COMPARATIVE GENOMICS OF AMINO ACID TANDEM REPEATS

PhD Thesis

Barcelona, June 2008

Loris Mularoni

TO MERITXELL, MY WIFE, WHOSE PATIENT LOVE ENABLED ME TO COMPLETE THIS WORK.

For the most curious, the background of the cover represents the binary codification of one of my wife's favorite poem. The original version of this poem by Miquel Martí i Pol is included at the end of this thesis.

COMPARATIVE GENOMICS

OF

AMINO ACID TANDEM REPEATS

Memòria presentada per

Loris Mularoni

per optar al grau de

Doctor en Biologia per la Universitat Pompeu Fabra

Aquesta Tesi Doctoral ha estat realitzada sota la direcció de la

Dra. **M.Mar Albà**

al Departament de Ciències Experimentals i de la Salut de la

Universitat Pompeu Fabra

M.Mar Albà
La Directora de Tesi

Loris Mularoni
L'Autor

Barcelona, Juny de 2008

Tu ne quaesieris, scire nefas, quem mihi, quem tibi
finem di dederint, Leuconoe, nec Babylonios
temptaris numeros. ut melius, quidquid erit, pati.
seu pluris hiemes seu tribuit Iuppiter ultimam,
quae nunc oppositis debilitat pumicibus mare
Tyrrhenum: sapias, vina liques et spatio brevi
spem longam reseces. dum loquimur, fugerit invida
aetas: **carpe diem quam minimum credula postero**.

Ask not - we cannot know - what end the gods have set for you, for me; nor attempt the Babylonian reckonings Leuconoe. How much better to endure whatever comes, whether Jupiter grants us additional winters or whether this is our last, which now wears out the Tuscan sea upon the barrier of the cliffs. Be wise, strain the wine; and since life is brief, prune back far-reaching hopes! Even while we speak, envious time has passed: seize the day, putting as little trust as possible in tomorrow!

ODE I-XI, QUINTUS HORATIUS FLACCUS (Venosa 65 BC - Roma 8 BC)

Contents

List of Figures

List of Tables

List of Abbreviations

bp	Base pairs
cDNA	Complementary DNA
CH	Codon homogeneity
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
Ka	Rate of non-synonymous substitutions
Ks	Rate of synonymous substitutions
RNA	Ribonucleic acid

Abstract

TANDEM amino acid repeats, also known as homopolimeric tract or homopeptides, are very common features of eukaryotic genomes and are present in nearly one-fifth of human encoded proteins. These structures have attracted much interest in the early 1990s when a number of neurological diseases associated with repeat expansion mutations were discovered in humans. Despite their abundance in coding proteins, little is known about their functional consequences. Two scenarios have been proposed. In one, tandem amino acid repeat is considered a neutral structure generated by slippage event and eventually tolerated in protein as long as it does not disrupt the protein function. However, an increasing number of studies proposed that tandem amino acid repeats may be involved in important functional or structural roles. For instance, tandem amino acid repeats had been found to be especially abundant in transcription factors and developmental proteins, where they can potentially modulate protein-protein interaction, exert an effect on gene transcriptional activity, or act as spacer between different protein domains. In addition, several studies have linked changes in repeat size to modification in developmental processes. Despite the advancement made in the last decade, little is known about the selective forces that shape their evolution. The aim of this thesis has been to gain further insight onto the evolutionary dynamics of tandem amino acid repeats by studying the different types of mutations that occur in the amino acid component of the human proteome, by studying the relationship between variability and abundance of amino acid tandem with the evolutionary constraints operating on the proteins, and by studying their conservation and distribution across various vertebrate genomes in both coding and non-coding sequences. The integration of these approaches enabled us to outline an evolutionary model of these structures.

Part I

INTRODUCTION

Introduction

Summary

The first part of this introductory chapter supplies a general background about repetitive elements in the genome. The second part is dedicated to amino acid tandem repeats and their relationship to coding microsatellites. The third part introduces the bioinformatics approaches used in the research works presented. More specific details can be found in the introductions of the articles included in this thesis.

1. Genome repetitive elements

1.1 The repetitive DNA content of genomes

DIFFERENT organisms show a wide range of genome sizes. Comparison of many major model organisms shows a correlation between genome size and the increase of complexity of the organism in terms of number of genes. In bacteria and archaea it is proportional to the genome size, but in eukaryotes, the gene number grows much slower, resulting in a large fraction of non-coding DNA (Figure 1). For instance, there is more than a 300-fold difference between the genome sizes of yeast and mammals, but only a modest 4 to 5 fold increase in overall gene number. The vast majority of DNA, about 97% in human, does not give rise to any proteins. The function of all this DNA has been obscure from the beginning and it was named "selfish DNA" (Orgel and Crick, 1980). In fact recent studies of complete mammalian genome sequences suggest that only about 5-6% of the genome is under selection pressure (Gibbs et al., 2004). The rest of the genome comprises many repetitive sequences (Figure 2). In mammals, repetitive elements reach about 44% of their DNA sequences (Venter et al., 2001; Harris et al., 2007).

Repetitive sequences can be divided into interspersed repeats, whose individual repeat units are distributed around the genome in a apparently random fashion, and tandem repeats, whose repeat units are placed next to each other.

1.2. Interspersed repeats

These mobile elements were first described by Barbara McClintock while doing classical genetic experiments in corn during the 1940s and 1950s, for which she was awarded a Nobel Prize in 1983. She characterized genetic entities that could move into and back out of genes, changing the phenotype

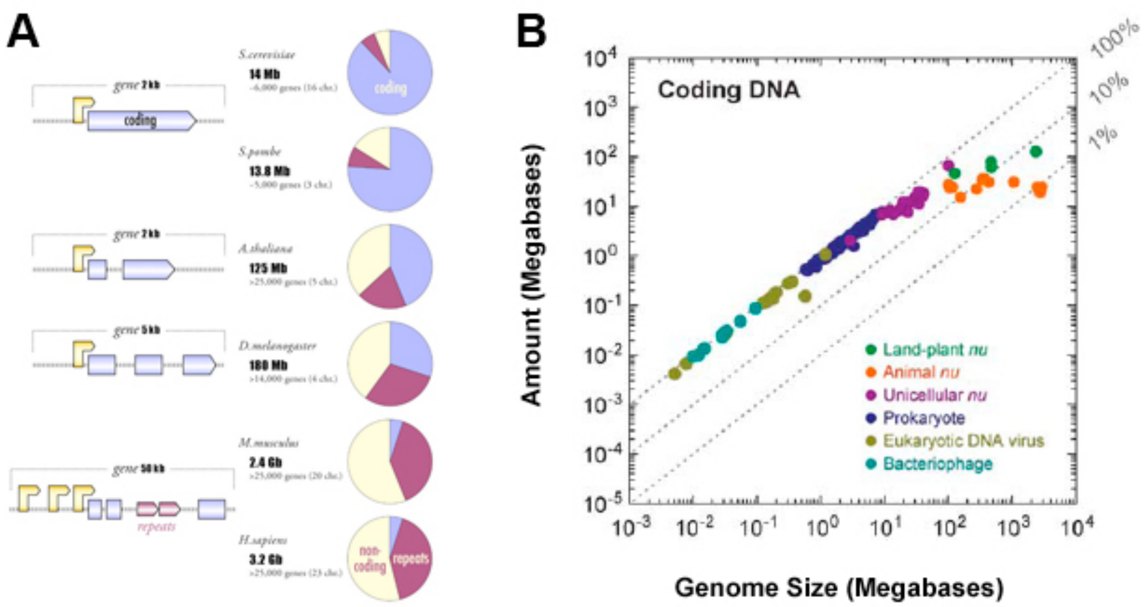


Figure 1. Genome sizes and amount of coding DNA. A) Genome sizes and corresponding composition of six major model organisms as pie charts. The increase in genome size correlates with the vast expansion of non-coding (i.e., intronic, intergenic, and interspersed repeat sequences) and repeat DNA (e.g., satellite, LINEs, Short interspersed nuclear element (SINEs), DNA (Alu sequence), in red) sequences in complex multicellular organisms. B) In general, the amount of coding DNA increases proportionally to the genome size, but in eukaryotes, the coding regions grow much slower than the genome size, resulting in a large fraction of non-coding DNA. Figure adapted from (Lynch, 2006).

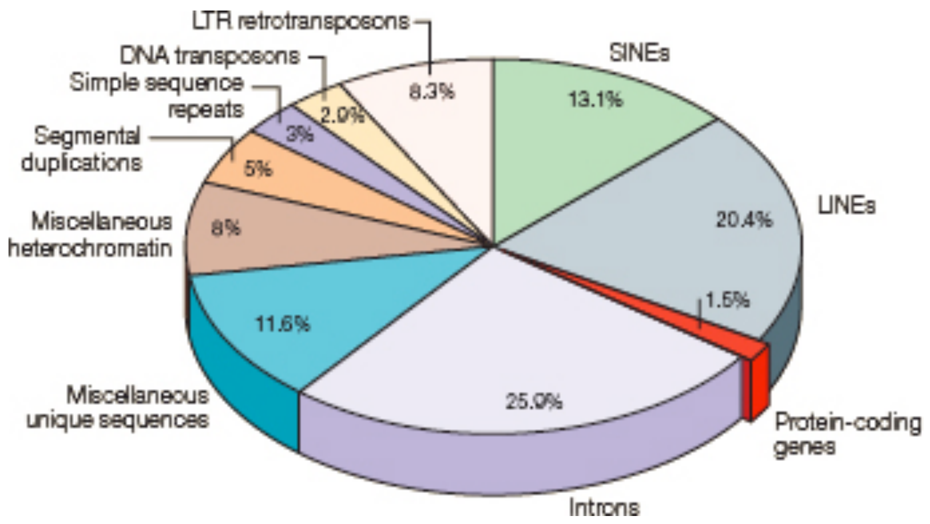


Figure 2. Components of the human genome. The figure provides a summary of the different components of the human genome. Less than 1.5% of the genome consists of protein–coding sequences. By contrast, a large majority is made up of non-coding sequences such as introns (almost 26%) and transposable elements (nearly 45%). Figure adapted from (Gregory, 2005).

of corn kernels. Interspersed repeats move around the genome by multiplying themselves via RNA or DNA intermediates. The most frequent way in which this occurs is by transposition, and most interspersed repeats have inherent transpositional activity. For this reason they are also known as transposable elements. Mobile elements that transpose through a DNA intermediate are generally referred to as transposons (or class I transposable elements), while those that transpose to new sites in the genome via an RNA intermediate are called retrotransposons (or class II transposable elements) because their movements are analogous to the infectious process of retroviruses (Figure 3). Both transposons and retrotransposons can be further classified based on their specific mechanism of transposition.

Retrotransposons can be subdivided into those elements that are autonomous, meaning that they encode their own replication machinery, and those that are nonautonomous. The former have long terminal repeats (LTR) at either ends, which play a role in the transposition process, and the latter do not have LTRs (non-LTR). In human about 10% of the genome is composed by LTR transposons (Lander et al., 2001). These mobile elements are very abundant in mammals and can be found in two different flavors, long interspersed element (LINEs) and short interspersed elements (SINEs).

Not all transposons require an RNA intermediate. Many are able to transpose in a more direct DNA-to-DNA manner, with a mechanism of cut-and-paste. DNA transposons move by two distinct transposition mechanisms, one involving direct interaction between the donor transposon and the target site, resulting in copying of the donor element (replicative transposition), and the second involving excision of the element and re-integration at a new site (conservative transposition)

1.3 Tandem repeats

Tandemly repeated DNA is a common feature of eukaryotic genomes. Tandem repeats can be divided in three groups. First come the major satellite

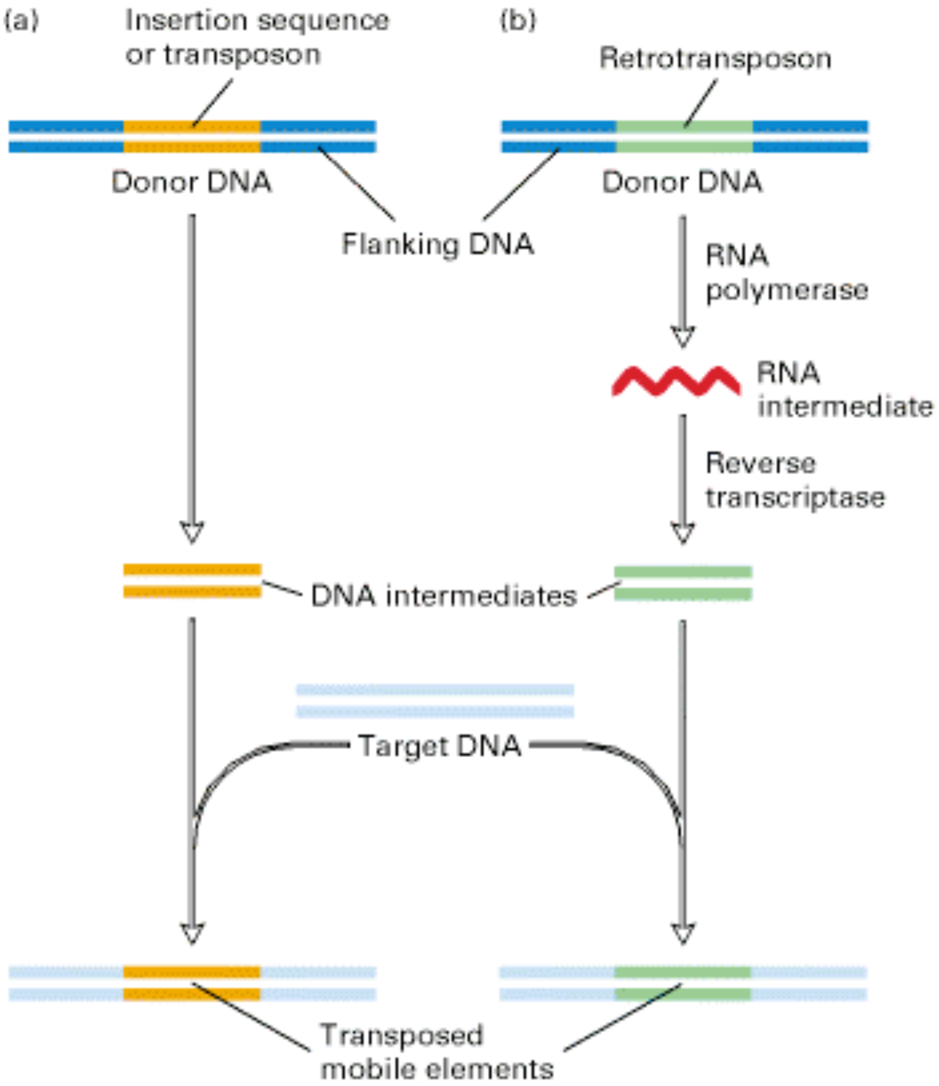


Figure 3. Interspersed repeats. Classification of mobile elements into two major classes. (a) Insertion sequences and transposons (orange) move via a DNA intermediate. (b) Retrotransposons (green) are first transcribed into an RNA molecule, which then is reverse-transcribed into double-stranded DNA. In both cases, the double-stranded DNA intermediate is integrated into the target-site DNA to complete movement. Figure taken from Molecular Cell Biology, chapter 9 (Lodish et al., 2000).

DNAs, which are long tracts, up to several Mb, of well-known families of repeat sequence elements. Next come minisatellites, which are blocks of repeated core elements of 10-100 nucleotides forming more or less uniform tracts of 10^2-10^5 long. The smallest class, generally known as microsatellites, includes short segments of 2-6 nucleotides repeated in more or less uniform tracts up to 10^2 nucleotides long (Tautz, 1993).

1.3.1 Satellites

Satellite DNAs were first discovered in the early 1960s as species of DNA, which due to their unusual base composition, band at densities distinct from bulk DNA upon equilibrium sedimentation (Kit, 1961). Satellite DNA consists of highly repetitive DNA, and stretch over almost all native centromeres and surrounding pericentromeric heterochromatin. Since no protein coding function could be primarily associated with satellite DNAs, early hypothesis considered them as useless genomic elements accumulated as junk (Ohno, 1972), or alternatively, as sequences that represent genomic parasites proliferating for their own sake (Orgel and Crick, 1980). An opposite view suggested the involvement of satellite DNAs in a series of functions ranging from chromosome organization and pairing to cell metabolism and speciation (John and Miklos, 1979). More recent studies supported these functionalist assumptions concerning the association of satellite DNAs with complex features of eukaryotic chromosomes (for instance, (Csink and Henikoff, 1998; Henikoff et al., 2001; Sullivan et al., 2001)). Although some satellites are scattered around the genome, most are located in the centromeric and pericentromeric heterocromatin, tow epigenetically determined regions responsible for correct pairing and disjunction of eukaryotic chromosomes in cell division (Arney and Fisher, 2004; Hall et al., 2004; Bloom, 2007). When present in centromeres, satellites may play a structural role, possibly as binding sites for one or more of the special centromeric proteins. Alternatively, the repetitive DNA content of the centromere might be a reflection of the fact that this is the last region of the chromosome to be replicated. In order to delay its replication until the very end of the cycle, the centromere DNA must lack sequences that can act as origins of replication. The repetitive nature of centromeric DNA may be a means of ensuring that such origins are absent (Csink and Henikoff, 1998).

1.3.2 Minisatellites

Minisatellites are tandemly repeated, highly variable DNA sequences found in many organisms. In fact they have been detected by DNA fingerprint analysis in most species including yeast (Andersen and Nilsson-Tillgren, 1997), fungi (Giraud et al., 1998), plants (Dallas, 1988) and higher eukaryotes (Jeffreys et al., 1985; Burke and Bruford, 1987; Jeffreys and Morton, 1987; Jeffreys et al., 1987a; Gilbert et al., 1990; Reeve et al., 1990; Buitkamp et al., 1991; Vergnaud et al., 1993). Minisatellites consist of repetitive, generally GC rich, variant repeats that range in length from 10 to over 100 base pairs. These variant repeats are tandemly intermingled, which makes minisatellites ideal for studying DNA turnover mechanisms. Although most minisatellites are GC rich, five AT rich loci have, however, been described in humans: the autosomal loci COL2A1 (Berg and Olaisen, 1993), ApoB (Desmarais et al., 1993; Ellsworth et al., 1995; Buresi et al., 1996), FRA16B (Yu et al., 1997) and FRA10B (Hewett et al., 1998) and the Y-chromosome-specific minisatellite MSY1 (Jobling et al., 1998).

Since the fortuitous discovery of the first human minisatellite (Wyman and White, 1980) and especially the discovery that the extreme polymorphism of minisatellites made them superb for DNA fingerprinting (Jeffreys et al., 1987b), these repeats have found several important applications, including as a powerful tool in forensic medicine, as markers for linkage studies in genetic analyses, and as a means for establishing kinship between individuals, including paternity determination. The minisatellites of the human genome are not evenly distributed but are primarily localized at the ends of the chromosomes, which implies a limitation in the use of these sequences in linkage analyses (Armour and Jeffreys, 1992).

It has long been known that some eukaryote genes contain minisatellites within their coding sequences. It has been suggested that some loci have evolved by concatemerisation of short tandem repeat sequences to create long open reading frames relatively resistant to point mutations which cause premature chain termination and reading frame shifts (Ohno, 1984; Ohno, 1987b; Ohno, 1987a; Haber and Louis, 1998). Coding minisatellites also have the capacity for rapidly generating new polymorphisms and therefore potential new functions. A striking example is the human apolipoprotein family. Each member of this family contains a coding minisatellite with a repeat unit of 33 or 66 base pairs (Boguski et al., 1986).

Apolipoproteins are capable of binding lipids, each being associated with a lipoprotein of specific density (Mahley et al., 1984). This specificity is due to the variable coding minisatellite region in each gene. Many other genes have been shown to contain minisatellites, such as the human epithelial mucin, involucrin, loricrin and small proline-rich (SPR) genes (Eckert and Green, 1986; Lancaster et al., 1990; Gibbs et al., 1993) and the D4 dopamine receptor (D4DR) (Van Tol et al., 1992; Lichter et al., 1993). The coding polymorphisms observed in the latter gene have been shown to affect the ligand binding affinity of D4DR (Van Tol et al., 1992; Asghari et al., 1994) and are associated with cognitive and emotional disorders, a consequence of the high level of D4DR expression in the limbic areas of the brain (Benjamin et al., 1996; Ebstein et al., 1996; Ebstein et al., 1997). In other occasions minisatellite instability can generate alleles that can disturb the expression of neighbouring genes. For instance the insulin (INS) minisatellite, located 600 base pairs upstream of the transcriptional start site of the INS gene is likely to be the insulin-dependent diabetes mellitus type 2 (IDDM2) locus (Bennett et al., 1995).

1.3.3 Microsatellites

Microsatellites are simple tandemly repeated DNA units (mono-, di-, trinucleotides, etc.), found in greater or less abundance in coding and non-coding Microsatellites are simple tandemly repeated DNA units (mono-, di-, trinucleotides, etc.), found in greater or less abundance in coding and non-coding regions of the genomes of just about every known organism and organelle. In particular these repetitive elements are well represented in the eukaryotic genome (Tautz and Renz, 1984; Tautz, 1989; Toth et al., 2000; Mrazek et al., 2007). They have come into prominence over the last decades because scientists have found them to be remarkably versatile molecular tools. Their applications range from estimation of the spatial relationships between chromosome segments to the elucidation of temporal relationships between origins of species and genera. The key to this versatility lies in the high levels of polymorphism (Bowcock et al., 1994), which are characteristically found at such loci. Unique DNA sequences in a genome exhibit a very low mutation rate (approximately 10^{-9} nucleotides per generation), whereas the mutation rates in microsatellite sequences are several orders of magnitude higher (Ellegren, 2000a), ranging from 10^{-6} to 10^{-2} nucleotides per generation. Indeed, if enough microsatellites are examinated then a unique genetic profile can be established for every person, with the only exceptions of genetically identical twins. The technology used for the molecular identification of individuals is generally known as 'DNA profiling' and perhaps the pre-eminent example of such work is the identification of the skeletons of the last Russian Tsar, Nicholas II, and other Romanov family members (Gill et al., 1994).

On the basis of different repeat units, microsatellites can be classified into different types. According to the length of the major repeat unit, they are classified into mono-, di-, tri, tetra-, penta- and hexanucleotide repeats. The total number of each type decreases as the size of the repeat unit increases. The most common microsatellites in the human genome are dinucleotides repeats (Lander et al., 2001).

Microsatellites are widely found in prokaryotes and eukaryotes. However, their distribution within chromosomes is not quite uniform, in fact they appear less frequently in sub-telomeric regions (Koreth et al., 1996). In humans, most microsatellites are found in the non-coding regions, while only about 8% locate in the coding regions (Ellegren, 2000a). Their densities also vary slightly among chromosomes. In humans, chromosome 19 has the highest density of repetitions (Subramanian et al., 2003). However on average, one microsatellite occurs per 2000 base pairs in the human genome (Lander et al., 2001). Despite microsatellites are more abundant in non-coding regions, many reports showed that a large number of repeats are located in transcribed region of genomes, including protein-coding genes and expressed sequence tags (ESTs) (Morgante et al., 2002). While in intergenic regions the most abundant microsatellites are di- or tetranucleotides repeats, in coding regions trinucleotide repeats predominate (Toth et al., 2000). Microsatellite occurrence in coding regions seems to be limited by non-perturbation of the reading frame. In many species, exons (unlike other genomic regions) rarely contain dinucleotide and tetranucleotide repeats, but have many more triplet and hexanucleotide microsatellites than other repeats (Field and Wills, 1996; Edwards et al., 1998; Metzgar et al., 2000; Wren et al., 2000; Young et al., 2000; Cordeiro et al., 2001; Morgante et al., 2002). Triplet repeats show approximately two fold greater frequency in exonic regions than in intronic and intergenic regions in all human chromosomes except the Y chromosome (Subramanian et al., 2003). This dominance of triplets over other repeats in coding regions may be explained on the basis of the suppression of nontrimeric repeats in coding regions, which would case frameshift mutations (Metzgar et al., 2000). In transcribed regions, UTRs harbor more microsatellites than the coding regions (Wren et al., 2000; Morgante et al., 2002). In particular 5'-UTRs in humans contains more monomers and trimers while 3'-UTR contains more monomer and dimmer motifs (Table 1). The distribution of microsatellites in introns is similar to that of genomic DNA; the majority of intronic repeats are monomers and/or dimmers depending on the species (Table 1).

Microsatellites were identified in eukaryotic DNA at the beginning of the 1970s. However, for long time, the mechanisms of their mutation have remained poorly understood. Up to now, two possible mechanisms have been proposed: unequal crossing (or gene conversion) over in meiosis and slippage-strand replication (Figures 4 and 5 respectively). Unequal crossing over is a well-known mechanism generating large blocks of satellite DNA. It is associated with the exchange of repeat units between homologous chromosomes. However, among these mutational mechanisms, the

Repeat	*H.sapiens*		Introns					
Unit (bp)	5-UTR	3-UTR	Primates	Rodents	Mammals	Vertebrata	Anthropoda	*C.elegans*
1	31,3	48,3	53,8	18,1	34,4	22,0	21,5	27,1
2	14,6	28,6	19,6	48,4	38,5	47,7	36,8	29,1
3	31,1	4,6	5,5	9,3	6,9	12,9	16,5	12,1
4	2,3	6,3	11,6	13,7	8,3	11,7	5,5	5,1
5	9,6	6,4	6,8	6,3	6,2	4,4	8,6	7,8
6	6,3	2,4	2,7	4,2	5,6	1,3	11,2	18,7

Table 1. Frequency (%) of microsatellites in non-translated regions of genes in eukaryotas. Adapted from (Toth et al., 2000; Wren et al., 2000).

predominant one is “slipped-strand mispairing” (Levinson and Gutman, 1987). When slipped-strand mispairing occurs within a microsatellite array during DNA synthesis, it can result in the gain or loss of one, or more, repeat units depending on whether the newly synthesized DNA chain loops out or the template chain loops out respectively. The relative propensity for either chain to loop out seems to depend in part on the sequences making up the array, and in part on whether the event occurs on the leading (continuous DNA synthesis) or lagging (discontinuous DNA synthesis) strand (Freudenreich et al., 1997). Experiments in vitro have demonstrated that DNA slippage occurs at very high rates (Schlotterer and Tautz, 1992). But in vivo, most of the DNA loops are recognized and eliminated by the mismatch repair system. It has been shown that a functional mismatch repair system reduces the slippage mutation rate between 100 and 1000 fold (Strand et al., 1993). However, the rate of mutation still remains several orders of magnitude higher than that of point mutations (Ellegren, 2000b). Several factors can affect the rate of slippage, among which the repeat unit is the most important. A negative correlation was suggested between the length of the repeat unit and the rate of slippage (Schlotterer and Tautz, 1992). The rate of slippage is highest in dinucleotide repeats and lowest in tetranucleotide repeats (Kruglyak et al., 1998). Besides the repeat unit, other factors such as the number, location and sequence of repeats are also likely to affect the rate and direction of slippage (Schlotterer, 1998). Recently some studies have suggested that equilibrium distributions of microsatellites length are a result of balance between slippage events and point mutation (Kruglyak et al., 1998). Replication slippage favors growth, whereas point mutations break down a long repeat array into two or more shorter ones.

Individual microsatellite arrays are often visualized as having a life cycle of sorts; they are born, they grow and ultimately they perish. These events may stretch over tens, or even hundreds of millions of years (Messier et al., 1996; Taylor et al., 1999) (Figure 6). They are two hypotheses, not mutually exclusive, to explain microsatellite genesis. Repeats could arise either spontaneously from or within unique sequences (Messier et al., 1996),

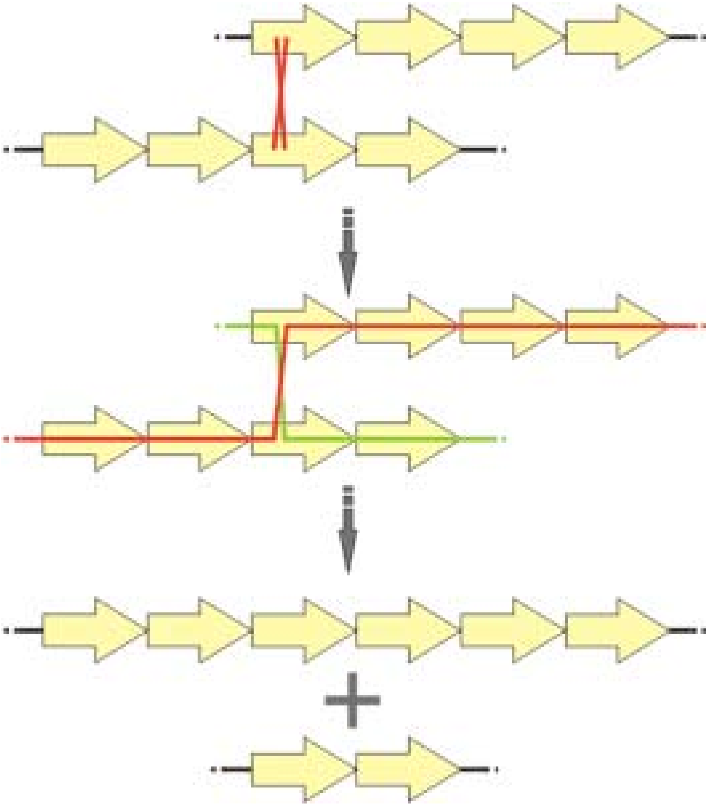


Figure 4. Unequal crossing-over. Tandem repeat polymorphisms can arise by unequal crossing-over. This process occurs during meiosis and it is associated with the exchange of repeat units between homologous chromosomes.

or could be brought about in a primal form into a receptive genomic location by mobile elements (Wilder and Hollocher, 2001). In the first case they are referred to as "de novo microsatellites" and in the second case they are referred to as "adopted microsatellites". De novo microsatellites are assumed to born from regions of "cryptic simplicity", i.e. regions in which variants of simple repetitive DNA sequence motifs are already over represented (Tautz et al., 1986). In the other hypothesis, microsatellites are adopted from other genomic regions via a number of transposable elements found in abundance in eukaryotes and thought to shape genome evolution (Kazazian, 2004). Once tandem duplications are generated, these short simple sequences may be prone to slippage (Rose and Falush, 1998). When slippage does occur, tandemly duplicated repeats will be added and will expand the array in length (Rose and Falush, 1998). Slippage generally results in a gain or loss of one repeat unit, depending on whether a loop is formed on the nascent or the

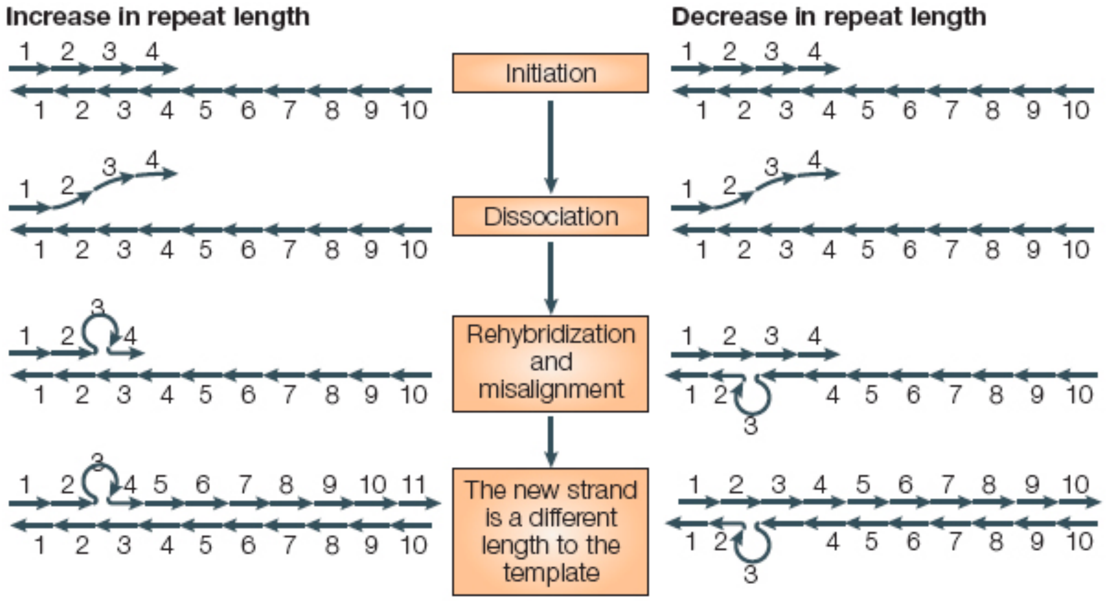

Increase in repeat length
Decrease in repeat length
Initiation
Dissociation
Rehybridization and misalignment
The new strand is a different length to the template
1 2 3 4
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11

Figure 5. Replication slippage. After the replication of a repeat tract has been initiated, the two strands might dissociate. If the nascent strand then realigns out of register, continued replication will lead to a different length from the template strand. If misalignment introduced a loop on the nascent strand, the end result would be an increase in repeat length. A loop that is formed in the template strand leads to a decrease in repeat length. Taken from (Ellegren, 2004).

template strand, respectively (Levinson and Gutman, 1987). Another mutational mechanism exists on these regions: unequal crossing over. This can lead to large-scale contractions and expansions in the repeat array (Richard and Paques, 2000) (Figure 4). However slippage is considered the major mechanism constantly affecting microsatellite length variability. Without an upper length constraint of some sort, expansion of microsatellites in eukaryotic genomes could be perpetual; instead, microsatellites most usually reach a pending state around a focal length. Expansion of long alleles seems to be restricted to a few tens of repeats although a few larger repeat arrays have been found, like some trinucleotide repeats in mammals (Pearson et al., 2005; Clark et al., 2006). Kruglyak et al. (1998) proposed that microsatellite growth reaches a finite upper limit since expansion by slippage is hindered by the introduction of imperfections in the repeat array. While expansions and contractions are in equilibrium and maintain the microsatellite at a focal length, interruptions can nevertheless still occur. Eventually the accumulation of interruptions breaks the repeat pattern and leads to a blend of unique DNA sequences that includes only short segments of the original repeat array. This event is known as “death” of the microsatellite (Taylor et al., 1999).

Since the last decade of the 20th century, scientists have been interested in the direct functions of microsatellites in some of their host organisms. Although microsatellites show high frequencies in genomes, most of them are thought to have no biological function and are regarded as “junk DNA". However, several interesting hypotheses suggest that microsatellites actually play an important role in many organisms. Certainly, some microsatellites have some function or influence on genomes (Li et al., 2002; Kashi and King, 2006). For instance at least one dinucleotide repeat has a role in regulating vasopressin gene expression altering mating behavior in a type of rodents (Hammock and Young, 2005). In addition one subfamily of repeats, composed of the basic hexamer TTAGGG, was found to have important capping functions in vertebrate telomeres (de Lange, 2005). Trinucleotide repeats have taken on special significance since genomic amplification of these tracts is the underlying genetic defect in a number of human diseases including neurodegenerative and neuromuscular diseases and mental retardation (Pearson et al., 2005) (Table 3). These trinucleotide expansion disorders are examples of repeat mutations that have detrimental effects, but there are also mutations at microsatellite loci that exert beneficial effects. In bacteria for example, microsatellite loci have a key role in the generation and maintenance of the high level of phenotypic diversity required for successful exploitation of variable environments. The human pathogen *Haemophilus influenzae*, which utilizes host hemoglobin and hemoglobin-haptoglobin as heme sources, has three hemoglobin binding protein genes (hgpA, hgpB and hgpC) that show variable expression patterns (Morton et al., 1999). All three genes contain lengths of tetrameric (CCAA)n repeats (Ren et al., 1999) and variation in the number of repeat units is a direct cause of phase-variation in expression (Morton et al., 1999; Ren et al., 1999).

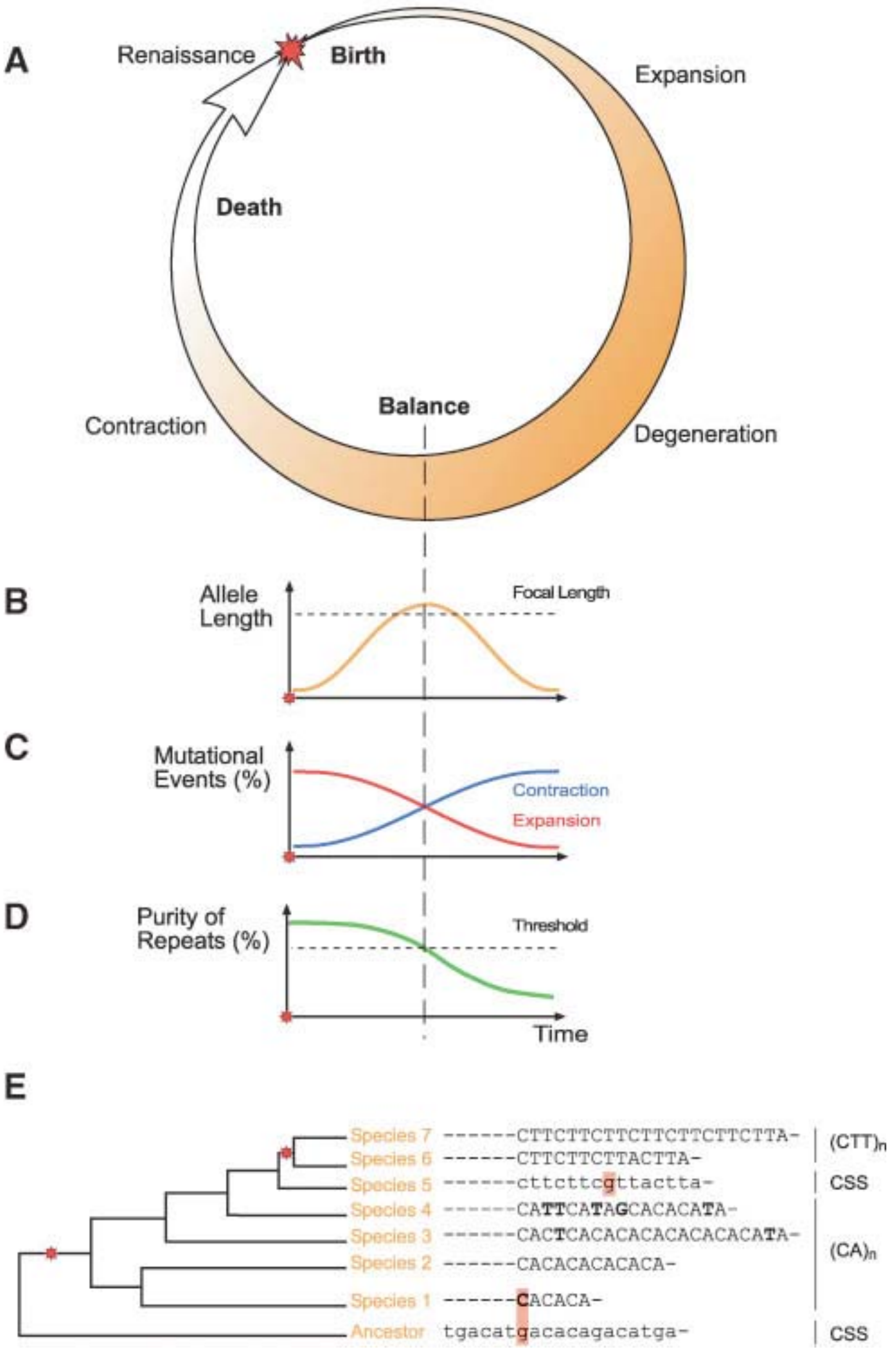
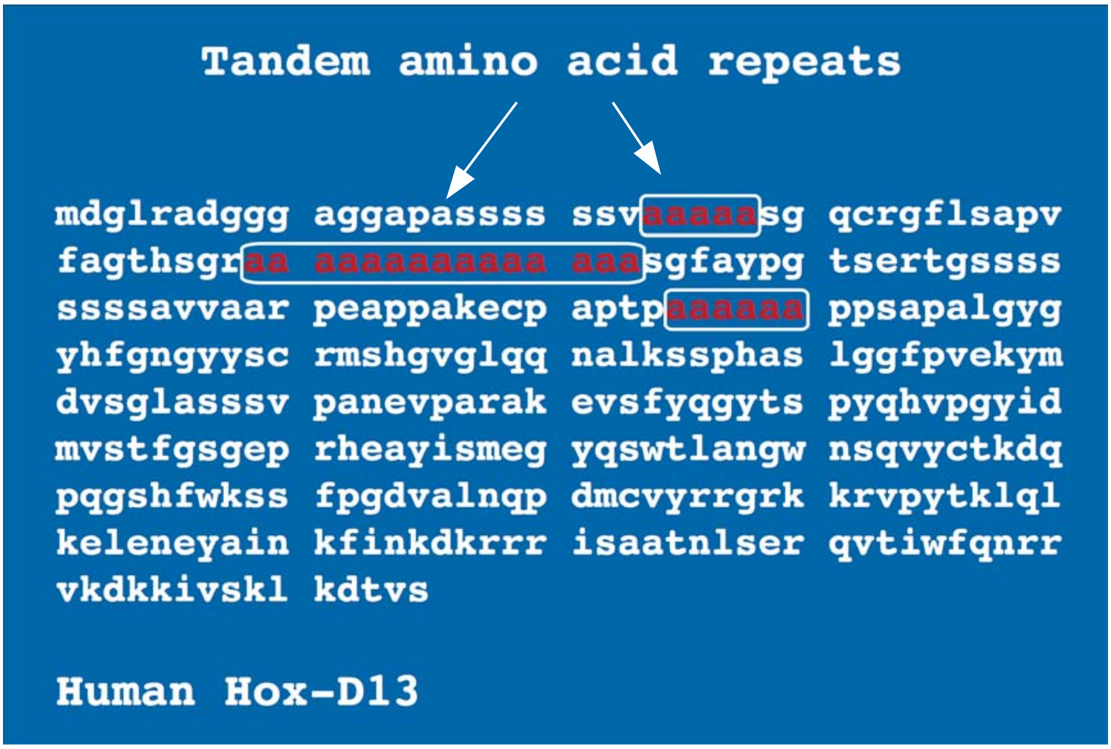

A
Renaissance
Birth
Expansion
Death
Contraction
Balance
Degeneration
B
Allele Length
Focal Length
C
Mutational Events (%)
Contraction
Expansion
D
Purity of Repeats (%)
Threshold
Time
E
Species 7 ------CTTCTTCTTCTTCTTCTTCTTA-
Species 6 ------CTTCTTCTTACTTA-
Species 5 ------cttcttcgttactta-
Species 4 ------CATTCATAGCACACATA-
Species 3 ------CACTCACACACACACACACATA-
Species 2 ------CACACACACACA-
Species 1 ------CACACA-
Ancestor tgacatgacacagacatga-
(CTT)n
CSS
(CA)n
CSS

Figure 6. **Hypothesised biology of a microsatellite locus**. (A) Schematic life cycle and (B–D) variations over time in: (B) average allele length, (C) relative occurrence of repeat expansion and contraction events and (D) purity of repeat (proportion of perfect repeats in the array). (E) Data superimposed on a phylogeny allow direct observation of a locus life cycle (CSS: Cryptically Simple Sequence). After initiation, a microsatellite expands and, with increasing length, interruptions by point mutations occur and affect the expansion mutation rate. The repeat array reaches an upper length limit where expansion and contraction mutations are in balance. While long deletions then occur predominantly and reduce the microsatellite in size, the continuing accumulation of imperfections breaks the array and decreases the rate of slippage. Both events lead to the fading of the microsatellite, but another locus may emerge from the remaining scramble of unique sequences. Time scale, upper allele length limit, dynamics of expansion versus contraction mutation events, and purity threshold are quantitatively unknown and are probably variable among loci. Taken from (Buschiazzo and Gemmell, 2006)

2. Tandem amino acid repeats

Tandem amino acid repeats are the subject of study of this thesis (Figure 7). These repetitive elements are also known as homopolymeric tracts or homopeptides. They are often embedded in low-complexity regions, or simple sequence, which include interrupted, non-tandem, repeats, and they are closely related with trinucleotide repeats, which represent a subset: while microsatellites are formed by the repetition in tandem of the same codon, tandem amino acid repeats could be formed by a mixture of synonymous codons. In fact the four nucleotides (adenine, guanine, thymine and cytosine) give rise to 61 out of 64 possible triplets, but only 20 amino acids are encoded (Figure 8). This phenomenon, named degeneracy of the genetic code, permits nucleotide sequences to vary without altering amino acid sequence. Indeed, in tandem amino acid repeats what results tandemly repeated is the amino acid encoded by the DNA sequence forming the repeats. For instance “CAG CAG CAG CAG CAG” and “CAA CAG CAG CAA CAG” both encode for the 5 residues of poly-Q, but only the former is composed by pure codon runs and represent a trinucleotide repeat. The codon composition of tandem amino acid repeats affects the dynamics of homopeptides, as will be further described later on.



Tandem amino acid repeats

```
mdglradggg aggapassss ssvaaaaasg qcrgflsapv
fagthsgraa aaaaaaaaaa aaasgfaypg tsertgssss
ssssavvaar peappakecp aptpaaaaaa ppsapalgyg
yhfgngyysc rmshgvglqq nalkssphas lggfpvekym
dvsglasssv panevparak evsfyqgyts pyqhvpgyid
mvstfgsgep rheayismeg yqswtlangw nsqvyctkdq
pqgshfwkss fpgdvalnqp dmcvyrrgrk krvpytklql
keleneyain kfinkdkrrr isaatnlser qvtiwfqnrr
vkdkkivskl kdtvs
```

Human Hox-D13

Figure 7. Tandem amino acid repeats. Example of alanine tandem amino acid repeats in the human protein Hox-D13. The 15 amino acid long polyA had been found to regulate the sesamoid bones structure in mouse (Anan et al., 2007). In humans, the expansion of this tract up to 22-29 residues is associated with the developmental disease Sympolydactyly type II (Muragaki et al., 1996).

2.1. Distribution of amino acid tandem repeats across genomes

Tandem amino acid repeats are very common feature of eukaryotic proteins (Green and Wang, 1994; Li et al., 2004; Hancock and Simon, 2005). For instance, in humans about one fifth of the encoded proteins contain amino acid tandem repeats of size 5 or longer (Karlin et al., 2002; Alba and Guigo, 2004). Instead, prokaryotes show a relatively low frequency of homopeptides (Karlin and Burge, 1996; Faux et al., 2005), about one order of magnitude lower than eukaryotes (Table 2). Among eukaryotes, the amount of proteins containing tandem amino acid repeats varies significantly. Homopeptides formed by 5 residues are present in about 13% of the proteins of *Caenorhabditis elegans*, 15% of the proteins of *Saccharomyces cerevisiae*,

	AGA								
	AGG								
GCA	CGA						GGA		
GCC	CGC						GGC		AUA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU
Ala	**Arg**	**Asp**	**Asn**	**Cys**	**Glu**	**Gln**	**Gly**	**His**	**Ile**
A	**R**	**D**	**N**	**C**	**E**	**Q**	**G**	**H**	**I**

UUA					AGC					
UUG					AUG					
CUA				CCA	UCA	ACA			GUA	
CUC				CCC	UCC	ACC			GUC	UAA
CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Leu	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**	**stop**
L	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**	

Figure 8. The genetic code. The genetic code is degenerated because some amino acids could be encoded by more than one possible codon, For instance serine, leucine, and arginine could be encoded by six different codons. Nucleotides are adenine (A), guanine (G), cytosine (C) and uracil (U). The complete name of amino acids are: Alanine (A),,Arginine (R), Asparagine (N), Aspartic acid (D), Cysteine (C), Glutamic acid (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y), Valine (V).

20% of the proteins of *Homo sapiens* and *Arabidopsis thaliana*, and 27% of the proteins of *Drosophila melanogaster* (Karlin et al., 2002). Moreover, among vertebrates, it had been show that homopeptides are more common in mammals (Faux et al., 2005). While in intergenic regions the most abundant microsatellites were represented by di- or tetranucleotides repeats, in coding regions tandem amino acid repeats predominate (Toth et al., 2000), probably as a consequence of the fact that they do not disrupt the open reading frame of the proteins.

Age groups				Species	No. of proteins	No of tandem AA repeats	Most abundant AA
Bacteria				*E.coli*	4.290	82(0.019)	L(29), A(21), G(8)
				B.subtilis	4.093	84 (0.020)	L(19), S(18), A(15)
Archaea				*A.fulgidus*	2.409	50 (0.020)	E(11), V(9), L/P(6)
				M.jannaaschii	1.773	27 (0.015)	E(7), K(6), L(3)
Eukarya				*S.cerevisiae*	6.680	1.246 (0.186)	S(298), Q(231), N(139)
				A.thaliana	25.761	6.694 (0.259)	S(1.746), G(832), E(798)
	Metazoa			*C.elegans*	26.032	5.569 (0.214)	S(804), Q(753), T(729)
				D.melanogaster	19.369	13.775 (0.711)	Q(4.299), A(2.094), S(1.606)
		Chordata		*C.intestinalis*	20.000	2.209 (0.110)	E(283), S(263), P(249)
				F.rubripes	22.102	3.200 (0.145)	S(674), P(518), E(364)
				G.gallus	24.268	2.986 (0.123)	S(588), E(519), P(434)
			Mammalia	*M.domestica*	26.943	4.143 (0.153)	P(747), S(661), E(609)
				M.musculus	36.471	7.601 (0.208)	P(1.247), E(1.235), S(982)
				H.sapiens	33.869	8.259 (0.244)	E(1.420), P(1.325), A(1.130)

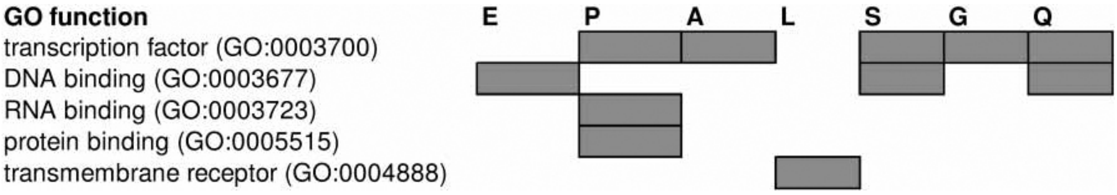
Table 2. Distribution of amino acid tandem repeats of size >= 5 in different species. Adapted from (Alba et al., 2007).

2.2. Relationship between GC content and repeat content

An interesting aspect is the relationship between GC content and repeat content. A number of studies reported a positive correlation between the richness of homopeptides and the GC content in mammalian genes (Sumiyama et al., 1996; Nakachi et al., 1997), indicating a predisposition to repeat generation in GC-rich contexts. These studies were focused on polyA, polyG and polyP, which are encoded by CG-rich codons. In a study of amino acid repeat size differences between mouse and rat it was also found that increased repeat length in one of the species was related to increased GC content in the rest of the gene, supporting an interrelation between the two processes (Alba and Guigo, 2004).

2.3. Coding repeat and functionality

Tandem amino acid repeats, which are usually formed by hydrophilic amino acids (such as serine, glutamine, glycine, alanine, leucine, etc.) (Green and Wang, 1994; Karlin and Burge, 1996; Mar Alba et al., 1999), can suffer repeat expansions or contractions as a consequence of trinucleotide slippage (Levinson and Gutman, 1987). This process is expected to produce an excess of pure codon repeat tracts in repeat coding sequences (Mar Alba et al., 1999; Alba et al., 2001). The most common amino acid repeats types vary between different organisms (Table 2). In human, which is the main species studies in this thesis, there is a predominance of alanine (A), glutamic acid (E), glycine (G), lysine (K), proline (P), glutamine (Q) and serine (S) (Mularoni et al., 2006; Mularoni et al., 2007; Mularoni et al., 2008). Despite the abundance of amino acid repeats in eukaryotic proteins, little is known about their functional consequences. This suggests that many tandem amino acid repeats could be just neutral structure, generated by slippage and then simply tolerated in protein as long as they do not disrupt the protein function. However recent findings point to the possibility that repeats may be involved in important functions or they may play structural roles. For instance, tandem repeats tend to be overrepresented in some classes of proteins, such as transcription factors and developmental proteins, (Nakachi et al., 1997; Richard and Dujon, 1997; Karlin et al., 2002; Cocquet et al., 2003; Alba and Guigo, 2004; Faux et al., 2005) (Figure 9). On the other hand, homopeptides are under-represented in yeast proteins involved in metabolic function (Young et al., 2000). A growing number of experiments show that some homopeptides, such as



GO function	E	P	A	L	S	G	Q
transcription factor (GO:0003700)		▨	▨		▨	▨	▨
DNA binding (GO:0003677)	▨				▨		▨
RNA binding (GO:0003723)		▨					
protein binding (GO:0005515)		▨					
transmembrane receptor (GO:0004888)				▨			

Figure 9. Repeats and functions. Gene ontology (GO) functions overrepresented in human proteins containing different amino acid repeat types. Taken from (Alba and Guigo, 2004).

polyA and polyQ can modulate interaction between proteins (Shimohata et al., 2000; Buchanan et al., 2004), exert an effect on gene transcriptional activity (Gerber et al., 1994; Lanz et al., 1995; Dunah et al., 2002; Galant and Carroll, 2002; Brown et al., 2005) or act as spacer between different protein domains (Newfeld et al., 1994). Finally some functions tend to be related to a particular amino acid repeat type: in mammals polyA had been found to be over-presented in RNA- and DNA binding proteins while polyL is characteristic of transmembrane receptors (Alba and Guigo, 2004). Tandem amino acid repeats, due to their high variability, have also been proposed to be an important source of variability, and may sometimes confer an adaptive evolutionary advantage (Kemp et al., 1987; Pizzi and Frontali, 2001; Kashi and King, 2006). A different type of evidence has recently been provided by Fondon and Garner (2004). The authors found that the length ratio of two adjacent repeats in the runt-related transcription factor *Runx-2*, encoding 18-20 glutamines followed by 12-17 alanines, was correlated with measures of facial shape across breeds, in particular with the midface length and clinorhynchy (dorsoventral nose bend) (Figure 10). Interestingly, the homologous gene in human (*CBFA1*), which encodes osteoblast-specific transcription factor *OSF2*, is known to influence craniofacial structure, and an expansion of the alanine stretch from 17 to 27 has been found in members of one family who have cleidocranial dysplasia (Mundlos et al., 1997). Another interesting example that demonstrates how repeats could be involved in functionality is given by Anan et al. (2007). The authors showed how a deletion of a 15 residues long polyA in the protein *Hox-D13* in transgenic mice resulted in a phenotypic effect on sesamoid bones: in the knock-out mice, lacking the poyA, the sesamoid bones were fused while in the wild-type mice the sesamoid bones were separated (Figure 11).

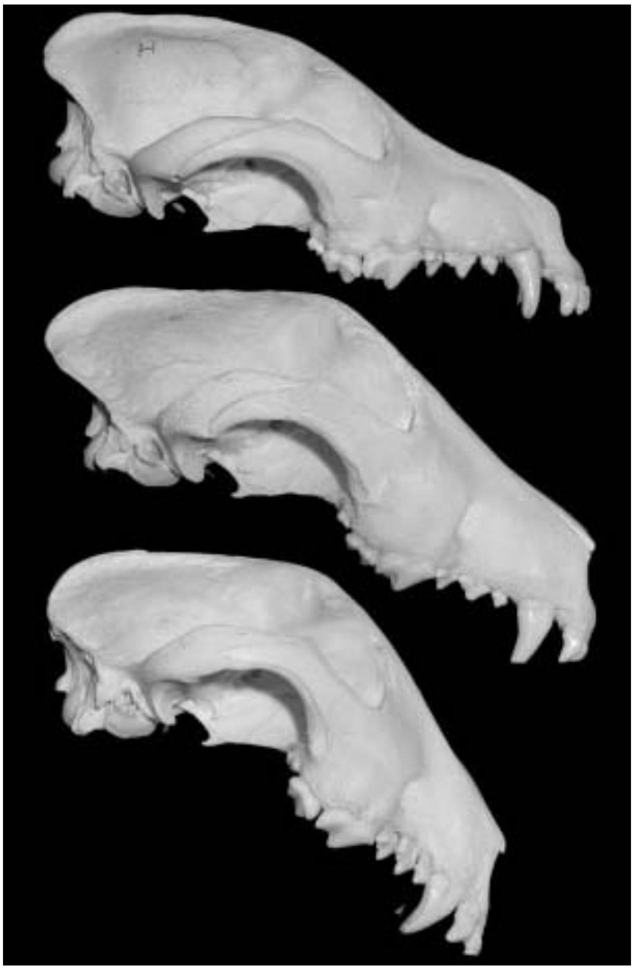


Figure 10. *Runx-2* and morphological evolution. Rapid and sustained evolution of dog breeds. Purebred bull terrier skulls from 1931 (*Top*), 1950 (*Middle*), and 1976 (*Bottom*) (24). Despite the lack of genetic diversity caused by population structure and history, these breeds are able to continually create new and more extreme morphological variations at a rapid and sustained pace. Analysis of the *Runx-2* repeats in the 1931 bull terrier reveals a more intermediate allele (Q19A14) than is present in the modern bull terrier (Q19A13). Taken from (Fondon and Garner, 2004).

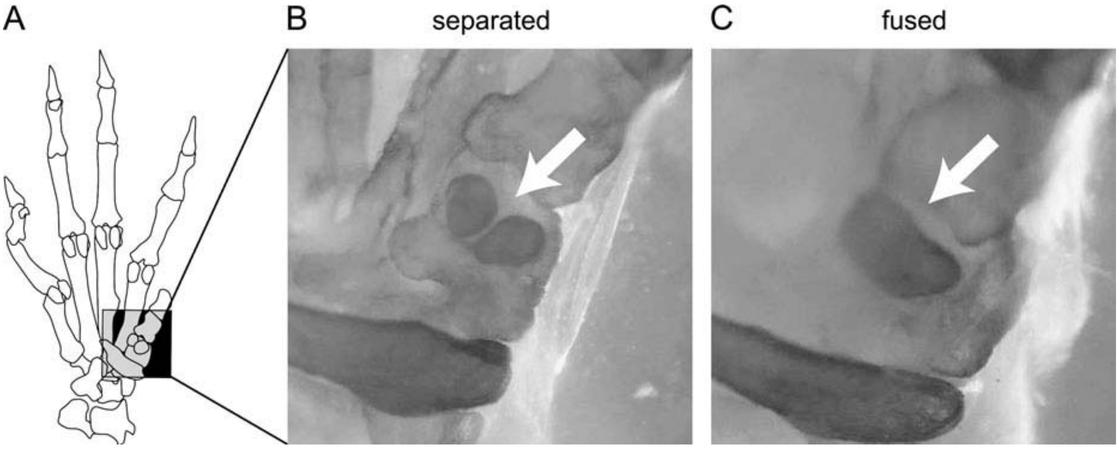


Figure 11. Hox proteins and development. Morphological variation in the sesamoid bones of the forelimb. (A) Location of the sesamoid bones in a forelimb. Separated (B) and fused (C) types of the sesamoid bones are indicated by arrows. Taken from (Anan et al., 2007).

2.4. Diseases associated repeat

Some repeat expansions may exert beneficial effects, but there are also expansions that have detrimental effects. One example is given by trinucleotide repeat associated diseases, that are characterized by trinucleotide repeats that expand far outside of their “normal” polymorphic range. To date, trinucleotide repeat associated diseases have only been identified in humans, although it was found that transgenic and knockin mouse models, that expressed the full length protein with the expanded repeats, recapitulated at least one human disease (Watase et al., 2002; Yoo et al., 2003). The first pathogenic mutations in humans were described in 1991, when abnormal expansion of arginine and glutamine were found to cause fragile X syndrome (FRAXA) and spinal and bulbar muscular atrophy (SBMA) respectively. From then, the list of disorders caused by trinucleotide expansion has been constantly growing and so far in humans 30 hereditary disorders have been found to be caused by tandem amino acid repeats instability (Table 3), Repeat instability has been associated with two main types of disorders: developmental disease, which are caused by uncontrolled expansion of polyA, and neurodegenerative disorders, which arise from abnormal expansion of polyQ. Pathogenic glutamine repeats are high polymorphic within human populations (Andres et al., 2003). Generally the severity of the disease is correlated with the extent of expansion. The polyglutamine diseases constitute a class of nine gain-of-function disorders that result from expansion of the codon CAG. The fact that a pure codon repeat tract encodes the polyQ suggests that slippage is the main contributor of the increased length of the repeat (Sinden et al., 2002). After a size threshold is overcome, further expansion become progressively more likely, leading to very long repeats in just a few generations. A common feature of polyglutamine diseases is the accumulation of mutant protein to form insoluble aggregates. The expanded polyQ tract alters protein conformation and recruits cellular defense mechanisms against protein misfolding and aggregation (Ciechanover and Brundin, 2003). Polyalanine coding repeats behave differently (Brown and Brown, 2004). In the first place, unlikely glutamine tracts, they are encoded by the imperfect codon $(GCN)_{n.}$ Additionaly, while polyQ tend to be polymorphic, alanine tracts are not. Related to this, expanded polyQ tracts tend to be meiotically and somatically unstable whereas expanded polyA tract are generally stable. Another contrast between glutamine and alanine tracts is that polyQ is thought to expand via slippage replication while unequal crossing-over (Figure 5 and 6, respectively) has been proposed as a mechanism that could explain most polyA expansion mutations (Warren, 1997). Again, pathologically expanded polyQ tracts can be very long while the threshold length for polyA expansion is extremely low (30-60 base), and they rarely expand more than 1,5-fold. Finally, in polyQ associated pathogenesis, the disease is triggered by a toxic accumulation of protein complexes, rendered insoluble by the glutamine expansion, that cause neuronal cell death late in life. On the contrary, in polyA associated diseases,a mutation, whether causing gain of a dominant negative activity or a loss-of-function, results in altered expression of downstream target genes, and this, in turn, results in abnormal development. So far, nine genes have been described to be subject to pathological polyA expansions (Table 3). Of these, eight are transcription factors that are expressed during development, and mutation in them result in abnormalities of the body plan. Aside from polyA and polyQ associated diseases, other disorders arise from expansion of different types of repeat units: Multiple Skeletal Dysplasias (COMP), myotonic dystrophy type 1 (DM1), Friedreich's ataxia (FRDA), Spinocerebellar ataxia 8 (SCA8), Spinocerebellar ataxia 12 (SCA12), Huntington's disease-like 2 (HDL2), fragile X syndromes (FRAXA and FRAXE), and Jacobsen syndrome (FRA11B). The pathogenic cause of the diseases listed above is not always by the mechanisms previously described. In summary, repeat associated diseases can be the result of expansion of non-coding repeats, which result in loss of protein function or altered RNA function; or expansion of coding repeats which result in altered protein function.

Disorder (nameHGNC Symbol)	Repeat unit	Repeat length	
		Normal	Pathogenic
Polyglutamine (CAG) tracts			
SBMA: Spinal and Bulbar Muscular Atrophy (AR)	(CAG)n•(CTG)n	9-36	40-55
HD: Huntington's disease (HD)	(CAG)n•(CTG)n	10-34	>35
DRPLA: Dentatorubral-pallidouysian atrophy (ATN1)	(CAG)n•(CTG)n	7-25	49-88
SCA1: Spinocerebellar ataxia 1 (ATXN1)	(CAG)n•(CTG)n	6-39	39-81
SCA2 Spinocerebellar ataxia 2 (ATXN2)	(CAG)n•(CTG)n	13-33	>34
SCA3: Spinocerebellar ataxia 3 (ATXN3)	(CAG)n•(CTG)n	13-44	>55
SCA6: Spinocerebellar ataxia 6 (CACNA1A)	(CAG)n•(CTG)n	4-18	20-29
SCA7: Spinocerebellar ataxia 7 (ATXN 7)	(CAG)n•(CTG)n	4 - 35	37-306
SCA17: Spinocerebellar ataxia 17 (TBP)	(CAG)n•(CTG)n	25-42	47-63
Polyalanine (CAN) tracts			
HOXD13: Synpolydactyly type II (HOXD13)	(GCG)n•(CGC)n	15	22 – 29
OPMD: Oculopharyngeal Muscular Dystrophy (PABPN1)	(GCG)n•(CGC)n	10	11-17
CBFA1: Cleidocranial dysplasia (RUNX2)	(GCG)n•(CGC)n	17	27
HPE5: Holoprosencephaly (ZIC2)	(GCG)n•(CGC)n	15	25
HOXA13: Hand-Foot-Genital Syndrome (HOXA13)	(GCG)n•(CGC)n	18	24 or 26
FOXL2: Blepharophimosis/Ptosis/Epicanthus inversus syndrome type II (FOXL2)	(GCG)n•(CGC)n	14	22–24
ARX: Infantile spasm syndrome (ARX)	(GCG)n•(CGC)n	16 12	18 or 23 20
SOX3: Mental retardation; X linked, with isolated growth hormone deficiency (SOX3)	(GCN)n•(NGC)n	15	26
CCHS: Congenital central hypoventilation syndrome (PHOX2B)	(GCN)n•(NGC)n	20	25-29

Table 3. Features of glutamine and alanine unstable repeat expansion disorders. Taken from (Brown and Brown, 2004; Gatchel and Zoghbi, 2005).

2.5. Mutational mechanisms of tandem amino acid repeats

The average mutation rate of DNA sequences is very low, and affects approximately 10^{-9} nucleotides each generation. However mutations seem to be much more common in tandem amino acid repeats (Ellegren, 2000a). This rate is even four to seven times increased if disease-associated trinucleotide repeats are considered. The increased mutational rate in tandem amino acid repeats could be explained with various mechanisms related to repeat evolution. Up to date, two main possible mechanisms have been proposed: replication slippage and unequal crossing-over.

2.5.1 Replication slippage

At present, replication slippage has been widely accepted as the main explanation of tandem amino acid repeats mutation processes. occurs during DNA replication, as a consequence of mispairing of one of more repeat units, between the nascent and template strand. When the nascent strand realigns out of register, renewed replication will lead to the insertion or deletion of repeat units relative to the template strand (Figure 5). Replication slippage occurs at very high rates but most of these primary mutations are corrected by the mismatch repair system, and only the small fraction that was not repaired ends up as microsatellite mutation events (Strand et al., 1993). Several factors could affect the rate of slippage events in tandem amino acid repeats, such as the number, location and sequence of repeats (Schlotterer, 1998). In addition, the internal architecture of the homopeptide influence the rate of slippage: repeats composed by pure codon tract are more likely to suffer slippage mutation respect to repeats composed by a mixture of synonymous codons. In this scenario, point mutation, breaking the homogeneity of the codons tracts by inserting synonymous substitutions, plays an important role in stabilizing the repeat. It has therefore been proposed that the length distribution of tandem amino acid repeats can be modeled as a balance between these two evolutionary forces: replication slippage and point mutation (Kruglyak et al., 1998).

2.5.2 Unequal crossing-over

Tandem amino acid repeats can also originate from unequal crossing-over in meiosis. This event is associated with the exchange of repeat units between homologous chromosomes in the process of DNA replication. However, this mechanism involves different chromosomes, and it probably plays a restricted role in tandem amino acid repeat mutation. Nevertheless, there evidence that some homopeptides, such as the 15 residues long polyA in HOXD13, suffer this type of mutation (Warren, 1997).

3. Bioinformatics approaches for the study of repeats

In the research works presented in this thesis various approaches has been used to study the features of tandem amino acid repeats. Expressed Sequence Tags (ESTs) were useful to investigate levels of variability of tandem amino acid repeats in human (see Chapter 1). For inter-specific repeat comparisons we used collections of orthologous genes (see Chapters 2 and 3). Repeats with a common origin were identified using protein sequence alignments. We made extensive use of sequence and genome databases, such as Genbank and Ensembl. The analysis of the data has implied the development of several computer programs, written in the Python programming language.

3.1 Expressed Sequence Tags (ESTs)

Expressed Sequence Tags, or simply ESTs, are small pieces of nucleotide sequence that are generated by sequencing either one or both ends (5'-EST and 3'-EST) of an expressed gene. The goal is to sequence bits of DNA that represent genes expressed in certain cells, tissue, or organs from different organisms. ESTs are produced by one-shot sequencing of cloned mRNAs, a method that allowed researches to sequence portion of vast numbers of complementary DNA (cDNA), which represent an expressed gene, isolated from various model organisms (Figure 12). These partial cDNA sequences are a relatively low quality fragments whose length is actually limited to approximately 500 to 800 nucleotides. Because these clones consist of DNA that is complementary to mRNA, the ESTs represent relatively short portions (tags) of expressed genes. They may be present in the database as either cDNA/mRNA sequence or as the reverse complement of the mRNA.

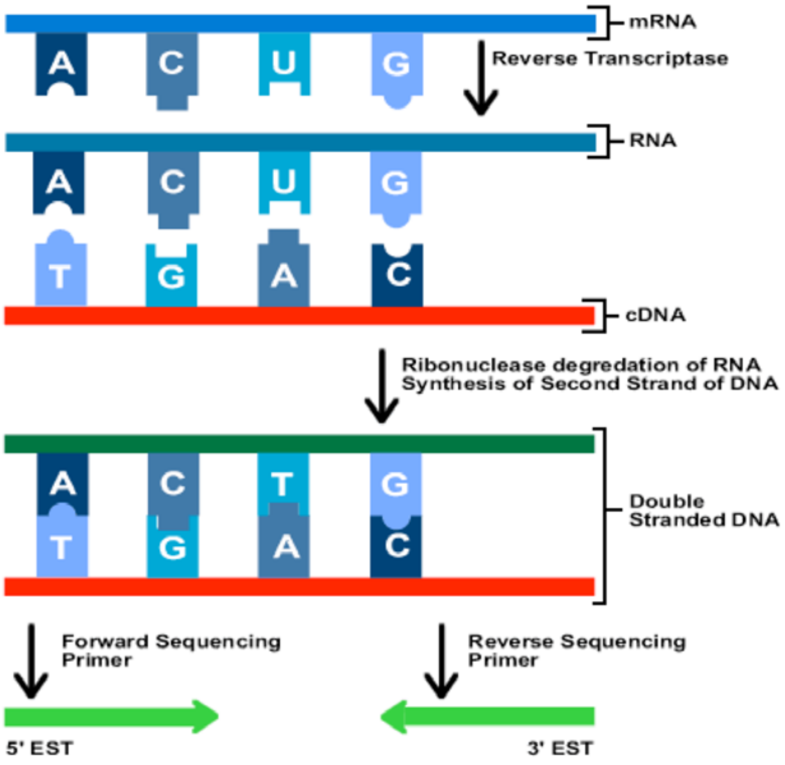

mRNA
A
C
U
G
Reverse Transcriptase
RNA
A
C
U
G
T
G
A
C
cDNA
Ribonuclease degredation of RNA
Synthesis of Second Strand of DNA
A
C
T
G
T
G
A
C
Double Stranded DNA
Forward Sequencing Primer
Reverse Sequencing Primer
5' EST
3' EST

Figure 12. **An overview of how ESTs are generated**. ESTs are generated by sequencing cDNA, which itself is synthesized from the mRNA molecules in a cell. The mRNAs in a cell are copies of the genes that are being expressed. mRNA does not contain sequences from the regions between genes, nor from the non-coding introns that are present within many interesting parts of the genome.

3.2 Orthologous genes

Two genes are said orthologous, or ortologs, if they diverged after a speciation event while are defined paralogous, or paralogs, if they diverged after a duplication event (Figure 13). Both orthology and paralogy belong to a most general definition, homology, which designates a relationship of common descent between any entities, without further specification of the evolutionary scenario. Accordingly, the entities related by homology, in particular, genes, are called homologs. The term homolog was introduced by Richard Owen in 1843 to designate *“the same organ in different animals under every variety of form and function”*. Owen clearly distinguished homologs from analogs, which he defined as a *“part or organ in one animal which has the same function as another part or organ in a different animal”* (Owen, 1848). The distinction between orthologs and paralogs and the terms themselves were introduced by Walter Fitch in 1970: *"Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact)"* (Fitch, 1970). In summary orthologs and paralogs are two types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication (Koonin, 2005).

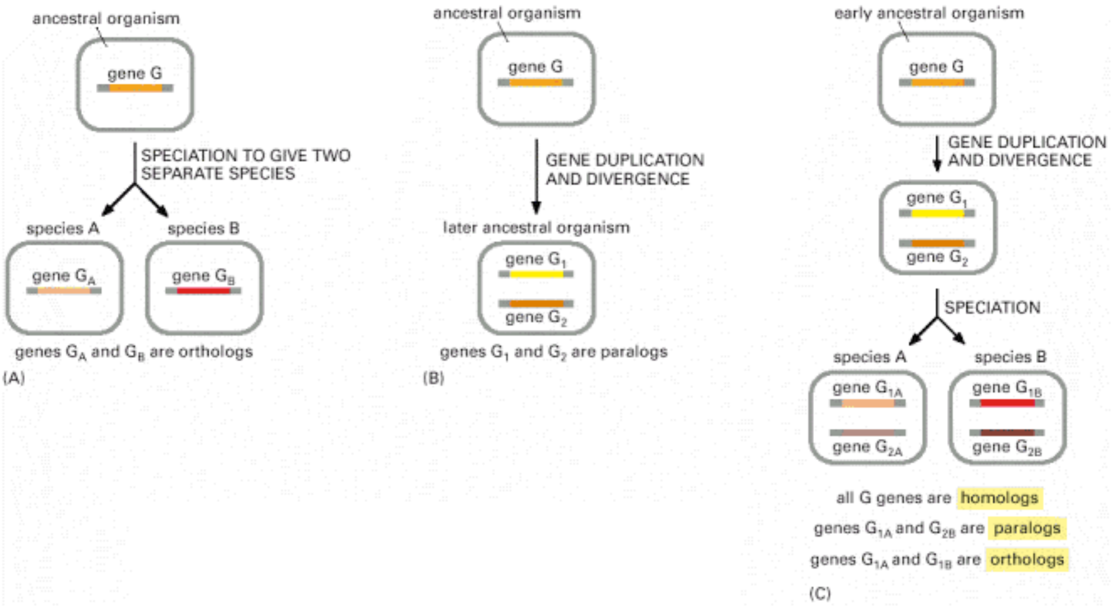


Figure 13. **Paralogous genes and orthologous genes**: two types of gene homology based on different evolutionary pathways. (A) orthology, (B) paralogy, (C) A more complex pattern of events that can occur. Taken from Molecular Biology of the Cell, chapter 1 (Alberts et al., 2002).

Part II

OBJECTIVES

Objectives

The objectives of this PhD thesis can be summarized as follows:

1. Compare the properties of amino acid tandem repeats composed of different amino acids.

2. Study the polymorphism associated with amino acid tandem repeats in the human proteome.

3. Analyze the relative contribution of slippage and point mutation mutational mechanisms in amino acid tandem repeat evolution.

4. Study the effect of the protein sequence selective constraints over the conservation, abundance and structure of amino acid tandem repeats.

5. Identify amino acid tandem repeats that may have a functional role on the basis of phylogenetic conservation..

6. Analyze the specific properties of polyQ and polyA disease-associated repeats in mammalian wild-type proteins.

Part III

RESULTS

Chapter 1

Summary

This chapter reports a study of the types of mutations occurring in the amino acid repeat component of the human proteome. We used a large collection of expressed sequence tags (ESTs) to estimate the intra-specific variability of tandem repeats and adjacent sequences in humans. We found that the number of gap polymorphisms in the regions surrounding repeat is five times lower than that within repeats, indicating a greatly reduced slippage activity outside the repeats. However the number of substitution is comparable to that within the repeats. We also found that non-synonymous substitution rate are lower than synonymous rate, both in repeats and adjacent sequences, indicating that selection plays a role in shaping the amino acid content of these regions.

The results of this work has been published in the following research article:

Mularoni, L., Guigo, R. and Alba, M.M.: **Mutation patterns of amino acid tandem repeats in the human proteome**. Genome Biol 7 (2006) R33.

Mularoni L, Guigó R, Albà MM.
Mutation patterns of amino acid tandem repeats in the human proteome.
Genome Biol. 2006;7(4):R33.

Chapter 2

Summary

This chapter reports two inter-specific studies based on the analyses of the conservation and of the mutational dynamics of tandem amino acid repeats in the human and mouse, and human and chimp, orthologous genes. We found a significant positive correlation between repeat size difference and protein non-synonymous substitution rate, which was observed for all the common amino acid repeat types. However, we also found that slowly evolving proteins, subject to stronger selective constraints, contain an unexpectedly large number of repeats, which tend to be conserved between the two species. As highly constrained proteins only rarely incorporate new repeat structures, there is evidence for a more important role than previously suspected of selection in the preservation of repeats.

The results of these works has been published in the following research articles:

Mularoni, L., Veitia, R.A. and Alba, M.M.: **Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats.** Genomics 89 (2007) 316-25.

Mularoni, L., Toll-Riera, M. and Albà, M.M.: **Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genome**s. Encyclopedia of Life Sciences (ELS). John Wiley&Sons, Ltd: Chichester. (2008).

Mularoni L, Veitia RA, Albà MM.
Highly constrained proteins contain anunexpectedly large number of amino acid tandem repeats.
Genomics. 2007Mar;89(3):316-25.

Mularoni L, Toll-Riera M, Albà MM.(April 2008)
Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genomes.
In: Encyclopedia of Life Sciences (ELS). John Wiley&Sons, Ltd:Chichester

Chapter 3

Summary

This chapter reports a study of the properties of amino acid tandem repeats in a dataset of orthologous proteins from 12 genomes. In order to better understand the selective forces that shape their evolution, we classified human repeats into seven groups, depending on the phylogenetic depth of conservation of the repeats. Then, we compared the distribution of coding repeats with that of repeats found in non-coding regions. We found an excess of repeats conserved across all vertebrate, suggesting a strong effect of negative selection in maintaining many repetitive structures. In this scenario, highly conserved repeats, especially if long, are likely to be functional. We also studied if there were an enrichment of developmental proteins among proteins containing mammals specific repeats, as reported by previous works. However, we found an approximately linear rate of repeats expansion and retention across vertebrate evolution.

Patterns of emergence of amino acid tandem repeats in the vertebrate phylogeny

Loris Mularoni[1], Alice Ledda[1], Macarena Toll-Riera[1], M.Mar Albà[1,2]

[1]Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Barcelona, Spain. Catalan Institution for Advanced and Research Studies, Barcelona, Spain.

ABSTRACT

Tandem amino acid repeats are evolutionary labile structures found in a large number of eukaryotic proteins. They are especially abundant in transcription factors and developmental proteins, where they can potentially modulate protein-protein interactions and/or influence transcriptional regulation. Given their variability, they may play important roles in adaptive processes. However, little is still known about the selective forces that shape their evolution. To gain further insights into this question, we present the first large-scale analysis of the properties of amino acid tandem repeats in a dataset of orthologous proteins from 12 vertebrate species. Intriguingly, we find that humans contain more repeats than any other species analyzed. The human repeats are then classified into seven groups depending on the phylogenetic depth of conservation of the repeat. The results show that there is a 3-fold excess of repeats which show very strong conservation (in all vertebrates) with respect to equivalent repeats from non-coding regions, suggesting a strong effect of negative selection in maintaining many repetitive structures. Contrary to previous observations based on several protein families, we do not find an enrichment of developmental proteins among proteins containing mammalian-specific repeats. The data indicates an approximately linear rate of repeat expansion and retention across vertebrate evolution.

INTRODUCTION

Tandem amino acid repeats, also known as homopolimeric tracts or homopeptides, are regions within proteins characterized by the consecutive repetition of a particular amino acid. These regions are very common in eukaryotic proteins (Green and Wang, 1994; Hancock and Simon, 2005; Li et al., 2004), being found in nearly one fifth of human proteins (Alba and Guigo, 2004; Karlin et al., 2002). Amino acid tandem repeats can suffer rapid expansions or contractions by trinucleotide slippage (Levinson and Gutman, 1987). When the repeats are short, expansion seems to dominate over contraction (Xu et al., 2000). Due to the high mutation rates associated with this process, of the order of 10^{-3} per locus and generation in humans (Weber and Wong, 1993), repeats are an important source of genetic variability (Mularoni et al., 2006; Wren et al., 2000) and may be involved in adaptive processes (Kashi and King, 2006; Kemp et al., 1987; Pizzi and Frontali, 2001). In humans, the uncontrolled expansion of tandem amino acid repeats is associated with several neurological or developmental diseases. The first type of disorders, which includes Huntington disease and several ataxias, involves very long expansions of poly-glutamine tracts (Gatchel and Zoghbi, 2005). Developmental diseases, in contrast, are associated to long poly-alanine tracts, for example in the HOXD13 protein causing Synpolydactyly type II (Brown and Brown, 2004).

The functional role of amino acid tandem repeats is still controversial. In non-coding regions, microsatellite-type sequences are considered to evolve in a neutral manner. In a similar manner, it is an extended belief that amino acid repeats are evolutionary neutral elements, which are tolerated in proteins as long as they do not alter the protein’s function (Lovell, 2003). However, there is increasing evidence that repeats can be of functional importance. First, amino acid tandem repeats are not found at random in different types of proteins, but there is a strong enrichment in transcription factors and developmental proteins (Alba and Guigo, 2004; Cocquet et al., 2003; Faux et al., 2005; Karlin et al., 2002; Mar Alba et al., 1999; Richard and Dujon, 1997). Second, experimental data based on alanine and glutamine homopeptides has shown that repeats can modulate protein-protein interactions (Buchanan et al., 2004; Shimohata et al., 2000), exert effects on gene transcriptional activity (Brown et al., 2005; Dunah et al., 2002; Gerber et al., 1994; Lanz et al., 1995) or confer an appropriate spacing between different protein domains (Newfeld et al., 1994). Third, several studies have linked changes in repeat size to modification in developmental processes. For example, poly-glutamine and poly-alanine repeat size variations in the RUNX-2 protein were found to correlate with skull morphology in purebred dogs, suggesting that changes in repeat size could be associated to very rapid phenotypic changes in this species (Fondon and Garner, 2004). In another study, elimination of a poly-alanine tract in HOX-D13 in transgenic mice resulted in fused sesamoid bones, as opposed to non-fused bones in wild-type mice (Anan et al., 2007). Fourth, many repeats in proteins are conserved across different species, suggesting the existence of selective pressures to maintain the repeat. In a previous work we found that highly constrained, slowly evolving, proteins, including many developmental factors, contained an expectedly high number of conserved repeats between human and mouse (Mularoni et al., 2007). Evolutionary conservation of alanine and glycine repeats across mammals has been reported for several transcription factors, including HOX proteins (Anan et al., 2007; Lavoie et al., 2003; Mortlock et al., 2000), GATA factors (Lavoie et al., 2003) and class III POU proteins (Sumiyama et al., 1996). Intriguingly, repeats in these proteins were almost completely absent from vertebrate homologues, which suggested increased rate of repeat expansion in more GC-rich environments (Nakachi et al., 1997; Sumiyama et al., 1996).

The above-mentioned studies have provided important insight into the evolutionary patterns of different amino acid tandem repeats but remain insufficient to comprehend the strength of the constraints that may have operated in these regions in general. For example, is mammalian-specificity a common feature of repeats found in developmental proteins? And, more generally, is the level of conservation of amino acid repeats generally compatible with a scenario of neutrality? Here we present a large-scale analysis of the distribution and properties of amino acid repeats in orthologous proteins from 12 vertebrate species, and the systematic identification of repeats which show taxon-specific conservation patterns. By comparing the results obtained in coding and non-coding sequences, as well as in different protein functional groups, we are able to address these questions.

RESULTS

Identification of amino acid tandem repeats in vertebrate orthologous proteins

We obtained a large dataset of vertebrate orthologous proteins from Ensembl (Hubbard et al., 2007), comprising 12 different species (human, chimpanzee, rhesus macaque, mouse, rat, dog, cow, opossum, chicken, frog, zebrafish and pufferfish). We identified all tandem amino acid repeats of size 4 or longer. The most common amino acids were glutamic acid, proline, serine, leucine, alanine, glyicine, lysine and glutamine, in agreement to previous studies (Alba and Guigo, 2004; Karlin et al., 2002). The number of repeats per protein was somewhat larger in mammalian proteins than in non-mammalian proteins, especially in humans, containing the largest number of repeats (Figure 1). Similar results were obtained using a repeat size cut-off of 5 (data not shown).

Definition of taxon-specific amino acid repeats

In order to gain an understanding on the origin and evolutionary constraints of repeats, we next focused our attention on repeats showing a well-defined phylogenetic distribution (Figure 2). Human repeats were classified in seven taxon-specific groups, according to the presence or absence of a similar repeat in other vertebrate lineages. The first group was ‘Primata’, containing repeats that, besides human, were found in chimpanzee and rhesus macaque orthologous proteins (size 4 or longer), but not in the rest of vertebrate orthologues. These repeats are likely to have originated, or expanded to a significant length, in a primate ancestor. Similarly, we defined the classes ‘Euarchontoglires’, ‘Eutheria’, ‘Mammalia’, ‘Amniota’, ‘Tetrapoda’ and ‘Vertebrata’ (Figure 3). In total, 2024 repeats could be classified, allowing the comparison of of the level of repeat gain at different periods of vertebrate evolution.

Comparison of repeats of different age

The most striking feature, especially considering the high inherent mutability of repeats, was the large size of the group Vertebrata (574 repeats; Figure 3, coding repeats). These repeats have been conserved for, at least, circa 450 Mya (Hedges and Kumar, 2002; Kumar and Hedges, 1998), so it seems likely that they contribute to some protein functionalities. A selection of proteins containing long very well conserved repeats is shown in Table 1. The number of repeats in the rest of groups followed a pattern that was largely compatible with an approximately linear rate of repeat gain (and retention), as the number of repeats was approximately proportional to the length of the phylogenetic branch (Miller et al., 2007). This led us to discard any sudden burst of repeat expansion during the period studied.

Interestingly, about one third of the human proteins analyzed contained more than one amino acid repeat (Figure 4). In most cases, the repeats belonged to different taxon-specifc classes. In particular, among proteins containing exactly 2 repeats, only 10,8% (45 proteins) contained repeats that were from the same group. This finding illustrates the fact that some repeat-prone proteins may continue to evolve by the continuous acquisition of new repeats.

Nearly one third of vertebrate conserved repeats have an earlier origin

The vast majority of vertebrate proteins in this analysis had homologues in non-vertebrate species (~95%), as determined by BLASTP searches against other eukaryotic genomes (see Methods). This implied that some of the repeats classified as Vertebrata could have arisen before the formation of the group. To investigate this, we examined the 3 best BLAST hits against C.intestinalis, D.melanogaster and C.elegans. About 23% were conserved in D.melanogaster and/or C.elegans (135 out of 574), and 7% only in C.intestinalis. Therefore, for about one third of repeats there were hints of an even earlier origin.

Comparison of coding and non-coding repeats

Could the observed patterns of amino acid repeat conservation have arisen by chance? To measure the deviation from am scenario of neutrality we computed the conservation of equivalent repeats from non-coding regions. We defined non-coding repeats as arrays of triplets that would translate into amino acid repeats if they were coding (for example an array of 3 GAA and 1 GAG would be equivalent to a repeat of 4 glutamic acid residues). Such regions are expected to evolve neutrally, or at least under non-protein related constraints. Most of the common amino acid repeat types were under-represented among non-coding repeats, glutamic acid and alanine showing the sharpest coding enrichment (Supl. Figure 1). More surprisingly, all common amino acid repeats except one (lysine) were significantly longer in human coding sequences than in non-coding sequences (Supl. Table 1). Therefore, long repeats of particular amino acids seem to have been favored in coding regions over the neutral expectation.

The phylogenetic conservation of human non-coding repeats in vertebrates was determined in a similar way as for protein sequences, using vertebrate genomic sequence alignments from UCSC Genome Database (Karolchik et al., 2003). In general, the non-coding repeats were much more poorly conserved (Figure 3, non-coding repeats, p-value < 0.001). A large amount of them were not conserved beyond the primates (41%), and only 9.7% were conserved at the Vertebrata level, in contrast to 23% for coding repeats. The patterns were generally consistent for different types of repeats (Sup. Table 1).

Functional analysis of proteins containing repeats

Next we studied the distribution of Gene Ontology functional terms (Ashburner et al., 2000) across the different groups. We created 3 non-redundant groups based on the most common words found in GO annotations of proteins containing repeats (see Methods). The group named 'Development and TF' contained 961 repeats that lied in proteins with development, genesis and transcription related functions. The group 'Membrane and Receptor' was composed by 1090 repeats found in proteins involved in membrane, receptor and transport related functions. The group 'Metabolism' was the smallest one (252 repeats) and comprised repeats that lied in proteins with metabolic and/or transferase activities. Although some types of amino acid repeats, such as poly-alanine or poly-glutamine, have generally been associated to transcription factors, we found that, except for leucine repeats, that were more abundant in the 'Membrane and receptor' group (p-value < 0.001), the rest of amino acids did not show strong differences (Figure 5).

We next determined whether the type of function had any impact on the level of conservation of the repeat. This was motivated by early observations indicating that developmental transcription factors contained many mammalian-specific repeats (Sumiyama et al., 1996, Nakachi et al., 1997, Mortlock et al., 2000). It was argued that pressure for increased GC content in mammalian genes would have triggered the expansion of A, G and P repeats, as they are encoded by GC-rich codons (Sumiyama et al., 1996). However, we did not observe any significant differences among the three functional groups (Figure 3). This is perhaps not surprising considering that such amino acid repeat are well represented in the three groups.

Discussion

Here we have presented the first large-scale comparison of amino acid tandem repeats in vertebrates using orthologous proteins. The data shows that there is a moderate enrichment of repeats in mammals with respect to non-mammalian vertebrates, which may be related to increased GC content in mammalian genes (Nakachi et al., 1997; Alba and Guigo, 2004). We have used the repeats in the human proteins to study the depth of repeat phylogenetic conservation. The main conclusion is that there is an excess of repeats conserved in all vertebrates compared to control non-coding regions. Highly conserved repeats, particularly if long, are likely to be functional (Table 1). This is illustrated by human protein kinase DYRK1A, listed in Table 1. This protein contains a 13 amino acids long histidine repeat, of similar length in other vertebrates. (11-13), and which is both necessary and sufficient to localize this protein to the nuclear speckle factor compartment (Alvarez et al., 2003).

Previous studies showed that poly-alanine repeats in some transcription factors, including HOX and GATA proteins, tended to be conserved in mammals but were absent from other vertebrate homologues (Anan et al., 2007; Mortlock et al., 2000; Nakachi et al., 1997). One of the aims of this work was to determine if this was a general trend. However, the data does not support a sudden increment of repeats at the base of the mammals, but rather it shows that repeats of different age usually coexist within the same protein. As just mentioned, most of the amino acid repeat types in vertebrate proteins are GC-rich. However, there is an interesting exception, lysine, encoded by AAA or AAG. It is revealing that the proportion of lysine repeats classified as Vertebrata is larger than for other amino acids (45% compared to 20-35%). One possible explanation is related to compositional differences in vertebrate or metazoan ancestral genome, maybe leading to a greater expandability of this amino acid in the past. However, it may also be that such highly charged amino acid stretches are normally highly disruptive, decreasing the chances that novel expansions are tolerated.

Trinucleotide slippage can result in expansions or contractions of the repetitive unit. When the repeats are short, expansion appears to dominate over contraction (Xu et al., 2000). Starting from short repeat “seeds” in coding sequences, slippage will continuously generated longer amino acid runs, which can be put to test. In case the repeat is irrelevant for the protein function, non-synonymous mutations will accumulate and the repeat will gradually “degenerate”. Alternatively, if the repeat provides a new, useful functionality, it will enter the selective path. In agreement with this model, it has been observed that novel, species-specific, repeats are often encoded by pure codon runs (for example CAGCAGCAGCAGCAG), whereas well-conserved repeats tend to include synonymous mutations (for example CAGCAGCAACAGCAG) (Alba et al., 1999). Similarly, in the current study we found that the percentage of amino acid repeats encoded by pure trinucleotide runs was superior in the younger groups (18.5% in Primata) than in older groups (10%-13% Eutheria or older). Interestingly, there is evidence that regions surrounding well-conserved repeats are generally better conserved than regions surrounding poorly-conserved repeats (lower non-synonymous to synonymous ratio, Ka/Ks), indicating that these regions are likely to be under selection (Hancock et al., 2001; Mularoni et al., 2007). So , there is a strong link between the type of evolution of the repeat and the characteristics of the region where it is located.

Some amino acid tandem repeats, such as alanine or glycine repeats, are know to be abundant in transcription factors and developmental proteins (Alba and Guigo, 2004; Faux et al., 2005; Nakachi et al., 1997), where they may influence transcriptional regulation (Buchanan et al., 2004; Dunah et al., 2002; Galant and Carroll, 2002). We were surprised to find that these types of amino acids are often also conserved in other kinds of proteins, such as receptors. It is unclear which function they may play in this context. In contrast, leucine repeats appear almost exclusively in membrane proteins. These repeats are typically located at the N-terminal end of proteins (Alba and Guigo, 2004), where they may act as signal peptides (Karlin et al., 2002).

Tandem amino acid repeats are probably the most conspicuous repetitive sequences in proteins, but they only represent a small part of the total amount of low complexity sequences. Current difficulties for the study of this type of sequences are the lack of appropriate evolutionary models and of accurate sequence alignment methods. More work in these areas will be required to obtain deeper insights into the evolutionary dynamics of repeats.

METHODS

Sequences

Protein and cDNA sequences for human (*Homo sapiens),* chimpanzee *(Pan troglodytes),* rhesus macque *(Macaca mulatta),* mouse *(Mus musculus),* rat *(Rattus norvegicus),* dog *(Canis familiaris),* cow *(Bos taurus),* opossum *(Monodelphis domestica),* chicken *(Gallus gallus),* frog *(Xenopus tropicalis),* zebrafish *(Danio rerio)* and pufferfish *(Takifugu rubripes)* were retrieved from Ensembl database release 46 (Hubbard et al., 2007). The genomic sequences were obtained from UCSC (golden path 200603) and the coordinates of coding regions were retrieved from Ensembl (based on NCBI36).

Identification of amino acid repeats in vertebrate proteins

We extracted orthologous gene pairs between human and the other vertebrate species listed above from Biomart at Ensembl (Kasprzyk et al., 2004). We obtained 6477 orthologous groups containing sequences from all groups of species listed above. Only 1:1 orthologous genes were considered with the exception of bonefishes, which underwent an whole-genome duplication about 350 million (Panopoulou and Poustka, 2005), and for which we considered both paralogous copies of each gene.

We used an in-house computer program for the detection of all amino acid tandem repeats of size 4 or longer, their positions, and the DNA sequences encoding the repeats, in all proteins from the orthologous protein dataset. We calculated codon homogeneity (CH) as the fraction of the repeat occupied by the longest pure codon run (Mularoni et al., 2007).

Identification of non-coding repeats

For comparison, we identified arrays of synonymous triplets in non-coding sequences. For example an array of 3 GAA and 1 GAG in a non-coding sequences would be equivalent to a repeat of 4 glutamic acid residues. In this work, this type of sequences have been labeled non-coding repeats. We obtained the current human genome sequence from UCSC (goldenpath 200603), identified all non-coding repeats, and discarded those that fell into protein coding regions according to annotations in Ensembl.

Amino acid repeat conservation

We next determined whether repeats present in one species were also present in other vertebrate orthologous protein sequences. We built pairwise sequence alignments by T-Coffee (Notredame et al., 2000), and identified equivalent repeats on the alignments by the procedure previously described (Mularoni et al., 2007). Briefly, for each repeat of the reference species (human) an equivalent repeat existed in the other sequences if it overlapped in the alignment and had a size above the threshold (4 in this study). In cases where the protein appeared incomplete we attempted to recover full protein sequences from Genbank using BLASTP (Altschul et al., 1997). Where there was multiple repeat overlap we selected the longest overlapping repeat. Repeats that contained, or were surrounded by, 35 gaps or more were not considered, as these regions of the alignment were of very poor quality. In addition, neither terminal non-conserved regions nor internal regions with very long gaps (>= 35 gaps) were considered, as these regions could represent parts of the gene that had not yet been annotated, such as missing exons.

Definition of taxon-specific repeats

Taxon-specific repeats were defined as those that were present in the species of the taxon but absent from more distant species. This allowed us to obtain repeats with a distribution compatible with a single event of expansion above the length cut-off (4 amino acids). The first group was ‘Primata’, containing repeats that, besides human, were found in chimpanzee and rhesus macaque orthologous proteins (size 4 or longer), but not in the rest of vertebrate orthologues. Similarly, we defined the classes ‘Euarchontoglires’ (Primata plus Rodents), ‘Eutheria’ (Euarchontoglires plus Laurasiatheria), ‘Mammalia’ (Eutheria plus *M.domestica*), ‘Amniota’ (Mammalia plus *G.gallus*), ‘Tetrapoda’ (Amniota plus *X.tropicalis*) and ‘Vertebrata’ (Tetrapoda plus *D.rerio*/*T.rubripes*). In the case of Rodents it was sufficient if one of the two species, mouse or rat, had the human repeat to consider that the repeat had been present in a common ancestor. The same was applied in the case of Laurasiatheria (dog and cow). This was a highly parsimonious approach that allowed us to circumvent possible problems in annotating repeats in some species, such as in dogs, containing a suspiciously low number of them. In the case of fishes the situation was sometimes complicated by the existence of two paralogous copies of the gene. We considered that the presence of the repeat in one of the copies, in at least one of the species (D.rerio or T.rubripes) was sufficient to assumed that the repeat was present in a vertebrate common ancestor. Following this criteria 2024 human repeats were classified in seven different taxon-specific group: Vertebrata (574), Eutheria (384), Tetrapoda (307), Mammalia (268), Amniota (239), Primata (157) and Euarchontoglires (95).

In the case of non-coding repeats we recovered genome multiple alignments from UCSC containing the same set of species described above. We obtained 846 repeats that could be classified in taxon-specific groups, following the same criteria than for amino acid repeats.

Homologous protein search

We searched for homologues in non-vertebrate eukaryotic species using BLASTP sequence similarity searches (E-value < 10^{-4}) of the human protein against species-specific protein libraries downloaded from Ensembl. A small percentage of proteins (~5%) lacked homologues in the six non-vertebrate species tested (A.gambiae, A.thaliana, C.elegans, D.melanogaster, S.cerevisiae, S.pombe), and therefore they may be vertebrate-specific proteins.

For proteins containing repeats classified as Vertebrata, we wished to determine if similar repeats existed in non-vertebrate homologues. As identification of orthology becomes more difficult for distantly-related species

(> 500 Mya) we analysed the three best BLASTP hits for C.intestinalis, E.elegans and D. melanogaster. In 36 cases a similar repeat existed only in C.intestinalis, suggesting it was a chordate-specific repeat, and in 135 cases in C.elegans and/or D.melanogaster, indicating the repeat was ancient (Metazoans).

Gene Ontology annotation

Repeat containing proteins were functionally annotated using Gene Ontology terms (Harris et al., 2004). The GO annotations for the human proteins were downloaded from Biomart at Ensembl. The most common words found in the GO annotation of the proteins studied were: “metabolic”, “response”, “genesis”, “transferase”, “receptor”, “transcription”, “transport”, “membrane”, “development”, and “differentiation”. We grouped together those terms that were functionally related and created three non-redundant datasets. The first dataset (‘Development and TF’) contained development, genesis and transcription related functions. The second dataset (‘Membrane and receptor’) contained membrane, receptor and transport related functions. The third dataset (‘Metabolism’) contained proteins involved in metabolic processes and/or with transferase activity. To avoid redundancy, genes already classified in the first group were eliminated from the rest. Genes initially classified in the second and third groups were kept in the second group only.

Statistical Analysis

Statistical analysis has been performed with the R statistical package (Team, 2007). The chi-square distribution has been used to calculated p-values unless stated.

FIGURE LEGENDS

Figure 1. Amino acid tandem repeats in vertebrate orthologous proteins. The average number of repeats per protein is provided.

Figure 2. Example of taxon-specific repeat.
Multialignment of transcription factor protein LHX2 (LIM homeobox 2). The red box enphasize the conservation of a polyA across mammals.

Figure 3. Amino acid repeats classified in different taxon-specific groups. In blue boxes the relative frequencies of coding and non-coding repeats (see text). In yellow boxes the relative frequencies of repeats in different protein functional groups. The tree has been built using the average number of substitutions per site reported in Miller et al., 2007.

Figure 4. Developmental transcription factors containing repeats. HOX and GATA factors containing multiple repeats. The age of repeats is indicated. ‘Likely’ is based on the most distant species with the repeat and not on well-defined taxon-specific groups.

Figure 5. Amino acid repeats and functional classes. Relative distribution of repeats composed of different amino acids in proteins with different functions.

SUPPLEMENTARY FIGURES AND TABLES

Supplementary Figure 1. Relative frequency of repeats of different composition in coding and non-coding sequences.

Supplementary Table 1. Average tandem repeat size for coding and non-coding repeats. Non-coding repeats were tandem repeats of any combination of codons for the amino acid. Only repeats with size 4 or longer were considered. p-value was obtained using Kolmogorov-Smirnov non-parametric test.

Supplementary Table 2. Taxon-specific human coding and non-coding repeats. Only repeats with size 4 or longer were considered.

Supplementary Data File. Mularoni_ et_al.xls contains the data used for this analysis. Available from http://evolutionarygenomics.imim.es/datasets.html

Table 1. Human repeats highly conserved across vertebrates.

Hs Ensembl Id	AA	Hs	Pt	Ma	Mu	Rn	Bt	Md	Gg	Xt	Dr	HUGO	Description
ENSP00000367057	A	9	9	9	9	7	9	9	9	9	9	DACH1	Dachshund homolog 1 (Dach1)
ENSP00000371160	A	9	9	9	9	9	9	9	9	9	9	PHOX2B	Paired mesoderm homeobox protein 2B
ENSP00000342831	A	8	8	8	8	8	8	8	8	8	9	RBM9	RNA-binding protein 9
ENSP00000282486	A	7	7	7	7	7	7	7	7	6	6	MBNL1	Muscleblind-like protein 1 (Triplet-expansion RNA-binding protein)
ENSP00000354370	A	7	7	7	8	8	7	5	5	5	5	PUM2	Pumilio homolog 2 (Pumilio-2)
ENSP00000363972	A	7	7	7	7	7	7	7	7	7	7	KIAA1166	Hepatocellular carcinoma-associated antigen 127
ENSP00000368563	A	7	7	7	7	7	7	7	7	7	7	CUGBP2	CUG triplet repeat, RNA binding protein 2 isoform 1
ENSP00000261366	E	8	8	8	9	9	8	9	6	5	4	LMNB1	Lamin-B1
ENSP00000353622	E	7	7	7	7	7	7	7	7	7	3	SIN3A	Paired amphipathic helix protein Sin3a (Transcriptional corepressor Sin3a)
ENSP00000297071	G	16	16	16	18	18	15	16	7	8	8	TRA2A	Transformer-2 protein homolog
ENSP00000216019	G	8	8	8	7	7	7	7	9	6	5	DDX17	Probable ATP-dependent RNA helicase DDX17
ENSP00000265165	G	8	8	8	6	6	6	10	5	4	6	LEF1	Lymphoid enhancer-binding factor 1 (LEF-1)
ENSP00000340373	H	13	13	13	13	13	13	13	13	11	12	DYRK1A	Dual specificity tyrosine-phosphorylation-regulated kinase 1A
ENSP00000351631	H	10	10	10	10	10	10	10	9	9	5	FAM76B	Protein FAM76B
ENSP00000334991	P	21	22	2	6	10	13	6	14	19	11	FMNL2	Formin-like protein 2
ENSP00000327442	P	11	9	7	8	8	8	7	4	8	8	FMNL1	Formin-like protein 1
ENSP00000333836	P	9	9	9	9	9	9	9	9	8	7	FIGN	Fidgetin
ENSP00000365411	P	8	9	8	8	8	8	7	8	10	11	APBB1IP	Amyloid beta A4 precursor protein-binding family B member 1-interacting protein
ENSP00000295851	P	7	7	7	7	7	7	7	7	7	6	ABI2	Abl interactor 2
ENSP00000360233	P	7	7	7	7	7	7	7	7	7	7	NFIA	Nuclear factor 1 A-type
ENSP00000216832	S	11	11	11	11	11	11	5	9	16	16	PNN	Pinin (140 kDa nuclear and cell adhesion-related phosphoprotein)
ENSP00000302916	S	10	10	0	10	10	10	10	10	10	10	JMJD6	JmjC domain-containing protein 6
ENSP00000369695	S	9	9	9	9	9	7	9	9	9	8	MLLT3	Protein AF-9
ENSP00000265814	S	7	7	7	7	7	7	7	7	7	7	RUNX1T1	Protein CBFA2T1
ENSP00000296003	S	7	7	7	7	7	7	6	7	7	7	MTMR14	Myotubularin-related protein 14 precursor
ENSP00000303928	S	7	7	7	7	7	7	7	7	7	6	KIAA0232	Uncharacterized protein KIAA0232

Table 1. Human repeats which are highly conserved across vertebrates.

AA = amino acid forming the repeat. Human (Hs), chimp (Pt), rhesus (Ma), mouse (Mu), rat (Rn), dog (Cf), cow (Bt), opossum (Md), chicken (Gg), frog (Xt), pufferfish (Tr) and zebrafish (Dr).

Figure 1.

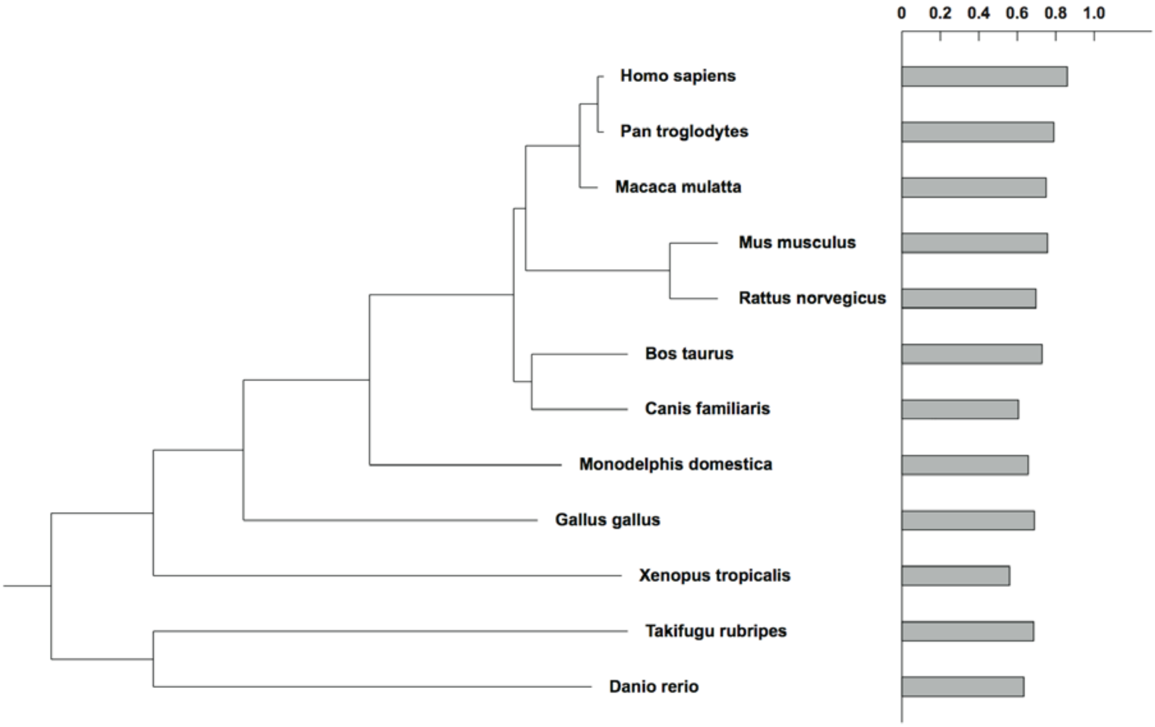
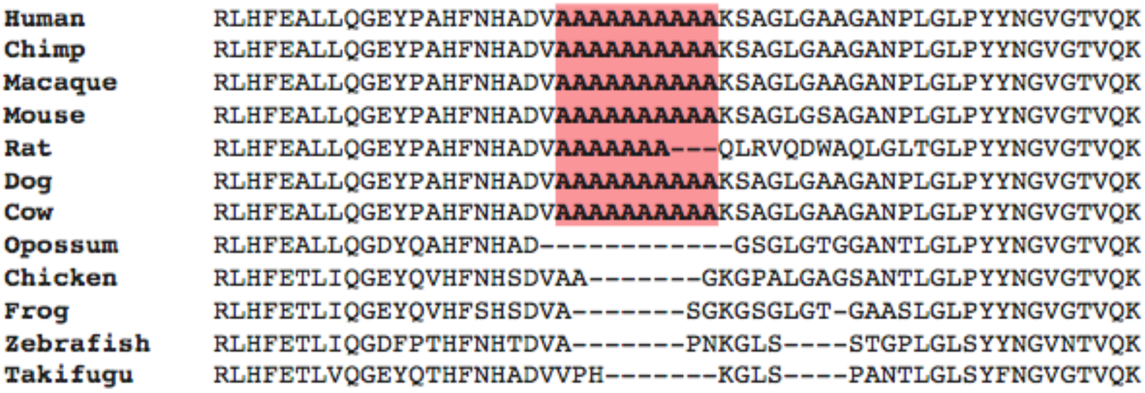

0
0.2
0.4
0.6
0.8
1.0
Homo sapiens
Pan troglodytes
Macaca mulatta
Mus musculus
Rattus norvegicus
Bos taurus
Canis familiaris
Monodelphis domestica
Gallus gallus
Xenopus tropicalis
Takifugu rubripes
Danio rerio

Figure 2.



```
Human        RLHFEALLQGEYPAHFNHADVAAAAAAAAAAKSAGLGAAGANPLGLPYYNGVGTVQK
Chimp        RLHFEALLQGEYPAHFNHADVAAAAAAAAAAKSAGLGAAGANPLGLPYYNGVGTVQK
Macaque      RLHFEALLQGEYPAHFNHADVAAAAAAAAAAKSAGLGAAGANPLGLPYYNGVGTVQK
Mouse        RLHFEALLQGEYPAHFNHADVAAAAAAAAAAKSAGLGSAGANPLGLPYYNGVGTVQK
Rat          RLHFEALLQGEYPAHFNHADVAAAAAAA---QLRVQDWAQLGLTGLPYYNGVGTVQK
Dog          RLHFEALLQGEYPAHFNHADVAAAAAAAAAAKSAGLGAAGANPLGLPYYNGVGTVQK
Cow          RLHFEALLQGEYPAHFNHADVAAAAAAAAAAKSAGLGAAGANPLGLPYYNGVGTVQK
Opossum      RLHFEALLQGDYQAHFNHAD------------GSGLGTGGANTLGLPYYNGVGTVQK
Chicken      RLHFETLIQGEYQVHFNHSDVAA-------GKGPALGAGSANTLGLPYYNGVGTVQK
Frog         RLHFETLIQGEYQVHFSHSDVA-------SGKGSGLGT-GAASLGLPYYNGVGTVQK
Zebrafish    RLHFETLIQGDFPTHFNHTDVA-------PNKGLS----STGPLGLSYYNGVNTVQK
Takifugu     RLHFETLVQGEYQTHFNHADVVPH-------KGLS----PANTLGLSYFNGVGTVQK
```

Figure 3.

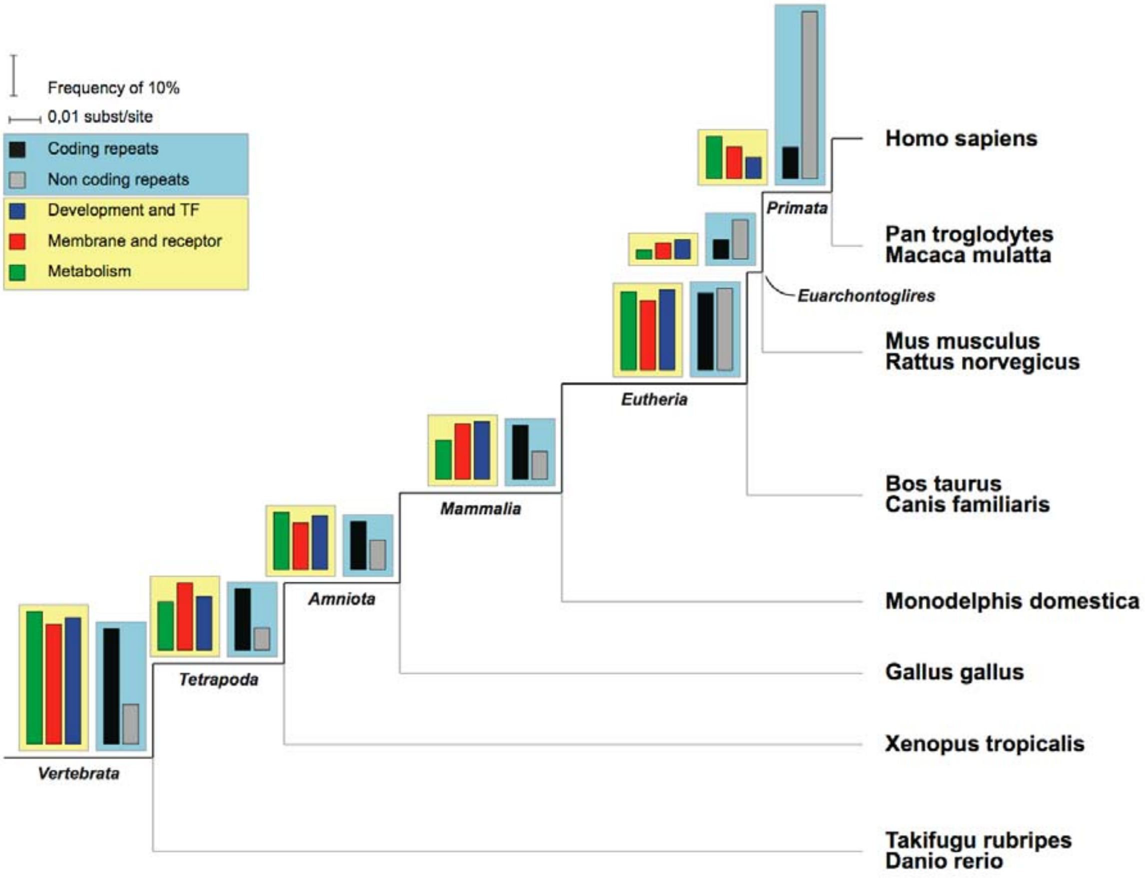

Frequency of 10%
0,01 subst/site
■ Coding repeats
■ Non coding repeats
■ Development and TF
■ Membrane and receptor
■ Metabolism
Primata
Euarchontoglires
Eutheria
Mammalia
Amniota
Tetrapoda
Vertebrata
Homo sapiens
Pan troglodytes
Macaca mulatta
Mus musculus
Rattus norvegicus
Bos taurus
Canis familiaris
Monodelphis domestica
Gallus gallus
Xenopus tropicalis
Takifugu rubripes
Danio rerio

Figure 4.

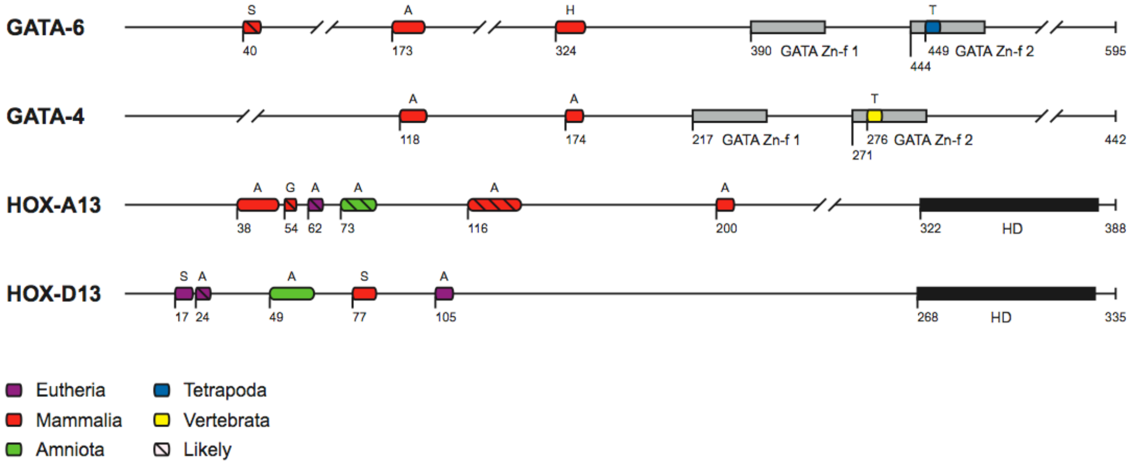
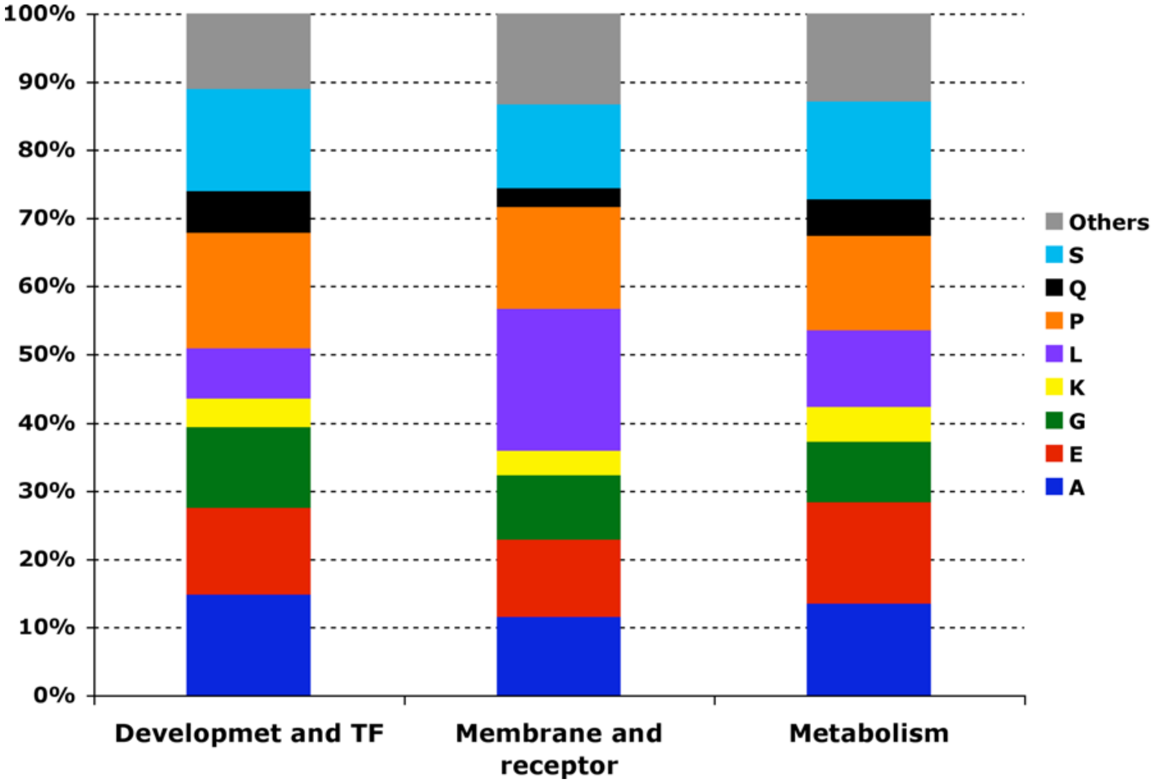
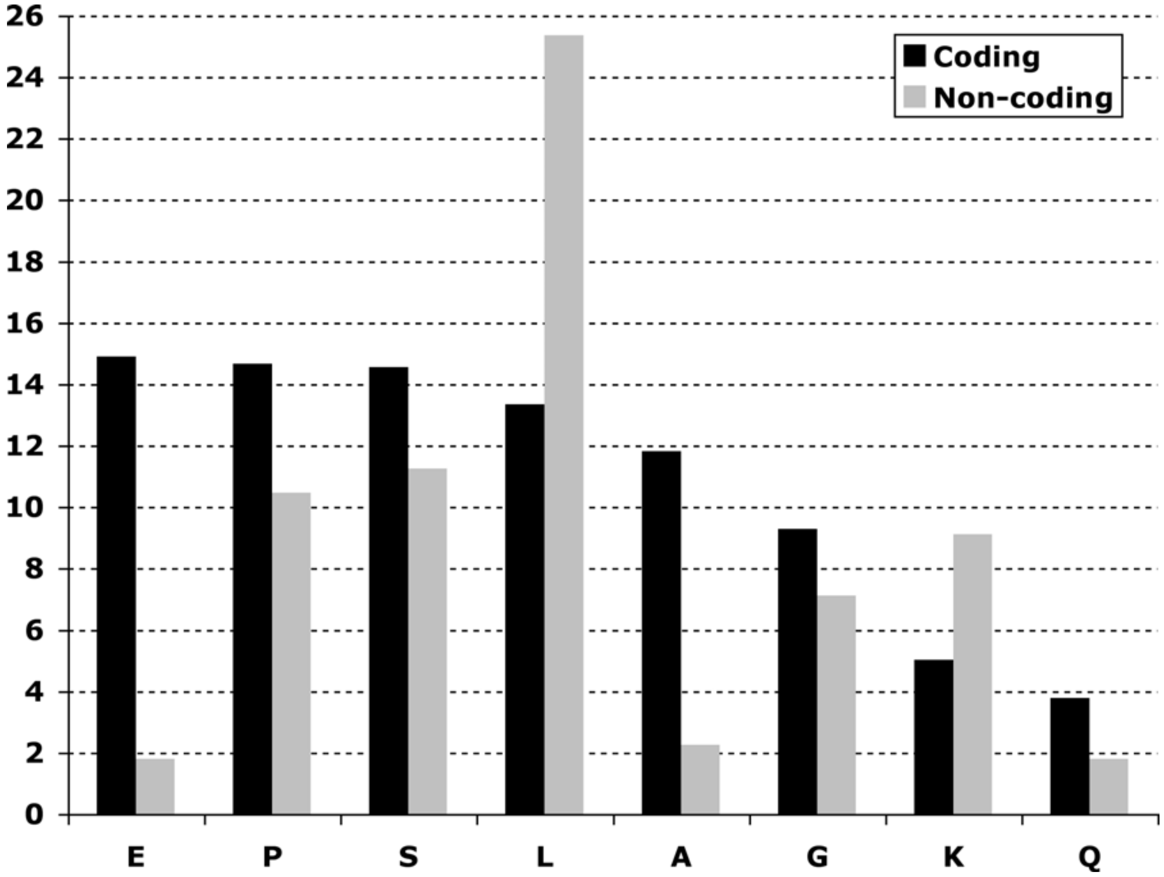

GATA-6
S
A
H
T
40
173
324
390 GATA Zn-f 1
449 GATA Zn-f 2
444
595
GATA-4
A
A
T
118
174
217 GATA Zn-f 1
276 GATA Zn-f 2
271
442
HOX-A13
A G A A
A
A
38 54 62 73
116
200
322 HD
388
HOX-D13
S A
A
S
A
17 24
49
77
105
268 HD
335
Eutheria
Mammalia
Amniota
Tetrapoda
Vertebrata
Likely

Figure 5.


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Developmet and TF
Membrane and receptor
Metabolism
Others
S
Q
P
L
K
G
E
A

Supplementary Figure 1.



Supplementary Table 1.

AA	Av. size coding (C)	Av. size non-coding (NC)	p-value (size C>NC)
A	5.18	4.45	$<10^{-5}$
E	5.09	4.81	$<10^{-5}$
G	5.011	4.25	$<10^{-5}$
K	4.59	4.31	N.A.
L	4.60	4.40	$<10^{-5}$
P	5	4.18	$<10^{-5}$
Q	6.35	4.80	$<10^{-5}$
S	4.75	4.30	$<10^{-5}$
all	4.919	4.600	$<10^{-5}$

Supplementary Table 2.

AA	Primata	Euarchontoglires	Eutheria	Mammalia	Amniota	Tetrapoda	Vertebrata	all
coding								
A	25	19	55	25	20	17	50	211
E	17	13	61	41	48	38	56	274
G	14	9	31	27	10	10	30	131
K	2	3	14	14	15	27	62	137
L	20	7	41	37	34	41	78	258
P	17	10	55	33	27	52	78	272
Q	1	6	12	9	14	17	32	91
S	41	20	70	42	40	55	91	359
all	157	95	384	268	239	307	574	2024
non-coding								
A	23	7	34	11	3	1	3	82
E	11	2	7	1	1	2	1	25
G	23	7	6	1	1	0	0	38
K	33	12	11	2	1	0	3	62
L	102	16	36	20	31	19	13	237
P	12	7	11	4	5	0	0	39
Q	4	1	2	0	0	5	1	13
S	50	12	38	5	8	11	39	163
all	329	77	161	55	58	44	78	802

REFERENCES

Alba, M. M., Guigo, R., 2004. Comparative analysis of amino acid repeats in rodents and humans. Genome Res. 14, 549-54.

Alba, M. M., et al., 1999. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. Mol Biol Evol. 16, 1641-4.

Altschul, S. F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-402.

Alvarez, M., et al., 2003. DYRK1A accumulates in splicing speckles through a novel targeting signal and induces speckle disassembly. J Cell Sci. 116, 3099-107.

Anan, K., et al., 2007. Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. Mol Biol Evol. 24, 281-7.

Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25, 25-9.

Brown, L., et al., 2005. In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. Hum Mol Genet. 14, 411-20.

Brown, L. Y., Brown, S. A., 2004. Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet. 20, 51-8.

Buchanan, G., et al., 2004. Structural and functional consequences of glutamine tract variation in the androgen receptor. Hum Mol Genet. 13, 1677-92.

Cocquet, J., et al., 2003. Compositional biases and polyalanine runs in humans. Genetics. 165, 1613-7.

Dunah, A. W., et al., 2002. Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. Science. 296, 2238-43.

Faux, N. G., et al., 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 15, 537-51.

Fondon, J. W., 3rd, Garner, H. R., 2004. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A. 101, 18058-63.

Galant, R., Carroll, S. B., 2002. Evolution of a transcriptional repression domain in an insect Hox protein. Nature. 415, 910-3.

Gatchel, J. R., Zoghbi, H. Y., 2005. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet. 6, 743-55.

Gerber, H. P., et al., 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science. 263, 808-11.

Green, H., Wang, N., 1994. Codon reiteration and the evolution of proteins. Proc Natl Acad Sci U S A. 91, 4298-302.

Hancock, J. M., Simon, M., 2005. Simple sequence repeats in proteins and their significance for network evolution. Gene. 345, 113-8.

Hancock, J. M., et al., 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. Mol Biol Evol. 18, 1014-23.

Harris, M. A., et al., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258-61.

Hedges, S. B., Kumar, S., 2002. Genomics. Vertebrate genomes compared. Science. 297, 1283-5.

Hubbard, T. J., et al., 2007. Ensembl 2007. Nucleic Acids Res. 35, D610-7.

Karlin, S., et al., 2002. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A. 99, 333-8.

Karolchik, D., et al., 2003. The UCSC Genome Browser Database. Nucleic Acids Res. 31, 51-4.

Kashi, Y., King, D. G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253-9.

Kasprzyk, A., et al., 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res. 14, 160-9.

Kemp, D. J., et al., 1987. Repetitive proteins and genes of malaria. Annu Rev Microbiol. 41, 181-208.

Kumar, S., Hedges, S. B., 1998. A molecular timescale for vertebrate evolution. Nature. 392, 917-20.

Lanz, R. B., et al., 1995. A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. Nucleic Acids Res. 23, 138-45.

Lavoie, H., et al., 2003. Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. Hum Mol Genet. 12, 2967-79.

Levinson, G., Gutman, G. A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 4, 203-21.

Li, Y. C., et al., 2004. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 21, 991-1007.

Lovell, S. C., 2003. Are non-functional, unfolded proteins ('junk proteins') common in the genome? FEBS Lett. 554, 237-9.

Mar Alba, M., et al., 1999. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol. 49, 789-97.

Miller, W., et al., 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 17, 1797-808.

Mortlock, D. P., et al., 2000. Evolution of N-terminal sequences of the vertebrate HOXA13 protein. Mamm Genome. 11, 151-8.

Mularoni, L., et al., 2006. Mutation patterns of amino acid tandem repeats in the human proteome. Genome Biol. 7, R33.

Mularoni, L., et al., 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics. 89, 316-25.

Nakachi, Y., et al., 1997. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol Biol Evol. 14, 1042-9.

Newfeld, S. J., et al., 1994. Drive-selection equilibrium: homopolymer evolution in the Drosophila gene mastermind. J Mol Evol. 38, 637-41.

Notredame, C., et al., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302, 205-17.

Panopoulou, G., Poustka, A. J., 2005. Timing and mechanism of ancient vertebrate genome duplications -- the adventure of a hypothesis. Trends Genet. 21, 559-67.

Pizzi, E., Frontali, C., 2001. Low-complexity regions in Plasmodium falciparum proteins. Genome Res. 11, 218-29.

Richard, G. F., Dujon, B., 1997. Trinucleotide repeats in yeast. Res Microbiol. 148, 731-44.

Shimohata, T., et al., 2000. Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. Nat Genet. 26, 29-36.

Sumiyama, K., et al., 1996. Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. J Mol Evol. 43, 170-8.

Team, R. D. C., 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Weber, J. L., Wong, C., 1993. Mutation of human short tandem repeats. Hum Mol Genet. 2, 1123-8.

Wren, J. D., et al., 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet. 67, 345-56.

Xu, X., et al., 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 24, 396-9.

Part IV

DISCUSSION

Discussion

1. Thesis overview

THE first part of this thesis places our work into context by providing a general overview of the tandem amino acid repeats by (i) reviewing the repetitive element of genomes, (ii) describing the known features of homopeptides, (iii) illustrating the different mutational processes responsible for tandem amino acid repeat evolution, and (iv) introducing the bioinformatics approaches used in the works presented. The second part enumerates the objectives while the third part describes the research works carried on during this thesis. Some of these studies have already been published as research articles (Chapter 1, (Mularoni et al., 2006) and Chapter 2, (Mularoni et al., 2007; Mularoni et al., 2008) while others are still in preparation (Chapter 3). Each work presented gained further insight about the evolutionary mechanisms of tandem amino acid repeats. In Chapter 1 we performed an intra-specific analysis of the types of mutations occurring in the amino acid repeat component of the human proteome, using a large collection of ESTs (Mularoni et al., 2006). The objectives of this work were to estimate the intra-specific variability of tandem repeats and adjacent sequences in humans, to compare the number of mutations in tandem repeats due to amino acid substitution and to slippage replication, to estimate the frequency of mutations inside and outside the repeats, and to identify differences in repeats composed of different amino acids. In Chapter 2 we performed inter-specific analyses of tandem amino acid repeats considering the orthologous proteins in human and mouse (Mularoni et al., 2007), and in human and chimp (Mularoni et al., 2008). The objectives of these studies were, first, to corroborate the hypothesis, suggested in a previous work (Hancock et al., 2001), that orthologous proteins that showed high non-synonymous substitution rates between two species tend to contain repeats which are more variable in size than proteins subject to stronger purifying selection. Second we analyzed the repeat content in proteins subjected to different selective constraints. Third, we studied specific properties of polyQ and polyA disease-associated repeats. In Chapter 3 we tried to better understand the evolutionary pattern of tandem amino acid repeats by analyzing their properties in a dataset of orthologous proteins from 12 genomes. The main objectives of this work were to compare the repeat content in proteins of different species, to study the effect of purifying selection in different phylogenetic groups, and study whether there was evidence of repeat taxon-specific functionality, as reported by other works (Sumiyama et al., 1996; Mortlock et al., 2000; Lavoie et al., 2003; Anan et al., 2007). In summary, the overall work presented in this thesis represents an effort to better understand the polymorphisms found within tandem amino acid repeats and the associated mutational mechanisms. The various articles included in this thesis are tightly connected to form a unique thread that puts together different perspectives of the evolution of amino acid tandem repeats.

2. Abundance of repeats

Tandem amino acid repeats are a very common feature of eukaryotic genomes and are present in a large number of coding proteins. Their abundance varies in the coding regions of different species, ranging from 13% in *Caenorhabditis elegans* to 27% in *Drosophila melanogaster* (Alba et al., 2007). First we studied their presence focusing on the human proteome. Then we extended our analysis to the orthologous proteins of human and mouse, and human and chimp. Finally we tried to gain better insight by considering the orthologous proteins from 12 vertebrate genomes. In general we found that mammalian proteins contained a slight larger number of repeats than non-mammalian proteins. This could be explained by the increased GC content of mammalian genomes, as previous studies have pointed to the existence of a relationship between GC content and the frequency of homopeptides in mammalian genes (Sumiyama et al., 1996; Nakachi et al., 1997; Alba and Guigo, 2004). Among the genomes analyzed, human showed the highest average number of repeats per protein.

3. Amino acid repeats features

According to the 4 research works presented in this thesis, the most common amino acid repeat types in mammals were alanine (A), glutamic acid (E), glycine (G), lysine (K), leucine (L), proline (P), glutamine (Q), and serine (S). These structures showed very diverse features. Glutamine repeats were the longest ones and showed a higher propensity to suffer expansions than other types of repeats. In the first place, an important fraction of the human repeats composed of glutamine are polymorphic and polyQ is among the most common structures that composed highly variable repeats between human and mouse. Second, in the human proteome, 88% of the polymorphic variants composed by glutamine repeats contained gaps. Third, glutamine repeats showed the highest codon homogeneity, which has been associated with the propensity to suffer slippage mutation (Alba et al., 1999; Wren et al., 2000). Notably, abnormal expansions of glutamine runs are associated with an increasing number of neurodegenerative disorders. Up to date, in humans, 30 hereditary disorders have been found to be triggered by tandem amino acid repeat instability and of these, 9 are caused by glutamine expansions (Table 3). Many of these disease-associated repeats are not only highly polymorphic in humans but they also showed a high variability in the orthologous proteins of human and mouse, and human and chimp. Intriguingly, disease-associated glutamine repeats are generally much shorter in rodent species than in humans, probably denoting a recent expansion in primates (Gibbs et al., 2004; Huang et al., 2004). Besides, a number of them show highest length variability in humans than in other apes, indicating increased expansion potential in our species (Andres et al., 2004).

In contrast to glutamine repeat instability, uncontrolled expansion of alanine has been associated with 9 developmental diseases (Table 3). Several of these disease-associated repeats were highly conserved between human and mouse, and human and chimp, and generally they have been found to be characterized by a low codon homogeneity. The structure of the repetitive tract, composed by a mixture of synonymous codons, is little prone to slippage events, and a different mutational mechanism, unequal crossing-over, has been proposed to explain their mutational dynamics (Warren, 1997).

Proline repeats appear to be little exposed to the action of slippage replication, and only 14% of its polymorphic variants contained gaps. It is interesting to note that, in spite of their low propensity to suffer size contraction or expansion, proline repeats in humans are not only very abundant but they are among the most common repeat types. The fact that this homopeptide is so abundant and at the same time it shows little polymorphism, could indicate a widespread role in mediating protein-protein interactions. In fact proline-rich regions are often found in protein-protein interaction surfaces (Kay et al., 2000) and proline tandem repeats have been found to be strongly associated with ‘protein binding’ functional annotation (Alba and Guigo, 2004). Another intriguing feature of proline homopeptides is that they show a non-synonymous substitution rate markedly higher than synonymous ones, suggesting that positive selection may have a role in the formation of these repeats.

Leucine repeats showed a number of interesting differences with respect to other amino acid repeat types. Fist, the analysis of the orthologous proteins in human and mouse showed that these kinds of repeats were the shortest ones, with an average repeat size of 5.8 residues. Second, in the human proteome, they showed the smallest polymorphism value across the most common amino acid repeat types. In addition, leucine repeats contained much more substitutions in the regions adjacent the repeat than in the repeat itself, as expected in the case of stronger functional constraints inside the repeats. Another intriguing feature is that leucine repeats, contrary to the other amino acid repeat types analyzed, did not show any correlation between non-synonymous substitution rate and the repeat content of the protein. In summary, leucine repeats are very abundant in fast evolving proteins but at the same time they show high conservation and stability, suggesting that these structures may be shaped by different types of functional constraints.. In fact, leucine repeats are often found at the amino terminus of transmembrane receptor proteins (Alba and Guigo, 2004), where it has been suggested that they could function as signal peptides (Karlin et al., 2002).

4. Balance between slippage replication and point mutation

It has been suggested that the evolutionary dynamics of tandem amino acid repeats can be the result of a balance between the action of slippage replication and that of point mutation (Kruglyak et al., 1998; Santibanez-Koref et al., 2001). The former would mainly promote the increase in length of the repetitive tract while the latter would lead to the accumulation of synonymous and non-synonymous substitutions in the repetitive structure. The inclusion of a synonymous substitution does not change the length of the repeat but does affect the propensity of the repetitive structure to suffer further slippage events. On the other hand, the inclusion of a non-synonymous substitution results in the interruption of the tandem repeats into two smaller tracts, an event that it is referred to as "death of the tandem amino acid repeat". The different frequencies of gaps and substitutions that can be observed in different types of repeats are, therefore, likely to reflect the different strength of these two evolutionary forces, coupled with the action of selection at the protein level. The analysis of tandem amino acid repeats polymorphism in humans showed that many of the variants containing gaps were encoded by a high rate of pure codon runs. These variants may have originated by trinucleotide slippage, as high level of codon homogeneity has been linked to increased repeat expansion (Kunst et al., 1997). This observation has been confirmed by the finding that highly conserved repeats in human and mouse (as well as in human and chimp) orthologous proteins, are generally encoded by more interrupted codon structure, containing mixture of synonymous codons, than highly variable repeats, which tend to be encoded by continuous runs of the same codon. Structures composed by a mixture of synonymous codon seem to be more stable, being much less prone to slippage event

(Schlotterer, 1998; Alba et al., 1999; Mularoni et al., 2006), and may represent relatively old repeats that have been maintained by selection. Therefore natural selection could have preserved functionally important repeats by shaping their structures towards mixed codons tracts. The accumulation of synonymous substitutions would have maintained the amino acid repeat while at the same time protecting it from further expansions or contractions.

5. Evolutionary dynamics of tandem amino acid repeats

As mentioned above, the evolutionary dynamics of trinucleotide structures can be explained as a balance between expansion by slippage and growth interruption by point mutation. We looked to see if there was any difference in the intra-specific variability of repeat structures and adjacent regions. Interestingly we found that the mutations that included gaps were generally more common in the repeat with respect to adjacent regions, suggesting an increased slippage activity in the repeat structures. On the other hand, the frequencies of substitutions were similar in both regions. Other important clues came from the analysis of the frequencies of synonymous and non-sysnonymous substitutions. As expected in the case of negative selection, the frequencies of non-synonymous substitution were lower than that of synonymous substitution, both in repeats and adjacent regions. However the distributions of synonymous and non-synonymous substitutions, inside the repeats and in adjacent regions, were very similar indicating that, at least at the intra-specific level, there are no major differences in the selective forces that act on substitutions in the two regions. Intriguingly, the resulting Ka/Ks ratio was higher than that typically observed in inter-specific comparisons (Ho et al., 2005), a fact that could be explained by the persistence in populations of slightly deleterious non-synonymous mutations that are yet to be lost (Penny, 2005).

Further insight in the understanding of repeat evolution came from the study of how protein selective constraints could affect the formation of new repeats or how it can induce changes in size of existing ones. It has been previously noted that regions adjacent to 28 poly-glutamine tracts in human and mouse proteins tend to show a significantly high non-synonymous substitution rate, particularly when repeats are not conserved between the two species (Hancock et al., 2001). We investigated this point considering the complete coding sequence of a larger dataset of human-mouse orthologous proteins. Our results showed a significant positive correlation between protein evolutionary rate and repeat size difference for each amino acid type considered, generalizing the trend previously found in glutamine repeats. This outcome indicates that the accumulation of new repeat structures is more frequent in proteins that evolve rapidly that in those that evolve slowly, and therefore, in the absence of selection, this should result in a lower number of repeats in slowly evolving proteins. Surprisingly that was not the case, as we found a general negative correlation between the protein content of repeats and the protein evolutionary rate. Therefore proteins that are subject to strong selective constraints and that are evolving at slow pace, tend to contain an unexpectedly large number of repeats, which tend to be well conserved between the two species. These repeats could represent old structures, as slowing evolving proteins are generally unlikely to incorporate new homopolimers. In addition, the results suggest a significant number of repeats are functionally important, even if these functions remain, for the most part, unknown.

With the results described above we detected several repeats that were conserved between human and mouse (or human and chimp) but we did not have any information about when the repeats had originated.. Further insight into this question could be obtained by analyzing the conservation of repeats across vertebrates. The aim was to distinguish recently originated repeats (for instance, those that appeared in primates), some of which could be involved in novel functions, from very old repeats (for instance, originated in an ancestral vertebrate), likely to have been maintained by selection and probably related to conserved repeat-associated functions. Previous studies had pointed out that alanine repeats in transcription factors and developmental proteins could have taxon-specific functions, as they tended to be conserved in mammals but is absent in other vertebrate species. Anan et al. (2007) investigated the conservation of a repetitive alanine tract in Hox-D13 across vertebrate species and found that the repeat was conserved in mammalia, but was truncated in amniota and absent in the other vertebrates. This indicated that the polyA repeat could have originated in chicken and eventually became longer and stable in size in mammalia. Similar results were obtained by Mortlock et al. (2000) on Hox-A13. This protein contains three long poly-alanine runs (of size 14, 18 and 6 residues in human) that are only present in eutherian and marsupial mammals, but not in the other vertebrates. Further evidence of alanine repeat taxon-specific functions in Hox family include Hox-A11 and Hox-D8. In both proteins the alanine homopeptide was present in human and mouse but was absent from chicken and other metazoans (Lavoie et al., 2003). Other developmental proteins seem to follow the same pattern. For example, GATA transcription factors 1, 4 and 6 contain poly-alanine runs that are mammalian specific (Lavoie et al., 2003). Finally, most of the poly-alanine and glycine repeats in class III POU transcription factors Brain-1, Brain-2 and Scip are absent from non-mammalian vertebrates (Nakachi et al., 1997). We studied whether this trend could be generalized using a large dataset of orthologous proteins from 12 vertebrate genomes. However we found that a number of proteins include repetitive structures of different age, without any significant evidence of a taxon-specific distribution of the repeats. In addition, we observed that lineage-specific repeats were distributed along the phylogenetic tree in a manner approximately proportional to the length of the corresponding branch, as measured by the nucleotide substitution rate. As a proxy to the expected distribution under neutral evolution we used similar repeats in non-coding sequences, measuring their level of lineage-specificity in genomic alignments containing all vertebrate groups. We found that non-coding repeats were much less conserved, indicating that the level of conservation of repeats in coding sequences was not due to chance. This strongly suggests that selective forces have acted to maintain a large number of amino acid repeats during vertebrate evolution. Although the function of many of these repeats remains unknown, they are strong candidates for future investigations into the functional impact of these highly prevalent sequences.

6. Future directions of research

The work developed in this thesis has helped us to gain further insight onto the evolutionary dynamics of tandem amino acid repeats. However interesting questions are still unanswered. Our analysis enabled us to date the time tandem amino acid repeats originated. An important outcome was the observation that a large number of human repeats are very ancient and are found in other vertebrate species. However, this study was focused on the conservation of human repeats, and we didn't consider those homopeptides that were present in other species but absent from the human lineage. A study considering all terminal branches of the tree would allow to assess other species-specific gains of repeats as well as to investigate events of repeat loss.

In addition, the above study was centered on the presence or absence of repeats but not on specific changes in the length of repeats. Interestingly, it has been previously suggested that pressure for increased GC content in mammalian genes would have triggered the expansion of alanine, glycine and proline repeats, as they are encoded by GC-rich codons (Sumiyama et al., 1996). This question could be further investigated by taking into account repeat length.

Equivalent repeats were detected by looking for repeats that overlapped in the protein alignments, therefore the accuracy of the alignments acquired a crucial role. While alignment software work appropriately with well-conserved proteins, problems arise with increased sequence divergence and the use of additional sequences, making the repeat detection not trivial. Improvements of sequence alignment methods with regards to repetitive sequences, with a better management of gap extensions, are highly desirable to further improve our understanding of the evolutionary dynamics of tandem amino acid repeats.

Tandem amino acid repeats are a particular class of low complexity, or cryptic, sequences. Low complexity sequences are defined as regions containing a statistically significant overrepresentations of single amino acids or short amino acid motifs (Tautz et al., 1986). These regions exhibit more complex patterns than amino acid tandem repeats, as they can contain amino acid interruptions of the repetitive tract, and the repetitive units may consist of two or more amino acids (for example RS repeats in some splicing factors). Many of these regions are structurally disordered, and this property has been linked to the formation of flexible interaction surfaces in proteins that act as network hubs (Haynes et al., 2006). Indeed, another interesting continuation of the work presented in this thesis would be to analyze the evolutionary dynamics of low complexity sequences in general.

Part V

CONCLUSIONS

Conclusions

The main points of the work presented in this thesis can be summarized as follows:

1. The intra-specific analysis of the tandem amino acid repeats in the human proteome showed increased slippage activity, and a comparable rate of substitution mutations, in the repeat region with respect to repeat surrounding regions.

2. The same study showed that in repeats and adjacent sequences, non-synonymous substitution frequencies are lower than synonymous ones, indicating that selection plays a role in shaping the amino acid content of these regions.

3. Variable tandem amino acid repeats tend to be encoded by more pure codon runs than conserved repeats, indicating that replication slippage is a major force in generating and expanding these structures.

4. Repeat size difference in orthologous human and mouse proteins correlates with the non-synonymous substitution rate of the coding sequence, indicating that proteins subject to low selective constraints are more prone to suffer repeat mutations.

5. Proteins subject to high evolutionary constraints contain an expectedly high number of repeats, which tend to be well conserved between the two species. As these proteins only rarely incorporate new repeat structures, many of their repeats appear to have been maintained by selection.

6. Neurodegenerative disease-associated poly-glutamine repeats tend to show extreme repeat size variability between human and mouse, whereas poly-alanine involved in developmental disorders tend to be highly conserved between the two species. Several other loci have been detected which have similar characteristics to trinucleotide-expansion diseases, and could be potentially pathogenic.

7. A significant fraction of human tandem amino acid repeats are conserved in other vertebrate species, indicating that they may be functionally important.

8. The coexistence of tandem amino acid repeats of different ages in some repeat-prone proteins supports a model of continuous evolutionary innovation linked to the formation of novel repetitive sequences.

Part VI

REFERENCES

References

Alba, M.M. and Guigo, R.: Comparative analysis of amino acid repeats in rodents and humans. Genome Res 14 (2004) 549-54.

Alba, M.M., Santibanez-Koref, M.F. and Hancock, J.M.: Conservation of polyglutamine tract size between mice and humans depends on codon interruption. Mol Biol Evol 16 (1999) 1641-4.

Alba, M.M., Santibanez-Koref, M.F. and Hancock, J.M.: The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and Drosophila. J Mol Evol 52 (2001) 249-59.

Alba , M.M., Veitia, R.A. and Tompa, P.: Amino Acid Repeats and the Structure and Evolution of Proteins. Genome Dyn. 3 (2007) 119-130.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P.: Molecular Biology of the Cell, Fourth ed. Garland Science, 2002.

Anan, K., Yoshida, N., Kataoka, Y., Sato, M., Ichise, H., Nasu, M. and Ueda, S.: Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. Mol Biol Evol 24 (2007) 281-7.

Andersen, T.H. and Nilsson-Tillgren, T.: A fungal minisatellite. Nature 386 (1997) 771.

Andres, A.M., Lao, O., Soldevila, M., Calafell, F. and Bertranpetit, J.: Dynamics of CAG repeat loci revealed by the analysis of their variability. Hum Mutat 21 (2003) 61-70.

Andres, A.M., Soldevila, M., Lao, O., Volpini, V., Saitou, N., Jacobs, H.T., Hayasaka, I., Calafell, F. and Bertranpetit, J.: Comparative genetics of functional trinucleotide tandem repeats in humans and apes. J Mol Evol 59 (2004) 329-39.

Armour, J.A. and Jeffreys, A.J.: Biology and applications of human minisatellite loci. Curr Opin Genet Dev 2 (1992) 850-6.

Arney, K.L. and Fisher, A.G.: Epigenetic aspects of differentiation. J Cell Sci 117 (2004) 4355-63.

Asghari, V., Schoots, O., van Kats, S., Ohara, K., Jovanovic, V., Guan, H.C., Bunzow, J.R., Petronis, A. and Van Tol, H.H.: Dopamine D4 receptor repeat: analysis of different native and mutant forms of the human and rat genes. Mol Pharmacol 46 (1994) 364-73.

Benjamin, J., Li, L., Patterson, C., Greenberg, B.D., Murphy, D.L. and Hamer, D.H.: Population and familial association between the D4 dopamine receptor gene and measures of Novelty Seeking. Nat Genet 12 (1996) 81-4.

Bennett, S.T., Lucassen, A.M., Gough, S.C., Powell, E.E., Undlien, D.E., Pritchard, L.E., Merriman, M.E., Kawaguchi, Y., Dronsfield, M.J., Pociot, F. and et al.: Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9 (1995) 284-92.

Berg, E.S. and Olaisen, B.: Characterization of the COL2A1 VNTR polymorphism. Genomics 16 (1993) 350-4.

Bloom, K.: Centromere dynamics. Curr Opin Genet Dev 17 (2007) 151-6.

Boguski, M.S., Birkenmeier, E.H., Elshourbagy, N.A., Taylor, J.M. and Gordon, J.I.: Evolution of the apolipoproteins. Structure of the rat apo-A-IV gene and its relationship to the human genes for apo-A-I, C-III, and E. J Biol Chem 261 (1986) 6398-407.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R. and Cavalli-Sforza, L.L.: High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368 (1994) 455-7.

Brown, L., Paraso, M., Arkell, R. and Brown, S.: In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. Hum Mol Genet 14 (2005) 411-20.

Brown, L.Y. and Brown, S.A.: Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet 20 (2004) 51-8.

Buchanan, G., Yang, M., Cheong, A., Harris, J.M., Irvine, R.A., Lambert, P.F., Moore, N.L., Raynor, M., Neufing, P.J., Coetzee, G.A. and Tilley, W.D.:

Structural and functional consequences of glutamine tract variation in the androgen receptor. Hum Mol Genet 13 (2004) 1677-92.

Buitkamp, J., Ammer, H. and Geldermann, H.: DNA fingerprinting in domestic animals. Electrophoresis 12 (1991) 169-74.

Buresi, C., Desmarais, E., Vigneron, S., Lamarti, H., Smaoui, N., Cambien, F. and Roizes, G.: Structural analysis of the minisatellite present at the 3' end of the human apolipoprotein B gene: new definition of the alleles and evolutionary implications. Hum Mol Genet 5 (1996) 61-8.

Burke, T. and Bruford, M.W.: DNA fingerprinting in birds. Nature 327 (1987) 149-52.

Buschiazzo, E. and Gemmell, N.J.: The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28 (2006) 1040-50.

Ciechanover, A. and Brundin, P.: The ubiquitin proteasome system in neurodegenerative diseases: sometimes the chicken, sometimes the egg. Neuron 40 (2003) 427-46.

Clark, R.M., Bhaskar, S.S., Miyahara, M., Dalgliesh, G.L. and Bidichandani, S.I.: Expansion of GAA trinucleotide repeats in mammals. Genomics 87 (2006) 57-67.

Cocquet, J., De Baere, E., Caburet, S. and Veitia, R.A.: Compositional biases and polyalanine runs in humans. Genetics 165 (2003) 1613-7.

Cordeiro, G.M., Casu, R., McIntyre, C.L., Manners, J.M. and Henry, R.J.: Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160 (2001) 1115-1123.

Csink, A.K. and Henikoff, S.: Something from nothing: the evolution and utility of satellite repeats. Trends Genet 14 (1998) 200-4.

Dallas, J.F.: Detection of DNA "fingerprints" of cultivated rice by hybridization with a human minisatellite DNA probe. Proc Natl Acad Sci U S A 85 (1988) 6831-5.

de Lange, T.: Shelterin: the protein complex that shapes and safeguards human telomeres. Genes Dev 19 (2005) 2100-10.

Desmarais, E., Vigneron, S., Buresi, C., Cambien, F., Cambou, J.P. and Roizes, G.: Variant mapping of the Apo(B) AT rich minisatellite. Dependence on nucleotide sequence of the copy number variations.

Instability of the non-canonical alleles. Nucleic Acids Res 21 (1993) 2179-84.

Dunah, A.W., Jeong, H., Griffin, A., Kim, Y.M., Standaert, D.G., Hersch, S.M., Mouradian, M.M., Young, A.B., Tanese, N. and Krainc, D.: Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. Science 296 (2002) 2238-43.

Ebstein, R.P., Nemanov, L., Klotz, I., Gritsenko, I. and Belmaker, R.H.: Additional evidence for an association between the dopamine D4 receptor (D4DR) exon III repeat polymorphism and the human personality trait of Novelty Seeking. Mol Psychiatry 2 (1997) 472-7.

Ebstein, R.P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Bennett, E.R., Nemanov, L., Katz, M. and Belmaker, R.H.: Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of Novelty Seeking. Nat Genet 12 (1996) 78-80.

Eckert, R.L. and Green, H.: Structure and evolution of the human involucrin gene. Cell 46 (1986) 583-9.

Edwards, Y.J., Elgar, G., Clark, M.S. and Bishop, M.J.: The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, Fugu rubripes: perspectives in functional and comparative genomic analyses. J Mol Biol 278 (1998) 843-54.

Ellegren, H.: Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet 24 (2000a) 400-2.

Ellegren, H.: Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 16 (2000b) 551-8.

Ellegren, H.: Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5 (2004) 435-45.

Ellsworth, D.L., Shriver, M.D. and Boerwinkle, E.: Nucleotide sequence analysis of the apolipoprotein B 3' VNTR. Hum Mol Genet 4 (1995) 937-44.

Faux, N.G., Bottomley, S.P., Lesk, A.M., Irving, J.A., Morrison, J.R., de la Banda, M.G. and Whisstock, J.C.: Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 15 (2005) 537-51.

Field, D. and Wills, C.: Long, polymorphic microsatellites in simple organisms. Proc Biol Sci 263 (1996) 209-15.

Fitch, W.M.: Distinguishing homologous from analogous proteins. Syst Zool 19 (1970) 99-113.

Fondon, J.W., 3rd and Garner, H.R.: Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 101 (2004) 18058-63.

Freudenreich, C.H., Stavenhagen, J.B. and Zakian, V.A.: Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol Cell Biol 17 (1997) 2090-8.

Galant, R. and Carroll, S.B.: Evolution of a transcriptional repression domain in an insect Hox protein. Nature 415 (2002) 910-3.

Gatchel, J.R. and Zoghbi, H.Y.: Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet 6 (2005) 743-55.

Gerber, H.P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S. and Schaffner, W.: Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263 (1994) 808-11.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R.A., Adams, M.D., Amanatides, P.G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C.A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C.L., Nguyen, T., Pfannkoch, C.M., Sitter, C., Sutton, G.G., Venter, J.C., Woodage, T., Smith, D., Lee, H.M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R.B., Dunn, D.M., Green, E.D., Blakesley, R.W., Bouffard, G.G., De Jong, P.J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C.M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W.C., Havlak, P.H., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K.C., Cooney, A.J., D'Souza, L.M., Martin, K., Wu, J.Q., Gonzalez-Garay, M.L., Jackson, A.R., Kalafus, K.J., McLeod, M.P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D.A., Zhang, Z., Bailey, J.A., Eichler, E.E., et al.: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428 (2004) 493-521.

Gibbs, S., Fijneman, R., Wiegant, J., van Kessel, A.G., van De Putte, P. and Backendorf, C.: Molecular characterization and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. Genomics 16 (1993) 630-7.

Gilbert, D.A., Lehman, N., O'Brien, S.J. and Wayne, R.K.: Genetic fingerprinting reflects population differentiation in the California Channel Island fox. Nature 344 (1990) 764-7.

Gill, P., Ivanov, P.L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E. and Sullivan, K.: Identification of the remains of the Romanov family by DNA analysis. Nat Genet 6 (1994) 130-5.

Giraud, T., Fortini, D., Levis, C. and Brygoo, Y.: The minisatellite MSB1, in the fungus Botrytis cinerea, probably mutates by slippage. Mol Biol Evol 15 (1998) 1524-31.

Green, H. and Wang, N.: Codon reiteration and the evolution of proteins. Proc Natl Acad Sci U S A 91 (1994) 4298-302.

Gregory, T.R.: Synergy between sequence and size in large-scale genomics. Nat Rev Genet 6 (2005) 699-708.

Haber, J.E. and Louis, E.J.: Minisatellite origins in yeast and humans. Genomics 48 (1998) 132-5.

Hall, A.E., Keith, K.C., Hall, S.E., Copenhaver, G.P. and Preuss, D.: The rapidly evolving field of plant centromeres. Curr Opin Plant Biol 7 (2004) 108-14.

Hammock, E.A. and Young, L.J.: Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308 (2005) 1630-4.

Hancock, J.M. and Simon, M.: Simple sequence repeats in proteins and their significance for network evolution. Gene 345 (2005) 113-8.

Hancock, J.M., Worthey, E.A. and Santibanez-Koref, M.F.: A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. Mol Biol Evol 18 (2001) 1014-23.

Harris, R.A., Rogers, J. and Milosavljevic, A.: Human-specific changes of genome structure detected by genomic triangulation. Science 316 (2007) 235-7.

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M. and Iakoucheva, L.M.: Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2 (2006) e100.

Henikoff, S., Ahmad, K. and Malik, H.S.: The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293 (2001) 1098-102.

Hewett, D.R., Handt, O., Hobson, L., Mangelsdorf, M., Eyre, H.J., Baker, E., Sutherland, G.R., Schuffenhauer, S., Mao, J.I. and Richards, R.I.: FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. Mol Cell 1 (1998) 773-81.

Ho, S.Y., Phillips, M.J., Cooper, A. and Drummond, A.J.: Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22 (2005) 1561-8.

Huang, H., Winter, E.E., Wang, H., Weinstock, K.G., Xing, H., Goodstadt, L., Stenson, P.D., Cooper, D.N., Smith, D., Alba, M.M., Ponting, C.P. and Fechtel, K.: Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol 5 (2004) R47.

Jeffreys, A.J. and Morton, D.B.: DNA fingerprints of dogs and cats. Anim Genet 18 (1987) 1-15.

Jeffreys, A.J., Wilson, V., Kelly, R., Taylor, B.A. and Bulfield, G.: Mouse DNA 'fingerprints': analysis of chromosome localization and germ-line stability of hypervariable loci in recombinant inbred strains. Nucleic Acids Res 15 (1987a) 2823-36.

Jeffreys, A.J., Wilson, V. and Thein, S.L.: Hypervariable 'minisatellite' regions in human DNA. Nature 314 (1985) 67-73.

Jeffreys, A.J., Wilson, V., Wong, Z., Royle, N., Patel, I., Kelly, R. and Clarkson, R.: Highly variable minisatellites and DNA fingerprints. Biochem Soc Symp 53 (1987b) 165-80.

Jobling, M.A., Bouzekri, N. and Taylor, P.G.: Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet 7 (1998) 643-53.

John, B. and Miklos, G.L.: Functional aspects of satellite DNA and heterochromatin. Int Rev Cytol 58 (1979) 1-114.

Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. and Gentles, A.J.: Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A 99 (2002) 333-8.

Karlin, S. and Burge, C.: Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci U S A 93 (1996) 1560-5.

Kashi, Y. and King, D.G.: Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22 (2006) 253-9.

Kay, B.K., Williamson, M.P. and Sudol, M.: The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. Faseb J 14 (2000) 231-41.

Kazazian, H.H., Jr.: Mobile elements: drivers of genome evolution. Science 303 (2004) 1626-32.

Kemp, D.J., Coppel, R.L. and Anders, R.F.: Repetitive proteins and genes of malaria. Annu Rev Microbiol 41 (1987) 181-208.

Kit, S.: Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. J Mol Biol 3 (1961) 711-6.

Koonin, E.V.: Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39 (2005) 309-38.

Koreth, J., O'Leary, J.J. and J, O.D.M.: Microsatellites and PCR genomic analysis. J Pathol 178 (1996) 239-48.

Kruglyak, S., Durrett, R.T., Schug, M.D. and Aquadro, C.F.: Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 95 (1998) 10774-8.

Kunst, C.B., Leeflang, E.P., Iber, J.C., Arnheim, N. and Warren, S.T.: The effect of FMR1 CGG repeat interruptions on mutation frequency as measured by sperm typing. J Med Genet 34 (1997) 627-31.

Lancaster, C.A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J. and Gendler, S.J.: Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. Biochem Biophys Res Commun 173 (1990) 1019-29.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al.: Initial sequencing and analysis of the human genome. Nature 409 (2001) 860-921.

Lanz, R.B., Wieland, S., Hug, M. and Rusconi, S.: A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. Nucleic Acids Res 23 (1995) 138-45.

Lavoie, H., Debeane, F., Trinh, Q.D., Turcotte, J.F., Corbeil-Girard, L.P., Dicaire, M.J., Saint-Denis, A., Page, M., Rouleau, G.A. and Brais, B.: Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. Hum Mol Genet 12 (2003) 2967-79.

Levinson, G. and Gutman, G.A.: Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4 (1987) 203-21.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A. and Nevo, E.: Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11 (2002) 2453-65.

Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E.: Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21 (2004) 991-1007.

Lichter, J.B., Barr, C.L., Kennedy, J.L., Van Tol, H.H., Kidd, K.K. and Livak, K.J.: A hypervariable segment in the human dopamine receptor D4 (DRD4) gene. Hum Mol Genet 2 (1993) 767-73.

Lodish, H., Berk, A., Zipursky, L.S., Matsudaira, P., Baltimore, D. and Darnell, J.: Molecular Cell Biology, Fourth ed. W. H. Freeman and Company, 2000.

Lynch, M.: Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60 (2006) 327-49.

Mahley, R.W., Innerarity, T.L., Rall, S.C., Jr. and Weisgraber, K.H.: Plasma lipoproteins: apolipoprotein structure and function. J Lipid Res 25 (1984) 1277-94.

Mar Alba, M., Santibanez-Koref, M.F. and Hancock, J.M.: Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol 49 (1999) 789-97.

Messier, W., Li, S.H. and Stewart, C.B.: The birth of microsatellites. Nature 381 (1996) 483.

Metzgar, D., Bytof, J. and Wills, C.: Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10 (2000) 72-80.

Morgante, M., Hanafey, M. and Powell, W.: Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30 (2002) 194-200.

Mortlock, D.P., Sateesh, P. and Innis, J.W.: Evolution of N-terminal sequences of the vertebrate HOXA13 protein. Mamm Genome 11 (2000) 151-8.

Morton, D.J., Whitby, P.W., Jin, H., Ren, Z. and Stull, T.L.: Effect of multiple mutations in the hemoglobin- and hemoglobin-haptoglobin-binding proteins, HgpA, HgpB, and HgpC, of Haemophilus influenzae type b. Infect Immun 67 (1999) 2729-39.

Mrazek, J., Guo, X. and Shah, A.: Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A 104 (2007) 8472-7.

Mularoni, L., Guigo, R. and Alba, M.M.: Mutation patterns of amino acid tandem repeats in the human proteome. Genome Biol 7 (2006) R33.

Mularoni, L., Toll-Riera, M. and Albà, M.M.: Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genomes. Encyclopedia of Life Sciences (ELS). John Wiley&Sons, Ltd: Chichester. (2008).

Mularoni, L., Veitia, R.A. and Alba, M.M.: Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics 89 (2007) 316-25.

Mundlos, S., Otto, F., Mundlos, C., Mulliken, J.B., Aylsworth, A.S., Albright, S., Lindhout, D., Cole, W.G., Henn, W., Knoll, J.H., Owen, M.J., Mertelsmann, R., Zabel, B.U. and Olsen, B.R.: Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell 89 (1997) 773-9.

Muragaki, Y., Mundlos, S., Upton, J. and Olsen, B.R.: Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. Science 272 (1996) 548-51.

Nakachi, Y., Hayakawa, T., Oota, H., Sumiyama, K., Wang, L. and Ueda, S.: Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol Biol Evol 14 (1997) 1042-9.

Newfeld, S.J., Tachida, H. and Yedvobnick, B.: Drive-selection equilibrium: homopolymer evolution in the Drosophila gene mastermind. J Mol Evol 38 (1994) 637-41.

Ohno, S.: So much "junk" DNA in our genome. Brookhaven Symp Biol 23 (1972) 366-70.

Ohno, S.: Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. J Mol Evol 20 (1984) 313-21.

Ohno, S.: Early genes that were oligomeric repeats generated a number of divergent domains on their own. Proc Natl Acad Sci U S A 84 (1987a) 6486-90.

Ohno, S.: Evolution from primordial oligomeric repeats to modern coding sequences. J Mol Evol 25 (1987b) 325-9.

Orgel, L.E. and Crick, F.H.: Selfish DNA: the ultimate parasite. Nature 284 (1980) 604-7.

Owen, R.: On the Archetype and Homologies of the Vertebrate Skeleton. London: Murray (1848).

Pearson, C.E., Nichol Edamura, K. and Cleary, J.D.: Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6 (2005) 729-42.

Penny, D.: Evolutionary biology: relativity for molecular clocks. Nature 436 (2005) 183-4.

Pizzi, E. and Frontali, C.: Low-complexity regions in Plasmodium falciparum proteins. Genome Res 11 (2001) 218-29.

Reeve, H.K., Westneat, D.F., Noon, W.A., Sherman, P.W. and Aquadro, C.F.: DNA "fingerprinting" reveals high levels of inbreeding in colonies of the eusocial naked mole-rat. Proc Natl Acad Sci U S A 87 (1990) 2496-500.

Ren, Z., Jin, H., Whitby, P.W., Morton, D.J. and Stull, T.L.: Role of CCAA nucleotide repeats in regulation of hemoglobin and hemoglobin-haptoglobin binding protein genes of Haemophilus influenzae. J Bacteriol 181 (1999) 5865-70.

Richard, G.F. and Dujon, B.: Trinucleotide repeats in yeast. Res Microbiol 148 (1997) 731-44.

Richard, G.F. and Paques, F.: Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1 (2000) 122-6.

Rose, O. and Falush, D.: A threshold size for microsatellite expansion. Mol Biol Evol 15 (1998) 613-5.

Santibanez-Koref, M.F., Gangeswaran, R. and Hancock, J.M.: A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. Mol Biol Evol 18 (2001) 2119-23.

Schlotterer, C.: Genome evolution: are microsatellites really simple sequences? Curr Biol 8 (1998) R132-4.

Schlotterer, C. and Tautz, D.: Slippage synthesis of simple sequence DNA. Nucleic Acids Res 20 (1992) 211-5.

Shimohata, T., Nakajima, T., Yamada, M., Uchida, C., Onodera, O., Naruse, S., Kimura, T., Koide, R., Nozaki, K., Sano, Y., Ishiguro, H., Sakoe, K., Ooshima, T., Sato, A., Ikeuchi, T., Oyake, M., Sato, T., Aoyagi, Y., Hozumi, I., Nagatsu, T., Takiyama, Y., Nishizawa, M., Goto, J., Kanazawa, I., Davidson, I., Tanese, N., Takahashi, H. and Tsuji, S.: Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. Nat Genet 26 (2000) 29-36.

Sinden, R.R., Potaman, V.N., Oussatcheva, E.A., Pearson, C.E., Lyubchenko, Y.L. and Shlyakhtenko, L.S.: Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. J Biosci 27 (2002) 53-65.

Strand, M., Prolla, T.A., Liskay, R.M. and Petes, T.D.: Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365 (1993) 274-6.

Subramanian, S., Mishra, R.K. and Singh, L.: Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol 4 (2003) R13.

Sullivan, B.A., Blower, M.D. and Karpen, G.H.: Determining centromere identity: cyclical stories and forking paths. Nat Rev Genet 2 (2001) 584-96.

Sumiyama, K., Washio-Watanabe, K., Saitou, N., Hayakawa, T. and Ueda, S.: Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. J Mol Evol 43 (1996) 170-8.

Tautz, D.: Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17 (1989) 6463-71.

Tautz, D.: Notes on the definition and nomenclature of tandemly repetitive DNA sequences. Exs 67 (1993) 21-8.

Tautz, D. and Renz, M.: Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12 (1984) 4127-38.

Tautz, D., Trick, M. and Dover, G.A.: Cryptic simplicity in DNA is a major source of genetic variation. Nature 322 (1986) 652-6.

Taylor, J.S., Durkin, J.M. and Breden, F.: The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Mol Biol Evol 16 (1999) 567-72.

Toth, G., Gaspari, Z. and Jurka, J.: Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10 (2000) 967-81.

Van Tol, H.H., Wu, C.M., Guan, H.C., Ohara, K., Bunzow, J.R., Civelli, O., Kennedy, J., Seeman, P., Niznik, H.B. and Jovanovic, V.: Multiple dopamine D4 receptor variants in the human population. Nature 358 (1992) 149-52.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al.: The sequence of the human genome. Science 291 (2001) 1304-51.

Vergnaud, G., Gauguier, D., Schott, J.J., Lepetit, D., Lauthier, V., Mariat, D. and Buard, J.: Detection, cloning, and distribution of minisatellites in some mammalian genomes. Exs 67 (1993) 47-57.

Warren, S.T.: Polyalanine expansion in synpolydactyly might result from unequal crossing-over of HOXD13. Science 275 (1997) 408-9.

Watase, K., Weeber, E.J., Xu, B., Antalffy, B., Yuva-Paylor, L., Hashimoto, K., Kano, M., Atkinson, R., Sun, Y., Armstrong, D.L., Sweatt, J.D., Orr, H.T., Paylor, R. and Zoghbi, H.Y.: A long CAG repeat in the mouse Sca1 locus replicates SCA1 features and reveals the impact of protein solubility on selective neurodegeneration. Neuron 34 (2002) 905-19.

Wilder, J. and Hollocher, H.: Mobile elements and the genesis of microsatellites in dipterans. Mol Biol Evol 18 (2001) 384-92.

Wren, J.D., Forgacs, E., Fondon, J.W., 3rd, Pertsemlidis, A., Cheng, S.Y., Gallardo, T., Williams, R.S., Shohet, R.V., Minna, J.D. and Garner, H.R.: Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet 67 (2000) 345-56.

Wyman, A.R. and White, R.: A highly polymorphic locus in human DNA. Proc Natl Acad Sci U S A 77 (1980) 6754-8.

Yoo, S.Y., Pennesi, M.E., Weeber, E.J., Xu, B., Atkinson, R., Chen, S., Armstrong, D.L., Wu, S.M., Sweatt, J.D. and Zoghbi, H.Y.: SCA7 knockin mice model human SCA7 and reveal gradual accumulation of mutant ataxin-7 in neurons and abnormalities in short-term plasticity. Neuron 37 (2003) 383-401.

Young, E.T., Sloan, J.S. and Van Riper, K.: Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. Genetics 154 (2000) 1053-68.

Yu, S., Mangelsdorf, M., Hewett, D., Hobson, L., Baker, E., Eyre, H.J., Lapsys, N., Le Paslier, D., Doggett, N.A., Sutherland, G.R. and Richards, R.I.: Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. Cell 88 (1997) 367-74.

Part VII

APPENDICES

Appendix 1

Summary

This appendix reports a study over the level of promoter conservation in orthologous genes. Gene transcription regulatory sequences contain a complex arrangement of motifs that are recognized by transcription factors, many of which are located upstream from transcription start sites (promoters). In this study has been shown that mammalian housekeeping genes show less conserved promoters than genes with tissue-specific expression, particularly upstream of position -500 with respect to the transcription start site. This suggests that genes with constitutive expression require shorter functional promoters.

The results of this work has been published in the following research article:

Farre, D., Bellora, N., Mularoni, L., Messeguer, X. and Alba, M.M.: **Housekeeping genes tend to show reduced upstream sequence conservation**. Genome Biol 8 (2007) R140.

Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM.
Housekeeping genes tend to show reduced upstream sequence conservation.
Genome Biol. 2007;8(7):R140.

Appendix 2

List of Publications

Loris Mularoni, Macarena Toll-Riera, M.Mar Albà. Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genomes. Encyclopedia of Life Sciences (ELS). John Wiley&Sons, Ltd: Chichester. (2008).

Domènec Farré, Nicolás Bellora, **Loris Mularoni**, Xavier Messeguer, M. Mar Albà. Housekeeping genes contain shorter promoters. Genome Biology (2007), 13;8(7):R140.

Loris Mularoni, Reiner A. Veitia, M. Mar Albà. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics (2007), 89:316-325.

Loris Mularoni, Roderic Guigó, M. Mar Albà. Mutation patterns of amino acid tandem repeats in the human proteome. Genome Biology (2006), 7:R33.

Appendix 3

Conferences Contributions

2008: Society for Molecular Biology and Evolution Meeting (SMBE), Barcelona (Spain)

Patterns of conservation of human amino acid tandem repeats in vertebrate proteins.

Authors: Loris Mularoni, Alice Ledda, M. Mar Albà

2007: The Biology of Genomes, Cold Spring Harbor Laboratory (USA)

Genome-wide analysis of evolutionary patterns of amino acid tandem repeats in mammalian proteins.

Authors: Loris Mularoni, M. Mar Albà

2005: 6th Anton Dohrn Workshop "Evolutionary Genomics", Ischia (Italy)

Evolutionary patterns of amino acid tandem repeats in mammalian proteins.

Authors: Loris Mularoni, Reiner A. Veitia, Roderic Guigó, M. Mar Albà

2005: Computacional Biology 2005 (ECCB 2005), Madrid (Spain)

Evolutionary constraints operating in sequences coding for tandem amino acid repeats.

Authors: Loris Mularoni, M. Mar Albà

LLETRA A DOLORS

Em costa imaginar-te absent per sempre.
Tants de records de tu se m'acumulen
que ni deixen espai a la tristesa
i et visc intensament sense tenir-te.
No vull parlar-te amb veu melangiosa,
la teva mort no em crema les entranyes,
ni m'angoixa, ni em lleva el goig de viure;
em dol saber que no podrem partir-nos
mai més el pa, ni fer-nos companyia;
però d'aquest dolor en trec la força
per escriure aquests mots i recordar-te.
Més tenaçment que mai, m'esforço a créixer
sabent que tu creixes amb mi: projectes,
il.lusions, desigs, prenen volada
per tu i amb tu, per molt distants que et siguin,
i amb tu i per tu somnio d'acomplir-los.
Te'm fas present en les petites coses
i és en elles que et penso i que t'evoco,
segur com mai que l'única esperança
de sobreviure és estimar amb prou força
per convertir tot el que fem en vida
i acréixer l'esperança i la bellesa.

Tu ja no hi ets i floriran les roses,
maduraran els blats i el vent tal volta
desvetllarà secretes melodies;
tu ja no hi ets i el temps ara em transcorre
entre el record de tu, que m'acompanyes,
i aquell esforç, que prou que coneixes,

de persistir quan res no ens és propici.
Des d'aquests mots molt tendrament et penso
mentre la tarda suaument declina.
Tots els colors proclamen vida nova
i jo la visc, i en tu se'm representa
sorprenentment vibrant i harmoniosa.
No tornaràs mai més, però perdures
en les coses i en mi de tal manera
que em costa imaginar-se absent per sempre.

MIQUEL MARTÍ I POL (Roda de Ter, Barcelona 1929 - Vic, Barcelona 2003)

COMPARATIVE GENOMICS OF AMINO ACID TANDEM REPEATS

Loris Mularoni

Universitat Pompeu Fabra