Megasatellite formation and evolution in vertebrate genes

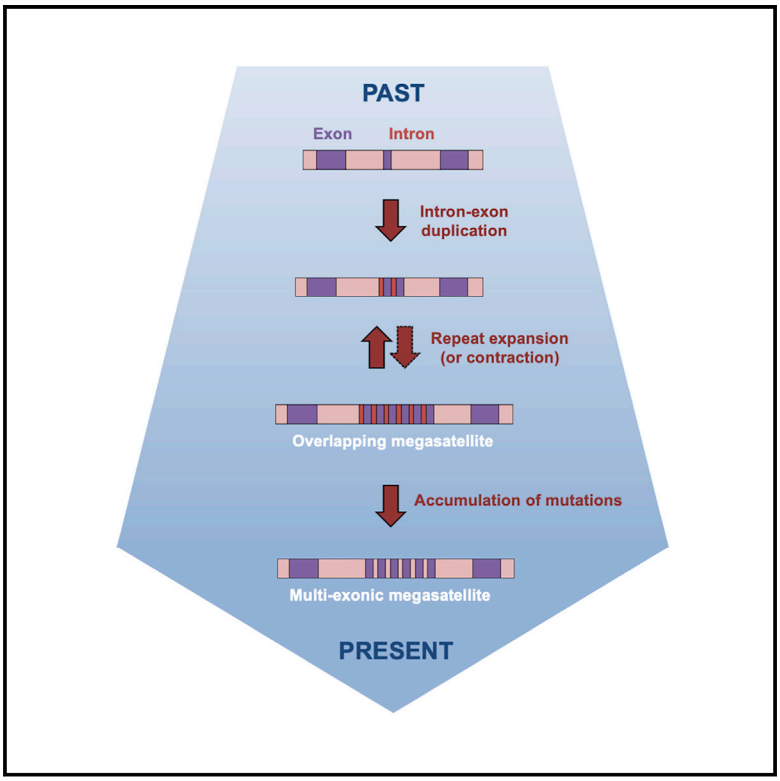
Stéphane Descorps-Declère, Guy-Franck Richard

Cell Reports

Article

Megasatellite formation and evolution in vertebrate genes

Graphical abstract



Highlights

- Megasatellite analysis in vertebrates shows two bursts of formation during evolution
- Megasatellites frequently encode transcription regulators and cell wall proteins
- Megasatellites frequently overlap exon-intron junctions and are erased with time in introns

Authors

Stéphane Descorps-Declère, Guy-Franck Richard

Correspondence

stephane.descorps-declere@pasteur.fr (S.D.-D.), gfrichar@pasteur.fr (G.-F.R.)

In brief

Descorps-Declère and Richard perform a genome-wide analysis of megasatellites in 58 vertebrate genomes. They find that they are enriched in subtelomeric regions and frequently encode proteins involved in cell wall homeostasis. Accumulation of mutations within intronic regions partially erases old megasatellites that can only be detected in exons.

Descorps-Declère & Richard, 2022, Cell Reports *40*, 111347
September 13, 2022
https://doi.org/10.1016/j.celrep.2022.111347

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Megasatellite formation and evolution in vertebrate genes

Stéphane Descorps-Declère[1,*] and Guy-Franck Richard[2,3,*]
[1]Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, 25 rue du Dr Roux, 75015 Paris, France
[2]Institut Pasteur, Université Paris Cité, CNRS UMR3525, Natural & Synthetic Genome Instabilities, 25 rue du Dr Roux, 75015 Paris, France
[3]Lead contact
*Correspondence: stephane.descorps-declere@pasteur.fr (S.D.-D.), gfrichar@pasteur.fr (G.-F.R.)
https://doi.org/10.1016/j.celrep.2022.111347

SUMMARY

Since formation of the first proto-eukaryotes, gene repertoire and genome complexity have significantly increased. Among genetic elements responsible for this increase are tandem repeats. Here we describe a genome-wide analysis of large tandem repeats, called megasatellites, in 58 vertebrate genomes. Two bursts occurred, one after the radiation between Agnatha and Gnathostomata fishes and the second one in therian mammals. Megasatellites are enriched in subtelomeric regions and frequently encoded in genes involved in transcription regulation, intracellular trafficking, and cell membrane metabolism, reminiscent of what is observed in fungus genomes. The presence of many introns within young megasatellites suggests that an exon-intron DNA segment is first duplicated and amplified before accumulation of mutations in intronic parts partially erases the megasatellite in such a way that it becomes detectable only in exons. Our results suggest that megasatellite formation and evolution is a dynamic and still ongoing process in vertebrate genomes.

INTRODUCTION

Eukaryotic genomes are characterized by an increase in complexity, often associated with a remarkable expansion of tandem repeat sequences, compared with prokaryotes. Tandem repeats include segmental duplications, Copy Number Variations (CNV), satellite DNA, microsatellites, minisatellites, and megasatellites (Richard et al., 2008). Segmental duplications have been studied in yeast (Koszul et al., 2004), mice (Bailey et al., 2004), brown rats (Tuzun et al., 2004) and humans (Bailey et al., 2002). Further analyses of other genome sequences showed that segmental duplications were more frequent in the great ape ancestor of the human lineage than in other primates (Marques-Bonet et al., 2009), with subtelomeric regions being hotspots of such polymorphisms (Linardopoulou et al., 2005). CNVs generally include segmental duplications but also encompass other structural variants. Their diversity and evolution have been studied in humans and great apes (Sudmant et al., 2013). It has been shown that 60% of nucleotides in human segmental duplications are CNVs (Zarrei et al., 2015). Microsatellites are short sequence repeats (SSR), whose base motif is less than 10 bp long. They are very frequent in all eukaryotic genomes and have been extensively studied in several completely sequenced organisms (Bachtrog et al., 1999; Dieringer and Schlötterer 2003; Hennequin et al., 2001; Innan et al., 1997; Malpertuy et al., 2003; Richard et al., 1999; Röder et al., 1998; Sakamoto et al., 2000; Dib et al., 1996). Minisatellites (or VNTRs [variable numbers of tandem repeats]) are tandem repeats whose base motif is at least 10 bp long. The distribution and length variability of minisatellites has been examined in *Saccharomyces cerevisiae* (Verstrepen et al., 2005; Richard and Dujon 2006; Bowen et al., 2005), *Tetraodon nigroviridis* (Roest Crollius et al., 2000), *Arabidopsis thaliana*, and *Caenorhabditis elegans* (Vergnaud and Denoeud 2008). The human genome contains roughly 12,000 minisatellites, with some of them exhibiting length polymorphism compared with their orangutan or chimpanzee orthologs, as expected (Sulovari et al., 2019). Longer tandem repeats are sometimes called megasatellites (Thierry et al., 2009) and were initially defined as direct and contiguous repeats of DNA sequences of three motifs or more, each of an individual length of at least 90 bp (Tekaia et al., 2013). They are widespread in fungus genomes but are particularly abundant in *Candida glabrata*, an opportunistic pathogenic yeast (Thierry et al., 2008). Their base motif length is always a multiple of three, and they are always found in frame within open reading frames. They are called megasatellites to distinguish them from minisatellites, made of smaller motifs and found mainly within intergenic regions. These properties distinguish them from other tandem repeats, such as segmental duplications, CNVs, or micro/minisatellites.

Vertebrates are much younger eukaryotes than fungi. They emerged as a monophyletic group from the Chordata phylum 550 million years ago during the pre-Cambrian explosion (Figure S1; Bromham et al., 1998; Erwin et al., 2011). They inhabit almost all ecological niches and are arguably the most successful of the chordates, with more than 66,000 species described (Genome 10K Community of Scientists 2009). We set out to

establish the first set of vertebrate megasatellites, based on the complete sequence of 61 vertebrate genomes, in addition to the yeast *S. cerevisiae*, whose mini- and megasatellite content had been determined previously and could serve as a positive control (Richard and Dujon 2006).

We detected more than 14,000 megasatellites unevenly distributed among the 12 clades studied. Two increases of formation were identified, the first one in the Gnathostomata, after the Agnatha radiation (jawless fish), and the second one in the therians (mammals with a uterus). Three-quarters of these megasatellites encode zinc-finger proteins, but other cellular features are highly represented, such as a cell membrane, intracellular trafficking, and RNA metabolism. Although most megasatellites are encoded in the exonic parts of genes, a significant fraction of them overlap exon-intron junctions in primates, suggesting recent formation.

RESULTS

Initially, a megasatellite was defined as a direct and contiguous DNA repeat of three motifs or more, each of an individual length of at least 90 bases (Tekaia et al., 2013). We initially planned to search whole-genome sequences with two programs: Tandem Repeat Finder (TRF) (Benson 1999) and MREPS (Kolpakov et al., 2003). Both were run on 61 vertebrate species genomes and on the well-described *S. cerevisiae* genome. These 61 species were chosen because, when this work started, their whole-genome sequences were available and annotated in the Ensembl database (release 101) (Yates et al., 2020). They are detailed in Table S1. *Danio rerio* (zebrafish), *Macropus eugenii* (wallaby), and *Tarsius syrichta* (tarsier) genomes contained an abnormally high number of tandem repeats compared with all other vertebrate genomes, suggesting that the sequence quality and/or its assembly were not acceptable, and they were excluded from subsequent studies.

Careful examination of tandem repeats detected in the 58 remaining genomes demonstrated the presence of many false negative and false positive results. The latter were typically tandem repeats of transposons, especially SINEs (*Alu* elements) in primate genomes. The former corresponded to genes known previously to contain large tandem repeats that were absent from our results. Analysis of these false negatives revealed that the presence of large introns precluded correct identification of exon-encoded megasatellites. This led to the conclusion that looking for megasatellites directly in DNA sequences was not possible to reliably achieve on large and complex vertebrate genomes with the available tools. We therefore switched to an alternative strategy.

Protein sequences were extracted from the 58 vertebrate and *S. cerevisiae* proteomes. The T-REKS software (Jorda and Kajava 2009) was run on these sequences, and the pipeline described in Figure 1 was followed (see STAR Methods for parameters). T-REKS identified 1,542,346 protein TRs (From now on, TR will exclusively refer to a protein tandem repeat), whose base motif ranged from 1–285 amino acids (Figure 1, step 1). Most of these repeats (94%) contain short base motifs, with 15% corresponding to microsatellites and 79% to minisatellites. Among the remaining TRs, 24,407 contain base motifs at least 30 amino acid long (1.6% of the total). However, a very large number of 28- and 29-amino-acid TRs were also found (4.8% of the total). These repeats mostly correspond to zinc-finger proteins (ZFPs), a widespread family of transcription regulators in vertebrates. We therefore decided to include these ZFPs in the present analysis. The 24,407 *bona fide* TRs exhibit base motif lengths ranging from 30–285 amino acids (Figure 2, inset).

Because repeated motifs of each TR can be very divergent from each other, containing many insertions or deletions, it was not possible to rely on their sequence to build families. Therefore, multiple sequence alignments were performed on TR motifs, and hidden Markov models (HMMs) were determined for each TR alignment (Figure 1, step 2). The HMM profiles were subsequently compared ''all against all'' using the HMMSEARCH software (HH-suite). To build the resulting graph, the log10 of the e-value (maximum 200) provided by HMMSEARCH was used as a distance. This value, although not a mathematical distance, was used to draw a complete graph between HMMs (Figure 1, step 3). From this graph, MCL, an unsupervised clustering method, identified TR clusters, each corresponding to a TR family (Figure 1, step 4). Proteins containing these TRs were identified, and all 1:1 orthologous proteins in which a TR was found were extracted from the Ensembl proteome (Figure 1, step 5). At this stage, HMMs were used to search TRs in these orthologs to ensure that no TR had been missed by the initial T-REKS search. At this stage, there are as many different orthologous protein families as there are TRs detected. Because it is likely that T-REX detects more than one TR in a particular family, many of these families are indeed identical. To merge them, a graph was constructed whose nodes are the families and whose edges represent a shared protein between two families (Figure 1, step 6). Thus constructed, the connected components of this graph constitute a set of TR orthologous families without duplicates. This method has the advantage of not being sensitive to TR phasing because the merging of families only depends on orthologous relationships based on similarity distributed over the whole protein sequence and not only on the TR. Mergings of identical families led to a final number of 257 families made of orthologous proteins in which at least one TR was detected (ORTHO FAM; Figure 1, step 7). Because, in the first place, we were interested in identifying megasatellites and not protein TRs, for each protein in which a TR was identified, the corresponding gene and transcript annotations were recovered from the Ensembl genome database. From these annotations, exon and intron coordinates were extracted, and DNA self-matrices were run, for each megasatellite-containing gene (Figure 1, step 8). These self-matrices were designed to manually check megasatellites to classify or discard them. Each of these 5,834 individual DNA matrices was visually inspected and validated or discarded from the database when no megasatellite was visible (Figure 1, step 9). At this stage, megasatellites were manually annotated as being contained in only one exon (MONOMEGA), more than one exon (MULTIMEGA), or overlapping at least one intron-exon junction (OVERMEGA). In some cases, the megasatellite was spread over several small exons separated by large intronic regions.

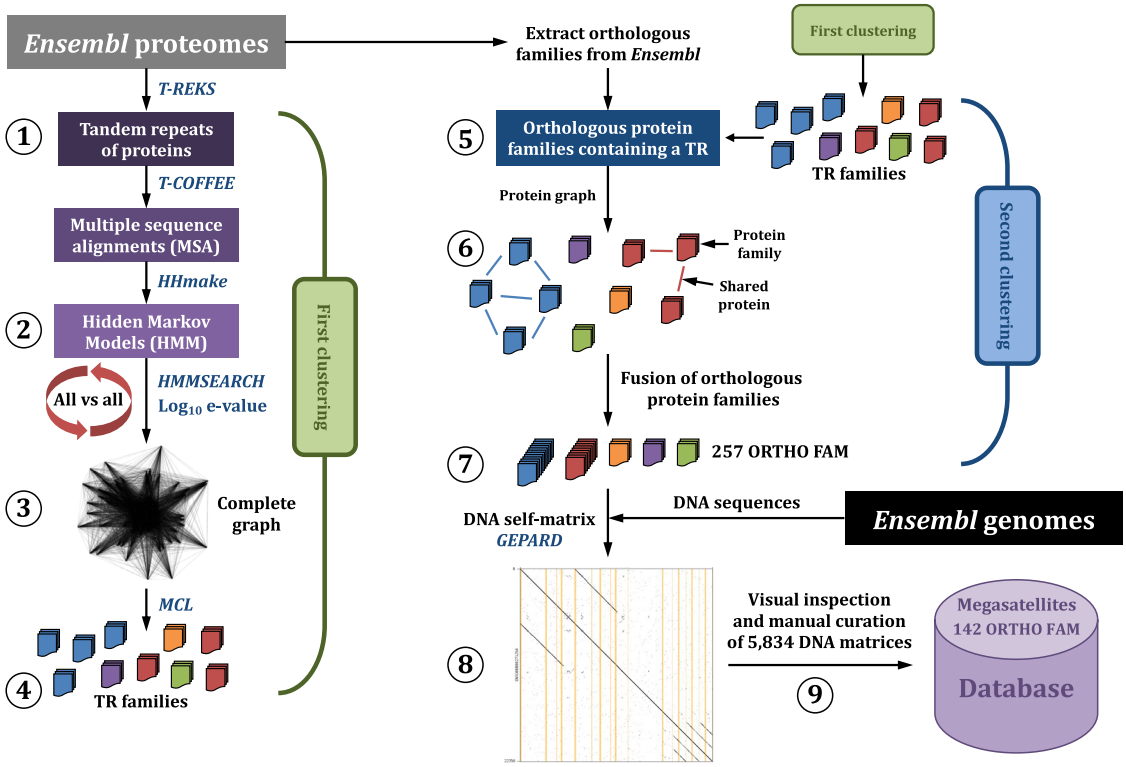


Figure 1. Pipeline used for the analysis

The T-REKS software was run on protein sequences extracted from the 58 vertebrate and *S. cerevisiae* proteomes. T-REKS identified 1,542,346 protein TRs whose base motif ranged from 1–285 amino acids (step 1). Multiple sequence alignments were performed on TR motifs, and hidden Markov models (HMMs) were determined for each TR alignment (step 2). HMM profiles were subsequently compared "all against all" using the HMMSEARCH software. To build the resulting graph, the log10 of the e-value (maximum 200) provided by HMMSEARCH was used as a distance. This value was used to draw a complete graph between HMMs (step 3). From this graph, MCL identified TR clusters, each corresponding to a TR family (step 4). All 1:1 orthologous proteins in which a TR was found were extracted from the Ensembl proteome (step 5). HMMs were used to search TRs in these orthologs to ensure that no TR had been missed by the initial T-REKS search. Identical families were merged using a graph whose nodes are the families and whose edges represent a shared protein between two families (step 6). Mergings of identical families led to a final number of 257 families made of orthologous proteins in which at least one TR was detected (ORTHO FAM, step 7). From Ensembl annotations, exon and intron coordinates were extracted, and DNA self-matrices were run, for each megasatellite-containing gene (step 8). Each of these 5,834 individual DNA matrices was visually inspected and validated or discarded from the database when no megasatellite was visible (step 9). This manual curation led to a final number of 3,982 megasatellites belonging to 142 megasatellite families.

These were called HIDDEN MULTIMEGA in the database but were considered MULTIMEGA in all subsequent analyses. Altogether, 3,982 megasatellites were identified, distributed among 142 families, including at least one member in one species. To this number must be added 10,575 ZFPs that were treated separately (see below). These families were called ORTHO FAM, the majority of them being MONOMEGA (65), MULTIMEGA (46), or OVERMEGA (26), and four contained a mix of more than one megasatellite type and were annotated as MIXMEGA (Table S2). Finally, one megasatellite was inadvertently found in an intronic sequence and conserved in 27 eutherian species. This suggests that it plays a functional role or that this gene annotation is wrong and this megasatellite is not purely intronic. Some examples of each category are shown in Figure 3. Because whole-genome duplications were frequent during vertebrate evolution (Dehal and Boore 2005; Jaillon et al., 2004), only these 142 ORTHO FAM families were considered for further analyses, and megasatellites encoded within paralogs were discarded at this stage.

Quality control of the pipeline

To determine whether our approach was exhaustive and discriminative, ORTHO FAM families were annotated using the Pfam database. Four well-described protein families known to possess large tandem repeat domains were examined: WD40 (Smith et al., 1999), leucine-rich repeats (LRRs; Kobe and Deisenhofer 1994), Ankyrin (Mosavi et al., 2004), and Kelch (Adams et al., 2000). In our analysis, WD40 and Kelch were each clustered into one single family (ORTHO FAM 189 and ORTHO FAM 156, respectively), proving that our pipeline correctly clusters members of these protein families that were detected by T-REKS. LRRs and Ankyrin were, respectively, found in three and five ORTHO FAMs, with HMM describing these ORTHO FAMs as significantly different because of the complexity and sequence variability of these repeats. In Pfam, Ankyrin repeats are also described by five different domains (Gasparini et al., 2017). Therefore, although our approach does not claim to be exhaustive, it seems sensitive and discriminative enough to identify well-known large tandem repeat families and describes 19

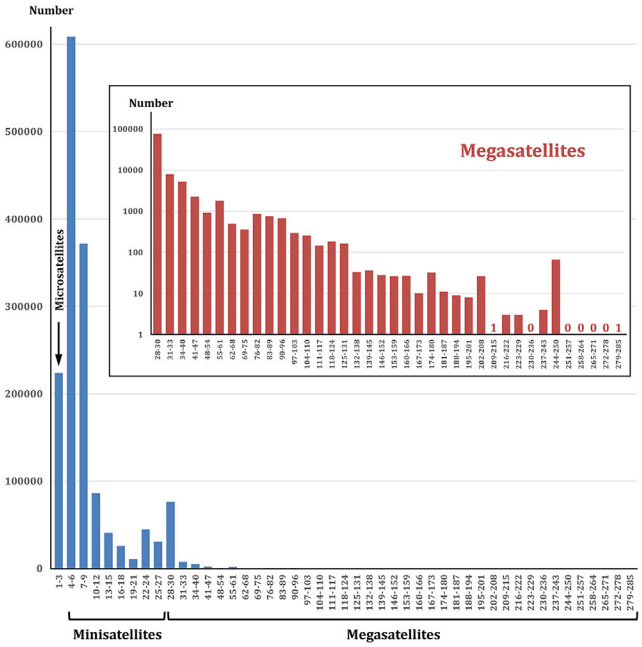


Figure 2. Length distribution of motifs detected by the T-REKS program

A bimodal distribution was observed, separating very short motifs (left of the red line) from longer motifs that were kept for further analyses (right of the red line). Motif lengths are in amino acids. Inset: size distribution of motifs at least 26 amino acids long.

that were not previously annotated as such (see "Distribution and function of megasatellites among species" below).

As a final validation of family homogeneity, all proteins in each family were annotated with a Panther identifier (STAR Methods). The result shows that 90% of the families contain only one single Panther ID, 8% contain two Panther IDs, and only 2% contain 3–5 identifiers (Figure S2). This proves the validity and consistency of our families.

Megasatellites are more frequent in subtelomeric regions

Minisatellites are unevenly distributed along eukaryotic chromosomes. They are enriched in subtelomeric regions in humans and *C. elegans*, whereas they are mainly located around the centromere in *A. thaliana* (Vergnaud and Denoeud 2008). Megasatellites are more frequent in subtelomeres in *S. cerevisiae* and *C. glabrata* (Richard and Dujon 2006; Thierry et al., 2008). To determine whether they exhibit a distribution bias in vertebrates, we extracted their position and compared it with a theoretical distribution if they were evenly distributed along each chromosome (STAR Methods). This was possible only for 22 species out of 58 because the other 36 genomes were available only as contigs. Each chromosome was cut into 10 segments of identical lengths, and the number of megasatellites present in the first and last 10% of the chromosome was compared with the theoretical number of a monotonous distribution. Comparisons of observed and expected values using a χ^2 test shows that, in all 22 species, subtelomeres are significantly enriched in megasatellites (Table S3).

We tried to perform the same analysis for centromeres. Mammalian centromeres are enriched in repeated elements, called satellite DNA, covering hundreds of kilobases (Ahmad et al., 2020). In the rest of the eukaryotic world, centromeres are made of various kinds of direct or inverted repeats and transposable elements, depending on the species considered (Muller et al., 2019). It is therefore a complicated task to identify centromeres, and they were indeed properly characterized in only six species of the above 22. In these species, we computed the number of megasatellites contained in the 10% of the chromosome length surrounding the centromere with the theoretical number using a χ^2 test. The results were not significant for two species (*Felis catus* and *Gallus gallus*) and showed a decrease of megasatellite density around centromeres in three primate species (*Gorilla gorilla*, *Homo Sapiens*, and *Pan troglodytes*). In *Mus musculus*, the opposite result was found, with a significant enrichment of megasatellites around centromeres (Table S3).

We concluded that megasatellites were significantly enriched in subtelomeres in all species and around centromeres in the mouse, whereas they tend to be less frequent in centromeric regions in primates.

Distribution of megasatellites among vertebrate clades

All vertebrate clades were found to contain megasatellites, although to different extents. Primates and eutherians contain, respectively, 116 and 107 of 142 families, whereas only 17 families were detected in Agnatha (Table S4). Eight of them are common to all vertebrate clades, Tenascin C (ORTHO FAM 1,130) and Tenascin R (ORTHO FAM 4,068), two related developmental proteins; Cortactin (ORTHO FAM 1,163), an actin polymerization protein; growth factor beta binding protein 1 (ORTHO FAM 1,746); Nebulin (ORTHO FAM 108) and Nebulette (ORTHO FAM 175), two proteins essential for regulation of the stability and length of actin filaments in skeletal and cardiac muscle fibers; CREB1 (ORTHO FAM 113), involved in cyclic AMP (cAMP) response; and Angiomotin (ORTHO FAM 4,295), involved in cell motility (Figure 4A). Nebulin and Nebulette have already been described as closely related repeat-containing genes present in all vertebrates (Björklund et al., 2010), and our results confirm this previous analysis.

Because the number of species studied and, hence, of proteomes varies among clades, family number was corrected according to proteome size (STAR Methods). It was found that all clades except eutherians exhibited fewer families than primates. Most clades, except Actinistia, Amphibia, Sauropsida and Monotremata, exhibited significant more families than Agnatha (Figure 4B). We concluded that there were probably two increases in megasatellite formation during vertebrate evolution, one at the root of the clade, after the Agnatha radiation, and another one much larger in mammals, after the Marsupiala radiation.

Three families were common to all Gnathostomata but were lost in Monotreme, and seven families common to all Sauropsida were also lost in Monotreme (Figure S3). It is unclear whether the significant loss of families in Monotreme reflects a biological reality or technical limitations in sequencing and/or assembly of the platypus genome, the only monotreme genome analyzed here. Finally, 11 families were common to therian mammals, 36

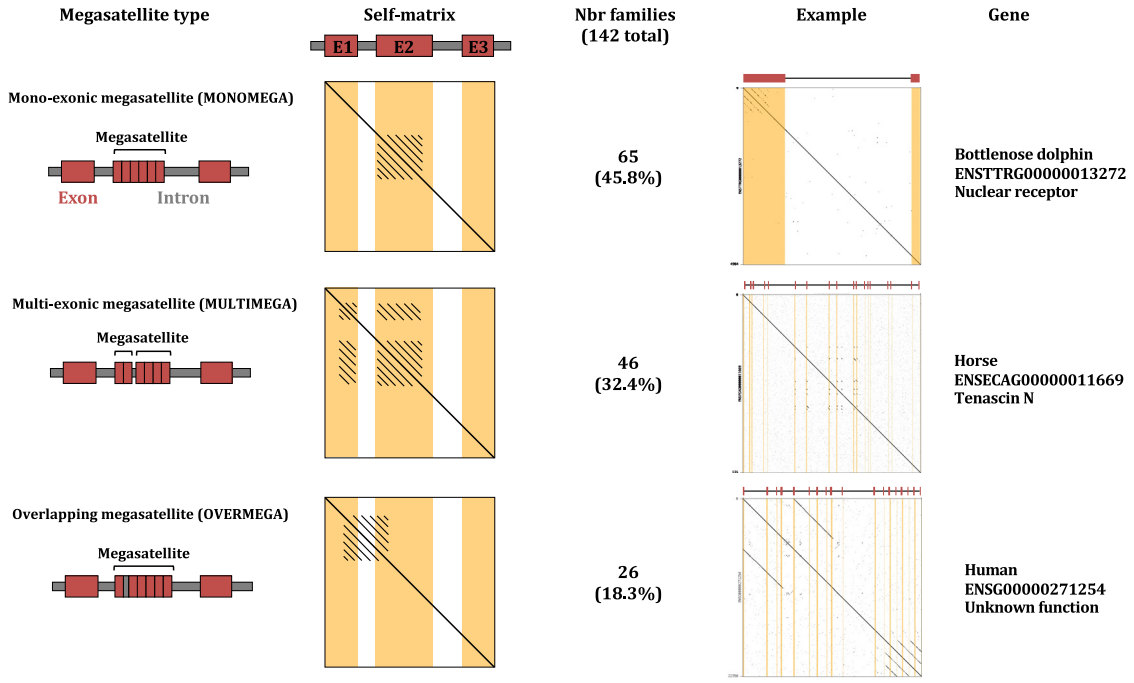


Figure 3. Megasatellite types

Left: MONOMEGA, MULTIMEGA, and OVERMEGA satellites are sketched. Red bars, exons; black line, introns. Center: schematic of self-matrix dot plots for each type. Right: One example of each megasatellite type is shown. Exons and introns are shown above each dot plot.

families were found to be conserved in all eutherians, and six were specific of non-primate eutherians, like mice, dolphins, and bats (Figure S3).

To check whether megasatellite distribution was homogeneous within each clade, their occurrences were compared in clades containing more than one species; i.e., birds, Marsupiala, teleostean fishes, eutherians, and primates. A significant excess of megasatellites was found in two birds (*Ficedula albicollis* and *Meleagris gallopavo*), two fishes (*Oryzias latipes* and *Poecilia formosa*), and two mammals (*Equus caballus* and *M. musculus*) (χ^2 $p < 0.01$). Besides these few cases, the number of megasatellites in each species was very homogeneous within a given clade.

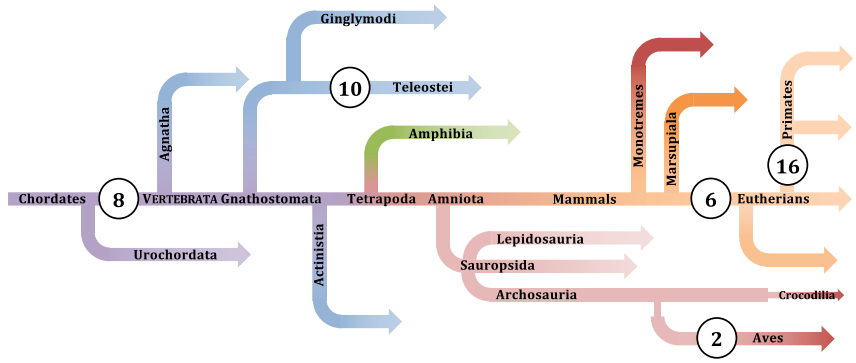
Distribution and function of megasatellites among species

Altogether, 3,982 megasatellites were detected in the 59 yeast and vertebrate families, not including ZFP-encoding genes (see below). Because paralogous genes were discarded during the analysis, only one member of each ORTHO FAM family was represented in each species; hence, the number of megasatellites per species is equal to the number of families present in that species. The species with the fewest repeats is *Petromyzon marinus* (19 megasatellites), whereas *M. musculus* has the highest number of megasatellites (133; Table S1).

For each megasatellite, the encoded protein was compared with the Pfam protein motif database, and the corresponding annotation (if any) was retrieved. Of 3,982 megasatellites, 205 did not match a protein motif in Pfam and therefore correspond to undescribed repeats. They were clustered in 19 different ORTHO FAMs. For each of these, an independent search in the Panther database (STAR Methods) was performed, and significant homology was found in 17 cases of 19 (Table S5).

The most frequent functions encoded by megasatellites were linked to cell membrane metabolism (15%); intracellular trafficking, including interactions with actin, myosin, and tubulin (11%); and RNA metabolism (10%). A significant fraction of them (8%) could not be associated with any known function based on sequence homology or protein motifs (Figure 5A). To assess whether these functions were significantly overrepresented compared with all cellular functions, we used the GO classification. GO terms (molecular functions) of megasatellite-encoded proteins were extracted and compared with all GO terms of the annotated vertebrate genomes using g:Profiler (Raudvere et al., 2019). Of 58 genomes, 26 did not show any significant enrichment. The remaining 32 genomes showed significant enrichment for functions related to protein binding, the cytoskeleton, calcium binding, receptor activity, fucose binding, protein complexes, and extracellular matrix (Figure S4). These functions are significantly overrepresented in megasatellite-encoded proteins and are present in all clades.

Because the GO terms collected here are rather vague to describe molecular functions or cellular pathways, we subsequently used Pfam protein motifs to infer the molecular function of megasatellite-encoded proteins. Most of the megasatellites (68%) encoded in the eight more frequent functional categories defined by Pfam are present in all clades, except for those involved in immune response, which are absent from Sauropsida and Monotreme (Figure 5B). Megasatellites in genes playing a function in protein metabolism are present only in Gnathostomata, with the exception of Lepidosauria and Monotreme.



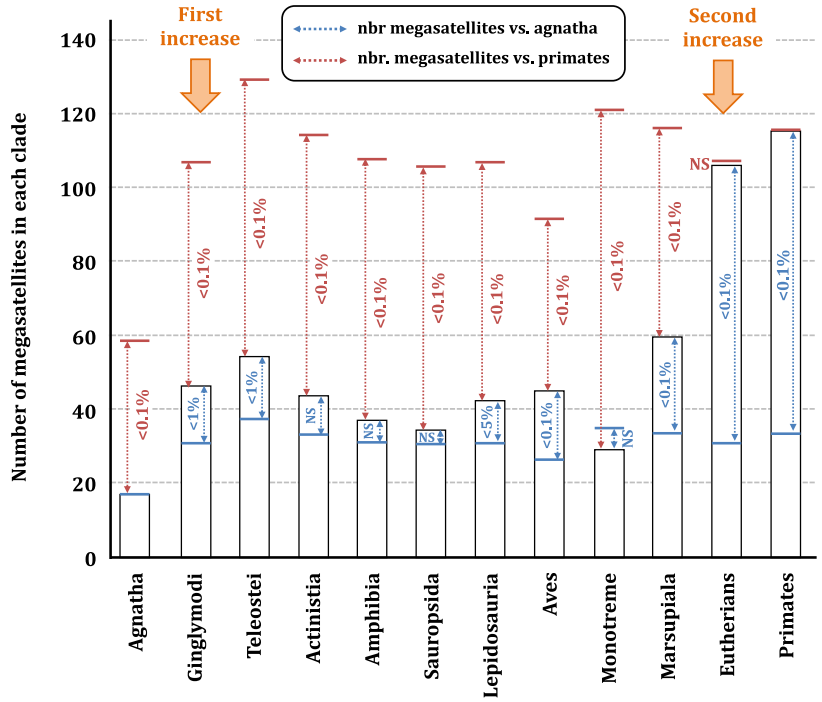


Figure 4. Cladogram of megasatellite distribution in vertebrates

(A) Clade-specific megasatellites. Branch lengths are arbitrary.

(B) Number of expected versus observed numbers of megasatellites in each clade. The observed number of megasatellites is shown by open bars. The expected numbers compared with Agnatha or primates are shown by blue and red horizontal bars, respectively. Corresponding χ^2 p values are indicated for each comparison. The two orange arrows point to the two statistical increases in megasatellite numbers during vertebrate evolution.

Several megasatellites are found in genes dedicated to a specific function and are unique to certain clades. It is the case of those related to the inflammatory response that are detected only in teleostei, and those involved in gametogenesis or keratin metabolism that are found only in eutherians (Figure 5B). These results show that, with a few exceptions, most megasatellites are encoded in genes that are common to all vertebrate clades and are therefore involved in conserved functions. Table S5 gives a complete list of all cellular functions encoded by megasatellites.

The *M. musculus* genome contains 59 megasatellites belonging to a gene encoding a product of unknown function, out of 133 in total. In comparison, the human genome contains only 32 megasatellites in genes of unknown function, out of 121 (26%). These proportions are statistically different (Fisher's exact test, p = 0.0038). Therefore, mouse genome megasatellites are more often encoded in genes of unknown function than human megasatellites. The possibility that they may encode mouse-specific functions that would be absent from other eutherians remains to be experimentally determined.

Megasatellites propagate by two different mechanisms

In the course of the present study, several orthologous families were identified as containing more than one member per species. Based on our analysis pipeline, this is highly unlikely because only one gene per species should be found in each family. A more thorough analysis of these 24 families showed that they included paralogous genes. When both paralogs carry the same megasatellite, independently identified at the beginning of the pipeline (Figure 1, step 1), they ended up in the same family after clustering (Figure 1, step 3). These 24 megasatellites were removed from the list of 142 orthologous families and considered paralogous families, which ended up with a final count of 118 orthologous families (Table S6).

These families showed that they could be classified in three different cases. Besides the 10 families whose sequence quality was too low to obtain reliable alignments, the simplest case included real paralogs containing the same megasatellite at the same position in each species. This was the status of six families (ORTHO FAM 143, 156, 1,163, 3,045, 3,121 and 3,340). An example of such a paralogous megasatellite is shown in Figure 6A. In four families (ORTHO FAM 113, 175, 1,130 and 3,163), more than two genes containing the same megasatellite were identified. Two of these genes are paralogs, and the third one is unrelated. This suggests that the megasatellite was transferred, or jumped, from one gene to another, although we cannot completely exclude that the third gene is a paralog that has

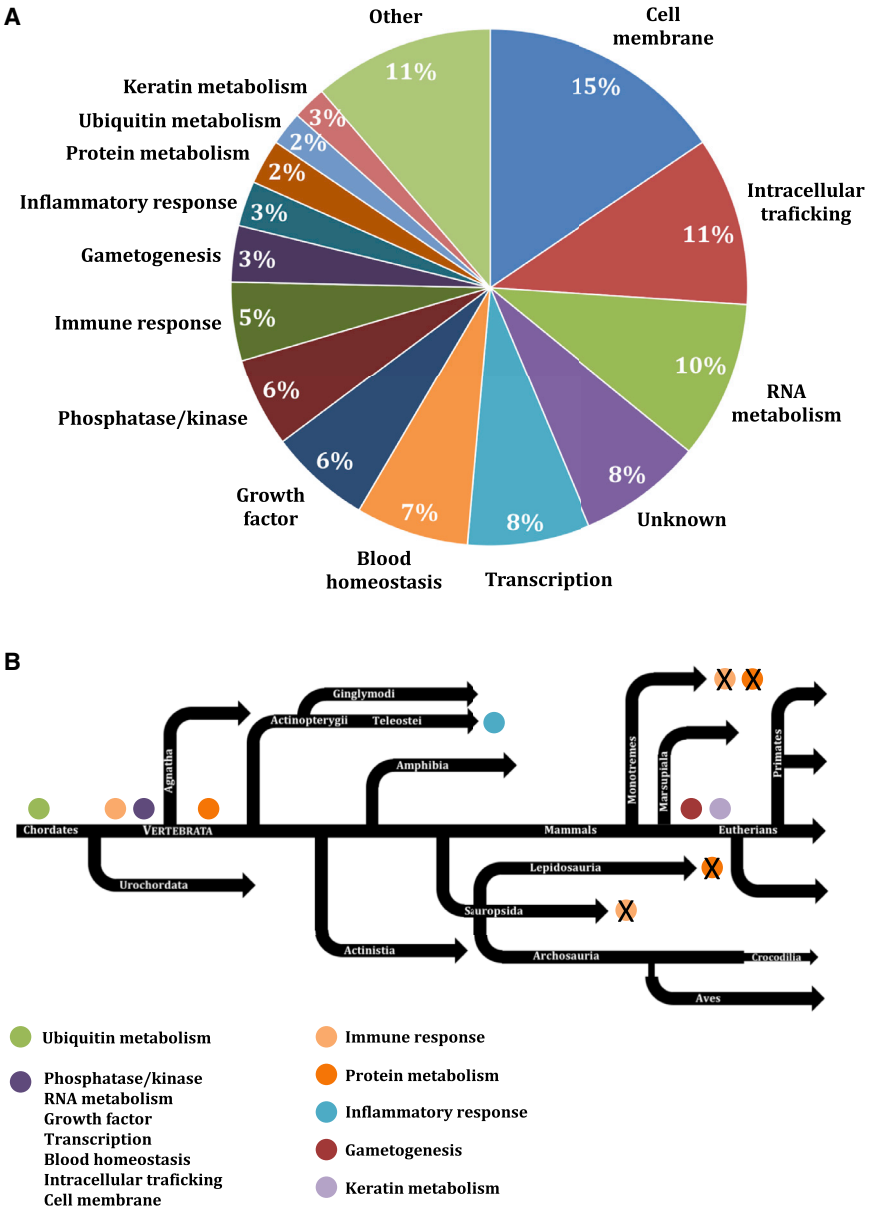


Figure 5. Function of megasatellite-encoding genes

(A) Pie chart showing the frequency of all identified functions. "Other" encompasses all functions associated with 1% or fewer genes.

(B) Repartition of the most frequent functions along a vertebrate cladogram. Branch lengths are arbitrary. Presence of a function is indicated by a colored disk and absence by a crossed colored disk.

diverged from the two others more rapidly than its megasatellite (Figure 6B). A third case corresponds to only one small family (ORTHO FAM 1,354) in which the same megasatellite is found in three different genes in one species. However, the megasatellite sequence is not totally conserved, and its length differs among the three genes, one of them containing only one motif, unrepeated (Figure 6C). Finally, three families contained short proteins, Ubiquitin C (UBC; ORTHO FAM 3,898) and two others too small to detect a reliable homology (ORTHO FAM 1,211 and 1,864) between their non-repeated parts. We concluded that megasatellites appear to propagate by two different modes, one involving duplication of an existing megasatellite-containing gene and another unexpected one, relying on transfer from one gene to another, phylogenetically unrelated gene.

Megasatellite copy number variability among species

Ubiquitin is a eukaryote-specific gene of archaeal origin encoding a tandemly repeated 76-amino acid polypeptide. It is post-translationally cleaved into the active 76-residue ubiquitin peptide, involved in regulating protein metabolism (Grau-Bové et al., 2015). In budding yeast, *UBI4* has been found to exhibit variability in the number of 76-amino acid tandem repeat units among different strains (Gemayel et al., 2017). In several eukaryotic lineages, *UBI4* is duplicated as two paralogous genes called UBB and UBC, with the latter exhibiting more repeat units than the former. Using the present approach, we detected polyubiquitin in 47 of 59 species studied (ORTH0 FAM 3,898). In some cases, two genes were identified as containing a polyubiquitin repeat. In these cases, it was assumed that the two copies correspond to UBB and UBC. When only one gene was detected, we assumed it was UBC, with UBB being the shorter one (Gemayel et al., 2017). The number of repeat units identified is generally in good accordance with previous reports. The average number of repeats is 5, but a large size variation is observed around this mean value (from 2–20 repeat units). There is no visible expansion of ubiquitin megasatellite length during vertebrate evolution. However, each clade exhibits large variability among species (Table S7).

Megasatellites are also widespread in fungal genomes (Tekaia et al., 2013) and are mainly found in genes involved in cell wall homeostasis and in cell-to-cell adhesion to other yeasts or to human epithelial cells. The *S. cerevisiae FLO1* gene, involved in yeast flocculation, contains one of the most well-studied megasatellites. Its length is positively correlated to flocculation. *FLO1* genes containing a long megasatellite flocculate more efficiently than those containing shorter megasatellites (Verstrepen et al., 2005). Similarly, it has been shown that the length of the megasatellite in the *FLO11* budding yeast gene was directly correlated to formation of buoyant biofilm at the surface of liquid cultures (Fidalgo et al., 2006).

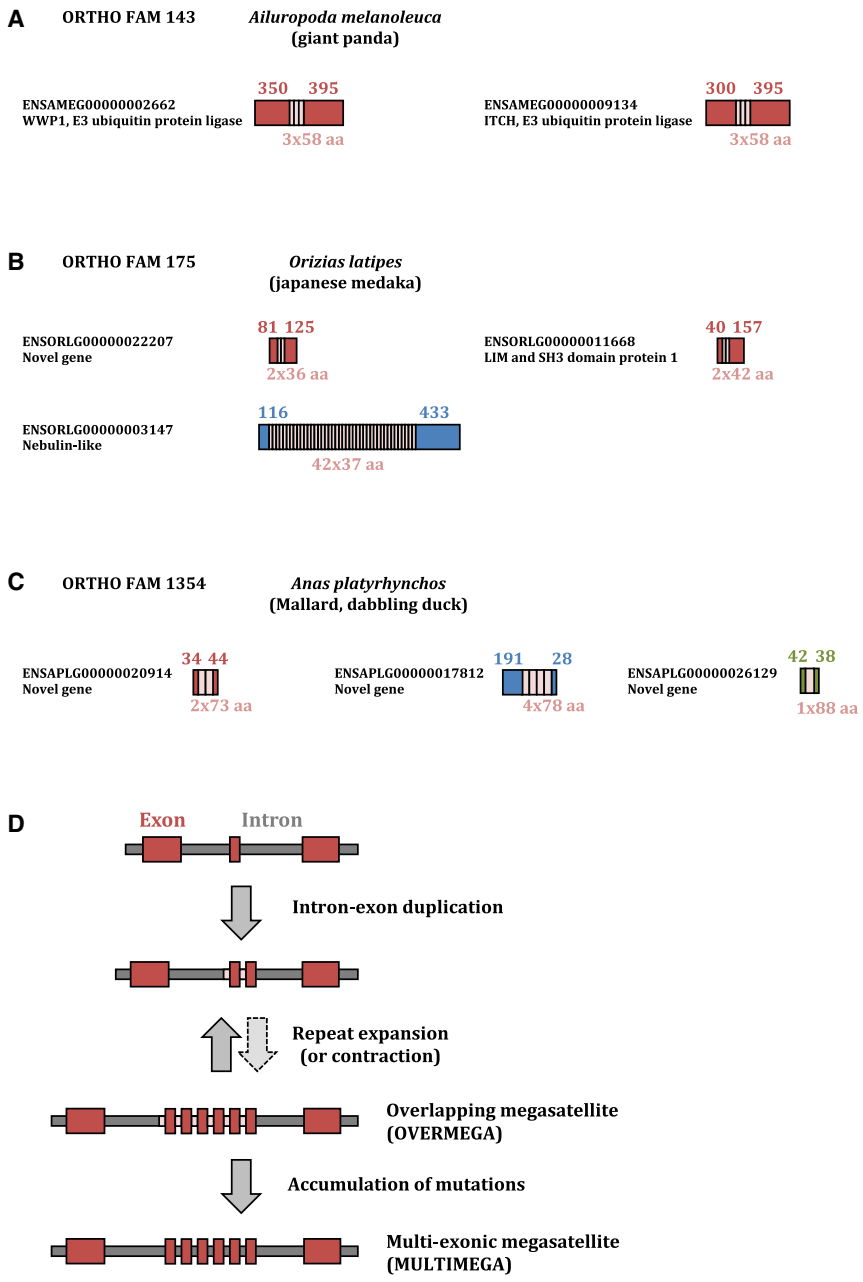


Figure 6. Different modes of megasatellite evolution

(A) An example of two paralogous genes containing the same megasatellite. The family as well as the species are indicated above the two paralogs. Gene names and supposed functions are shown next to the protein. Lengths of each protein part are indicated in amino acids.

(B) An example of a three-member family containing two paralogs. Homologous protein parts are shown in red, and the third protein is shown in blue. The megasatellite is present as a duplication in the two paralogs and is largely expanded in the third member.

(C) An example of a three-member family containing three unrelated genes (red, blue, and green). The megasatellite is present as a single motif in the green gene.

(D) A model for megasatellite evolution in vertebrates. See text for details.

Megasatellites are frequent in the opportunistic pathogen *C. glabrata* (Thierry et al., 2008). They exhibit length variability, but it is unclear whether they play an important role in cellular adhesion. *Candida albicans*, another opportunistic pathogenic yeast, also contains megasatellites in the paralogous agglutinin-like sequence (ALS) gene family, involved in yeast adhesion to human cells. The number of repeat motifs varies between 6 and 19 in the *ALS3* adhesin (Oh et al., 2005) and between 1 and 33 in the *ALS7* gene (Zhang et al., 2012) among different isolates. However, there is no known correlation between repeat length and adhesion or pathogenicity.

In *Aspergillus fumigatus*, several megasatellites are found in genes encoding cell wall proteins. The number of repeats in each megasatellite varies among natural isolates of this pathogenic fungus. However, it is unclear whether this variability is directly involved in the pathogenic process or in escaping the host immune system (Levdansky et al., 2007).

When all other megasatellite families were considered, the average number of motifs per tandem repeat was remarkably constant, around 7 per megasatellite in most clades, ranging from 6.3 in Gingly-modi to 10.9 in Agnatha (mean = 7.7, 99% confidence interval [CI] = 6.4–9.0). This indicates that megasatellites do not tend to increase in length with evolutionary time.

Abundance of ZFP families in eutherians, primates, and Lepidosauria

Although their base motif is slightly smaller than 90 bp (30 amino acids), ZFPs were analyzed in the present study. ZFPs contain a repetitive structure, the so-called "zinc finger," a 26- to 28-amino acid (78–84 nt) repeat, containing two histidine and two cysteine residues coordinating a zinc atom, hence their name. ZFPs are DNA-binding proteins containing a variable number of fingers, each of them binding 3 nt according to a specific code. Zinc-finger genes are part of very large paralogous families in vertebrates, involved in gene regulation networks. Zinc-finger polymorphism generates target diversity. For example, the product of the *PRDM9* gene is a ZFP recognizing a 13-mer DNA sequence present at mouse meiotic hotspots. Allelic variants of this gene are associated with

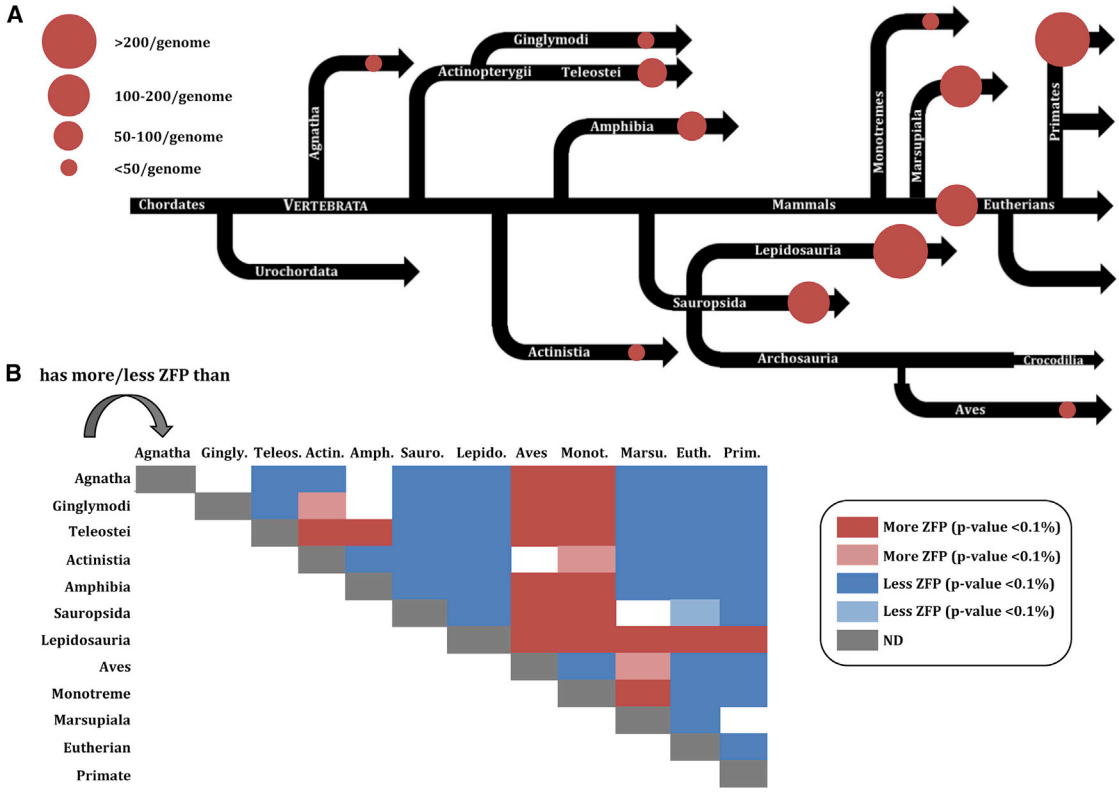


Figure 7. Zinc-finger genes in vertebrates

(A) Cladogram of zinc-finger gene distribution in vertebrate genomes. Branch lengths are arbitrary. The size of the colored disk represents the average number of ZFP in each genome.

(B) Comparisons of ZFP among clades. Clades exhibiting statistically less ZFP are shown in blue, and more ZFP are shown in red.

significant variations in hotspot usage among humans (Baudat et al., 2010).

ZF genes are highly represented in our set of megasatellites. Altogether, 10,575 ZF genes were found, unevenly distributed in all clades. There is a general increase in their number with time, with younger vertebrates encoding more of them than older clades (Figure 7A). When the proteome size of each clade was taken into consideration, ZFP number was significantly higher in Sauropsida, Lepidosauria, marsupials, eutherians, and primates (Figure 7B). Comparisons of these five clades showed that they were more frequent in primates and in Lepidosauria than in the three others and more frequent in Lepidosauria than in primates (χ^2 $p < 0.1\%$). We concluded that megasatellite-encoded ZFPs were expanded in Lepidosauria and more frequent than in any other vertebrate clade. This suggests that gene expression regulation may rely more frequently on ZFPs in this clade compared with other clades.

Krüppel-associated box domain-containing ZFPs (KZFPs) are a subclass of ZFPs that regulate transposable elements by repressing their expression. Here, KZFPs were restricted to four families, ORTHO FAM 1, 3,191, 4,452, and 5,539, with the three last families containing only five KZFPs. They are less frequent in Aves than in any other vertebrate clade, confirming a former study (Imbeault et al., 2017).

DISCUSSION

In the present study, we identified 3,982 megasatellites encoded in 58 vertebrate and budding yeast genomes (Table S6). It is, by far, the most complete description to date of large tandem repeats in eukaryotic genomes. It is remarkable that 18% of non-ZFP megasatellites are overlapping several introns (Figure 3), with these cases being more frequent in primates, which are the youngest species studied here, having diverged from other eutherians less than 10 million years ago (Figure S1). This is less frequent in older vertebrates, suggesting that megasatellite formation in vertebrates may start with duplication of an exon-intron DNA segment that subsequently becomes amplified to form an overlapping megasatellite. Mutations accumulate over time within the intronic part, erasing the tandem repeat and leading to what is detected as a multi-exonic megasatellite (Figure 6D). Intron size tends to increase over time during eukaryotic evolution, with ancestral eukaryotes exhibiting smaller introns than more recent ones (Csuros et al., 2011). If a tandem repeat is found overlapping one or more introns, then two mutational forces will tend to erase its presence over time: (1) intron size increases by accumulation of transposons or other virus-like elements, which will make megasatellite detection impossible with current software, and (2) accumulation of mutations that

will slowly erase the repeat in the intronic part of the megasatellite to retain only its intronic portion.

Vertebrate megasatellites exhibit two modes of propagation

In a former study on *S. cerevisiae* and *C. glabrata* megasatellites, it was suggested that these tandem repeats propagate by three different mechanisms: (1) duplication of an existing megasatellite-containing gene, (2) a megasatellite "jump" between two unrelated genes, and (3) gene conversion between paralogs (Rolland et al., 2010). Here we show that at least two of these mechanisms are recapitulated in vertebrates. Six families were found in which all species contained two paralogs encoding the same megasatellite (Figure 6A). These paralogous genes may come from whole-genome duplications, common in vertebrate evolution (Jaillon et al., 2004; Ohno 1970; Sacerdot et al., 2018), or from segmental duplications or CNVs (Sudmant et al., 2013; Zarrei et al., 2015). The second mechanism suggests that a given megasatellite may transfer its genetic information from one gene to another unrelated one, like a transposable element. Several examples compatible with this hypothesis have been discovered in the present study (Figures 6B and 6C). However, formal experimental proof of such a mechanism is lacking at the present time.

Megasatellites are frequently found in cell membrane genes

In yeast and fungi, most of the time megasatellite-encoding genes are involved in cell wall metabolism and function (Tekaia et al., 2013). In *S. cerevisiae*, it is the case of the *FLO* gene family important for cell-cell adhesion and flocculation (Verstrepen et al., 2005; Smukalla et al., 2008; Fidalgo et al., 2006). The large ALS gene family also contains a megasatellite whose role in adhesion has been demonstrated in *C. albicans* (Oh et al., 2005). In *C. glabrata*, the megasatellite-containing epithelial adhesin (EPA) gene family is also essential for cellular adhesion to human epithelial cells (Cormack et al., 1999). Here we show that a frequent function associated with megasatellite-encoding genes in vertebrates is cell membrane metabolism (Figure 5A). GO term analysis showed that megasatellite-encoded proteins are enriched in molecular functions related to the cytoskeleton, calcium binding, receptor activity, fucose binding, protein complexes, and extracellular matrix, functions important for membrane homeostasis (Figure S4). This suggests a universal role of long tandem repeats in cell membrane function and integrity.

Limitations of the study

The *S. cerevisiae* genome contains five megasatellites encoded in *UBI4*, *FLO1*, *FLO5*, *FLO9* (which are three paralogs), and *NUM1* (Verstrepen et al., 2005; Richard and Dujon 2006). With the present approach, only *UBI4*, encoding ubiquitin, was identified. This is due to a known limitation of our approach. To study megasatellite evolution, it was decided to build families of orthologous megasatellite-containing genes. *NUM1* and *FLO* megasatellites were detected by T-REKS, but because they have no orthologs in any vertebrate, they were discarded from the analysis. This limitation could be alleviated in the future by considering all megasatellites, not only those present in more than one genome.

A recurrent limitation of all studies on tandem repeats is sequence length and quality. The vertebrate genomes analyzed here were sequenced using first-generation Sanger sequencing or second-generation short read (Illumina) sequences. In both cases, read length is not sufficient to encompass a whole megasatellite. Therefore, read assembly is an obligate step to reconstitute the complete sequence, with all caveats linked to tandem repeats. For this reason, three genomes that did not seem to comply with high-quality sequence standards for repeats were discarded from the present analysis (Table S1). The absence of several common megasatellites from the Monotreme may be due to incomplete assembly of the platypus genome sequence (Warren et al., 2008), although we cannot exclude that these megasatellites may have been missed in this species.

Nevertheless, it is probable that megasatellite length was often underestimated here. Third-generation sequencing using Nanopore or PacBio long reads should help overcome this problem when sequence quality of such reads will be improved to the level of alternative technologies. Another issue is allelic polymorphism because all vertebrate genomes are diploid but only one set of chromosome sequence was available in the database. This issue is, of course, not specific to tandem repeat analyses and will be improved in the coming years with more efficient and thorough sequence analyses and database storage.

Finally, tandem repeat detection is another limitation because it is known that no algorithm is able to correctly detect all tandem repeats of a given genome (Leclercq et al., 2007). For this reason, we focused the present work on megasatellites found in protein-coding genes, but non-coding regions also contain large tandem repeats, like CNVs (Sudmant et al., 2013; Zarrei et al., 2015), and probably many megasatellites. By analogy with the distribution of minisatellites in the *S. cerevisiae* genome (Richard and Dujon 2006), 35% of which are found in non-coding regions, we may infer that a similar proportion of megasatellites will be found in non-coding regions in vertebrates. However, this comparison is limited by the fact that the yeast genome is much more compact than vertebrate genomes. Megasatellite identification is such a large task, and complex genomes will require additional tools that are not currently available. Progress should therefore also be made in this area, using more sophisticated software based on alternative detection methods. This is probably the next frontier of tandem repeat identification in complex genomes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Detection of TR-bearing proteins
  - TR modeling
  - TR clustering

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111347.

ACKNOWLEDGEMENTS

Work in the G.-F. Richard laboratory is generously supported by the Institut Pasteur and by the Centre National de la Recherche Scientifique (CNRS).

AUTHOR CONTRIBUTIONS

S.D.-D. and G.-F.R. designed the analysis. S.D.-D. wrote the pipeline, set up the database, and extracted the results. S.D.-D. and G.-F.R. analyzed the results and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: November 23, 2021
Revised: April 28, 2022
Accepted: August 23, 2022
Published: September 13, 2022

REFERENCES

Adams, J., Kelso, R., and Cooley, L. (2000). The kelch repeat superfamily of proteins: propellers of cell function. Trends Cell Biol. *10*, 17–24.

Ahmad, S.F., Singchat, W., Jehangir, M., Suntronpong, A., Panthum, T., Malaivijitnond, S., and Srikulnath, K. (2020). Dark matter of primate genomes: satellite DNA repeats and their evolutionary dynamics. Cells *9*, 2714.

Anisimova, M., Pečerska, J., and Schaper, E. (2015). Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. Front. Bioeng. Biotechnol. *3*, 31. https://www.frontiersin.org/articles/10.3389/fbioe.2015.00031/full.

Bachtrog, D., Weiss, S., Zangerl, B., Brem, G., and Schlötterer, C. (1999). Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. Mol. Biol. Evol. *16*, 602–610.

Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M., and Eichler, E.E. (2004). Analysis of segmental duplications and genome assembly in the mouse. Genome Res. *14*, 789–801.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836–840.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Björklund, A.K., Light, S., Sagit, R., and Elofsson, A. (2010). Nebulin: a study of protein repeat evolution. J. Mol. Biol. *402*, 38–51.

Bowen, S., Roberts, C., and Wheals, A.E. (2005). Patterns of polymorphism and divergence in stress-related yeast proteins. Yeast *22*, 659–668.

Bromham, L., Rambaut, A., Fortey, R., Cooper, A., and Penny, D. (1998). Testing the Cambrian explosion hypothesis by using a molecular dating technique. Proc. Natl. Acad. Sci. USA *95*, 12386–12389.

Cormack, B.P., Ghori, N., and Falkow, S. (1999). An adhesin of the yeast pathogen Candida glabrata mediating adherence to human epithelial cells. Science *285*, 578–582.

Csuros, M., Rogozin, I.B., and Koonin, E.V. (2011). A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput. Biol. *7*, e1002150.

Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. *3*, e314.

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., et al. (1996). A comprehensive genetic map of the human genome based on 5, 264 sequences. Nature *380*, 152–154.

Dieringer, D., and Schlötterer, C. (2003). Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. *13*, 2242–2251.

Erwin, D.H., Laflamme, M., Tweedt, S.M., Sperling, E.A., Pisani, D., and Peterson, K.J. (2011). The cambrian conundrum: early divergence and later ecological success in the early history of animals. Science *334*, 1091–1097.

Fidalgo, M., Barrales, R.R., Ibeas, J.I., and Jimenez, J. (2006). Adaptive evolution by mutations in the FLO11 gene. Proc. Natl. Acad. Sci. USA *103*, 11228–11233.

Gasparini, A., Tosatto, S.C.E., Murgia, A., and Leonardi, E. (2017). Dynamic scaffolds for neuronal signaling: in silico analysis of the TANC protein family. Sci. Rep. *7*, 6829.

Gemayel, R., Yang, Y., Dzialo, M.C., Kominek, J., Vowinckel, J., Saels, V., Van Huffel, L., van der Zande, E., Ralser, M., Steensels, J., et al. (2017). Variable repeats in the eukaryotic polyubiquitin gene ubi4 modulate proteostasis and stress survival. Nat. Commun. *8*, 397.

Genome 10K Community of Scientists (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. J. Hered. *100*, 659–674.

Grau-Bové, X., Sebé-Pedrós, A., and Ruiz-Trillo, I. (2015). The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin. Mol. Biol. Evol. *32*, 726–739.

Hennequin, C., Thierry, A., Richard, G.-F., Lecointre, G., Nguyen, H.V., Gaillardin, C., and Dujon, B. (2001). Microsatellite typing as a new tool for identification of Saccharomyces cerevisiae strains. J. Clin. Microbiol. *39*, 551–559.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554.

Innan, H., Terauchi, R., and Miyashita, N.T. (1997). Microsatellite polymorphism in natural populations of the wild plant Arabidopsis thaliana. Genetics *146*, 1441–1452.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature *431*, 946–957.

Jorda, J., and Kajava, A.V. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics *25*, 2632–2638.

Kobe, B., and Deisenhofer, J. (1994). The leucine-rich repeat: a versatile binding motif. Trends Biochem. Sci. *19*, 415–421.

Kolpakov, R., Bana, G., and Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. *31*, 3672–3678.

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. (2004). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J. *23*, 234–243.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics *23*, 1026–1028.

Leclercq, S., Rivals, E., and Jarne, P. (2007). Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinf. *8*, 125.

Levdansky, E., Romano, J., Shadkchan, Y., Sharon, H., Verstrepen, K.J., Fink, G.R., and Osherov, N. (2007). Coding tandem repeats generate diversity in Aspergillus fumigatus genes. Eukaryot. Cell *6*, 1380–1391.

Linardopoulou, E.V., Williams, E.M., Fan, Y., Friedman, C., Young, J.M., and Trask, B.J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature *437*, 94–100.

Makałowski, W., Mitchell, G.A., and Labuda, D. (1994). Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. *10*, 188–193.

Malpertuy, A., Dujon, B., and Richard, G.-F. (2003). Analysis of microsatellites in 13 hemiascomycetous yeast species: mechanisms involved in genome dynamics. J. Mol. Evol. *56*, 730–741.

Marques-Bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L.A., et al. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. Nature *457*, 877–881.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., and Thomas, P.D. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat. Protoc. *14*, 703–721.

Millot, G. (2011). Comprendre et réaliser les tests statistiques à l'aide de R, 2nd edition (de boeck).

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419.

Mosavi, L.K., Cammett, T.J., Desrosiers, D.C., and Peng, Z.Y. (2004). The ankyrin repeat as molecular architecture for protein recognition. Protein Sci. *13*, 1435–1448.

Muller, H., Gil, J., and Drinnenberg, I.A. (2019). The impact of centromeres on spatial genome architecture. Trends Genet. *35*, 565–578.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton. J. Mol. Biol. *302*, 205–217.

Oh, S.H., Cheng, G., Nuessen, J.A., Jajko, R., Yeater, K.M., Zhao, X., Pujol, C., Soll, D.R., and Hoyer, L.L. (2005). Functional specificity of Candida albicans Als3p proteins and clade specificity of ALS3 alleles discriminated by the number of copies of the tandem repeat sequence in the central domain. Microbiology *151*, 673–681.

Ohno, S. (1970). Evolution by Gene Duplication (Springer).

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198.

Richard, G.-F., and Dujon, B. (2006). Molecular evolution of minisatellites in hemiascomycetous yeasts. Mol. Biol. Evol. *23*, 189–202.

Richard, G.-F., Hennequin, C., Thierry, A., and Dujon, B. (1999). Trinucleotide repeats and other microsatellites in yeasts. Res. Microbiol. *150*, 589–602.

Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. *72*, 686–727.

Röder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.-H., Leroy, P., and Ganal, M.W. (1998). A microsatellite map of wheat. Genetics *149*, 2007–2023.

Roest Crollius, H., Jaillon, O., Dasilva, C., Ozouf-Costaz, C., Fizames, C., Fischer, C., Bouneau, L., Billault, A., Quetier, F., Saurin, W., et al. (2000). Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. Genome Res. *10*, 939–949.

Rolland, T., Dujon, B., and Richard, G.-F. (2010). Dynamic evolution of megasatellites in yeasts. Nucleic Acids Res. *38*, 4731–4739.

Sacerdot, C., Louis, A., Bon, C., Berthelot, C., and Roest Crollius, H. (2018). Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. *19*, 166.

Sakamoto, T., Danzmann, R.G., Gharbi, K., Howard, P., Ozaki, A., Khoo, S.K., Woram, R.A., Okamoto, N., Ferguson, M.M., Holm, L.-E., et al. (2000). A microsatellite linkage map of rainbow trout (Oncorhynchus mykiss) characterized by large sex-specific differences in recombination rates. Genetics *155*, 1331–1345.

Smith, T.F., Gaitatzes, C., Saxena, K., and Neer, E.J. (1999). The WD repeat: a common architecture for diverse functions. Trends Biochem. Sci. *24*, 181–185.

Smukalla, S., Caldara, M., Pochet, N., Beauvais, A., Guadagnini, S., Yan, C., Vinces, M.D., Jansen, A., Prevost, M.C., Latgé, J.P., et al. (2008). FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. Cell *135*, 726–737.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. Bioinformatics *21*, 951–960.

Sudmant, P.H., Huddleston, J., Catacchio, C.R., Malig, M., Hillier, L.W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R.E., Persengiev, S., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. Genome Res. *23*, 1373–1382.

Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Human Genome Structural Variation Consortium; Warren, W.C., Pollen, A.A., Chaisson, M.J.P., et al. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. USA *116*, 23243–23253.

Tekaia, F., Dujon, B., and Richard, G.-F. (2013). Detection and characterization of megasatellites in orthologous and nonorthologous genes of 21 fungal genomes. Eukaryot. Cell *12*, 794–803.

Thierry, A., Bouchier, C., Dujon, B., and Richard, G.-F. (2008). Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Res. *36*, 5970–5982.

Thierry, A., Dujon, B., and Richard, G.F. (2009). Megasatellites: a new class of large tandem repeats discovered in the pathogenic yeast Candida glabrata. Cell. Mol. Life Sci. *67*, 671–676. https://doi.org/10.1007/s00018-009-0216-y.

Tuzun, E., Bailey, J.A., and Eichler, E.E. (2004). Recent segmental duplications in the working draft assembly of the brown Norway rat. Genome Res. *14*, 493–506.

van Dongen, S. (2014). Clustering similarity graphs encoded in BLAST results. https://micans.org/mcl/man/clmprotocols.html.

van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. In Bacterial Molecular Networks: Methods and Protocols, J. Van Helden, A. Toussaint, and D. Thieffry, eds. (Springer), pp. 281–295. https://doi.org/10.1007/978-1-61779-361-5_15.

Vergnaud, G., and Denoeud, F. (2008). Minisatellites: mutability and genome architecture. Genome Res. *10*, 899–907.

Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. Nat. Genet. *37*, 986–990.

Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T., et al. (2008). Genome analysis of the platypus reveals unique signatures of evolution. Nature *453*, 175–183.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.

Zhang, N., Harrex, A.L., Holland, B.R., Fenton, L.E., Cannon, R.D., and Schmid, J. (2012). Sixty alleles of the ALS7 open reading frame in Candida albicans: ALS7 is a hypermutable contingency locus. Genome Res. *13*, 2005–2017.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Code used to classify and cluster megasatellites	Github	https://github.com/sdeclere/tandem_repeats
Code used to classify and cluster megasatellites	Zenodo	https://zenodo.org/badge/latestdoi/366383351
Self-matrices	Figshare	https://doi.org/10.6084/m9.figshare.19668807
Ailuropoda melanoleuca, release 04-sept-18	*Ensembl* database	N/A
Anas platyrhynchos, release 04-sept-18	*Ensembl* database	N/A
Anolis carolinensis, release 05-sept-18	*Ensembl* database	N/A
Astyanax mexicanus, release 05-sept-18	*Ensembl* database	N/A
Bos taurus, release 05-sept-18	*Ensembl* database	N/A
Callithrix jacchus, release 05-sept-18	*Ensembl* database	N/A
Canis familiaris, release 04-sept-18	*Ensembl* database	N/A
Cavia porcellus, release 04-sept-18	*Ensembl* database	N/A
Chlorocebus sabaeus, release 05-sept-18	*Ensembl* database	N/A
Choloepus hoffmanni, release 05-sept-18	*Ensembl* database	N/A
Danio rerio, release 05-sept-18	*Ensembl* database	N/A
Dasypus novemcinctus, release 05-sept-18	*Ensembl* database	N/A
Dipodomys ordii, release 05-sept-18	*Ensembl* database	N/A
Echinops telfairi, release 04-sept-18	*Ensembl* database	N/A
Equus caballus, release 05-sept-18	*Ensembl* database	N/A
Erinaceus europaeus, release 04-sept-18	*Ensembl* database	N/A
Felis catus, release 04-sept-18	*Ensembl* database	N/A
Ficedula albicollis, release 04-sept-18	*Ensembl* database	N/A
Gadus morhua, release 05-sept-18	*Ensembl* database	N/A
Gallus gallus, release 05-sept-18	*Ensembl* database	N/A
Gasterosteus aculeatus, release 05-sept-18	*Ensembl* database	N/A
Gorilla gorilla, release 04-sept-18	*Ensembl* database	N/A
Homo sapiens, release 05-sept-18	*Ensembl* database	N/A
Ictidomys tridecemlineatus, release 05-sept-18	*Ensembl* database	N/A
Latimeria chalumnae, release 05-sept-18	*Ensembl* database	N/A
Lepisosteus oculatus, release 04-sept-18	*Ensembl* database	N/A
Loxodonta Africana, release 04-sept-18	*Ensembl* database	N/A
Macaca mulatta, release 05-sept-18	*Ensembl* database	N/A
Macropus eugenii, release 23-nov-16	*Ensembl* database	N/A
Meleagris gallopavo, release 05-sept-18	*Ensembl* database	N/A
Microcebus murinus, release 05-sept-18	*Ensembl* database	N/A
Monodelphis domestica, release 04-sept-18	*Ensembl* database	N/A
Mus musculus, release 05-sept-18	*Ensembl* database	N/A
Mustela putorius furo, release 05-sept-18	*Ensembl* database	N/A
Myotis lucifugus, release 05-sept-18	*Ensembl* database	N/A
Nomascus leucogenys, release 05-sept-18	*Ensembl* database	N/A
Ochotona princeps, release 04-sept-18	*Ensembl* database	N/A
Oreochromis niloticus, release 04-sept-18	*Ensembl* database	N/A
Ornithorhynchus anatinus, release 04-sept-18	*Ensembl* database	N/A
Oryctolagus cuniculus, release 05-sept-18	*Ensembl* database	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oryzias latipes, release 05-sept-18	*Ensembl* database	N/A
Otolemur garnettii, release 04-sept-18	*Ensembl* database	N/A
Ovis aries, release 05-sept-18	*Ensembl* database	N/A
Pan troglodytes, release 04-sept-18	*Ensembl* database	N/A
Papio Anubis, release 05-sept-18	*Ensembl* database	N/A
Pelodiscus sinensis, release 05-sept-18	*Ensembl* database	N/A
Petromyzon marinus, release 04-sept-18	*Ensembl* database	N/A
Poecilia Formosa, release 05-sept-18	*Ensembl* database	N/A
Pongo abelii, release 05-sept-18	*Ensembl* database	N/A
Procavia capensis, release 04-sept-18	*Ensembl* database	N/A
Pteropus vampyrus, release 05-sept-18	*Ensembl* database	N/A
Rattus norvegicus, release 05-sept-18	*Ensembl* database	N/A
Saccharomyces cerevisiae, release 05-sept-18	*Ensembl* database	N/A
Sarcophilus harrisii, release 05-sept-18	*Ensembl* database	N/A
Sorex Araneus, release 05-sept-18	*Ensembl* database	N/A
Sus scrofa, release 04-sept-18	*Ensembl* database	N/A
Taeniopygia guttata, release 04-sept-18	*Ensembl* database	N/A
Tarsius syrichta, release 24-nov-16	*Ensembl* database	N/A
Tetraodon nigroviridis, release 05-sept-18	*Ensembl* database	N/A
Tursiops truncates, release 05-sept-18	*Ensembl* database	N/A
Vicugna pacos, release 04-sept-18	*Ensembl* database	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be addressed to Guy-Franck Richard (gfrichar@pasteur.fr).

Materials availability

No material was generated in the course of this study.

Data and code availability

This study did not generate any unique datasets. All self-matrices are available at Figshare:10.6084/m9.figshare.19668807. The code generated during this study is available at Github (https://github.com/sdeclere/tandem_repeats). Code and data are available at Zenodo: https://zenodo.org/badge/latestdoi/366383351. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

We developed an original method able to capture the genomic structure of orthologous Tandem Repeat (TR)-bearing protein families. As it is difficult to detect megasatellites directly in vertebrate genomes, our analysis consisted of searching for TR-rich orthologous protein families and studying the genomic structures of the genes encoding the members of these families using dotplots. The whole pipeline of analysis is schematized in Figure 1.

Detection of TR-bearing proteins

We executed T-REKS (Jorda and Kajava 2009) on all proteins of all proteomes of our dataset. To keep only the most relevant results, length distribution of all detected tandem repeats was studied. The threshold retained for a *bona fide* megasatellite was set at 30 amino acids (90 nucleotides) and at least three repeated motifs (Figure 2). Since only protein-coding genes were considered, no filter for transposable element was set up. The reasoning was that if a transposon was translated as part of an exon, it was part of the megasatellite, hence should not be excluded (Makałowski et al., 1994).

TR modeling

To capture inter-motif sequence diversity, we formalized protein TR as a Hidden Markov Model (HMM) profile (Anisimova et al., 2015). To do this, each tandem repeat was subdivided into its individual motifs and their sequences were saved in a multi-sequence file. Those files were then aligned with T-COFFEE (Notredame et al., 2000) and resulting Multi Sequence Alignments (MSA) were filtered to retain only columns with less than 50% gaps.

TR clustering

Distances between MSA were calculated with HH-suite (Söding 2005). In order to group related TR (e.g. belonging to orthologous proteins), we instantiated a TR similarity graph from which clusters of strongly similar TR were extracted using the MCL method (van Dongen and Abreu-Goodger 2012). Practically, we first transformed each MSA into an HMM using the hhmake tool, using default parameters. For each megasatellite cluster, only one HMM was generated.

Then we concatenated all produced HMM profiles into a giant HMM dataset. In order to generate similarity matrix of all TR, each profile was compared to all others, running a HHsearch against the full dataset. All results were then transcoded from hhr formatted results to ''abc'' format (https://micans.org/mcl/man/clmprotocols.html). The e-value was retained as the distance between two HMM, and will be used to draw a complete similarity graph using MCL, which implements a fast and scalable unsupervised cluster algorithm. It was run with Inflation parameters set to 1.4 in order to cut the graph into clusters, each of them representing ancestrally related TR.

In order to verify that the clustering step did not fail to group related HMM profiles, the heatmap of the distance of each family against all families was drawn (Figure S5). For each comparison, the $-\log_{10}$(e-value) is used as the measure of distance between two HMM. The best pair is always the HMM against itself, as expected if the clustering is correct. Note that only the 118 families that exclusively contain orthologous members (see above) out of the 142 ORTHO FAM, were used at this step.

Constitution of ORTHO FAM families

From each MCL cluster, TR-carrying proteins composing this cluster were extracted. For each protein, the API Rest of Ensembl (Yates et al., 2020) was used to retrieve orthologues in the 58 vertebrate genomes eventually kept. These orthologues were added to the protein lists to make protein families. These consolidated families, called ORTHO FAM, are composed of TR-bearing proteins and of orthologous proteins that may or may not contain a TR.

Merging of differently-phased TR

This step addresses the problem of TR phasing. Depending on the first amino acid used to define the TR, motif sequences may not be identical in the end. For example, the sequence ABCD ABCD ABCD may be phased as a TR whose motif is ABCD, or BCDA, CDAB, or DABC. Therefore, they could possibly be gathered in different TR clusters. To circumvent this problem, a script producing an ORTHO_FAM inclusion graph was developed. Each node is a protein family and each edge symbolizes a shared protein. Using this graph, families containing strongly connected components (homologous proteins) were merged. This allowed to merge TR-containing proteins which were differently phased by T-REKS.

Database and dotplots

These new ORTHO FAM were deposited in a SQLite database. For each protein component of these families, genomic sequence and gene model corresponding to each protein were retrieved. Using this information and a modified version of Gepard (Krumsiek et al., 2007), dotplots for all the genes encoding these proteins were generated. Each of the 5,834 dot plots was visually inspected and manually annotated, using the following criteria. If the megasatellite was detected in only one exon, it was annotated as MONOMEGA (Figure 3, top). If it was overlapping two or more exons, it was annotated as MULTIMEGA (Figure 3, middle). If the megasatellite was overlapping at least one exon-intron junction, it was annotated as OVERMEGA (Figure 3, bottom). These cases were the simple ones and with some experience can be quickly annotated. More complex cases include megasatellites that are spread over several small exons separated by large intronic regions (for example ORTHO FAM 1746, Supplemental Material S1). In such rare cases, the self-matrix was insufficient to visually detect the megasatellite. In these cases, the ambiguity came from the resolution of the self-matrix, which did not allow to clearly visualize diagonals. This was due to several unusually long introns cutting the megasatellite in several undetectable small pieces. These were called HIDDEN MULTIMEGA in the database, but were considered as MULTIMEGA in all subsequent analyses. Finally, when no megasatellite could be detected by eye on the matrix, the protein self-matrix was checked. In all cases, a TR was found on the matrix, showing that T-REKS correctly identified protein TR but the megasatellite was erased by subsequent mutations following its formation. This manual curation led to a final number of 3,982 megasatellites belonging to 142 megasatellite families. All megasatellites are described in Figure S4 and all dot plots are in Supplemental Material S1 (available at 10.6084/m9.figshare.19668807).

Determination of expected family numbers

The number of species in each clade varies from one to 26. Therefore, in order to compare family numbers between different clades, corrections were necessary. First, the average proteome length in each clade was calculated by dividing the total proteome length by the number of species (Table S8). Second, based on this average proteome length, the expected number of families was calculated

for each clade, using the frequency observed in agnatha (which shows the lowest frequency). Third, the expected number of families was calculated for each clade, using the frequency observed in primates (which shows the highest frequency). Finally, the observed number of families was compared using a Chi2 test to the expected number of families based on agnatha and primate frequencies (Figure 4B).

The same approach was used when comparing zinc-finger proteins between clades. In that case, the expected number in each clade was compared to the observed number in each clade (Figure 7B).

Similarly, the expected number of families in each species was calculated by taking into consideration respective proteome sizes of each species.

Megasatellite distribution along chromosomes

In order to study the distribution of megasatellites along the chromosomes, we first extracted the identifiers of all megasatellite-containing genes in our database. The chromosomal position of all extracted genes was then obtained by querying the *Ensembl* REST API using the "lookup/id" resource (https://rest.ensembl.org/). In parallel, chromosome lengths were retrieved from *Ensembl* using the REST API by querying the information related to the assembly through the "info/assembly" resource. At this stage, were kept only chromosomes properly identified with numbers (contigs were therefore excluded). With this information, all chromosomes were cut into 10 bins of equal lengths. Finally, the center of each megasatellite-containing gene was assigned to a chromosomal bin according to its location on the chromosome. Since information about centromere positions is unfortunately missing from the *Ensembl* annotations, we extracted the 'gap' table (containing centromeres' positions) from all genomes present in the UCSC public SQL database available at "genome-mysql.cse.ucsc.edu". This information was available for only six species.

Function of megasatellite-encoded proteins

Protein sequences were extracted from Ensembl (Yates et al., 2020) and compared to protein motifs listed in the Pfam database, a widely used resource for classifying protein sequences into families and domains (Mistry et al., 2021). When no hit was found, the classification retrieved from the PANTHER database was used instead (when available). The PANTHER classification system (http://www.pantherdb.org) is a comprehensive system that combines genomes, gene function classifications, pathways and statistical analysis tools (Mi et al., 2019).

QUANTIFICATION AND STATISTICAL ANALYSIS

For the first clustering, we used the log10 of the e-value (maximum 200) provided by HMMSEARCH as a proxy of the distance, as recommended by the author (van Dongen 2014). For the second clustering, families are based on the orthologous set of protein families provided by *Ensembl*. We did not recalculate these families, they were just used to fuse our own TR families (Figure 1). Note that any mistake in the first clustering will be subsequently fixed during the second clustering using the *Ensembl* families.

To determine the expected number of megasatellite families (ORTHO FAM) in each clade (Figure 4B), we first calculated the mean number of proteins per clade. Then, based on the observed number of ORTHO FAM in agnatha, we calculated an expected of ORTHO FAM in each clade. We compared these expected numbers to observed number using a Chi2 test. Similarly, based on the observed number of ORTHO FAM in primates, we calculated an expected of ORTHO FAM in each clade, and compared them to observed numbers with a Chi2 test (Table S8). This gave us upper and lower values of expected ORTHO FAM in each clade. A statistical increase was found after the agnatha and after the marsupiala (Figure 4B).

To determine whether a species contained a significant increase of megasatellites within a clade, we calculated the expected number of megasatellites in each species using proteome sizes. These expected numbers were compared to the observed numbers using a Chi2 test (Table S1).

For Zinc-finger proteins, expected numbers of ZFP were calculated for each clade, according to numbers observed in other clades and on the number of proteins in each clade. These expected numbers were compared to the observed numbers of ZFP in each clade, using a Chi2 test (Figure 7B).

For GO terms, the significance threshold was set at 0.05 and calculated using the g:SCS tailor-made algorithm of the g:Profiler database, using a correction for multiple testing (biit.cs.ut.ee/gprofiler/gost). All statistical tests were performed using 'R' (Millot 2011), and all of the statistical details can be found in STAR Methods.