

# Genomes and evolution

## Editorial overview

Mark S Boguski\*, David R Cox† and Richard M Myers‡

### Addresses

\*National Center For Biotechnology Information, National Library of Medicine, Bldg 38A, Room 8S-810, 8600 Rockville Pike, Bethesda, Maryland 20894; e-mail: boguski@ncbi.nlm.nih.gov

†Department of Genetics and ‡The Stanford Human Genome Center, Stanford University School of Medicine, Stanford, California 94305-5120, USA

†e-mail: cox@shgc.stanford.edu

‡e-mail: myers@shgc.stanford.edu

**Current Opinion in Genetics & Development** 1996, 6:683–685

© Current Biology Ltd ISSN 0959-437X

### Abbreviations

**EST** expressed sequence tag

**mtDNA** mitochondrial DNA

**NRPY** non-recombining portion of the Y chromosome

“What is true for *E. coli* is true for the elephant”, asserted Jacques Monod [1] during the heroic age of molecular biology when it was first imagined that all of the complexities of living systems could be derived from a few basic principles and mechanisms. Of course, noted Francis Crick, there were many biochemical details to be filled in, so that what was known in outline could also be known in detail. In this issue of *Current Opinion in Genetics & Development*, readers will not find anything about elephants or, sadly, very much about *E. coli*, but they will find outlines of life expressed in the DNA sequences of entire genomes, as well as a sampling of the types of details that Erwin Chargaff described as an “immensely diversified phenomenology” of cells and organisms [1].

The issue starts off with a review by Leipe (pp 686–691) that places the subsequent articles into a context of global biodiversity. Unfortunately, representation of the overall diversity of life at the molecular level is poor and plagued by misconceptions. The community’s selection criteria for “model organisms” to study at the genome level reflects a combination of anthropocentrism, culture, historical circumstances and economic implications, as well as basic research considerations. There is no lack of interest among biologists in expanding our horizons to study the genomes of a much larger variety of living things; doing so would only require that high-throughput DNA sequencing become much cheaper! While you are reading this article and others in the issue, we recommend that you have handy Leipe’s Figure 4 (page 689) and the taxonomy browser page on the World Wide Web [2].

The next four articles are progress reports on the mapping and sequencing of several eukaryotic genomes. The 100 Mb sequence of the genome of *Caenorhabditis elegans* is nearing completion in 1998 and intermediate data

products have stimulated much research. Apart from the intrinsic biological interest of the nematode genome, the strategies and technologies involved in its sequencing are serving as models for projects involving other organisms. Specifically, physical mapping in parallel with cDNA sequencing is an efficient and cost-effective strategy to develop genome sequencing infrastructure as well as to provide grist for biologists’ mills as early as possible. Blumenthal and Spieth (pp 692–698) review some of the biology revealed by the nematode sequence. Interestingly, ~25% of the genes are contained in operons, an observation with important implications for researchers working on other organisms and studying genes that have nematode homologues.

Donelson (pp 699–703) reviews genome research in *Trypanosoma brucei* and *Trypanosoma cruzi* and illustrates the utility of studying things in pairs. These two species are highly divergent evolutionarily, cause different human diseases (trypanosomiasis and Chagas disease) and are biologically dissimilar in that *T. cruzi* is an intracellular parasite whereas *T. brucei* operates extracellularly by antigenic variation and immune evasion. An interesting peculiarity of trypanosome biology is the extensive RNA editing that occurs in the expression of kinetoplast gene expression. This may be an example of a primitive function on its way to extinction. Ivens and Blackwell (pp 704–710) bring us up to date on *Leishmania* genome research, which is further along than trypanosome studies in the sense that the cosmid sequences of genomic DNA are already present in GenBank. Expressed sequence tag (EST) surveys have already provided genes whose protein products are potential candidates for chemotherapeutic targets and/or vaccines. Trypanosomiasis, Chagas disease and leishmaniasis are three of seven major parasitic diseases targeted for genome studies by the World Health Organization and other agencies. The other four diseases and organisms include: schistosomiasis (*Schistosoma mansoni*, *S. japonicum*), filariasis (*Brugia malayi*), toxoplasmosis (*Toxoplasma gondii*) and malaria (*Plasmodium falciparum*).

Rice (*Oryza sativa*) and other cereals are the subject of the article by Havukkala (pp 711–714), who points out that conservation of synteny is extensive among cereal genomes and that this phenomenon is of great utility for interspecies comparisons and the positional cloning of agriculturally important genes. Two highlights in this field over the past year are the positional cloning from rice of the bacterial resistance gene *Xa21* and the demonstration of convergent domestication of cereal crops through independent selection of mutations in corresponding genes controlling quantitative traits in different species.

The next two articles deal with the evolution of the 'hardware' and 'software' of developmental regulatory systems and the comparative mapping of mammalian genomes. Sidow (pp 715–722) explains how genome-wide duplications might have allowed complexity gains through the evolution of developmental regulatory pathways, how the balance may be tipped from entropic decay to pseudogenes to selective advantage, and concludes that "junk [DNA] is useful". He argues for mini-genome projects of amphioxus, hagfish, lamprey, echinoderm and tunicate to better understand early vertebrate evolution. As before, the only obstacle is cheaper genomic sequencing! Eppig (pp 723–730) reviews the current status of mammalian comparative maps and electronic access to comparative mapping data. She stresses that, for comparative purposes, homology (orthology) is more than just sequence similarity. This is advice well-taken, given the genome duplications, the gene families, and the unpredictable loss of paralogues that Sidow describes. Eppig shows how the technique called ZOO-FISH might rapidly expand our knowledge of comparative genome organization among mammalian orders.

Stoneking and Soodyall (pp 731–736), and Mitchell and Hammer (pp 737–742) review molecular studies of human female and male origins based on sequence variations in mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRPY), respectively. Human origins assessed by mtDNA, the genetic properties of mtDNA, the history of specific populations and the analysis of ancient remains are also discussed. The use of Y chromosome sequences for studying human lineages is a more recent approach compared with the use of mtDNA but some estimates on the time of the most recent common Y ancestor overlap with dates commonly cited for the mtDNA ancestor. Such studies are controversial and Mitchell and Hammer wisely point out that further sequence data on many more males is required to confirm this timing and to provide information on the geographical location of the ancestral Y chromosome. Nevertheless, studies of Y chromosome polymorphisms in even modest numbers of individuals are increasing our understanding of more recent events, such as human migrations and colonizations.

The next two articles by Smit (pp 743–748) and McClure (pp 749–756) deal with aspects of retroviral genomes, their evolution in 'real time', and the retroviral (and other) origins of interspersed repeats in the human genome. Even though the complete genomic sequences of HIV and other retroviruses have been available for over a decade, our understanding of their biology and evolution, particularly with respect to medical interventions for their control, is incomplete. Interspersed repeats in vertebrate genomes, of which vestigial retroviruses are one important class, belong to a small number of well-defined categories and their origins and mechanisms of amplification are understood.

As such elements make up a very large fraction of human genomic DNA, knowledge of them is critical for the analysis and interpretation of high-throughput sequence data, a hundred million nucleotides of which is anticipated over the next two years.

Over the past two years, the complete sequences of two eubacterial genomes have become available; Koonin and Mushegian (pp 757–762) have performed extensive analysis of their genes and devised a "minimal gene set" common to both organisms and the 75% of the *E. coli* genome that is currently available. This set may indicate the biochemical functions and pathways sufficient for sustaining cellular life. Furthermore, these authors speculate that this set—along with information derived from the complete sequences of a eukaryote (*Saccharomyces cerevisiae*) and an archeon (*Methanococcus jannaschii*)—may permit the reconstruction of the genome of the last common ancestor of the three biological kingdoms. This work takes us to the threshold of fulfilling what Monod probably considered only an intellectual exercise when he said: "The secret of life is in large part known—in principle if not in all details. For a simple living creature to be synthesized, there is no further principle that would need to be discovered" [1]. In addition to providing a framework for theoretical evolutionary genomics, the minimal gene set is of practical importance for the analysis of dozens of additional prokaryotic genomic sequences that are expected to be finished over the next two years.

So what does one do in a post-genome sequence era? The answer to this question is becoming clear for yeast, as Bassett *et al.* (pp 763–766) describe. There are systematic efforts to link genes with phenotypes, to study large-scale gene expression in different biological states, and to use sequence information to aid in the design and interpretation of protein–protein interaction studies. By systematically cross-referencing yeast genes with their counterparts in other organisms, the experimental analysis of their functions can also be greatly accelerated.

As always, evolution gets the last word, and this volume ends with an appraisal by Takahata (pp 767–772) of the current status of Kimura's neutral theory of molecular evolution. It is interesting to note that, when Kimura published his theory in 1968, GenBank would not be created for another 14 years and even the development of rapid DNA sequencing technology was still eight years in the future. Takahata describes reformulation of neutral theory in terms of gene genealogy that has been made possible by the wealth and detail of DNA sequence data. Every time we align sequences for the purpose of identifying conserved and non-conserved regions, or when we make an assertion about human origins on the basis of sequence variation or polymorphism, the neutral theory of molecular evolution is implicit in our analysis.

What topics and developments of the past year are missing from this issue? There are many, to be sure, and it was not our intention to slight research on *Arabidopsis*, *Drosophila*, *fugu*, zebrafish, or any other species. Also, nothing is included about the new map of 16 000 human genes [3], nor is there a description of the near 100 000 mouse ESTs currently available. We also wonder what Jacques Monod would have thought about the possibility of bacterial life on Mars? There is only so much room, so much time, and so much one can do to convince authors to take time away from their research to write a review. Nevertheless, there is much to learn in the present set of papers about

the outlines and details of our current knowledge about genomes and evolution.

## References

1. Judson HF: *The Eighth Day of Creation: Makers of the Revolution in Biology. Expanded Edition*. New York, Cold Spring Harbor Laboratory Press; 1996.
2. The National Center for Biotechnology Information Homepage on World Wide Web URL: <http://www.ncbi.nlm.nih.gov>
3. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tomé P, Aggarwal E, Bajorek E *et al.*: **A gene map of the human genome**. *Science* 1996, **274**:540–546.

## Other *Current* articles of interest to readers of this issue

### Research papers

**Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*** by Roman L Tatsuov, Arcady Mushegian, Peer Bork, Nigel P Brown, William S Hayes, Mark Borodovsky, Kenneth E Rudd and Eugene V Koonin. *Current Biology* 1996, **6**:279–291

**Codon usage limitation in the expression of HIV-1 envelope glycoprotein** by Jürgen Haas, Eun-Chung Park and Brian Seed. *Current Biology* 1996, **6**:315–324

**Genomic differentiation among natural populations of orang-utan (*Pongo pygmaeus*)** by Lu Zhi, William B Karesh, Dianne N Janczewski, Harmony Frazier-Taylor, Dondin Sajuthi, Francis Gombek, Mahedi Andau, Janice S Martenson and Stephen J O'Brien. *Current Biology* 1996, **6**:1326–1336

### Reviews and dispatches

**Invasion, control and persistence of *Leishmania* parasites** by Christian Bogdan, André Gessner, Werner Solbach and Martin Rölinghoff. *Current Opinion in Immunology* 1996, **8**:517–525

**Evolution and plasticity of CTL responses against HIV** by Brigitte Autran, Fabienne Hadida and Gaby Haas. *Current Opinion in Immunology* 1996, **8**:546–553

**Mitochondrial DNA: Molecular fossils in the nucleus** by Nicole T Perma and Thomas D Kocher. *Current Biology* 1996, **6**:128–129

**Sequencing and analysis of bacterial genomes** by Eugene V Koonin, Arcady Mushegian and Kenneth E Rudd. *Current Biology* 1996, **6**:404–416

**Genome sequencing: The complete code for a eukaryotic cell** by Mark Johnston. *Current Biology* 1996, **6**:500–503

**Molecular evolution: A difficult phase for introns-early** by Laurence D Hurst and Gilean T McVean. *Current Biology* 1996, **6**:533–536

**Evolution: When was life's first branchpoint?** by Brian Golding. *Current Biology* 1996, **6**:679–682

**Mimicry meets the mitochondrion** by James Mallet, Chris D Jiggins and W Owen McMillan. *Current Biology* 1996, **6**:937–940

**Phylogenetic trees: Whither microbiology?** by Carl R Woese. *Current Biology* 1996, **6**:1060–1063

**Speciation: More than the sum of its parts** by NH Barton. *Current Biology* 1996, **6**:1244–1246

**Genomes: *Methanococcus janaschii* and the golden fleece** by Roger A Garrett. *Current Biology* 1996, **6**:1377–1380

All these articles are also available online in the BioMedNet library

<http://BioMedNet.com/>