A3855

ENCYCLOPEDIA OF LIFE SCIENCES

December 1999

Genome Organization: Human

Advanced

Structures and Processes

Genetics and Molecular Biology

human genome#genome organization#DNA sequence complexity#gene families#chromosomes

Kass, David H

David H Kass

Eastern Michigan University, Ypsilanti, Michigan, USA

Batzer, Mark A

Mark A Batzer

Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA

The human nuclear genome is a highly complex arrangement of two sets of 23 chromosomes, or DNA molecules. There are various types of DNA sequences and chromosomal arrangements, including single-copy protein-encoding genes, repetitive sequences and spacer DNA.

Introduction

The human nuclear genome contains about 3000 million base pairs (bp) of DNA, of which only an estimated 1.5–2% possess protein-encoding sequences. As shown in Figure 1, the DNA sequences of the eukaryotic genome can be classified into several types, including single-copy protein-encoding genes, DNA that is present in more than one copy (repetitive sequences) and intergenic (spacer) DNA. The most complex of these are the repetitive sequences, some of which are functional and some of which are without function. Functional repetitive sequences are classified into dispersed and/or tandemly repeated gene families that either encode proteins (and may include noncoding pseudogenes) or are important noncoding transcribed sequences such as the ribosomal RNA genes. Repetitive sequences with no known function include the various highly repeated satellite families, and the dispersed, moderately repeated transposable element families. The remainder of the genome consists of spacer DNA, which is simply a broad category of undefined DNA sequences.

The human nuclear genome consists of 23 pairs of chromosomes, or 46 DNA molecules, of differing sizes (Table 1).

Table 1 DNA content of chromosomes, extrapolated from Stephens *et al*. (1990), based on the length of chromosome as a percentage of the total length of the genome

Chromosome	Amount of DNA (Mb)	Chromosome	Amount of DNA (Mb)
1	249	13	108
2	237	14	105
3	192	15	99
4	183	16	84
5	174	17	81
6	165	18	75
7	153	19	69
8	135	20	63
9	132	21	54
10	132	22	57
11	132	X	141
12	123	Y	60

Sequence Complexity

The human genome contains various levels of complexity as demonstrated by reassociation kinetics. This involves the random shearing of DNA into small fragments averaging about 500 bp, heat denaturation to separate the strands of the double helix, and slow cooling. During cooling, complementary sequences anneal; the more copies there are of a particular sequence, the greater the chance of finding a complement to anneal to. Therefore, the reassociation is dependent on time (t), as well as the initial concentration of that sequence (C_0) yielding what is referred to as a C_0t value. Analysis of the human genome estimates that 60% of the DNA is either single copy or in very low copies; 30% of the DNA is moderately repetitive; and 10% is considered highly repetitive.

Various staining techniques demonstrate alternative banding patterns of mitotic chromosomes referred to as karyograms. Although the three broad classes of DNA are scattered throughout the chromosome, chromosomal banding patterns reflect levels of compartmentalization of the DNA. Using the C-banding technique yields dark-staining regions of the chromosome (or C bands), referred to as heterochromatin. These regions are highly coiled, contain highly repetitive DNA and are typically found at the centromeres, telomeres and on the Y chromosome. They are composed of long arrays of tandem repeats and therefore some may contain a nucleotide composition that differs significantly from the remainder of the genome (approximately 40–42% GC). This means that they can be separated from the bulk of the genome by buoyant density (caesium chloride) gradient centrifugation. Gradient centrifugation results in a major band and three minor bands referred to as satellite bands, hence the term satellite DNA.

The G-banding technique yields a pattern of alternating light and dark bands reflecting variations in base composition, time of replication, chromatin conformation and the density of genes and repetitive sequences. Therefore, the karyograms define chromosomal organization and allow for identification of the different chromosomes. The darker bands, or G bands, are comparatively more condensed, more AT-rich, less gene-rich and replicate later than the DNA within the pale bands, which correspond to the R bands by an alternative staining technique. More recently these alternative banding patterns have been correlated to the level of compaction of scaffold-attachment regions (SARs).

The human genome may also be compartmentalized into large segments of DNA with distinctive GC richness referred to as 'GC content domains' (Lander *et al*., 2001). There is a distinct association between GC-richness and gene density. This is consistent with the association of most genes with CpG islands, the 500–1000-bp GC-rich segments flanking (usually at the 5' end) most housekeeping and many tissue-specific genes. The clustering of CpG islands, as demonstrated by fluorescence *in situ* hybridization (Craig and Bickmore, 1994), further depicts gene-poor and gene-rich chromosomal segments.

Additionally, there are five human chromosomes (13, 14, 15, 21, 22) distinguished at their terminus by a thin bridge with rounded ends referred to as chromosomal satellites. These contain repeats of genes coding for rRNA and ribosomal proteins that coalesce to form the nucleolus and are known as the nucleolar organizing regions.

Single-copy Sequences

Although originally defined as a functional unit of heredity, a gene may be defined as an expressed segment of DNA containing transcriptional regulatory sequences. Venter *et al*. (2001) estimated that there are 26 588 protein-encoding transcripts with an additional 12 000 computationally derived genes. This is consistent with the 30 000–40 000 estimate of the international human genome sequencing consortium (Lander *et al*., 2001). In addition, over 700 noncoding RNA genes have been identified with over 5000 related genes, of which most are pseudogenes (see below).

The proportion of the genome consisting of genes would be estimated at 15–20% assuming an average gene size of 15 kb. However, approximately 90% of the DNA from protein-encoding genes are noncoding, including upstream and downstream regulatory sequences, and introns. Therefore, only 1.5–2% (45–60 Mb) of DNA has coding function. Regulatory sequences include common promoters recognized by transcription factors located at specific upstream distances from the transcription start site. These sequences include the TATA, CCAATT and GC boxes. There are also tissue-specific promoter sequences. Enhancers and silencers are *cis*-acting elements that function in various orientations and locations within or near a gene and that upregulate and downregulate gene expression, respectively. Many coding sequences may be included as members of gene families as described below. In addition, there may be single-copy sequences in the spacer DNA with no known (determinable) function.

Repetitive Sequences

In the human genome various sized stretches of DNA sequences exist in variable copy numbers. These repetitive sequences may be in a tandem orientation and/or dispersed throughout the genome. Repetitive sequences may be classified by function, dispersal patterns and sequence relatedness. Satellite DNA typically refers to highly repetitive sequences with no known function; gene families are DNA sequences, with at least one functional gene, related by sequence homology and/or function; and interspersed repeat sequences are typically the products of transposable element integrations, but may include retropseudogenes of a functional gene.

Macrosatellites, Minisatellites and Microsatellites

Macrosatellites are very long arrays, up to hundreds of kilobases, of tandemly repeated DNA. The three satellite bands observed by buoyant density centrifugation represent sections of the human genome containing highly repeated DNA that in effect alter the proportion of nucleotides from the rest of the genome. However, not all satellite sequences are resolved by density gradient centrifugation. Alpha satellite DNA or alphoid DNA constitutes the bulk of centromeric heterochromatin on all chromosomes. The interchromosomal divergence of the alpha satellite families allows the different chromosomes to be distinguished by fluorescence *in situ* hybridization (FISH).

Minisatellites are tandemly repeated sequences of DNA, yielding a total length from less than 1 kbp to 15 kbp. One subset of minisatellites comprises the highly polymorphic arrays of short tandem repeats with no known function that serve as useful DNA markers referred to as variable number tandem repeats (VNTRs). These sequences generally contain 1–5 kbp of DNA of repeating units of 15–100 nucleotides. Several minisatellites share enough sequence homology to be analysed by a single probe yielding DNA fingerprints. An example is a 10–15-bp core sequence of myoglobin minisatellites, which includes an almost invariant core sequence (GGGCAGGANG) among several polymorphic VNTR loci.

Telomeric DNA sequences contain another subset of minisatellites. The telomeric sequences contain 10–15 kb of hexanucleotide repeats, most commonly TTAGGG in the human genome, at the termini of the chromosomes. These sequences are added by telomerase to ensure complete replication of the chromosome. Telomeres of somatic cells are generally shorter than in germ cells, illustrated by their decreasing size within human B cells and skin cells with increasing age. In humans, it has been postulated that telomeric loss is associated with ageing and tumorigenesis.

Microsatellites are small arrays of short simple tandem repeats, primarily 4 bp or less. Different arrays are found dispersed throughout the genome, although dinucleotide CA/TG repeats are most common, yielding 0.5% of the genome. Runs of As and Ts are common as well. Microsatellites have no known functions. However, CA/TG dinucleotide pairs can form the Z-DNA conformation *in vitro*, which is possibly indicative of function. Repeat unit copy number variation of microsatellites apparently occurs by replication slippage yielding highly polymorphic DNA markers referred to as short tandem repeat polymorphisms (STRPs). STRPs are commonly used in commercial DNA fingerprinting kits. The expansion of trinucleotide repeats within

genes has been associated with genetic disorders such as Huntington disease, myotonic muscular dystrophy, Friedreich ataxia and fragile-X syndrome.

Gene Families

Gene families generally consist of a set of genes with high sequence homology over their entire length, primarily in the exons for protein-encoding gene families. Members of gene families, or possibly separate clusters of the same gene family, are considered paralogous, and are derived from an ancestral gene or locus by duplication, and are therefore evolutionarily and functionally related. The duplication of a gene, however, may yield a nonfunctional pseudogene (see below). Additionally, there are genes yielding products with weak overall sequence homology, but that are homologous at functionally conserved domains or short amino acid motifs, collectively forming an additional type of gene family. A group of genes with functionally and structurally related products with weak sequence homologies and lacking conserved amino acid motifs may be referred to as a gene superfamily. Only a limited number of examples will be discussed.

Gene families with essentially identical products

If the cell warrants numerous proteins or RNA molecules, one solution might be the production of multiple functional copies of a gene. The human genome, and eukaryotic genomes in general, have amplified a number of genes whose products are responsible for general purpose functions such as DNA replication and protein synthesis.

Histone genes are highly conserved among eukaryotes and have a fundamental role in chromatin structure. The histone family consists of five genes that tend to be linked, although in differing arrays of variable copy numbers dispersed in the human genome. The individual genes of a particular histone family encode essentially identical products (i.e. H4 genes yield the same H4 protein). Analysis of individual human genomic clones has identified isolated histone genes (e.g. H4), clusters of two or more histone genes, or clusters of all histone genes (e.g. H3-H4-H1-H3-H2A-H2B) (Hentschel and Birnstiel, 1981). A majority of histone genes form a large cluster on human chromosome 6 (6p21.3) and a small cluster at 1q21. Additionally, histone genes lack introns; a rare feature for eukaryotic genes.

Genes that encode ribosomal RNA (rRNA), inclusive of the spacer units, total about 0.4% of the DNA in the human genome. The individual genes of a particular rRNA family are essentially identical. The 28S, 5.8S and 18S rRNA genes are clustered with spacer units (ETS (external transcribed spacer), ITS (internal transcribed spacer)), in tandem arrays of approximately 60 copies each yielding about 2 Mbp of DNA. These clusters are present on the short arms of five acrocentric chromosomes and form the nucleolar organizing regions, hence approximately 300 copies. These three rRNA genes are transcribed as a single unit (yielding 41S rRNA) and then cleaved. 5S rRNA genes are clustered on chromosome 1q.

There are an estimated 30 human transfer RNA (tRNA) genes. tRNA genes and their pseudogenes are dispersed on at least seven chromosomes (McBride *et al*., 1989). In

addition, tRNA genes have been found in various clusters, i.e. cloned genomic fragments have been isolated containing several tRNA genes. Dispersal of tRNA pseudogenes may have occurred by RNA-mediated retroposition (McBride *et al*., 1989). This is consistent with the postulation that various SINE families (see below) have been derived from tRNA genes (Deininger and Batzer, 1993).

Small nuclear RNA (snRNA) molecules are thought to function in RNA processing. There are six families of related snRNA genes, termed U1 to U6, that are dispersed among the chromosomes. However, differing cluster patterns have been observed for these genes on different chromosomes. For example 35–100 functional U1 genes, all sharing 20 kb of nearly identical 5' and 3' flanking sequences, are loosely clustered in chromosome 1p36, and contain over 44 kb of intergenic sequences, whereas 10–20 U2 genes are clustered in a tight, virtually perfect 6-kb repeat unit on 17q21–q22 (Lindgren *et al*., 1985). In addition, more than one subfamily of a U snRNA has been identified; U3 comprises at least two subfamilies, which differ in the flanking sequences. Also, pseudogenes of snRNA have been identified and are thought to be dispersed in the genome by retroposition. tRNA genes are also found clustered with U RNA genes; for example chromosome band 1p36 contains 15–30 copies each of U1, Glu-tRNA and Asn-tRNA genes (van der Drift *et al*., 1994).

Gene families with high sequence homologies

There are numerous families of genes in the human genome sharing extensive intrafamily homology. These are generally dispersed, but may contain linked members. One of the most comprehensively studied gene families is the haemoglobin family. Human haemoglobin is a tetrameric protein consisting of two α-globin and two β-globin subunits. There are several possible polypeptides constructing the haemoglobin molecule with differing physiological properties and ontological regulation. This probably occurred as a result of gene duplication allowing for divergence of sequences for procuring new function. The two globin families exist as clusters of genes and pseudogenes on separate chromosomes. The α-globin gene cluster is on human chromosome 16 and the β-globin cluster is on 11. Although related in sequence, there is greater intra- than intercluster homology. Therefore, intracluster duplications postdate the duplication of the ancestral gene yielding a- and β-globin. The ontological regulation is apparently coordinated on each cluster by upstream sequences, providing the expression of the gene best suited for the oxygen need (a fetus, for example, exists in a relatively hypoxic environment). Predating haemoglobin divergence is the divergence of haemoglobin and myoglobin from an ancestral gene. Myoglobin is a monomeric protein encoded by a single gene on human chromosome 22 and stores oxygen in muscle, whereas haemoglobin is the oxygen carrier in blood.

Proto-oncogenes are also gene family members. These genes contribute to neoplasia when their expression (sequence or level) is altered. The gene products have normal cellular functions such as secreted growth factors (e.g. *Wnt* gene family), cell surface receptors (e.g. *erb*B gene family), intracellular signal transducers (e.g. *ras* gene family) and DNA-binding proteins (e.g. *myc* gene family). The *Wnt* gene family consists of at least 15 structurally related genes functioning in various aspects of growth and differentiation. They contain an N-terminal secretory signal peptide, a

short domain of low sequence conservation, and a highly conserved block (ranging from 40 to 95%) of about 300 amino acids with highly conserved motifs, and conservation of spacing of 22 cysteine residues. The *Wnt* genes map to different chromosomes, some demonstrating conservation of synteny (Bergstein *et al*., 1997).

There are four *erb*B genes. These are epidermal growth factor receptors (EGFR) grouped, as are other receptor tyrosine kinases, into a family based on the sequence homology of their kinase domains, their structure and the structural similarity of their ligands. The *myc* genes are members of the basic helix–loop–helix family of transcription factors. Functional members, including c-*myc*, L-*myc* and N-*myc*, are not linked genetically, with the latter two demonstrating more restricted patterns of expression, but they share a three exon, two intron structure. A detailed sequence analysis of *myc* genes suggests that the progenitor of the N-*myc* and L-*myc* genes was a duplicated c-*myc* gene (Atchley and Fitch, 1995). The *ras* genes represent a subfamily of guanosine triphosphate (GTP)-binding proteins, found dispersed in the genome. N-*ras*, H-*ras* and K-*ras* are closely related genes encoding for a $p21^{ras}$ product. Additional members of this family include TC21 and R-*ras*.

There are many other examples of multigene families in the human genome and some of these are listed in Table 2.

Table 2 Examples of interspersed multigene families in the human genome

Gene	Number of functional genes	Estimated number of pseudogenes
Actin	4	>16
Aldolase	3	2
Arginosuccinate synthetase	1	14
β-Tubulin	2	15–20
Cytochrome *c*	2	20–30
Ferritin heavy chain	1	>14
Glyceraldehyde-3-phosphate dehydrogenase	1	25
Ribosomal protein L32	1	20
Triose-phosphate isomerase	1	5–6

Gene families with low sequence homology but functionally conserved domains

Some sequences in the human genome share highly conserved amino acid domains with weak overall homologies. These often have developmental function. There are nine dispersed paired box (*Pax*) genes that contain highly conserved DNA-binding domains with six α helices. The homeobox or *Hox* genes share a common 60 amino

acid encoding sequence. In humans there are four *Hox* gene clusters, on different chromosomes. However, the individual genes in the cluster demonstrate greater homology to a counterpart gene on another cluster than to the other genes on the same cluster.

Gene families with different products but conserved short amino acid motifs

Some genes are considered families based not on entire-length sequence homology, but on conserved short amino acid motifs. DEAD box genes encode products with RNA helicase activity, and share eight short amino acid motifs, including the DEAD box (Asp-Glu-Ala-Asp). However, there are other gene families with conserved amino acid motifs, such as the WD box, that provide different functions. The WD box genes are characterized by between four and eight tandem repeats of a core sequence of fixed length terminating in a WD dipeptide.

Gene Superfamilies

DNA sequences that yield functionally and structurally related products with weak sequence homology and lacking significantly conserved amino acid motifs may be grouped as a gene superfamily. However, different families of genes may comprise a superfamily. Genes of the immunoglobulin superfamily encode proteins that form dimers consisting of extracellular variable domains at the N-terminus and constant domains at the C-terminus. Members of the immunoglobulin superfamily include immunoglobulin, human leucocyte antigen (HLA), T-cell receptor (TCR), T4 and T8 genes. Another example includes three superfamilies of growth factor receptors: (1) proteins with a core structure of seven transmembrane α-helical sequences; (2) large glycoproteins generally possessing a single transmembrane sequence and tyrosine kinase activity (includes the EGFR/*erb*B family described above); and (3) single transmembrane proteins lacking kinase activity.

Transposable Elements

The human genome contains interspersed repeat sequences that have largely amplified in copy number by movement throughout the genome. These sequences are referred to as transposable elements. Almost all transposition has occurred via an RNA intermediate yielding classes of sequences referred to as retrotransposons or retroposons (Figure 2). However, there is also evidence of an ancient DNA-mediated transposon family (*pogo*) in the human genome (Robertson, 1996).

Short and long interspersed DNA elements (SINEs and LINEs, respectively) are the primary families of transposable elements in the human genome. These are referred to as retroposons, since they lack the long terminal repeats (LTRs) of retroviruses, and are amplified via an RNA intermediate. LINEs are sometimes referred to as non-LTR retrotransposons because sequences in the elements code for enzymes utilized in the retroposition process.

The *Alu* element is estimated at over one million copies in the human genome representing the primary SINE family (Lander *et al*., 2001). Sequence comparisons suggest that *Alu* repeats were derived from the 7SL RNA gene. Each *Alu* element is about 280 bp with a dimeric structure, contains RNA polymerase III promoter sequences, and typically has an A-rich tail and flanking direct repeats (generated during integration). Approximately 5000 *Alu* elements have integrated within the human genome subsequent to the divergence of humans from the great apes. About 25% of the more recent *Alu* integrations have yielded presence/absence insertion polymorphisms that are useful as DNA markers for the study of forensics and human population genetics (Deininger and Batzer, 1993). Recent germline *Alu* element integrations have resulted in pathogenic phenotypes. *Alu* elements predominate in chromosomal R bands and preferentially insert into A-rich sequences including the A-tails of previous *Alu* integrations. *Alu* elements and LINEs appear to amplify by the activity of a few master loci, leaving the vast majority of these repeats as inactive pseudogenes.

LINEs (or L1 elements) are estimated at over 500 000 copies, and are predominately found in chromosomal G bands. A full-length LINE is approximately 6.1 kbp, although most are truncated pseudogenes (see below) with various 5' ends due to incomplete reverse transcription. About 1–2% of the estimated 3500 full-length LINEs have functional RNA polymerase II promoter sequences along with two intact open reading frames necessary to generate new L1 copies. Individual LINEs contain a poly(A) tail and are flanked by direct repeats. LINE mobilization activity has been verified in both germinal and somatic tissues.

SINEs and LINEs additionally contribute to the evolution of the genome by yielding sites for unequal homologous recombination. Within the low-density lipoprotein (LDL) receptor gene alone there have been several alterations attributed to recombination at various *Alu* sites resulting in familial hypercholesterolaemia.

The human genome also contains families of retroviral-related sequences. These are characterized by sequences encoding enzymes for retroposition and contain LTRs. However, most of these sequences are defunct truncated and mutated retrovirus-like elements. The endogenous retroviruses may have originally been incorporated into the genome following retroviral infection of the germ cells. In addition, solitary LTRs of these elements may be located throughout the genome. There are several low abundant (10–1000 copies) human endogenous retrovirus (HERV) families, with individual elements ranging from 6 to 10 kb. Overall, LTR elements encompass approximately 8% of the genome.

The transposon-like human element (THE-1) contains the long terminal repeats of integrated retroviral genomes, but lacks sequences encoding for enzymes involved in retroposition. Therefore, this interspersed DNA family is tentatively characterized as a retrotransposon. The 2.3-kb THE-1 sequence is estimated at 10 000 copies with an additional 10 000 solitary LTRs.

There are pseudogenes (see below) that are the result of retroposition (retropseudogenes). These pseudogenes lack introns and the flanking DNA sequences of the functional locus and therefore are not products of gene duplication. The

generation of these types of elements is dependent on the reverse transcriptase of other elements such as LINEs.

Medium reiteration frequency repetitive sequences (MERs) represent a broad group of families of uncharacterized interspersed sequences. The mechanisms for amplification of these sequences are unclear, and therefore may or may not warrant inclusion as a transposable element. However, it has been suggested that some of these sequences are replicated and disseminated by DNA viruses, indicated by the presence of MER sequences in SV40 recombinants. Copy numbers of MER families range from 200 to 10 000, collectively yielding 100 000–200 000 copies.

Pseudogenes

Pseudogenes are nonfunctional copies of a gene containing part or all of the original sequence. Pseudogenes may arise by tandem duplications, accumulating mutations as a result of the lack of selection pressure, and are usually recognizable by a lack of an open reading frame. An example is the globin pseudogenes. A processed pseudogene or retropseudogene is derived by an RNA intermediate. The characteristic features of these sequences are that they lack regulatory sequences, and therefore are normally incapable of expression, and they lack introns that have been spliced during RNA processing. There may be as many as 20–30 retropseudogenes that have arisen from a parental functional gene, e.g. ribosomal protein L32 and glyceraldehyde-3-phosphate dehydrogenase. However, SINEs and LINEs represent the most abundant families of retropseudogenes. Processed pseudogenes may be derived from RNA genes as well. Evidence for tRNA ‘retro’ pseudogenes are the CCA sequences at the 3' end which are added posttranscriptionally to the functional tRNA.

Mitochondrial Genome

The mitochondrion contains an autonomously replicating genome. The human mitochondrial genome contains 16 569 bp encoding for 37 genes, including tRNA and rRNA genes used for mitochondrial protein synthesis. Mitochondrial DNA (mtDNA) is maternally inherited and generally there are thousands of copies of mtDNA in a cell.

Genome Evolution

The fact that genomes vary considerably among organisms is indicative of the highly dynamic nature of the genome. However, comparisons of human and chimpanzee DNA demonstrate 98–99% sequence identity. This poses an interesting question as to what makes us human. By FISH and karyograms, it is evident that a shuffling of different segments of the genome between humans and chimpanzees has occurred, although there is conservation of synteny of genes, i.e. genes that are linked in humans are also linked in chimpanzees. Genomic rearrangements may alter temporal or spatial expression of genes, as a result of the chromosomal location of the gene(s), or possibly as a function of a shift in gene imprinting, consequently yielding phenotypic variation. Small deoxyribonucleotide changes may alter the biochemical nature of the protein product, also contributing to phenotypic differences. Alterations in regulatory sequences, possibly by the incorporation of retroposed sequences, may also contribute

to phenotypic variation. Although chimpanzees and humans share nearly all integrated *Alu* elements, there has been dispersal of human-specific *Alu* sequences. It is possible that exon shuffling has played a major role. Exon shuffling occurs by unequal homologous recombination in introns, and may be a reason for the existence and maintenance of introns (Gilbert, 1978), providing a source for the generation of new proteins/gene families that are composites of differing functional domains as outlined in this review.

Acknowledgements

DHK was supported by an Eastern Michigan University Spring/Summer Research Award and a University Research in Excellence Fund. MAB was supported by award 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position of the US Department of Justice.

Revised and updated: February 2003

References

Atchley WR and Fitch WM (1995) Myc and Max: molecular evolution of a family of proto-oncogene products and their dimerization partner. *Proceedings of the National Academy of Sciences of the USA* **92**: 10217–10221.

Bergstein I, Eisenberg LM, Bhalerao J *et al*. (1997) Isolation of two novel WNT genes, WNT14 and WNT15, one of which (WNT15) is closely linked to WNT3 on human chromosome 17q21. *Genomics* **46**: 450–458.

Craig JM and Bickmore WA (1994) The distribution of CpG islands in mammalian chromosomes. *Nature Genetics* **7**: 376–382.

Deininger PL and Batzer MA (1993) Evolution of retroposons. *Evolutionary Biology* **27**: 157–196.

van der Drift P, Chan A, van Roy N, Laureys G, Westerveld A, Speleman F and Versteeg R (1994) A multimegabase cluster of snRNA and tRNA genes on chromosome 1p36 harbours an adenovirus/SV40 hybrid virus integration site. *Human Molecular Genetics* **3**: 2131–2136.

Gilbert W (1978) Why genes in pieces? *Nature* **271**: 501.

Hentschel CC and Birnstiel ML (1981) The organization and expression of histone gene families. *Cell* **25**: 301–313.

Lander ES, Linton LM, Birren B, Busbaum R *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lindgren V, Ares M Jr, Weiner AM and Francke U (1985) Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. *Nature* **314**: 115–116.

McBride OW, Pirtle IL and Pirtle RM (1989) Localization of three DNA segments encompassing tRNA genes to human chromosomes 1, 5, and 16: proposed mechanism and significance of tRNA gene dispersion. *Genomics* **5**: 561–573.

Robertson HM (1996) Members of the pogo superfamily of DNA-mediated transposons in the human genome. *Molecular and General Genetics* **252**: 761–766.

Stephens JC, Cavanaugh ML, Gradie MI, Mador ML and Kidd KK (1990) Mapping the human genome: current status. *Science* **250**: 237–244.

Venter JC, Adams MD, Myers EW *et al*. (2001) The sequence of the human genome. *Science* **291**: 1304–1351.

Further Reading

Li W-H and Graur D (1991) *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Strachan T and Read AP (1999) *Human Molecular Genetics*, 2nd edn. Oxford: BIOS Scientific.

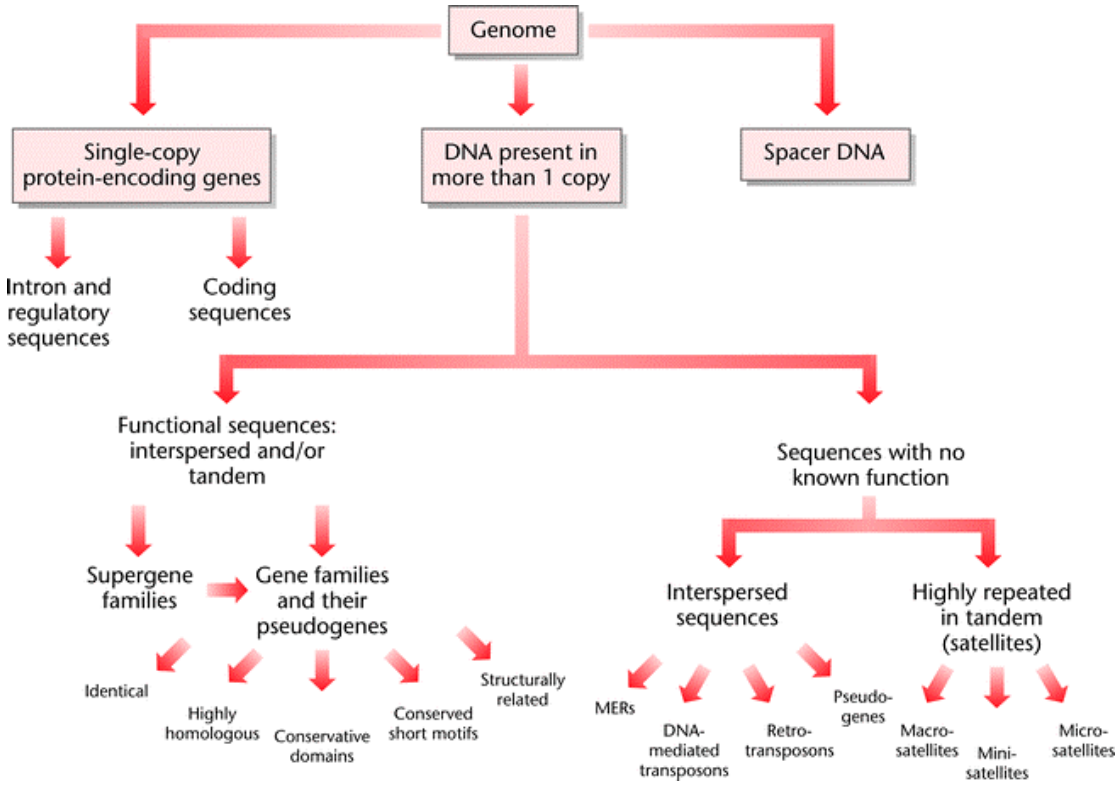

Figure 1#Broad classification of DNA sequences in the human genome. MER, medium reiteration frequency repetitive sequence.

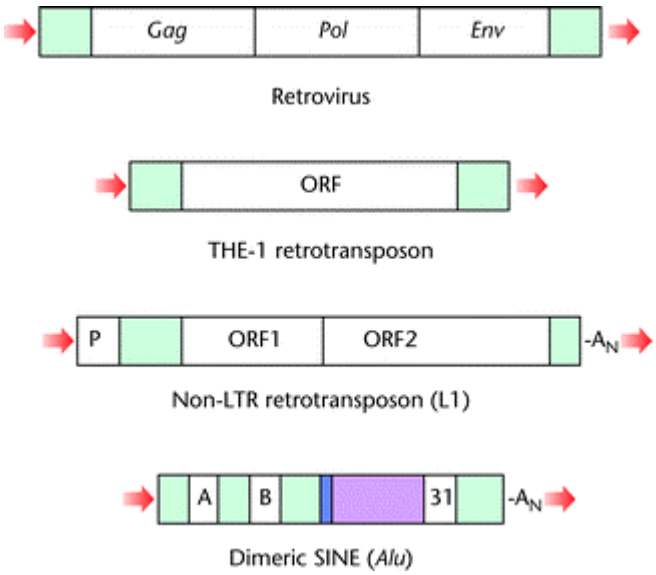

Figure 2#RNA-mediated transposable elements in the human genome. Each contains the characteristic flanking direct repeats (arrows). The human endogenous retrovirus containing long terminal repeats (LTRs) (pale green regions), *gag* (group-specific antigen gene), *pol* (polymerase gene) and *env* (envelope gene). The THE-1 retrotransposon consists of an open reading frame (ORF) and LTRs. The non-LTR retrotransposon (L1) contains internal RNA polymerase II promoter sequences (P), two open reading frames, and a poly(A) tail. The *Alu* element has a dimeric structure of homologous halves separated by a middle A-rich region (blue). The left half contains A- and B-box RNA polymerase III promoter sequences, and the right half contains an additional internal 31 bp. For L1 and *Alu* elements, pale green and mauve regions are sequences unique to these elements.

Glossary

Chromosome#A single DNA molecule consisting of genes, regulatory sequences and additional classifications of sequences, compacted with proteins and RNA that help define its structure and level of activity.

CpG island#Guanine plus cytosine-rich stretches of DNA (500–1000 nucleotides) usually flanking housekeeping genes and most tissue-specific genes.

Genome#The total DNA in a cell.

Karyogram#Representation of an entire chromosome set that has been stained by one of several possible methods yielding discrete banding patterns.

Polymorphism#The existence of two or more alleles of a gene at significant frequencies in a population.

