Comparative Genomics

I King Jordan, *National Institutes of Health, Bethesda, Maryland, USA*
Eugene V Koonin, *National Institutes of Health, Bethesda, Maryland, USA*

Comparative genomics involves the comparison of features of completely sequenced (or nearly so) genomes. Comparative sequence analyses can facilitate both the functional annotation of genomes and whole-genome approaches to evolutionary issues.

Introductory article

Article contents

doi: 10.1038/npg.els.0005296

Comparative Methods in Biology and Genomics

The comparative approach in biology is probably as old as this science itself. Systematic comparative analysis of organisms can be traced at least to the eighteenth century naturalist Carl Linnaeus, who developed the hierarchical classification of plant and animal species on the basis of detailed observations of comparative morphology. The first half of the nineteenth century was marked by outstanding achievements of comparative anatomy, as exemplified by the work of George Cuvier and Richard Owen. Owen coined the term 'homology' to describe similar, albeit modified, 'organs' with common underlying structures that are shared between different species. These scientists lacked the evolutionary perspective, however. It was left to Darwin and his followers to emphasize that homologous structures result from descent with modification from a common ancestor. For more than a century after the publication of Darwin's *Origin of Species*, comparative morphology flourished as the preeminent method of discerning evolutionary relationships among species and ordering the vast organismic diversity of the natural world.

The last 40 years of the twentieth century witnessed the burgeoning of molecular evolutionary studies pioneered by Emile Zuckerkandl and Linus Pauling in 1962. Molecular evolution can be considered to consist mainly of two subdisciplines: (1) the use of molecules as markers to reconstruct the evolutionary histories of organisms and (2) the study of the nature of evolutionary change of the molecules (genes and proteins) themselves. Molecular evolutionary studies have resulted in previously unimaginable advances in our understanding of fundamental evolutionary events and processes. For example, phylogenetic analysis, primarily by Carl Woese and colleagues, of ribosomal ribonucleic acid (RNA) sequences resulted in the abandonment of the established five-kingdom taxonomic system. Their discovery of a completely phylogenetically distinct form of life, the archaea, precipitated the reorganization of the biosphere into primary domains: bacteria, archaea and eukarya. The theoretical foundation of molecular evolution, the neutral theory, developed primarily by Motoo Kimura, provided a new understanding of the mode of genome evolution – that is, that the vast majority of changes at the molecular level were neutral, or nonadaptive, with respect to organismic fitness. The neutral theory of molecular evolution is at present widely used as a null hypothesis against which to test results bearing on the molecular basis of adaptation. A corollary of the neutral theory that is critical for functional inferences from comparisons of molecular sequences is that certain portions of these sequences are conserved because they are subject to strong purifying selection and, accordingly, are functionally important.

Molecular evolution studies flourished in the pre-genomics era. However, the rules of the game were forever altered in 1995, with the availability of multiple complete genome sequences. At that time, it became possible to employ whole-genome comparisons to address functional and evolutionary questions. Comparative genomics, in the modern sense, was born. Once the complete genome sequence of an organism is available, it becomes possible, at least in principle, to deduce the entire sets of genes and proteins and to obtain comprehensive information on the linear order of genes on chromosomes. Thus, it became possible to confidently (with some limitations, owing to imperfect methods of gene identification and comparison, but, on the whole, with a high reliability) infer the presence or absence of a given gene, gene family or functional class of gene from a genome. It was also possible to address higher-order global questions about the evolution of gene order, through the comparison of complete genomes.

Comparative analysis of complete genomes has facilitated the development of a robust natural classification system for genes (or proteins). Such a

classification system relies on the distinction between two different classes of homologous gene: orthologs versus paralogs. Orthologs are genes that have diverged from a common ancestor as a result of speciation, whereas paralogs are genes that have diverged as a result of gene duplication. Reliable and transparent distinctions between these two classes of homolog can be made only when complete predicted proteomes are compared. This delineation has important functional implications. Proteins encoded by orthologous genes tend to perform the same or very similar functions in different species; in contrast, proteins encoded by paralogous genes often diverge in function. Thus, the distinction between orthologs and paralogs is critical when the function of an uncharacterized gene (e.g. from a newly sequenced genome) is predicted on the basis of a known function of a homolog from another species.

The importance of the comparative approach for functional annotation of the sequenced genomes cannot be overestimated. In most cases, experimentally derived functional characterization of genes lags far behind genome sequencing. This gap is expected only to widen in the foreseeable future. Thus, functional annotation of genomes relies primarily on information transfer from experimentally characterized genes to their homologs in other genomes. Such information transfer is possible only because the majority of protein sequences encoded in each genome show high levels of sequence conservation across a broad range of species (e.g. for bacterial and archaeal genomes, it has been observed that 70–85% of the proteins are conserved in at least three taxonomically distant lineages). The validity of the conclusions reached on the basis of information transfer critically depends on the correct interpretation of homologous relationships, in terms of orthology versus paralogy.

This article briefly describes the approaches and some major findings of comparative genomics, with an emphasis on its contributions to evolutionary biology. The specific results of comparative genomics that are important for the analyses and understanding of the draft sequence of the human genome are also discussed.

Approaches and Findings in Comparative Genomics

Levels of analysis

Comparative genomics entails comparisons of various features of (nearly) completely sequenced genomes. Such comparisons can be performed at several different levels of biological organization. For instance, nucleotide sequences can be compared between homologous genes within and between species. Because nucleotide sequences consist of only four characters, nucleotide variation rapidly becomes saturated, owing to multiple substitutions, as sequences diverge over time. Once the changes between sequences are saturated ($>50\%$ divergence), it becomes difficult if not impossible to accurately recognize and align homologous sequences. The comparison of nucleotide sequences is therefore limited to relatively closely related organisms, in the case of orthologous genes, or to recently duplicated paralogs. Genomic studies that involve such comparisons of closely related sequences (at or below the species level) can be referred to as 'microevolutionary genomics'.

Comparative analysis can also be done with amino acid sequences of homologous proteins. Amino acid sequence comparisons are much more useful for the detection and characterization of distant evolutionary relationships. Amino acid sequences consist of 20 different characters, and saturation of substitutions occurs much more slowly than for nucleotide sequences. Using recently developed sensitive methods and statistical procedures, amino acid sequences with as much as 80% (or even greater) divergence can be often reliably identified as homologous and aligned.

Higher-order comparisons between genomes often focus on the presence or absence of genes, the composition and distribution of paralogous gene families, and the relative gene order across genomes. Such higher-order comparisons, as well as amino acid sequence comparisons, are most frequently used in macroevolutionary genomics, that is, evolutionary studies that focus on relationships above the species level.

Macroevolutionary Genomics

Lineage-specific expansions

Gene duplication is a crucial evolutionary force that often results in functional diversification between paralogous genes. The presence and extent of duplicate genes in a genome can be inferred by performing a series of sequence similarity searches (e.g. with Basic Local Alignment Search Tool, or BLAST) using each protein encoded by the genome as a query in searches against the entire predicted proteome. Proteins from the same genome that show significant similarity to one another are likely to be encoded by genes that were duplicated at some point in the evolutionary history of the genome being studied. When these types of comparative analysis are performed, it becomes readily apparent that genomes are

full of duplicated genes. For example, in bacterial species, paralogous gene families typically make up approximately 50% of the genes in the genome, and eukaryotic genomes are even more enriched in duplicated genes.

From an evolutionary perspective, a particularly interesting class of duplicated genes is made up of lineage-specific expansions. These expansions result from gene duplications that have occurred along one evolutionary lineage subsequent to the diversification from the last common ancestor shared with other analyzed lineages. Such paralogous groups are delineated by identifying all genes (or proteins) from the same genome (or a group of genomes of related species) that are more closely related to each other than to any proteins encoded by any of the other genomes being compared. Lineage-specific expansions often contribute substantially to species-specific coding repertoires and are likely to have special adaptive significance. For example, pathogenic bacterial genomes, such as *Mycobacterium tuberculosis* and *Helicobacter pylori*, encode lineage-specific expansions of membrane proteins that are thought to be involved in the interaction with the target cells of their host organisms. The enhanced cell surface variability provided by these expansions is likely to be involved in the avoidance and escape from host immune surveillance. Each of the three currently available animal genomes, those of the nematode *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster* and *Homo sapiens*, shows massive, lineage-specific expansions of chemoreceptors that facilitate the ability of each animal to process and respond to environmental cues in a species-specific manner. The plant *Arabidopsis thaliana* genome and other, partially sequenced plant genomes encode lineage-specific expansions of ATPases that are homologous to animal ATPases involved in programmed cell death and are involved in the plants' resistance to various pathogens. These are just a few of the many examples of lineage-specific expansions that have shaped the evolution of the organisms that encode them.

Gene loss

Gene loss along specific phylogenetic lineages is the evolutionary converse of lineage-specific expansion. Comparative genomics has shown that gene loss, like lineage-specific expansion, is extremely common. Careful examination of gene loss can also yield important clues about organismic adaptation and phenotype.

The extent of lineage-specific gene loss was first noticed upon comparative analysis of bacterial genomes. There are a number of parasitic bacteria that have substantially reduced genome sizes compared with their free-living relatives. For example, the smallest known cellular genome, that of *Mycoplasma genitalium*, contains a mere 480 genes, as opposed to the roughly 4000 genes in the free-living bacterium *Bacillus subtilis*, which belongs to the same bacterial lineage. Such a drastic reduction of genes is facilitated by the parasitic lifestyle of the organism. *M. genitialium* and other bacterial parasites are able to assimilate many of the metabolites from their hosts that nonparasitic organisms have to produce themselves. As parasites maintain the intimate relationship with their hosts over time, genes encoding metabolic enzymes that are involved in the reactions that yield the same products that are provided by the hosts are lost.

Lineage-specific gene loss is not limited to parasitic bacteria. Comparative genomic analyses have also revealed evidence of substantial gene loss in eukaryotes. The baker's yeast, *Saccharomyces cerevisiae*, has lost approximately 300 genes since its divergence from the fission yeast *Schizosaccharomyces pombe*. This represents about 5% of the genome and is likely to be an underestimate of the total gene loss in *S. cerevisiae*, as it does not reflect gene losses that occurred before the divergence of *S. cerevisiae* and *S. pombe*. The gene loss in *S. cerevisiae* seems to have been remarkably coordinated from an adaptive perspective, as there was a marked coelimination of genes that interact in the same functional pathways. For example, there was a parallel loss of many of the components of the spliceosome that are involved in removing introns from pre-messenger RNA (pre-mRNA). This is entirely consistent with the virtual absence of spliceosomal introns in *S. cerevisiae*. An evaluation of the genes lost in *S. cerevisiae* not only confirms and enhances previous knowledge of several aspects of its biology, it can also be used to make predictions about groups of genes with the potential to functionally interact. That coelimination of functionally interacting genes is so prevalent suggests that the absence of a group of genes in any lineage may indicate a previously unappreciated functional connection between them.

Nonorthologous gene displacement

Convergent evolution is the independent evolution of similar features from distinct ancestral states, and it implies a strong adaptive significance for convergently derived features. Nonorthologous gene displacement is a striking example of this phenomenon. Numerous cases have been demonstrated where unrelated, independently evolved (nonorthologous) proteins can carry out the same biochemical function in different species or evolutionary lineages. Complete genome

sequences are essential for identifying cases of nonorthologous gene displacement. This is because, without complete sets of predicted proteins, it may be the case that a homologous gene exists, encoding the same function in all species considered, that has not yet been detected in one or more of the species. With complete genome sequence comparison, one can be confident that an ortholog that encodes an essential cellular function is missing from some of the compared species.

Perhaps the most striking example of nonorthologous gene displacement involves several components of the deoxyribonucleic acid (DNA) replication machinery. DNA replication is essential to all cellular organisms, and there are obvious functional similarities among the replication machineries of all three domains of life. However, there is evidence of non-orthologous gene displacement for several of the core components of DNA replication where unrelated or distantly related, but apparently not orthologous, proteins perform the same function in bacteria and in archaea/eukarya. For example, the main catalytic subunits of DNA polymerases and primases in bacteria appear to be unrelated to their functional counterparts in the archaea/eukarya. In addition, ATPases involved in the initiation of DNA replication are homologous, but not orthologous, between bacteria and archaea/eukarya. It is thought that non-orthologous ATPase domains were independently recruited to function in DNA replication in the different groups. Cases of nonorthologous gene displacement highlight the adaptive significance of the function in question and underscore the ability of evolution to solve the same problem independently using different genetic materials.

Horizontal gene transfer

Horizontal (lateral) gene transfer is the nonsexual transmission of genetic material across species boundaries. In the pregenomics era, it was widely assumed that horizontal transfer was a rare and, more or less, inconsequential occurrence. Comparative genomic analyses, particularly among prokaryotes, have shown horizontal transfer to be astonishingly widespread.

Striking examples of the significance of horizontal transfer in genome evolution were uncovered by comparison of the genome sequences of bacterial and archaeal hyperthermophiles. The bacteria *Aquifex aeolicus* and *Thermotoga maritima* are exceptional in that they live in hyperthermophilic environments that tend to be dominated by archaeal species. The complete genome sequences of these bacterial hyperthermophiles contain literally hundreds of genes that appear to have been acquired via horizontal transfer from archaea, as indicated by their anomalously high sequence similarity to archaeal orthologs. The genes acquired from archaeal hyperthermophiles via horizontal transfer might have contributed to the adaptation of these bacterial species to their extreme environments. However, the high numbers of horizontally transferred genes may also partly reflect the fact that these organisms share the same environment and thus have ample opportunities for genetic exchange.

Horizontal transfer of genes among bacterial and archaeal genomes is so common that it challenges basic ideas about the topology and even the existence of the tree of life for prokaryotes. The finding that a substantial fraction of each prokaryotic genome is derived from horizontal transfer – in other words, that prokaryotic genomes are fundamentally chimeric – indicates that their evolutionary history cannot be accurately represented as a bifurcating tree. The question remains whether a core of genes not subject to frequent horizontal transfers can be identified and used to construct a tree that would reflect major evolutionary trends. There is also considerable evidence of horizontal gene transfer involving eukaryotic genomes, particularly the transfer of genes from endosymbiotic organelles, mitochondria, and chloroplasts, into the eukaryotic nuclear genome. Other bacterial symbionts and parasites also might have contributed genes to the eukaryotic genomes, but the extent of these contributions remains unclear.

Comparative genomics and reconstruction of the history of life

One of the most ambitious goals of macroevolutionary genomics is to provide a comprehensive understanding of the history of life from a molecular perspective. The construction of phylogenetic trees that reflect the evolutionary history and relationships among all species, a monumental task in and of itself, is only the starting point of this endeavor. Detailed and comprehensive comparative genomic studies that attempt to classify and to order the entire 'universe' of protein domains are simultaneously being undertaken. This information is then combined with structural and functional considerations of paralogous protein families, superfamilies and folds, considered with respect to organismic phylogeny, and finally used to trace the evolution of the fundamental biochemical processes that characterize cellular life. These types of study have the potential to shed light on various aspects of ancient cellular machinery and to elucidate the molecular underpinnings of major evolutionary transitions. For instance, comparative analysis of

DNA replication and translation machineries among the three domains of life indicates that the last common ancestor of all cellular life might not have contained a double-stranded DNA genome typical of extant cellular life forms. Instead, this ancestor possibly contained a genetic system with both RNA and DNA, where DNA was replicated by reverse transcription of RNA. Further analysis of the distribution of protein families involved in translation among different life forms, combined with the data on the catalytic activities of ribozymes (RNA enzymes), suggests that ancient life forms carried out translation using catalytic RNAs that functioned together with protein cofactors and subsequently with crude protein enzymes that had low specificity. Over time, the protein enzymes involved in translation experienced a tremendous expansion and functional diversification that allowed them to usurp most of the catalytic roles originally fulfilled by RNA.

Human Comparative Genomics

Human gene number

Perhaps the single greatest surprise from the Human Genome Project was the estimate of roughly 30 000 for the number of human genes. This number is quite low when considered with respect to previous estimates for the human gene number, many of which hovered around 100 000. The estimate is even more startling when one considers how close it is to the gene number estimates of other completely sequenced eukaryotic genomes: about 20 000 for *C. elegans* (worm) genes; about 13 000 for *D. melanogaster* (fly); and about 26 000 for *A. thaliana* (mustard plant). It seems fairly obvious and intuitive that humans are the most complex among these species. If so, the low estimate for the human gene number provides a direct challenge to the notion that the organismic complexity of the human lineage can be accounted for directly by an increase in the number of genes. However, organismic complexity is a difficult concept to pin down. For example, one may evaluate genomic, regulatory, developmental and behavioral complexity separately. A general way to measure organismic complexity is by the number of different parts and, by extension, the number of interactions between parts. So, in terms of genomic complexity (gene number), humans are not appreciably more complex than many invertebrate organisms. When it comes to different cell or tissue types (a widely used measure of organismic complexity), however, humans are in fact far more complex than invertebrates.

The low estimate of the human gene number, which was consistently produced by the two independent groups that sequenced the human genome, calls for a reconsideration of the genetic determinants of human biological complexity. One possibility is that human genes, on average, produce more protein variants than those of less complex eukaryotes. The production of multiple proteins from a single gene can be achieved by the alternative splicing of transcribed RNAs resulting in the production of mature mRNAs with different combinations of exons. Comparative analysis of a subset of human and worm mRNA sequences revealed that alternative splicing was far more prevalent for human genes. However, more work needs to be done to verify whether this result is representative and extends to other species.

It is also possible that posttranslational modification of proteins may be more extensive in complex eukaryotes. A new field, proteomics, aims to study the protein products of genomes directly, as opposed to relying on protein sets predicted from the genome sequence. Proteomic analyses confirm that the total number of human proteins is far greater than the number of human genes, but it is not clear whether this effect is relatively greater for humans than for other, less complex eukaryotes.

A final possibility is that individual genes themselves are more complex in the human genome than in less complex eukaryotes. Indeed, the human genome encodes substantially more proteins that can be classified into more than one functional category than do the genomes of other completely sequenced eukaryotes. This suggests that human proteins may be more multimodal in their functional capacity than are proteins of the less complex eukaryotes. Perhaps even more importantly, human genes on average tend to contain more domains than do genes encoded by the complete invertebrate genomes. This results from a process known as domain accretion that seems to have been a major contribution to the increased biological complexity of humans, as discussed next.

Domain evolution and accretion

Protein sequences can be broken down into discrete domains, distinct structural and functional units that have partially independent evolutionary trajectories, as indicated by the formation of diverse domain combinations (domain architectures). Multidomain proteins whose architecture tends to change during evolution are considerably more common in eukaryotes than in prokaryotes. Comparative analysis of eukaryotic genomes must therefore focus on domains, rather than entire proteins, as the fundamental unit of study. A comparative census of domains encoded in eukaryotic genomes was performed as part of the analysis of the draft human sequence. Compared

with fungi, plants and invertebrates, the human genome shows remarkable expansions of paralogous families of domains involved in transcription regulation and cytoskeletal/structural and defense/immunity functions. Similar, but not as pronounced, proliferation of paralogs was detected in other eukaryotes. Lineage-specific expansion of paralogous families seems to be one of the major routes for increasing complexity and may be an important adaptive mechanism, particularly in eukaryotes.

Along with the delineation of lineage-specific expansions, comparative analysis of the eukaryotic proteomes allowed for the characterization of protein domains that are specific to the vertebrate lineage. These are recently evolved domains for which homologs cannot be found in other eukaryotes. The invention of new protein domains was a rare event in the evolution of the vertebrate lineage, with only about 7% of the detected domains identified as vertebrate-specific. Furthermore, only one of the vertebrate-specific domains is an enzyme. Thus, almost all human enzymes have relatively ancient evolutionary origins. The vertebrate-specific set of domain families is enriched for proteins involved in defense and immunity as well as those that function in the nervous system. These classes of recently and/or rapidly evolving proteins probably have a special role in determining the unique biological characteristics of vertebrates.

While the evolution of new protein domains along the lineage that led to humans appears to have been rare, there was a substantial evolutionary increase in new proteins through the 'invention' of new domain architectures. The predicted human proteome was shown to contain almost twice as many distinct domain architectures as the fly or the worm, and almost six times as many as the yeast. In many cases, new domain architectures in vertebrates evolved by the process of domain accretion, whereby new domains are added to the ends of preexisting proteins. For example, a number of human chromatin-associated proteins have domain architectures that are identical in the central region to those of the fly, worm and yeast but differ markedly at the ends, owing to the addition of extra domains. Domain accretion seems to be an important mode of increasing biological complexity without increasing the actual number of genes.

Transposable elements

The results from comparative analyses of the human genome and other eukaryotic genomes discussed above centered on the protein-coding genes. However, protein-coding sequences make up only about 1.5% of the entire human genome sequence. The most abundant class of human genomic sequence is made up of transposable elements (TEs). These elements are repetitive genomic sequences that are able to move (transpose) from one genomic location to another. Almost 50% of the human genome can be unequivocally demonstrated to be related to TE sequences. Most of these sequences are relatively ancient insertions that probably no longer have the capacity to transpose. The figure of 50% is probably an underestimate for the fraction of the genome made up of TE sequences, as many of these evolve rapidly and have probably changed beyond recognition. More than anything else, the human genome is a vast collection of TE-derived sequences, which are only sparsely interspersed with 'real' genes.

When TEs move, they often replicate themselves in the genome. This gives the TEs an enhanced transmission rate relative to nonmobile, protein-coding genes that are transmitted in a strict vertical fashion. The presence and abundance of TEs in eukaryotic genomes can be explained solely by their ability to out-replicate the nontransposing parts of the genome. In this sense, the TEs are often considered to be 'selfish', or 'junk', DNA whose contribution to organismic evolution is negligible. However, a growing body of evidence demonstrates many ways in which TEs have contributed to the phenotypic evolution of their hosts. Analysis of the human genome revealed more than 500 cases of protein-coding sequences that appear to be derived from TEs. In addition to contributing to protein-coding sequence evolution, TEs seem to provide regulatory sequences to thousands of human genes.

As with domain accretion, these data are consistent with the notion, first articulated by François Jacob, that evolution works as a tinkerer. Instead of creating new features from scratch, evolution will more often rearrange whatever materials are at hand to create novelty. Transposable elements are extraordinarily abundant in eukaryotic genomes and contain both regulatory and protein-coding sequences. Thus, they seem to be readily available genetic building blocks with which evolution can tinker to create new genes and modify existing ones.

Prospects and Challenges

By combining the conceptual tools of molecular evolution with the abundance of data made available by genome projects, comparative genomics has opened up new vistas on fundamental evolutionary processes. However, much more remains to be done. With every new genome that is completed, there remains a substantial fraction of genes, for example roughly 30% of human genes, that are not evolutionarily

conserved, do not belong to any characterized family and for which no functional prediction can be made. For practical reasons, these 'orphan' proteins are often neglected in comparative genomic analyses. In many cases, these proteins are likely to have evolved so rapidly that the similarity to their relatives can no longer be recognized. Such rapid evolution characterizes the response to adaptive selection pressure that results in the evolution of new functions. Thus, the genes that encode this overlooked set of proteins may be the most responsible for evolutionary diversification between species. A systematic comparative treatment of these genes has the potential to yield many new insights into the molecular basis of adaptation.

Furthermore, the demonstration that TEs contribute regulatory and coding sequences to many vertebrate genes gives us only the tip of the proverbial iceberg of evolution of new genes. Understanding where new genes come from is one of the crucial goals of comparative genomics, and currently we are far from reaching this goal.

Even with respect to evolutionarily conserved genes, the current results of comparative genomics can only be viewed as preliminary. New genome sequences from diverse branches of life, and careful, genome-wide phylogenetic analysis, are required to reconstruct the evolutionary scheme for each conserved family and to eventually come up with a satisfying picture of life's evolution.

See also

Bacterial DNA in the Human Genome
Fugu: The Pufferfish Model Genome
Homologous, Orthologous and Paralogous Genes
Orthologs, Paralogs and Xenologs in Human and Other Genomes
Transposable Elements: Evolution

Further Reading

Bork P (ed.) (2000) *Advances in Protein Chemistry*, vol. 54. New York, NY: Academic Press.

Dacks JB and Doolittle WF (2001) Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* **107**: 419–425.

Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* **284**: 2124–2129.

Fitch WM (2000) Homology: a personal view on some of the problems. *Trends in Genetics* **16**: 227–231.

Fraser CM, Eisen J, Fleischmann RD, Ketchum KA and Peterson S (2000) Comparative genomics and understanding of microbial biology. *Emerging Infectious Diseases* **6**: 505–512.

Green P and Koonin EV (eds.) (1999) Genomes and evolution. *Current Opinion in Genetics and Development* **9**: 621–722.

Henikoff S, Greene EA, Pietrokovski S, *et al*. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**: 609–614.

Koonin EV, Aravind L and Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. *Cell* **101**: 573–576.

Koonin EV, Makarova KS and Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* **55**: 709–742.

Kreitman M and Gaasterland T (eds.) (2001) Genomes and evolution. *Current Opinion in Genetics and Development* **11**: 601–690.

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

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science* **278**: 631–637.

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

Waterston RH, Lindblad-Toh K, Birney E, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.