Primer

Comparative Genomics

Ross C. Hardison

A complete genome sequence of an organism can be considered to be the ultimate genetic map, in the sense that the heritable characteristics are encoded within the DNA and that the order of all the nucleotides along each chromosome is known. However, knowledge of the DNA sequence does not tell us directly how this genetic information leads to the observable traits and behaviors (phenotypes) that we want to understand. Finding all the functional parts of genome sequences and using this information to improve the health of individuals and society are the focus of the next phase of the Human Genome Project (Collins et al. 2003). Comparative analyses of genome sequences will be a major part of this effort.

The major principles of comparative genomics are straightforward. Common features of two organisms will often be encoded within the DNA that is conserved between the species. More precisely, the DNA sequences encoding the proteins and RNAs responsible for functions that were conserved from the last common ancestor should be preserved in contemporary genome sequences. Likewise, the DNA sequences controlling the expression of genes that are regulated similarly in two related species should also be conserved. Conversely, sequences that encode (or control the expression of) proteins and RNAs responsible for differences between species will themselves be divergent.

Different questions can be addressed by comparing genomes at different phylogenetic distances (Figure 1). Broad insights about types of genes can be gleaned by genomic comparisons at very long phylogenetic distances,

Primers provide a concise introduction into an important aspect of biology that is of broad and current interest.

Ross C. Hardison is at the Center for Comparative Genomics and Bioinformatics at The Pennsylvania State University in University Park, Pennsylvania, United States of America. E-mail: rch8@psu.edu

DOI: 10.1371/journal.pbio.0000058

Glossary

Conserved: Derived from a common ancestor and retained in contemporary related species. Conserved features may or may not be under selection.

Evolutionary drift: The accumulation of sequence differences that have little or no impact on the fitness of an organism; such neutral mutations are not under selection. Sequence polymorphisms arise randomly in a population, most of which have no effect on function. Stochastic processes allow a small fraction of these to increase in frequency until they are fixed in a population; these are detectable as neutral substitutions in interspecies comparisons.

Homologs: Features (including DNA and protein sequences) in species being compared that are similar because they are ancestrally related.

Negative selection: The removal of deleterious mutations from a population; also referred to as **purifying selection**.

Nonredundant protein sets: The set of proteins from which similar proteins, derived from duplicated genes, have been removed.

Orthologs: Homologous genes (or any DNA sequences) that separated because of a speciation event; they are derived from the same gene in the last common ancestor. Orthologs are distinguished from **paralogs**, which are homologous genes that separated because of gene duplication events.

Phylogenetic distances: Measures of the degree of separation between two organisms or their genomes, expressed in various terms such as number of accumulated sequences changes, number of years, or number of generations. The distances are often placed on phylogenetic trees, which show the deduced relationships among the organisms.

Positive selection: The retention of mutations that benefit an organism; also referred to as **Darwinian selection**.

Synteny: The property of being on the same chromosome. Conserved synteny means that genes that are on the same chromosome in one species are also on the same chromosome in the comparison species; these are also referred to as **homology blocks**.

e.g., greater than 1 billion years since their separation. For example, comparing the genomes of yeast, worms, and flies reveals that these eukaryotes encode many of the same proteins, and the nonredundant protein sets of flies and worms are about the same size, being only twice that of yeast (Rubin et al. 2000). The more complex developmental biology of flies and worms is reflected in the greater number of signaling pathways in these two species than in yeast. Over such very large distances, the order of genes and the sequences regulating their expression are generally not conserved. At moderate phylogenetic distances (roughly 70–100 million years of divergence), both functional and nonfunctional DNA is found within the conserved DNA. In these cases, the functional sequences will show a signature of purifying or negative selection, which is that the functional sequences will have changed less than the nonfunctional or neutral DNA (Jukes and Kimura 1984). Not only does comparative genomics aim to discriminate conserved from divergent and functional from nonfunctional DNA, this approach is also contributing to identifying the general functional class of certain DNA segments, such as coding exons, noncoding RNAs, and some gene regulatory regions. Examples of analyses at this distance include comparisons among enteric bacteria (McClelland et al. 2000), among several species of yeast (Cliften et al. 2001, 2003; Kellis et al. 2003), and between mouse and human (International Mouse Genome Sequencing Consortium 2002). The new comparison of the genomes of *Caenorhabditis briggsae* and *Caenorhabditis elegans* (Stein et al. 2003) falls in this category. In contrast, very similar genomes, such as those of humans and chimpanzees (separated by about 5 million years of evolution), are particularly apt for finding the key sequence differences that may account for the differences in the organisms. These are sequence changes under positive selection. Comparative genomics is thus a powerful and burgeoning discipline that becomes more and more informative as genomic sequence data accumulate.

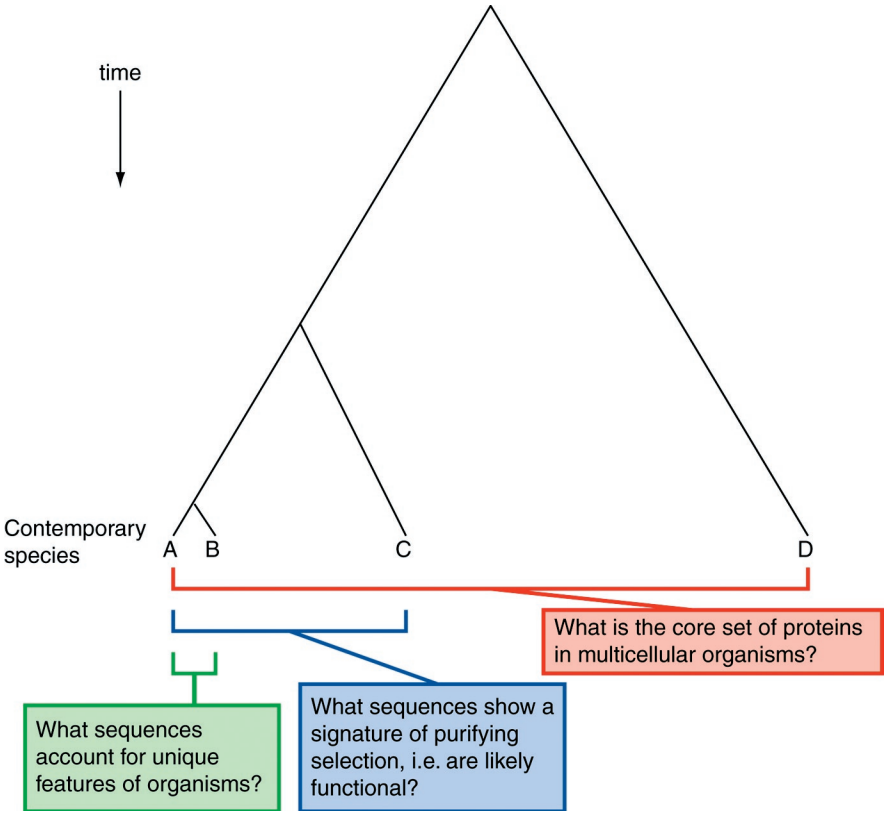
Alignment of DNA sequences is the core process in comparative genomics. An alignment is a mapping of the nucleotides in one sequence onto the nucleotides in the other sequence, with gaps introduced into one or the other sequence to increase the number of positions with matching nucleotides. Several powerful alignment algorithms have been developed to align two or more sequences.

However, the computational power required to align billions of nucleotides between two or more species vastly exceeds what is normally available in individual laboratories. Thus, several research groups make available precomputed alignments of genome sequences through servers or browsers (Table 1). An early example is EnteriX, for enteric bacteria (Florea et al. 2000; McClelland et al. 2000). Aligned human, mouse, and rat genomes can be accessed at several sites, including VISTA (Mayor et al. 2000; Couronne et al. 2003), the conservation tracks at the University of California at Santa Cruz (UCSC) genome browser (Kent et al. 2002), Ensembl (Clamp et al. 2003), and GALA (Giardine et al. 2003).

What Can You Learn about Genome Evolution?

The basic observation in comparative genomics is a description of the matches between genomes. For example, in the roughly 75–80 million years since humans diverged from mouse, the large-scale gene organization and gene order have been preserved (International Mouse Genome Sequencing Consortium 2002). About 90% of the human genome is in large blocks of homology with mouse. These regions of conserved synteny have many genes from one human chromosome that match genes on a mouse chromosome, often in very similar orders.

Sequences with no obvious function, such as relics of transposons that were last active in the common ancestor of human and mouse, can still align in mammalian comparisons; thus, not all the aligning DNA is functional. By evaluating the quality of the alignments genome-wide, the proportion that scores significantly higher than alignments in the ancestral (and presumed nonfunctional) repeats can be determined. This analysis leads to an estimate that about 5% of the human genome is under purifying selection and thus is functional. This portion of the human genome under selection is about three times larger than the portion coding for protein. Within the noncoding sequences under selection, one expects to find noncoding RNA



DOI: 10.1371/journal.pbio.0000058.g001

Figure 1. Comparisons of Genomes at Different Phylogenetic Distances Are Appropriate to Address Different Questions

A generalized phylogenetic tree is shown, leading to four different organisms, with A and D the most distantly related pairs. Examples of the types of questions that can be addressed by comparisons between genomes at the different distances are given in the boxes.

genes, sequences involved in regulation of gene expression, and other critical components of the genome.

Virtually all (99%) of the protein-coding genes in humans align with homologs in mouse, and over 80% are clear 1:1 orthologs. In most cases, the intron-exon structures are highly conserved. This extensive conservation in protein-coding regions may be expected, because many biochemical functions of humans should also be found in mouse. However, it is not seen in all comparisons over an equivalent amount of phylogenetic separation. Only about 60% of the *C. elegans* genes encoding proteins have clear homologs in *C. briggsae* (Stein et al. 2003). The two worms are difficult to distinguish morphologically and probably have similar patterns of development, but they achieve these similarities with some significant differences in the gene sets. Detailed comparisons of the similarities and differences in the relevant genes in these organisms will therefore provide useful insights into developmental processes.

At the nucleotide level, about 40% of the human genome aligns with the mouse genome (International Mouse Genome Sequencing Consortium 2002). The other 60% is composed of at least two classes of sequences, resulting from lineage-specific insertions, deletions, and possibly other mechanisms. One class, occupying about 24% of the genome, is comprised of the repetitive elements that arose by transposition only on the human lineage (International Mouse Genome Sequencing Consortium 2002). These particular insertions did not occur in mice, and thus they cannot align between human and mouse. Likewise, rodent- and mouse-specific retrotransposons independently expanded to occupy about 33% of the mouse genome. The lineage-specific and ancestral repeated elements occupy a substantial portion of the genome of all multicellular organisms, averaging about 50% in mammalian genomes and expanding even higher in the maize genome (San Miguel et al. 1996).

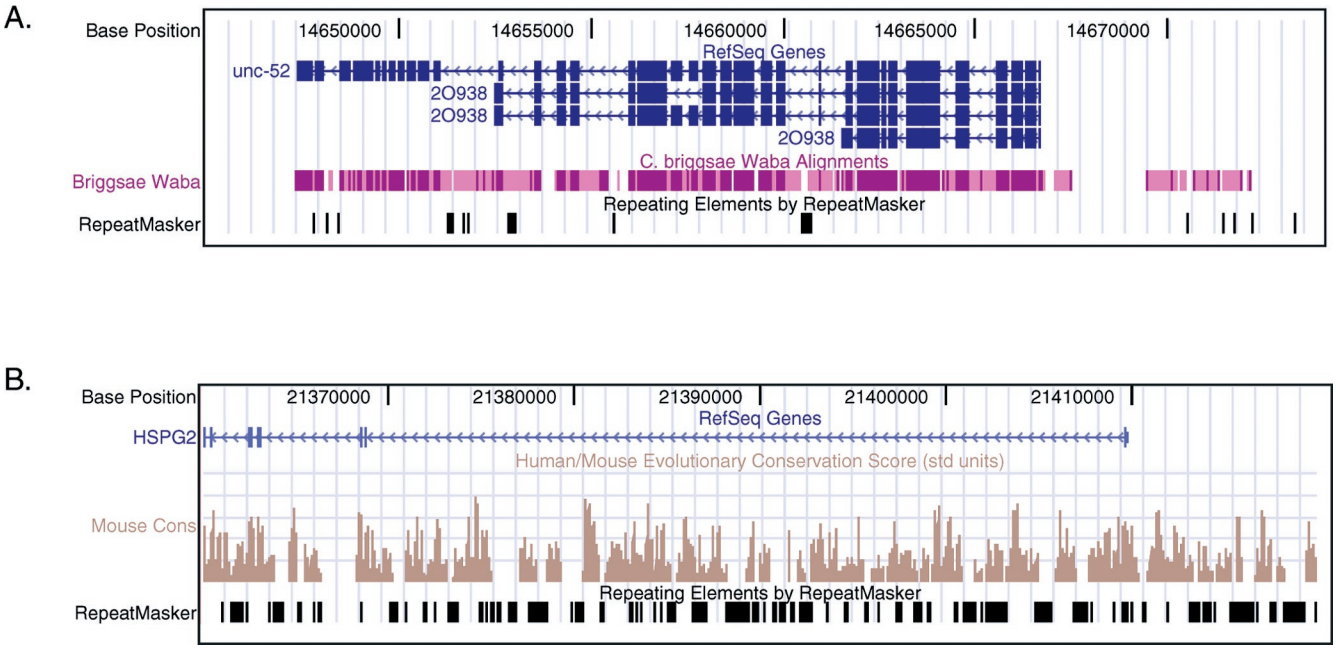
The remaining 36% of the human genome currently cannot be accounted for unambiguously. Some of it could be explained by limitations in the sensitivity of the alignment procedures; i.e., some of the nonaligning DNA could be orthologous DNA that has changed so much that current programs cannot recognize that the sequence has evolved from a common ancestor sequence. However, the homologs to some of the nonaligning DNA in human could be deleted in mouse (International Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003). Given the large expansion of mammalian genomes by transposable elements, one would expect that a compensatory amount of the ancestral DNA would be deleted from the genome. As genome sequences from additional species are determined, the various possible explanations for this nonaligning, nonrepetitive DNA can be tested.

Table 1. URLs for Accessing Precomputed Whole-Genome Alignments and Their Analysis

Server or Browser	Genomes Covered	URL
EnteriX	enteric bacteria	http://bio.cse.psu.edu/
VISTA Genome Browser	human, mouse, rat	http://pipeline.lbl.gov/
UCSC Genome Browser	mammals, worms, zoo[a]	http://genome.ucsc.edu/
Ensembl	mammals, fish, insects, worms	http://www.ensembl.org/
GALA	human, mouse, rat	http://bio.cse.psu.edu/

[a] Data from multiple alignments of 13 vertebrate genome sequences homologous to the human *CFTR* region (Thomas et al. 2003) are included.

DOI: 10.1371/journal.pbio.0000058.t001



DOI: 10.1371/journal.pbio.0000058.g002

Figure 2. Examples of UCSC Genome Browser Views of Genes and Alignments
The unc-52 gene in *C. elegans* (A) and part of its homolog HSPG2 in human (B) are shown, with rectangles for exons and lines for introns; arrows along the introns show the direction of transcription. Both genes encode a chondroitin sulfate proteoglycan. The gene in *C. elegans* is much smaller (about 29 kb) than the gene in humans (about 180 kb; only the 5' portion is shown in [B]). The positions of alignments between *C. elegans* and *C. briggsae* are shown by the purple rectangles in (A). The probability that alignments between human and mouse result from purifying selection are plotted along the Human Cons track in (B). Note that in both comparisons, substantial amounts of intronic and flanking regions align, and several peaks of likely-selected DNA are seen for the human-mouse alignments in the noncoding regions. Among these are candidates for regulatory elements.

Not only do the rates of evolution vary along phylogenetic lineages, but also they are also highly variable within genomes (Wolfe et al. 1989). With the whole-genome alignments between mammals, which encompass many sites that are highly likely to have no function, it is clear that the neutral rate varies significantly in large, megabase-sized regions along mammalian chromosomes. The rates of insertion of certain classes of retrotransposons, inferred large deletions, and meiotic recombination vary along with the neutral substitution rates (International Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003). These observations indicate that large segments of mammalian chromosomes have an inherent tendency to change by any of several processes, such as nucleotide substitution, insertion of transposons, and recombination.

Another major source of change in genomes is segmental duplications, which are particularly prominent in primate genomes (Bailey et al. 2002). These large duplications of tens to thousands of kilobases are revealed by intraspecies comparisons. These regions of genomic instability may play a role in expanding the diversity of the proteins encoded in the genome.

What Can You Learn about Genome Function?

Information on sequence similarity among genomes is a major resource for finding functional regions and for predicting what those functions are. One of the best examples is the improvement in identification of protein-coding genes. Software that incorporates interspecies similarity into gene prediction (Batzoglou et al. 2000; Korf et al. 2001; Wiehe et al. 2001; Alexandersson et al. 2003) is being used to analyze large genomes. Several of the novel genes predicted in mammals using these programs have been verified experimentally, adding about 1,000 new genes to the mammalian set (International Mouse Genome Sequencing Consortium 2002). Comparisons of the worm genomes led to 1,275 well-supported suggestions for new genes in *C. elegans*, adding significantly to the roughly 20,000 known and predicted genes. Predicting and verifying noncoding RNA genes is a current challenge in genomics and bioinformatics (Rivas and Eddy 2001), and it is likely that interspecies comparisons will also help in this analysis.

Regions of noncoding DNA with a particularly high similarity among species have long been recognized as good candidates for functional regions (Hardison et al. 1997; Pennacchio and Rubin 2001), and several have been confirmed as gene regulatory sequences (e.g., Loots et al. 2000). However, the appropriate threshold for the level of sequence similarity that is diagnostic for functional sequences has not been established, and investigators use a variety of such thresholds. What is needed is a robust assessment of the likelihood that a particular alignment results from purifying selection rather than evolutionary drift. The analysis is complicated by the variable rate of neutral evolution within species, but solutions have been developed and are being improved. Comparison of the rates of within-species polymorphism and between-species divergence has proven effective for monitoring selection in nucleotides sequences from *Drosophila* species (Hudson et al. 1987). This method

uses the intraspecies polymorphism measurements as a monitor of neutral evolution, and deviations from neutrality, measured as significantly less interspecies change than expected, are indicators of selection. For the human-mouse genome comparisons, the local neutral rate was estimated from the divergence of aligned ancestral repeats, and similarity scores were adjusted accordingly. By evaluating the distribution of these similarity scores in likely-neutral DNA and in DNA inferred as being under selection, a probability that any human-mouse alignment reflects purifying selection can be computed (Figure 2), and such scores are available genome-wide on the UCSC Genome Browser.

Predicting exactly what the function is of these noncoding sequences under selection is a major challenge. One promising approach is to collect good training sets of alignments within sequences of known functions, such as gene regulatory sequences, and use those alignments to develop statistical models for estimating a likelihood that any given alignment could be generated by that model (e.g., Elnitski et al. 2003). This type of approach could be applied to any functional category in which the conserved DNA sequence is critical to the function. For instance, it is still not clear whether conserved DNA sequences are critical to the function of replication origins; if they are not, then this analytical model will not successfully predict this important functional category, and other methods will need to be developed.

Prospects

The past year has brought the genome sequences of species that are close relatives of many model organisms. The list includes several yeast species to compare with *Saccharomyces* cerevisiae, another *Drosophila* species and *Anopheles* (Holt et al. 2002) to compare with *Drosophila melanogaster*, mouse to compare with human, and now *C. briggsae* to compare with *C. elegans* (Stein et al. 2003). Fully harvesting the information in these comparative analyses and integrating it across the many comparisons will be a continuing and fruitful exercise.

Comparing more than two genomic sequences provides even more resolving power. The efficacy of multiple sequences for functional prediction is shown dramatically by the analyses of 13 genomic sequences from species ranging from fish to humans (Thomas et al. 2003). Other approaches using multiple sequences from more closely related species substantially improve the resolving power of comparative genomics (Gumucio et al. 1992; Boffelli et al. 2003). The Human Genome Project recognizes the power of this broad comparative analysis (Collins et al. 2003). Researchers may reasonably expect in the near future to have results of this comparative analysis readily available. By calibrating these results, such as estimated likelihoods of being under selection, likelihood of being a coding exon, etc., against known functional elements, the power of the comparative approaches should improve. The critical next stage is large-scale experimental tests of the predictions, which should prove exciting and challenging. ■

References

Alexandersson M, Cawley S, Pachter L (2003) SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res 13: 496–502.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007.

Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES (2000) Human and mouse gene structure: Comparative analysis and application to exon prediction. Genome Res 10: 950–958.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391–1394.

Clamp M, Andrews D, Barker D, Bevan P, Cameron G, et al. (2003) Ensembl 2002: Accommodating comparative genomics. Nucleic Acids Res 31: 38–42.

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, et al. (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. Genome Res 11: 1175–1186.

Cliften PF, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71–76.

Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422: 835–847.

Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, et al. (2003) Strategies and tools for whole-genome alignments. Genome Res 13: 73–80.

Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. (2003) Distinguishing regulatory DNA from neutral sites. Genome Res 13: 64–72.

Florea L, Riemer C, Schwartz S, Zhang Z, Stojanovic N, et al. (2000) Web-based visualization tools for bacterial genome alignments. Nucleic Acids Res 28: 3486–3496.

Giardine BM, Elnitski L, Riemer C, Makalowska I, Schwartz S, et al. (2003) GALA: A database for genomic sequence alignments and annotations. Genome Res 13: 732–741.

Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, et al. (1992) Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. Mol Cell Biol 12: 4919–4929.

Hardison R, Oeltjen J, Miller W (1997) Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. Genome Res 7: 959–966.

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, et al. (2003) Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. Genome Res 13: 13–26.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129–149.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.

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

Jukes TH, Kimura M (1984) Evolutionary constraints and the neutral theory. J Mol Evol 21: 90–92.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.

Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17 Suppl 1: S140–S148.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, et al. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288: 136–140.

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, et al. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046–1047.

McClelland M, Florea L, Sanderson K, Clifton SW, Parkhill J, et al. (2000) Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, *Typhimurium*, *Typhi* and *Paratyphi*. Nucleic Acids Res 28: 4974–4986.

Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2: 100–109.

Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2: 8.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. (2000) Comparative genomics of the eukaryotes. Science 287: 2204–2215.

San Miguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765–768.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. (2003) The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. PLoS Biol 1: DOI: 10.1371/journal.pbio.0000044.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788–793.

Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R (2001) SGP-1: Prediction and validation of homologous genes based on sequence alignments. Genome Res 11: 1574–1583.

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337: 283–285.