average Purity of the AARs in Positively Selected and Non-Positively Selected Genes

	Df	Sum Sq	Mean Sq	F value	*p-value*	Var. (%)[6]
Residuals	3616	20.5105	0.0057			97.351
GCcontext[1]	1	0.3351	0.3351	59.078	**1.94E-14**	**1.590**
Species[2]	5	0.1154	0.0231	4.0684	**0.001101**	**0.548**
ω[3]	1	0.0872	0.0872	15.3805	**8.95E-05**	**0.414**
LRT[4]	1	0.0183	0.0183	3.2305	0.072362	0.087
P. length (aa)[5]	1	0.002	0.002	0.3525	0.552754	0.009
Total	3625	21.0685	0.4714			

[1]GC content excluding the stretch containing AARs; [2]species containing the AAR(s); [3]omega (d_N/d_S) of the most significant evolutionary model; [4]significant test for positive selection at any branch of the tree; [5]protein length in aminoacids; [6]proportion of variance explained.

Table 4: ANOVA of Linear Model to Explain the Number of AARs in Positively Selected and Non-Positively Selected Mammalian Genes

	Df	Sum Sq	Mean Sq	F value	*p-value*	Var. (%)[6]
Residuals	82096	6806.8	0.1			96.879
P. length (aa)[1]	1	168.1	168.1	2027.45	**>2.20E-16**	**2.392**
GCcontext[2]	1	48.9	48.9	590.12	**>2.20E-16**	**0.696**
LRT[3]	1	1.4	1.4	16.3141	**>3.71E-06**	**0.020**
Species[4]	5	0.8	0.2	1.9078	0.0894	0.011
ω[5]	1	0.048	0.04798	0.5787	0.4468	0.001
Total	82105	7026.01	218.748			

[1]Protein Length in aminoacids; [2]GC content excluding the stretch containing AARs; [3]significant test for positive selection at any branch of the tree; [4]species containing the AAR(s); [5]d_N/d_S of the most significant evolutionary model; [6]proportion of variance explained.

For the 1,057 human and 1,009 mouse genes that contain at least one AAR, we performed an analysis of variance including the expression levels in 5 representative organs as factors. The result shows that expression level has no impact on the expansion of AARs, measured as average purity or as number of repeats in the hosting gene (Additional file 1, Tables S5-S8), neither in mouse nor human.

Conversely, the number of AARs proximal to the translation start for human and mouse does not explain, in any of the 5 organs, the observed variance in the expression levels. For simplicity we show only the results obtained for the human brain (Table 5).

In conclusion, we can reject any simple relation between the presence of AARs or their age, and the expression level of human and mouse genes.

Molecular function of genes hosting amino acid repeats

We studied the relation between AARs and the Gene Ontology terms (GO), for Molecular Function, Biological Process and Cell Component, of all human and mouse protein-coding genes. As very similar results were obtained for both species we will report only those obtained for human.

Genes containing AARs are enriched in a wide variety of molecular functions, mainly involved in binding, transcription and nuclear structures (Table 6); analyses accounting for purity or Biological Process of genes with AARs support these results (data not shown). Including these molecular function terms in the linear model to explain the number of AARs per gene, the total percentage of variance explained by significantly enriched GO terms is 13.9% for human and 15.2% for mouse (see Table 7 for human and Table S9 for mouse). This is not the case for average purity of AARs, for which GC context remains the main explanatory factor in human (2.73% of variance explained, Table S10). Finally, the cellular compartment nucleus is also enriched in genes with AARs, and in genes with purer AARs (GO:0005634, $p < 6.19 \cdot 10^{-12}$).

The ice binding molecular function (GO:0050825) is overrepresented. But this excess disappears after excluding the Alanine repeats. This appears to be an annotation bias, as genes containing alanine-rich repeats are attributed this function by partial sequence similarity with the InterPro entry IPR000104 (Antifreeze protein, type I), a special glycoprotein identified in marine teleosts from polar oceans[25].

Discussion

In mammals, a positive correlation between d_N and repeat length is weak but statistically significant. This result is congruent with previous analyses in smaller datasets of human and mouse genomes [13,15]. The purity of the AARs per gene or exon shows a similar trend. But these weak correlations can be explained by the influence of the GC context surrounding the repeat. High GC content can generate a sequence context more prone to slippage[21,26-28] and thus expansion of AARs. Indeed we found an example of this in exons that have experienced GC-biased gene conversion in primates. Similarly, while there is an increase in the amount of recent AARs in mammalian PSGs, these recent expansions are better explained by GC content than by positive selection acting on codons. Therefore it seems that, in contradiction to previous reports [15], the expansion of AARs is not causally associated with substitution rates. While purifying selection limits the expansion of AARs[e.g. [29]], this appears to be distinct from the selective pressure on individual (aligned) amino acid sites. That means that these repeats are experiencing not only different mutational processes, but also particular selective constraints, leading to a more complex scenario of evolution.

Our analyses, even of individual exons, suggest that increased substitution rates are not usually linked to the

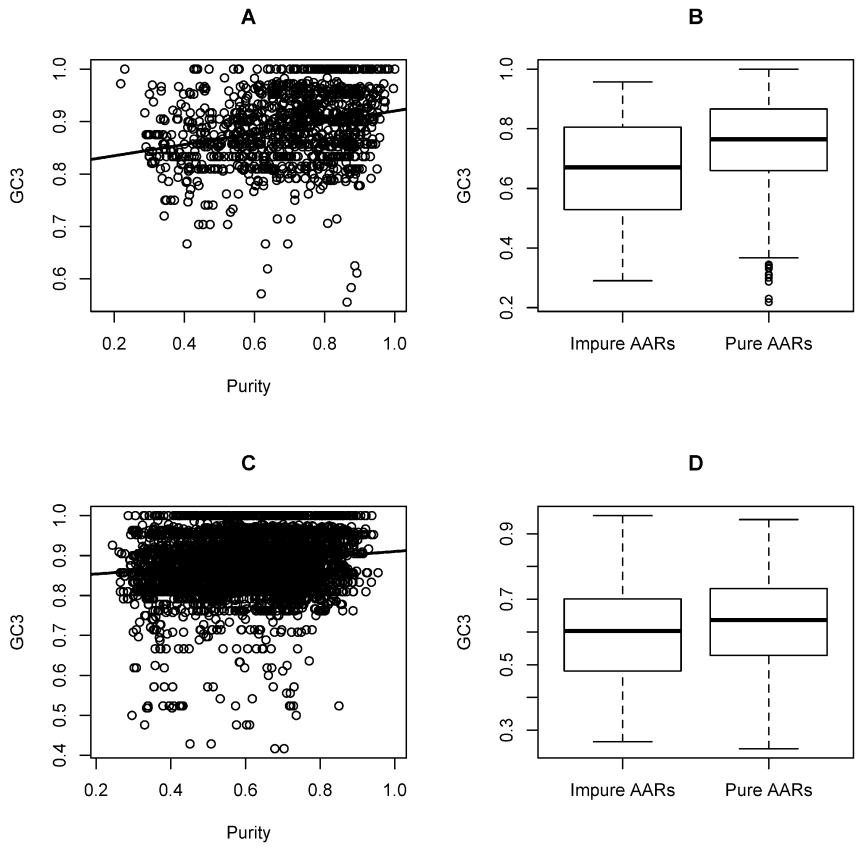

Figure 1
Influence of GC content at 3rd codon position on AAR purity. GC3, GC at 3rd codon positions in the sequence context of the repeats. (A) positive correlation and regression line (using least squares) between GC3 and purity in orthologous mammalian exons; (B) Average GC3 in Impure and Pure AARs in orthologous mammalian exons ($p < 2.16 \cdot 10^{-16}$; Welch's t-test); (C) positive correlation between GC3 and purity in mammalian genomes and regression line (using least squares); (D) Average GC3 in Impure and Pure AARs in mammalian genomes ($p < 2.16 \cdot 10^{-16}$; Welch's t-test).

presence of AARs. However, it is possible that in some particular cases, as has been suggested for *Drosophila*, the expansion of AARs can produce compensatory changes on the neighbouring sites to accommodate the perturbation generated by the repeat[30]. We also cannot exclude the existence of adaptive evolution related with AARs[7,8], in the absence of a good reference neutral model for tri-nucleotide expansions in proteins. But our results do show that the selective pressure as measured by codon models is not related with putative adaptive evolution of AARs.

AARs in mammalian genes do not seem to affect gene expression significantly. Unlike repeats which disrupt the reading frame, and have a strong effect on replication and transcription stability[31], the tri-nucleotide repeats might be constrained in a different way. It seems that repeats located in the promoter region[32] have a stronger influence on transcription than do AARs, even those near the transcription start.

The analyses of molecular function confirmed an enrichment in the transcription factor, DNA binding, molecular transducers and binding categories that is consistent with previous studies of polymorphic repeats [26,33,34]. The overrepresentation of transcription factor categories supports the existence of *trans* effects, as these repeats might alter the expression of the target genes and end up producing dramatic changes on the phenotype[7]. However, while the ice-binding protein is involved in hypothermic resistance in some antartic fishes vertebrates[25,35], its overrepresentation in alanine-rich mammalian genes is probably due to an annotation bias.

In general, we found that AARs are located in proteins that interact with DNA, RNA, ligands or other proteins, so it is likely that they contribute to adapt or modulate the inter-

Table 5: ANOVA of a Linear Model to Explain the Expression Level of Human Genes in the Brain

	Df	Sum Sq	Mean Sq	F value	*p-value*
P. length (aa)[1]	1	2.5	2.5	0.6648	0.4151
GCcontext[2]	1	0.1	0.1	0.0178	0.894
N° AARs[3]	1	0.1	0.1	0.0226	0.8805
AARs $_{+30\ nt}$[4]	1	1	1	0.2669	0.6055
AARs $_{+60\ nt}$[5]	1	1.3	1.3	0.3386	0.5608
AARs $_{+90\ nt}$[6]	1	5.5	5.5	1.4469	0.2293
d_N[7]	1	10.1	10.1	2.6413	0.1045
Average Purity[8]	1	0.4	0.4	0.114	0.7357
Residuals	893	3416.8	3.8		

[1]GC content excluding the stretch containing AARs; [2]protein length in aminoacids; [3]Number of AARs; [4-6]Number of AARs in a window of [4]+30 nt, [5]+60 nt and [6]+90 nt from translation start; [7]Non-synomymous substitution rate; [8]Average Purity of the AARs.

action capacity of these proteins. Longer proteins and repeat-rich proteins tend to have a higher connectedness within interaction networks, suggesting that they contribute to an enlarged interaction surface and constitute more flexible subunits[36]. Some AAR have been recently associated to the presence of repeats to specific domains, such as signal peptides or transmembrane regions[16], pointing to their role in facilitating molecular interactions of extreme importance. For example, in the Drosophila ARC 70 cofactor complex, the -130 and -230 subunits contain an expansion of glutamine residues, a prevalent feature of sequence-specific activators in *Drosophila*[37].

Conclusions

Despite the appealing idea of an adaptive role of the expansion of amino acid repeats, we can rule out a link with adaptive evolution in mammalian protein-coding genes as measured by codon models. Genome-wide, GC content is more relevant to amino acid repeat expansions than substitution rates. Amino acid repeats are under strong functional constraints and expand preferentially in transcription factors and nuclear genes involved in DNA and/or protein interactions. Why some genes accumulate more and most recent amino acid repeats requires further study in a network context, to shed light on the evolutionary dynamics and function of these mutations.

Methods

Positively Selected Genes (PSGs)

A recent study in mammals[17] performed a thorough analysis for detecting positive selection in six mammalian genomes. A likelihood ratio test for positive selection on any branch of the phylogeny reported 400 Positively Selected Genes (PSGs), and 16,129 genes that have not experienced any detected positive selection in mammals (non-PSGs). Alignments for these genes were downloaded from the author's website http://compgen.bscb.cornell.edu/projects/mammal-psg/lrtall.txt and screened for repeats.

High-quality Mammalian Genomes

To study the relationship of multiple factors that could be influencing the expansion of repeats in mammalian genomes, we used mammalian assemblies with high cov-

Table 6: Enrichment of Molecular Functions of Genes containing AARs

GO.ID	Term[1]	Corrected p-value[2]
GO:0050825	**ice binding**	< 1E-26
GO:0003677	**DNA binding**	4.01E-15
GO:0003700	**transcription factor activity**	1.26E-13
GO:0043565	**sequence-specific DNA binding**	5.79E-13
GO:0005199	structural constituent of cell wall	1.00E-08
GO:0004879	**ligand-dependent nuclear receptor activity**	3.15E-07
GO:0003682	**chromatin binding**	2.54E-06
GO:0003723	RNA binding	7.63E-05
GO:0008270	zinc ion binding	0.000303826
GO:0004969	histamine receptor activity	0.0008013
GO:0045735	nutrient reservoir activity	0.0008013
GO:0003702	RNA polymerase II transcription factor activity	0.001116964
GO:0003676	nucleic acid binding	0.001580342
GO:0003705	RNA polymerase II transcription factor activity, enhancer binding	0.009862154
GO:0003735	structural constituent of ribosome	0.02671
GO:0005249	voltage-gated potassium channel activity	0.049858667
GO:0004386	helicase activity	0.065105625
GO:0016563	transcription activator activity	0.13355
GO:0003714	transcription corepressor activity	0.13355
GO:0005179	hormone activity	0.199622105

[1] In bold terms overrepresented also for genes hosting the highest average purity of their AARs; [2] FDR < 20%.

Table 7: Percentage of Explained Variance of the Number of Aminoacid Repeats

Factor	Pr(>F)	Var. (%)
ice binding	<2.20E-16	5.869336006
P. length	<2.20E-16	2.718369933
structural constituent of cell wall	<2.20E-16	1.965991088
DNA binding	<2.20E-16	1.544242393
GC context	<2.20E-16	0.754548334
structural constituent of ribosome	<2.20E-16	0.597911216
Transcription factor activity	<2.20E-16	0.575348528
hormone activity	<2.20E-16	0.554521432
histamine receptor activity	<2.20E-16	0.553219739
nucleic acid binding	<2.20E-16	0.547145169
Voltage-gated potassium channel activity	<2.20E-16	0.488135064
ligand-dependent nuclear receptor activity	2.33E-12	0.348853859
sequence-specific DNA binding	3.01E-09	0.249491255
RNA binding	1.70E-07	0.193952332
d_S	1.25E-06	0.166616768
chromatin binding	3.29E-06	0.153165936
RNA polymerase II transcription factor activity, enhancer binding	3.63E-06	0.151864242
d_N	6.19E-06	0.144921877
nutrient reservoir activity	0.0004664	0.086779567
transcription corepressor activity	0.0054142	0.054671127
ω	0.0134962	0.043389783
RNA polymerase II transcription factor activity	0.0240022	0.03601352
helicase activity	0.1667501	0.013450833
zinc ion binding	0.198911	0.011715242
transcription activator activity	0.4614908	0.003905081

In italics GO Terms that remain significant after Bonferroni Correction. In Bold functions enriched in pure AARs.

erage (ranging from 6-11×) and their corresponding Ensembl 50 Genes[38]. We compared the genomes of 2 primates (*Homo sapiens* NCBI36 and *Pan troglodytes* CHIMP2.1), 2 rodents (*Mus musculus* NCBIM37 and *Rattus norvegicus* RGSC3.4) and 2 domestic species (*Bos taurus* Btau_3.1 and *Canis familiaris* Canfam 2.0).

For each mammalian genome, we downloaded all the known protein coding genes, with exception of dog and chimp genomes where, in order to gather the largest accurate dataset, we used the "known by projection" set. The repeat analyses are restricted to non-redundant one-to-one orthologues to an equidistant outgroup, dog in the case of rodents and primates, and human for the domestic species. We filtered the genes by keeping the protein corresponding to the longest transcript and excluding all coding sequences that did not begin with a start codon. Finally the number of genes that were screened for repeats in each species was 13,926 human, 11,120 chimpanzee, 13,921 mouse, 10,360 rat, 7,073 cow and 7,834 dog genes.

Orthologous Exons

We downloaded 1,168 orthologous exons alignments including 9 to12 mammalian species, from the OrthoMam database [39]. This is a curated database that contains the amino acid and coding sequence alignments for each particular exon. The inclusion of these alignments allowed studying local AAR expansions without biases due to regional differences in substitution rates and GC context along the whole gene. The exon trees were built using PHYML (substitution model = JTT, estimated proportion of invariable sites, four categories, estimated gamma, initial tree with BIONJ) [40]. Evolutionary rates for each branch where obtained running the *free-ratios* model in PAML 4.1 [41] and keeping d_N, d_S and ω convergent values of 5 replicate runs. Non-convergent or 999 values were not considered in further analyses.

Homo-polymeric Amino-acid Repeats and Purity

As in many previous studies we focused on perfect homopolymeric amino-acid repeats, where we assume that the expansion of a tri-nucleotide by slippage gave birth to the repetition of a single amino-acid motif within the protein. To consider that an amino-acid repeat appeared by polymerase slippage a minimum threshold of 5 units was frequently used in the literature [e.g. [8,26]]. We used a minimum number of 7 units. The reasons for this are, first, to increase the significance level[6] and, second, to increase the chance that a repeat locus shows length polymorphism [42,43].

The *purity* of the nucleotide sequence coding the amino-acid repeat was calculated following the method described by Laidlaw et al. in 2007[8]that is summarized in the equation below;

$$Purity = \frac{(m-n)}{m} \quad (1)$$

where m is the total number of nucleotides coding the amino-acid repeat and n the number of interruptions or nucleotide changes with respect to the canonical codon (the most frequent or most likely to have experienced expansion by slippage). The presence of AARs was considered for each species independently of the presence of that repeat in orthologues.

Summary of parameters and estimates

Each gene was screened for homo-polymeric amino acid repeats within the corresponding protein sequence. The following parameters were calculated:

i) Weighted Average Purity of the Repeats of a Gene: the weighted average of the *purity* estimates of every amino-acid repeat in the protein sequence of a gene. The weighting is based on the length of the coding sequence of each individual repeat, as described in the following equation;

$$Av.Purity = \sum_{i}^{n} \frac{l_i \cdot P_i}{L} \quad (2)$$

where n is the total number of AARs on the protein-coding gene, l is the length in bp of each individual repeat, P is the corresponding Purity and L the sum in bp of the length of all AARs in the gene. This measure allowed us to compare if certain genes contain purer AARs than others. Note that the vast majority of the cases correspond to genes hosting only one AAR. (Additional file 2, Figures SF1 and SF2)

ii) d_N and d_S: sitewise maximum likelihood estimates of d_N and d_S for each orthologous pair were downloaded from Ensembl [38].

iii) GC context (%GC): the GC content of gene after excluding all regions encoding repeats[44]. Similarly, we estimated GC3 as the GC content on third codon positions of the full repeat-free coding dna of the gene or exon. These parameters depict the sequence context in which repeats are born.

Gene Expression data

Microarray data of mouse and human tissues were downloaded from ArrayExpress (E-AFMX-4 and E-AFMX-5)[45]. E-AFMX-4 uses an Affymetrix Custom Array - Novartis Mouse (A-AFFY-39) and E-AFMX-5 uses an Affymetrix Custom Array - Novartis Human (A-AFFY-40). Mapping of Ensembl gene on Affymetrix probesets from these chips was taken from http://biogps.gnf.org/downloads/. E-AFMX-5 also uses an Affymetrix GeneChip Human Genome HG-U133A (A-AFFY-33), whose mapping to Ensembl genes was downloaded from BioMart [46].

We extracted expression data for 5 organs in mouse (cerebral cortex, liver, kidney, testis and heart) and human (brain, liver, kidney, testis and heart). Raw CEL files were renormalized using the package gcRMA [47] of Bioconductor version 2.2[48]. We used the "affinities" model of gcRMA, which uses mismatch probes as negative control probes to estimate the non-specific binding of probe sequences. The normalized values of expression are in log2 scale, which attenuates the effect of outliers. Expression values were averaged between replicates and between multiple probes mapped to a same gene. Probes mapping to more than one gene were discarded.

GO term enrichment

Over and under representation of GO terms [49] was tested by means of a Fisher exact test, using the Bioconductor package topGO version 1.8.1 [50]. The reference set was all Ensembl genes used in the repeats analysis. The GO annotation of Ensembl genes was downloaded from BioMart. The "elim" algorithm of topGO was used, allowing to decorrelate the graph structure of the gene ontology, reducing non-independence problems. Gene ontology categories with a FDR < 20% were reported.

Authors' contributions

FC designed the project, gathered the genomic data, performed most of the evolutionary and statistical analyses, and wrote the original manuscript. JR gathered the expression and gene ontology data, and performed the analysis of GO Term enrichment. MRR supervised the work, provided critical comments about the statistical analyses and the biological discussion, and revised thoroughly the manuscript. All authors have read and approved the final manuscript.

Additional material

Additional file 1

Supplementay Tables. *A PDF file containing additional Tables S1-S11. These tables contain analyses of variance with factors sorted by percentage of explained variance. Further details are provided as footnotes accompanying each table.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-619-S1.PDF]

Additional file 2

A PDF file containing additional Figures SF1-SF5. Further details are provided as footnotes accompanying each Figure.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-619-S2.PDF]

Acknowledgements

We acknowledge funding from Etat de Vaud, Swiss National Science Foundation grant 116798, and the Swiss Institute for Bioinformatics. We would like to thank Nicolas Galtier and Emmanuel Douzery for helping with the exon alignments. We also thank Carolin Kosiol, Nicolas Salamin, Matthew T. Webster, Carles Vilà and anonymous referees for their helpful comments.

References

1. Ellegren H: **Microsatellite mutations in the germline: implications for evolutionary inference.** *Trends in Genetics* 2000, **16(12):**551.
2. Kashi Y, King DG: **Simple sequence repeats as advantageous mutators in evolution.** *Trends in Genetics* 2006, **22(5):**253-259.
3. Kruglyak S, Durrett RT, Schug MD, Aquadro CF: **Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95(18):**10774-10778.
4. Ellegren H: **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5(6):**435-445.
5. Beena T, Koshy HYZ: **The CAG/Polyglutamine Tract Diseases: Gene Products and Molecular Pathogenesis.** *Brain Pathology* 1997, **7(3):**927-942.
6. Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes and disease associations.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(1):**333-338.
7. Fondon JW III, Garner HR: **Molecular origins of rapid and continuous morphological evolution.** *Proceedings of the National Academy of Sciences* 2004, **101(52):**18058-18063.
8. Laidlaw J, Gelfand Y, Ng K-W, Garner HR, Ranganathan R, Benson G, Fondon JW III: **Elevated Basal Slippage Mutation Rates among the Canidae.** *J Hered* 2007, **98(5):**452-460.
9. Zamorzaeva I, Rashkovetsky E, Nevo E, Korol A: **Sequence polymorphism of candidate behavioural genes in *Drosophila melanogaster* flies from 'Evolution canyon'.** *Molecular Ecology* 2005, **14(10):**3235-3245.
10. Nevo E, Beharav A, Meyer RC, Hackett CA, Forster BP, Russell JR, Powell W: **Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel.** *Biological Journal of the Linnean Society* 2005, **84(2):**205-224.
11. Fahima T, Röder MS, Wendehake K, Kirzhner VM, Nevo E: **Microsatellite polymorphism in natural populations of wild emmer wheat, Triticum dicoccoides, in Israel.** *TAG Theoretical and Applied Genetics* 2002, **104(1):**17-29.
12. Hammock EAD, Young LJ: **Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits.** *Science* 2005, **308(5728):**1630-1634.
13. Mularoni L, Veitia RA, Albà MM: **Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats.** *Genomics* 2007, **89(3):**316-325.
14. Alba MM, Santibanez-Koref MF, Hancock JM: **Conservation of polyglutamine tract size between mice and humans depends on codon interruption.** *Mol Biol Evol* 1999, **16(11):**1641-1644.
15. Hancock JM, Worthey EA, Santibanez-Koref MF: **A Role for Selection in Regulating the Evolutionary Emergence of Disease-Causing and Other Coding CAG Repeats in Humans and Mice.** *Mol Biol Evol* 2001, **18(6):**1014-1023.
16. Simon M, Hancock J: **Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins.** *Genome Biology* 2009, **10(6):**R59.
17. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of Positive Selection in Six Mammalian Genomes.** *PLoS Genetics* 2008, **4(8):**e1000144.
18. Preacher KJ: **Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software].** [http://www.quantpsy.org].
19. Studer RA, Penel S, Duret L, Robinson-Rechavi M: **Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes.** *Genome Research* 2008, **18(9):**1393-1402.
20. Galtier N, Duret L, Glémin S, Ranwez V: **GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates.** *Trends in Genetics* 2009, **25(1):**1-5.
21. Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S: **Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors.** *Mol Biol Evol* 1997, **14(10):**1042-1049.
22. Oma Y, Kino Y, Toriumi K, Sasagawa N, Ishiura S: **Interactions between homopolymeric amino acids (HPAAs).** *Protein Science* 2007, **16(10):**2195-2204.
23. Drummond DA, Wilke CO: **Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution.** 2008, **134(2):**341-352.
24. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-Sequence Determinants of Gene Expression in Escherichia coli.** *Science* 2009, **324(5924):**255-258.
25. Sicheri F, Yang DSC: **Ice-binding structure and mechanism of an antifreeze protein from winter flounder.** *Nature* 1995, **375(6530):**427-431.
26. Mularoni L, Guigo R, Albà MM: **Mutation patterns of amino acid tandem repeats in the human proteome.** *Genome Biology* 2006, **7(4):**R33.
27. Brock GJ, Anderson NH, Monckton DG: **Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands.** *Hum Mol Genet* 1999, **8(6):**1061-1067.
28. Jurka J, Pethiyagoda C: **Simple repetitive DNA sequences from primates: Compilation and analysis.** *Journal of Molecular Evolution* 1995, **40(2):**120-126.
29. Loire E, Praz F, Higuet D, Netter P, Achaz G: **Hypermutability of genes in Homo sapiens due to the hosting of long mono-SSR.** *Mol Biol Evol* 2008:111-121.
30. Huntley MA, Clark AG: **Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species.** *Mol Biol Evol* 2007, **24(12):**2598-2609.
31. Ackermann M, Chao L: **DNA Sequences Shaped by Selection for Stability.** *PLoS Genet* 2006, **2(2):**e22.
32. Riley DE, Jeon JS, Krieger JN: **Simple repeat evolution includes dramatic primary sequence changes that conserve folding potential.** *Biochemical and Biophysical Research Communications* 2007, **355(3):**619-625.
33. Legendre M, Pochet N, Pak T, Verstrepen KJ: **Sequence-based estimation of minisatellite and microsatellite repeat variability.** *Genome Research* 2007, **17(12):**1787-1796.
34. O'Dushlaine C, Edwards R, Park S, Shields D: **Tandem repeat copy-number variation in protein-coding regions of human genes.** *Genome Biology* 2005, **6(8):**R69.
35. Hirano Y, Nishimiya Y, Matsumoto S, Matsushita M, Todo S, Miura A, Komatsu Y, Tsuda S: **Hypothermic preservation effect on mammalian cells of type III antifreeze proteins from notched-fin eelpout.** *Cryobiology* 2008, **57(1):**46-51.
36. Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P: **Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution.** *Journal of Proteome Research* 2006, **5(11):**2985-2995.
37. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424(6945):**147-151.
38. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, *et al.*: **Ensembl 2009.** *Nucleic Acids Research* 2009, **37(suppl_1):**D690-697.
39. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M-K, Douzery E: **OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics.** *BMC Evolutionary Biology* 2007, **7(1):**241.
40. Guindon Sp, Gascuel O: **A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood.** *Systematic Biology* 2003, **52(5):**696-704.
41. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** *Mol Biol Evol* 2007, **24(8):**1586-1591.
42. Pupko T, Graur D: **Evolution of Microsatellites in the Yeast Saccharomyces cerevisiae: Role of Length and Number of Repeated Units.** *Journal of Molecular Evolution* 1999, **48(3):**313-316.
43. Weber JL: **Informativeness of human (dC-dA)n·(dG-dT)n polymorphisms.** *Genomics* 1990, **7(4):**524-530.

44. Albà MM, Guigo R: **Comparative Analysis of Amino Acid Repeats in Rodents and Humans.** *Genome Res* 2004, **14(4):**549-554.
45. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, *et al.*: **ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression.** *Nucl Acids Res* 2009, **37(suppl_1):**D868-872.
46. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart - biological queries made easy.** *BMC Genomics* 2009, **10(1):**22.
47. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.** *Journal of the American Statistical Association* 2004, **99(468):**909-917.
48. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5(10):**R80.
49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25(1):**25-29.
50. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22(13):**1600-1607.