Published in final edited form as:

Bioessays. 2007 March ; 29(3): 262–270.

The origins of polypeptide domains

Edward E. Schmidt^{1,2,*} and **Christopher J. Davies^{2,3}**

¹*Molecular Biosciences, Montana State University, Bozeman, MT.*

²*Center for Reproductive Biology, Washington State University, Pullman, WA.*

³*Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA.*

Summary

Three decades ago Gilbert posited that novel proteins arise by re-shuffling genomic sequences encoding polypeptide domains. Today, with numerous genomes and countless genes sequenced, it is well established that recombination of sequences encoding polypeptide domains plays a major role in protein evolution. There is, however, less evidence to suggest how the novel polypeptide domains, themselves, arise. Recent comparisons of genomes from closely related species have revealed numerous species-specific exons, supporting models of domain origin based on “exonization” of intron sequences. Also, a mechanism for the origin of novel polypeptide domains has been proposed based on analyses of insertion-based polymorphisms between orthologous genes across broad phylogenetic spectra and between allelic variants of genes within species. This review discusses these processes and how each might participate in the evolutionary emergence of novel polypeptide domains.

Introduction

One evolutionarily important process is the genesis of novel proteins.(1) Analyses of protein and genome sequences reveal the traces of two predominant mechanisms for creation of novel proteins: (1) gene, genomic region, or genome duplication followed by divergence(2-4) and (2) recombination of sequences encoding portions of existing proteins, such as by exon shuffling.(1,3,5–9) Gene duplication can produce a second copy of a gene in one generation; however, considerable subsequent mutation and genetic drift may be required before the two resultant genes become functionally distinct. Conversely, recombination can yield a functionally novel protein in a single event.(5,9,10) Recombination can be likened to junkyard creativity, where components salvaged from existing genes are assembled into a new useful product. But how do novel exons or novel polypeptide domains arise in the first place?

Definitions of terms

For this discussion, the term “polypeptide domain” refers to the recognizable evolutionary units or modules from which proteins are composed.(9) Some small proteins may consist of a single polypeptide domain but most proteins comprise multiple domains.(11,12) Importantly, sequences encoding a polypeptide domain can have an evolutionary history and genomic distribution that is distinct from any individual protein in which this domain is found. As modular building blocks, polypeptide domains will frequently also have some degree of structural or functional autonomy.(11,13) The term “oligopeptide” is used here to refer to stretches of amino acids that are too simple to determine whether they are homologous to other similar sequences. Thus, the occurrence of distinct regions of DNA encoding similar

*Correspondence to: Edward E. Schmidt, Molecular Biosciences, Montana State University, 960 Technology Blvd, Bozeman, MT 59717.
E-mail: eschmidt@montana.edu

oligopeptides could result from either shared ancestry or independent re-emergence. A third more-inclusive term, “motif”, is used to refer to both polypeptide domains and oligopeptides. All instances of the emergence of additional DNA sequences in a gene are referred to here as “insertions”. When repeats of a DNA sequence are found in proximity in a gene, we infer that an ancestral monomer of that sequence templated the more recently inserted copy, and therefore this class of insertions is termed “duplications”.

Genesis of novel polypeptide domains by “exonization” of intronic sequences

Recent comparisons of genomes from closely related species have revealed evolutionarily new exons.(14,15) Most of these are alternative exons that can be optionally included in an mRNA, but the ancestral mRNA lacking this exon can also still be produced (Fig. 1). The implication is that new exons can be assembled over time by accumulation of random mutations within introns of existing genes, that is, by “exonization” of intronic sequences.(15-17)

Mechanistic evidence of exonization, including evidence for piecemeal assembly of a novel protein-coding region and evidence for acquisition of splicing signals, comes from a study of the multicopy *Sdic* gene in *Drosophila* species.(18) *Sdic* arose very recently by a fusion of representatives of two other multicopy genes, *AnnX* and *Cdic*. The recent origin of *Sdic* and retention of copies of both parental genes allows alignment of both protein-coding and non-coding sequences. The novel first protein-coding exon of *Sdic* is homologous to intron 3 sequences of the *Cdic* gene, and arose by an exonization process that included insertion of a 10 bp splice-donor site.(18)

Exonization can also allow assimilation of adjacent intron sequences into an existing exon. Thus, by inactivation of a 5' splice donor site and activation or acquisition of a downstream site, an exon could be extended into more 3' sequences. Similarly, by intronic acquisition of 3' splice acceptor signals, sequences upstream of an existing exon could be assimilated into that exon. An example of the former case was very recently demonstrated for the primate *SETMAR* gene, in which a 27 bp deletion both eliminated the stop codon for the ancestral SET histone methyl transferase gene and created a 5' splice donor site. In higher primates, this new splice donor now splices to a novel 3' acceptor site, resulting in production of a unique mRNA encoding a fusion of SET and the transposase open reading frame of a proximal downstream mariner-family transposon.(19)

Because exonization can rapidly result in novel protein-coding sequences, it has received attention. Theoretically similarly prevalent, but to our knowledge not yet described, might be “intronization”, whereby mutation of splicing signals could lead to rapid deletion of protein-coding sequences from mRNAs. Exon expansion and contraction events such as these may contribute, in part, to the observed evolutionary plasticity of exon-bordering regions.(20)

Are there other origins for polypeptide domains? Insights from TBP

The TATA-binding protein, TBP, is a component of the basal transcription machinery involved in transcription initiation by all three eukaryotic RNA polymerases and by *Archaeobacterial* RNA pol.(21,22) Accordingly, the 181 amino acid C-terminal core of TBP (TBP_{CORE}) is found in life forms from archaea to mammals; however the N terminus is more variable (Fig. 2). The ~ 120–150 amino acid mammalian TBP N terminus(23) arose at the pre-vertebrate–vertebrate transition, roughly 540 million years ago, and is now found uniquely and universally in vertebrates.(24) Between this N-terminal region and the TBP_{CORE} is a (PXT)_n repeat (proline, X, threonine, repeated n times) that is shared among most metazoans, including pre-vertebrate taxa.

Most of the N-terminal domain in mammals lies in a single large exon (exon 3),(25,27) which initially suggested that this domain was “shuffled” onto the TBP_{CORE} of the protein at the pre-vertebrate–vertebrate transition as a single exon.(24) However, further analyses disfavored this model. Besides containing most of the vertebrate-specific region of the protein, exon 3 also encodes the 15 amino acid (PXT)_n region, which is shared with pre-vertebrate metazoans, and 8 universally conserved amino acids of the TBP_{CORE} (Fig. 2).(27) Also, a conserved part of the vertebrate-specific N terminus is encoded by exon 2.(26) Thus, two exons would have had to “shuffle” into the *tbp* locus at the vertebrate transition, the second of which would have had to obtain about 70 nucleotides of novel (PXT)_n and TBP_{CORE} amino acid-coding sequence to replace the equivalent sequence in the pre-vertebrate exon. The exon 2/3 and exon 3/4 junctions are in different phases (0 and 2, respectively), which is rare in shuffled exons.(28) Further arguing against exon shuffling, cyclostome vertebrates (lampreys and hagfishes, the most primitive living vertebrates) possess a clade-specific intron (phase 0), splitting the N-terminal protein-coding region into three exons.(24) Lastly, outside of a central glutamine (Q)-repeat sequence, the mammalian TBP N terminus bears no recognizable similarity to any other known genomic sequences in any species, raising the question of what the pre-existing exonic source of these sequences might have been. These subtle complexities in the primary and intron/exon structure of the vertebrate TBP N terminus suggested that the origins of this polypeptide domain were independent of the unit-exon.

Repeat-based mutations in the genome

Natural selection determines the relative reproductive success of individuals. For natural selection to discriminate between individuals in a population, the population must have sufficient genetic diversity to cause fitness differences between individuals. Thus, one should consider what types of genetic differences commonly exist between individuals within a population that might provide a spectrum of potential fitness advantages for niche adaptation. Not surprisingly, the most-common types of genetic variation between individuals are also the types most frequently used in genetic marker analyses. One of these is the number of small repeats at a specific genomic location, in particular repeats of 1–6 nucleotides in length, termed microsatellites.(29) Microsatellite repeat number polymorphisms have been estimated to be up to five orders of magnitude more common than point mutations between individuals;(3) roughly 8% of all microsatellite repeats in the human genome are protein-coding.(31) Thus, at least in some species, repeat-based polymorphisms are likely the greatest source of protein-coding diversity between individuals within a population.

Small sequence repeats can expand or contract in a single generation. For example, the (CAG)_n trinucleotide repeat in the *Huntingtin* gene can expand to severely pathological lengths in a single gamete.(32) Proposed mechanisms involve polymerase slipping during DNA replication and perhaps error-prone repair processes;(33–35) however, the parameters governing the lengths of the repeat units and the frequencies and extents of tandem duplication and deletion of genomic sequences are largely unknown.

Long before the phenotypic consequences of a point mutation can be subjected to natural selection, these errors are screened by replication-associated proofreading activities and generally corrected.(36) Conversely, the extent to which proofreading activities antagonize insertions and deletions of small sequences is unclear.(37,38) In *Drosophila melanogaster*, a species that generally carries only small numbers of repeats, disruption of the mismatch repair gene *Spellchecker1* leads to expansion of repeats, suggesting that a proofreading-like editing mechanism might prevent expansion of small repeats in this species.(39) However, in mammalian cells, disruption of mismatch repair function does not affect the lengths or stabilities of either endogenous or introduced repeats.(40,41) The relative abundance of repeat-

based polymorphisms in most species suggests that proofreading activities are at least more permissive of repeat-based errors.

Repeated motifs in proteins

Although perhaps the best-known protein repeats are encoded by microsatellite “trinucleotide repeats”, for example the (CAG)_n encoding the Q-repeat in Huntingtin protein,(32,34) repeated motifs are not limited to homo-polymerization of a single amino acid. Common repeat-units vary in length from one to hundreds of amino acids. Familiar examples include Q-, ankyrin-, BZIP-, collagen-, EGF-, immunoglobulin-, WD40-, protease-, protein kinase-, SH-, and fn1-repeats, and others.(10) Moreover, proteins often have semisymmetrical substructures, for example β -barrels, composed of multiple iterations of a common or somewhat diverged motif.

Advanced programs for recognizing repeated protein motifs indicate such repeats are far more common than previously thought. Ponting et al. analyzed 14,226 predicted proteins from *D. melanogaster* for repeated domains and detected 1,656 that contained at least one substantial internal repeat pair.(42) The repeated domains included 224 novel repeat “groups”, that is a domain in a protein that was shared with at least one other protein, and 455 repeat “orphans”, that is, an internal repeat for which the domain was not found in other proteins.(42) Ponting et al.'s study sought polypeptide repeats that would likely form autonomously structured domains and, therefore, repeats shorter than 30 amino acids were not further studied. Twenty-seven previously unrecognized repeat families were identified; many of these were shared with proteins in other phyla.(42) A more-recent study analyzed over 200,000 proteins in the predicted proteomes of 24 species across all phyla and reported that 13% of these proteins contained repeated domains.(43)

Repeat-based polymorphisms in simple oligopeptides

Polymorphisms in microsatellite-encoded amino acid repeats have traditionally been considered to have little potential as evolutionarily determinant mutations. Excessive expansion of Q-repeats in various otherwise unrelated human proteins results in similar neuropathologies;(34) however, no functional role for Q-repeats in normal proteins has been genetically established and we are aware of no proposal in which Q-repeat expansion would result in a fitness benefit or positive selection. Nevertheless, recent reports indicate that intragenic expansions or contractions of protein-coding repeats are correlated to vertebrate evolution of amelogenin, a protein that participates in tooth development,(44) and to the diverse and rapidly achieved phenotypic differences that have been selected between breeds of domestic dogs.(45) These studies show that small sequence duplications and deletions are frequent events that might commonly modify the activities of existing proteins.(46,47) Importantly, the polymorphisms reported to date are simple microsatellite-based repeats that alter the length of homo-amino acid stretches in their respective proteins. In the case of amelogenin, these subtle modifications likely alter the higher-order assembly of protein/hydroxyapatite complexes that determine the physical characteristics of tooth enamel.(44) In dogs, several polymorphisms have been mapped to regulatory transcription factors—the implication being that these mutations alter downstream gene regulation.(45,46)

Additional evidence that single-generation sequence insertions and deletions might substantially alter protein function comes from analyses on chordate *thp* genes. Importantly, some of these studies suggest that these modifications can alter the lengths of both homopolymeric and more complex oligopeptide regions, and that, even in the case of homopolymeric repeats, the causal mechanism can involve duplication/deletion of longer-than-trinucleotide sequence elements. A clinical case report by Shatunov et al. described a patient exhibiting severe neuropathology in which the Q-repeat region of one of the patient's *thp* genes had expanded from encoding 37 to encoding 55 consecutive Qs.(48) Subsequent analyses of

the patient's relatives indicated that this mutation had occurred de novo in production of one of the gametes that fused to yield the patient. Interestingly, the Q-repeat in the *thp* gene, unlike those in many other poly-Q genes, is composed of both Q-encoding codons (CAG and CAA; the *Huntingtin* poly-Q tract is purely CAG). Although other studies have reported Q-repeat variations in the *thp* gene between individuals,(24,49) Shatunov et al. were able to use this codon heterogeneity and familial Q-region sequence determinations to show that expansion of the TBP Q-repeat in this patient did not result from consecutive trinucleotide-expansion steps, but rather, resulted from single-step germline tandem duplication of the 54-base pair sequence CAACAGCAA(CAG)₁₅.(48) Could similar multi-codon duplicative processes also contribute to production of novel complex polypeptide domains?

Evolution of TBP suggests insertion-based genesis of novel complex polypeptide domains

For investigations into the phylogeny of the TBP N terminus (see above), one of us (EES) isolated multiple TBP clones each from amphioxus (*Branchiostomata floridae*), Pacific hagfish (*Eptatretus stoutii*), and Atlantic sea lamprey (*Petromyzon marinus*).⁽²⁴⁾ Comparisons of clones from multiple individuals of these species revealed interesting polymorphisms that suggested an alternative model for genesis of the N terminus.

Sequences of multiple *thp* clones from each of these species revealed almost no intra-species point mutations, synonymous or non-synonymous, protein-coding or non-coding. Conversely, many *thp* cDNAs differed by insertions or deletions of small sequence elements. For example, hagfish clones encoded 11, 12, 13, 14 and 15 Q residues in the repeat region.⁽²⁴⁾ Hagfish also had one more PXT repeat than other vertebrates. However, the greatest insights came from the sequence polymorphisms outside the Q- and (PXT)_n-repeat regions in the *thp* genes of hagfish and amphioxus.

The hagfish TBP N terminus exhibited insertions of two oligopeptides not found in any other vertebrate, one of 7- and one of 9-amino acids.⁽²⁴⁾ Amphioxus has a long TBP N terminus that lacks a Q-repeat and cannot be aligned with the vertebrate N terminus. TBP clones from amphioxus differed by insertion of a six nucleotide N-terminal protein-coding element, which was a tandem repeat of adjacent sequences.⁽²⁴⁾ Moreover, three different isoforms of the 3'-untranslated (UTR) region of amphioxus TBP cDNAs were found; each differed by multiple insertions or deletions of small sequences.⁽²⁴⁾ In other words, although point mutations, including synonymous mutations, were rare, small sequence insertions and deletions, including some that altered protein-coding, were common.

All insertion-based differences in protein-coding sequences necessarily retained the correct reading frame (multiples of three nucleotides) such that the downstream TBP_{CORE} region would still be translated; however in the 3' UTR, small sequence insertions and deletions were even more common and were of arbitrary sizes (4, 7, 8, 18, and 23 nucleotides).⁽²⁴⁾ Thus, unlike point mutations that are constrained by proofreading and natural selection, our data suggested a much more predominant role for natural selection (i.e. maintenance of protein production/function) in constraining insertions and deletions.

Finally, although amino acid alignments show that all vertebrate N termini are homologous through a common ancestor that existed at the pre-vertebrate-vertebrate transition,⁽²⁴⁾ the inferred progression from this ancestor to higher vertebrates is intriguing. The short complex oligopeptide-encoding insertions in hagfish (see above) are tandem duplications of adjacent sequences (Fig. 2), lending emphasis to the repetitive ancestry of the domain.⁽²⁴⁾ The hagfish TBP N terminus exhibits a strikingly repetitive character both because it possesses these additional iterations of two repeat units as compared to other vertebrates and because the

tandem repeat units are more similar to each other (Fig. 2).(24) Since diverging from an ancestor shared with hagfish, the amino acid substitutions that have accumulated in the lineage leading to higher vertebrates result in the domain in modern tetrapods, outside of the poly-Q region, retaining little evidence of its historical repetitive character.

Based on these observations, it was posited that the TBP N terminus evolved de novo by insertions, deletions and divergence of small sequence elements.(24) At the same time, an equivalent model was proposed as a general mechanism for the evolution of intrinsically unstructured proteins.(50) By this model, one or more small regions within the protein-coding portion of an existing gene could undergo multiple duplications and deletions, eventually resulting in a novel polypeptide sequence that, like the TBP N terminus in primitive vertebrates, (1) would have a repetitive character, and (2) may appear unrelated to the ancestral version. Because a functional protein serves as the starting template, successful modifications should be in sequences that are more permissive of structural alterations, while more constrained domains are conserved. For example, we predict that the complex oligopeptide insertions in the hagfish N terminus (see above) are, on their own, functional subdomains that are at least partially redundant with adjacent copies of the repeat unit (unpublished). As such, their duplication would not alter protein function, but rather, would provide a redundant copy, which could subsequently undergo functional drift while the parental copies of the domain retain the ancestral function.

Nested between the highly polymorphic N terminus-encoding region and the highly polymorphic 3' UTR region of the *thp* locus lie exons encoding one of nature's most-immutable sequences: the TBP_{CORE}.(23) Estimates of purifying selection against amino acid substitutions in the TBP_{CORE} of vertebrates based on rates of non-synonymous to synonymous nucleotide substitutions (d_N/d_S ; loci not under selection should have a d_N/d_S value of 1.0) indicate this is an extremely highly conserved protein domain ($d_N/d_S = 0.006$, for comparison vertebrate histone H2B $d_N/d_S = 0.07$). (24) Thus, *thp* is not simply a hot locus for accumulation of insertion-based mutations. The close juxtaposition of the highly invariant TBP_{CORE} between two highly polymorphic sequences poses a striking paradox that more likely reflects a focused action of natural selection on the TBP_{CORE} than a fickle targeting of mutagenesis to flanking sequences.

Small sequence insertion-based polymorphisms in other proteins

Small insertion-based polymorphisms are common and widely dispersed;(29,30,34) however there are few reports of the insertion of small sequence elements, either by duplication or other means, contributing to the successful genesis of novel polypeptide domains.(24,42,43,50) To determine whether insertion-based modification of functional domains was common, we examined the genes of the major histocompatibility complex (MHC).

MHC genes are multi-gene families in higher vertebrates that participate in adaptive immune defenses. MHC class I (MHC-I) proteins are expressed on the surface of almost all nucleated cells and present proteasome-generated fragments of all proteins being made in a cell, including cancer antigens, viral proteins, or other intracellular pathogen-associated proteins, to circulating cytotoxic Tcells.(51) Each MHC-I protein will preferentially present protein fragments (antigens) with different physical and chemical properties, so diversity in MHC-I genes allows better defense against a broader range of intracellular pathogens.(52,53)

In studying the correlation between pathogen resistance and MHC-I gene diversity in cattle and bison, one of us (CJD) recognized a novel allele of a classical MHC-I molecule (*Bibi-N*00501*; NCBI accession #DQ989113) in bison that had a direct adjacent seven amino acid sequence duplication in the antigen-binding region of the protein that was not found in other alleles (Fig. 3; Table 1). Further analyses identified the parent allele (*Bibi-N*00502*; NCBI

accession #DQ989114) in the bison population. Alleles *Bibi-N*00502* and *Bibi-N*00501* are identical except for the 21 nucleotide duplication in *Bibi-N*00501*. The novel allele is prevalent in the population, suggesting that it is under selective pressure, likely due to its augmenting resistance to an endemic pathogen (unpublished).

We searched the more exhaustive collection of mouse MHC-I alleles available in NCBI and found other examples of haplotype diversity arising from small sequence insertions, either as a result of proximal duplication or from some other unidentified source (Table 1). These observations suggest that small insertion-based mutagenesis is an important participant in providing the sequence diversity that pathogens select at MHC-I loci.

Larger duplications

The documentation of expansion of the Q-repeat in a human TBP protein occurring not via multiple consecutive trinucleotide duplications, but rather as a single duplication of a 54 bp sequence(48) (discussed above), demonstrates that sequence duplications are not limited to microsatellite-sized sequences. Thus, it is reasonable to propose that larger semi-symmetrical protein motifs, such as β -barrels, could be born as duplications followed by repeat divergence.

Compelling evidence of duplication and deletion of large and complex protein-coding regions can be inferred from the distribution of one of Ponting et al.'s novel repeats.(42) A duplicated polypeptide domain (repeat unit ~60 amino acids) in the fly Timeless-2 (TIM-2) protein(42) is also found as a duplicated domain in the rat TIM protein. However, the fly TIM protein contains only a monomer of this domain, and the round worm ortholog of TIM-2 contains three diverged tandem iterations of the domain.(42) The difference in numbers of diverged repeat units within this family of proteins indicate that the mechanisms of altering the numbers of repeats in TIM proteins must occur at a greater frequency than the gene duplicative mechanisms that led to the paralogous gene family.

Bjorklund et al. addressed the possibility that intragenic domain repeats might arise by exon shuffling.(43) Their analysis of over 5,000 domain repeat families across all phyla indicated that, only in the case of a few extracellular repeat domains in higher eukaryotes, most notably the immunoglobulin- and EGF-repeat families, is the intron/exon structure expected to be compatible with an exon shuffling-based mechanism for repeat expansion.(43) Thus, even in the case of large or complex polypeptide domains, like the TIM domain discussed above,(42) a likely mechanism of altering their numbers within a gene would be by single-step tandem duplications of the repeat unit in the germ-line.

Conclusions

Recent studies provide support for the decades-old prediction that novel polypeptide domains can arise by exonization of intron sequences.(6) The abundance of novel exons not shared between closely related species suggests that this is a common source of unique protein-coding sequence.(14,15)

Studies on simple and complex oligopeptide insertions and deletions in proteins reveal an alternative means of creating a novel polypeptide domain within an existing protein-coding sequence. Perhaps the most-abundant type of genetic variability, both protein-coding and non-coding, between individuals within populations results from insertions and deletions of small sequence elements.(30) The lengths of protein-coding tri-nucleotide repeat regions have long been known to differ between individuals, occasionally with pathological consequences. Recently, species-specific differences in tooth structure and physiological differences between dog breeds have been correlated to alterations in the lengths of homo-amino acid repeats in

specific proteins.(44,45) This suggests that, besides being either neutral or deleterious, some modifications of small repeat elements can result in positive (Darwinian) selection.(46,47)

Analysis of one pathological incident of a Q-repeat expansion in a human family showed that homo-amino acid repeat expansion can occur by single-step duplications of complex larger-than-trinucleotide sequences.(48) The same mechanism might duplicate other complex nucleic acid sequences. Repetitive substructures are common in proteins, suggesting that sequences encoding the compositional motifs might be re-used repeatedly in the genesis of a protein. These observations lend insights into understanding the origins of more complex repetitive domains and substructures in proteins. They also suggest that this is a highly active process that has the potential to rapidly and drastically alter protein-coding sequences within existing genes, resulting in the appearance of entirely novel polypeptide domains.

Acknowledgments

The authors thank O. Lucas, J. Langeland, Y.-s. Piao, A. Bondareva, and J. Prigge for their comments on the manuscript. We thank Drs. Hong Li and Donald L. Traul for permission to include data from the USDA, Animal Disease Research Unit, Bison Malignant Catarrhal Fever Research project (USDA, ARS grant #CWU-5348-32000-018-00D).

Funding agency: E.E.S. was supported by a new investigator award (#5-FY00-520) and a research grant (#6-FY03-61) from the March of Dimes Foundation, a career development award (#0446536) and a research grant (#0090884) from the National Science Foundation, and a research grant (#1R01-AI55739) from the National Institutes of Health. C.J.D. was supported by the Washington State University, Safe Food Initiative and USDA Animal Health Formula Funds.

Abbreviations

bp, base pairs; C, carboxyl; EGF, epidermal growth factor; MHC, major histocompatibility complex; MHC-I, MHC Class I; N, amino; Q, glutamine; RNA pol, RNA polymerase; TBP, TATA-binding protein; TBP_{CORE}, the widely conserved 181 amino acid TBP C-terminal region; TIM, Timeless protein; UTR, untranslated region.

References

1. Gilbert W. Why genes in pieces? *Nature* 1978;271:501. [PubMed: 622185]
2. Greer JM, Puetz J, Thomas KR, Capecchi MR. Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature* 2000;403:661–665. [PubMed: 10688203]
3. Li WH, Gu Z, Cavalcanti AR, Nekrutenko A. Detection of gene duplications and block duplications in eukaryotic genomes. *J Struct Funct Genomics* 2003;3:27–34. [PubMed: 12836682]
4. Hoegg S, Meyer A. Hox clusters as models for vertebrate genome evolution. *Trends Genet* 2005;21:421–424. [PubMed: 15967537]
5. de Souza SJ, Long M, Gilbert W. Introns and gene evolution. *Genes Cells* 1996;1:493–505. [PubMed: 9078380]
6. Gilbert W, de Souza SJ, Long M. Origin of genes. *Proc Natl Acad Sci USA* 1997;94:7698–7703. [PubMed: 9223251]
7. Long M, de Souza SJ, Rosenberg C, Gilbert W. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc Natl Acad Sci USA* 1996;93:7727–7731. [PubMed: 8755543]
8. Liu M, Grigoriev A. Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet* 2004;20:399–403. [PubMed: 15313546]
9. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. *Science* 2003;300:1701–1703. [PubMed: 12805536]
10. Tordai H, Nagy A, Farkas K, Banyai L, Patthy L. Modules, multidomain proteins and organismic complexity. *Febs J* 2005;272:5064–5078. [PubMed: 16176277]
11. Han JH, Kerrison N, Chothia C, Teichmann SA. Divergence of interdomain geometry in two-domain proteins. *Structure* 2006;14:935–945. [PubMed: 16698554]

12. Teichmann SA, Park J, Chothia C. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci USA* 1998;95:14658–14663. [PubMed: 9843945]
13. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 2004;14:208–216. [PubMed: 15093836]
14. Nekrutenko A. Identification of novel exons from rat-mouse comparisons. *J Mol Evol* 2004;59:703–708. [PubMed: 15693626]
15. Wang W, Zheng H, Yang S, Yu H, Li J, et al. Origin and evolution of new exons in rodents. *Genome Res* 2005;15:1258–1264. [PubMed: 16109974]
16. Gilbert W. Genes-in-pieces revisited. *Science* 1985;228:823–824. [PubMed: 4001923]
17. Kondrashov FA, Koonin EV. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet* 2003;19:115–119. [PubMed: 12615001]
18. Ranz JM, Ponce AR, Hartl DL, Nurminsky D. Origin and evolution of a new gene expressed in the *Drosophila* sperm axoneme. *Genetica* 2003;118:233–244. [PubMed: 12868612]
19. Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci USA* 2006;103:8101–8106. [PubMed: 16672366]
20. Liu M, Walch H, Wu S, Grigoriev A. Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids Res* 2005;33:95–105. [PubMed: 15640447]
21. DeDecker BS, O'Brien R, Fleming PJ, Geiger JH, Jackson SP, et al. The crystal structure of a hyperthermophilic archaeal TATA-box binding protein. *J Mol Biol* 1996;264:1072–1084. [PubMed: 9000631]
22. Hernandez N. TBP, a universal eukaryotic transcription factor? *Genes Dev* 1993;7:1291–1308. [PubMed: 8330735]
23. Tamura T, Sumita K, Fujino I, Aoyama A, Horikoshi M, et al. Striking homology of the 'variable' N-terminal as well as the 'conserved core' domains of the mouse and human TATA-factors (TFIID). *Nucleic Acids Res* 1991;19:3861–3865. [PubMed: 1861978]
24. Bondareva AA, Schmidt EE. Early vertebrate evolution of the TATA-binding protein, TBP. *Mol Biol Evol* 2003;20:1932–1939. [PubMed: 12885957]
25. Ohbayashi T, Schmidt EE, Makino Y, Kishimoto T, Nabeshima Y, et al. Promoter structure of the mouse TATA-binding protein (TBP) gene. *Biochem Biophys Res Commun* 1996;225:275–280. [PubMed: 8769130]
26. Schmidt EE, Ohbayashi T, Makino Y, Tamura T, Schibler U. Spermatid-specific overexpression of the TATA-binding protein gene involves recruitment of two potent testis-specific promoters. *J Biol Chem* 1997;272:5326–5334. [PubMed: 9030607]
27. Sumita K, Makino Y, Katoh K, Kishimoto T, Muramatsu M, et al. Structure of a mammalian TBP (TATA-binding protein) gene: isolation of the mouse TBP genome. *Nucleic Acids Res* 1993;21:2769. [PubMed: 8332475]
28. Kaessmann H, Zollner S, Nekrutenko A, Li WH. Signatures of domain shuffling in the human genome. *Genome Res* 2002;12:1642–1650. [PubMed: 12421750]
29. Bennett P. Demystified. microsatellites. *Mol Pathol* 2000;53:177–183. [PubMed: 11040939]
30. Ellegren H. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* 2000;16:551–558. [PubMed: 11102705]
31. Wren JD, Forgacs E, Fondon JW 3rd, Pertsemliadis A, Cheng SY, et al. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* 2000;67:345–356. [PubMed: 10889045]
32. Pearson CE. Slipping while sleeping? Trinucleotide repeat expansions in germ cells. *Trends Mol Med* 2003;9:490–495. [PubMed: 14604827]
33. Lahue RS, Slater DL. DNA repair and trinucleotide repeat instability. *Front Biosci* 2003;8:s653–s665. [PubMed: 12700078]
34. Parniewski P, Staczek P. Molecular mechanisms of TRS instability. *Adv Exp Med Biol* 2002;516:1–25. [PubMed: 12611433]

35. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 1987;4:203–221. [PubMed: 3328815]
36. Kunkel TA, Erie DA. DNA mismatch repair. *Annu Rev Biochem* 2005;74:681–710. [PubMed: 15952900]
37. Ellegren H. Mismatch repair and mutational bias in microsatellite DNA. *Trends Genet* 2002;18:552. [PubMed: 12414181]
38. Sainudiin R, Durrett RT, Aquadro CF, Nielsen R. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* 2004;168:383–395. [PubMed: 15454551]
39. Harr B, Todorova J, Schlotterer C. Mismatch repair-driven mutational bias in *D. melanogaster*. *Mol Cell* 2002;10:199–205. [PubMed: 12150919]
40. Tomlinson IP, Hampson R, Karran P, Bodmer WF. DNA mismatch repair in lymphoblastoid cells from hereditary non-polyposis colorectal cancer (HNPCC) patients is normal under conditions of rapid cell division and increased mutational load. *Mutat Res* 1997;383:177–182. [PubMed: 9088350]
41. Yamada NA, Smith GA, Castro A, Roques CN, Boyer JC, et al. Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. *Mutat Res* 2002;499:213–225. [PubMed: 11827714]
42. Ponting CP, Mott R, Bork P, Copley RR. Novel protein domains and repeats in *Drosophila melanogaster*: insights into structure, function, and evolution. *Genome Res* 2001;11:1996–2008. [PubMed: 11731489]
43. Bjorklund AK, Ekman D, Elofsson A. Expansion of Protein Domain Repeats. *PLoS Comput Biol* 2006;2
44. Delgado S, Girondot M, Sire JY. Molecular evolution of amelogenin in mammals. *J Mol Evol* 2005;60:12–30. [PubMed: 15696365]
45. Fondon JW 3rd, Garner HR. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci USA* 2004;101:18058–18063. [PubMed: 15596718]
46. Caburet S, Cocquet J, Vaiman D, Veitia RA. Coding repeats and evolutionary “agility”. *Bioessays* 2005;27:581–587. [PubMed: 15892112]
47. Sire JY, Delgado S, Fromentin D, Girondot M. Amelogenin: lessons from evolution. *Arch Oral Biol* 2005;50:205–212. [PubMed: 15721151]
48. Shatunov A, Fridman EA, Pagan FI, Leib J, Singleton A, et al. Small de novo duplication in the repeat region of the TATA-box-binding protein gene manifest with a phenotype similar to variant Creutzfeldt-Jakob disease. *Clin Genet* 2004;66:496–501. [PubMed: 15521976]
49. van Roon-Mom WM, Reid SJ, Faull RL, Snell RG. TATA-binding protein in neurodegenerative disease. *Neuroscience* 2005;133:863–872. [PubMed: 15916858]
50. Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 2003;25:847–855. [PubMed: 12938174]
51. Margulies, D. The major histocompatibility complex. In: Paul, W., editor. *Fundamental Immunology*. Fourth. Lippincott-Raven; Philadelphia: 1999. p. 263-285.
52. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK. The nature of selection on the major histocompatibility complex. *Crit Rev Immunol* 1997;17:179–224. [PubMed: 9094452]
53. Wegner KM, Kalbe M, Schaschl H, Reusch TB. Parasites and individual major histocompatibility complex diversity—an optimal choice? *Microbes Infect* 2004;6:1110–1116. [PubMed: 15380781]
54. Nikolov DB, Hu SH, Lin J, Gasch A, Hoffmann A, et al. Crystal structure of TFIID TATA-box binding protein. *Nature* 1992;360:40–46. [PubMed: 1436073]
55. Ellis SA, Morrison WI, MacHugh ND, Birch J, Burrells A, et al. Serological and molecular diversity in the cattle MHC class I region. *Immunogenetics* 2005;57:601–606. [PubMed: 16086171]

Figure 1.

Genesis of novel polypeptide domains by “exonization” of intron sequences. At top is shown cartoon of a simple gene (exons green, introns uncolored, simplified splice donor and acceptor signals diagramed in blue and red, respectively, promoter sequences in yellow, initiation site as a bent arrow, polyadenylation signals as orange, and splicing pattern by bent lines). Below, acquisition of splice donor and acceptor signals within an existing intron can generate a novel exon (labeled “N”); however, the phase of the exon–exon junctions must be preserved by this exon. Alternative splicing can produce either the original protein or the modified version with the novel polypeptide. This novel exon might subsequently shuffle to other parts of the genome.

Figure 2.

Exon- and primary amino acid-structures of vertebrate TBP. At top is depicted the protein-coding exon arrangement of higher vertebrate and cyclostome TBP mRNAs.(27) Heavy lines depict the relevant region of each mRNA, triangles indicate exon–exon junctions, numbers are exon designations. For cyclostomes, the exon structure has only been determined for the N-terminal region; the C-terminal region is labeled “???”. A cyclostome-specific intron divides the region homologous to higher vertebrate exon 3 into separate left and right exons (designated 3L and 3R).(24) Below is the primary amino acid structure of vertebrate TBP with subregions and repeat units emphasized. Below this is the linear distribution of phylogenetic histories of each region along the sequence. Note that boundaries in phylogenetic history do not correlate to exon boundaries. The TBP_{CORE} domain is encoded by an inverted repeat (arrows) in all phyla, resulting in the conserved symmetrical “saddle” structure of the protein;(21,54) other recognized repeat units are tandem.(24) At the bottom is a key of the various oligopeptide units recognizable in hagfish TBP, indicating the number of repeats of each unit seen in human and hagfish TBPs.(24) Due to mutation and genetic drift, some repeat units are less obvious in higher vertebrates.(24) The two regions having no recognizable pattern (*) are dissimilar to each other and are not obviously derived by duplication and deletion of adjacent sequences; however, like the rest of the N terminus, these regions might have arisen by insertion/deletion of sequences whose ancestry is not recognized. Q repeat lengths are given as number of Q-codons (**) because, even though this region can expand and contract as larger multi-codon units, this has only been documented once.(48) Normal human *tbp* genes have 36–38 Qs in the repeat region; however patients with SCA17 neuropathologies have severely expanded TBP Q-domains.(48,49)

Figure 3.

Alignment of predicted amino acid sequences of the alpha-2 domain of two bison MHC-I alleles. Allele *Bibi-N*00501* is a novel allele with a seven amino acid duplication following position 163 of the parent allele, *Bibi-N*00502*. Numbering corresponds to amino acid positions of normal bovine or bison MHC-I proteins starting from the first amino acid of the mature protein. Dots in the *Bibi-N*00501* sequence indicate amino acid identity with *Bibi-N*00502*. The donor sequence is highlighted in green and the duplicated peptide in allele *Bibi-N*00501* is highlighted in blue. The symbols above the *Bibi-N*00502* sequence indicate amino acids of typical MHC-I proteins (e.g. *Bibi-N*00502*) that are predicted to contact either the bound oligopeptide or the T-cell receptor: * predicted to contact the peptide, + predicted to contact the T-cell ~ receptor, predicted to contact both.(55) The seven amino acid duplication in *Bibi-N*00501* is in the middle of the α -helix that contacts with both the peptide antigen and the T-cell receptor.

Table 1
Small sequence insertions in MHC-I family genes(¹)

Identification(²)	Insertion unit(³)	Context(⁴)
XM_001003869 (mouse)	GTGTCCCGGCCCGGC	family
		new(⁵)
NM_010390 BC080812 (mouse)	AGTCATTGGAGCTGTGGT	family
XM_001003852 XM_001003837 XM_001003842 XM_001003826 XM_001003817 (mouse)	CATGGCTCC	new
XR_005037 NM_010380 AY496687 (mouse)	AGCCTCCTC	family
DQ989113 (bison)	AGGAACTACCTGGAGGGCACG	new
XM_903939 (mouse)	GATCCTTTCCTATCTGTG	Novel insertion, ancestral template not obvious.
NM_010394 AY989883 (mouse)	GGGGCAAGG	Novel insertion, ancestral template not obvious.
numerous (mouse)	TGTTTCTGGAGGCCTTCAAAG or GCCCTGTGTCCCTCCATAG(⁶)	Novel insertion, ancestral template not obvious.

¹ Fifty-four mouse MHC-I family cDNA sequences or forty-two bison and cattle MHC-I family cDNAs were aligned using ClustalW. Regions of insertion in one or a few genes of each family were examined for similarity to adjacent or nearby sequences. Although numerous other repeated sequences can be found in MHC-I family genes, duplications shared among all alleles and genes within a species predate paralogous gene duplication events, whereas insertions of additional copies in a subset of genes or alleles demonstrates duplication as a recently occurring mechanism of inter-allelic diversity.

² NCBI accession numbers.

³ Although some point mutation-based degeneracy is found between some repeat units, for simplicity, it is not shown.

⁴ Context is diagrammatic representation of repeats in majority of gene family or allele types in each species.

⁵ In mouse XM_001003869, the duplicated region is separated from the ancestral region by 30 bp of conserved non-duplicated sequence. This 30 bp sequence shows evidence of increased evolutionary plasticity, in that it is partially deleted in XM_001003817 and XM_001003837, and it is completely absent in AY989843, AY989839, AY989845, AY989822, AY989854, AY989853, XM_903941, and XM_903942.

⁶ These are likely older insertions because roughly 1/3 of the alleles analyzed had no insertion, 1/3 had a version of the upper insertion, and 1/3 had a version of the lower insertion.