Anais da Academia Brasileira de Ciências (2005) 77(4): 627-650
(Annals of the Brazilian Academy of Sciences)
ISSN 0001-3765
www.scielo.br/aabc

# Evolutionary change – patterns and processes

FRANCISCO M. SALZANO*

Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul
Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brasil

*Manuscript received on March 30, 2005; accepted for publication on April 6, 2005*

**ABSTRACT**

The present review considered: (a) the factors that conditioned the early transition from non-life to life; (b) genome structure and complexity in prokaryotes, eukaryotes, and organelles; (c) comparative human chromosome genomics; and (d) the Brazilian contribution to some of these studies. Understanding the dialectical conflict between freedom and organization is fundamental to give meaning to the patterns and processes of organic evolution.

**Key words:** molecular evolution, principles of evolution, evolution, plants, human evolution.

## IN THE BEGINNING – FROM NON-LIFE TO LIFE

It seems that everything started with a marvelous explosion. Although there are divergent views, the universe should have emerged from a single, unbelievably small, dense, hot region, 10 billion or more years ago (Halliwell 1991, Adams and Laughlin 2001). Everything occurred in very, very short periods of time and at extremely high temperatures. The corpuscles of matter present at that moment are called *quarks* [presumably from a sentence of Johann Wolfgang von Goethe (1749-1832) who would have said "Poor God, he puts the nose in every quark", meaning the smallest particle of matter]. Eight types of quarks have been identified (*up, down, charm, strange, bottom*, and *top*), the last one being clearly established at the beginning of 1995 only (Nascimento 1995).

An interpretation of what happened in the next chain of events depends on the hypotheses established to explain the universe's structure. Three of them were that the universe: (a) is *open*, and its expansion is indefinite; (b) *closed*, the period of expansion being finite – it is postulated that a new contraction, similar to that which existed before the Big Bang, will occur; and (c) it is *flat*, a situation intermediate between the other two. Presently it is believed that the universe does not contain sufficient energy to be closed, and that it is in permanent expansion (Adams and Laughlin 2001).

The next big transition was the origin of life. Since our solar system, and with it the earth, was formed at about 4.5 billion years ago, and we have evidence of life in fossils dated around 3.5 years ago, it is clear that life started early in the history of our planet. The fact that ribonucleic acid (RNA) can duplicate without the help of the other molecules, and that ribozymes can perform enzyme-like reactions, led to the idea that in the beginning there was a RNA World. Joyce and Orgel (1999) then described how this World could be formed, by what they called "The Molecular Biologists' Dream", using extrapolations from prebiotic chemistry and directed RNA evolution.

*Member, Academia Brasileira de Ciências
E-mail: francisco.salzano@ufrgs.br

The sequence of events could be envisaged as follows: (a) nucleoside bases and sugars, either developed on earth or received from the outside space were formed; (b) nucleotides were then synthesized and stored in pools; (c) a mineral catalyst at the bottom of the pool (for example, montmorillonite) afterwards participated in the formation of long single-stranded polynucleotides, some of which were then converted to complementary double strands by template-directed synthesis; (d) at least one of these double-stranded RNAs on melting would yield a single-stranded rybozime capable of copying itself and its complement; (e) repeated copying would lead to an exponentially growing population; and (f) the RNA pre-organism would be enclosed by a membrane (Orgel 2004). From the very beginning, natural selection was a key factor in the whole process (De Duve 2005).

Alternative hypotheses (protein-first), however, exist. Berger (2003) developed the idea that at the beginning primitive enzymatic sites would have been formed by abiotic amino acids specifically gathered around different substrates. These proteinoids would have then been transferred to messenger-like RNAs by a mechanism reverse to that of the present protein synthesis, and after that to DNA.

Other questions: was hot or cool the most appropriate environment for the origin of life? Bada and Lazcano (2002) favored the second alternative; they also suggested that life may have originated several times. Cavalcanti et al. (2004) considered the origin and evolution of the genetic code. They maintained that an initial version, containing fewer amino acids than the present 20, was modified by the incorporation of new ones after the duplication and divergence of previous synthetases and tRNA molecules.

If evolution is a fact, who was the universal ancestor? Extensive comparisons of genome sequences from widely different modern organisms identified only 60 genes (mostly involved in the translation process) that appear to be universal. The minimum number needed by a self-sufficient organism, however, is one-order of magnitude higher (600) (Whitfield 2004). How can this be explained? Woese (1998, 2002) suggested that the universal ancestor was not a discrete entity, but a community of cells (a supramolecular aggregate) that survived and evolved as an unit. Initially both mutation rate and lateral gene transfer levels were elevated. As increasingly complex and precise biological structures evolved, the two above-indicated processes diminished in number. Modern cells began to evolve with the origin of the translation process. Vertically generated novelty assumed greater importance, until a threshold was reached, called by Woese (2002) the Darwinian threshold. After that, the more orthodox processes of mutation and selection that are identified today became the key evolutionary factors.

Wonderful, but how all of this reasoning can be tested? A large number of researchers are now trying to synthesize artificial cells in the laboratory. Two approaches are being followed. The top-down type of study tries to create artificial cells by simplifying and genetically reprogramming existing cells with simple genomes; while the bottom-up experiments start from the very beginning, using nonliving organic and inorganic materials. Presently the second approach is the most explored (Rasmussen et al. 2004).

## GENOME STRUCTURE AND COMPLEXITY

### Data Availability and Methods

The enormous progress made in relation to the understanding of the DNA, RNA, and protein changes at the molecular, cellular, tissue, and organism levels is remarkable. This was due, to a great extent, to the development of techniques of laboratory and informatic analyses. A general review of them is provided by Saccone and Pesole (2003) who considered in detail: (a) DNA sequencing; (b) gene expression; and (c) proteomic techniques. They also provided much information about databases and the computational tools that are being used for the analysis of the results. Other more restricted reviews can be found in the Pennacchio and Rubin (2003), Snyder and Gerstein (2003), Ureta-Vidal et al. (2003), Haubold and Wiehe (2004), and Miller et al. (2004) contributions. The latter also considered what future developments can be expected in this area. Soltis et al. (2004), on the other hand, stressed that independently of the techniques, adequate taxon sampling is vital for correct evolutionary implications.

The dialectic dilemma quantity or quality is pervasive in all the history of biological evolution, and has been variously considered in relation to genomics in recent years. Independently of the differences in size and complexity of the genomes of diverse organisms, all DNA has to be divided in small pieces for sequence determination. And the first decision that has to be made is whether a clone-by-clone shotgun (CCS) or a whole-genome shotgun (WGS) method should be employed. The latter (WGS) is simpler than the former, and has been adopted not only for the establishment of one of the versions of the human genome, but also for the investigation of a wide array of other organisms. Venter et al. (2003) provided an overview of these results. But there are doubts about the WGS efficiency in providing a really reliable minute version of a genome. The controversy about the quality (and independence) of the human genome sequences generated by the public consortium and the Celera Company is not new, and can be assessed in Waterston et al. (2003) and Adams et al. (2003) papers. Schmutz et al. (2004a) and She et al. (2004) considered the question of the quality of the human genome data in detail. The latter authors verified that large (greater than 15 thousand bases, or kb) and highly similar (greater than 97%) duplications are not adequately resolved by WGS assembly. A mixed strategy, using a targeted clone-by-clone approach to resolve duplications was proposed.

The fact is that database information is increasing in logarithmic progression, and much of these data have not been adequately and comprehensively analyzed. An attempt in this direction was provided by Driskell et al. (2004). They considered the phylogenetic potential of about 300 thousand proteins sequences stored in Swiss-Prot and GenBank, and stressed that more than 100 thousand species (about 6% of all those described so far) have at least one molecular sequence archived in public databases. The evaluation they made involved the identification of clusters of putative protein homologs, construction of supertrees and supermatrices. A surprising finding is that combining many genes is a robust procedure, and that a relatively large amount of missing data is tolerable. They are optimistic that with these tools they may eventually identify the Tree of Life. Crandall and Buhay (2004), favorably discussing their results, expressed however some cautionary views. They also mentioned that we are loosing 27 thousand species each year, due to environmental degradation.

## Genome Variation in Prokaryotes

Prokaryotes are unicellular, and are the most numerous organisms on earth. Species delineation among them is complicated, due to the absence, in many, of sex; about four thousand species have been described, but some scholars suggest that there should be around four million of them.

Their genetic systems are not as simple as was imagined some time ago. Although most of them have circular genomes, others present linear structures. Also, while the majority has just one large chromosome, several have two or three large replicons; and in addition, many possess extrachromosomal DNA segments. Haploidy is not universal; during the exponential growth phase as much as 10 copies of the main chromosome can occur per cell.

Since the 1970's, molecular investigations indicated that the prokaryotes could be separated in two distinctive domains, named Archaea, and Bacteria. The main characteristics of the genomes of completely sequenced prokaryotes listed by Saccone and Pesole (2003) and described by others are given in Table I. The numbers of species considered are not high (18 Archaea, 76 Bacteria in the most general comparison), but some first generaliza-

**TABLE I**

**Main characteristics of completely sequenced prokaryotic genomes.**[1]

| Characteristics | Archaea | Bacteria |
|---|---|---|
| *Completely sequenced genomes* | | |
| No. of species considered | 18 | 77 |
| Size of the main chromosome | | |
| Range (Mb)[2] | 0.5-5.8 | 0.6-8.7 |
| Average (Mb)[2] | 2.2 | 3.0 |
| *Content of protein coding genes* | | |
| No. of species considered | 11 | 69 |
| No. of ORFs[2] (thousands) | | |
| Range | 0.5-3.0 | 0.5-6.7 |
| Average | 2.0 | 2.8 |
| Percent functionally identified (%) | | |
| Range | 24-67 | 40-88 |
| Average | 43 | 60 |
| Protein coding regions (%) | | |
| Range | 84-92 | 50-95 |
| Average | 89 | 87 |
| *Nucleotide composition (% GC)* | | |
| No. of species considered | 12 | 60 |
| Range | 31-68 | 22-67 |
| Average | 44 | 45 |

[1] Sources: Saccone and Pesole (2003) for a general review, plus Brazilian National Genome Project Consortium (2003); Brüggemann et al. (2003); Buell et al. (2003); Dufresne et al. (2003); Garnier et al. (2003); Gil et al. (2003); Ivanova et al. (2003); Kleerebezem et al. (2003); Read et al. (2003); Suerbaum et al. (2003); Waters et al. (2003); Xu et al. (2003); Bell et al. (2004); Holden et al. (2004); Moran et al. (2004); Nascimento et al. (2004a, b); Niermann et al. (2004); Seshadri et al. (2004). [2] Mega (one million) bases; ORFs: Open Reading Frames.

tions can be made. Similarities are more conspicuous than dissimilarities. The range of genome sizes is not much different, although the averages (Archaea: 2.2 Mb; Bacteria: 3.0 Mb) are somewhat distinct. The smallest fully sequenced archaeal genome (0.5 Mb; 491 kb) was the one from *Nanoarcheum equitans*, an obligate symbiont of *Ignicoccus*. It is also the most compact, with 95% of the DNA predicted to encode proteins or stable RNAs. On the other hand, *Methanosarcine acetivorans* presented the highest completely sequenced archeal genome (5.8 Mb). The corresponding Bacteria species were *Buchnera* sp., with only 0.6 Mb in its main chromosome, which however has two not very small extrachromosomal elements of respectively 7.3 and 7.8 kb; and *Streptomyces coelicolor* (8.7 Mb on the main chromosome; 356 kb and 31 kb, respectively, in two extra chromosomes). The number of ORFs follow the same trend (somewhat higher values in the Bacteria; averages of 2.0 and 2.8 thousands, respectively).

Functionally identified ORFs are higher in Bacteria (60%) than Archaea (43%); but the proportion of loci codifying for proteins are similar in the two groups (averages of 89% and 87%). The percentages of GC dinucleotides were also similar (respectively 44% and 45%).

We are still far away from functionally identifying a significant proportion of the microbial genes by functional categories, but data compiled by Saccone and Pesole (2003) indicated that some of the most frequent were related to energy metabolism, transport and binding proteins, protein synthesis, and cellular processes. Graham et al. (2000) identified 351 clusters of signature proteins, that have no recognizable bacterial or eukaryal homologs. They are involved in key energetic systems, cofactor biosynthesis and other functions. These unique genes, which are present in around 15% of the archaeal genomes, suggest that they really constitute an anciently diverged major lineage.

## VIRUSES

Viruses generally have much smaller genomes than bacteria. For instance, the virus responsible for the severe acute respiratory syndrome (SARS) recently identified in China and which spread to many countries has a 30 kb genome (Marra et al. 2003). The poxvirus genomes range from 145 kb to 290 kb in size, and each genome contains about 200 genes only. McLysaght et al. (2003) conducted a study in 20 of these viruses. They stressed four characteristics of their genome evolution. First, its structure is highly conserved among the chordopox (vertebrate-infecting) viruses; second, gene loss and gain varied markedly among the gene families considered; third, many of these acquisitions occur due to horizontal transfer events; and fourth, both conservative and positive selection could be identified, that undoubtedly are related to characteristic features of infection, replication, and virulence. Proteins that may be targeted for drug design, to be used for their control, were identified.

The genomic structure of a parasitoid wasp (*Cotesia congregate*) bracovirus (CcBV) was found to possess 568 kb in 30 DNA circles, which together contain 156 coding DNA sequences. In their organization they resemble more an eukaryote genomic region than a viral one. Many CcBV genes contain introns (69%) and 42% of their putative coding DNA sequences have no similarity to previously described genes (Espagne et al. 2004).

The largest known virus genome described so far is from a double-stranded DNA of a mimivirus which grows in amoeabae. It has 1.18 Mb, with 1,262 putative open reading frames, 10% of which exhibit a similarity to proteins of known function. Unexpectedly, genes for the codification of central protein-translation components, DNA repair pathways, and polysaccharide synthesis enzymes are present, as well as those coding for six tRNAs. By their structure mimivirus blurs the recognized frontier between viruses and the parasitic cellular organisms with small genomes. Raoult et al. (2004) suggest that these large DNA viruses could have emerged before the establishment of the three domains of life (Archaea, Bacteria, Eukarya).

## GENOME VARIATION IN EUKARYOTES

The term eukaryote refers to an enormous range of organisms whose common feature is that their cells have nuclei bounded by a membrane, separating them from the cytoplasm. There are many unicellular eukaryotes, and the question is whether their genomes are similar or dissimilar from those of the prokaryotes. Table II presents the characteristics of seven completely sequenced unicellular eukaryotes. Comparison with the numbers of Table I indicates that they possess much larger genomes (excluding *Encephalitozoon cuniculi*), their sizes ranging from 9 Mb to 34 Mb, against averages of 2.2 Mb and 3.0 Mb found in Archaea and Bacteria, respectively. *Encephalitozoon cuniculi* is a special case. This intracellular parasite microsporidium infects a wide range of hosts, from protozoans to humans (where it was identified in opportunistic infections after the emergence of AIDS and immunosuppressive therapies for organ transplantation). Its genome comprises 11 chromosomes, and it has about two thousand genes. The absence of genes for some biosyn-

**TABLE II**

**Main characteristics of completely sequenced unicellular eukaryotes.**[1]

| Organism | Size (Mb) | % (GC) | No. of genes | Mean gene length[2] (bp) | Gene density (bp per gene) | % coding |
|---|---|---|---|---|---|---|
| **Algae** | | | | | | |
| *Thalassiosira pseudonana* | 34.5 | 47 | 11,242 | 992 | 3,500 | NA[3] |
| **Protozoa** | | | | | | |
| *Cryptosporidium parvum* | 9.1 | 30 | 3,807 | 1,795 | 2,382 | 75 |
| *Cryptosporidium hominis* | 9.2 | 32 | 3,994 | 1,576 | 2,293 | 69 |
| *Encephalitozoon cuniculi* | 2.5 | 47 | 1,997 | NA[3] | 1,256 | 90 |
| *Plasmodium falciparum* | 22.9 | 19 | 5,268 | 2,283 | 4,338 | 53 |
| **Fungi** | | | | | | |
| *Saccharomyces cerevisae* | 12.5 | 38 | 5,770 | 1,424 | 2,088 | 70 |
| *Schizosaccharomyces pombe* | 12.5 | 36 | 4,929 | 1,426 | 2,528 | 57 |

[1] Sources: Abrahamsen et al. (2004); Armbrust et al. (2004); Xu et al. (2004). The optical map of *Leishmania major* was reported by Zhou et al. (2004), but besides the protozoa's total size (34.7 Mb) no other information about the variables listed were given. [2] Excluding introns. [3] NA: Not available in the sources consulted.

thetic pathways, like the tricarboxilic acid cycle, indicates that this organism has a strong host dependence.

As for the other information given in Table II, two of the organisms listed there, *Saccharomyces cerevisae* and *Plasmodium falciparum*, have been extensively studied, the first due to its economic importance, and the second because it is the agent of one of the most severe forms of malaria. *P. falciparum*'s genome is almost 2× higher than that of *S. cerevisae*, but it carries about the same number (five to six thousand) of genes; this, naturally, conditions diversity in relation to gene density and length. In addition, GC content is quite diverse in the two species (38% in *S. cerevisae*; 19% in *P. falciparum*).

Let us now consider in more detail a group of eukaryotic organisms, the plants. Some say that a plant is something green that doesn't move around very much. Using a phylogenetic approach, we can separate the green plants from other photosynthetic organisms and focus in the land plants. Land conquest was achieved by them some 430 million years ago, with subsequent wide diversification. The bryophytes and the ancestor of vascular plants (tracheophytes) diverged early in land plant evolution. Mosses are bryophytes and their morphologies and life cycles differ significantly from those of flowering plants. For instance, the gametophyte (haploid) generation is dominant in mosses, while the sporophyte (diploid) generation is the one dominant in flowering plants.

Taking into consideration these differences, Nishiyama et al. (2003) investigated in what way the transcriptome (the total product of DNA transcription) of a bryophite, the moss *Physcomitrella patens* differed from that of *Arabidopsis thaliana*, the first vascular plant which had its genome fully sequenced. A total of 15,883 transcripts were assembled, and at least 66% of the genes in *P. patens* had homologues in *A. thaliana*.

There is wide variation in the genome sizes of the vascular plants; for instance, these sizes vary in the major crop plants from 0.4 giga (one billion) bases (Gb) in rice to 16 Gb in wheat (Messing et al. 2004). Within the vascular plants, the angiosperms are by far the most numerous group; they have from 250 thousand to 300 thousand species; while the other plants number just around 53 thousand (Eguiarte et al. 2003). Traditionally the angiosperms have been classified in monocots (plants with a single cotyledon and other characteristic features) and dicots (with two cotyledons). In phylogenetic analyses only the monophyly of the monocots is supported; but although dicots are apparently nonmonophyletic, a large number of species traditionally considered within this group form well-supported clades (Judd et al. 1999). Be as it may, only a single species from each of these large subdivisions had their genome completely sequenced: *Arabidopsis thaliana* (a weed; dicot) and *Oryza sativa* (rice; monocot).

*Arabidopsis thaliana*'s genome is much smaller (125 Mb) than the *Oryza sativa* genome (420 Mb in the subspecies *indica*; 466 in the subspecies *japonica*). They also differ in GC content (36% in *Arabidopsis*; 43% to 44% in *Oryza*) and in number of genes (*Arabidopsis*: 25 thousand; *Oryza*: 32-55 thousand). As for homology, 80-85% of the *Arabidopsis* genes have a rice homolog, but only 49% of rice genes are represented in *A. thaliana* (Eguiarte et al. 2003, Saccone and Pesole 2003).

*Zea mays*, or corn, is one of the most important crops. Its genome has 2.3 Gb, and therefore it would be difficult to sequence it entirely. However, Messing et al. (2004) have generated 307 Mb of its sequence, estimating that repeat sequences occur in 58%, and genic regions in 7.5% of the genome. Two striking aspects of it were emphasized by them. First, although the ancestor of maize arose by tetraploidization, fewer than half of the genes appear to be present in two orthologous copies, suggesting significant gene loss in the diploidization process. Second, the remaining gene number has increased dramatically due to tandemly amplified gene families.

Information about the genomes of 11 completely sequenced representatives of two other kingdoms of life (Fungi and Animalia) are presented in Table III. The method of study varied in the different investigations, but the tendency seems to be to combine the speed of the Whole Genome Shotgun with the precision of additional Clone-by-Clone investigations. There is only one fungus species (*Neurospora crassa*) that presents a much smaller (38.6 Mb) genome size and estimated number of genes (10.1 thousand), as well as a more compact (average of two introns per gene) genetic system, as compared to the other organisms. But even within the Animalia, the range of values, especially in relation to genome sizes, is wide (97.0 Mb in *Caenorhabiditis elegans*; 2.9 Gb in *Homo sapiens*; the latter value is about 30× higher than the first). The interval in relation to number of genes is much smaller (13.6 thousand in *Drosophila melanogaster*; 31.1 thousand in *Fugu rubripes*, a 2× difference), while the number of introns per gene varies from three (*D. melanogaster*) to nine (*Rattus norvegicus*). Notice the much larger genome size of *Bombix mori* (428.7 Mb), as compared to that of *Drosophila melanogaster* (180.0 Mb), a 2.4× difference. Larger genes were also found in *B. mori*, due to the insertion of transposable elements, an event that occurred in relatively recent times. The number of estimated genes is also 1.4× higher in *B. mori* (18.5 thousand) as compared to *D. melanogaster* (13.6 thousand).

Other characteristics of these and related organisms deserve mention. Thus, *Neurospora crassa* presents the widest array of genome defense mechanisms observed in any eukaryotic organism, including a unique mechanism, named repeat-induced point mutation. This is a process that efficiently detects and mutates both copies of a sequence duplication, determining the methylation and silencing of repetitive DNA. Thus the organism establishes a protection against selfish or mobile DNA, but with a price. It prevents gene innovation through gene duplication, and the result is a genome with an unusually low proportion of closely related genes.

This mechanism does not exist in *Ciona intestinalis*, and therefore it did not prevent the devel-

TABLE III

**Selected information about the genomes of completely sequenced multicellular eukaryotes.**[1]

| Organisms and their taxonomic classification | Genome size[2] | Protein-coding genes (thousands) | Introns per gene[2] | Method of study[2] |
|---|---|---|---|---|
| Kingdom Fungi | | | | |
| Phylum Ascomycota | | | | |
| Species *Neurospora crassa* | 38.6 Mb | 10.1 | 2 | WGS |
| Kingdom Animalia | | | | |
| Phylum Loricifera | | | | |
| Subphylum Nematoda | | | | |
| Species *Caenorhabditis elegans* | 97.0 Mb | 19.1 | 5 | CCS |
| Phylum Anthropoda | | | | |
| Subclass Insecta | | | | |
| Species *Anopheles gambiae* | 260.0 Mb | 15.2 | NA | CCS |
| *Bombyx mori* | 428.7 Mb | 18.5 | NA | WGS |
| *Drosophila melanogaster* | 180.0 Mb | 13.6 | 3 | WGS/CCS |
| Phylum Chordata | | | | |
| Branch Urochordata | | | | |
| Species *Ciona intestinalis* | 160.0 Mb | 15.8 | 6 | WGS |
| Class Osteichthyes | | | | |
| Species *Fugu rubripes* | 365.0 Mb | 31.1 | NA | WGS |
| Class Aves | | | | |
| Species *Gallus gallus* | 1.6 Gb | 17.7 | NA | WGS |
| Class Mammalia | | | | |
| Order Rodentia | | | | |
| Species *Mus musculus* | 2.6 Gb | 30.0 | 8 | WGS/CCS |
| *Rattus norvegicus* | 2.7 Gb | 21.0 | 9 | WGS/CCS |
| Order Primates | | | | |
| Species *Homo sapiens* | 2.9 Gb | 26.4 | 8 | WGS/CCS |

[1] Sources: The *C. elegans* Sequencing Consortium (1998); Adams et al. (2000); International Human Genome Sequencing Consortium (2001); Venter et al. (2001); Aparicio et al. (2002); Dehal et al. (2002); Holt et al. (2002); Mouse Genome Sequencing Consortium (2002); Galagan et al. (2003); Rat Genome Sequencing Project Consortium (2004); Biology Analysis Group (2004); International Chicken Genome Sequencing Consortium (2004). [2] Abbreviations: Mb: Megabases; NA: Not available in the sources consulted; Gb: Gigabases; WGS: Whole Genome Shotgun; CCS: Clone-by-Clone Shotgun.

opment of *Ciona*'s unique genes for making cellulose. These genes, that since they were never observed in other animals may appear to have come out of nowhere, actually should have been received through horizontal gene transfer from a bacteria (Pennisi 2002).

What is the origin of the central nervous system? To investigate this question Mineta et al. (2003) determined the nucleotide sequence of expressed sequence tags (ESTs) from the head portion of the planaria *Dugesia japonica*, a basal bilateran animal. Out of 3,101 nonredundant EST clones they found 116 clones that had significant similarity to known genes related to the nervous system, and 110 of them (95%) were shared with *H. sapiens, D. melanogaster*, and *C. elegans*, indicating considerable conservation. Even more interesting, 35 of them (30%) had homologous sequences with those of *A. thaliana* and *S. cerevisae*, which do not have a nervous system! Therefore, nervous system-related genes greatly predated the origin of the nervous system.

Other recent genome studies involving domestic animals include a 1-Mb resolution radiation hybrid map of the canine genome (Guyon et al. 2003); and a detailed physical map of the horse Y chromosome (Raudsepp et al. 2004).

Are there marked differences in the rates of change that occur along different eukaryotic lineages? Table IV presents some information concerning this point. As is shown there, the rate of substitutions, and the number of bases involved in the events, is two-times higher in the rodent lineage that originated *M. musculus* and *R. norvegicus* than in the human lineage. Independently of the lineages, deletion events are two to three times as common as insertion events, and the number of bases involved are three to four times higher in deletions as compared to insertions.

### Genome Variation in Organelles

Mitochondria are cytoplasmic organelles that originated from a symbiosis between bacteria that led to the formation of the primitive eukaryotic cell. They are mainly related to processes of oxidative phosphorylation; and as a matter of fact the number of mitochondria per cell is strictly correlated with energy requirements. Their structure also varies widely during the cell cycle. Not less than 292 mitochondrial genomes (mtDNAs) had been completely sequenced around 2003, and their sizes (ranges and averages) are listed in Table V according to different taxonomic categories. The largest genomes are found in plants (average of 207 kb), and the smallest among Protists (38 kb). In Animalia the variability is not extensive. Despite the extreme values of 13.5 kb found in *Taenia crassipes*, a platyhelminth, and 22.7 kb found in *Venerupis philippinarum*, a mollusk, the averages in the several phyla generally occur around 16 kb, with no clear connection with phylogeny. The value for *Homo sapiens* is 16.6 kb.

The mitochondrial genetic code shows several differences from the universal genetic code, and these changes are diverse from phylum to phylum. Protist mtDNAs generally resemble plant rather than animal or fungal mtDNAs. Despite the wide variation in size (19.4-100.3 kb) the informational content of the mitochondrial genome in fungi is quite constant. Impressive diversity occurs in plants; their genome may be formed by a single large circular molecule (the master chromosome) as well as by a heterogeneous population of linear, circular, and more complex structures. Animal mtDNAs, on the other hand, are characterized by a compact arrangement, constancy of gene content, and the presence of a single noncoding region about 1 kb long (Saccone and Pesole 2003).

Plastids are cellular organelles that have also arisen through symbiotic processes, involving specifically a primitive eukaryote and a cyanobacterium-like ancestor. Among plastids the chloroplast is undoubtedly the best studied, since the organelle is the central site of the photosynthetic process. It has a generally flat and lens-shaped structure, and is limited by a double membrane. Much less information is available, as compared to mitochondria, for the plastid genome, but the sizes of the complete genome of 21 species are given in Table VI. The smallest (35 kb) is *Toxoplasma gondii*'s plastid. It is similar to chloroplast genomes and its small size is probably related to the fact that this microorganism is a parasite. *Euglena gracilis* has a much (4.1×) larger cpDNA (143.2 kb). Algae generally have large chloroplast genomes (average: 150.3 kb), and the three Poales species studied (all

**TABLE IV**

**Midpoint values and their variation in the branches of a phylogenetic tree connecting humans, the common rodent line, mice, and rats.**[1]

| Characteristic | Rate (values and standard deviations) | | | |
|---|---|---|---|---|
| | Human | Rodent lineage | Mouse | Rat |
| Substitutions per site | 0.11 ± 0.00 | 0.24 ± 0.00 | 0.07 ± 0.00 | 0.08 ± 0.01 |
| Substitutions in neutral sites only | 0.13 ± 0.01 | 0.28 ± 0.03 | 0.08 ± 0.01 | 0.09 ± 0.01 |
| Insertion events per kb | 2.7 ± 0.9 | 4.7 ± 1.0 | 1.5 ± 0.8 | 1.4 ± 0.7 |
| Deletion events per kb | 5.3 ± 0.5 | 12.0 ± 1.2 | 3.8 ± 0.2 | 4.5 ± 0.1 |
| Inserted bases per kb | 6.4 ± 2.9 | 9.4 ± 1.6 | 3.6 ± 1.5 | 3.2 ± 1.3 |
| Deleted bases per kb | 18.0 ± 2.0 | 40.0 ± 4.9 | 11.0 ± 0.5 | 13.0 ± 0.1 |

[1] Source: Rat Genome Sequencing Project Consortium (2004).

economically important) show little variation (average: 136.5 kb). Notable is the low size (70.0 kb) of *Epifagus virginiana*, an Euasterid.

## Unicellularity, Multicellularity, and Genome Complexity

Everything in life has its price. In organic evolution the price payed to allow more diversification was death. Raff and Kaufman (1983) asserted "By the account of the Old Testament book of Genesis, death was the price of knowledge; more prosaically, it was the price of multicellularity". The most primitive metazoans should have had just two cell types, *somatic* and *germinative*. The latter retained most of the properties of their basically immortal unicellular ancestors. The somatic ones, however, showed adhesiveness and could only perform a finite number of cell divisions. The interaction between these two types, determined by a balance between growth factors and cell death (apoptosis), led to organisms of different sizes and shapes.

Multicellularity has arisen many times in eukaryotic evolution, with independent origins in the lineages leading to animals, green plants, fungi, cellular slime moulds, and several other taxa. Three factors are shared by them: division of labor, differentiation, and cell communication (Brooke and Holland 2003).

A curious organism, that presents intermediate properties between the unicellular and multicellular conditions, was described by Keim et al. (2004a, b). It is a spherical prokaryotic magnetotactic multicellular living being whose components align themselves along magnetic fields, swim, and divide by binary fission as a unit. Its life cycle is therefore completely multicellular, differing therefore from the latter because it does not present, as part of the life cycle, a unicellular stage.

How is this reflected in genome complexity? Lynch and Conery (2003) argued that there is an inverse relationship between population density per unit of area and average individual body mass within a species, and also an inverse relationship between organism size and the product of its effective population (the number that effectively matters in evolution) and the mutation rate ($N_e u$). Therefore, the enormous long-term effective population sizes of prokaryotes would impose a strong barrier to the evolution of complex genomes and morphologies.

A closer examination of this matter involves first the question of *quantity*; what determines genome expansion? The following agents can be listed: (a) gene duplication; (b) segmental genomic duplication (mainly by the mechanism of unequal crossing-over in gene families); (c) polyploidy; and (d) lateral gene transfer. All of them create redun-

**TABLE V**

**Size of completely sequenced mitochondrial genomes present in different organisms.**[1]

| Taxonomic characterization | No. of species tested | Range (kb) | Average (kb) |
|---|---|---|---|
| Kingdom Protoctista | 28 | 5.9 - 69.0 | 38.0 |
| Kingdom Fungi | 14 | 19.4 - 100.3 | 49.4 |
| Kingdom Plantae | 3 | 186.6 - 368.8 | 307.4 |
| Kingdom Animalia | | | |
| Phylum Cnidaria | 2 | 17.4 - 18.3 | 17.8 |
| Phylum Platyhelminthes | 6 | 13.5 - 16.7 | 14.4 |
| Phylum Loricifera | 10 | 13.6 - 15.0 | 14.0 |
| Phylum Brachiopoda | 2 | 14.0 - 15.4 | 14.7 |
| Phylum Mollusca | 7 | 14.1 - 22.7 | 16.6 |
| Phylum Annelida | 2 | 15.0 - 15.6 | 15.3 |
| Phylum Arthropoda | 30 | 14.5 - 19.5 | 15.6 |
| Phylum Echinodermata | 5 | 15.6 - 16.3 | 15.9 |
| Phylum Chordata | | | |
| Subphylum Acraniata | 4 | 14.8 - 15.7 | 15.2 |
| Phylum Craniata | | | |
| Class Cyclostomata | 4 | 16.2 - 18.9 | 17.1 |
| Class Condrichthyes | 6 | 16.7 - 18.6 | 17.0 |
| Class Osteichthyes | 67 | 15.6 - 17.3 | 16.5 |
| Class Amphibia | 4 | 16.6 - 17.8 | 17.2 |
| Class Reptilia | 9 | 16.5 - 17.9 | 17.0 |
| Class Aves | 23 | 15.3 - 18.7 | 16.9 |
| Class Mammalia | 66 | 15.4 - 17.7 | 16.7 |

[1] Source: Saccone and Pesole (2003).

dancy, and the possibility of evolutionary plasticity. On the other hand, factors exist that can reduce the genome size, especially deletions of different sizes. But quantity is not enough. The way this new material was introduced and its content certainly matters (*quality*).

It is clear, for instance, that multicellular organisms would need an expansion of regulatory domain families involved in extracellular adhesion or signal transduction, as well as DNA-binding transcription factors. In addition to the invention and proliferation of specific protein domains, new architectures created by the juxtapositions of previously existent domains would also be needed (Saccone and Pesole 2003). Other factors are intron number and size. The average number of introns per gene in most multicellular species is between four and seven, while for unicellular eukaryotes is less than two. As a matter of fact, there seems to be a threshold genome size of about 10 Mb below which introns are very rare. Intron length is also important; there is a relationship between it and the amount of recombination,

**TABLE VI**

**Sizes of completely sequenced chloroplast and related plastid genomes.**[1]

| Taxonomic characterization and species | Size (kb) |
|---|---|
| Kingdom Protoctista | |
| *Astasia longa* | 73.3 |
| *Euglena gracilis* | 143.2 |
| *Toxoplasma gondii* | 35.0 |
| Algae | |
| *Chlorella vulgaris* | 150.6 |
| *Cyanidium caldarium* | 164.9 |
| *Cyanophora paradoxa* | 135.6 |
| *Guillardia theta* | 121.5 |
| *Mesostigma viride* | 118.4 |
| *Nephroselmis olivacea* | 200.8 |
| *Porphyra purpurea* | 191.0 |
| *Odontella sinensis* | 119.7 |
| Bryophyta | |
| *Marchantia polymorpha* | 121.0 |
| Conifer | |
| *Pinus thumbergii* | 119.7 |
| Monocots, Poales | |
| *Oryza sativa* | 134.5 |
| *Triticum aestivum* | 134.5 |
| *Zea mays* | 140.4 |
| Eurosids | |
| Fabales | |
| *Lotus japonicus* | 150.5 |
| Myrtales | |
| *Oenothera elata* | 163.9 |
| Brassicales | |
| *Arabidopsis thaliana* | 154.5 |
| Euasterids | |
| Solanales | |
| *Nicotiana tabacum* | 155.9 |
| Lamiales | |
| *Epifagus virginiana* | 70.0 |

[1] Source: Saccone and Pesole (2003).

probably modulated by selection (Lynch and Conery 2003).

Other aspects have been considered by Wolfe and Li (2003), such as: (a) the origins of new genes (*de novo* formation, mosaic); (b) asymmetric directional mutation pressure; and (c) effects of genome location on mutation rates. Certainly *structure matters*, as is indicated by the fact that adjacent genes are co-regulated more often than is expected by chance, and in comparative genomics extensive conservation of gene order occurs. Vinogradov (2004), on the other hand, discussed especially the problem of variation in the amount of noncoding DNA in terms of multilevel selection, mutation bias, and negative and/or positive feedbacks to genome size changes.

## Comparative Human Chromosome Genomics

The draft sequence of the human genome was presented with much fanfare in February, 2001 (International Human Genome Sequence Consortium 2001, Venter et al. 2001), but detailed analysis of its contents is continuing quietly in the laboratories. A series of papers are appearing giving information for each chromosome of our genome, and four of the main characteristics of those most completely studied are given in Table VII. Chromosome 19 is notable, since it presents the highest numbers in three of the characteristics compared, namely the highest GC (48%) and the highest repeat contents (55.7%), as well as the highest gene density (26.0 genes per Mb). It is also medically important due to the presence there of genes involved in familial hypercholesterolemia and insulin-resistant diabetes (Grimwood et al. 2004). Chromosome 13, the largest acrocentric of the set, is notable conversely for the low percentage of GC (38.5%) and low density (6.5 genes per Mb). As a matter of fact, it contains a central region of 38 Mb where the gene density drops to only 3.1 genes per Mb. This chromosome is also important because it carries several genes involved with cancer (breast cancer, retinoblastoma, B-cell chronic lymphocytic leukemia) (Dunham et al. 2004).

**TABLE VII**

**Characteristics of ten of the most completely studied chromosomes of the human genome.**[1]

| Chromosome | Sequence length (Mb) | % GC | % repeat | Gene density ($Mb^{-1}$) |
|---|---|---|---|---|
| 5 | 177.7 | 39.5 | 46.3 | NA |
| 6 | 166.8 | 40.0 | 43.9 | 9.2 |
| 7 | 153.8 | 41.0 | 45.0 | 7.5 |
| 13 | 95.5 | 38.5 | 42.3 | 6.5 |
| 14 | 87.4 | 40.9 | 46.2 | 10.0 |
| 16 | 78.9 | 44.7 | 47.8 | NA |
| 19 | 55.8 | 48.0 | 55.7 | 26.0 |
| 20 | 59.2 | 44.1 | 42.0 | 12.2 |
| 21 | 33.5 | 40.9 | 40.1 | 6.7 |
| 22 | 33.5 | 47.8 | 41.9 | 16.3 |
| Genome | 2,862.7 | 41.0 | 44.8 | 10.0 |

[1] Source: Dunham et al. (2004); Martin et al. (2004); Schmutz et al. (2004b). NA: Not available in the sources consulted.

Chromosome 22, as chromosome 19, also shows a high percentage of GC (47.8%), and high gene density (16.3 genes per Mb). This chromosome and chromosome 21 were subjected to a close analysis by Chen et al. (2002), who found that: (a) the number of gene structures containing untranslated exons exceeds 25%; and (b) the terminal exon and initial intron tend to be the largest in their categories.

Respectively low and high percentages of repeats have been found in chromosome 21 (40.1%) and 16 (47.8%). The former is of course involved in the etiology of Down syndrome and one form of Alzheimer's disease, and has been extensively studied by Stylianos E. Antonarakis and his group in Geneva. For instance, Dermitzakis et al. (2002) found 2,262 non-genic conserved blocks in this chromosome that are potentially functional. As for chromosome 16, it has many segmental duplications, that are particularly clustered along its short arm (Martin et al. 2004).

Chromosome 5 harbor the important protocadherin and interleukin gene families, as well as the duplicated region associated with various forms of spinal muscular atrophy (Schmutz et al. 2004b). Chromosome 6 has the largest transfer RNA gene cluster in all the genome, and within its major histocompatibility complex presents HLA-B, the most polymorphic of our loci (Mungall et al. 2003). Chromosome 7, the seat of the cystic fibrosis gene, exhibits an unusual amount of segmentally duplicated sequences (8.2%), that might have led to a series of chromosome rearrangements that were evolutionary important, and included 440 breakpoints associated with disease (Hillier et al. 2003, Scherer et al. 2003, Müller et al. 2004).

As for the remaining chromosomes listed in Table VII, chromosome 14 has a heterochromatic short arm that contains essentially ribosomal RNA genes, while its euchromatic long arm presents most, if not all, of the protein-coding genes. Two loci of special importance for the immune system (the alpha/delta T-cell receptor and the immunoglobulin heavy chain), as well as more than 60 disease genes, have been localized in it (Heillig et al. 2003). Chromosome 20 generally has intermediate figures for the characteristics listed in Table VII. It is best known for harboring the genes that cause Creutzfeldt-Jakob disease and severe combined immunodeficiency (Deloukas et al. 2001).

Not listed in Table VII but also extensively characterized is the Y chromosome. Its male-specific region comprises 95% of its length, and is a mosaic of heterochromatic and euchromatic sequences. The latter have been classified in three classes: X-transposed, X-degenerate and ampliconic. These classes present all 156 known transcription units, which include 78 protein-coding genes that collectively encode 27 distinct proteins. The most prominent features of the ampliconic region are eight massive palindromes, at least six of which contain testis genes (Skaletsky et al. 2003).

## BRAZILIAN CONTRIBUTION TO GENOMICS

Genome research in Brazil initiated nine years before the release of the draft sequence of the human genome. All started in Belo Horizonte in 1992, as a collaborative effort between Sergio D.J. Pena, from the Federal University of Minas Gerais, and Andrew J. Simpson, at the time in the René Rashou Research Center of the Oswaldo Cruz Foundation. They decided to obtain expressed-sequence tags (ESTs) (namely those that uniquely identify genes) of *Schistosoma mansoni*, a trematode worm responsible for schistosomiasis, an endemic parasitic disease that affects about 12 million Brazilians, especially in the rural area. In this endeavor they received initial logistic and financial support from J. Craig Venter and its Institute for Genomic Research in the USA (Pena 1996).

This project continues to be developed through consortia coordinated both in Belo Horizonte and São Paulo. But the real turning point in relation to Brazilian genetics science in general occurred with the complete sequencing (using a combination of clone-by-clone shotgun and whole genome shotgun) of *Xylella fastidiosa*. This microorganism is responsible for a serious disease of orange trees, citrus variegated chlorosis. When Andrew J. Simpson moved from Belo Horizonte to the Ludwig Institute for Cancer Research in São Paulo, an opportunity appeared for the development of a project, enthusiastically endorsed by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for the genome sequencing of this bacterium. The results were published in the 13 July, 2000 issue of Nature as a cover article, and with very favorable comments.

The success of the *Xylella* project (which led to the sequencing, afterwards, of an USA strain – Van Sluys et al. 2003) stimulated the development of several other genome investigations. A list of those for which information is available is given in Table VIII. All of them concentrated: (a) in organisms or subjects of major economical or medical interest (agricultural pests, organisms that could produce substances for industrial use, important crops, agents or vectors of diseases, the cancer problem in humans); and (b) in the functional fraction of the genomes, through the EST approach.

As can be seen in Table VIII, all kingdoms of life are being considered, namely 11 species of Bacteria (*C. violaceum, G. diazotrophicus, H. seropedicae, L. xyli, L. interrogans, M. hyopneumoniae, M. synoviae, R. tropici, X. axonopodis, X. campestris, X. fastidiosa*), two of Protozoa (*L. chagasi, T. cruzi*), two of Fungi (*C. perniciosa, P. brasiliensis*), four of Plants (*C. arabica, E. grandis, P. cupana, S. officinalis*) and five of Animalia (*S. mansoni, L. vannamei, A. aegypti, B. taurus, H. sapiens*). Also, all of them are being developed through consortia of laboratories from different institutions, an intelligent way to optimize personnel and equipment resources, providing also opportunity for less developed centers to integrate with more sophisticated units in important investigations.

Other results involved the identification of short interrupted palindromes (a sequence that reads the same, 5' to 3', on complementary strands) on the extragenic DNA of three bacterial genomes (Vasconcelos et al. 2000); and a System for Automated

TABLE VIII

**Selected information about the Brazilian contribution to genome projects.**[1]

| Organism | Consortium[1] | Stage of development and/or reference[2] |
|---|---|---|
| *Aedes aegypti* | Institut Pasteur/FAPESP | In course |
| *Bos taurus* | AEG-FAPESP<br>ESALQ | Sequencing finished<br>Annotation in progress (1) |
| *Chromobacterium violaceum* | BRGENE | (2) |
| *Coffea arabica* | AEG-FAPESP/EMBRAPA<br>IAC/EMBRAPA/CENARGEN | Sequencing finished<br>Annotation in progress |
| *Crinipellis perniciosa* | FAPESP/MCT | In course |
| *Eucalyptus grandis* | GENOLYPTUS<br>FORESTS-AEG-FAPESP | Sequencing and annotation finished (3) |
| *Gluconacetobacter diazotrophicus* | RIOGENE/FAPESP/MCT | In course (4) |
| *Herbaspirilum seropedicae* | UFPR/MCT | In course |
| *Homo sapiens* | Brazilian ONSA/Head and Neck Annotation | (5-9) |
| *Leifsonia xyli* | FAPESP | (10) |
| *Leishmania chagasi* | PROGENE | In course |
| *Leptospira interrogans* | AEG-FAPESP/ONSA | (11, 12) |
| *Litopenaeus vannamei* | CNPq/ABCC | In course |
| MMO (Magnetotactic multicellular organism) | CNPq/FAPESP | Starting |
| *Mycoplasma hyopneumoniae* | PIGS/BRGENE | (13) |
| *Mycoplasma synoviae* | | |
| *Paracoccidioides brasiliensis* | Rede do Centro-Oeste | (14, 15) |
| *Paullinia cupana* | REALGENE/CNPq/MCT | In course |
| *Rhizobium tropici* | EMBRAPA/UFPR/UEL/LNCC | In course |
| *Saccharum officinalis* | Brazilian ONSA/FAPESP/ COPERSUCAR | (16-19) |
| *Schistosoma mansoni* | CNPq/FAPEMIG, IQ-USP/ FAPESP | (20-25) |
| *Trypanosoma cruzi* | IBMP/FIOCRUZ | In course |
| *Xanthomonas axonopodis* | Brazilian ONSA/FAPESP | (26-28) |
| *Xanthomonas campestris* | | |
| *Xylella fastidiosa* | Brazilian ONSA/FAPESP | (29-31) |

[1] Abbreviations: ABCC: Associação Brasileira de Criadores de Camarão; AEG: Agronomical and Environmental Genomes; BRGENE: Brazilian National Genome Project Consortium; CENARGEN: EMBRAPA Recursos Genéticos e Biotecnologia; COPERSUCAR: Cooperativa dos Produtores de Açúcar e Álcool do Estado de São Paulo; CNPq: Conselho Nacional de Desenvolvimento Científico e Tecnológico; EMBRAPA: Empresa Brasileira de Pesquisa Agropecuária; ESALQ: Escola Superior de Agricultura Luiz de Queiroz; FAPEMIG: Fundação de Amparo à Pesquisa do Estado de Minas Gerais; FAPESP: Fundação de Amparo à Pesquisa do Estado de São Paulo; FIOCRUZ: Fundação Oswaldo Cruz; FORESTS: FAPESP's *Eucalyptus* Genome Project; GENOLYPTUS: Rede Brasileira de Pesquisa do Genoma de *Eucalyptus*; IAC: Instituto Agronômico de Campinas; IBMP: Instituto de Biologia Molecular do Paraná; IQ-USP: Instituto de Química da Universidade de São Paulo; LNCC: Laboratório Nacional de Computação Científica; MCT: Ministério de Ciência e Tecnologia; ONSA: Organization for Nucleotide Sequencing and Analysis; PIGS: Southern Network for Genome Analysis; PROGENE: Projeto Genoma do Nordeste; REALGENE: Rede da Amazônia Legal de Pesquisas Genômicas; RIOGENE: Rede Genoma do Estado do Rio de Janeiro; UEL: Universidade Estadual de Londrina; UFPR: Universidade Federal do Paraná. [2] References: 1. Anonymous (2003); 2. Brazilian National Genome Project Consortium (2003); 3. Anonymous (2004); 4. Fernandes (2003); 5. Dias Neto et al. (2000); 6. de Souza et al. (2000); 7. Camargo et al. (2001); 8. Brentani et al. (2003); 9. Reis et al. (2005); 10. Monteiro-Vitorello et al. (2004); 11. Nascimento et al. (2004a); 12. Nascimento et al. (2004b); 13. Vasconcelos et al. (unpublished data); 14. Felipe et al. (2003); 15. Goldman et al. (2003); 16. Arruda et al. (2001); 17. Vettore et al. (2003); 18. Fioravanti (2003); 19. Junior et al. (2004); 20. Pena et al. (1995); 21. Franco et al. (1997); 22. Santos et al. (1999); 23. Prosdocimi et al. (2002); 24. Verjovski-Almeida et al. (2003); 25. De Marco et al. (2004); 26. Fioravanti (2000); 27. Mateos (2002); 28. Silva et al. (2002); 29. *Xylella fastidiosa* Consortium (2000); 30. Moura (2000); 31. Van Sluys et al. (2003).

Bacterial (genome) Integrated Annotation, SABIA (Almeida et al. 2004).

## STUDIES OF THE PORTO ALEGRE GROUP

Our research group is working in different aspects of molecular evolutionary change, and information about the most recent results is provided in Tables IX and X. The plant studies have two foci of inquire: (a) a particularly interesting group of substances, the pathogenesis-related proteins (PRs); and (b) relationships among taxa in two contrasting genera, one (*Passiflora*) much variable, while the other (*Petunia*) shows a restricted degree of speciation.

PRs are coded by plants as a response to pathological or related situations and provide a good model for the testing of mechanisms of positive selection. We (Scherer et al. 2005) encountered evidence for the action of this type of selection at specific sites of five of the 13 PR families investigated. The challenge, now, is to establish relationships between structure and function, and a beginning was made by Thompson et al. (2005) by modeling four PR-5 proteins.

After a first overall phylogenetic analysis of 61 species of *Passiflora* (Muschner et al. 2003) our group is working with special situations. One of them is the possible differentiation of *P. elegans* from *P. actinia* under the influence of changes that occurred in the Atlantic forest in southern Brazil (Lorenz-Lemke et al. 2005); the other is the documentation and details as how *P. alata* is invading unoccupied areas of this region (Koehler-Santos et al. 2005a, b). As for *Petunia*, despite the fact that the 11 species studied do not show marked molecular differences, they can be separated in two complexes, their representatives living respectively in high and low altitude levels. *Petunia* plus its closely allied genus *Calibrachoa* should have diverged from other clades at about 25 million years before present (Kulcheski et al. 2005).

The studies on human populations are listed in Table X. A particularly interesting genomic region occurs in chromosome 19, that codes for the low-density lipoprotein receptor (*LDLR*) and contains a 3' untranslated region (3' UTR) that has multiple copies of the *Alu* family of repetitive DNA. Its upstream portion presented the highest mutation rate estimated thus far for an autosomal locus in humans (0.632% per million years), possibly due to the action of selection (Fagundes et al. 2005). The results suggest a single origin for the first colonizers of the American continent (Heller et al. 2004), as was observed in other data set, involving L1 and *Alu* insertions (Mateus Pereira et al. 2005).

Other findings: (a) Native North American Indian mtDNA haplogroup X was not present in 991 individuals from 25 South American Indian groups, indicating that while present in the north, it should be rare or absent in the south (Dornelles et al. 2005); (b) spatial gradients of the APO E alleles are compatible with a directional demographic expansion which occurred in northeastern Asia and much of the New World (Demarchi et al. 2005); and (c) Data from 404 STRs and 17 2-9-site haplotypes indicate that although the early colonization of the Americas may have conditioned some loss of genetic variability, the range of differences found among five Native American populations was two times higher than those found between the most variable Amerindian (Maya) and a control African Yoruba sample (Salzano and Callegari-Jacques 2005a).

## UNDERLYING PRINCIPLES

Genomic evolution can be viewed as just a part of the universe's evolution, made possible by the origin of life. The whole process can be envisaged as a permanent struggle between the dialectical agents of change and order. Disordered chaos prevailed in the beginning of life, but the building of structures and channeling of processes soon established limits to variation. Natural selection was the primary agent responsible for the evolution of a genetic region or organism in one or the other direction, through mechanisms of positive (emphasizing novelty) or negative (protecting the *status quo*) se-

**TABLE IX**
**Recent investigations (papers published or in press) on plant molecular evolution by the Porto Alegre group.**

| Organisms | Subject of the investigation | No. of indiv./ sequences | Systems or DNA regions considered[1] | References[2] |
|---|---|---|---|---|
| Five species of *Passiflora* | Comparison between pathogenesis-related proteins (PRs) | 35 | Bet v 1 homologues | 1 |
| Fifty-four species | Evidence for positive selection | 194 | 13 classes of PRs | 2 |
| Five species of Rosaceae and Fagaceae | PR 5 modelling | 26 | PR 5 structure | 3 |
| *Passiflora actinia* and *P. elegans* | Intra and interspecific variability, relationship between the two | 53 | ITS, *trnL-trnF, psbA-trnH* | 4 |
| *Passiflora alata* | Characteristics of an invasive process | 85 | ITS, *G3pdh, LEAFY* | 5 |
| Eleven species of *Petunia*, five of three other species | Interspecies comparison, divergence time of *Petunia* + *Calibrachoa* from other genera | 127 | ITS, *trnL* intron, *TrnL-trnF, psbA-TrnH, psbB-psbF, TrnG-trnS, nad1* intron | 6 |

[1] Bet v 1: major allergenic compound of the pollen of Betulaceae; ITS: nuclear ribosomal internal transcribed spacer region; *trnL-trnF, psbA-trnH, psbB-psbF, trnG-trnS*: plastid intergenic spacers; nad1 intron: mitochondrial DNA nicotinamide-adenine dinucleotide intron. [2] 1. Finkler et al. (2005); 2. Scherer et al. (2005); 3. Thompson et al. (2005); 4. Lorenz-Lemke et al. (2005); 5. Koehler-Santos et al. (2005a, b); 6. Kulcheski et al. (2005).

lection. Mutation itself was influenced by this factor, which determined rates of change and forbidden options.

The structure of life involved the assemblage of genes in chromosomes, the establishment of longitudinal differentiation along them of coding and non-coding (regulatory) regions, and the determination of the gene order that would best suit the process of protein coding. In relation to the latter, the division in domains furnished the possibility of a degree of flexibility impossible to be obtained otherwise.

Early in the process, the fusion of different kinds of organisms provided the possibility of the development of transitions, changes that opened completely new avenues of life exploration. Mitochondria and plastids originated in this way, and presently modulate the whole process of energy formation and use.

Small or large? Simple or structured? The option was in large part determined by population sizes, modes of reproduction, and the immediate physical and biotic environment. The result is the marvelous assemblage of life forms that are presently seen in the world.

The importance of the direct study of the genetic material in the proportions that have been reviewed here cannot be overemphasized. We are starting to understand whole genomes, their struc-

**TABLE X**

**Recent investigations (papers published or in press) on human population molecular diversity by the Porto Alegre group.**

| Ethnic populations | Subject of the investigation | No. of indiv./ sequences | Systems or DNA region considered[1] | References[2] |
|---|---|---|---|---|
| Two Mongolian, two Siberian, 25 South Amerindian | Presence or absence of haplogroup X | 1,159 | mtDNA control and coding regions | 1 |
| Two Mongolian, two Siberian, 10 Native American | Intra and intercontinental variability | 103 | 3'UTR, *LDLR* gene | 2 |
| African, Asian European, Amerindian | Worldwide variability, presence of selection | 111 | 3'UTR, *LDLR* gene | 3 |
| 11 Amerindian, Five Asian | Intra and inter-population relationships, migration inferences | 678 | L1 and *Alu* insertions | 4 |
| Four Amerindian | Genetic distance and gene diversity analyses | 196 | 15 STRs | 5 |
| Nine Amerindian | Geographical gradients, presence of selection | 315 | APOE variants | 6 |
| Eight Amerindian | Interpopulation variation | 241 | CCR5d32, TCRBV3S1, TCRBV18 | 7 |
| African, Asian, European, Amerindian | World wide variability, America's colonization | 656 | 404 STRs, 17 2-9-site haplotypes | 8 |
| Members of four different Amerindian linguistic families | Congruence between genetic and linguistic classifications | 17,083 | 37 classical systems and 13 STRs | 9 |
| Four admixed neo-Brazilians | Interethic variability | 221 | CFTR gene | 10 |
| Pooled European-derived Brazilian | Interethnic variability | 119 | HVSI/mtDNA, seven Y-chromosome loci | 11 |

[1] mtDNA: mitochondrial DNA; 3'UTR, *LDLR* gene: 3' untranslated region, low density lipoprotein receptor gene; L1: Long interspersed elements, family 1; *Alu*: short interspersed elements identified by this restriction enzyme; STRs: short tandem repeats; APOE: apolipoprotein E; CCR5d32: chemokine receptor 5 delta 32; TCRBV3S1, TCRBV18: T cell receptor gene segments; HVSI: hypervariable segment I. [2] 1. Dornelles et al. (2005); 2. Heller et al. (2004); 3. Fagundes et al. (2005); 4. Mateus Pereira et al. (2005); 5. Kohlrausch et al. (2005); 6. Demarchi et al. (2005); 7. Hünemeier et al. (2005); 8. Salzano and Callegari-Jacques (2005a); 9. Salzano et al. (2005b); 10. Heckman et al. (2005); 11. Marrero et al. (2005).

ture, and how they function to produce these differentiated organisms that always amazed us. Exciting times! Let us enjoy them!

### ACKNOWLEDGMENTS

Our research is sponsored by Programa de Apoio a Núcleos de Excelência (PRONEX), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), and Pró-Reitoria de Pesquisa da Universidade Federal do Rio Grande do Sul (PROPESQ-UFRGS).

### RESUMO

A presente revisão considerou: (a) os fatores que condicionaram a transição inicial entre não-vida e vida; (b) a estrutura e complexidade genômica em procariotos, eucariotos e organelas; (c) a genômica comparada dos cromossomos humanos; (d) a contribuição brasileira a alguns desses estudos. A compreensão do conflito dialético entre liberdade e organização é fundamental para dar significado aos padrões e processos da evolução orgânica.

**Palavras-chave:** evolução molecular, princípios evolutivos, evolução, plantas, evolução humana.

### REFERENCES

ABRAHAMSEN MS ET AL. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304: 441–445.

ADAMS F AND LAUGHLIN G. 2001. Uma biografia do universo. Do big bang à desintegração final. Rio de Janeiro, RJ, Brasil: Jorge Zahar 290 p.

ADAMS MD ET AL. 2000. The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

ADAMS MD, SUTTON GG, SMITH HO, MYERS EW AND VENTER JC. 2003. The independence of our genome assemblies. Proc Nat Acad Sci USA 100: 3025–3026.

ALMEIDA LGP, PAIXÃO R, SOUZA RC, COSTA GC, BARRIENTOS FJA, TRINDADE DOS SANTOS M, ALMEIDA DF AND VASCONCELOS ATR. 2004. A system for automated bacterial (genome) integrated annotation – SABIA. Bioinformatics 20: 2832–2833.

ANONYMOUS. 2003. A vez do boi. Pesq FAPESP 87: 22–23.

ANONYMOUS. 2004. A genética lucrativa do eucalipto. Pesq & Estrat 3 (Suppl. FINEP): 6–7.

APARICIO S ET AL. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297: 1301–1310.

ARMBRUST EV ET AL. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306: 79–86.

ARRUDA P, GOLDMAN MHS, LEITE A, MENCK CFM, NOBREGA FG AND SILVA AM. 2001. Sugarcane transcriptome. A landmark in plant genomics in the tropics. Genet Mol Biol 24: 1–296.

BADA JL AND LAZCANO A. 2002. Some like it hot, but not the first biomolecules. Science 296: 1982–1983.

BELL KS ET AL. 2004. Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. Proc Nat Acad Sci USA 101: 11105–11110.

BERGER G. 2003. Deterministic hypotheses on the origin of life and its reproduction. Med Hypoth 61: 586–592.

BIOLOGY ANALYSIS GROUP. 2004. A draft sequence for the genome of the domesticated silkworm (*Bombix mori*). Science 306: 1937–1940.

BRAZILIAN NATIONAL GENOME PROJECT CONSORTIUM. 2003. The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. Proc Nat Acad Sci USA 100: 11660–11665.

BRENTANI H ET AL. 2003. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Nat Acad Sci USA 100: 13418–13423.

BROOKE NM AND HOLLAND PWH. 2003. The evolution of multicellularity and early animal genomes. Curr Op Genet Developm 13: 599–603.

BRÜGGEMANN H ET AL. 2003. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc Nat Acad Sci USA 100: 1316–1321.

BUELL CR ET AL. 2003. The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. Tomato DC 3000. Proc Nat Acad Sci USA 100: 10181–10186.

C. ELEGANS SEQUENCING CONSORTIUM. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012–2018.

CAMARGO AA ET AL. 2001. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Nat Acad Sci USA 98: 12103–12108.

CAVALCANTI ARO, LEITE ES, NETO BB AND FERREIRA R. 2004. On the classes of aminoacyl-tRNA synthetases, amino acids and the genetic code. Orig Life Evol Biosph 34: 407–420.

CHEN C, GENTLES AJ, JURKA J AND KARLIN S. 2002. Genes, pseudogenes, and *Alu* sequence organization across human chromosomes 21 and 22. Proc Nat Acad Sci USA 99: 2930–2935.

CRANDALL KA AND BUHAY JE. 2004. Genomic databases and the tree of life. Science 306: 1144–1145.

DE DUVE C. 2005. The onset of selection. Nature 433: 581–582.

DE MARCO R ET AL. 2004. Saci-1, -2, and -3 and perere, four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*. J Virol 78: 2967–2978.

DE SOUZA SJ ET AL. 2000. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. Proc Nat Acad Sci USA 97: 12690–12693.

DEHAL P ET AL. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298: 2157–2167.

DELOUKAS P ET AL. 2001. The DNA sequence and comparative analysis of human chromosome 20. Nature 414: 865–871.

DEMARCHI DA, SALZANO FM, ALTUNA ME, FIEGENBAUM M, HILL K, HURTADO AM, TSUNETO LT, PETZL-ERLER ML AND HUTZ MH. 2005. APOE polymorphisms distribution among Amerindians and related populations. Ann Hum Biol 32: 351–365.

DERMITZAKIS ET ET AL. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 420: 578–582.

DIAS NETO E ET AL. 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Nat Acad Sci USA 97: 3491–3496.

DORNELLES CL, BONATTO SL, FREITAS LB AND SALZANO FM. 2005. Is haplogroup X present in extant South American Indians? Am J Phys Anthropol 127: 439–448.

DRISKELL AC, ANÉ C, BURLEIGH JG, MCMAHON MM, O'MEARA BC AND SANDERSON MJ. 2004. Prospects for building the tree of life from large sequence databases. Science 306: 1172–1174.

DUFRESNE A ET AL. 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Nat Acad Sci USA 100: 10020–10025.

DUNHAM A ET AL. 2004. The DNA sequence and analysis of human chromosome 13. Nature 428: 522–528.

EGUIARTE LE, CASTILLO A AND SOUZA V. 2003. Evolución molecular y genómica en angiospermas. Interciencia 28: 141–147.

ESPAGNE E, DUPUY C, HUGHET E, CATTOLICO L, PROVOST B, MARTINS N, POIRIÉ M, PERIQUET G AND DREZEN JM. 2004. Genome sequence of a polydnavirus: insights into symbiotic virus evolution. Science 306: 286–289.

FAGUNDES NJR, SALZANO FM, BATZER MA, DEININGER PL AND BONATTO SL. 2005. Worldwide genetic variation at the 3'-UTR region of the *LDLR* gene: possible influence of natural selection. Ann Hum Genet 69: 389–400.

FELIPE MSS ET AL. 2003. Transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis* by EST analysis. Yeast 20: 263–271.

FERNANDES T. 2003. Código decifrado. C Hoje 33 (194): 50.

FINKLER C, GIACOMET C, MUSCHNER VC, SALZANO FM AND FREITAS LB. 2005. Molecular investigations of pathogenesis-related Bet v 1 homologues in *Passiflora* (Passifloraceae). Genetica 124: 117–125.

FIORAVANTI C. 2000. As descobertas se multiplicam. Pesq FAPESP 60: 32–37.

FIORAVANTI C. 2003. Farta colheita. Pesq FAPESP 91: 44–47.

FRANCO GR ET AL. 1997. Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). DNA Res 4: 231–240.

GALAGAN JE ET AL. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422: 859–868.

GARNIER T ET AL. 2003. The complete genome sequence of *Mycobacterium bovis*. Proc Nat Acad Sci USA 100: 7877–7882.

GIL R ET AL. 2003. The genome analysis of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Nat Acad Sci USA 100: 9388–9393.

GOLDMAN GH ET AL. 2003. Expressed sequence tag analysis of the human pathogen *Paracoccidioides brasiliensis* yeast phase: identification of putative homologues of *Candida albicans* virulence and pathogenicity genes. Eukar Cell 2: 34–48.

GRAHAM DE, OVERBEEK R, OLSEN GJ AND WOESE CR. 2000. An archaeal genomic signature. Proc Nat Acad Sci USA 97: 3304–3308.

GRIMWOOD J ET AL. 2004. The DNA sequence and biology of human chromosome 19. Nature 428: 529–535.

GUYON R ET AL. 2003. A 1-MB resolution radiation hybrid map of the canine genome. Proc Nat Acad Sci USA 100: 5296–5301.

HALLIWELL JJ. 1991. Quantum cosmology and the creation of the universe. Scient Am 265(6): 76–85.

HAUBOLD B AND WIEHE T. 2004. Comparative genomics: methods and applications. Naturwissenschaften 91: 405–421.

HECKMAN MIO, MENDES-JUNIOR CT, TADA MS, SANTOS MG, CABELLO GMK, SALZANO FM, SIMÕES AL AND ENGRACIA V. 2005. CFTR haplotype distribution in the Brazilian Western Amazonian Region. Hum Biol (in press).

HEILLIG R ET AL. 2003. The DNA sequence and analysis of human chromosome 14. Nature 421: 601–607.

HELLER AH ET AL. 2004. Intra- and intercontinental molecular variability of an *Alu* insertion in the 3' untranslated region of the *LDLR* gene. Hum Biol 76: 591–604.

HILLIER LW ET AL. 2003. The DNA sequence of human chromosome 7. Nature 424: 157–164.

HOLDEN MTG ET AL. 2004. Genomic plasticity of the causative agent of melioidosis, *Burkhholderia pseudomallei*. Proc Nat Acad Sci USA 101: 14240–14245.

HOLT RA ET AL. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129–149.

HÜNEMEIER T, NORNBERG I, HILL K, HURTADO AM, CARNESE FR, GOICOECHEA AS, HUTZ MH, SALZANO FM AND CHIES JAB. 2005. T cell and chemokine receptor variation in South Amerindian populations. Am J Hum Biol 17: 515–518.

INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695–716.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001. Initial sequencing and analysis of the human genome. Nature 409: 861–921.

IVANOVA N ET AL. 2003. Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. Nature 423: 87–91.

JOYCE GF AND ORGEL LE. 1999. Prospects for understanding the origin of the RNA world. In: GESTELAND RF, CECH TR AND ATKINS JF (eds.) The RNA world. Cold Spring Harbor: Cold Spring Harbor Press, p. 49–77.

JUDD WS, CAMPBELL CS, KELLOGG EA AND STEVENS PF. 1999. Plant systematics. A phylogenetic approach. Sunderland, MA: Sinauer.

JUNIOR TC, CARRARO DM, BENATTI MR, BARBOSA AC, KITAJIMA JP AND CARRER H. 2004. Structural features and transcript-editing analysis of sugarcane (*Saccharum officinarum* L.) chloroplast genome. Curr Genet 46: 366–373.

KEIM CN, ABREU F, LINS U, BARROS HL AND FARINA M. 2004a. Cell organization and ultrastructure of a magnetotactic multicellular organism. J Struct Biol 145: 254–262.

KEIM CN, MARTINS JL, ABREU F, ROSADO AS, BARROS HL, BOROJEVIC R, LINS U AND FARINA M. 2004b. Multicellular life cycle of magnetotactic prokaryotes. FEMS Microb Lett 240: 203–208.

Kleerebezem M et al. 2003. Complete genome sequence of *Lactobacillus plantarum* WC FS1. Proc Nat Acad Sci USA 100: 1990–1995.

Koehler-Santos P, Lorenz-Lemke AP, Salzano FM and Freitas LB. 2005a. Ecological-evolutionary relationships in *Passiflora alata* from Rio Grande do Sul, Brazil. Braz J Biol (in press).

Koehler-Santos P, Lorenz-Lemke AP, Muschner VC, Bonatto SL, Salzano FM and Freitas LB. 2005b. Molecular genetic variation in *Passiflora alata* (Passifloraceae), an invasive species in southern Brazil. Biol J Linnean Soc (in press).

Kohlrausch FB, Callegari-Jacques SM, Tsuneto LT, Petzl-Erler ML, Hill K, Hurtado AM, Salzano FM and Hutz MH. 2005. Geography influences microsatellite polymorphism diversity in Amerindians. Am J Phys Anthropol 126: 463–470.

Kulcheski FR, Muschner VC, Lorenz-Lemke AP, Stehmann JR, Bonatto SL, Salzano FM and Freitas LB. 2005. Molecular phylogenetic analysis of *Petunia* Juss (Solanaceae). Genetica (in press).

Lorenz-Lemke AP, Muschner VC, Bonatto SL, Cervi AC, Salzano FM and Freitas LB. 2005. Phylogeographic inferences concerning evolution of Brazilian *Passiflora actinia* and *P. elegans* (Passifloraceae) based on ITS (nrDNA) variation. Ann Bot 95: 799–806.

Lynch M and Conery JS. 2003. The origins of genome complexity. Science 302: 1401–1404.

Marra MA et al. 2003. The genome sequence of the SARS-associated coronavirus. Science 300: 1399–1404.

Marrero AR, Leite FPN, Carvalho BA, Peres LM, Kommers TC, Cruz IM, Salzano FM, Ruiz-Linares A, Silva Jr WA and Bortolini MC. 2005. Heterogeneity of the genome ancestry of individuals classified as white in State of Rio Grande do Sul-Brazil. Am J Hum Biol 17: 496–506.

Martin J et al. 2004. The sequence and analysis of duplication-rich human chromosome 16. Nature 432: 988–994.

Mateos SB. 2002. Três novas rotas de ataque à praga. Pesq FAPESP 76: 40–42.

Mateus Pereira LH, Socorro A, Fernandez I, Masleh M, Vidal D, Bianchi NO, Bonatto SL, Salzano FM and Herrera RJ. 2005. Phylogenetic information in polymorphic L1 and *Alu* insertions from East Asians and Native American populations. Am J Phys Anthropol (in press).

McLysaght A, Baldi PF and Gaut BS. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. Proc Nat Acad Sci USA 100: 15655–15660.

Messing J et al. 2004. Sequence composition and genome organization of maize. Proc Nat Acad Sci USA 101: 14349–14354.

Miller W, Makova KD, Nekrutenko A and Hardison RC. 2004. Comparative genomics. Annu Rev Genom Hum Genet 5: 15–56.

Mineta K, Nakazawa M, Cebrià F, Ikeo K, Agata K and Gojobori T. 2003. Origin and evolutionary process of the CNS elucidated by comparative genomics analysis of planarian ESTs. Proc Nat Acad Sci USA 100: 7666–7671.

Monteiro-Vitorello CB et al. 2004. The genome sequence of the gram-positive sugarcane pathogen *Leifsonia xyli* subsp. *Xyli*. Mol Plant-Microb Inter 17: 827–836.

Moran MA et al. 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432: 910–913.

Moura M. 2000. O novo produto brasileiro. Pesq FAPESP 55: 8–15.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Müller S, Finelli P, Neusser M and Wienberg J. 2004. The evolutionary history of human chromosome 7. Genomics 84: 458–467.

Mungall AJ et al. 2003. The DNA sequence and analysis of human chromosome 6. Nature 425: 805–811.

Muschner VC, Lorenz AP, Cervi AC, Bonatto SL, Souza-Chies TT, Salzano FM and Freitas LB. 2003. A first molecular phylogenetic analysis of *Passiflora* (Passifloraceae). Am J Bot 90: 1229–1238.

Nascimento ALTO et al. 2004a. Genome features of *Leptospira interrogans* serovar Copenhageni. Braz J Med Biol Res 37: 459–478.

NASCIMENTO ALTO ET AL. 2004b. Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. J Bacteriol 186: 2164–2172.

NASCIMENTO S. 1995. Origem do universo ganha uma explicação. Zero Hora (Porto Alegre), 24 de março.

NIERMANN WC ET AL. 2004. Structural flexibility in the *Burkholderia mallei* genome. Proc Nat Acad Sci USA 101: 14246–14251.

NISHIYAMA T ET AL. 2003. Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc Nat Acad Sci USA 100: 8007–8012.

ORGEL LE. 2004. Prebiotic chemistry and the origin of the RNA world. Crit Rev Biochem Mol Biol 39: 99–123.

PENA HB, SOUZA CP, SIMPSON AJG AND PENA SDJ. 1995. Intracellular promiscuity in *Schistosoma mansoni*: nuclear transcribed DNA sequences are part of a mitochondrial minisatellite region. Proc Nat Acad Sci USA 92: 915–919.

PENA SDJ. 1996. Third World participation in genome projects. Tr Biotechnol (TIBTECH) 14: 74–77.

PENNACCHIO LA AND RUBIN EM. 2003. Comparative genomic tools and databases: providing insights into the human genome. J Clin Invest 111: 1099–1106.

PENNISI E. 2002. Tunicate genome shows a little backbone. Science 298: 2111–2112.

PEREZ JF. 2003. Genoma: um balanço preliminar. Pesq FAPESP Especial: 20–23.

PROSDOCIMI F, FARIA-CAMPOS AC, PEIXOTO FC, PENA SDJ, ORTEGA JM AND FRANCO GR. 2002. Clustering of *Schistosoma mansoni* mRNA sequences and analysis of the most transcribed genes: implications in metabolism and biology of different developmental stages. Mem Inst Oswaldo Cruz 97 (Suppl. 1): 61–69.

RAFF RA AND KAUFMAN TC. 1983. Embryos, genes and evolution. New York: MacMillan.

RAOULT D, AUDIC S, ROBERT C, ABERGEL C, RENESTO P, OGATA H, LA SCOLA B, SUZAN M AND CLAVERIE J-M. 2004. The 1.2 megabase genome sequence of mimivirus. Science 306: 1344–1350.

RASMUSSEN S, CHEN L, DEAMER D, KRAKAUER DC, PACKARD NH, STADLER PF AND BEDAU MA. 2004. Transitions from nonliving to living matter. Science 303: 963–965.

RAT GENOME SEQUENCING PROJECT CONSORTIUM. 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521.

RAUDSEPP T, SANTANI A, WALLNER B, KATA SR, REN C, ZHANG H-B, WOMACK JE, SKOW LC AND CHOWDHARY BP. 2004. A detailed physical map of the horse Y chromosome. Proc Nat Acad Sci USA 101: 9321–9326.

READ TD ET AL. 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 423: 81–86.

REIS EM ET AL. 2005. Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. Cancer Res 65 (5): 1–7.

SACCONE C AND PESOLE G. 2003. Handbook of comparative genomics. Hoboken, NJ USA: J Wiley & Sons.

SALZANO FM AND CALLEGARI-JACQUES SM. 2005a. Amerindian and non-Amerindian autosome molecular variability – a test analysis. Genetica (in press).

SALZANO FM, HUTZ MH, SALAMONI SP, ROHR P AND CALLEGARI-JACQUES SM. 2005b. Genetic support to proposed patterns of relationship among Lowland South American languages. Curr Anthropol (in press).

SANTOS TM ET AL. 1999. Analysis of the gene expression profile in the *Schistosoma mansoni* cercariae using Expressed Sequence Tags approach. Mol Biochem Parasitol 103: 79–97.

SCHERER NM, THOMPSON CE, FREITAS LB, BONATTO SL AND SALZANO FM. 2005. Patterns of molecular evolution in pathogenesis-related proteins. Genet Mol Biol (in press).

SCHERER SW ET AL. 2003. Human chromosome 7: DNA sequence and biology. Science 300: 767–772.

SCHMUTZ J ET AL. 2004a. Quality assessment of the human genome sequence. Nature 429: 365–368.

SCHMUTZ J ET AL. 2004b. The DNA sequence and comparative analysis of human chromosome 5. Nature 431: 268–274.

SESHADRI R ET AL. 2004. Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. Proc Nat Acad Sci USA 101: 5646–5651.

SHE X, JIANG Z, CLARK RA, LIU G, CHENG Z, TUZUN E, CHURCH DM, SUTTON G, HALPERN AL AND EICHLER EE. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431: 927–930.

SILVA ACR ET AL. 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature 417: 459–463.

SKALETSKY H ET AL. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.

SNYDER M AND GERSTEIN M. 2003. Defining genes in the genomics era. Science 300: 258–260.

SOLTIS DE ET AL. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Tr Pl Sci 9: 477–483.

SUERBAUM S ET AL. 2003. The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. Proc Nat Acad Sci USA 100: 7901–7906.

THOMPSON CE, FERNANDES CL, SOUZA ON, SALZANO FM, BONATTO SL AND FREITAS LB. 2005. Molecular modeling of pathogenesis-related proteins of Family 5. Cell Biochem Biophys (in press).

URETA-VIDAL A, ETTWILLER L AND BIRNEY E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4: 251–262.

VAN SLUYS MA ET AL. 2003. Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J Bacteriol 185: 1018–1026.

VASCONCELOS AT, MAIA MAGM AND ALMEIDA DF. 2000. Short interrupted palindromes on the extragenic DNA of *Escherichia coli* K-12, *Haemophilus influenzae* and *Neisseria meningitides*. Bioinformatics 16: 968–977.

VENTER JC ET AL. 2001. The sequence of the human genome. Science 291: 1304–1351.

VENTER JC, LEVY S, STOCKWELL T, REMINGTON K AND HALPERN A. 2003. Massive parallelism, randomness and genomic advances. Nat Genet 33 (Suppl.): 219–227.

VERJOVSKI-ALMEIDA S ET AL. 2003. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nat Genet 35: 148–157.

VETTORE AL ET AL. 2003. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res 13: 2725–2735.

VINOGRADOV AE. 2004. Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? Curr Op Genet Developm 14: 620–626.

WATERS E ET AL. 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc Nat Acad Sci USA 100: 12984–12988.

WATERSTON RH, LANDER ES AND SULSTON JE. 2003. More on the sequencing of the human genome. Proc Nat Acad Sci USA 100: 3022–3024.

WHITFIELD J. 2004. Born in a watery commune. Nature 427: 674–676.

WOESE CR. 1998. The universal ancestor. Proc Nat Acad Sci USA 95: 6854–6859.

WOESE CR. 2002. On the evolution of cells. Proc Nat Acad Sci USA 99: 8742–8747.

WOLFE KH AND LI W-H. 2003. Molecular evolution meets the genomics revolution. Nat Genet 33 (Suppl.): 255–265.

XU J, BJURSELL MK, HIMROD J, DENG S, CARMICHAEL LK, CHIANG HC, HOOPER LV AND GORDON JI. 2003. A genomic view of the human-*Bacteroides thetaiotamicron* symbiosis. Science 299: 2074–2076.

XU P ET AL. 2004. The genome of *Cryptosporidium hominis*. Nature 431: 1107–1112.

XYLELLA FASTIDIOSA CONSORTIUM. 2000. The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406: 151–159.

ZHOU S ET AL. 2004. Shotgun optical mapping of the entire *Leishmania major* Friedlin genome. Mol Biochem Parasitol 138: 97–106.