

Vertebrate genome sequencing: building a backbone for comparative genomics

James W. Thomas and Jeffrey W. Touchman

The human genome sequence provides a reference point from which we can compare ourselves with other organisms. Interspecies comparison is a powerful tool for inferring function from genomic sequence and could ultimately lead to the discovery of what makes humans unique. To date, most comparative sequencing has focused on pair-wise comparisons between human and a limited number of other vertebrates, such as mouse. Targeted approaches now exist for mapping and sequencing vertebrate bacterial artificial chromosomes (BACs) from numerous species, allowing rapid and detailed molecular and phylogenetic investigation of multi-megabase loci. Such targeted sequencing is complementary to current whole-genome sequencing projects, and would benefit greatly from the creation of BAC libraries from a diverse range of vertebrates.

In the midst of the excitement surrounding the completion of a draft sequence of the human genome [1,2] and the steady progression of other vertebrate sequencing projects, comparative sequencing is justly gaining the attention of the genetics community. The comparison of genomic sequence of two or more species has the extraordinary power to highlight evolutionary changes that have shaped genome structure and content, and to reveal specific sequences that have been conserved throughout evolution. Vertebrate comparative sequencing, in particular, can be very useful for confirming functional annotations or computational gene-finding results and for identifying novel genes in the human genome [2,3]. It also affords the unique potential to identify conserved sequences that reside outside of coding regions that could control gene expression [4,5], and possibly those elements involved in gene imprinting, chromosome packaging and chromosome pairing.

Although some putative regulatory elements identified by interspecies sequence comparisons have been confirmed experimentally [6–8], we are only just beginning to identify and characterize these

sequences. As additional genomic sequence data is generated from other vertebrates, thousands of potentially functional elements might be identified by computational methods. However, determining whether identical sequences have been actively conserved over time, or whether they are identical simply because of shared common ancestry, is a major challenge facing investigators. Additionally, different rates of evolutionary change observed between pairs of orthologous sequences can confound interpretation even further [9]. It is increasingly clear, however, that using sequence from multiple species in a single comparison adds significant power to the identification of functionally conserved sequences [6,10,11].

There are currently six vertebrate organisms whose genomes are undergoing systematic sequencing: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), and two pufferfish species (*Fugu rubripes* and *Tetraodon nigroviridis*) (Table 1). These species have been selected for sequencing primarily for their intrinsic value, although the two pufferfish were chosen largely because of their comparative value to the human genome [12,13]. The evolutionary relationships of these vertebrates are illustrated in Fig. 1, which emphasizes the very limited range of evolutionary distances among the vertebrates currently being

sequenced. There are large portions of vertebrate evolutionary history with virtually no genomic sequence data (for instance, there are no marsupials, reptiles or amphibians represented). The trio of placental mammals (human, mouse and rat) are important, but could have limited comparative value by themselves [11]. In the future, sequencing vertebrates from interspersed evolutionary time points may be necessary to understand fully the evolution of genomes and the relevance (if any) of conserved orthologous sequences.

There is no shortage of candidate species for systematic genomic sequencing. The decisions facing scientists and funding agencies about how and where to proceed beyond organisms that have already been selected are complex. Addressing this issue, objective criteria to prioritize future mammalian sequencing targets have been proposed [14], many of which can be applied to vertebrates as well. These criteria include (but are not limited to) the overall cost of a project, the relevance to human health, the size of the corresponding research community, the suitability of an organism for experimentation, the presence of existing resources, and whether the results will increase our basic knowledge substantially across many scientific disciplines. To address and prioritize these criteria, a peer review system is being planned in the USA to evaluate individual

Table 1. Vertebrate genome sequencing, as of December 2001

Species (size in millions of base pairs)	Relevant World Wide Web sites
Human (~3200)	http://www.nghri.nih.gov/genome_hub.html http://www.ensembl.org
Human (~3200)	http://www.celera.com
Mouse (~3000)	http://www.informatics.jax.org http://www.nih.gov/science/models/mouse http://www.ncbi.nlm.nih.gov/genome/seq/MmHome.html
Rat (~3000)	http://www.hgsc.bcm.tmc.edu/rat http://www.celera.com http://www.nih.gov/science/models/rat http://www.rgd.mcv.edu
Zebrafish (~1700)	http://www.sanger.ac.uk/Projects/D_rerio http://zfinfo.org/ZFIN
Pufferfish (<i>Fugu rubripes</i>) (~400)	http://www.jgi.doe.gov/programs/fugu.htm
Pufferfish (<i>Tetraodon nigroviridis</i>) (~400)	http://www.genoscope.cns.fr/externe/tetraodon

proposals [15]. However, in addition to these mainly intrinsic criteria, additional criteria that assess the importance of a genome specifically for its comparative value might be considered. For instance, comparing the human genome to primates could provide unique genotype-to-phenotype correlations, whereas comparisons to more distantly related species might allow the identification of functional sequence elements conserved over hundreds of millions of years. Such reciprocal studies lend unique information to the analysis of all species in a comparison.

Whole-genome comparative sequencing

Perhaps as important as selecting an organism to sequence is the decision regarding the sequencing strategy. The strategy will influence the cost of the project and the quality of the resulting sequence. The most common approaches to systematic sequencing of vertebrate genomes are shotgun sequencing of mapped bacterial artificial chromosome (BAC) clones ('clone-by-clone') and whole-genome shotgun (WGS) sequencing [16]. Both strategies have their strengths for comparative sequence analysis. WGS sequencing provides rapid and global sequence information that is immediately useful in comparisons with other finished genomes [2,3]. The clone-by-clone approach, with its associated clone map, allows the generation of local high-quality sequence and has the potential to sort out the intrinsic complexities of vertebrate genomes, such as high-copy repetitive sequences and segmental duplications [17].

In either case, the effort of sequencing vertebrate genomes is great, requiring a significant investment of resources. Even WGS projects necessitate a commitment of many millions of sequence reads to approach a useful sampling of the target genome [18]. As such, there will be a limited number of whole-genome sequences available for comparative analyses, at least for the foreseeable future.

Targeted comparative sequencing

As an alternative to generating global genomic sequence from a vertebrate genome, there are robust clone-by-clone methods available for more focused studies of orthologous regions of the human genome. In fact, the availability of a complete draft human sequence creates new opportunities for the comparative mapping and sequencing of any other

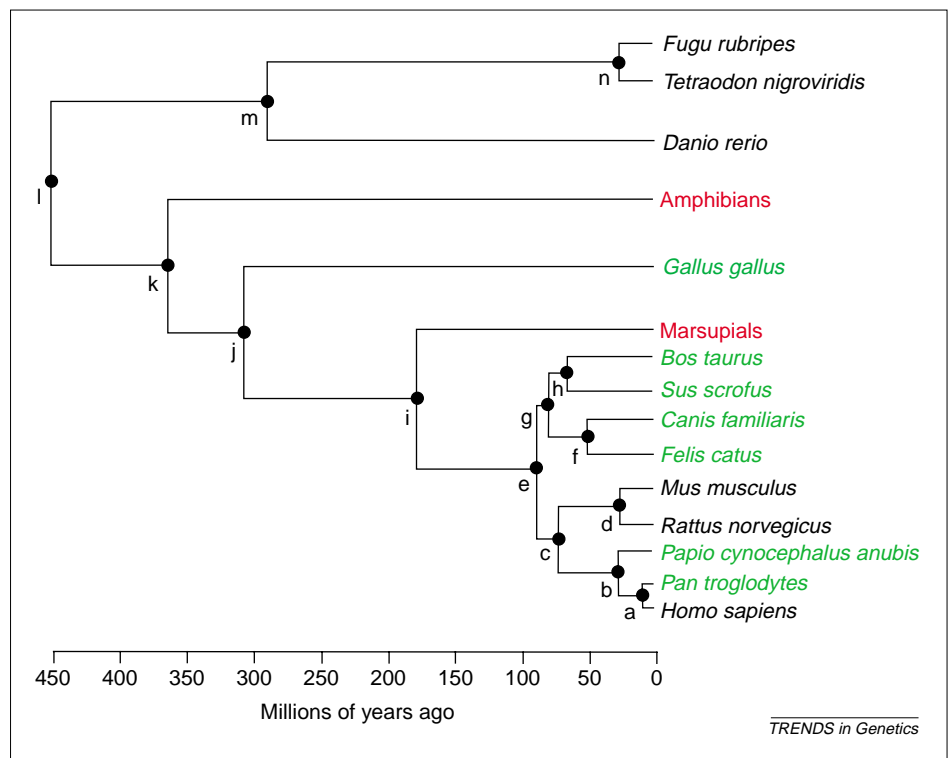


Fig. 1. Evolutionary relationships of select vertebrates. The evolutionary relationships among a set of vertebrates including those selected for whole-genome sequencing (black; *Fugu rubripes*, *Tetraodon nigroviridis*, *Danio rerio*, *Mus musculus*, *Rattus norvegicus* and *Homo sapiens*); two representative groups (amphibians and marsupials) for which virtually no genomic sequence is currently available (red) and those included in a comparative sequencing project at the US National Institutes of Health (green; *Gallus gallus*, *Sus scrofa*, *Bos taurus*, *Canis familiaris*, *Felis catus*, *Papio cynocephalus anubis* and *Pan troglodytes*). The phylogenetic tree was constructed using data from several sources [28,31-33]. Estimated times of divergence from the last common ancestor were based on the following: a, 6 million years (Myr) [33]; b, 25 Myr [33]; c, 64-74 Myr [34]; d, 14-41 Myr [35,36]; e, 68-102 Myr [34]; f, 40-52 Myr [36]; g, 76-96 Myr [34]; h, 62-67 Myr [36]; i, 161-185 Myr [36]; j, 310 Myr [36]; k, 356-374 Myr [36]; l, 415-485 Myr [36]; m, 256-312 Myr [32]; n, 18-30 Myr [37].

vertebrate. This is made possible by the local conservation of genome organization among vertebrates, as revealed by the observed similarities in gene order along the chromosomes of different species [19].

Specifically, recent studies demonstrate that human genomic sequence greatly facilitates the targeted comparative mapping and sequencing of other species [20,21]. As illustrated in Fig. 2, human genomic sequence can be used to identify regions of orthologous sequence by searches of various sequence databases, including those containing expressed sequence tags (ESTs), known genes, BAC end sequences, or WGS sequence from another species, such as mouse. The identified orthologous sequences can then be used to design small 'overgo' DNA probes [22] for screening a BAC library spotted in high density on nylon filters. The identified clones can be ordered on the basis of probe content and restriction enzyme fingerprints, and a set of overlapping clones selected and sequenced. The resulting sequence from these clones provides the basis for comparative sequence

analysis of the targeted region. Already, new insights into chromosome evolution have been produced using this method [23], and because this approach is applicable to any vertebrate with a partial sequence database, it offers a potentially powerful and practical means for targeted comparative sequencing in many other species.

Computational tools for multi-species sequence comparisons

Alignment of orthologous sequences defines the basis of studying molecular evolution and inferring phylogenetic relationships. For short sequences that are highly conserved and co-linear, common sequence alignment tools (e.g. Clustal X [24]) are sufficient for informative sequence comparisons. However, as the length and the divergence between sequences increase, single definitive alignments become more difficult to generate and harder to visualize. Several graphical computer tools are now available for interspecies sequence alignments of long genomic regions (Table 2).

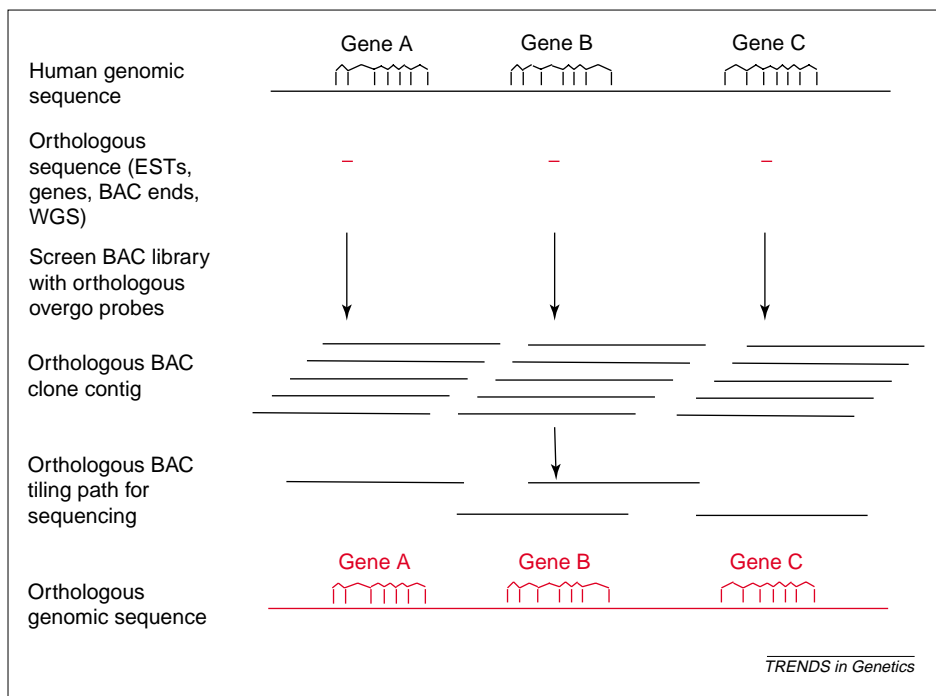


Fig. 2. Targeted comparative mapping and sequencing. Ordered and orientated human genomic sequence from a particular region of interest is used to search a sequence database from a second species for orthologous sequences. The comparative species sequence database can include, but is not restricted to, genes, expressed sequence tags (ESTs), bacterial artificial chromosome (BAC) ends, or whole-genome shotgun (WGS) sequences. Once orthologous sequences are identified, 'overgo' probes can be designed from this sequence and used to screen a BAC library from the comparative species to isolate the entire orthologous genomic region. A set of overlapping BAC clones can then be selected as the sequencing substrate in the comparative species.

Figure 3 shows an example of the graphical output from MultiPipMaker generated by aligning human genomic sequence containing a portion of the *CFTR* gene to orthologous BAC-derived genomic sequence from baboon (*Papio cynocephalus anubis*), cow (*Bos taurus*), mouse and fugu. The pairwise alignment of human and baboon sequence in this 20-kilobase interval illustrates a high degree (~94%) of sequence identity. This sequence conservation includes both unique sequence and repetitive elements, such as the primate specific LINE-1 element, L1PB2 [25], in intron 11. However, there are also significant differences in this region between human and baboon. For example, in intron 15 there is an endogenous retrovirus (ERV) pTR5 [26,27] that is present in human, but apparently absent in baboon. It could therefore be proposed that the integration of this repetitive element occurred in the ancestral human lineage after the split from the lineage leading to baboon. Sequencing more primates could elucidate the exact origin of this repetitive element, and other such differences between these two species.

The sequence comparison between human and cow reveals a much lower, but still extensive, amount of sequence

conservation outside of the exons (which are notably conserved with human). Specifically, there is generally a uniform pattern of conservation across this region that is only interrupted by the two primate repetitive elements mentioned above. Comparison between human and mouse shows a further reduction in sequence conservation. Although recent phylogenetic analysis of placental mammals places human and mouse with a more-recent common ancestor than human and cow [28], the lesser degree of sequence conservation seen between human and mouse is consistent with the higher mutation rate in the rodent lineage owing to an increased number of generations, referred to as the 'generation-time effect hypothesis' [29].

Comparison of the alignments using mouse BAC sequence and unassembled mouse WGS sequences aligned to this region (<http://www.ensembl.org>) highlight the important but limited utility of detecting orthologous sequences using individual WGS reads. Both sources of mouse sequence are able to reveal the conservation of exons 12–15 of *CFTR*, yet the BAC-derived data reveals more aligned segments in the introns of the gene. The decreased sequence conservation seen

Table 2. Graphical computational tools for comparative sequence analysis

Program	World Wide Web site	Refs
PipMaker	http://bio.cse.psu.edu/	[40]
MultiPipMaker	pipmaker	
VISTA	http://sichuan.lbl.gov/vista/	[41]
Alfresco	http://www.sanger.ac.uk/Software/Alfresco/	[42]
SynPlot	http://www.sanger.ac.uk/Users/jjrg/SynPlot/	[6]
GLASS	http://plover.lcs.mit.edu/	[43]

between human and mouse using WGS reads as opposed to complete BAC sequence is probably due to the fragmentary state of the unassembled mouse WGS data and limited computational power to identify poorly conserved orthologous sequences in a whole-genome comparison. Finally, pairwise-sequence comparison of human and fugu genomic sequence shows very little sequence conservation between these species. Indeed, some of the exons are not even conserved to a high enough degree to be aligned under these conditions.

From this set of interspecies sequence comparisons to the human genome, it is clear that the information that can be extrapolated from such analyses is dependent on the sequence divergence between the species in the comparison. In the human–baboon comparison, the most interesting and important use of the alignments would be to detect the differences between these two closely related species to begin to understand the molecular basis for the differences in phenotypes between these two primates. In comparisons between distantly related species, such as human and fugu, the primary interest is in all regions that have been actively conserved over hundreds of millions of years. Less clear is the interpretation of interspecies alignments between species of intermediate divergence, such as human and mouse or human and ox. This is because some fraction of the conserved sequences are being actively selected for, and are thus of potential functional significance, whereas the rest of the conserved sequences are detected simply because of a lack of sufficient divergence from an ancestral sequence. Thus, with respect to interspecies sequence comparison, two major challenges present themselves: determining what species are most informative for a given comparative study, and effectively evaluating sequence alignments to infer genuine function from background 'noise'. These challenges might only be met by

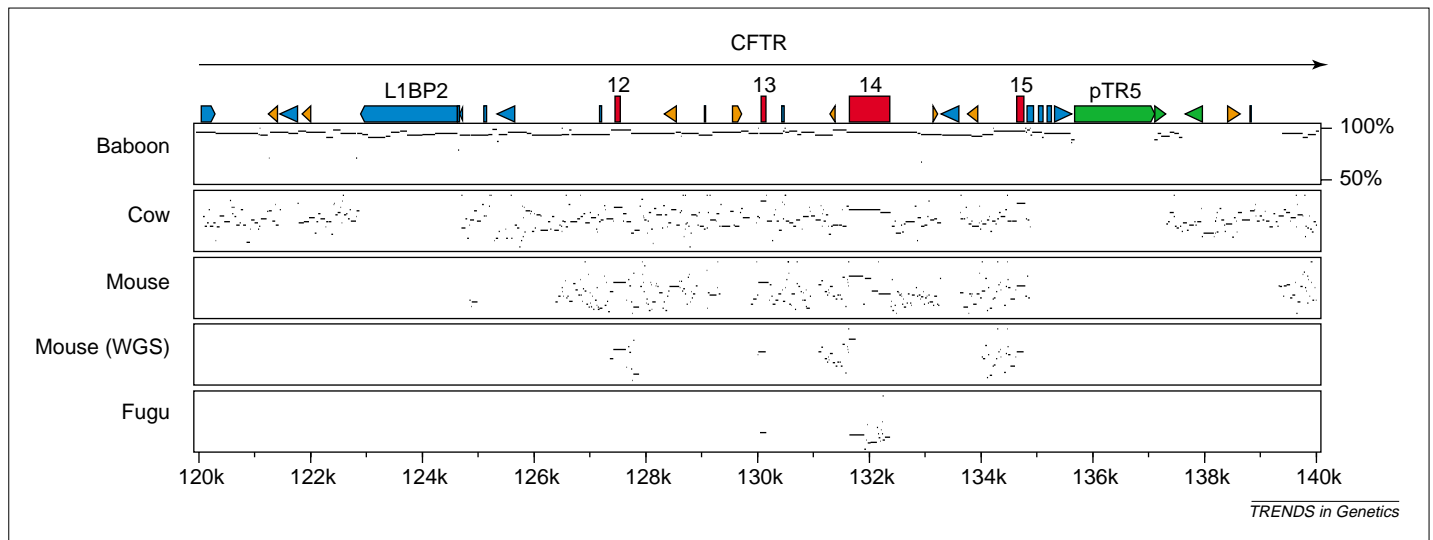


Fig. 3. Comparison of human genomic sequence to multiple species. Human genomic sequence containing the *CFTR* gene (GenBank no. AC000111) was compared with baboon (GenBank no. AC091381), cow (GenBank no. AC089993), mouse (GenBank no. AF162137 [38]) and fugu (GenBank no. AJ271361 [39]). BAC-derived genomic sequence using MultiPipMaker (<http://bio.cse.psu.edu/pipmaker/>). In addition, mouse whole-genome shotgun (WGS) sequence reads (~threefold sequence coverage) aligned to this region were retrieved from Ensembl (<http://www.ensembl.org/>) and re-aligned to the human genomic sequence with MultiPipMaker. Exons 12–15 of *CFTR* are indicated by red boxes with the corresponding exon number listed above. Repetitive elements are indicated by a series of additional symbols along the top. The L1BP2 and pTR5 repeats are indicated. For each species, pairwise alignments with human are graphically displayed by plotting the length and percent identity of each ungapped alignment between 50 and 100% identity. The numbers at the bottom indicate the scale in kilobases along the human reference sequence.

adding genomic sequence from several other organisms to the analysis.

Future resources

Currently at the NIH Intramural Sequencing Center (<http://www.nisc.nih.gov>), a comparative sequencing project funded by the National Human Genome Research Institute is underway to establish a dataset of genomic sequence from multiple vertebrates. This is being accomplished by the mapping and sequencing of orthologous genomic intervals in 11 vertebrates that correspond to six regions on human chromosome 7. Together, these regions include more than 15 megabases of the mammalian genome that were previously characterized in large-scale mapping efforts in human and mouse [20,30]. These regions were also chosen because they have a range of gene and repeat content, G+C content, and known evolutionary breakpoints. The current list of species being mapped and sequenced are indicated in Fig. 1, including eight mammals (chimpanzee [*Pan troglodytes*], baboon, cow, pig [*Sus scrofa*], dog [*Canis familiaris*], cat [*Felis catus*], mouse and rat), one bird (chicken [*Gallus gallus*]) and two fish (fugu and zebrafish). These data, in addition to being a unique and important resource for evolutionary biology, will immediately allow the

evaluation of existing computational tools for multi-species sequence comparisons, provide a basis for the design of new computational tools, and help guide decisions about future sequencing efforts. However, the scope of this project and others like it is limited at this time by the significant lack of available BAC libraries from a broad range of organisms. In fact, the only criterion for choosing the 11 vertebrate species in this pilot project was the availability of a BAC library. Fortunately, both the National Science Foundation and the National Institutes of Health are initiating efforts to generate larger collections of BAC libraries (<http://www.nsf.gov/pubs/2001/nsf01145.html>; <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-01-002.html>). Because BAC libraries can support both individual gene-cloning projects and whole-genome sequencing efforts, these proposed plans will benefit individual labs as well as genome centers and allow greater access to studies of the evolutionary history of the vertebrates.

Comparative mapping and sequencing is rapidly becoming an essential component in the analysis of a genome, adding depth of knowledge about all species used in any given comparison. Modern methods of genomics have opened myriad opportunities for the sequencing of

other genomes, both in whole-genome and targeted fashions, and together these are being used to build a backbone for future studies in comparative genomics. By comparing the genomes of other species to ourselves, we are better able to understand how the overall structures of genes and genomes have evolved and how they function today, thereby providing a unique perspective into the human genetic blueprint.

Acknowledgements

We thank Robert Blakesley and Eric Green for critical review of the manuscript.

James W. Thomas

Jeffrey W. Touchman*

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.

*e-mail: jefft@nhgri.nih.gov

References

- 1 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- 2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
- 3 Roest, C.H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238
- 4 Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16, 369–372
- 5 Pennacchio, L.A. and Rubin, E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2, 100–109
- 6 Gottgens, B. *et al.* (2000) Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* 18, 181–186
- 7 Loots, G.G. *et al.* (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288, 136–140
- 8 Touchman, J.W. *et al.* (2001) Human and mouse alpha-synuclein genes: comparative genomic

- sequence analysis and identification of a novel gene regulatory element. *Genome Res.* 11, 78–86
- 9 Johnson, M.E. *et al.* (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413, 514–519
 - 10 Dubchak, I. *et al.* (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* 10, 1304–1306
 - 11 Frazer, K.A. *et al.* (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* 11, 1651–1659
 - 12 Brenner, S. *et al.* (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 366, 265–268
 - 13 Crollius, H.R. *et al.* (2000) Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* 10, 939–949
 - 14 O'Brien, S.J. *et al.* (2001) Genomics. On choosing mammalian genomes for sequencing. *Science* 292, 2264–2266
 - 15 Gewolb, J. (2001) Genomics: animals line up to be sequenced. *Science* 293, 4409–4410
 - 16 Green, E.D. (2001) Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.* 2, 573–583
 - 17 Eichler, E.E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* 17, 661–669
 - 18 Bouck, J.B. *et al.* (2000) Shotgun sample sequence comparisons between mouse and human genomes. *Nat. Genet.* 25, 31–33
 - 19 O'Brien, S.J. *et al.* (1999) The promise of comparative genomics in mammals. *Science* 286, 458–481
 - 20 Thomas, J.W. *et al.* (2000) Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Res.* 10, 624–633
 - 21 Kim, J. *et al.* (2001) Homology-driven assembly of a sequence-ready mouse BAC contig map spanning regions related to the 46-Mb gene-rich euchromatic segments of human chromosome 19. *Genomics* 74, 129–141
 - 22 Vollrath, D. (1999) DNA markers for physical mapping. In *Mapping Genomes* (Birren, B. *et al.*, eds), pp. 187–215, Cold Spring Harbor Laboratory Press
 - 23 Dehal, P. *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 293, 104–111
 - 24 Jeanmougin, F. *et al.* (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403–405
 - 25 Smit, A.F. *et al.* (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* 246, 401–417
 - 26 La Mantia, G. *et al.* (1989) Identification of new human repetitive sequences: characterization of the corresponding cDNAs and their expression in embryonal carcinoma cells. *Nucleic Acids Res.* 17, 5913–5922
 - 27 Costas, J. and Naveira, H. (2000) Evolutionary history of the human endogenous retrovirus family ERV9. *Mol. Biol. Evol.* 17, 320–330
 - 28 Murphy, W.J. *et al.* (2001) Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614–618
 - 29 Ohta, T. (1993) An examination of the generation-time effect on molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 90, 10676–10680
 - 30 Bouffard, G.G. *et al.* (1997) A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* 7, 673–692
 - 31 Venkatesh, B. *et al.* (2001) Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* 98, 11382–11387
 - 32 Kumazawa, Y. *et al.* (1999) Mitochondrial molecular clocks and the origin of euteleostean biodiversity: familial radiation of perciforms may have predated the cretaceous/tertiary boundary. In *The Biology of Biodiversity* (Kato, M. *et al.*, eds), pp. 35–52, Springer-Verlag
 - 33 Goodman, M. *et al.* (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* 9, 585–598
 - 34 Eizirik, E. *et al.* (2001) Molecular dating and biogeography of the early placental mammal radiation. *J. Hered.* 92, 212–219
 - 35 Jacobs, L.L. and Pilbeam, D. (1980) Of mice and men: fossil-based divergence dates and molecular 'clocks'. *J. Hum. Evol.* 9, 551–555
 - 36 Kumar, S. and Hedges, S.B. (1998) A molecular timescale for vertebrate evolution. *Nature* 392, 917–920
 - 37 Crnogorac-Jurcevic, T. *et al.* (1997) *Tetraodon fluviatilis*, a new puffer fish model for genome studies. *Genomics* 41, 177–184
 - 38 Ellsworth, R.E. *et al.* (2000) Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1172–1177
 - 39 Davidson, H. *et al.* (2000) Genomic sequence analysis of *Fugu rubripes* CFTR and flanking genes in a 60 kb region conserving synteny with 800 kb of human chromosome 7. *Genome Res.* 10, 1194–1203
 - 40 Schwartz, S. *et al.* (2000) PipMaker – a web server for aligning two genomic DNA sequences. *Genome Res.* 10, 577–586
 - 41 Mayor, C. *et al.* (2000) VISTA: VISualization Tool for Alignments. *Bioinformatics* 16, 1046–1047
 - 42 Jareborg, N. and Durbin, R. (2000) Alfresco – a workbench for comparative genomic sequence analysis. *Genome Res.* 10, 1148–1157
 - 43 Batzoglou, S. *et al.* (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* 10, 950–958

Electronic tools to manage gene expression data

Dale A. Begley and Martin Ringwald

The scientific community is generating an ever-growing amount of gene expression data of an increasingly diverse and complex nature. This proliferation of scientific results raises questions of how to manage, integrate and analyze these data. Even within an individual laboratory it is becoming difficult to manage the data produced. The Gene Expression Database and the Gene Expression Notebook are publicly available electronic tools developed to address these problems.

The laboratory mouse is an important animal model in biomedical research. It is closely related to humans, and tissues from many strains and mutants are available that allow detailed and diverse gene expression studies. To assist researchers in coping with these data we have developed the Gene Expression Database (GXD) as a community resource [1–4]. We have also

developed the Gene Expression Notebook (GEN) as a laboratory management tool. See Box 1 for information on how to access GXD and obtain GEN.

The Gene Expression Database (GXD)

GXD is a community resource designed to provide integrated access to different types of expression data and to place data in the context of other biological information. GXD is integrated with the Mouse Genome Database (MGD) [5] to foster close links to genotype and phenotype data, as well as to provide interconnections with sequence databases and databases for other species. Genes are classified in conjunction with the Gene Ontology project [6–8], according to biological processes, molecular functions and cellular components, providing additional search parameters.

The focus of GXD is on endogenous gene expression information, and it holds data from both wild-type and mutant laboratory mice, including mutants generated by homologous recombination. In future, we plan to include data from knock-in experiments. However, expression information from other types of transgenic embryos is not included in GXD, because it might not reflect the endogenous expression patterns due to position effects or multiple insertions. Data from different assay types are described in detail using controlled vocabularies of standardized terms. Expression patterns are described using an extensive anatomical dictionary developed in conjunction with the Mouse Atlas and Gene Expression Database Project in Edinburgh [9,10]. This allows comparison of data from multiple assay types that have different spatial