

I dedicate this review in memory of Professor Jacek Augustyniak, who introduced me to the world of genes and genomes

Review

The human genome structure and organization[★]

Wojciech Makalowski[✉]

*National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, U.S.A.*

Received: 22 January, 2001; accepted: 26 February, 2001

Genetic information of human is encoded in two genomes: nuclear and mitochondrial. Both of them reflect molecular evolution of human starting from the beginning of life (about 4.5 billion years ago) until the origin of *Homo sapiens* species about 100 000 years ago. From this reason human genome contains some features that are common for different groups of organisms and some features that are unique for *Homo sapiens*. 3.2×10^9 base pairs of human nuclear genome are packed into 23 chromosomes of different size. The smallest chromosome - 21st contains 5×10^7 base pairs while the biggest one - 1st contains 2.63×10^8 base pairs. Despite the fact that the nucleotide sequence of all chromosomes is established, the organisation of nuclear genome put still questions: for example: the exact number of genes encoded by the human genome is still unknown giving estimations from 30 to 150 thousand genes. Coding sequences represent a few percent of human nuclear genome. The majority of the genome is represented by repetitive sequences (about 50%) and noncoding unique sequences. This part of the genome is frequently wrongly called "junk DNA". The distribution of genes on chromosomes is irregular, DNA fragments containing low percentage of GC pairs code lower number of genes than the fragments of high percentage of GC pairs.

[★]Presented at the XXXVI Meeting of the Polish Biochemical Society, Poznań, 13 September 2000, Poland.

[✉]Mailing address: NCBI/NLM/NIH, 45 Center Drive, MSC 6510, Bldg. 45, Room 6As.47A, Bethesda, MD 20892-6510, U.S.A., phone: (301) 435 5989; fax: (301) 480 2918; e-mail: makalowski@ncbi.nlm.nih.gov

Abbreviations: CDS, coding DNA sequence; EST, expressed sequence tag; FISH, fluorescence *in situ* hybridization; HERV, human endogenous retrovirus; LINE, long interspersed repetitive element; LTR, long terminal repeat; SAR, scaffold-attachement region; SINE, short interspersed repetitive element; TE, transposable element; UTR, untranslated region.

INTRODUCTION – HISTORICAL PERSPECTIVE

From the beginning of humanity, people have been interested in themselves. They were well aware of two aspects of living nature: an immense variability within each species and the tendency for characteristics of parents to be transmitted to their offspring. Already pre-Socratic philosophers noticed that people shared some characteristics, e.g. had usually, with some exceptions, two hands, a nose, large forehead, in other words they were alike. On the other hand, everybody was different and nobody should have a problem to distinguish those two gentlemen by such characteristics as eyes, cheeks, or shirts. Ancient people were also aware that the above was true for both intra- and inter-species comparison.

The question arises: how does it happen that our children are more similar to parents than to monkeys? The problem already intrigued pre-Socratic philosophers. Probably the first person who publicly expressed his thoughts on the subject was Anaxagoras of Clazomenae. According to his teaching, seed material is carried from all parts of the body to reproductive organs by the humors. Fertilization is the mixing of the seed material of father and mother. That all parts of the body participate in the production of seed material is documented by the fact that blue-eyed parents have blue-eyed children and baldheaded men have sons that become baldheaded – not a very good prospect for my own children. The idea of panspermy or pangenesis was adapted and taught by the famous physician Hippocrates (about 460–377 B.C.) and was widely accepted until the end of the nineteenth century, also by Charles Darwin. One of the greatest scientists of all time, Aristotle of Stagira had a different view on the problem. Aristotle's theory of inheritance, as described in one of his major works *De generatione animalium*, was holistic. He held that the contributions by males and females

were not equal. The semen of the male contributes the form-giving principle, *eidos*, while the menstrual blood, *cantemina*, of the female is the unformed substance shaped by the *eidos* of the semen. "The female always provides the material, the male provides that which fashions the material into shape; this in our view, is the specific characteristic of each sex: that is what it means to be male or to be female." (Aristotle, 1965).

The twentieth century witnessed accelerated development of biology and with it the nature of the inheritance process was understood. Consequently, an effort to decipher the blueprint of our species has started. Several biological discoveries were especially important to decipher the human genome. Everything started with the rediscovery of Mendel's laws by Hugo Marie De Vries (1900), followed by discovery of chromosomes by Thomas H. Morgan in 1910 (Morgan, 1910). In 1953, James D. Watson and Francis H.C. Crick unraveled the structure of DNA (Watson & Crick, 1953a; Watson & Crick, 1953b). Four years later, Johan H. Matthaei and Marshall Nirenberg performed experiments which enabled deciphering the genetic code. With the development of the fast methods of DNA sequencing in the mid-seventies (Maxam & Gilbert, 1977; Sanger *et al.*, 1977), followed by automation of cloning and sequencing in the nineties, the way to understand our blueprint became clear. By now, many complete genomes of both prokaryotic and eukaryotic organisms have been sequenced. For up-to-date tables with completed genomes, go to <http://www.ebi.ac.uk/genomes/>. On June 26, 2000, virtually all news agencies in the world announced completion of a working draft of the human genome. This accomplishment was so important for humankind that instead of announcing it at a scientific conference or in a scientific journal, as used to be with a scientific milestones, a special press conference was organized in *The White House* in Washington, D.C. In several days faces of major players from both private and public

sectors appeared on journals' covers around the world, including the Polish weeklies *Polityka* and *Wprost*. It is worth pointing out that the public genome project already completed sequence of two chromosomes: 22 (December, 1999) (Dunham *et al.*, 1999) and 21 (May, 2000) (Hattori *et al.*, 2000). The working draft of the human genome was published by both projects last January.

HUMAN GENOME – GENERAL INFORMATION

Our genetic material is stored in two organelles: nucleus and mitochondria. This review is focused on the nuclear genome in which 3.2 miliard bp are packed in 22 pairs of autosomes and two sex chromosomes, X and Y. Human chromosomes are not of equal sizes; the smallest, chromosome 21, is 54 mln bp long; the largest, chromosome 1, is almost five times bigger with 249 mln bp (see Table 1).

Genomic sequences can be divided in several ways. From the functional point of view we can distinguish genes, pseudogenes, and non-coding DNA (Fig. 1). Only a minute fraction of the genome – about 3% – codes for pro-

Table 1. Physical sizes of human chromosomes

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

transposable elements as well but with time they have mutated beyond recognition.

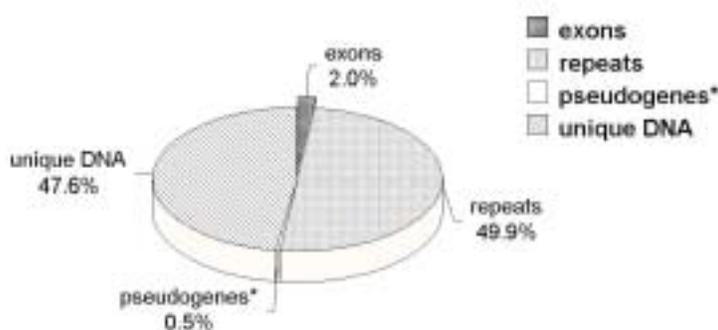


Figure 1. Fractions of different sequences in the human genome.

teins. There are many pseudogenes in the human genome (0.5%) but most of the genome consists of introns and intergenic DNA. Almost half of these sequences consist of different transposons; moreover, the remaining non-coding DNA most likely originated from

SEQUENCE COMPLEXITY

The human genome contains various levels of complexity as demonstrated by reassociation kinetics. Such analyses of the human genome estimate that 60% of the DNA is ei-

ther 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). That 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 (> 300 kb) segments of DNA that are homogeneous in base composi-

tion referred to as isochores (Bernardi, 2000), based on sequence analysis and compositional mapping. L1 and L2 are GC-poor (or 'light') isochore families representing about 62% of the genome. The H1, H2 and H3 (heavy) isochore classes are increasingly GC-rich. There is some correlation between isochores and chromosomal bands. G bands are almost exclusively composed of GC-poor isochores, with a minor contribution from H1. R bands can be classified further into T bands (R banding at elevated temperatures), which are composed mainly of H2 and H3 isochores, and R' (non-T R bands) which are comprised of nearly equal amounts of GC-rich (primarily H1) and GC-poor isochores.

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.

HUMAN GENE NUMBER

It is interesting that the number of genes coded by our genome is not known and probably will not be known long after completion of the human genome sequencing. Nevertheless, in the last decade, several groups tried to answer this question using different methods (see Table 2). Unfortunately, the estimations differ very much with prediction as low as 28 000 up to 80 000 genes per human haploid. The whole genomic community is so excited with this mysterious number that they decided to organize the Gene Sweepstake. The Gene Sweepstake will run between 2000 and 2003 and its detailed rules may be found at: <http://www.ensembl.org/Genesweep/>. As of January 2001, 165 bets were made with gene number between 27 462 and 153 478 and a mean value of 61 710.

Table 2. Estimation of human gene number using different methods

Gene number	Method	Reference
80 000	CpG islands	(Antequera & Bird, 1994)
64 000	ESTs	(Fields <i>et al.</i> , 1994)
35 000	ESTs	(Ewing & Green, 2000)
28 000–34 000	Comparative genomics	(Roest Crolius <i>et al.</i> , 2000)
30 000	Gene punctuation	(Yang <i>et al.</i> , 2001)

EXONS CHARACTERISTIC

In most human genes, coding sequences are interrupted by stretches of non-coding sequences, which are spliced out during mRNA maturation. Using nomenclature introduced by Walter Gilbert (Gilbert, 1978), the human genes look like mosaics, consisting of series of *exons* (DNA sequences that can be subsequently found in the mature mRNA) and *introns* (silent DNA sequences that are absent from the final mRNA). As nothing in nature is simple, some of the introns carry significant information and even code for other complete genes (see description of *nested genes* below). Initially, it was thought that introns occurred only in untranslated parts of mRNA and coding sequences (CDS) were not interrupted. However, it soon became clear that introns could be found in all domains of mRNA molecule. Therefore, exons can be classified as follows: 5' UTR exons, coding exons, 3' UTR exons, and all possible combinations of those three main types, including single exons that cover the whole mRNA. The latter are very interesting from the evolutionary biology point of view, because in most cases they are retroposed copies of "regular" genes with introns. Michael Zhang of Cold Spring Harbor Laboratory analyzed 4731 human exons (Zhang, 1998). It appears that human exons are relatively short with median value of 167 bp and mean equal to 216 bp. The shortest exon was only 12 bp while the longest one 6609 bp. These numbers have to be taken with some caution because they are based on GenBank annotation, which sometimes is not very precise. Mixed (including coding and

non-coding sequences) exons tend to be longer than single type exons, especially those at the end of the message; not surprisingly so, since 3' UTRs are relatively long in mammalian mRNAs. In our analysis of over 2000 human mRNA sequences the median and mean sizes of human message domains were as follow: 118 nt and 191 nt for 5' UTR, 1191 and 1424 for CDSs, and 534 and 576 for 3' UTRs, respectively (Makalowski *et al.*, 1996; Makalowski & Boguski, 1998).

GENE DISTRIBUTION

Genes may be transcribed from either the same or from the opposite strand of the genome, i.e. they may lie in the same (tail-to-head) or opposite orientation (head-to-head or tail-to-tail). Although the vast majority of the human genome accounts for non-exonic sequences, a surprisingly large number of genes occupy the same genomic space. About 6% of human genes reside in introns of other genes (Wong *et al.*, 2000). For example, intron 27th of NF1 gene hosts three other genes that have small introns on their own, suggesting that they are not products of retroposition (see Fig. 2). Additionally, over 100 gene pairs are overlapping at 3' end, i.e. their 3' UTRs occupy the same region though different strands (I. Makalowska, personal communication). TPR and MSF genes map to the same region of chromosome 1. The last exon of the TPR gene is 872 nt long and overlaps completely with the last exon of the MSF gene (200 nt). Interestingly, the very end of

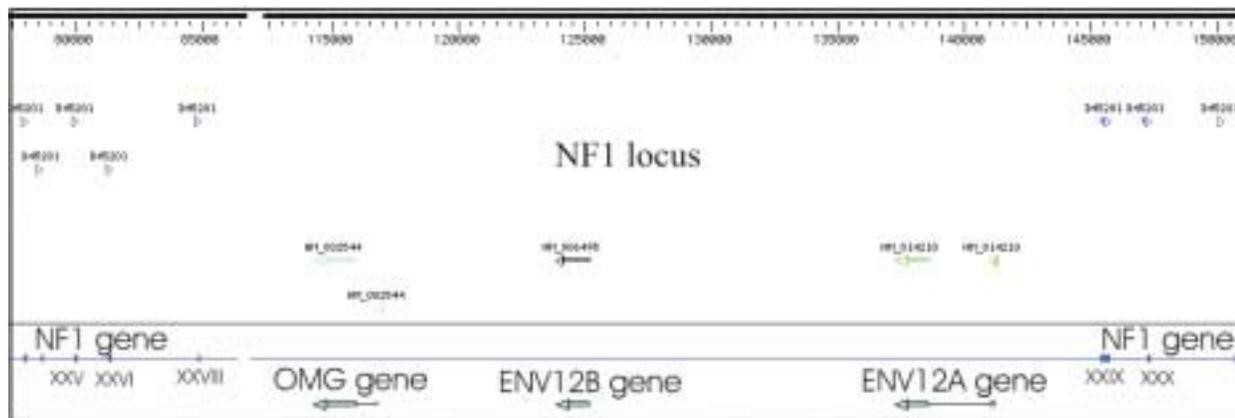


Figure 2. An example of nested genes.

The human sequence from chromosome 1 (GenBank accession number AC004526) was analyzed using GeneMachine (Makalowska *et al.*, 2001). Connected closed boxes represent gene models as predicted by GenScan software (Burge & Karlin, 1997) and boxes with arrows represent results of BLASTn search; AC004526 was used as a query against nr database.

the MSF gene overlaps with the intron of the TPR gene (see Fig. 3).

Unlike in plant genomes, most of non-exonic sequences in human genome account for introns (Wong *et al.*, 2000). However, genes are not equally distributed throughout the genome. 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 seg-

ments 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 further depicts gene-poor and gene-rich chromosomal segments (Craig & Bickmore, 1994). As a consequence, more than half of human genes locate in the so-called “genomic core” (isochores H2 and H3) comprising only 12% of the human genome (see Table 3).

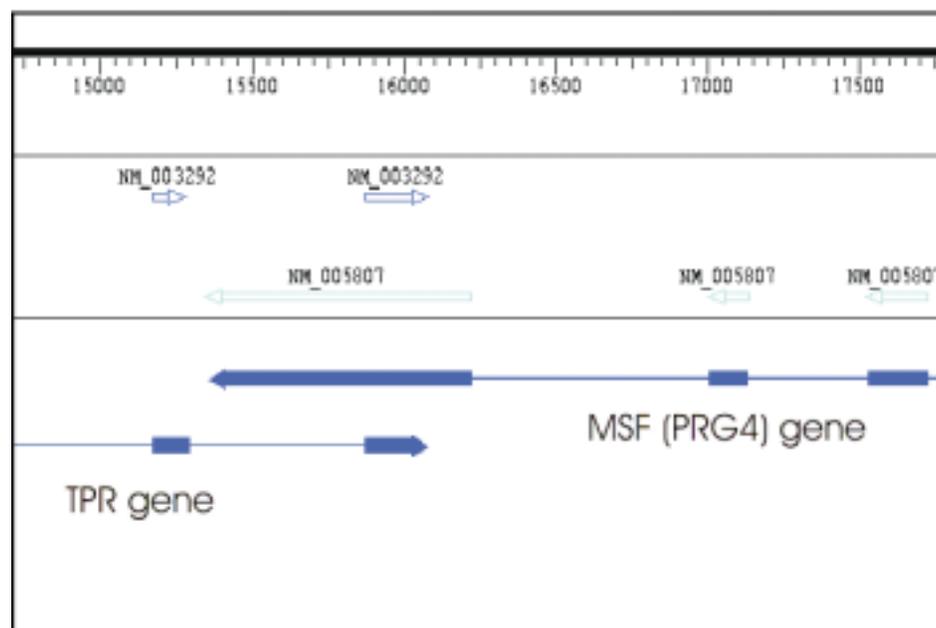


Figure 3. An example of overlapping genes.

The human sequence from chromosome 1 (GenBank accession number AL13-3533) was analyzed using GeneMachine (Makalowska *et al.*, 2001). Connected closed boxes represent gene models as predicted by GenScan software (Burge & Karlin, 1997) and open boxes with arrows represent results of BLASTn search; AL13-3533 was used as a query against nr database.

Table 3. Gene density in different isochores

	Genomic core	"Empty" space
Isochore type	H2 and H3	L, H1
Genome fraction	12%	88%
Gene fraction	54%	46%
Gene density	1/10 kbp	1/100 kbp

GENE FAMILIES

Many genes can be clustered in groups of different sizes based on sequence similarity. The similarity between two genes varies from genes coding identical products to genes in which product similarity is barely detectable and/or limited to short sequence stretches called sequence motives. Gene families arose during the evolution by gene duplications over the different periods of time as reflected in sequence similarity. In general, more similar genes shared a common ancestor later (in nearer past) than genes with a weaker similarity, although gene conversion can result in very similar or identical gene copies regardless of gene duplication time. Gene duplication can occur by different mechanisms, like unequal recombination or retroposition. Not all duplicated genes remain active, some of them end up in genomic oblivion and are called pseudogenes. Some of the pseudogenes can be rescued from the genomic death by capturing a promoter and regulatory elements in the course of evolution as happened with Θ -globin gene which was rescued by an Alu element after 200 mln years of silent existence (see discussion in Makalowski, 1995).

The histone gene family is an example of very similar genes. It 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. all H4 genes code for the identical 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 & Birnstiel, 1981). A majority of histone genes form a large cluster on human chromosome 6 (6p21.3) and a small cluster at 1q21. Interestingly, histone genes lack introns; a rare feature for eukaryotic genes.

Genes that encode ribosomal RNA (rRNA) 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 in tandem arrays of approximately 60 copies each yielding about 2 million bp 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 and then cleaved. 5S rRNA genes are clustered on chromosome 1q.

Some genes in the human genome share highly conserved amino-acid domains with weak overall similarity. 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 sequence. In humans there are four Hox gene clusters, each on a different chromosome. However, the individual genes in the cluster demonstrate greater similarity to a counterpart gene in another cluster than to the other genes in the same cluster.

There are pseudogenes that are the result of retroposition (retropseudogenes). The 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 retroelements such as LINES.

REPETITIVE SEQUENCES

The human genome is occupied by stretches of DNA sequences of various length that exist in variable copy number. These repetitive sequences may be in a tandem orientation or they may be 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 and interspersed repeat sequences are typically the products of transposable element integration, including retrogenes and retropseudogenes of a functional gene. For the up-to-date list of human repetitive elements visit the RepBase at <http://www.girinst.org/>.

GENOMIC DUPLICATIONS

Thirty years ago, Suzumu Ohno put forth a hypothesis about two duplications of the whole genome in the early stages of vertebrate evolution (Ohno, 1970). According to his hypothesis, most vertebrate gene families should give three or four well-defined branches, as presented in Fig. 4. Unfortunately, analysis of over 10 000 vertebrate gene families does not support Ohno's hypothesis (Makalowski, unpublished observation). Nevertheless, duplications in human genome do exist and they play a significant role in genes and the genome evolution. Although sometimes very large, they appear to be on a local, not a global scale. For example, the comparison of the complete human chromosome 21 sequence with both itself and other human sequences revealed many large duplications

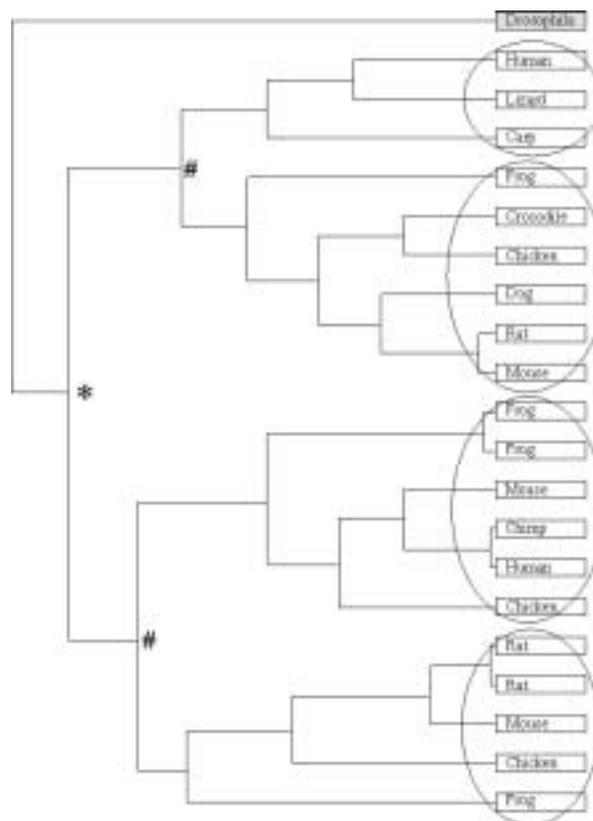


Figure 4. A hypothetical phylogenetical tree of vertebrate gene family under Ohno's hypothesis about two genome duplications in early vertebrate evolution.

Drosophila gene represents an outgroup and four clusters of a gene family are encircled. An asterisk (*) marks first genome duplication and a hash sign (#) marks points of second genome duplication. Different branch lengths suggest different evolutionary rates after ancestral gene duplication.

with the largest intra-chromosomal duplication being 189 kb (position 188–377 and 14795–15 002 in q arm) and the largest detected inter-chromosomal duplication of over 100 kb region from q arm of chromosome 21 (position 646–751) duplicated in chromosome 22 (position 45–230) (RIKEN, 2000).

MICROSATELLITES, MINISATELLITES, AND MACROSATELLITES

Microsatellites are small arrays of short simple tandem repeats, primarily 4 bp or less. Dif-

ferent 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 may indicate some function. Repeat unit copy number variation of microsatellites apparently occurs by replication slippage. The expansion of trinucleotide repeats within genes has been associated with genetic disorders such as Huntington disease or fragile-X syndrome.

Minisatellites are tandemly repeated sequences of DNA of lengths ranging from 1 kbp to 15 kbp. For example, telomeric DNA 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.

Macrosatellites are very long arrays, up to hundreds of kilobases, of tandemly repeated DNA. There are three satellite bands observed by buoyant density centrifugation. However, not all satellite sequences are resolved by density gradient centrifugation, e.g. alpha satellite DNA or alphoid DNA that constitute 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).

TRANSPOSABLE ELEMENTS

The human genome contains interspersed repeat sequences that have largely amplified in copy number by movement throughout the genome. Those sequences (transposable elements or TEs) can be divided into two classes based on the mode of transposition (Finnegan, 1989). The Class I elements are TEs which transpose by replication that involves

an RNA intermediate which is reverse transcribed back to DNA prior to reinsertion. These are called retroelements and include LTR transposons, which are structurally similar to integrated retroviruses, non-LTR elements (LINEs and SINEs), and retrogenes (see Fig. 5). Class II elements move by a conservative cut-and-paste mechanisms, the excision of the donor element is followed by its reinsertion elsewhere in the genome. Integration of Class I and Class II transposable elements results in the duplication of a short sequence of DNA, the target site. There are about 500 families of such transposons. Most of transposition has occurred *via* an RNA intermediate, yielding classes of sequences referred to as retroelements (more than 400 families, e.g. Alu, L1, retrogenes, MIR). However, there is also evidence of an ancient DNA-mediated transposition (more than 60 families of class II (DNA) transposons, e.g. THE-1, Charlie, Tigger, mariner).

RETROELEMENTS

Short interspersed repetitive elements (SINEs) and long interspersed repetitive elements (LINEs) are the two most abundant classes of repeats in human, and represent the two major classes of mammalian retrotransposons. Structural features shared by LINEs and SINEs include an A-rich 3' end and the lack of long terminal repeats (LTRs); these features distinguish them from retroviruses and related retroelements.

A full-length LINE (or L1 element) is approximately 6.1 kbp although most are truncated pseudogenes with various 5' ends due to incomplete reverse transcription. There are about 100 000 copies of L1 sequences in our genome. Approximately 1% 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

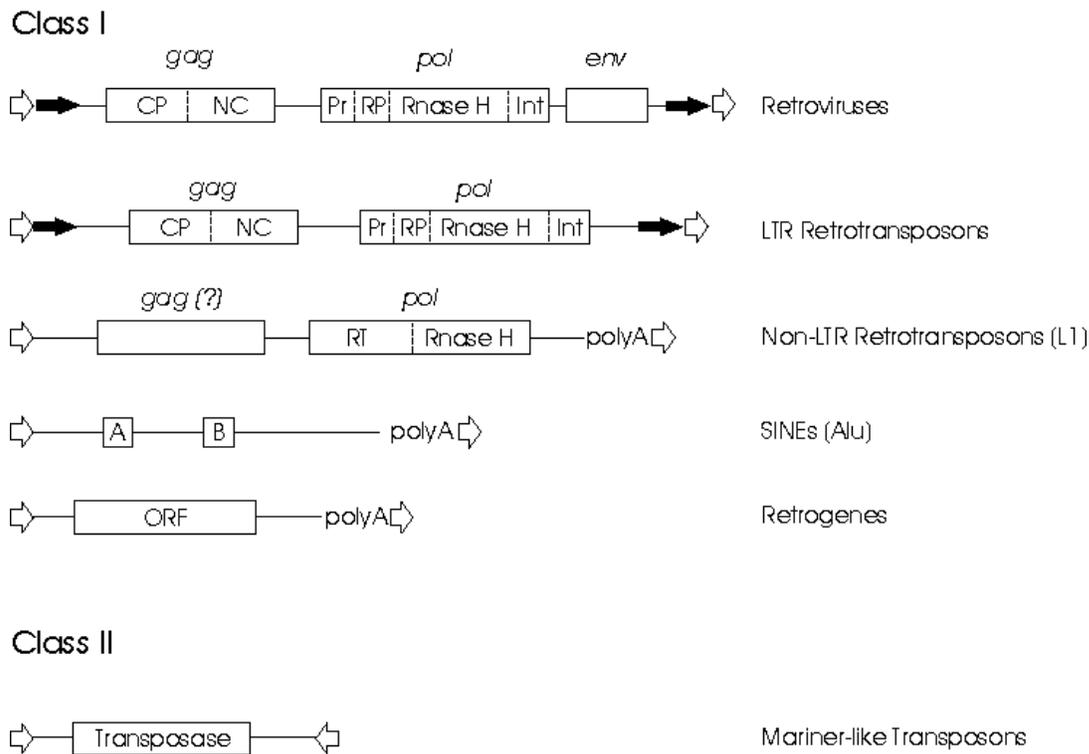


Figure 5. The structure of different human transposable elements.

Open arrows denote duplicated target sites and closed arrows denote long terminal repeats (LTRs). The following abbreviations are used: CP, capsid; NC, nucleocapsid; Pr, proteinase; RT, reverse transcriptase; Int, integrase; ORF, open reading frame, and A and B denote polymerase III internal promoter.

repeats. LINE mobilization activity has been verified in both germinal and somatic tissues.

The Alu element is estimated at 500 000–900 000 copies in the human genome representing the primary SINE family, the most successful transposon in any genome. 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). Although Alu elements are present in all primate genomes, more than 2000 Alu elements have integrated within the human genome subsequent to the divergence of humans from the great apes.

The human genome also contains families of retroviral-related sequences. These are characterized by sequences encoding enzymes for retroposition and contain LTRs. In addition, solitary LTRs of these elements may be located throughout the genome. There are sev-

eral low abundant (10–1000 copies) human endogenous retrovirus (HERV) families, with individual elements ranging from 6 to 10 kb, collectively encompassing about 1% of the genome.

CLASS II ELEMENTS

Class II elements contain inverted repeats (10–500 bp) at their termini and encode a transposase that catalyses transposition. They move by excision at the donor site and reinsertion elsewhere in the genome by a non-replicative mechanism. The human genome hosts a number of repeated sequences originated in more than 60 different DNA transposons.

The mariner ‘fossils’ present in our genome closely resemble members of three sub-families identified in insects, adding to the already extensive evidence that horizontal transfer between genomes has been impor-

tant in genomic evolution. Other human DNA transposon remains also show high similarity to sequences in distantly related organisms. Nevertheless, the level of sequence divergence suggests that activity of all identified elements predates human evolution.

CONCLUSIONS

The 3.2 billion bp of our genetic blueprint is packed into 23 pairs of chromosomes, or 46 DNA molecules. Only a fraction of the genome is occupied by protein-coding exons and the majority of non-exonic sequences consists of repetitive elements. Functional exons contribute merely 2% of a genome, up to 50% of a genome is occupied by repetitive element, the remaining 48% is called unique DNA, most of which probably originated in mobile elements diverged over time beyond recognition. Different evolutionary forces shape the human genome composition and structure. It appears that different mobile elements play a significant role in this process (reviewed recently in Makalowski, 2000). The human genome is a dynamic entity, new functional elements appear and old ones become extinct as genes that evolve according to birth and death rule (Ota & Nei, 1994) similarly to species evolution. This confirms that the theory of evolution is truly universal and applies not only to all organisms but to all levels of life as well.

I would like to thank Izabela Makalowska for sharing unpublished data and Jakub Makalowski for preparing Fig. 5.

REFERENCES

- Antequera, F. & Bird, A. (1994) Predicting the total number of human genes. *Nat. Genet.* **8**, 114.
- Aristotle (1965) *De generatione animalium*. Oxonii, E Typographeo Clarendoniano.
- Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17.
- Burge, C. & Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94.
- Craig, J.M. & Bickmore, W.A. (1994) The distribution of CpG islands in mammalian chromosomes. *Nat. Genet.* **7**, 376–382.
- Dunham, I., Shimizu, N. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature* **402**, 489–495.
- Ewing, B. & Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25**, 232–234.
- Fields, C., Adams, M.D., White, O. & Venter, C.O. (1994) How many genes in the human genome? *Nat. Genet.* **7**, 345–346.
- Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**, 103–107.
- Gilbert, W. (1978) Why genes in pieces? *Nature* **271**, 501.
- Hattori, M. *et al.* (2000) The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium (see comments). *Nature* **405**, 311–319.
- Hentschel, C.C. & Birnstiel, M.L. (1981) The organization and expression of histone gene families. *Cell* **25**, 301–313.
- Makalowska, I. *et al.* (2001) GeneMachine: A tool for sequence analysis and annotation. submitted.
- Makalowski, W. (1995) SINEs as a genomic scrap yard: An essay on genomic evolution; in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (Marais, R.J. & Austin, R.G., eds.) pp. 81–104, Landes Company.
- Makalowski, W. (2000) Genomic scrap yard: How genomes utilize all that junk. *Gene* **259**, 61–67.
- Makalowski, W. & Boguski, M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9407–9412.

- Makalowski, W., Zhang, J. & Boguski, M.S. (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6**, 846-857.
- Maxam, A.M. & Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560-564.
- Morgan, T.H. (1910) Chromosomes and heredity. *Amer. Nat.* **44**, 449-496.
- Ohno, S. (1970) *Evolution by Gene Duplication*. Springer Verlag, New York.
- Ota, T. & Nei, M. (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol. Biol. Evol.* **11**, 469-482.
- RIKEN (2000) <http://hgp.gsc.riken.go.jp/chr21/annotation.htm>
- Roest Crolius, H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* **25**, 235-238.
- Sanger, F., Nicklen, S. & Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463-5467.
- Watson, J.D. & Crick, F.H.C. (1953a) Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964-967.
- Watson, J.D. & Crick, F.H.C. (1953b) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738.
- Wong, G.K., Passey, D.A., Huang, Y., Yang, Z. & Yu, J. (2000) Is "Junk" DNA mostly intron DNA? *Genome Res.* **10**, 1672-1678.
- Yang, C.Z. *et al.* (2001) Gene Punctuation. Submitted.
- Zhang, M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**, 919-932.