

# Simple sequence in brain and nervous system specific proteins

Melanie A. Huntley, Sanaa Mahmood, and G. Brian Golding

**Abstract:** We examined sequences expressed in the brain and nervous system using EST data. A previous study including sequences thought to have neurological function found a deficiency of simple sequence within such sequences. This was despite many examples of neurodegenerative diseases, such as Huntington disease, which are thought to be caused by expansions of polyglutamine tracts within associated protein sequences. It may be that many of the sequences thought to have neurological function have other additional, non-neurological roles. For this reason, we examined sequences with specific expression in the brain and nervous system, using EST expression data to determine if they too are deficient of simple, repetitive sequences. Indeed, we find this class of sequences to be deficient. Unexpectedly, however, we find sequences expressed in the brain and nervous system to be consistently enriched for histidine-enriched simple sequence. Determining the function of these histidine-rich regions within brain-specific proteins requires more experimental data.

*Key words:* amino acid repeats, homopeptides, simple sequence, triplet repeat diseases, nervous system proteins, brain-specific proteins.

**Résumé :** Les auteurs ont examiné les séquences exprimées dans le cerveau et le système nerveux en exploitant les données EST. Une étude antérieure portant entre autres sur des séquences supposées avoir une fonction neurologique avait révélé l'absence de microsatellites au sein de telles séquences. Cette constatation allait à l'encontre de plusieurs cas de maladies neurodégénératives, telles que la maladie de Huntington, que l'on croit dues à l'expansion de suites de glutamines au sein des protéines impliquées. Il se pourrait que plusieurs de ces séquences, auxquelles on attribue une fonction neurologique, pourraient avoir, en plus, des rôles non-neurologiques. C'est pourquoi les auteurs ont examiné des séquences exprimées spécifiquement dans le cerveau ou le système nerveux, sur la base des données EST, afin de déterminer si elles aussi seraient dépourvues de microsatellites. Il s'avère que cette classe de séquences est en effet dépourvue de microsatellites. De façon inattendue, cependant, les auteurs ont noté au sein des séquences exprimées dans le cerveau et le système nerveux un enrichissement pour des suites d'histidines. La fonction de telles régions riches en histidine au sein de protéines exprimées dans le cerveau nécessitera davantage de travaux pour en déterminer la signification.

*Mots clés :* suites d'acides aminés répétés, homopeptides, microsatellite, maladies associées à des répétitions trinuécléotidiques, protéines du système nerveux, protéines spécifiques du cerveau.

[Traduit par la Rédaction]

## Introduction

Protein sequences are far from a random arrangement of amino acids. Function and history have a great influence on the composition of peptide sequences. It seems intuitive that proteins involved in similar functions would share similar features in their amino acid sequence. Likewise, orthologous sequences show sequence conservation. Even much broader groups, such as developmental proteins, can be shown to possess patterns characteristic of that group (Karlin and Burge 1996; Huntley and Golding 2004).

One of these unusual sequence features is the presence of excess simple sequence in proteins (Wootton and Federhen 1993). Simple sequences can range from highly biased homopolymer tracts and regions enriched primarily for one amino acid, to larger, more complex repetitive structures. Simple repetitive protein sequences are found in all domains of life; however, they are particularly abundant within eukaryotic proteins (Karlin and Burge 1996; Marcotte et al. 1999; Huntley and Golding 2000; Sim and Creamer 2002).

Within the eukaryotes, most homopolymer tracts have been investigated more thoroughly owing to their relative ease of detection. Karlin and Burge (1996) found that both short and long homopeptides are more frequent in developmental proteins than in other classes of proteins. They also found that many proteins containing multiple long homopeptide sequences were involved in nervous system disease and development. Indeed, Huntington disease (Duyao et al. 1993; Snell et al. 1993; Kiebertz et al. 1994), Kennedy disease (also known as spinal and bulbar muscular atrophy (La Spada et al. 1991)), dentatorubral pallidolusian atrophy (Li et al. 1993; Burke et al. 1994; Koide et al. 1994; Nagafuchi

Received 30 July 2004. Accepted 3 December 2004.  
Published on the NRC Research Press Web site at  
<http://genome.nrc.ca> on 6 April 2005.

Corresponding Editor: T.E. Bureau.

**M.A. Huntley, S. Mahmood, and G.B. Golding,<sup>1</sup>**  
Department of Biology, McMaster University, 1280 Main  
Street West, Hamilton, ON L8S 4K1, Canada.

<sup>1</sup>Corresponding author (e-mail: [golding@mcmaster.ca](mailto:golding@mcmaster.ca)).

et al. 1994), and several spinocerebellar ataxias (Banfi et al. 1994; Kawaguchi et al. 1994; Pulst et al. 1996; David et al. 1997; Zhuchenko et al. 1997; Nakamura et al. 2001; Silveira et al. 2002) contain CAG repeats, which encode polyglutamine tracts.

To investigate this association with repetitive sequence, we previously conducted a survey of neurological and developmental proteins from *Homo sapiens* and *Drosophila melanogaster* (Huntley and Golding 2004). Our results confirmed that developmental proteins are indeed enriched for simple sequences but that sequences with neurological function are not. However, many of those sequences considered to be neurological proteins may not be specific to the brain and nervous system. Further, many of the proteins involved in the neurodegenerative disorders may have a normal, non-pathogenic function that remains elusive.

Therefore, to study sequences specific to the brain and nervous system we used EST expression data. In this study, we examine ESTs from the brain and nervous system, which may not have a known function, to determine whether sequences expressed specifically in these tissues are enriched for simple sequences.

## Materials and methods

All human (*Homo sapiens*), mouse (*Mus musculus*), frog (*Xenopus laevis*), and zebrafish (*Danio rerio*) EST expression data were collected from NCBI using the UniGene database. Only entries with associated protein sequences were used for further analysis. The expression data for many entries within UniGene include multiple tissues. Often genes expressed in the brain also have expression elsewhere. To examine proteins specific to the brain and nervous system we sampled those UniGene entries with expression data exclusively in the target tissues.

In this way, tissue-specific databases were created for the brain and also for the brain and nervous system (heart, kidney, and testis databases were created as controls). The brain-specific database for humans contained only protein sequences expressed solely in the brain (based on the expression data provided in the UniGene database, which may change with the addition of data from future gene expression studies). Likewise, the brain and nervous system database contained only sequences expressed in the brain and (or) nervous system. Sequences expressed both in the target tissues and other tissues were excluded.

Each database was then filtered to remove redundant duplicates, isozymes, and ancient duplications. This was done by performing a BLAST search (Altschul et al. 1997) to screen for similar proteins within the tissue-specific database. All proteins that had a BLAST expect value less than 0.75 were then pairwise aligned using ALIGN (Myers and Miller 1988). The smaller of any 2 sequences with a percent identity greater than 20% (e.g., the percent identity between hemoglobin and myoglobin) was thrown away, as it was considered to be too recently evolutionarily related. In this way, we retained the larger protein of any related pair of sequences.

Proteins that did not have brain-specific expression were used to create 100 comparison databases for the brain-specific database. Fifty of these comparison databases con-

tained sequences with lengths within 5% of the brain-specific protein lengths; the other half were within 10%. These comparison databases were created by sampling from the non brain-specific proteins until a sequence within 5% (or 10%) of the length of each individual brain-specific sequence was found. Thus, each comparison database had nearly identical composition to the brain-specific database based on the number and length of sequences. This was done for each tissue-specific database, such that the comparison databases served as a statistical measure of patterns observed in the tissue-specific databases.

Following the methodology of Huntley and Golding (2004), we performed local BLAST searches using 100-residue-long homopolymers of each amino acid to determine how common simple repetitive regions similar to homopolymers were in these databases. The number of BLAST hits with expect values less than or equal to 0.01 in the tissue-specific databases were compared with those found from 100 comparison databases and the corresponding distributions of hits were visualized by box plots for each tissue-specific comparison (Figs. 1–5).

The frequencies of all the amino acids were calculated for each tissue-specific database, and in the case of *H. sapiens* and *M. musculus* for all of the non-redundant protein sets. This was done to determine if a bias in amino acid composition might be associated with the presence or absence of repetitive regions within the databases (Tables 1–4).

## Results

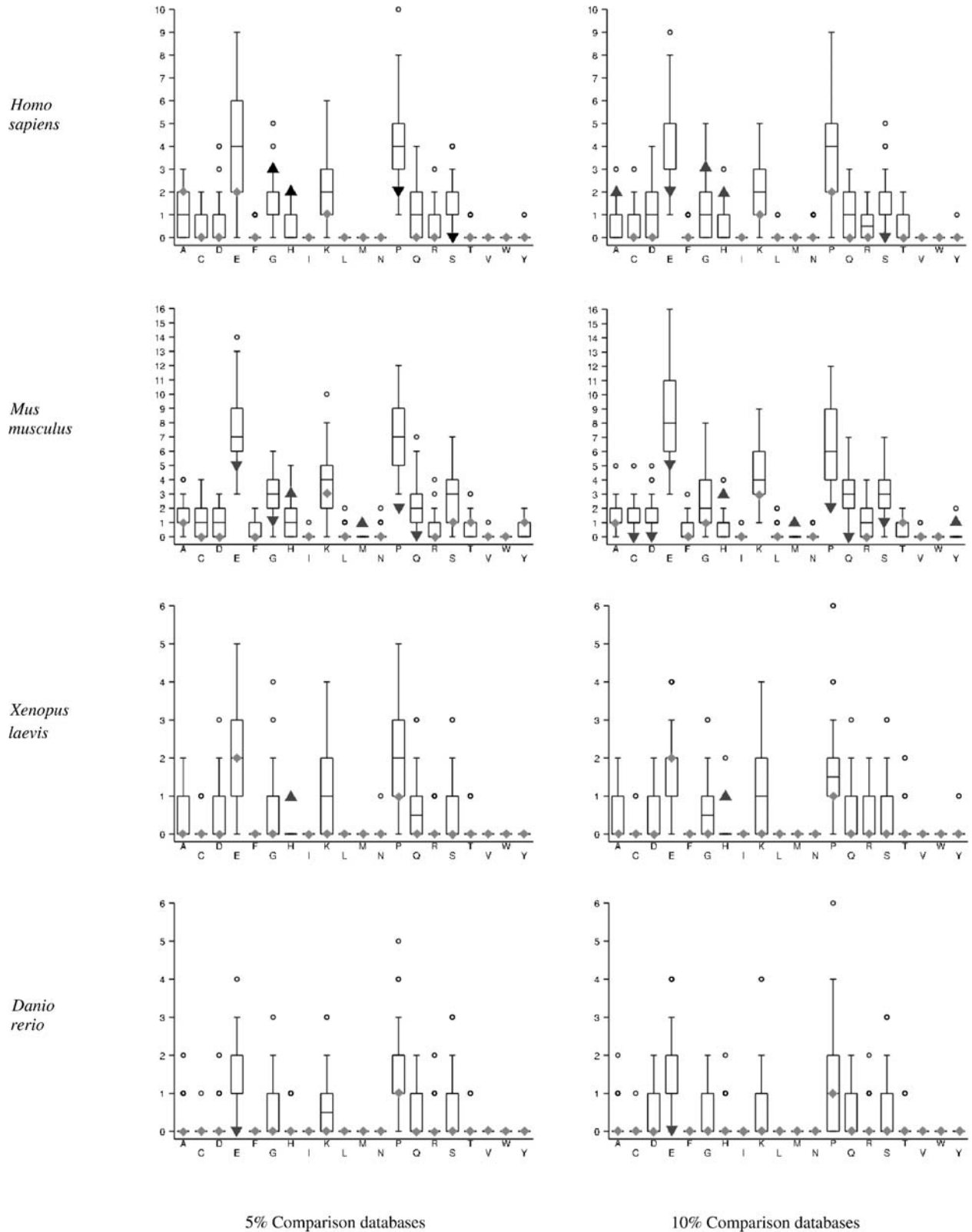
Tables 1–4 display the amino acid frequencies in the tissue-specific databases for *H. sapiens*, *M. musculus*, *X. laevis*, and *D. rerio*, respectively. Despite the smaller sizes of the *X. laevis* and *D. rerio* tissue-specific databases, all 4 species show consistency in the 2 most frequent amino acids in the brain and neurological databases: leucine and serine. Indeed, this pattern is observed broadly except in the *X. laevis* heart-specific database, where glutamic acid and serine are most frequent.

Figure 1 shows the results for the brain-specific proteins. Sequences similar to polyhistidine were over-represented in the brain-specific proteins of *H. sapiens*, *M. musculus*, and *X. laevis*. This was the only enrichment consistent for 3 out of the 4 species. Interestingly, histidine is not overly abundant within brain-specific proteins based on amino acid frequencies. However, it is slightly more frequent in brain-specific sequences than overall for *H. sapiens* and *M. musculus*, as seen in Tables 1 and 2.

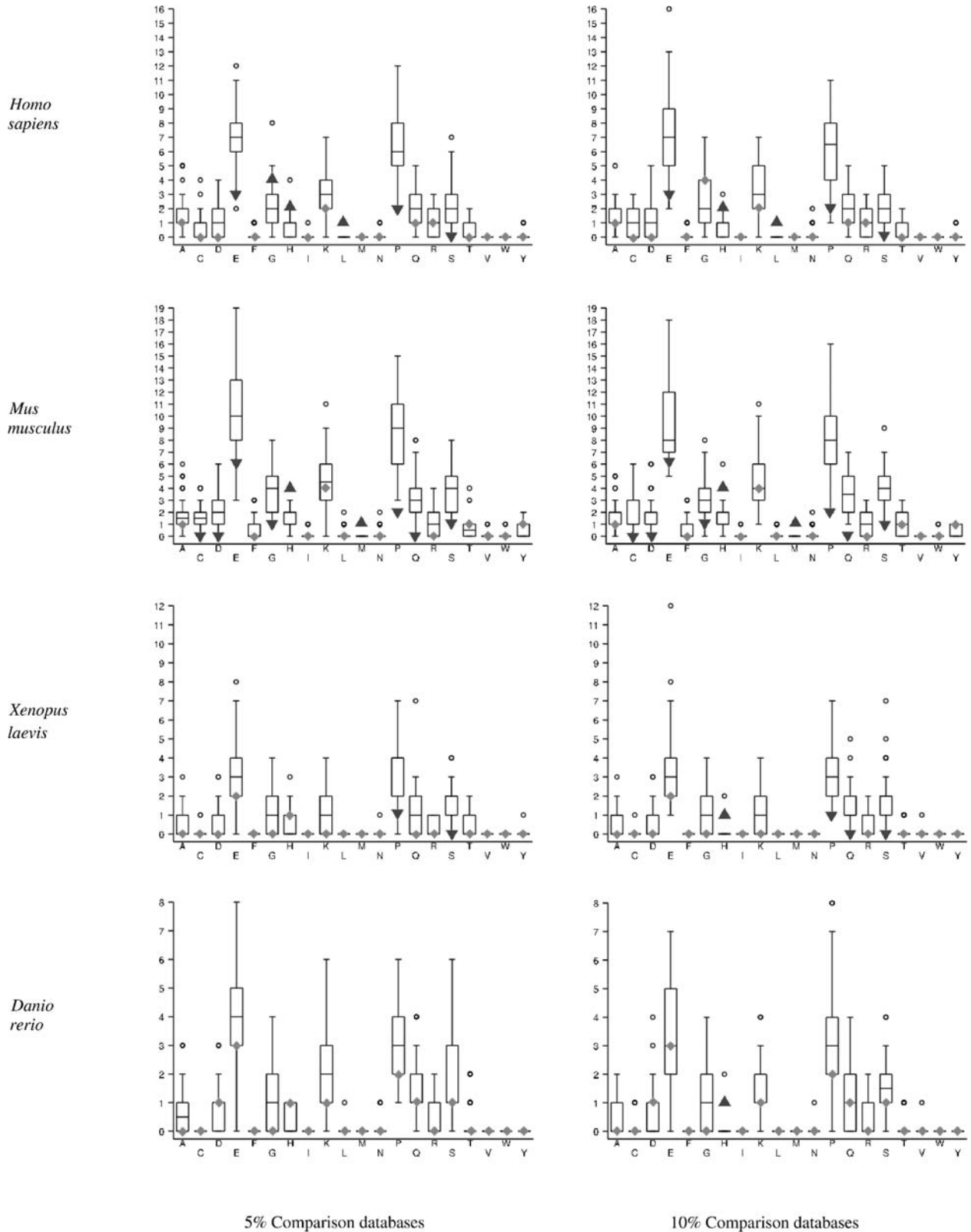
In contrast, sequences similar to polyglutamic acid are under-represented in brain-specific proteins of *H. sapiens*, *M. musculus*, and *D. rerio*. The 2 mammalian species also show under-representation for polyproline and polyserine. This is despite serine being the most frequent amino acid both among brain-specific proteins and overall in these 2 organisms.

A similar trend is found for the neurological proteins in Fig. 2. Polyhistidine is over-represented in the nervous system proteins of all 4 species. Polyproline and polyserine are again depleted in *H. sapiens* and *M. musculus*, and also in *X. laevis*. Polyglutamine is depleted in *M. musculus* and

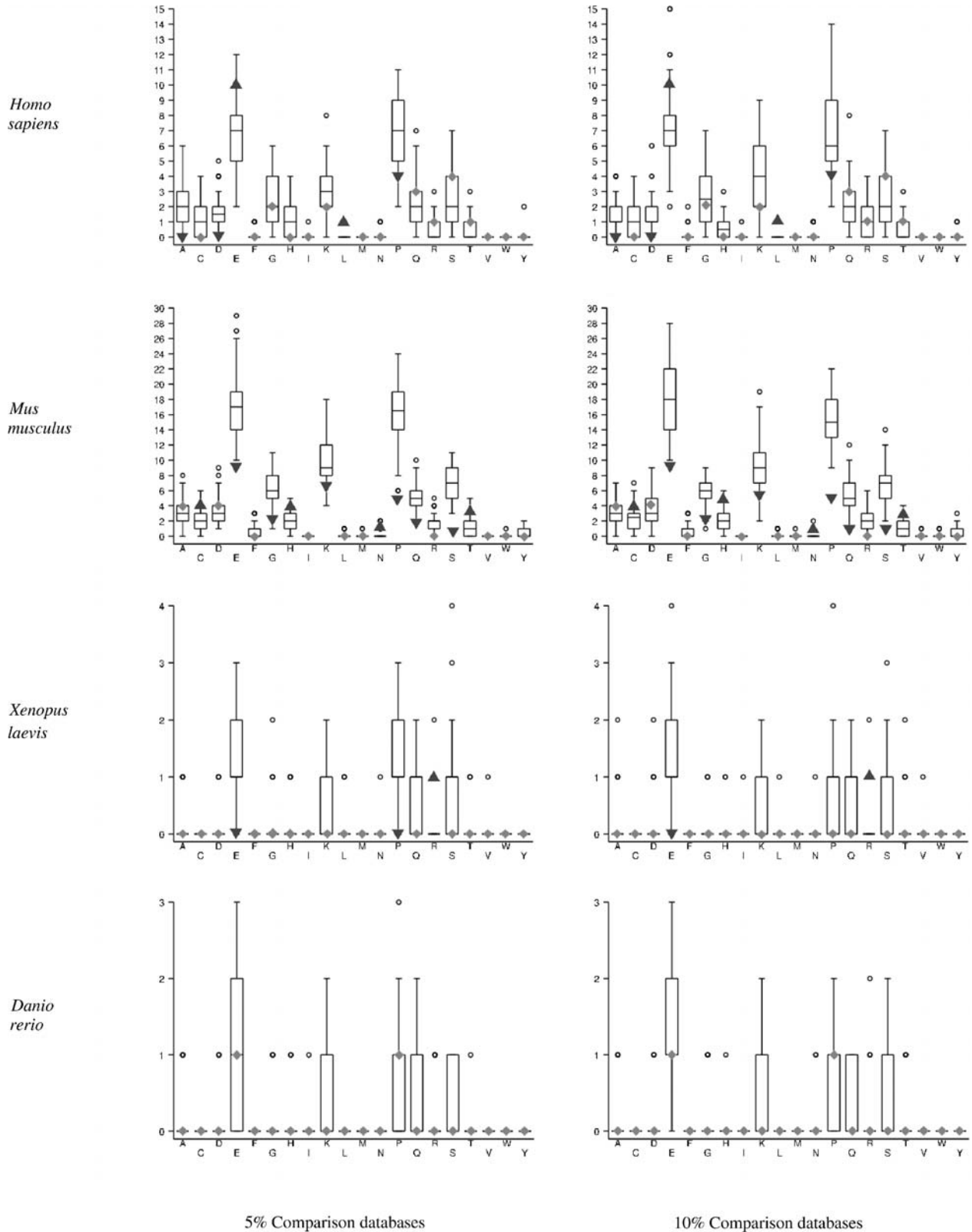
**Fig. 1.** The distribution of significant BLAST hits among brain-specific and comparison databases (5% and 10%). Diamonds and arrowheads denote results for the brain-specific database; box and whiskers represent the distribution of hits for the comparison databases. Arrowheads pointing up highlight when the number of significant BLAST hits within the brain database was above the inner-quartile range of the comparison databases; arrowheads pointing down highlight when they were below the inner-quartile range.



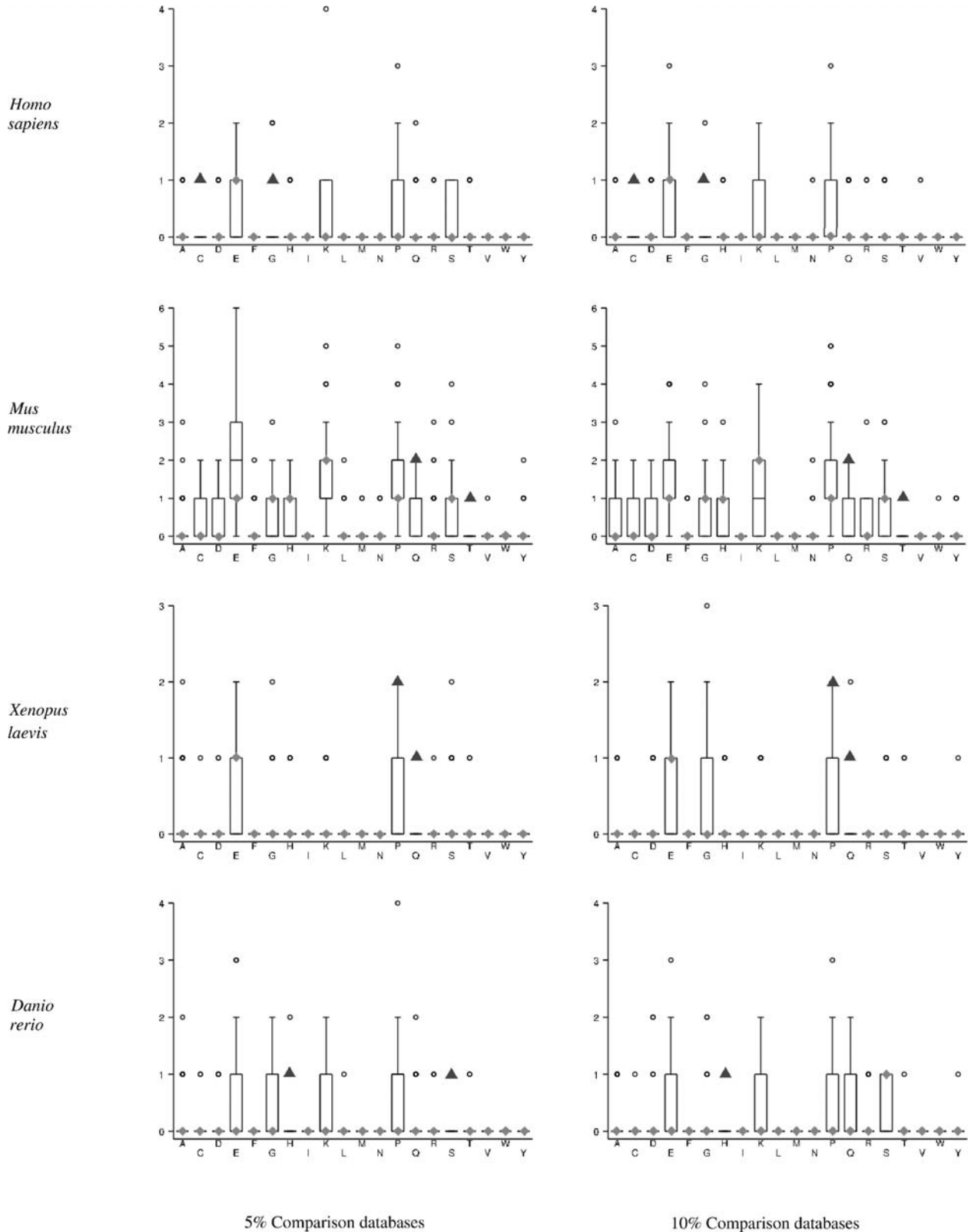
**Fig. 2.** The distribution of significant BLAST hits among nervous system specific and comparison databases (5% and 10%). Diamonds and arrowheads denote results for the nervous system specific database; box and whiskers represent the distribution of hits for the comparison databases. Arrowheads pointing up highlight when the number of significant BLAST hits within the nervous system data-base was above the inner-quartile range of the comparison databases; arrowheads pointing down highlight when they were below.



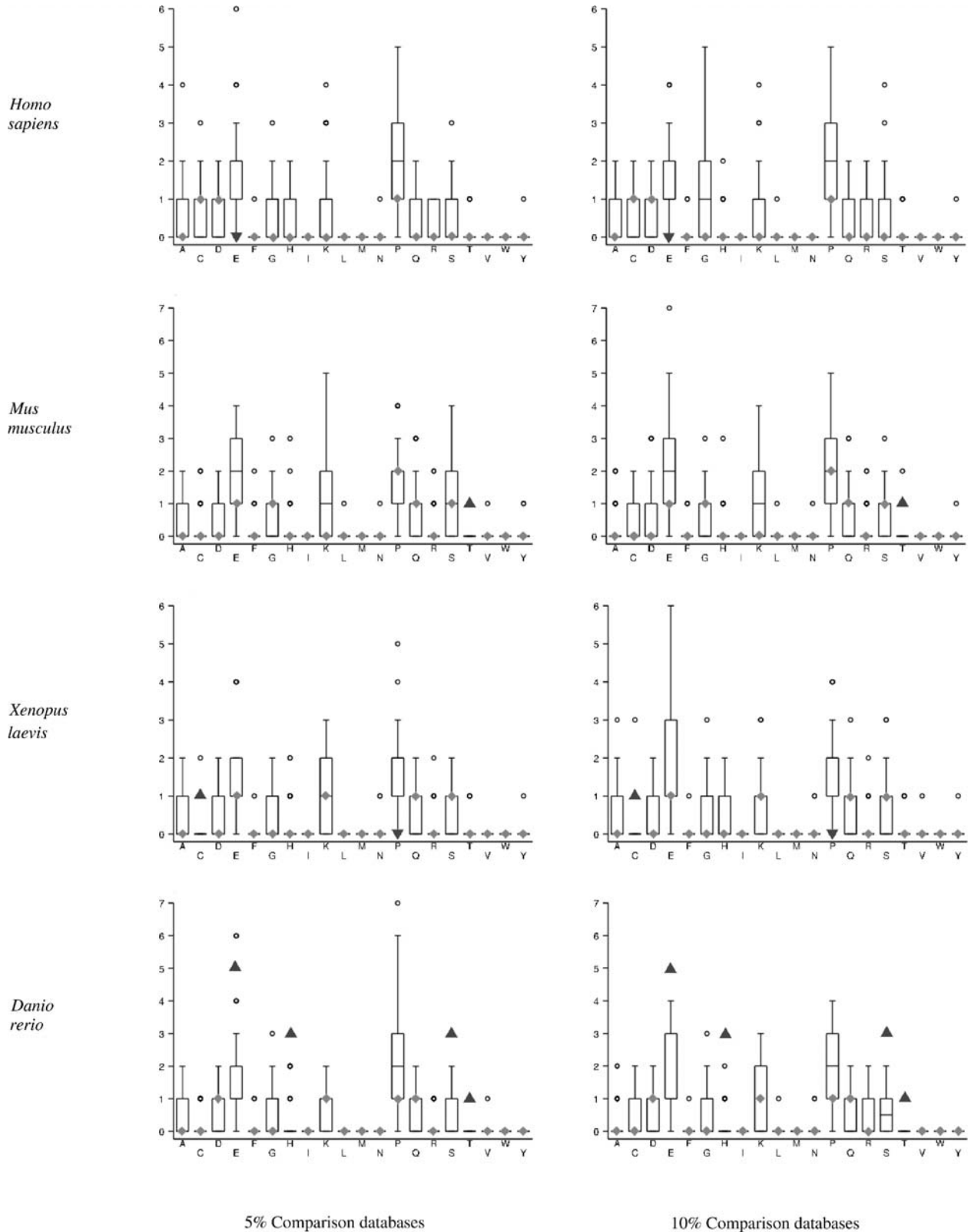
**Fig. 3.** The distribution of significant BLAST hits among testis specific and comparison databases (5% and 10%). Diamonds and arrowheads denote results for the testis specific database; box and whiskers represent the distribution of hits for the comparison databases. Arrowheads pointing up highlight when the number of significant BLAST hits within the testis database was above the inner-quartile range of the comparison databases; arrowheads pointing down highlight when they were below.



**Fig. 4.** The distribution of significant BLAST hits among heart-specific and comparison databases (5% and 10%). Diamonds and arrowheads denote results for the heart-specific database; box and whiskers represent the distribution of hits for the comparison databases. Arrowheads pointing up highlight when the number of significant BLAST hits within the heart database was above the inner-quartile range of the comparison databases; arrowheads pointing down highlight when they were below.



**Fig. 5.** The distribution of significant BLAST hits among kidney-specific and comparison databases (5 and 10%). Diamonds and arrowheads denote results for the kidney-specific database; box and whiskers represent the distribution of hits for the comparison databases. Arrowheads pointing up highlight when the number of significant BLAST hits within the kidney database was above the inner-quartile range of the comparison databases; arrowheads pointing down highlight when they were below.



**Table 1.** The frequency of amino acids among *Homo sapiens* non-redundant proteins.

Amino acid	Tissue					
	All	Brain	Neurological	Testis	Heart	Kidney
A	0.070	0.076	0.073	0.064	0.070	0.070
C	0.022	0.026	0.027	0.028	0.043	0.029
D	0.048	0.037	0.040	0.042	0.040	0.044
E	0.071	0.059	0.059	0.066	0.066	0.057
F	0.035	0.038	0.039	0.037	0.035	0.046
G	0.064	0.074	0.072	0.061	0.064	0.064
H	0.026	0.030	0.029	0.027	0.026	0.028
I	0.042	0.040	0.041	0.046	0.042	0.046
K	0.058	0.050	0.049	0.061	0.049	0.047
L	0.099	0.103	0.104	0.097	0.097	0.109
M	0.022	0.022	0.021	0.023	0.019	0.023
N	0.035	0.030	0.032	0.039	0.030	0.038
P	0.063	0.065	0.068	0.064	0.063	0.065
Q	0.048	0.043	0.044	0.049	0.059	0.041
R	0.057	0.066	0.063	0.058	0.059	0.061
S	0.083	0.090	0.087	0.088	0.086	0.082
T	0.053	0.052	0.053	0.055	0.049	0.051
V	0.060	0.059	0.059	0.054	0.060	0.058
W	0.013	0.017	0.016	0.014	0.016	0.017
Y	0.026	0.022	0.023	0.026	0.029	0.027
<b>Total amino acids</b>	5078841	34045	61522	67874	4174	13905
<b>Total distinct proteins</b>	13078	148	232	239	14	58

**Table 2.** The frequency of amino acids among *Mus musculus* non-redundant proteins.

Amino acid	Tissue					
	All	Brain	Neurological	Testis	Heart	Kidney
A	0.068	0.066	0.066	0.059	0.056	0.063
C	0.022	0.028	0.027	0.025	0.018	0.025
D	0.049	0.045	0.045	0.045	0.040	0.040
E	0.069	0.062	0.061	0.066	0.062	0.054
F	0.038	0.040	0.039	0.038	0.032	0.043
G	0.061	0.067	0.067	0.058	0.061	0.066
H	0.027	0.028	0.028	0.028	0.030	0.027
I	0.044	0.045	0.045	0.049	0.049	0.046
K	0.056	0.051	0.052	0.062	0.052	0.052
L	0.104	0.101	0.101	0.102	0.094	0.100
M	0.022	0.023	0.023	0.025	0.024	0.022
N	0.036	0.033	0.033	0.039	0.035	0.038
P	0.058	0.063	0.062	0.060	0.063	0.062
Q	0.048	0.043	0.044	0.049	0.043	0.040
R	0.055	0.058	0.059	0.054	0.046	0.051
S	0.084	0.088	0.090	0.088	0.112	0.091
T	0.054	0.054	0.055	0.056	0.086	0.070
V	0.063	0.061	0.061	0.058	0.061	0.066
W	0.013	0.015	0.015	0.014	0.014	0.017
Y	0.027	0.027	0.027	0.026	0.024	0.027
<b>Total amino acids</b>	2412332	84224	113585	198477	24406	18437
<b>Total distinct proteins</b>	3240	271	365	643	68	54

*X. laevis*, whereas polyglutamic acid is depleted only in *H. sapiens* and *M. musculus*.

The testis-specific database had no consistent enrichment for repetitive sequences among the 4 taxa (see Fig. 3). How-

ever, it did have more instances of enrichment than the brain and nervous system databases (14 cases where the number of significant BLAST hits within the testis database was above the inner-quartile range of the comparison databases;

**Table 3.** The frequency of amino acids among *Xenopus laevis* non-redundant proteins.

Amino acid	Tissue				
	Brain	Neurological	Testis	Heart	Kidney
A	0.057	0.059	0.046	0.054	0.063
C	0.024	0.024	0.039	0.009	0.020
D	0.056	0.055	0.060	0.067	0.047
E	0.075	0.072	0.060	0.132	0.061
F	0.043	0.042	0.042	0.037	0.045
G	0.065	0.066	0.064	0.046	0.066
H	0.025	0.028	0.024	0.021	0.025
I	0.053	0.050	0.059	0.045	0.060
K	0.066	0.063	0.048	0.067	0.064
L	0.086	0.085	0.090	0.082	0.096
M	0.025	0.026	0.028	0.023	0.026
N	0.042	0.041	0.056	0.057	0.046
P	0.050	0.053	0.041	0.040	0.052
Q	0.036	0.038	0.043	0.056	0.041
R	0.053	0.054	0.046	0.042	0.045
S	0.077	0.080	0.080	0.084	0.084
T	0.054	0.056	0.061	0.049	0.053
V	0.065	0.063	0.067	0.059	0.061
W	0.013	0.012	0.010	0.005	0.012
Y	0.034	0.033	0.038	0.025	0.033
<b>Total amino acids</b>	10338	12048	2067	3230	21522
<b>Total distinct proteins</b>	25	30	6	9	46

**Table 4.** The frequency of amino acids among *Danio rerio* non-redundant proteins.

Amino acid	Tissue				
	Brain	Neurological	Testis	Heart	Kidney
A	0.070	0.064	0.054	0.067	0.063
C	0.023	0.021	0.018	0.025	0.023
D	0.046	0.049	0.058	0.049	0.051
E	0.059	0.068	0.076	0.053	0.069
F	0.047	0.037	0.036	0.040	0.040
G	0.070	0.064	0.048	0.058	0.058
H	0.025	0.026	0.026	0.031	0.031
I	0.049	0.045	0.044	0.049	0.049
K	0.055	0.060	0.056	0.050	0.059
L	0.096	0.094	0.117	0.090	0.095
M	0.024	0.024	0.023	0.029	0.026
N	0.037	0.038	0.033	0.047	0.041
P	0.055	0.056	0.057	0.058	0.056
Q	0.034	0.041	0.037	0.044	0.045
R	0.052	0.060	0.064	0.061	0.050
S	0.083	0.088	0.098	0.084	0.088
T	0.054	0.055	0.052	0.056	0.057
V	0.070	0.067	0.061	0.060	0.059
W	0.014	0.012	0.009	0.013	0.011
Y	0.037	0.031	0.032	0.036	0.029
<b>Total amino acids</b>	5066	12664	1800	4549	26146
<b>Total distinct proteins</b>	14	31	3	10	53

12 and 11 such cases for the brain and nervous system databases, respectively). All but *D. rerio* showed an under-representation of polyproline in the testis-specific sequences. Fig. 4 shows that the heart-specific databases were most en-

riched for repetitive sequences, though the only shared excess was polyglutamine in *M. musculus* and *X. laevis*. Finally, the kidney-specific databases in Fig. 5 were not consistently enriched for sequences in more than 2 taxa.

## Discussion

Unlike developmental proteins, which are known to be enriched for repetitive simple sequences, brain and nervous system specific proteins are not found to be enriched for repetitive simple sequences. In particular, they are not enriched for polyglutamine, despite the many neurodegenerative diseases associated with polyglutamine (La Spada et al. 1991; Snell et al. 1993; Banfi et al. 1994; Kawaguchi et al. 1994; Nagafuchi et al. 1994; Pulst et al. 1996; David et al. 1997; Zhuchenko et al. 1997; Nakamura et al. 2001). This confirms previous results of Huntley and Golding (2004).

Interestingly, there is an over abundance of sequences enriched with histidine simple sequence in brain and nervous system specific sequences. A previous study using only *H. sapiens* and *D. melanogaster* proteins thought to have neurological function also noted enrichment for histidine simple sequence, as well as alanine simple sequence (Huntley and Golding 2004). Alanine simple sequence was not found to be enriched in brain or nervous system sequences in any of the 4 vertebrate taxa used in this study. This may suggest that enrichment for histidine simple sequence is a characteristic of the nervous system and neurological class of proteins.

Studies have identified testis- and sperm-specific sequences to be rapidly evolving in comparison to other tissue-specific sequences (Torgerson et al. 2002; Winter et al. 2004). While it has been demonstrated that repetitive regions enjoy a higher rate of evolution than non-repetitive regions within a protein (Huntley and Golding 2000), the rapid rate of evolution seen in testis-specific sequences must be facilitated by some other mechanism, since our results indicate an underrepresentation of repetitive sequence within the testis-specific sequences of *H. sapiens* and *M. musculus*.

Similarly, several groups have identified sequences with brain and nervous system specific expression to be among the most slowly evolving sequences (Kuma et al. 1995; Hurst and Smith 1999; Duret and Mouchiroud 2000; Winter et al. 2004; Zhang and Li 2004). Hence, it might not be surprising to find a lack of enrichment for repetitive sequences in this class of proteins. However, the consistent enrichment for histidine across all 4 taxa is intriguing.

Typically it is thought that even when simple sequence is conserved, the predominant amino acid can still vary (Huntley and Golding 2000; Sim and Creamer 2004). In this case, it seems that some characteristic peculiar to histidine is causing its repetitive abundance within the repetitive sequences of brain and nervous system specific proteins. Furthermore, histidine is not among the more frequent amino acids within these proteins, so their aggregation within repetitive regions is even more curious.

The imidazole sidechain of histidine allows it to change from neutral to positive charge. The flexibility this confers results in histidine taking up positions both on the surface and within the protein. This also makes histidine residues ideal for catalytic function and binding sites. For these reasons histidine has been likened to being ambidextrous (Creighton 1993). Studies have shown histidine to have a major role in the binding of phosphate groups (Loomis et al. 1997), hormones (Cugini et al. 1992), and iron (Rogers et al. 1977). Additionally, histidines have been noted to be essential for the function, typically including binding, of several

types of proteins (Kery et al. 1986; Pelton and Ganzhorn 1992; Samuel et al. 1993; Rittig et al. 2002). The transcription factors of the class III *POU* genes, which are thought to be important in neural development, also contain histidine-rich simple sequences (Sumiyama et al. 1996). However, this enrichment was only conserved within the mammalian sequences. Indeed, of all the brain and nervous system specific sequences containing histidine-enriched regions in this study, those with known function were typically transcription factors, or otherwise involved in DNA binding.

The essential binding properties of histidine may be a possible explanation for the enrichment of histidine within brain and nervous system specific sequences; however, in most studies, only one or a few histidine residues were present or necessary.

Another possible explanation for the overrepresentation of histidine-rich regions in proteins of the brain may be histamine biosynthesis. The neurotransmitter histamine is produced from the decarboxylation of histidine. Because of the blood brain barrier, histamine in the brain is produced locally. Proteins with histidine-rich regions might be a good source of histidine for histamine biosynthesis, since such simple regions tend to be structurally disordered and more frequently undergo protease digestion (Fontana et al. 1986; Iakoucheva et al. 2001). However, this liberation of histidine residues seems rather indirect. The presence of numerous histidine residues within brain-specific proteins is strongly demonstrated by the data presented here but more experimental data is required to determine their function.

## Acknowledgements

This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to G.B.G. and an NSERC graduate scholarship to M.A.H.

## References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Banfi, S., Servadio, A., Chung, M.Y., Kwiatkowski, Jr., T.J., McCall, A.E., Duvick, L.A., Shen, Y., Roth, E.J., Orr, H.T., and Zoghbi, H.Y. 1994. Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat. Genet.* **7**: 513–520.
- Burke, J.R., Wingfield, M.S., Lewis, K.E., Roses, A.D., Lee, J.E., Hulette, C., Pericak-Vance, M.A., and Vance, J.M. 1994. The Haw River syndrome: dentatorubropallidolusian atrophy (DRPLA) in an African-American family. *Nat. Genet.* **7**: 521–524.
- Creighton, T.E. 1993. *Proteins: structures and molecular properties*. W.H. Freeman and Company, New York, N.Y.
- Cugini, Jr., C.D., Leidy, Jr., J.W., Chertow, B.S., Berard, J., Bradley, W.E., Menke, J.B., Hao, E.H., and Usala, S.J. 1992. An arginine to histidine mutation in codon 315 of the cerbA beta thyroid hormone receptor in a kindred with generalized resistance to thyroid hormones results in a receptor with significant 3,5,3'-triiodothyronine binding activity. *J. Clin. Endocrinol. Metab.* **74**: 1164–1170.
- David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E., et al. 1997. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nat. Genet.* **17**: 65–70.

- Duret, L., and Mouchiroud, D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**: 68–74.
- Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., et al. 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* **4**: 387–392.
- Fontana, A., Fassina, G., Vita, C., Dalzoppo, D., Zamai, M., and Zamboni, M. 1986. Correlation between sites of limited proteolysis and segmental mobility in thermolysin. *Biochemistry*, **25**: 1847–1851.
- Huntley, M., and Golding, G.B. 2000. Evolution of simple sequence in proteins. *J. Mol. Evol.* **51**: 131–140.
- Huntley, M.A., and Golding, G.B. 2004. Neurological proteins are not enriched for repetitive sequences. *Genetics*, **166**: 1141–1154.
- Hurst, L.D., and Smith, N.G. 1999. Do essential genes evolve slowly? *Curr. Biol.* **9**: 747–750.
- Iakoucheva, L.M., Kimzey, A.L., Masselon, C.D., Bruce, J.E., Garner, E.C., Brown, C.J., Dunker, A.K., Smith, R.D., and Ackerman, E.J. 2001. Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci.* **10**: 560–571.
- Karlin, S., and Burge, C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. U.S.A.* **93**: 1560–1565.
- Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., et al. 1994. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.* **8**: 221–228.
- Kery, V., Both, V., Sevcik, J., and Zelinka, J. 1986. The number and role of histidine residues in the active site of guanyloribonuclease Sa. *Gen. Physiol. Biophys.* **5**: 405–414.
- Kiebertz, K., MacDonald, M., Shih, C., Feigin, A., Steinberg, K., Bordwell, K., Zimmerman, C., Srinidhi, J., Sotack, J., Gusella, J., et al. 1994. Trinucleotide repeat length and progression of illness in Huntington's disease. *J. Med. Genet.* **31**: 872–874.
- Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., et al. 1994. Unstable expansion of CAG repeat in hereditary dentatorubral-pallidolusian atrophy (DRPLA). *Nat. Genet.* **6**: 9–13.
- Kuma, K., Iwabe, N., and Miyata, T. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Mol. Biol. Evol.* **12**: 123–130.
- La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., and Fischbeck, K.H. 1991. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature (London)*, **352**: 77–79.
- Li, S.H., McInnis, M.G., Margolis, R.L., Antonarakis, S.E., and Ross, C.A. 1993. Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms. *Genomics*, **16**: 572–579.
- Loomis, W.F., Shaulsky, G., and Wang, N. 1997. Histidine kinases in signal transduction pathways of eukaryotes. *J. Cell Sci.* **110**: 1141–1145.
- Marcotte, E.M., Pellegrini, M., Yeates, T.O., and Eisenberg, D. 1999. A census of protein repeats. *J. Mol. Biol.* **293**: 151–160.
- Myers, E.W., and Miller, W. 1988. Optimal alignments in linear-space. *Comput. Appl. Biosci.* **4**: 11–17.
- Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama, N., et al. 1994. Dentatorubral and pallidolusian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nat. Genet.* **6**: 14–18.
- Nakamura, K., Jeong, S.Y., Uchihara, T., Anno, M., Nagashima, K., Nagashima, T., Ikeda, S., Tsuji, S., and Kanazawa, I. 2001. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum. Mol. Genet.* **10**: 1441–1448.
- Pelton, P.D., and Ganzhorn, A.J. 1992. The effect of histidine modification on the activity of myo-inositol monophosphatase from bovine brain. *J. Biol. Chem.* **267**: 5916–5920.
- Pulst, S.M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunke, A., et al. 1996. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat. Genet.* **14**: 269–276.
- Rittig, S., Siggaard, C., Ozata, M., Yetkin, I., Gregersen, N., Pedersen, E.B., and Robertson, G.L. 2002. Autosomal dominant neurohypophyseal diabetes insipidus due to substitution of histidine for tyrosine(2) in the vasopressin moiety of the hormone precursor. *J. Clin. Endocrinol. Metab.* **87**: 3351–3355.
- Rogers, T.B., Gold, R.A., and Feeney, R.E. 1977. Ethoxyformylation and photooxidation of histidines in transferrins. *Biochemistry*, **16**: 2299–2305.
- Samuel, M., Samuel, E., and Villanueva, G.B. 1993. Histidine residues are essential for the surface binding and autoactivation of human coagulation factor XII. *Biochem. Biophys. Res. Commun.* **191**: 110–117.
- Silveira, I., Miranda, C., Guimaraes, L., Moreira, M.C., Alonso, I., Mendonca, P., Ferro, A., Pinto-Basto, J., Coelho, J., Ferreirinha, F., et al. 2002. Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)<sub>n</sub> allele at the SCA17 locus. *Arch. Neurol.* **59**: 623–629.
- Sim, K.L., and Creamer, T.P. 2002. Abundance and distributions of eukaryote protein simple sequences. *Mol. Cell. Proteomics*, **1**: 983–995.
- Sim, K.L., and Creamer, T.P. 2004. Protein simple sequence conservation. *Proteins*, **54**: 629–638.
- Snell, R.G., MacMillan, J.C., Cheadle, J.P., Fenton, I., Lazarou, L.P., Davies, P., MacDonald, M.E., Gusella, J.F., Harper, P.S., and Shaw, D.J. 1993. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat. Genet.* **4**: 393–397.
- Sumiyama, K., Washio-Watanabe, K., Saitou, N., Hayakawa, T., and Ueda, S. 1996. Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. *J. Mol. Evol.* **43**: 170–178.
- Torgerson, D.G., Kulathinal, R.J., and Singh, R.S. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol. Biol. Evol.* **19**: 1973–1980.
- Winter, E.E., Goodstadt, L., and Ponting, C.P. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14**: 54–61.
- Wootton, J.C., and Federhen, S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**: 149–163.
- Zhang, L., and Li, W.H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21**: 236–239.
- Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D.W., Amos, C., Dobyns, W.B., Subramony, S.H., Zoghbi, H.Y., and Lee, C.C. 1997. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat. Genet.* **15**: 62–69.