[Frontiers in Bioscience 10, 335--344, January 1, 2005]

GENOME ORGANIZATION AND THREE KINDS OF HERITABLE CHANGES: GENERAL DESCRIPTION AND STOCHASTIC FACTORS (A REVIEW)

Michael Golubovsky and Kenneth G. Manton

Center for Demographic Studies, Duke University, 2117 Campus Drive, Durham, NC 27708

TABLE OF CONTENTS

1. ABSTRACT

Due to the increased knowledge of genome architecture, topology, and the mechanisms of hereditary variability, the list of genetic components has grown. This review outlines the general features and principles of genome organization in diverse organisms. The genome codes, stores, and transfers information in both structurally and dynamically. The genome includes two subsystems of genetic elements: obligatory (genes and gene families) and various types of facultative elements which are predominant the content of the human genome. The features of three kinds of heritable changes: mutations, variations (changes in the number or topography of facultative elements), and epigenetic alterations are described. Facultative elements are the first to react to environmental challenges. Together with epigenetic changes, they implement the operational genomic memory. This review discusses both the role of stochastic factors and the transient features of DNA components.

2. INTRODUCTION

The genome, which is the cell's hereditary system, codes, stores, and transfers information in diverse ways. We describe three of these ways: (i) mutations, (ii) variations, and (iii) epigenetic alterations. Mutations are changes in genes or their order in chromosomes. Variations are changes in the number and topography of FE (in the nucleus and cytoplasm). Mutations and variations differ in incidence and stability. Variations reflect the operational memory of the genome. The dynamic aspects of genome organization and function are manifested in epigenetic changes (1-3, 15, 16, 17).

The genome can be subdivided into Obligatory (genes and gene families) and facultative elements (FE). In humans, 5-10% of genome DNA is Obligate. The rest are facultative elements (fractions of highly repetitious DNA, retroelements and transposons, pseudogenes, and amplicons) (3).

Genetic analysis depends on advances in our understanding of the genome structure and of the stochastic molecular mechanisms of hereditary variability. This means that there are determinate structural elements of both heredity and the environment (including cellular thermodynamics), which impose uncertainty about how structures evolve over time. The heritable changes, apart from mutations, include FE variations and epigenetic alterations. This approach requires evaluating FE profiles for heterogeneity and stability (especially for senescence and age-related diseases). Enhancement of knowledge about epigenetic programming and reprogramming during embryonic development and gametogenesis requires new types of data sets (3, 5-8) and an analysis of three generations in biodemographic studies.

3. GENOME ORGANIZATION

3.1. Genome and the cell information system

The term "genome" was coined by cytologist H. Winkler in 1920 to designate the haploid chromosome set for a species. This term was used when speaking about either analysis of allopolyploids or genomic mutations (such as changes in chromosome numbers). Its meaning has widened to include the entire hereditary constitution of the

cell (i.e., the structural and dynamic aspects of coding, storage, and transfer of species-specific hereditary information).

The necessity of including in the genome concept of stochastic dynamics comes from the discovery of regulatory genes that control the level and stability of expression of structural genes. The genome contains not only blueprints, but a coordinated program of protein synthesis (1, 2). Genome and genome-controlled development involves: (i) structural elements, (ii) their dynamics, and (iii) "emergent" properties of the system (3). The holistic aspects of species-specific hereditary systems might be viewed metaphorically as the architectural design of a temple that cannot be understood by studying separate bricks – genes –at a fixed point in time (3, 8).

The discoverers of regulator genes - and the operon - entitled their paper "Teleonomic mechanisms in cellular metabolism, growth and differentiation" (2). To preserve intracellular homeostasis and the adaptive response of the genome to environmental challenges, they emphasized the biological purposefulness (or teleonomy) of the cell's regulatory systems. Recent molecular discoveries have led to significant conceptual changes in genome organization. Theory has shifted from genes as units of inheritance and function to the genome as a complex and dynamic stochastic information (in Shannon's theory) system (6-8).

Genome organization includes regulatory and coding motifs shared among many loci. They do not code proteins, but act as signals determining genome functions such as transcription, translation RNA processing, DNA replication, chromatin condensation, and packaging. Genetic loci are organized into interconnected genome-wide networks which function dynamically. Diverse genetic loci are active in different tissues and developmental stages with interconnections at the DNA, protein, and chromosomal levels (5-9, 13).

The cell's ability to analyze external and internal conditions (and to control growth, movement, and differentiation) can be compared to an information computing network and check-points. By means of signal transduction pathways, a cell receives external signals and transmits, amplifies, and directs them internally. Seventeen signal transduction pathways were discovered - plus at least two stress response pathways, which are evolutionary and conservative, through which cells receive external signals, transmit, amplify, and direct them internally (14). The pathways are highly conserved in nematodes, flies and all vertebrates. Each pathway includes signal accepting receptor, membrane, or cytosolic proteins including kinases and phosphatases to convey the signal and key transcription factors capable of switching its state, activating, or suppressing transcription of definite genes. Protein kinases participating in this cascade are activated by chemical signals (calcium ions, cyclic AMP, mitogene molecules) and transfer phosphate groups to pathway proteins. Phosphorylation of cytoplasmic proteins amplifies signals, resulting in differential gene expression (14).

The normal cycle of cell division is regulated in accordance with the 'trade-off' principle or a counterbalance between stimulatory and inhibitory pathways of the cells. Cell division control includes check-point functions that stop the cell cycle until after the DNA repair process has been completed. From this aspect cancer growth originates as deviations from normal regulatory circuits that govern cell proliferation and homeostasis (14). If repair is impossible, programmed cell death (apoptosis) is activated (apoptosis is necessary for normal differentiation). For example, about 12% of cells formed during the development of a C elegans die due to a genetically controlled suicide program (9).

DNA repair systems remove damage. Multiple proofreading mechanisms stochastically recognize and remove errors that occur during DNA replication or due to mutagens. Repair systems allow cells not to be passive victims of random physico/chemical forces. They control the level of mutability by modulating cell repair system activity (6, 7). DNA sequence affects enzymes that copy, move, and repair DNA (10).

Tumor origin is associated with dysregulation of cell cycle control: an overexpression of stimulatory pathways (oncogenes) or, conversely, as defects of the inhibitory pathways (tumor-supressors). This concept explains the involvement of both the same primary genes in many kinds of cancer and the multigenic nature of the each type of cancer. Thus, hyperactive proteins encoding by mutant ras-oncogenes, (they transmit stimulatory signals from growth factors receptors to other proteins) are found in about 25% of all human tumors (including carcinomas of the colon, pancreas and lung). Tumor cell growth often results from the malfunction of the checkpoint control system or dysfunction of elements of decision-making program. (15).

Mobile genetic elements (ME), discovered by B. McClintock in corn and found in all genomes, can move from one chromosomal position to another and induce DNA rearrangement (11). ME contain genetic punctuations (promoters, enhancers, stop signals of transcription, etc.), regulate expression of genes, and promote the appearance of new constructs. The term "Natural Genetic Engineering" coined by Shapiro (4, 5) emphasizes that biotechnology uses the same enzymes (nucleases, ligases, reverse transcriptases and polymerases) that living cells use to reshuffle the genome in ME activation. Though ME are repetitive and dispersed on different chromosomes, they can be activated simultaneously by cell signal (e.g., HSP90, the genetic "capacitor") resulting in outbreaks of non-random genetic variability. The basic structure of amino acids to be formed is the result of selecting a statistical topology during evolution. Genetic engineering has been extended beyond this to study multiplex base codon arrangements and to form more than 20 basic amino acids (16).

Mobile elements - from insertion segments to retroelements, transposons, plasmids, and retroviruses - promote transfer of genetic elements between species.

Table 1. Genome size, gene numbers of some prokaryotes and eukaryotes

	Organisms	Source	Genome size (Mb)	Total genes [1]
Phage/viruses, organelle				
1	mitochondrion	Human mitochondrion	16.6 x10^3	40
2	SV40	Simian virus 40	5.2 x 10^3	8
3	TMV	Tobacco Mosaic Virus (RNA)	6.4 x10^3	4
4	HIV	AIDS virus (RNA)	9.3 x10^3	10
5	adenovirus 2	Human adenovirus	35.9 x10^3	11
6	λ phage	lysogenic phage	48.5 x10^3	50
7	T4 phage	DNA virus E. coli	169 x10^3	300
Bacteria				
8	Mycoplasma genitalium		580 x10^3	470
9	Rickettsia prowazeki		1.11x10^6	834
10	Streptococcus pyogenes		1.85 x10^6	1752
11	Bacillus subtilis		4.21 x10^6	4100
12	Esherichia coli		4.64 x10^6	4288
13	Mycobacterium tuberculesis		4.41 x10^6	3924
14	Helicobacter pylori		1.668 x10^6	1590 [1]
			1.644 x10^6	1495 [1]
Eucaryotes				
15	Saccharomyces cerevisiae	budding yeast	12 x10^6	5885
16	Caenorhabditis elegans	worm, Nematoda	97 x10^6	19100
17	Arabidopsis thaliana	plant, Angiospermae	120 x 10^6	25500
18	Drosophila melanogaster	fruit fly, Diptera	165 x10^6	13000
19	Mus musculus	mouse	2.5 x10^9	30000
20	Homo sapiens	humans	3.2 x10^9	31000

[1] Two strains isolated from two geographically separated regions of USA (32).

Evidence of horizontal transfers between prokaryotes and eukaryotes was found in sequences of various species. Mice, rats, cats, pigs, and humans carry common rudiments of endogenous retroviruses. Thus, due to ME, the gene pools of all organisms can be viewed as an integrated world-wide gene pool (13).

Dynamic, determinate, and stochastic heritable alterations can occur without changes in DNA structure. This was discovered in the 1960s using the operon concept and combining them into circuits. The resulting cyclic self regulatory systems can switch from one state to another and maintain it for generations. State changes in such systems closely mimic transmissible alteration of genetic material (2). Reversible changes of genetic states transmitted through cell generations are designated as epigenetic (17-18).

Current research focuses on two mechanisms mediating epigenetic alterations: methylation and histone modifications (19-21). There is a spectrum of epigenetic changes in mammalian development (including X-chromosome inactivation, parent-of-origin chromosomal and gene imprinting, alterations of ME, and metastable alleles [22-26]). Patterns of DNA methylation and chromatin structure may modify gene expression in cancer (27). Due to the dynamics of inheritance, epigenetic information established during gametogenesis cannot be restored after nuclear transfer. Cloned animals suffer from transcriptional and developmental dysregulation (28, 29).

3.2. Comparative Genomics: DNA size and gene assemblage

Sequencing of prokaryotic and eukaryotic organisms allows us to examine the diversity of genome organization and function. Table 1 presents features of sequenced genomes. Viral genomes vary from 5.2 to 170x10^3 base pairs (bp) and from 4 to 300 genes. Size is related to complexity. Viruses with a capsid (protein structures covering DNA or RNA) and envelopes have a larger genome (e.g., phages λand T4). Mitochondria are semi-autonomous self-reproducing organelle believed to have arisen from aerobic bacteria in symbiotic relations with prokaryotes (9). Mitochondria contain genes involved in oxidative phosphorylation to produce **adenosine triphosphate** (ATP), the cell's energy machine. The mitochondrion (genome of the organelle) does not contain introns. It has the highest gene density: 40 genes on 16 600 bp (in humans) or one gene per 400 bp (37).

The average gene size of bacteria is 1000 bp – large enough to code proteins with 300 amino acids. The smallest bacterial genome is the Mycoplasma genus, which lacks a cell wall. M. genitalium, a parasite of the human genital and respiratory tracts, has only 580x 10^3 bp. Its 470 genes (average 1040 bp) comprise 88% of the genome. It has the minimal set of genes (~ 500) necessary for independent life. All 470 genes from M. genitalium are found in larger bacteria. M. pneumoniae has 820 kilo-base (kb) pairs and 679 genes. Proteins of the two species have 67% of their DNA sequences in common (8).

The genome size of bacteria varies significantly, as confirmed by analysis of two unrelated strains of H. pylori (Table 1) that are linked to gastro duodenal diseases including peptic ulcer and cancer (31).Almost 50% of the world population, approaching 100% in some countries, is infected. The genomic size of the two strains differ about 24 kb, between 6 to 7 % of the genes were specific to each strain (31).

Eukaryotes encompass uni- and multi-cell organisms, have a true nucleus bounded by envelopes, and undergo mitosis and meiosis. Oxidative enzymes are packaged in mitochondria. Eukaryotes include four kingdoms: Protista, Fungi, Plantae and Animalia. The genome of budding yeast, S.cerevisiae, has 12.07 million bp (12.07Mb) and 17 chromosomes. Genes occupy 70% of the genome. Only 4% contain introns. The fission yeast Schizosaccharamices pombe genome has 13.8 Mb in three chromosomes. 12.5 Mb occupies unique sequences with 4830 protein-coding genes. 43 % of genes of Sch. pombe contain introns. Regions upstream of genes are longer than in budding yeasts (32-33).

C. elegans has 816 cells, of which 302 are neurons. Its genome contains 97Mb and 19,100 protein-coding genes - 5000 bp for a a gene. Exons and introns constitute 24 % of the genome; each gene has an average of 5 introns, as shown in Table 1. Humans have 31,000 genes: 1.5-2 times more than C. elegans and D. melanogaster. This suggests that much of human DNA does not code proteins and is involved in gene/cell regulatory functions or is repeated to different degrees. The origin and role of redundant DNA are unclear.

3.3. E. coli genomics as a model

Sequencing of the circular chromosome of E. coli helps to understand the genome of higher organisms. The length of E. coli is 4,600 kb and includes 4288 genes (31) defined by protein coding sequences (starting from a definite initiation codon, subsequent amino acid coding triplets and ending with a termination codon, which stops translation). Protein coding genes occupy 87.8% of this genome. Intergenic regions occupy 11%. Of 4288 genes, 38% have unknown functions - despite 50 years of study (31).

3.3.1.Template and basic genetic processes

The 4288 genes of E. coli were classified into 22 functional groups, one of which controls essential genetic processes. The cell/genome self-reproduction and transfer of hereditary information is realized by two genetic processes obligate for all cells: (i) template and (ii) basic genetic processes. Template processes include replication (DNA-DNA), transcription (DNA-RNA), and translation (RNA-polypeptide chain). Basic genetic processes include recombination, repair, and segregation of replicated genomes between two daughter cells. Thus we have two triads which are invariant for all living organisms:

1. Template processes – Replication, Transcription, Translation
2. Basic genetic processes – Repair, Recombination, Segregation.

To be hereditary, any DNA change requires cell generation to be capable of maintenance and transfer through the template and basic genetic processes. Mutation is a multi-step, dynamic process. In E. coli, the following functions and the number of genes involved are (31):

- replication, recombination and DNA repair - 115 (2,7%);
- transcription, synthesis and modification of RNA – 55 (1,3%);
- translation and posttranslational protein modification – 182 (4,2 %)
- ribosomal r-RNA – 21 (0.5%)
- t-RNA – 86 (2.0%)

A group of gene/enzyme factors is involved simultaneously in the "three R" processes - replication, recombination and repair. In humans, mitotic DNA replication proceeds rapidly; each second 10 - 20 nucleotides are added at each site. Since $3x10^9$ nucleotides are copied in each division, mistakes frequently occur (9, 36). DNA polymerase, in addition to replication, has proof-reading functions preventing incorporation of the wrong nucleotides. Other biochemical systems recognize and repair DNA damage. Exposure to ultra violet (UV) causes pairing of adjacent thymine molecules in the same DNA strand, forming dimers, which locally distort the DNA and interfere with replication. Dimers can be corrected by DNA repair mechanisms (8, 9, 36).

DNA repair is a powerful source of genomic stability eliminating 90% of DNA damage. Genes encoding DNA repair proteins are "caretakers." Inactivation of a caretaker gene results in instability, an increased mutation rate, and promotion of carcinogenesis and cell senescence (34). Caretaker proteins repair DNA damage caused by mutagenic agents (due to, for example, UV, radiation, drugs) or mistakes in replication (35).

Several disorders are caused by mutations in repair genes. Xeroderma pigmentosa (XP) is an autosomal recessive disorder with a frequency of 1 in 250,000. Affected individuals are sensitive to UV light. Short UV exposure causes dry flaking skin and pigmented spots that can develop into skin cancer at 1000 times the normal rate. Mutations of eight different genes can induce XP due to an inability to repair UV damage (36). Other disorders caused by mutations in repair genes include: Ataxia telangiectasia (hypersensitivity to X-rays, lymphoid cancer), Fanconi anemia (X ray sensitivity, leukemia), Bloom syndrome (UV hypersensitivity, skin cancer) (35).

3.3.2.Cell metabolism, transporters, and paralogous genes

Twenty-five percent of the resources in bacteria are devoted to small-molecule metabolism. These resources include nucleotide biosynthesis and metabolism – 58 genes; amino acid biosynthesis and metabolism - 131 genes. Energy metabolism needs 243 genes or 5.7 % of the genome. One hundred thirty genes participate in carbon compound catabolism; 12% of the genes are involved in large-molecule metabolism; 20% are associated with cell structure and processes and 180 genes control regulatory proteins (31).

If two or more genes in the same species are so similar in nucleotide sequences that they could have originated from a single ancestral gene, they are called paralogs. Human alpha and delta hemoglobin chain loci are examples. A paralogous family is composed of proteins with similar, though not necessarily identical, functions. Paralogs share at least 30% sequence identity over more

than 60% of their length. In E. coli, 1345 proteins have at least one paralogous sequence (31). The largest paralogous family is ATP binding cassette (ABC) transporters - proteins that span plasma membranes and transport molecules in (or out of) the cell. They use ATP energy to pump substrates across the membrane against a concentration gradient. Substrates can be inorganic ions, amino-acids, sugars, or polypeptides (9, 36)

In humans, a transporter mutation leads to the most common hereditary disease of Caucasians – Cystic fibrosis (CF). About 1 in 25 Americans of European descent, 1 in 46 Hispanics, 1 in 60-65 African-Americans, and 1 in 150 Asian Americans is a carrier for an abnormal CF gene located on the long arm of chromosome 7. Its 27 exons encode a protein of 1408 amino acids named CF trans-membrane conducting regulator (CFTR). It inserts into the plasma membrane of exocrine gland cells and regulates chloride ion flow.

A defective CFTR protein decreases secretions of exocrine glands. The CF gene is expressed in mucus-secreting epithelial cells (e.g., in bronchi, salivary glands, sweat glands, testes, and intestines). For example, CF causes the production of thick mucus that clogs ducts carrying digestive enzymes from the pancreas to the small intestine and reduces digestive efficiency. CF children suffer from malnutrition despite increased appetite and food intake. Cysts form in the pancreas and it degenerates. Because of thick mucus in the lungs, CF patients also develop obstructive lung diseases and are susceptible to infections. In males, mucus blocks ducts that carry sperm, producing infertility. In CF women thick mucus may plug the entrance to the uterus, reducing fertility (9, 36).

4. HUMAN GENOME STRUCTURE

4.1. Main classes of DNA sequences

The human genome is 20-30 times larger than that of Drosophila or C. elegans. It contains 30,000 – 40,000 protein-coding genes - twice as many as Drosophila or C. elegans. Coding sequences in humans are less than 5% of the genome; 15-20% of the genome is connected with gene/chromosome activity regulation. Facultative elements (50% of the genome) include highly repeated DNA, duplications of genes and chromosomal segments, and mobile elements or ME. ME can be one of four types: three types transpose through RNA intermediates (retroposons) and one transposes as DNA (transposons). Retroelements are labeled LINE (long-interspersed elements) and SINE (short-interspersed elements). LINE families are 6kb long and include two open reading frames (ORFs) with 10,000 copies per genome. SINE is shorter than 500 bp and is present as millions of Alu elements. Long Terminal Reprints (LTRs) are the third retroposon. Most retroviruses have LTRs on their ends and are transposable. The human genome contains at least seven classes of DNA transposons. Currently recognized LINEs, SINEs, LTR retroposons, and DNA transposons comprise 20%, 13%, 8% and 3% of the genome. Hundreds of human genes resulted from horizontal transfer from bacteria during vertebrate evolution (37).

Segmental duplications – transfer of 1-200 kb blocks to one or more locations in the genome - are a remarkable feature of the human genome. These duplications can be divided into inter- and intra-chromosomal types and comprise about 3.3% of human DNA. Many inter-chromosomal duplications map near the centromere or telomere regions. A 9500 bp segment containing an adrenoleukodystrophy locus from X-chromosome Xq28 has been duplicated near the centromeres of chromosome 2, 10, 16, and 22 (37).

Intra-chromosomal duplications include repeats associated with genetic disease. Chromosome 17 contains three copies of a 200-kb repeat and two copies of 24- kb repeats - each separated by 1.5 Mb segments. Recombination between paralogous duplications is associated with Smith-Magenis, Charcot-Marie-Tooth or Prader Willi/Angelman syndromes (9, 36).

The human genome has thousands of genes producing non-coding RNA (ncRNA). There are several classes of ncRNA: 1) transfer RNA (tRNA) ; 2) ribosomal RNA (rRNA), which is important in translation; 3) small nucleolar RNA required for RNA processing; and 4) small nuclear RNA (snRNA), a critical component of large nucleoprotein complexes (spliceosomes) that splice introns out of pre-mRNA in the nucleus. Other ncRNAs include nucleoproteins, such as telomerase RNA or signal transduction particles (STP). All cells have STP that select proteins to be either secreted or integrated into the plasma membrane and target them to the endoplasmic reticulum (or plasma membrane in prokaryotic cells). There are also ncRNAs of enigmatic function (such as large Xist transcript from active X-chromosome in females connected with X-dosage compensation (37)

Other repetitive FE are simple sequence repeats (SSRs), which are perfect or slightly imperfect tandem repeats. SSRs are important in human genetic studies because they show a high degree of length polymorphism in populations and promote molecular localization. SSRs are the “workhorses” of human disease studies. SSRs with a short repeat unit (1-13) are called microsatellites. Longer SSRs (14-500) are minisatellites. SSRs comprise 3% of the human genome, with the greatest number of dinucleotide repeats. SSRs occur because of slippage during DNA replication. There is one SSR per 2 kb (37).

4.2. Gene sizes, exons, and introns

The size of genes and introns varies, with both distributions having long tails. Many human genes have more than 100 kb. The dystrophin gene, which results in X-linked recessive Duchenne muscular dystrophy (DMD), has 2.4 million bp (comparable to the whole E. coli genome). It is the largest known gene with 79 exons, of which 99% are introns.

RNA polymerase II needs 16 hours to transcript this gene. Dystrophin contains 3,685 amino-acids and represents 0.002% by weight of muscle. It attaches to the cytoplasmic side of the plasma membrane in muscle cells and stabilizes the membrane during contraction. If

dystrophin is absent or defective, the plasma membrane breaks down, causing the death of the muscle tissue. Less severe mutations produce too little dystrophin, resulting in a milder disease called "Becker muscular dystrophy."

The transcript of the DMD gene is formed by 70 exons. Shorter isoforms of dystrophin are found in brain, liver and other tissues. These isoforms result from promoters residing in DMD genes and alternative splicing. There is a mutational hot spot between exons 44 and 45. Fifty percent of patients have deletions in this 2kb region (9, 36). Progressive muscle weakness is an early sign of DMD (35).In the USA DMD affects 1 in 3500 males and presents between 1 and 6 years of age

The variation in coding sequence size (exons) is less extreme. The titin gene, the longest size (at 80 780 bp), is located on the long arm of chromosome 2 and codes the largest protein, with 27,000 amino acids. Each molecule is larger than one micrometer. Titin acts as a spring pulling the muscle fiber back into shape after being stretched and has a variety of isoforms generated by alternative splicing. It has the largest number of exons (178) and the longest single exon 17 106 bp.(37).

A comparison of human genes with those from C. elegans and drosophila shows that their exon lengths are similar (1311 bp for C. elegans, 1497 bp for drosophila and 1340 bp for humans). The average length of internal exons is 50-200 bp. Intron size is more variable in humans with a peak at 87 bp but a long tail resulting in a mean of more than 3300 bp (37).

The set of human proteins (the "proteome") is more complex than in drosophila or C. elegans. This is due not only to vertebrate-specific protein domains and motifs, 7% of the proteome, but also because vertebrates develop rich gene/protein networks. An alternative splicing was found for 59% of human genes with an average of 2.6 transcripts per gene. These figures are higher than for drosophila and C. elegans (22% of genes). 70% of alternative splice forms affect coding sequences; terminal exon splicing was seen for 20% of genes. Protein variety is greater than gene variety in humans (37).

5. THREE TYPES OF HERITABLE CHANGES

5.1. Obligate and facultative elements: mutations and variations

Absence of the strict correlation of DNA content in the haploid genome with taxonomic status of species and 2-to-5-fold differences in genome size in closely related species is called the "C-paradox", (9, 36) i.e. violations of classical genetics in which: (i) all genetic material of chromosomes consists of genes (all DNA has information function) and (ii) the list of genes with alleles corresponds to genotypes.

Many structural aspects of eukaryotic genomes appear paradoxal: mosaic gene structure; programmed developmental DNA rearrangements; ability of local DNA segments for autonomous replication both intra-chromosomally or forming the linear or circular DNA amplified repeats (amplicons) in cytoplasm; abundance of transposable elements to diverse intra-cell incarnations; and presence of viruses and cytobionts (37-38). Simplified notions of genetic determinants of senescence and longevity such as the Hayflick (1961) limit are insufficient to describe stochastic inheritance of such complex traits as longevity in humans. That is, instead of a determinate genetic structure causing genetic changes, it it is possible that loss of fidelity in gene-protein translation may be the primary "genetic" factor in senescence (due to accumulation of stochastic errors).

The discovery that protein coding loci in the human genome look like islands in an ocean of other DNA sequences was unexpected (3, 13,14). One explanation is to view redundant DNA as "selfish DNA," (9) which occurs due to mistakes in the template and basic genetic processes. Some DNA sequences or definite chromosomal segments may be advantaged in reduplication and amplify. Repeated DNA sequences may "automatically" increase in number due to unequal recombination. One way DNA length increases is by means of transposable elements inserted at different sites of the host genome. Retroviruses are LTR containing transposable elements. (3, 5-9, 13).

Every genome carries a subsystem of facultative elements. It is instructive to separate Obligate Elements (OE) from FE (38-39). FE includes the hierarchy of intra and extra chromosomal elements. In the nucleus FE comprise a highly repeated and satellite DNA, pseudogenes and retrotranscripts, transposons, amplicons and additional chromosomes (B-chromosome). In cytoplasm FEs include plasmids, amplified segments, endosymbionts (3, 5-9, 13, 38-39).

OE and FE exhibit different patterns of heritable alterations. Point mutations, classically, acorrespond to OE, (i.e., changes in structure, position and number of genes, in definite chromosomes sites). The situation with FE is different. The LINE-1 retrotransposon, L1, has in the human genome 3000-5000 full length copies and 500,000 truncated copies -- 15-17% (39). Active L1 elements are transposed in male germ cells and result in insertion mutations and disease. A million Alu retrotransposon elements comprise 10 to 12% of the human genome (40). One of every 100-200 human births has a de novo Alu insertion. A similar rate is found for the L1 retro-element (41-42). Thirty-three insertion mutations mediated with L1 elements were observed, leading to diseases such as hemophilia A and B, thalassemia and DMD. L1- and Alu-mediated insertion mutations occur once in every 50-100 human cells. (39-40). A large number of homologous L1 and Alu sequences lead to mutation through mis-pairing and unequal cross-over (42).

Alterations of FE intra-genome population and their response to external factors are different from gene mutations so Jacob and Wollman called them variations (44). Variations usually occur in the intra-cell population of FE (e.g., hybrid dysgenesis in Drosophila). P-M dysgenesis in F1 hybrids from crosses of paternal P-stock containing

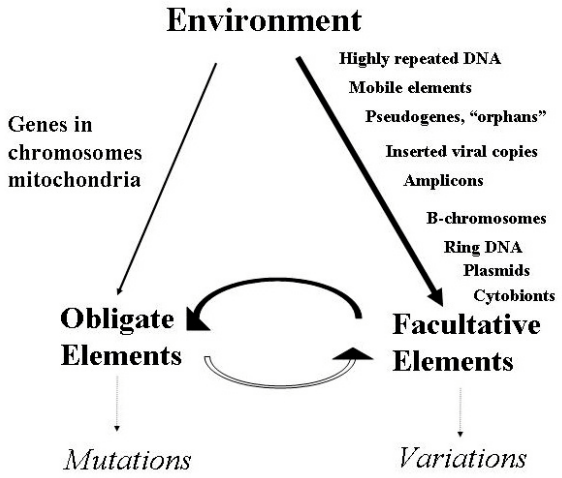

Figure 1. Obligate and Facultative elements of the genome and two types of hereditary changes. Associated with them are two types of hereditary changes: mutations and variations. Arrows indicate the direction of the links, while their width corresponds to the intensity of their force. Some facultative elements are more sensitive to the environment, and their activation leads to gene/chromosome insertions and mutations.

active P-transposons with females of M-stock devoid of P-active copies and the cytoplasmic transposition repressor generates P-element mobility. This situation produces insertion mutations and chromosomal rearrangements, damaging germ cells and causing sterility. Their incidence in F1 progeny of dysgenic crosses may reach 10%. Chromosomal breaks are ordered and site-specific: They occur near P-elements. Multi-site inversions occur as often as single ones (43). Activation of ME may result in mutation outbreaks in populations (45).

Variations of FE hereditary changes: occur (i) simultaneously in many individuals, (ii) are site-specific, and (iii) may be induced by non-mutagenic factors as temperature fluctuationsor interline crosses. Interactions are illustrated in Figure 1. There is a two-step mechanism of spontaneous mutation occurrence. First, diverse FE are activated by environmental fluctuations. Second, changes in number and location of FE result in insertion mutations and chromosomal rearrangements (3-11, 13, 14, 38, 46).

Another example of OE and FE differences is the localized amplification of chromosomal segments during the adaptation of somatic cells to drugs that block cell divisions, (such as cytostatics). DNA loci containing the gene conferring resistance to a cytostatic agent may amplify. Amplified segments (amplicons) can be located in tandem chromosomal blocks or transformed into plasmid or mini-chromosomes capable of autonomous replication in cytoplasm. Both the number and topography of amplicons varies over cells and cell lines. Amplification of chromosomal segments containing the myc oncogene occurs in 25 % of human neuroblastomas. Amplified copies can be mini-chromosomes or micronuclei in the cytoplasm. The number of amplified segments can not be determined even in daughter-cloned cells (13, 43, 44).

5.1.1. Stochastic approach to the genome

Finch and Kirkwood (47) in their book discussed the role of chance events and stochasticity in living processes, including development and aging. They analyzed why genetically identical human twins and highly inbred laboratory animals show wide individual differences in patterns of aging over life spans. An additional previously non-discussed stochastic force consists of the existence in the genome of many families of diverse FE with varying number, size and cell topography (both in nucleus and cytoplasm) indicating a stochastic aspect of its structure and hereditary variability. The behavior of FE in the human genome might be associated with complex stochastic processes.

In place of discrete categories of integral heritable elements we would use lower dimensional fuzzy classifications of elements that allow for continuous evolution of the crisp set of traits for prototype elements (e.g., 48). More specifically, we can represent observed multi-dimensional phenotypic traits as functions of vectors of the probabilities of specific genetic traits and the K dimensional stochastic state vector g_{ik},

$\text{Prob}(x_{ij\ell} = 1) = (\sum_{k=1}^{K} g_{ik}\lambda_{kj\ell})$,

where $0 \le g_{ik} \le 1.0$ and $\sum g_{ik} = 1.0$ and $\sum \lambda_{kj\ell} = 1.0$ which are evaluated by maximum likelihood (48). The simplest case is to assume that the algebra of the $\lambda_{kj\ell}$s is determined by classical genetic principles and that g_{ik}s are individual scores indicating how stochastic factors and thermodynamics mix the K sets of classically determined trait profiles for each individual. This would also involve simplification of the plethora of integral genetic components (especially FE) by filtering stochastic factors through the K dimensional filter of g_{ik}s. If the g_{ik}s show "drift" with the passage of chronological time, this would be a way of distinguishing the effects of aging on genetic expressions (i.e., the $g_{ik} \to g_{ik}(t)$ where t is a function of time/age). Changes in $\lambda_{kj\ell}$ reflect changes in genetic contribution, and changes in $g_{ik}(t)$ represent intra and extra-cellular factors affecting gene expression.

We could use such models to simplify the theoretical model of heritability with a plethora of new molecular products of intra-cellular dynamics that classical genetics is not capable of handling. This is especially true when the stochastic process of intra-cellular mechanics is non-stationary (as in aging human cells).

5.2. Epigenetic inheritance

Classic genetic analysis of DNA structural mutations is "just part of the story". There exist both

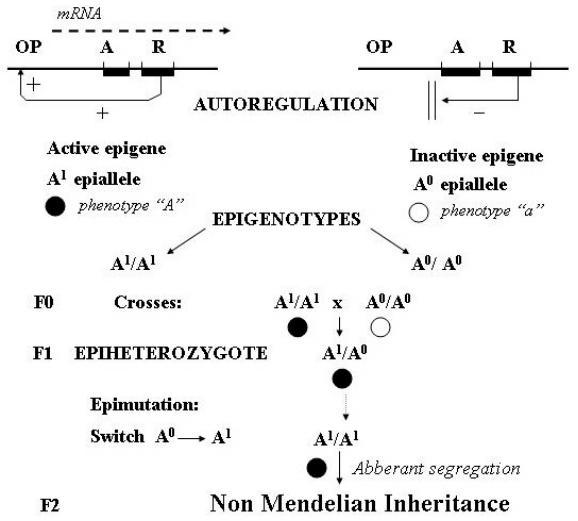

Figure 2. A scheme of epigenes. Epigenes constitute autoregulatory feed back links via DNA binding proteins. The figure shows the positive auto-reregulation at the transcription level . A^1 – active state of the epigene, A^o – non-active state. Crosses result in the epigenotype where, due to the transition of one epiallele (A^0 from non-active to active state), "epimutation" occurs via the process of trans-activation.

structural and dynamic modes of coding, storage, and transfer of hereditary information. Dynamic aspects are called epigenetic. The spectrum of epigenetic inheritance is wide: from the transformation of serotypes in the paramecium to chromosomal and genomic imprinting. Terminology has not stabilized. "Epigenotype" was used in the 1960s; Holliday used "epimutation" and indicated DNA methylation as the main source (17-18). In the 1970s, Tchuraev used "epigene" to suggest its artificial synthesis and validated this idea in 2000 (49, 50).

The epigene is an auto-regulatory hereditary unit, a genetic system with cyclic links or feedback, having two or more functional states and capable of maintaining each state over cell generations. A one-component system is shown in Figure 2. Feedback might be positive as in the auto-regulatory c1 gene where state is determined by the λ phage regime (lysogenic or lythic). Auto-regulation may be negative as in transposon Tn3 in E coli. Transposons P in Drosophila and Ac and SpM are organized as epigenes with positive auto-regulation (51, 52). Thus the integral element of heredity here is not a gene element, but a stochastic subsystem.

Figure 2 shows stochastic switching from inactive to active state in positive auto-regulation in cell epiheterozygotes A^1/A^o (where upper indices '1' means active and '0' means inactive). Switching is an epimutation. Epimutation in multi-cell organisms may occur in somatic and germ cells - with non-mendelian inheritance resulting as the influence of heterozygous conditions on allelic structure and function of the next generation of epiheterozygotes.

Also important are independent epigenes. If we have 10 epigenes each with two states we have 2^{10} or 1024 states without changing DNA (3, 14). "Epigenetics" - as well as variations - suggest the need for a stochastic fuzzy state principle to develop a parsimonious theoretical model of heredity.

5.3 Methodological Implications

The growing complexity of the monitoring and repair functions of the genome-protein complex means that we must use different types of mathematical models to analyze hereditary changes and genome function. The increased complexity of genetic elements resembles the proliferation in the number of sub-particles thought to exist in high energy physics to preserve the myth of a closed system. It is unclear whether the continued fracturing of this genetic structure identifies truly distinct structural elements or whether it reflects molecular uncertainty and stochasticity in genome mapping and partitioning functions. With this new complexity, classical determinate mechanics, as used in the macro physical world, do not apply. The extent which the new units identified, especially those without identified functions, are "integral" (or are an artificial partitioning of chance molecular aggregates generated by a stochastic process of micro-molecular thermodynamics in cell partitions) is unclear. An elegant and manageable mathematical theory must deal with this stochastic uncertainty.

6. BIODEMOGRAPHIC IMPLICATIONS

The discussed generalized approach to the genome organization and function suggests three types of heritable changes: mutations, variation and epigenetic alterations . This has important epidemiological and biodemographic data collection implications. Heritable FE changes (variations) may occur in many individuals, is site-specific and stochastically induced by weak, non-mutagenic factors such as temperature, inter-population crosses, or nutritional shifts ("genotrophes" in plants). The same concerns epigenetic alterations. Consider the two-step pattern of mutation occurrence in nature. Usually FE - and especially diverse ME - are sensitive to weak "non-mutagenic" changes of external and genotypic milieu. Activation increases mobility with insertion mutations and mutability outbreaks. This was studied in geographically isolated Drosophila populations (37, 45). Insertions in Drosophila are 70 % of spontaneous visible mutations and chromosomal rearrangements (3, 7-9, 13). In epidemiology and biodemography this can be manifested by unpredictable differences in incidence of phenotypic and heritable disorders between cohorts.

Second, it is well known that an outbreak of mobility occurs in hybrid crosses of wild stock Drosophila with active P-transposons with females of M-stock that are devoid of cytoplasmic repressor of transposition (9, 13). This results in frequent P-insertions in many loci, multi-site chromosomal rearrangement and, as sequelae to increased variability, damage to germ cells and hybrid sterility.(13) A similar phenomenon may occur in inter-ethnic or inter-racial crosses. The first case of multiple mutations in the

DMD sex-linked locus was in an extended inter-ethnic family. Molecular studies showed that one boy has de novo duplication; another, deletion; and a third, point mutation (53).

The third implication is the association of senescence with non-controlled transposition and multiplication of ME. Senescence of eukaryotic organisms may be connected with rapid chaotic propagation of egoistic FE. Elimination of these elements during cell cycle, or meiosis, could be viewed as "rejuvenation" (54).

The fourth implication is associated with diverse epigenetic changes and regulation of DNA silencing in normal development and oncogenic cell transformation. Developmentally induced epigenetic changes (epimutations) may be propagated through meiosis. DNA silencing (due to methylation of promoter regions of both oncogenes and tumors) is common and regular events (55) as summarized in "Epigenetics in Cancer Prevention" (56). In one example, 12 genes selected from major signal transduction pathways in the cell (including cell cycle, DNA repair, cell adherence, apoptosis) all had control gene areas that are methyltransferase targets (57). Silencing these genes is associated with cancer risk.

Finally, we argue that studies need a three-generation approach because: (i) individual development begins before fertilization and includes all gametogenesis and meiosis processes started in the female grandmother and (ii) germ line extraction and maturation include two genome reprogramming events: erasure of parental imprints and epigenetic materialization. Each egg physically and genetically links three generations. On the epidemiological and biodemographic levels, pregnancy conditions of the F (n-2) generation (grandmother) may influence the phenotypic manifestation and health status of the F(n) generation (grandchild). This is analyzed in (58).

7. ACKNOWLEDGEMENTS

The work in this paper was supported by the following grants: K12-AG-00982 - Scientific recruitment for the demography of aging and 5U01-AG-007198-16A1-Functional and health changes of the elderly.

8. REFERENCES

1. Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3, 318-356 (1961)
2. Mono J, F.Jacob: General conclusion: teleonomic mechanisms in cellular metabolism, growth and differentiation. In: Cold Spring Harbor *Symp Quant Biol* 26, 389-401 (1961)
3. Golubovsky M D: A century of genetics: evolution of ideas and concepts. Borey (publisher), St. Petersburg, Russia (2000)
4. Shapiro J.A: Genome organization, natural genetic engineering, and adaptive mutation. *Trends in Genetics* 13, 98-104 (1997)
5. J A Shapiro: Natural genetic engineering, adaptive mutation & bacterial evolution. In: Microbial Ecology and Infectious Disease. Eds: Rosenberg E. ASM Press, Washington, pp. 259-275 (1999)
6. Shapiro J A: Genome system architecture and natural genetic engineering in evolution. *Annal N Acad Sci* 870, 23-35 (1999)
7. Shapiro JA: Transposable elements as the key to a 21st century view of evolution. *Genetica* 107 (1/3), 171-179 (1999)
8. Shapiro J A: Genome Organization and Reorganization in Evolution: Formatting for Computation and Function. *Ann NY Acad Sci* 981, 111-134 (2002)
9. King R C, W D Stansfield: A dictionary of genetics (sixth edit.). Oxford Univ Press (2002)
10. Caporale L.H: Mutation is modulated:implications for evolution. *BioEssays* 22, 388-395 (2000)
11. Fedoroff, Nina: The Dynamic Genome:Barbara McClintock's ideas in the century of genetics. Ed.: Bolstein D., Cold Spring Harb.Lab.Press (1991)
12.. Ryan A., J.Mehl, Christopher Anderson, Stephen W. Santoro, Lei Wang, Andrew B. Martin, David S. King, David M. Horn, and Peter G. Schultz: Generation of a Bacterium with a 21 Amino Acid Genetic Code. *J Am Chem So* 125 (4), 935 -939 (2003)
13. Golubovsky M.D.: Genome Inconstancy by Roman B.Khesin in terms of conceptual history of genetics. *Molec Biol* 36, 2, 259-266 (2002)
14. Nebert D.W.Transcription factors and cancer:an overview. *Toxicology* 181-182, 131-141 (2002)
15. Hanahan D., R.Weinberg. The Hallmarks of cancer. *Cell* 100, 57-70, (2000)
16. Chin J.W, A.Cropp, C.Anderson, M.Mukherji, Z.Zhang, P.Schultz. An expanded eukaryotic genetic code. *Science* 301, 964-967, (2003)
17. Holliday R: The inheritance of epigenetic defects. *Science* 238, 163-170 (1987)
18. Holliday R: Mechanisms for the control of gene activity during development. *Biol Rev* 65, 431-470 (1990)
19. Bird A: DNA methylation patterns and epigenetic memory. *Genes Dev* 16, 6-21 (2001)
20. Jones PA: Takai D: The role of DNA methylation in mammalian epigenetics. *Science* 293, 1068-1074 (2001)
21. Jenuwein T, Allis D: Translating the histone code. *Science* 293, 1074-1080, (2001)
22 Ferguson-Smith A.C, Surani A: Imprinting and epigenetic asymmetry between parental genomes. *Science* 293, 1086-1087 (2001)
23. Surani A: Reprogramming of genome function through epigenetic inheritance. *Nature* 414, 122-128 (2001)
24. Reik W: Walter J. Genomic imprinting:parental influence on the genome. *Nature Rev* 2, 21-32 (2001)
25. Kierszenbaum A L: Genomic imprinting and epigenetic reprogramming: unearthing the garden of forking paths. *Mol Reprod Dev* 63, 269-272, (2002)
26. Jaenish R.,Bird A: Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet Suppl* 33, 245-254 (2003)
27. Reik W: Mammalian epigenomics:reprogramming the genome for development and therapy. *Theriogenology* 59, 1-22, (2003)
28. Jones PA, Baylin S: The fundamental role of epigenetic events in the cancer. *Nat Rev Genet* 3, 416-428 (2002)

29. Jaenisch R, Wilmut I: Don't clone humans. *Science* 291, 2552 (2001)
30. Rideout W.M., Kelvin E., Jaenisch R: Nuclear cloning and epigenetic reprogramming of the genome. *Science* 293, 1093-1098 (2000)
31. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: The complete genome sequence of Escherichia coli K12. *Science* 227, 1453-1462 (1997)
32. Alm R.A., Trust T.J.: Analysis of the genetic diversity of Helicobacter pylori:the tale of two genomes. *J Mol Med* 77, 834-846 (1999)
33. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, Hornsby T, Howarth S, Huckle EJ, Hunt S, Jagels K, James K, Jones L, Jones M, Leather S, McDonald S, McLean J, Mooney P, Moule S, Mungall K, Murphy L, Niblett D, Odell C, Oliver K, O'Neil S, Pearson D, Quail MA, Rabbinowitsch E, Rutherford K, Rutter S, Saunders D, Seeger K, Sharp S, Skelton J, Simmonds M, Squares R, Squares S, Stevens K, Taylor K, Taylor RG, Tivey A, Walsh S, Warren T, Whitehead S, Woodward J, Volckaert G, Aert R, Robben J, Grymonprez B, Weltjens I, Vanstreels E, Rieger M, Schafer M, Muller-Auer S, Gabel C, Fuchs M, Dusterhoft A, Fritzc C, Holzer E, Moestl D, Hilbert H, Borzym K, Langer I, Beck A, Lehrach H, Reinhardt R, Pohl TM, Eger P, Zimmermann W, Wedler H, Wambutt R, Purnelle B, Goffeau A, Cadieu E, Dreano S, Gloux S, Lelaure V, Mottier S, Galibert F, Aves SJ, Xiang Z, Hunt C, Moore K, Hurst SM, Lucas M, Rochet M, Gaillardin C, Tallada VA, Garzon A, Thode G, Daga RR, Cruzado L, Jimenez J, Sanchez M, del Rey F, Benito J, Dominguez A, Revuelta JL, Moreno S, Armstrong J, Forsburg SL, Cerutti L, Lowe T, McCombie WR, Paulsen I, Potashkin J, Shpakovski GV, Ussery D, Barrell BG, Nurse P, Cerrutti L: The genome sequence of Schizosaccharomyces pombe. *Nature*, 415, 871-880 (2002)
34. Deccottignies A, Sanchez-Perez I, Nurse P.: Schizosaccharomyces pombe essential genes. A pilot study. *Genome Res* 13, 399-401 (2003)
35. Kinzler K.W., Vogelstein B. Gatekeeprs and caretakers. *Nature* 386,761-763 (1997)
36. M R Cummings: Human Heredity.Principles and issues. Fifth edit. Brooks/Cole Press, CA (2000)
37. International Human Genome Sequencing Concortium: Initial sequencing and analysis of the human genome. *Nature* 409, 6822, 860-921 (2001)
38. Golubovsky M.D.: Organization of the genotype and forms of hereditary variability in eukaryotes. *Adv Curr Biol* (Moscow) 100, 323-330 (1985)
39. Golubovsky M.D.: Mobile elements and forms of heritable changes in eukaryotes. *Biopolymeres and cells* 11 (2), 29-38 (1995)
40. Ostertag F.M., Kazazian H.H.: Biology of mammalian L1 retrotransposons. *Annual Rev Genet* 35, 501-538 (2001)
41. Prack E.T.,Dodson A,W., Farkash E.A., Kazazian H.H.: Tracking an embryonic L1 retrotransposition event. *Proc Natl Acad Sci* 100, 1832-1837 (2003)
42. Tomilin N.V.: Control of genes by mammalian retroposons. *Intern Rev Cytol* 186, 1-48, (1999)
43. Deininger P.L., Batzer M.: Mammamlian retroelements. *Genome Res* 12, 1455-1465 (2002)
44. Jacob F., Wollman E.: Sexuality and genetics of bacteria. Acad Press, New York (1961)
45. Berg R.L., Engels W.R., Kreber A.: Site specific X-chromosome rearrangements from hybrid dysgenesis in Drosophila melanogaster. *Science* 210, 427-429 (1980)
46. Zakharenko LP, Gracheva EM, Romanova OA, Zakharov IK, Voloshina MA, Kochieva EZ, Simonova OB, Golubovsky MD, Georgiev P.Hobo induced rearrangements are responsible for mutation bursts at the yellow locus in a natural population of Drosophila melanogaster. *Mol Gen Genet* 263, 335-341 (2000)
47. Finch C.E, T.Kirkwood . Chance, Development and Aging.Oxford Univ.Press (2000)
48. Manton K.G., M.A.Woodbury and H.D.Tolley. Statistical Application Using Fuzzy Sets. New York, Wiley (1994)
49. Tchuraev R.N.: The hypothesis on epigene. Studies on mathematical genetics. *Novosibirsk* (1975)
50. Tchuraev R.N., Stupak I.V., Tropunina T.S., Stupak E. Epigenes: design and construction of new hereditary units. *FEBBS* 486 (3) (2000)
51. Rio D.: Regulation of Drosophila P element transposition. *Trends Genet* 9, 282-287 (1991)
52. Gierl A.: How maize transposable elements escape negative selection. *Trends Gene* 5, 155-158 (1990)
53. Miciak A., Keen, A, Jadayel, D, Bundey S.: Multiple mutations in an extended Duchenne muscular dystrophy family. *J Med .Genet* 29, 2, 123-126 (1992)
54. Vakhtin Yu.B: Senescence as metagenic process: *Tsytologia* (*Cytology*, Russ) 44 (2) 218-223 (2002)
55. Jones P.A., Baylin S.: The role of DNA methylation in mammalian epigenetics. *Science* 293, 1068-1074 (2001)
56. Epigenetics in Cancer Prevention: Early detection and risk assessment. *Ann NY Acad Sci* v. 983. N.Y. (2003)
57. Kling J.: Put the blame on methylation. *The Scientist* 17 (12) 29-30 (2003)
58. Golubovsky M.D., K.Manton. A three-generation approach in biodemography is based on developmental profiles and the epigenetics of females. *Front Biosci* 10, 187--191, (2005)

Key Words: Genome, Facultative Elements, Mutation, Variations, Epigenetics, Heritable Changes, Epigenetic Inheritance, Reivew

Send correspondence to: Michael Golubovsky, Ph.D. Duke University, Center for Demographic Studies, 2117 Campus Drive Box 90408, Durham, NC 27708, Tel: 919-682-2949, extension 223, Fax: 919-684-3861, E-mail: mgolub@cds.duke.edu

http://www.bioscience.org/current/vol10.htm