Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains

Hugo Lavoie¹, François Debeane¹, Quoc-Dien Trinh¹, Jean-François Turcotte², Louis-Philippe Corbeil-Girard¹, Marie-Josée Dicaire¹, Anik Saint-Denis¹, Martin Pagé¹, Guy A. Rouleau³ and Bernard Brais¹,*

¹Laboratoire de Neurogénétique, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, Canada, ²Département de Biochimie, Université de Montréal, Montréal, Canada and ³Montreal General Hospital, McGill Health Center, McGill University, Montréal, Canada

Received July 14, 2003; Revised and Accepted September 18, 2003

Mutations causing expansions of polyalanine domains are responsible for nine hereditary diseases. Other GC-rich sequences coding for some polyalanine domains were found to be polymorphic in human. These observations prompted us to identify all sequences in the human genome coding for polyalanine stretches longer than four alanines and establish their degree of polymorphism. We identified 494 annotated human proteins containing 604 polyalanine domains. Thirty-two percent (31/98) of tested sequences coding for more than seven alanines were polymorphic. The length of the polyalanine-coding sequence and its GCG or GCC repeat content are the major predictors of polymorphism. GCG codons are over-represented in human polyalanine coding sequences. Our data suggest that GCG and GCC codons play a key role in polyalanine-coding sequence appearance and polymorphism. The grouping by shared function of polyalanine-containing proteins in Homo sapiens, Drosophila melanogaster and Caenorhabditis elegans shows that the majority are involved in transcriptional regulation. Phylogenetic analyses of HOX, GATA and EVX protein families demonstrate that polyalanine domains arose independently in different members of these families, suggesting that convergent molecular evolution may have played a role. Finally polyalanine domains in vertebrates are conserved between mammals and are rarer and shorter in Gallus gallus and Danio rerio. Together our results show that the polymorphic nature of sequences coding for polyalanine domains makes them prime candidates for mutations in hereditary diseases and suggests that they have appeared in many different protein families through convergent evolution.

INTRODUCTION

Since 1996, mutations causing expansions of polyalanine tracts have been documented in nine human diseases and in one mouse strain (1–10). Polyalanine expansions have been found in: synpolydactily (from 15 to 22–29 alanines in HOXD13) (3,11), cleidocranial dysplasia (from 17 to 27 alanines in RUNX2) (5), oculopharyngeal muscular dystrophy (from 10 to 11–17 alanines in PABPN1) (1), familial holoprosencephaly (from 15 to 25 alanines in ZIC2) (2), hand–foot–genital syndrome (from 18 to 26 alanines in HOXA13) (4), blepharo-phimosis/ptosis/epicanthus inversus syndrome type II (from 15 to 30 alanines in FOXL2) (6), X-linked mental retardation and epilepsy (from 12–16 to 20–23 alanines in ARX) (8), X-linked mental retardation with growth hormone deficiency (from 15 to 26 alanines in SOX3) (9) and congenital central hypoventilation syndrome (from 25 to 29 alanines in PHOX2B) (10). The shortest disease-causing expansion of a polyalanine domain was observed in recessive OPMD, in which the addition of a single alanine residue caused a recessive late-onset disease (1). Interestingly, this polymorphism is present in 2% of the general population and, when inherited in a compound heterozygote fashion with a dominant expansion, causes a more severe phenotype (1). Non-pathological polyalanine polymorphisms...
were reported in polyalanine coding sequences in seven other human genes prior to this study: MICA, GPX1, TGFBR1, RPL14, MSH3 and FOXE1 (12–17). Two mechanisms leading to lengthening of polyalanine coding sequences have been proposed: expansions of stretches of GCN codons in PABPN1, ARX, MICA, GPX1, TGFBR1, RPL14 and FOXE1 and duplications of mixed (GCN)_n polyalanine-coding sequences in PABPN1, HOXD13, HOXA13, RUNX2, ZIC2, FOXL2, ARX, SOX3, PHOX2B and MSH3 (1–6, 12–17). These observations prompted us to identify human proteins with domains of five alanines or more and to investigate if variations could be found in their polyalanine-coding sequences.

Knowledge of the functional roles of polyalanine domains in proteins is very limited. Studies have suggested that polyalanine may play a role in transcriptional repression in drosophila kruppel, engrailed and even-skipped, and in human glucocorticoid receptor and octamer binding protein 1 (18–23). In contrast, the deletion of the polyalanine domains in SCIP did not alter its transcriptional activity (24). Together, these observations suggest that polyalanine stretches in these proteins play a secondary role in transcription. Polyalanine domains in silo are believed to form β-pleated sheets in dragline spider silk, moth silk and mollusk shell and tendons (25–27). The interaction between numerous polyalanine domains is known to play a major role in ensuring fiber tensile strength in spider’s silk (25). The biophysical characteristics of homopolymeric alanine stretches in vitro and in silico have been extensively studied but are still somewhat controversial since polyalanine polymers have served as models in protein folding research for both β-sheet and α helical conformations (28–30). Polyalanine peptides of 10 residues or more form stable fibrillar macromolecules resistant to chemical denaturation and enzymatic degradation in vitro (29,31). Although the structure of polyalanine peptides in solution is still debated, a recent study shows that β-sheet structures are favored in hydrophobic physiological conditions and α helical conformation in hydrophobic medium (28).

The appearance and conservation of polyalanine domains has not been extensively studied. The polyalanine domains of HOXA13 and class III POU transcription factors appeared only in mammals (32,33). In the case of HOXA13, it has been shown that the human protein is 35% longer than its zebrafish ortholog. These differences in size are primarily due to accumulation of polyalanine and flanking regions rich in proline and glycine (33). Similarly, the Drosophila ortholog of PABPN1 lacks a polyalanine domain and a large portion of the protein’s N-terminus (34). It was suggested that polyalanine domains in arthropod silk and mollusk shell and tendon have appeared by convergent evolution (35), implying that these domains appeared independently in orthologs and were conserved for functional reasons.

The complete sequence of the human genome was a prerequisite to this study’s identification by data mining of all human proteins with polyalanine domains and allows the screening for polymorphisms of a large set of sequences coding for polyalanine. The increasing number of sequenced eukaryotic genomes further permitted the comparative analyses of species differences in size and differential appearance of polyalanine domains through evolution. This study establishes that polyalanine-coding sequences are frequently polymorphic in the human population, that polyalanine domains are more frequent in specific functional categories, especially in transcription regulators, that they likely appeared by convergent molecular evolution in some protein families and that they are longer in mammals than in other vertebrates.

RESULTS

Sequences coding for polyalanine domains in human are frequently polymorphic

We identified 494 known human proteins with polyalanine domains longer than four alanines in the human refseq protein database (Supplementary Material) (36). These 494 proteins contained a total of 604 polyalanine domains. We selected the 124 proteins with domains longer than seven alanines to investigate if the polyalanine coding sequences are polymorphic in humans. We retrieved genomic coding sequences for 135 polyalanine domains (117 genes) and 98 were successfully amplified by PCR on 42 unrelated DNA samples. Amplification products were run on denaturing gels and samples showing two or more alleles were sequenced. Thirty-one of the 98 sequenced regions (32%) were found to have at least two sequenced-proven alleles. Table 1 lists the genes with polymorphic sequences sorted by the frequency of the most common allele observed. Four of these polymorphisms were previously reported in the literature (Table 1).

Table 1. Examples of polyalanine domains coding sequences that are polymorphic.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Polyalanine Domain</th>
<th>Nucleotide Repeat</th>
<th>Common Allele</th>
<th>Minor Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHOX2B</td>
<td>15</td>
<td>GCN 4</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>FOXE1</td>
<td>10</td>
<td>GCN 3</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>MICA</td>
<td>12</td>
<td>GCN 5</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>GPX1</td>
<td>16</td>
<td>GCN 6</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>TGFBR1</td>
<td>18</td>
<td>GCN 7</td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

In 48% of the polymorphic loci, the minor alleles have a frequency inferior to 5% (Table 1). Codon repeat expansions/contractions and deletions/insertions of larger sequences are both responsible for rare and frequent alleles. It is noteworthy that we also identified new polymorphisms in two genes containing small polyalanine domains of seven alanines (GDF1 and CHD4, data not shown). This raises the possibility that sequences coding for even shorter polyalanine domains could also be polymorphic. Polymorphisms with high frequencies could be predicted by aligning Unigene EST sequences for nine genes (Table 1). We found that four of the nine genes with known pathogenic lengthening of a polyalanine tract are polymorphic in the general population: PABPN1, RUNX2, ZIC2 and HOXA13 (1,2,4,5) (Table 1). This further underlines that sequences encoding a polymorphic polyalanine domain are prime candidates for diseases mapped to the same chromosomal region.

The size of the polyalanine coding sequence and its repeat content appear as the major predictors of polymorphism. Most of the polymorphic sequences (62%) and very few non-polymorphic sequences (11%) code for more than 13 alanines or contain repeats of more than five codons (Fig. 1A). Polymorphism of polyalanine domains coded by mixed GCN codons seems to be mainly influenced by the length of the sequence. In fact, the mixed (GCN)_n polyalanine-coding
sequences are more frequently polymorphic when they code for more than 13 alanines (Fig. 1A, arrows).

**GCG codons are over-represented in polyalanine-coding sequences**

There is a statistically significant over-representation of GCG codons in polyalanine-coding sequences compared with its average relative occurrence in a control set of open reading frames (57/63% in polyalanine compared with 11/12% of GCG usage in 597 control sequences) and according to the codon usage database (11% of GCG codon usage) (37) (Fig. 1B). This contrasts with GCC codons that are as frequently used in polyalanine-coding sequences as in control sequences (Fig. 1B). GCA and GCT codons are overall under-represented in polyalanine-coding sequences (Fig. 1B). A noticeable increase in GCG and GCC codons usage is observed as the size of the polyalanine domains increase while GCA and GCT codon usage tend to decrease with length (Fig. 1B).

The analysis of the contribution of codon repeats in polyalanine-coding sequences demonstrates that GCG and GCC repeats become predominant at a length of three codons or more (Fig. 1C). Most of the sequences are not coded by identical repeated codons. Of the 154 sequences coding for eight or more alanines, only 12% are entirely or in part coded by tracts of eight or more identical codons (Fig. 1C). Therefore, long polyalanine domains are usually coded by mixed (GCN) sequences. Since GCGGCG and GCCGCC are complementary, their over-representation in polyalanine-coding sequences points toward a particular involvement of these GC-rich sequences in the appearance and lengthening of sequences coding for polyalanine domains.

**The longest polyalanine domains are found in eukaryotes**

To establish if polyalanine domain-containing proteins share functional roles, we identified all such proteins in three near complete eukaryotic genomes: Homo sapiens, Drosophila melanogaster, and Caenorhabditis elegans (Supplementary Material). Figure 2A, C and E presents the distributions of polyalanine domains according to their size in each species. The number of polyalanine domains encoded in each genome is variable (Fig. 2A, C and E). Based on previous estimates of gene number in these species, polyalanine-containing proteins account approximately for 1.4, 3.9 and 0.80% in H. sapiens, D. melanogaster and C. elegans, respectively (38–42).

### Table 1. Polymorphisms observed in gene sequences coding for domains of eight or more alanines.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>ID*</th>
<th>Polyalanine length</th>
<th>Variation type†</th>
<th>Number of alleles</th>
<th>Frequency of the most frequent allele</th>
<th>Polymorphic sequence</th>
<th>Number of repeats‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRP14δ</td>
<td>6727</td>
<td>9</td>
<td>E</td>
<td>2</td>
<td>99%</td>
<td>GCA</td>
<td>9, 10</td>
</tr>
<tr>
<td>SOX21</td>
<td>11166</td>
<td>13</td>
<td>E</td>
<td>2</td>
<td>99%</td>
<td>GCC</td>
<td>6, 8</td>
</tr>
<tr>
<td>FOXF2</td>
<td>2295</td>
<td>9</td>
<td>E</td>
<td>2</td>
<td>99%</td>
<td>GCC</td>
<td>9, 10</td>
</tr>
<tr>
<td>C14orf4</td>
<td>64207</td>
<td>10</td>
<td>C</td>
<td>2</td>
<td>99%</td>
<td>GCTGCCGCGCGCGCCG</td>
<td>1, 0</td>
</tr>
<tr>
<td>ZIC2</td>
<td>7546</td>
<td>15</td>
<td>d</td>
<td>2</td>
<td>99%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>4, 0</td>
</tr>
<tr>
<td>ABT1δ</td>
<td>463</td>
<td>13</td>
<td>C</td>
<td>2</td>
<td>99%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>4, 0</td>
</tr>
<tr>
<td>HOXA13</td>
<td>3209</td>
<td>18</td>
<td>d</td>
<td>2</td>
<td>98%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>1, 0</td>
</tr>
<tr>
<td>PABPN1</td>
<td>8106</td>
<td>10</td>
<td>E</td>
<td>2</td>
<td>98%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>6, 7</td>
</tr>
<tr>
<td>TBL1°</td>
<td>6907</td>
<td>8</td>
<td>C</td>
<td>2</td>
<td>98%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>5, 4</td>
</tr>
<tr>
<td>FBS1</td>
<td>64319</td>
<td>19</td>
<td>i</td>
<td>2</td>
<td>98%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>1, 2</td>
</tr>
<tr>
<td>HOXD11δ</td>
<td>3237</td>
<td>13</td>
<td>E</td>
<td>2</td>
<td>97%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>6, 7</td>
</tr>
<tr>
<td>SLC12A2</td>
<td>6558</td>
<td>15</td>
<td>C</td>
<td>2</td>
<td>97%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>5, 7</td>
</tr>
<tr>
<td>ZIC3</td>
<td>7547</td>
<td>10</td>
<td>E</td>
<td>2</td>
<td>97%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>8, 9</td>
</tr>
<tr>
<td>PHOX2B</td>
<td>8929</td>
<td>20</td>
<td>C</td>
<td>2</td>
<td>97%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>4, 3</td>
</tr>
<tr>
<td>SGC8</td>
<td>6443</td>
<td>8</td>
<td>E</td>
<td>2</td>
<td>97%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>4, 5</td>
</tr>
<tr>
<td>MZF3</td>
<td>4150</td>
<td>9</td>
<td>d</td>
<td>3</td>
<td>95%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>1, 0</td>
</tr>
<tr>
<td>ZNF358</td>
<td>55136</td>
<td>17</td>
<td>d</td>
<td>2</td>
<td>95%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>0, 1</td>
</tr>
<tr>
<td>MAP3K4δ</td>
<td>4216</td>
<td>9</td>
<td>E</td>
<td>2</td>
<td>93%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>9, 10</td>
</tr>
<tr>
<td>SOX1</td>
<td>6656</td>
<td>8</td>
<td>C</td>
<td>2</td>
<td>92%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>3, 5</td>
</tr>
<tr>
<td>RPL14δ</td>
<td>9045</td>
<td>8</td>
<td>E</td>
<td>2</td>
<td>91%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>8, 9</td>
</tr>
<tr>
<td>HLBXBδ</td>
<td>3110</td>
<td>16</td>
<td>C</td>
<td>2</td>
<td>90%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>9, 11</td>
</tr>
<tr>
<td>RUNX2</td>
<td>860</td>
<td>17</td>
<td>E</td>
<td>2</td>
<td>88%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>1, 0</td>
</tr>
<tr>
<td>FOXD1</td>
<td>2297</td>
<td>8</td>
<td>E</td>
<td>2</td>
<td>87%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>4, 5</td>
</tr>
<tr>
<td>DMRTA2</td>
<td>63950</td>
<td>9</td>
<td>C</td>
<td>2</td>
<td>69%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>6, 8</td>
</tr>
<tr>
<td>KCNN2</td>
<td>3781</td>
<td>8</td>
<td>E</td>
<td>2</td>
<td>66%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>4, 5</td>
</tr>
<tr>
<td>EOMES</td>
<td>8320</td>
<td>14</td>
<td>E</td>
<td>2</td>
<td>65%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>7, 9</td>
</tr>
<tr>
<td>CCI3δ</td>
<td>9584</td>
<td>12</td>
<td>C</td>
<td>3</td>
<td>63%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>5, 6, 8</td>
</tr>
<tr>
<td>TGFBR1ε</td>
<td>7046</td>
<td>9</td>
<td>E/C</td>
<td>3</td>
<td>63%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>6, 9, 10</td>
</tr>
<tr>
<td>PRDM12</td>
<td>59335</td>
<td>12</td>
<td>E/C</td>
<td>4</td>
<td>55%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>5, 6, 8</td>
</tr>
<tr>
<td>FOXE1δ</td>
<td>2304</td>
<td>14</td>
<td>E/C</td>
<td>3</td>
<td>55%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>7, 9, 11</td>
</tr>
<tr>
<td>MSH3ε</td>
<td>4437</td>
<td>12</td>
<td>D/d</td>
<td>4</td>
<td>53%</td>
<td>GCCGCGCGCGCGCGC</td>
<td>1, 4, 5, 6</td>
</tr>
</tbody>
</table>

*LocusLink ID; †Expansion/contraction (E/C) and deletion/duplication (D/d); ‡number of units of repeated sequence responsible for each allele; ‡gene sequence predicted to be polymorphic based on EST sequence alignment; °region unable to sequence; and ‡gene sequence previously found to be polymorphic.
Interestingly, no domain longer than 20 alanines was uncovered in the analyzed genomes. This maximum length was observed in D. melanogaster ovo and asx and in H. sapiens PHOX2B. In C. elegans, the length of polyalanine domains never exceeded 14 alanines (Fig. 2E).

We also established that the genome of S. cerevisiae contains only 39 polyalanine-containing proteins (about 0.65% of S. cerevisiae genes) and none had a domain of more than 10 alanines (data not shown). The number of polyalanine domains was even smaller in prokaryotes. Polyalanine-coding genes account for a mean of 0.74% of archael genomes and 0.84% of eubacterial genomes. In fact, 10 completely sequenced archael and 31 completely sequenced eubacteria contained respectively 162 and 746 polyalanine-coding genes, and the size of the domains never exceeds nine alanines. Therefore, domains longer than nine alanines appear to be exclusively found in eukaryotes. Although 85% of human proteins contain a single polyalanine domain, some contain more: 48 proteins have two; 15 have three; four have four; and five contain five polyalanine domains. Proteins bearing more than one domain are also found in D. melanogaster (563 have one, 108 have two, 49 have three, 18 have four, 18 have five, one has seven and one contains 10 polyalanine domains) and in C. elegans (122 have one, 12 have two, two have three and two contain five polyalanine domains).

Proteins with polyalanine domains are often involved in transcription regulation

The polyalanine-containing proteins from C. elegans, D. melanogaster and H. sapiens were grouped by high level molecular function categories according to Gene Ontology Consortium functional classification (www.geneontology.org; Fig. 2B, D and F). Human, Drosophila and worm proteins were divided for analysis based on a size threshold of eight alanines (Fig. 2B, D and F). For all three organisms the ‘binding’ molecular function is the most represented in polyalanine-containing proteins followed by ‘transcription regulator’ and ‘enzyme’ annotations. The patterns of functions distribution obtained for H. sapiens, D. melanogaster and C. elegans polyalanine-containing proteins are very similar (Fig. 2B, D and F). The significant enrichment of polyalanine-containing proteins in ‘binding’ molecular function is mostly explained by the presence of DNA binding transcription regulators (Fig. 2B, D and F). The organization of the ontology tree leads to this observation. In fact, the ontologies are organized so that one low level function could be linked to more than one high level annotation. Transcription regulators account for 7, 9 and 10% of H. sapiens, D. melanogaster and C. elegans complete...
Strikingly, 36% of *H. sapiens*, 35% of *D. melanogaster* and 25% of *C. elegans* polyalanine-containing proteins bearing a domain of five to seven alanines are annotated as ‘transcription regulators’. The latter functional group is significantly enriched in both the five to seven alanines and more than seven alanines polyalanine-containing proteins compared with whole proteomes for the three organisms (Fig. 2B, D and F). The groups of proteins with more than seven alanines domains contain an even greater proportion of transcription regulators (Fig. 2B, D and F), but this increased proportion is only significant for *D. melanogaster* polyalanine-containing proteins. It is also noteworthy that polyalanine-containing proteins having a protein binding function often bind transcription factors (data not shown). Finally, the number of genes annotated as ‘enzyme’ is significantly smaller in five to seven alanines and more than seven alanines polyalanine-containing proteins from *H. sapiens* and *D. melanogaster* (Fig. 2B and D).
Polylanine domains appeared by convergent evolution in different protein families

Polylanine domains are more frequent in certain protein families. For example, polylanine domains are present in 17% of proteins from the homeobox superfamily (35 out of 212 homeobox proteins retrieved with PSIBLAST in the human refseq database) (43,44). To investigate the appearance and conservation of polylanine domains within protein lineages we first established which families had many members in our set of proteins. There are 21 protein families represented by three or more members in the human data set, 13 of which are transcription factor families: CLCN, DMRT, GATA, FOX, GPR, HOX, IRX, KCN, MAPK, NRG, PBX, POU, PPP1R, PRDM, RPL, SLC, SMARC, SOX, TBX, TLE and ZIC. The over-representation of polylanine domains in certain protein families raises the possibility that they may have either appeared in a common ancestral protein or been acquired independently.

Polylanine domains have been studied mostly in proteins of the HOX family (33). It has been suggested that all homeobox DNA-binding proteins may have a common origin and form a single superfamily (43,45,46). In vertebrates 4 HOX gene clusters (A–D) are believed to have arisen by duplications of a single ancestral cluster followed by divergence (47–50). Phylogenetic analysis suggests that polylanine domains appeared recently and in only some orthologs within paralogous groups of the HOX family (Fig. 3A). For example, phylogenetic analysis of the paralogous groups HOXA13 and HOXD13 does not support a shared common ancestral origin for their polylanine domains but an independent appearance specific to the mammalian lineage (Fig. 3A, in red and green). The same phenomenon is observed for the paralogs HOXA11 and HOXD11 (Fig. 3A, yellow and purple). In addition to these paralogs, HOXD8 has also acquired a polylanine domain. This phenomenon of independent appearance is also observed in the GATA-zinc-fingers family where mammalian GATA6, GATA4 and GATA1 have acquired a polylanine sequence independently (Fig. 3C). Convergence is not only observed at the level of paralogous groups, it is also seen between orthologs as is exemplified by even-skipped proteins from various species (Fig. 3E). In this particular case, a domain appeared in all vertebrate evx2 (Fig. 3E, blue branches) and a second domain is unique to mammalian evx2 (Fig. 3E, red). A third event occurred only in drosophila eve of all known insects even-skipped proteins (Fig. 3E, green). Finally, a fourth event of polylanine appearance is specific to nematodes VAB-7 (Fig. 3E, yellow). The same phenomenon of evolutionary convergence within orthologous proteins was also observed in runt and engrailed (data not shown). This further supports the independent convergent appearance of polylanine domains in proteins sharing similar functional role. It is noteworthy that polylanine domains appeared only in mammals in HOXD8, HOXA11, HOXD11, HOXA13, HOXD13, GATA1, GATA4 and GATA6 phylogenetic trees.

The possible convergent evolutionary appearance of polylanine domains in proteins of the same family is further supported by their different positions relative to functional domains or to other conserved regions. Figure 3B depicts the position of polylanine insertions (arrowheads) within sequence alignments of HOX13 and HOX11 paralogous groups. The same was observed in an alignment of GATA4 with GATA6 (Fig. 3D). This phenomenon is even more clearly illustrated by the N-terminal position of the polylanine domain in VAB-7, the ortholog of even-skipped in nematodes, compared with C-terminal insertions in vertebrates evx2 and drosophila eve (Fig. 3F). The relative position of polylanine domains relative to the DNA binding domain in different transcription factor superfamilies is clearly variable. For example, 89% of polylanine are located in N-terminus in homeobox-containing proteins, 71% are in C-terminus in SOX-containing proteins and 73% are in C-terminus in forkhead box domain-containing proteins.

Polylanine domains are longer in mammals than in other vertebrates

The independent appearance of polylanine domains in related proteins and the observation that polylanine domains arose mainly in mammals (Fig. 3A and C) led us to assess polylanine conservation between human and mouse (Mus musculus), human and chicken (Gallus gallus) and human and zebrafish (Danio rerio). Figure 4A depicts the high level of conservation of polylanine domains length between H. sapiens and M. musculus. Figure 4B and C shows that polylanine domains are poorly conserved between H. sapiens and G. gallus and H. sapiens and D. rerio. The Pearson correlation coefficient (r) between human and the three species polylanine length decreases from M. musculus to D. rerio (Fig. 4A–C). The percentage of identity of pairwise blast alignments seems not to correlate with the conservation of polylanine tracts. Even highly conserved protein segments (80–100% identity; open triangles) do not share the same polylanine length (Fig. 4A–C). Polylanine size is clearly larger on average in mammals than in other vertebrates. In fact, the mean polylanine length difference between H. sapiens and M. musculus, H. sapiens and G. gallus and H. sapiens and D. rerio was respectively —1.06±2.22 alanines, —4.53±3.79 alanines and —6.10±2.22 alanines. Figure 4D illustrates the changes in size of 28 polylanine domains in 23 proteins sharing orthologs in human, chicken and zebrafish. Polylanine domains in chicken and zebrafish proteins are shorter overall than their orthologs in human. Only two out of the 28 domains studied are conserved in the three organisms (NRF1 and MAF) and only one is longer in zebrafish than in human and chicken (PBX1) (Fig. 4D).

DISCUSSION

This study identified by data mining 494 known or predicted human proteins containing a total of 604 polylanine domains longer than four alanines. We demonstrate that sequences coding for polylanine domains are frequently polymorphic in the human genome. In fact, 31 out of 98 tested polylanine-coding sequences were polymorphic. We established that the presence of a GCN repeat longer than five or a sequence coding for polylanine stretches longer than 13 alanines is a good predictor of polymorphism. Based on these criteria, 82 polylanine-coding sequences of our set are likely to be polymorphic in human. Of these, 28 were not tested because either their genomic sequence was not available or they fell under our
eight alanines size threshold for screening. Polymorphism of polyalanine-coding sequences is probably more widespread because some shorter sequences were also found to have more than one allele and rarer polymorphisms may have been missed because of the limited size of our DNA panel (84 chromosomes). The observation that rare polymorphisms exist in four of the nine genes with known disease-causing mutations further emphasizes the importance of screening similar sequences in diverse pathologies. Furthermore, the finding that the minor allele of \( P A B P N 1 \) carried by 2% of the population can act as both a modulator of severity for dominant OPMD and a recessive OPMD mutation raises the possibility that the rare

![Figure 3](https://academic.oup.com/hmg/article-abstract/12/22/2967/606641/2973)
polymorphisms observed in this study may play similar roles in other traits (1).
The selection of these polyalanine-coding genes as candidates for numerous conditions is expedited by their known chromosomal location in most cases. It has been shown that other types of mutations in \textit{HOXD13}, \textit{HOXA13}, \textit{RUNX2}, \textit{ZIC2} and \textit{FOXL2} cause phenotypes similar to expansions of the polyalanine domain in these proteins (2,4–6,51). In this light, of particular interest are 10 genes with polyalanine domains in which other types of mutations have been identified in human: \textit{APC}, \textit{ELN}, \textit{EMX2}, \textit{FOXE1}, \textit{ZIC3}, \textit{AGPS}, \textit{AR}, \textit{IRS2}, \textit{OFD1} and \textit{TBX3} (www.ncbi.nlm.nih.gov/omim). Interestingly, mice knock-out for \textit{Hoxd13}, \textit{Hoxa13} and \textit{Runx2} show phenotypes similar to human polyalanine expansions in these genes (52–55). This phenotypic overlap between knock-out mice and human disease allows the flagging of at least 34 human genes with sequences coding for domains of more than seven alanines with knock-out models as prime candidate genes for

Figure 4. Conservation of polyalanine domain length between human and three vertebrate species. Polyalanine domain length in \textit{H. sapiens} proteins was plotted against corresponding polyalanine domain length in \textit{M. musculus} (A), \textit{G. gallus} (B) and \textit{D. rerio} (C). Each polyalanine domain pair was associated with a protein pair. The polyalanine domains were ranked by their associated protein pair percentage of identity in blast alignments and they were given a distinct symbol corresponding to the category they fell into (50–59, 60–69, 70–79, 80–89 and 90–100% of identity). \textit{N} corresponds to the number of polyalanine domain pairs between two species and \textit{R} is the Pearson correlation coefficient between human and species of interest polyalanine sizes. (D) Changes in size of 28 polyalanine domains found in 23 different proteins that share an ortholog in \textit{H. sapiens}, \textit{G. gallus} and \textit{D. rerio}. Series are labeled according to human protein name.
human diseases with phenotypic overlap. For example, this raises the possibility that polyalanine expansions in SLC12A2 (17 alanines in man) may cause deafness in humans as point mutations have been found in the Shaker-with-syndactylism (sy) deaf mouse strain (56). Interestingly, this strain also has syndactyly, an observation found in two of the already described polyalanine diseases (56). The ages of onset of the diseases caused by expansions of polyalanine domains found to date suggests that such mutations may be involved in traits and diseases that span human life. Further studies will establish the functional importance of these relatively rare polymorphisms, but they already allow the identification of a large set of mapped genes as prime candidates for disease-causing mutations considering that the chromosomal localizations are available for the majority of human genes with sequences coding for polyalanine domains.

The mechanism responsible for polyalanine-coding sequence polymorphisms or mutations is still debated. Two hypotheses have been proposed: expansion/contractions of repeated sequences through slippage during replication and deletions/duplications of mixed (GCN) ₙ sequences due to unequal crossovers during meiosis. Although our sequence data demonstrate that most changes in size are due to changes in the number of repeated codons (77%, 24/31), only unequal crossover can explain both the presence of expanded repeat size and mixed (GCN) ₙ sequence mutations at the same locus (57).

This work clearly shows that GCG codons are significantly over-represented in polyalanine-coding sequences. In fact, GCG codons are on average three times more frequent in polyalanine-coding sequences than in control coding sequences. Furthermore, GCG and GCC repeats are more frequently represented in polyalanine coding sequences than repeats of GCA or GCT. These results together with the observation that most of polyalanine polymorphisms are associated with GCG or GCC repeats suggest that these codons may be particularly important in the appearance and polymorphism of polyalanine domains. GCGGCC and GCGGCC being complementary, they can form stable hairpins or other higher order structures that may play a key role in this process and explain the relative abundance of GCG/GGC and GCC/GCG repeats in polymorphic sequences coding for polyalanine. It has been shown by many groups that biophysical properties of GCC/GCG repeats observed in fragile X mental retardation makes them good substrate for replication slippage. In fact, oligos consisting of CGG repeats (corresponding to GCG polyalanine repeats) fold in vitro as either stable Hogsteen-bounded tetrahelical structures or as hairpins (58–62). It is also noteworthy that, except for Friedrich ataxia, diverse frames of GCN/CGN triplet reiteration are the substrates for all other pathogenic triplet repeat mutations (58). The observation that CGG, CGG and CGA repeats are in fact prone to proliferation in protein coding sequences also points to their importance in the appearance and expansion of amino acid reiterations (63).

Using Gene Ontology molecular function classes we grouped polyalanine-containing proteins from three phylogenetically distant organisms. It is striking that polyalanine proteins in the three eukaryotes studied are mostly involved in transcription regulation. Proteins with polyalanine domains more than seven alanines are even more frequently involved in transcription than proteins with shorter domains. Transcription regulators are at least five times more abundant in polyalanine-containing proteins than in the human proteome. This enrichment of polyalanine domains in proteins sharing the same molecular function in distantly related organisms indicates that polyalanine probably participates in function. Previous studies have suggested that, indeed, polyalanine domains play a role in transcriptional repression (18–23,64,65). It is, however, unknown how the hydrophobic polyalanine domains participate in transcription. It is also noteworthy that the relative importance of each molecular function was similar between human and drosophila.

This study also demonstrates that polyalanine domains have appeared independently in paralogous genes. Homeobox containing proteins account for the largest single functional group of polyalanine proteins: 49/494 in human and 43/758 in drosophila. It is believed that all homeoboxes descend from the same ancestor gene (43,45,46). We have shown that polyalanine domains have appeared independently in many of these proteins often in a different position relatively to their homeobox domain. The repeated observation of this phenomenon in many other protein families lead us to conclude that evolutionary pressure has led to this convergence. Convergent molecular evolution has been postulated as an explanatory principle for the presence of polyalanine domains in spider silks and mollusk shell and tendon (35). In these two examples polyalanine domains have appeared in very distant species, suggesting that the acquired functional properties were responsible for the evolutionary pressure and led to convergence at the molecular level. Molecular convergence was also invoked in the appearance of a repetitive sequence in antifreeze glycoproteins shared between phylogenetically distant Antarctic notothenioid fishes and Arctic cod (66). Another example of molecular convergence was documented for glycine-rich domains in cyanobacterial and eukaryotic RNA binding proteins (67). These various examples illustrate that repeated codons may be good substrates for convergent molecular evolution. These sequences probably promote their own variation, allowing for the appearance of new domains.

The evolutionary pressure responsible for the convergent appearance of polyalanine domains in eukaryotes is unknown. As previously discussed, polyalanine domains are frequent in transcription regulators and they may play a role in protein–protein interactions (3). Our hierarchical classification of polyalanine-containing proteins involved in ‘binding’ uncovered that many were transcription factors while another important group was involved in transcription factor binding. These observations led us to search the literature for interaction data involving polyalanine-containing proteins. We found that many proteins with polyalanine domains interact with each other. For example, GATA4 and GATA6 interact to regulate cardiac development (68). The most impressive polyalanine protein network uncovered is the gro/TLE transcription corepressor family (TLE1, 3 and 4) that interact with polyalanine-containing transcription factors RUNX2, FOXD1/2/3, EN1, EVX2, PAX7, IRX3, NKX6.1, NKX2D, HEY2 and UTX1 to form a regulatory network involved in various aspects of vertebrate development (69–79). It was further found that HEY2 homo and heterodimerize with HAND1 and HAND2 that also contain a polyalanine-domain (80). This polyalanine...
protein network can be enlarged with the interaction of EN1 with PBX1, 2 and 3 (81). TLEs interact with the eh1 motif of many of these proteins and the role of the polyalanine domains in these interactions has not been studied. Stable polyalanine hydrophobic protein–protein interactions could play a role, considering the known biophysical properties of polyalanine stretches (28–31). It is possible that polyalanine protein–protein interfaces may stabilize interactions between transcription regulators. Thus, new or more stable protein–protein interactions introduced by the appearance of polyalanine domains contribute to the evolutionary convergence observed in numerous protein families.

In most cases, when convergent evolution is observed, polyalanine domains are exclusively found in the mammalian lineage. Fast evolution of polyalanine domains in mammals has previously been reported for HOXA13 and POU class III genes (32,33). We thus carried out a study of polyalanine length conservation among vertebrates. We comprehensively assessed the conservation of polyalanine domains between human and three vertebrate species and we observed good conservation between human and mouse but poor conservation between human and chicken and human and zebrafish. We also followed the evolution of 28 individual polyalanine domains and found that, except for three of them, all were shorter in chicken and zebrafish with respect to human domain lengths. The clear trend is one of increase in the polyalanine domain sizes between H. sapiens > G. gallus > D. rerio. Together, this data argues for a lengthening of most of polyalanine domains in terrestrial vertebrates and especially in the mammalian lineage.

In conclusion, the roles of the polyalanine domains in evolution, development and diseases clearly need to be further investigated. The boundaries between monogenic and multifactorial traits are becoming blurred by the identification of genetic modifiers such as the (GCG)_7 recessive PABPN1 allele in OPMD (1,82,83). Mutations causing polyalanine expansions in some of the 494 human proteins identified are likely to contribute to monogenic or polygenic traits having onsets at all stages of life. Therefore, gene and protein databases should be annotated to identify the occurrence of such domains and their variations. This study emphasizes the importance of screening sequences coding for polyalanine domains for mutations and functional polymorphisms and design strategies to establish the function and evolutionary role of these domains.

MATERIALS AND METHODS

Identification of proteins with polyalanine domains and alanine codon usage analysis

We used a program consisting of a regular expression search to retrieve the ID and polyalanine length of proteins containing one or more polyalanine domains (ActivePerl 5.6.1; www.activeperl.com). The protein databases of H. sapiens, M. musculus, G. gallus, D. rerio, D. melanogaster and C. elegans available through NCBI’s taxonomy resource were analyzed (updated 1 September 2001; www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Sequence redundancy was eliminated by automatic means using cd-hit and the data set was manually curated by using unique identifiers provided by LocusLink and Refseq for H. sapiens (www.ncbi.nlm.nih.gov/RefSeq/; www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/), FlyBase for D. melanogaster (http://flybase.bio.indiana.edu/) and WormBase for C. elegans (www.wormbase.org/) (36,84–86). The polyalanine size threshold of five was based on the observation that the smallest polymorphic sequences previously observed were the five alanine domains of MICA and GPX1 and on a study of polyglutamine size evolution (12,13,87). Chromosomal localization, Unigene identifiers, representative protein and mRNA sequences for human loci coding for polyalanine were retrieved by using Stanford SOURCE retrieval tool (88); http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceSearch). Contigs of Unigene ESTs for tested genes were constructed using LASERGENE2000 Seqman software. All the data were stored in a Microsoft Access database and most are provided as Supplementary Material at www.genome.org. Alanine codon usage was determined by using a perl script (ActivePerl 5.6.1; www.activeperl.com) extracting ORF from mRNA sequences and counting alanine codons in ORF sequence. The control gene set was obtained through RefSeq database (www.ncbi.nlm.nih.gov/RefSeq/; Supplementary Material). A one-way ANOVA was used to test the null hypothesis of no difference between polyalanine and control alanine codon usage.

Detection of polymorphisms in human sequences coding for polyalanine domains

The genomic DNA from 42 unrelated participants in our OPMD study were genotyped. All had signed an informed consent. Twenty-six were carriers of a dominant PABPN1 OPMD mutation while the others were healthy family members. The 124 human proteins with polyalanine domains of eight or more alanines were chosen for polymorphism search. The genomic sequence was retrieved by tBLASTn for 117 of these genes (135 domains). Primers based on these genomic sequences were designed using LASERGENE2000 PrimerSelect software. Fragments were amplified by PCR with either 7-deaza-dGTP or betaine at 1M concentration (1,89). Amplicons were run on a 5% polyacrylamide denaturing gel. Gels were dried and autoradiograms were obtained (1). Samples that demonstrated two alleles were reamplified and the purified PCR products were sequenced using an ABIPrism automated DNA sequencer (Applied Biosystems).

Functional annotations of proteins, phylogenetic analyses and pairwise comparisons of polyalanine domains

Polyalanine-containing proteins were functionally classified by using molecular function annotations available through the Gene Ontology Consortium web page (www.geneontology.org/). These annotations were based on EBI annotations for human, Flybase for D. melanogaster and Wormbase for C. elegans (86,90,91). Given the structure of the ontology tree, one single protein could have more than one annotation (www.geneontology.org/). Chi-square calculations were used to test for statistically significant differences in molecular function distributions between whole proteomes and proteins with polyalanine domains. The identification of all known
proteins from a specific family was achieved by searching the human refseq database with PSI-BLAST (blastpgp) and the consensus sequence for the motifs of interest as a query sequence (BLAST version 2.1.2) (43,44,92).

Known protein sequences from families of interest were retrieved by running BLASTp against the GenBank database. Sequences obtained were aligned with the CLUSTAL W program and the alignments were manually edited (93). For tree construction, a distance approach was used (PHYLIP 3.6) (93). PROTDIST was used to calculate a distance matrix (94). A gamma distribution was applied to correct for unequal rates of change at different amino acid positions. The tree topology was finally inferred using NEIGHBOUR (95). The final unrooted trees were edited using Treeview. Bootstrapping was carried out (1000 replicates).

Pairwise comparisons of polyalanine domains between vertebrate species was done using BLASTp with a threshold E value of 1E – 20 and a threshold alignment length of 200 amino acids (BLAST version 2.1.2) (44). We assessed the conservation of the polyalanine tract by blasting a query file containing all human polyalanine proteins against a database of all proteins from the species of interest (mouse, chicken and zebrafish). The query was also done by running blast with three files containing all polyalanine proteins of the three species against a database of all available human proteins. The resulting pairwise alignments were manually curated for overall protein percent identity and presence of corresponding polyalanine domains between aligned protein pairs. Polyalanine domains were considered orthologous if protein sequence identity was observed on at least one side of the aligned polyalanine domains. If more than one domain was present in a pairwise BLAST alignment, they were considered separately.

SUPPLEMENTARY MATERIAL
Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS
We especially want to thank Franz Lang for his help and suggestions concerning philogenies. H.L. received a scholarship from the AFM and the Neuromuscular Research Partnership Program of the Canadian Institute of Health Research (CIHR), the Muscular Dystrophy Association of Canada (MDAC) and the Canadian Amyotrophic Lateral Sclerosis Association.

REFERENCES
2978 Human Molecular Genetics, 2003, Vol. 12, No. 22


