Evolutionarily Conserved Noncoding DNA

Anil G Jegga, *Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA*
Bruce J Aronow, *Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA*

The availability of increasingly complete collections of genomic sequence data from evolutionarily separated species has created an outstanding opportunity to identify and examine the functions served by DNA sequences that have been highly conserved through evolution. These conserved sequence domains occur widely throughout the genomes of higher organisms.

Advanced article

Article contents

doi:10.1002/9780470015902.a0006126

Introduction

After decades of ignominy as 'junk', 'parasite' or 'selfish', noncoding deoxyribonucleic acid (DNA) present in introns and intergenic regions has recently been recognized as symbolizing a rich fossil record of DNA sequences subject to variation and conservation throughout evolution. In addition to clusters of evolutionarily conserved genes in close proximity to one another, human chromosomes also harbor stretches of noncoding DNA that has withstood the effects of evolution with much less variability in their underlying DNA sequence than is observed in the majority of non-protein-encoding regions of the genome. Although there is still no definitive count of human genes, a pertinent question is, why are humans much more complex organisms than the worm or fly, when the number of human genes is only several-fold greater those of the worm or fly? Some of the clues to this puzzle have been suggested through transgenic and classical genetic analyses. These studies have identified highly complex gene regulatory regions in multiple domains within and surrounding the genes whose expression pattern they determine.

Gene orthologs (a pair of genes encoding equivalent protein products in evolutionarily separated organisms) show a high level of conservation within the protein-coding exons as well as in the organization of exons in chromosomal DNA. However, what about similarities in noncoding regions of the genes? Since human, mouse and other mammals shared a common ancestor approximately 80 million years ago, the genomes of all mammals are similar. For example, syntenic regions of chromosomes contain many of the same genes arranged in virtually identical order over millions of base pairs. The protein-coding regions of the mouse and human genomes are about 85% identical. This represents an overall estimate of how conserved other regions might be if a similar functional requirement for conservation were acting on a particular sequence domain. In contrast, the noncoding regions are much less similar at an overall level (about 50%). However, there are many islands of noncoding DNA sequences, located within similar relative positions in orthologous genes, that can be well over 90% identical. Thus, when we encounter highly conserved segments in noncoding regions of DNA in species separated by many millions of years, it implies that these sequences are responsible for some crucial function. A comparison of the noncoding regions of the same DNA sequences from humans and other species easily detects these domains. A variety of experimental approaches, including transgenic, knockout and classical genetic studies, have proven that conserved noncoding sequences frequently harbor critical regulatory information responsible for gene or chromosomal activity.

Noncoding DNA includes all sequences within the genome except exon sequences that encode protein-reading frames. Noncoding sequences thus harbor a wide variety of genomic features and elements (**Table 1**). A popular and misleading metaphor from earlier concepts of the genome was that barren deserts of intergenic regions were junk and bereft of meaning. We now recognize these regions as highly prolific research terrains of features that are critical for the maintenance and function of complex genomes. We also know that hidden in these genetic neighborhoods are keys to a better understanding of elements that are essential for gene and chromosomal function. Studying these regions will yield valuable clues that would otherwise lay undetected or too complicated to initially track down through functional assays. With this improved insight into these important functional elements of the genome, we are in a better position to dissect their precise contributions and the mechanisms by which they function. Ironically, it is this junk DNA that has helped scientists to come to terms with one of the human genome's most mystifying paradoxes, the C-value paradox: the lack of correspondence between

Table 1 Structural and functional classification of genomic DNA sequence features

Structural and functional features of genes
5′ and 3′ flanking
Promoter
Exonic 5′ and 3′ untranslated regions
Exonic protein-coding
Splicing: 5′ and 3′ splice sites, splicing enhancer, other regulatory
Transcription control elements: enhancer, repressor, coregulator/modifier, attenuator/pausing
Structural and functional features of chromosomal DNA
Matrix/scaffold attachment region
Origin of replication
Centromere
Telomere
CpG island
Nucleosome phasing elements
Pseudogene/gene fragment
Repetitive elements
Unique or low-copy number repetitive element
Moderate to highly abundant repetitive element
Simple repeat expansions (e.g. single base, doublet, triplet)
Tandemly repeated or clustered repeats
Satellite DNA, minisatellite, microsatellite, megasatellite DNA
Interspersed repetitive elements
Retroposons
SINEs (short interspersed elements): *Alu*, MIR (mammalian interspersed repeats)
LINEs (long interspersed elements): LINE1, LINE2
RLEs (retrovirus-like elements): HERVs (human endogenous retroviruses): MaLRs (mammalian apparent LTR-retrotransposons), others
DNA transposons: mariner, others
Unclassified elements

genome size and biological complexity. Our genome is 200 times larger than that of yeast, but 200 times smaller than that of amoeba!

Pseudogenes

Pseudogenes are DNA sequences that are closely related to functional genes but are incapable of encoding proteins as a result of deletions, insertions and nonsense mutations that abolish the reading frame or otherwise prevent gene expression. They are detected as the inactive orthologs of their still-active counterparts in the genomes of lower vertebrates. There are two major types of pseudogenes. The first arises through the duplication and subsequent inactivation of a gene (pseudogenes in the α- and β-globin clusters). The second type contains only the exons of the parental gene, dispersed randomly in the genome. Pseudogenes are relatively common in the human genome and may be especially prevalent in multigene families like β-globin, actin, interferons, keratins, T-cell receptors and immunoglobulin gene clusters. However, single-copy genes may also have multiple pseudogenes: *PHB* (*prohibitin*), four pseudogenes; *ASS* (*argininosuccinate synthetase*), 14 pseudogenes (Cooper, 1999).

Are pseudogenes conserved?

A considerable number of pseudogenes, similar in sequence as well as in positional relationship to other genes, have been reported in humans, chimpanzee and other mammals. The best-known example of a shared pseudogene is the psi eta-globin pseudogene, a member of the β-globin gene family. One question, considering what pseudogenes are, is why do they have any tendency to be conserved whatsoever? The case of pseudogenes is reminiscent of the history of vestigial organs, in which an apparent lack of function actually indicated a lack of knowledge about the function. There is still much about pseudogenes that is not well understood. Pseudogenes are not evolutionary dead ends. They may influence the evolution of other functionally significant sequences by mediating recombination events or acting as sequence donors in gene

conversion (Cooper, 1999). Interestingly, there have also been reports of natural reactivation of pseudogenes in humans (*CRYGEP1* (*crystalline γ-E pseudogene 1*) and *HBZP* (*hemoglobin, ζ-pseudogene*)).

Pseudoexons

Pseudoexons are gene regions that are functional in a gene from one species but have been conserved in a nonfunctional form in another species. The presence of pseudoexons in a number of human genes has so far provided evidence for the occurrence of exon inactivation during mammalian evolution. The frequency of occurrence of pseudoexons in the human genome is usually difficult to estimate, because they can be detected only through a careful comparison of a pair of orthologous genes. Several instances of pseudoexons have been reported (*CRYAA* (*crystalline, α-A*), *GYPB* (*glycophorin B*), *TKT* (*transketolase*), *CR2* (*complement component receptor 2*), *AMAC* (*acyl-malonyl condensing enzyme*), *PHEX* (*phosphate regulating gene with homologies to endopeptidases on the X chromosome*), *KIR2DL3* (*killer cell immunoglobulin receptor, two domains, long cytoplasmic tail 3*) (Cooper, 1999).

Introns and Evolution

Introns are interruptions in the coding sequences of the genes of multicellular animals. They can vary in size from as small as 24 base pairs (bp), as in the case of the *parvalbumin* (*PVALB*) gene, to more than 600 kilobases (kb), as in the case of the *collagen, type V, α-1* (*COL5A1*) gene. Although introns are largely, but not necessarily, confined to the eukaryotes, a close correlation between intron density and developmental complexity is not wholly implausible. Evolutionarily conserved introns have been shown to be important in regulating gene expression by harboring gene enhancers or gene silencers. After the upstream promoter regions, they are the second most important sites of predilection for gene regulatory keys that control tissue-specific expression. For instance, one of the first tissue-specific enhancers identified, the immunoglobulin κ-enhancer, was at first distinguished as a highly conserved region within an intron.

Introns-early versus introns-late theory

When *Fischerella*, a photosynthetic cyanobacterium, was found to have introns, the argument about the appearance of introns began.

Introns-early theory

As per the introns-early theory, or the exon theory of genes, exons started as minigenes before being assembled to make whole genes at a later stage of evolution. Introns were the 'functionless' pieces that held these exons together. If genes evolved this way, then each exon or minigene would encode a unit of protein. The earlier assumption that each exon encodes a separate protein domain is not always true. Introns sometimes occur in the middle of a domain. The breaks could even be within a single codon. This is particularly detrimental to the exon theory of genes, because any minigenes that started in the middle of a codon would be disturbed by the frameshift error.

Introns-late theory

According to this theory, introns probably originated to avoid the problem of the arbitrary distribution of stop codons in random primeval sequences. However, the striking similarity among the introns found in species that diverged too long ago makes us think the contrary.

One of the principal differences between orthologous genes from different species at different evolutionary strata is often the number of introns they possess. Usually, as we go up the evolutionary hierarchy, there is an increase in the number of introns. For instance, the version of a gene possessed by lower eukaryotes such as yeast has no introns, while the versions in more highly organized creatures such as humans have multiple introns. The number of introns could be related to the number of times a gene has been transported across a species boundary during evolution.

Untranslated Regions

Untranslated regions (UTRs) are regions at either end (5′ or 3′) of a mature transcript (preceding the initiation codon and following the stop codon respectively) that are not translated into a protein. The earliest-discovered highly conserved region (HCR) reported has been conserved at least since the echinoderm/chordate divergence. A comparison of nucleotide sequences from different classes of vertebrates that diverged more than 300 million years ago not only revealed the distribution of highly conserved noncoding regions within genes, but also showed that functional constraints are generally much stronger in 3′ noncoding regions than in promoters or introns. The 3′ HCRs are particularly rich in adenine (A) and thymine (T) and are always located in the transcribed UTRs of genes, which suggests that they are involved in posttranscriptional processes. In other words, since the HCRs occur relatively infrequently within the introns, the evolutionary constraints would seem to operate at the level of the mature messenger ribonucleic acid (mRNA). A comparative study of 77 orthologous mouse and human gene pairs revealed

that, of the noncoding regions, 3′ UTRs were most conserved. The longest conserved element covers nearly 2000 bp in the 3′ UTR of the Δ-EF1 transcriptional repressor among chicken, mouse, hamster and human (Duret *et al*., 1993; Jareborg *et al*., 1999). The 5′ and 3′ UTRs show sequence identities of 67% and 69% respectively (Makalowski *et al*., 1996).

Repetitive Elements

With the notable exception of amoeba, the amount of noncoding DNA we have accumulated in our genome far surpasses that collected by any of our early evolutionary cousins. The human genome has a greater percentage (50%) of repetitive DNA than the mustard weed (11%), the worm (7%) or the fly (3%). Unexpectedly, however, there seems to have been a significant decrease in transpositional activity over the past 50 million years. Equally surprising is the fact that there seems to be no such reduction in the incidence of repeats in rodents. The extinct or near-extinct repeats called DNA transposons and long-terminal repeat (LTR) retroposons respectively remain active in the mouse genome. This remarkable disparity in the activity of such a sizeable portion of two mammalian genomes is intriguing, conceivably pointing to fundamental forces in the evolution of these different mammals. This contrast also suggests that the extinction or near extinction of these repeat elements may be accounted for by some basic differences between hominids and rodents.

The repeats have amended the genome by reorganizing it, creating entirely new genes and revising and rearranging the existing genes. Contrary to the earlier common assumption that insertion of repetitive elements into genes would impair the protein's ability to function, a surprisingly large number of these elements are found in translated proteins also. The repetitive elements seem to insert into noncoding regions of a gene and be incorporated into protein through alternative splicing. Since the elements contain splicing sites, new proteins may be created as a result of the shuffling, elongation or truncation of coding regions of the old gene. Thus, the location and distribution of the human repetitive elements may provide insight into their role in gene evolution and species differentiation.

Repeat elements in the human genome are, broadly, of four categories: the extinct type (DNA transposons), the near-extinct type (the LTR retroposons) and the still active type (long interspersed (LINE) elements and short interspersed (SINE) elements). Striking differences in the distribution of repetitive DNA are seen across the human genome. At one extreme are *HOX* gene clusters, which contain less than 2% interspersed repeats within 100 kb; by contrast, a 525-kb region of chromosome Xp11 has a repeat density of 89%. Distribution of these elements by GC content (or gene-rich neighborhoods) seems to defy any logic. The majority of the repeat elements end up in less desirable regions of the genome that are AT-rich and guanine/cytosine (GC)-poor, whereas SINE elements seem to have settled in the gene-rich regions of the genome. It has been surmised that either the SINEs somehow trick their way into the GC-rich regions, or that most SINEs land in GC-poor regions to begin with, and evolution favors the SINEs that happened to land in GC-rich regions. The latter conjecture was further strengthened by studies of *Alu* repeats, the most conspicuous human SINEs. Young *Alu* elements dwell in the AT-rich regions, whereas older *Alu* elements tend to move up to the GC-rich regions. Thus, the hypothesis that evolution tends to place SINEs near genes sounds true.

Molecular studies of the human genome sequence aided in bringing to light remnants of an ancient exodus that occurred within our primitive vertebrate ancestors. These ancestors, with few defense systems against invading parasites, became easy targets for bacteria to take residence inside the vertebrate host. During the cohabitation of host and parasite, ancient genes were presumably exchanged between the two. Scientists conjecture that the genes may have been left behind by the bacterial invaders or transported into the genome by viral intermediaries. It also cannot be ruled out that the bacteria possibly 'stole' genes from the vertebrate ancestors.

Tandem repeats

Widespread throughout the human genome, tandemly repeated DNA sequences show sufficient variability among individuals in a population that they have become important in several fields comprising genetic mapping, linkage analysis and human identity testing. Tandemly repeated regions of DNA are classified into several groups based on the size of the repeat region. Minisatellites (variable number of tandem repeats, VNTRs) have core repeats with 9–80 bp, while microsatellites (short tandem repeats, STRs) contain repeats of 2–5 bp.

Minisatellite and microsatellite DNA

Minisatellites are tandemly repeated, highly variable DNA sequences often prone to instability. They are reported to be involved in recombination and transcriptional and translational control of gene expression. The minisatellite located in intron 9 of the human *glucose phosphate isomerase* (*GPI*) gene is present in seven other species of mammals and also chicken. Telomeric minisatellite DNA is conserved throughout

the vertebrates, probably because of selection pressure to ensure continued recognition by the telomerase enzyme.

Microsatellites, the tandem repeats of very simple motifs, are found frequently throughout eukaryotic genomes. They are repetitive sequences of one to six nucleotide patterns that can be used as genetic markers for a wide range of applications, from genome mapping to forensic testing to population studies. They can also serve as regulatory elements, and some are conserved at orthologous positions in the genomes of different species. For example, consider the $(TCAT)_n$ repeat element in the fifth intron of the human *tyrosine hydroxylase* (*TH*) gene. This repeat is similar to the consensus thyroid response element (TRE) present in the human and rat *TH* genes. Several examples of apparent conservation of intragenic microsatellites located within orthologs in human, mouse and rat have been reported (Stallings, 1995).

Interspersed repeats

These are repeated DNA sequences located at dispersed regions in a genome and are also known as mobile elements or transposable elements. A stretch of DNA sequence may be copied to a different location through DNA recombination. Following several generations, such repeat units could spread over to several regions. First discovered by Barbara McClintock in the 1940s from the studies of corn, and found in all organisms, the most common mobile elements in mammals are SINEs and LINEs.

SINEs

SINEs are ubiquitous in mammalian genomes. Remarkable forms of these repeats among placental orders indicate that most of them augmented in each lineage separately, subsequent to mammalian radiation. SINEs, reportedly present in all vertebrates and also mollusks include various types of DNA transposable elements, such as tiggers and mariners, which are characterized by the possession of terminal inverted repeats and target site duplications. The tigger element closely resembles pogo, a DNA transposon in *Drosophila melanogaster*.

Alu

The *Alu* family of SINEs are present in all primates. *Alu* repeats are reported to have evolved as processed pseudogenes from the transcripts of the *RNA, 7SL, cytoplasmic* (*RN7SL*) gene and are specific to primates. The B1 repeat in mouse, which also generated from an RNA 7SL-like gene, is an *Alu* counterpart. During their existence for the last 50–100 million years, *Alu* sequences have contributed to the function of many useful genes. As a source of mutation and variation, they have strongly influenced primate evolution.

MIRs

Mammalian interspersed repeats (MIRs) are an ancient family of repeats whose sequence divergence and common occurrence among placental mammals, marsupials and monotremes represent a 'fossilized' record of a major genetic event preceding the radiation of placental orders (Jurka *et al*., 1995). The high divergence, and their presence at orthologous sites in different mammals, indicates that MIRs, at least in part, amplified before the mammalian radiation. Next to *Alu* repeats, MIRs are the most common interspersed repeat in primates, with an estimated 300 000 copies still discernible, accounting for 1–2% of our DNA. Interestingly, a small, central region of MIR appears to be much better conserved in the genomic copies than is the rest of the sequence (Smit and Riggs, 1995).

Mariners

Mariners are small (about 1.3 kb), DNA-mediated transposable elements with 25–30 bp of inverted terminal repeat sequences. They have been shown to be extremely widespread throughout the metazoa, present in organisms as diverse as humans and coelenterates.

Regulatory Regions

Of the noncoding proportion of the human genome, an indeterminate fraction has a decisive role in regulating gene expression. It is widely appreciated that comparisons among genome sequences are key to identifying functional regions of noncoding DNA by virtue of the conservation of their primary sequences. Furthermore, analysis of noncoding regions with high percentage of identity has shown that they are also frequently conserved in other mammals and unique within the human genome, two common features of long-range regulatory elements.

Seeking out 'phylogenetic footprints', clusters of invariant or slowly changing positions in the aligned sequences of related but divergent organisms, has now become a standard approach to examine those DNA segments flanking and interrupting the coding regions. Phylogenetic footprints have been defined as noncoding sequence motifs that show 100% conservation in several species over a region of six or more contiguous base pairs (Gumucio *et al*., 1996).

Exploring and analyzing the phylogenetic footprints of regulatory elements has been fruitful

Table 2 Examples of comparative analysis of human genomic sequence facilitating identification of regulatory regions

Gene(s)	Species	Reference
Immunoglobulin heavy locus	Human versus mouse	Ravetch *et al*. (1980)
T cell receptor complex	Human versus mouse	Hood *et al*. (1995)
Globin genes	Human versus primate	Gumucio *et al*. (1996)
Adensosine deaminase	Human versus mouse	Brickner *et al*. (1999)
BTK (Bruton's tyrosine kinase)	Human versus mouse	Oeltjen *et al*. (1997)
Human 12p13	Human versus mouse	Ansari-Lari *et al*. (1998)
α-Globin cluster	Human versus mouse versus chicken versus pufferfish	Flint *et al*. (2001)

Table 3 Tools on the web for identifying regulatory regions in genomic sequences using the comparative sequence analysis approach

Name	Web address	Reference
Cister	http://zlab.bu.edu/mfrith/cister.shtml	Boston University, USA
rVISTA	http://pga.lbl.gov/rvista.html	Loots *et al*. (2002)
Theater	http://www.hgmp.mrc.ac.uk/Registered/Webapp/theatre	Human Genome Mapping Project Resource Centre, UK
Trafac	http://trafac.chmcc.org	Jegga *et al*. (2002)

(**Table 2**). Sparking off the postulation that sequence conservation involves functional constraint is an old concept in the theory of molecular evolution: the rates of substitution may vary among sites, depending on constraint. Tools (**Table 3**) have been developed taking into account the conserved noncoding sequences, and initial results in the identification of regulatory regions have been promising. When comparing genomic orthologs, looking for conserved *cis* elements in the context of sequence similarity makes the overall sequence space to be searched highly manageable. And, in investigations of regulatory regions, any approach that substantially reduces the size of the sequence space to be searched can be very valuable.

Strong evolutionary conservation of promoter and other transcription control regions

Sequence conservation in promoter regions is a good indicator of functional importance and is thus used as a reliable guide to locate *cis*-acting regulatory elements that bind the *trans*-acting transcription factors. Phylogenetic footprinting of promoter regions has been successful in identifying several regulatory regions in several human and mouse genes. For instance, in comparing the promoter regions of the gene *SRY* in 10 different mammalian species, a total of 10 regulatory elements have been identified with apparent differences in the presence or absence of specific binding sites, spacing of these sites, relative location and orientation (Margarit *et al*., 1998). However, the regulatory elements need not necessarily always be conserved between orthologs. For example, the 83-bp intron 1 of human *CD68* contains a macrophage-specific enhancer, while the corresponding intronic region of the mouse ortholog does not, even though there is about 80% sequence similarity (Greaves *et al*., 1998). Orthologous promoters also differ with respect to the presence or absence of specific *cis*- elements or, alternatively, in terms of the number of such *cis*- elements. This may lead to the different expression profile for an orthologous gene, as in the case of *GPR1*, which is hippocampus-specific in humans but not in rats (Marchese *et al*., 1994).

MARs/SARs

Matrix (MARs) or scaffold attachment (SARs) regions are the regions of the nuclear matrix where the chromatin is attached. The organization of chromatin with respect to the nuclear scaffold is thought to determine the chromosome architecture in terms of its functional domains, which in turn influences gene activity. MARs do not usually share extensive sequence homology, but often comprise 200 bp of AT-rich DNA. MARs also appear to be preferentially associated with enhancer-type elements. A *cis*-acting regulatory element 3′ of the gene *HBG1*, known to be associated with the nuclear matrix, has been shown to bind specifically to an AT-rich binding protein (SATB1) that binds to MARs (Cunningham *et al*., 1994). However, this region is not conserved between human and mouse.

See also

mRNA Untranslated Regions (UTRs)
Long Interspersed Nuclear Elements (LINEs): Evolution

Pseudogenes and their Evolution
Retrosequences and Evolution of Alu Elements
Short Interspersed Elements (SINEs)

References

Ansari-Lari MA, Oeltjen JC, Schwartz S, *et al*. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Research* **8**: 29–40.

Brickner AG, Koop BF, Aronow BJ and Wiginton DA (1999) Genomic sequence comparison of the human and mouse adenosine deaminase gene regions. *Mammalian Genome* **10**: 95–101.

Cooper DN (1999) Pseudogenes and their formation. In: Cooper DN *Human Gene Evolution*, pp. 265–296. Oxford: BIOS Scientific.

Cunningham JM, Purucker ME, Jane SM, *et al*. (1994) The regulatory element 3′ to the A γ-globin gene binds to the nuclear matrix and interacts with special A-T-rich binding protein 1 (SATB1), an SAR/MAR-associating region DNA binding protein. *Blood* **84**: 1298–1308.

Duret L, Dorkeld F and Gautier C (1993) Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Research* **21**: 2315–2322.

Flint J, Tufarelli C, Peden J, *et al*. (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α-globin cluster. *Human Molecular Genetics* **10**: 371–382.

Greaves DR, Quinn CM, Seldin MF and Gordon S (1998) Functional comparison of the murine macrosialin and human CD68 promoters in macrophage and nonmacrophage cell lines. *Genomics* **54**: 165–168.

Gumucio DL, Shelton DA, Zhu W, *et al*. (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes. *Molecular Phylogenetics and Evolution* **5**: 18–32.

Hood L, Rowen L and Koop BF (1995) Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. *Annals of the New York Academy of Sciences* **30**: 390–412.

Jareborg N, Birney E and Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research* **9**: 815–824.

Jegga AG, Sherwood SP, Carman JW, *et al*. (2002) Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Research* **12**(9): 1408–1417.

Jurka J, Zietkiewicz E and Labuda D (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Research* **23**: 170–175.

Loots GG, Ovcharenko I, Pachter L, Dubchak I and Rubin EM (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research* **12**: 832–839.

Makalowski W, Zhang J and Boguski M (1996) Comparative analysis of the 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Research* **6**: 846–857.

Marchese A, Cheng R, Lee MC, *et al*. (1994) Mapping studies of two G protein-coupled receptor genes: an amino acid difference may confer a functional variation between a human and rodent receptor. *Biochemical and Biophysical Research Communications* **205**: 1952–1958.

Margarit E, Guillen A, Rebordosa C, *et al*. (1998) Identification of conserved potentially regulatory sequences of the *SRY* gene from 10 different species of mammals. *Biochemical and Biophysical Research Communications* **245**: 370–377.

Oeltjen JC, Malley TM, Muzny DM, *et al*. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Research* **7**: 315–329.

Ravetch JV, Kirsch IR and Leder P (1980) Evolutionary approach to the question of immunoglobulin heavy chain switching: evidence from cloned human and mouse genes. *Proceedings of the National Academy of Sciences of the United States of America* **77**: 6734–6738.

Smit AF and Riggs AD (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Research* **23**: 98–102.

Stallings RL (1995) Conservation and evolution of (CT)*n*/(GA)*n* microsatellite sequences at orthologous positions in diverse mammalian genomes. *Genomics* **25**: 107–113.

Further Reading

Clark MS (1999) Comparative genomics: the key to understanding the human genome project. *BioEssays* **21**: 121–130.

Cooper DN (1999) *Human Gene Evolution*. Oxford: BIOS Scientific.

Stojanovic N, Florea L, Riemer C, *et al*. (1999) Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research* **27**: 3899–3910.

Web Links

glucose phosphate isomerase (*GPI*); LocusID: 2821.LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 2821

hemoglobin, γ A (*HBG1*); LocusID: 3047.LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 3047

RNA, 7SL, cytoplasmic (*RN7SL*); LocusID: 6029.LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 6029

tyrosine hydroxylase (*TH*); LocusID: 7054.LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 7054

glucose phosphate isomerase (*GPI*); MIM number: 172400.OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?172400

hemoglobin, γ A (*HBG1*); MIM number: 142200.OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?142200

tyrosine hydroxylase (*TH*); MIM number: 191290.OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?191290