The mouse that roared

Mark S. Boguski

The laboratory mouse has become an indispensable tool for investigators in many areas of biomedical research. The availability of the full mouse genome sequence will immeasurably advance both the character and the pace of discovery.

"Know then thyself, presume not God to scan; The proper study of mankind is man," wrote Alexander Pope in 1733. What better reason could there have been to sequence the human genome? But the planners of the Human Genome Project realized that the data could not be fully understood, or used to advance biomedicine, in isolation. Indeed, many of the "lessons learned and promises kept"[1] have been derived from the study of model organisms. *Mus musculus*, a species of mouse, has been one of the five key model organisms sequenced since the beginnings of the Human Genome Project. In 1998–99 the US National Institutes of Health published an action plan for mouse genomics[2] which, among other things, called for a working draft sequence of the mouse genome by 2003. An international Mouse Genome Sequencing Consortium[3] has now achieved this goal. The particular mouse strain concerned is C57BL/6J (Fig. 1), and the consortium's findings are reported on page 520 of this issue.

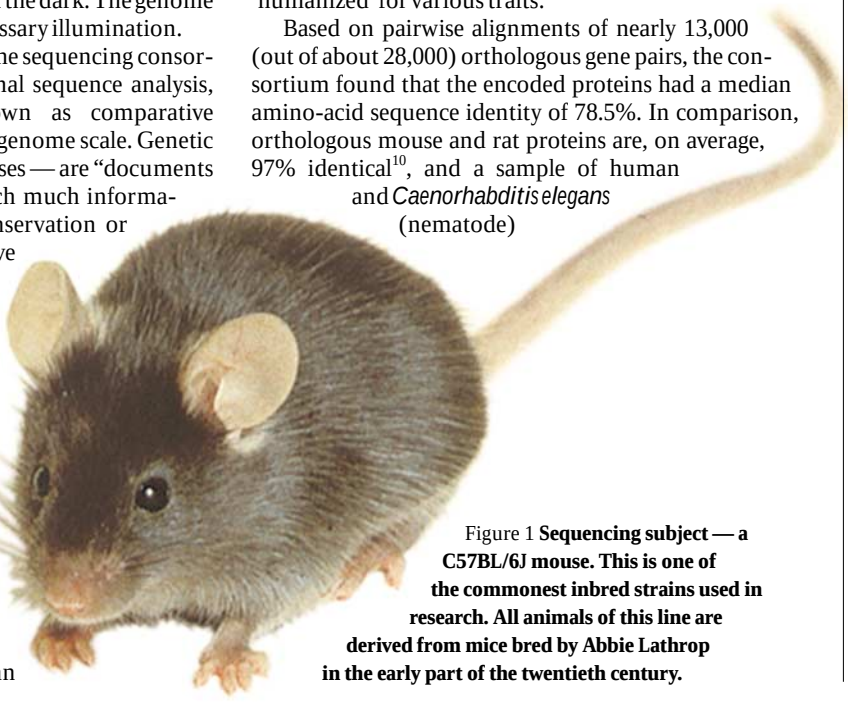
Why is this so important? It is because there can scarcely be a major area of mammalian biology or medicine to which mouse studies have not contributed in some way, often as surrogates for human studies. For genetics and development, for immunology and pharmacology, for cancer and heart disease, even for behaviour, learning and memory and psychiatric disorders[4,5], the laboratory mouse has become an indispensable tool. Much of this power has come from technologies to manipulate the mouse genome, but until now we have in effect been shooting in the dark. The genome of *Mus musculus* will provide the necessary illumination.

Most of the findings reported by the sequencing consortium[3] were revealed by computational sequence analysis, the overall approach being known as comparative genomics[6–8] when performed on the genome scale. Genetic sequences — strings of nucleotide bases — are "documents of evolutionary history"[9], from which much information can be inferred from their conservation or divergence and rearrangement relative to a common ancestor. Genome analysts have applied this approach at many levels, from multi-megabase rearrangements reflected in chromosome structure down to single nucleotide changes between orthologous genes — two genes are orthologous if they diverged after a speciation event, when a new species forms from an existing one; two genes are paralogous if they diverged after a gene duplication event. One of the outcomes of a comparative genomic analysis is an enumeration of the total number of genes shared by two species (see mammalian gene count, below).

An unexpected finding to emerge from the sequence of the human genome is that we appear to have far fewer protein-coding genes than previously suspected — fewer than 30,000, compared with the figure of 80,000–100,000 frequently cited in textbooks published before 2001. Analysis of the mouse genome backs up this finding. The sequencing consortium estimates that it contains 27,000–30,500 protein-coding genes. Ninety-nine per cent of these genes have a sequence match in the human genome and 96% of these lie within 'syntenic' regions of mouse and human chromosomes.

The consortium was able to align long segments of mouse and human chromosomes in which the linear orders of genes in the common ancestral sequences were conserved. Although this so-called 'conservation of synteny' between mouse and human chromosomes has long been recognized, the comprehensiveness and precision afforded by the genome sequences will allow effective cross-reference of the locations of any genetically mapped traits in the mouse with genes in the orthologous regions of the human genome (and vice versa). This will greatly accelerate the isolation of disease genes. It will also be important for precise deletion (knockout) of mouse genes to study their functions and for targeting human sequences to their syntenic locations in the mouse genome, allowing the mice to be 'humanized' for various traits.

Based on pairwise alignments of nearly 13,000 (out of about 28,000) orthologous gene pairs, the consortium found that the encoded proteins had a median amino-acid sequence identity of 78.5%. In comparison, orthologous mouse and rat proteins are, on average, 97% identical[10], and a sample of human and *Caenorhabditis elegans* (nematode)



JACKSON LABORATORY

Figure 1 **Sequencing subject — a C57BL/6J mouse. This is one of the commonest inbred strains used in research. All animals of this line are derived from mice bred by Abbie Lathrop in the early part of the twentieth century.**

proteins had an average of 49% of their amino acids in common[11]. For many of the unaligned human and mouse genes[3], it was not possible to confidently identify one-to-one orthologous pairs because of differential gene expansion — contraction or loss in various gene families that results in one-to-many or even many-to-many pairings. It is undoubtedly in the divergent evolution of these gene families that some of the differences between humans and rodents will be found.

There is a highly practical aspect to this kind of study. For example, the consortium describes the expansion of four subfamilies of cytochrome P450 (CYP) genes in mice compared with humans. These genes encode enzymes that function in drug metabolism, and biologically are part of a general xenobiotic 'sensory system' for chemical compounds in the environment. In the pharmaceutical industry, one stage of drug development is known as 'ADME/Tox' — shorthand for studies of the absorption, distribution, metabolism, excretion and toxicology of drug candidates. Mice and humans have been exposed to different types and amounts of environmental substances (including toxins) during their evolutionary divergence, and one might expect natural selection to have altered the numbers and activities of these enzymes in the two species. Indeed, humans and rodents do often have different responses to drugs and other chemicals. However, using transgenic technologies — which will be improved with the availability of the genome — mice can be 'humanized' to metabolize drugs more as we do[12].

The free availability of mouse sequence data, throughout the project, has already allowed the identification of novel genes through comparative genomics. For instance, a gene called *APOAV* encodes a previously unknown member of the apolipoprotein gene family. A remarkable aspect of this story is that the last member of this gene family, *APOAIV*, was cloned nearly two decades ago[13] in the now well-trodden fields of lipid metabolism and cardiovascular disease. After years of thinking that all of the members had been catalogued, the discovery of *APOAV* was quite surprising[14].

The *APOAIV*, *APOCIII* and *APOAI* genes occur within a 20-kilobase stretch of DNA on human chromosome 11 (ref. 15). Through comparative analysis of human and mouse genomic sequences, Pennacchio *et al.*[14] identified a region of conserved sequence, some 25 kilobases from *APOAIV*, that proved to contain the *APOAV* gene. Because the *AIV/CIII/AI* gene locus was known to influence plasma lipid levels in humans, Pennacchio *et al.* studied lipid levels in knockout and transgenic mice and found that the APOAV protein has a strong inverse correlation with plasma triglyceride levels — a risk factor for coronary artery disease. Subsequent studies[16] showed that genetic variation in the *APOAV* locus influences plasma triglyceride levels in humans.

This example is but one harbinger of the impact that the mouse genome will have on human biology, and more systematic efforts will undoubtedly be made to cross-reference known and novel genes in humans and rodents with their biological effects. For instance, the Alliance for Cellular Signaling[17] aims to identify all of the proteins that are involved in signalling pathways in B lymphocytes (antibody-producing cells) and cardiac myocytes (heart muscle cells) in the mouse. By cataloguing all of the orthologous proteins encoded by the human genome, one will eventually be able to move almost effortlessly back and forth between clinical observations in humans and experiments with mouse models of disease. For physiological and pharmacological studies, the rat (not the mouse) has been the long-standing model organism, primarily because of its larger size. A working draft of the rat genome has just become available[18]. A proposed 'triangulation' strategy[19] should powerfully leverage the advantages of all three organisms (mice, rats and humans) for studies of human disease.

Finally, what might the mouse genome teach us about mammalian biology and evolution? About two years ago, an article reporting results of the sequencing and analysis of the *Drosophila melanogaster* (fruitfly) genome was boldly entitled "Comparative genomics of the eukaryotes"[20] — eukaryotes, loosely, being organisms whose cells have nuclei, in contrast to the Bacteria and Archaea. This was despite the fact that genome-scale data for only four eukaryotes (*Drosophila, C. elegans*, the yeast *Saccharomyces cerevisiae* and human) out of several million eukaryotic species on the planet were then available. We must take care not to over-generalize from small sample sizes. Likewise, analyses of the genomes of humans and mice have led to the notion of a 'mammalian gene count' (currently standing at about 28,000) derived from only two out of an estimated 4,600–4,800 extant mammalian species on Earth. But energetic programmes[21] to obtain more generally representative data are under way (Box 1). These are likely to deliver an ultimately more satisfying picture of what the late Stephen Jay Gould called the "full house" of biological variation, from cabbages to kings[22]. ■

Box 1 Sequencing priorities

Much genome-sequencing capacity is still devoted to finishing the human, mouse and rat genomes. But as that work is completed, laboratories will become free to take on new projects. The US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland, one of the main funding agencies for sequencing, has called for 'white paper' proposals for future priorities among candidate organisms. One stipulation for NHGRI support is that data should be released rapidly and without restriction.

The first review of proposals by an expert panel, announced in May of this year, gave high priority to sequencing the genomes of the chicken, chimpanzee and honey bee, as well as several species of fungi, a sea urchin and a protozoan, *Tetrahymena thermophila*. Sequencing work on plants and the Bacteria and Archaea is supported by different agencies.

In a second announcement, in September, the cow and dog were made high priorities, along with *Oxytricha*, a single-celled organism that is of especial interest to geneticists and evolutionary biologists. The cow is evidently of great importance in agriculture, whereas dogs provide good models for certain human diseases, such as narcolepsy and obsessive-compulsive disorder. Along with a sequenced genome of the chimpanzee, the data on both animals will also provide grist for comparative studies among mammals.

Among the criteria for selection are the likely demand for the sequence data from particular research communities, the genome size and projected cost of the project, and whether a complete sequence is necessary.

Mark S. Boguski is at the Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North D4 100, PO Box 19024, Seattle, Washington 98109, USA.
e-mail: mboguski@fhcrc.org

1. Tilghman, S. M. *Genome Res.* **6,** 773–780 (1996).
2. Battey, J., Jordan, E., Cox, D. & Dove, W. *Nature Genet.* **21,** 73–75 (1999).
3. Mouse Genome Sequencing Consortium *Nature* **420,** 520–562 (2002).
4. Tarantino, L. M. & Bucan, M. *Hum. Mol. Genet.* **9,** 953–965 (2000).
5. Bucan, M. & Abel, T. *Nature Rev. Genet.* **3,** 114–123 (2002).
6. Tugendreich, S., Boguski, M. S., Seldin, M. S. & Hieter, P. *Proc. Natl Acad. Sci. USA* **90,** 10031–10035 (1993).
7. Bassett, D. E. Jr *et al. Trends Genet.* **11,** 372–373 (1995).
8. Womack, J. E. & Kata, S. R. *Curr. Opin. Genet. Dev.* **5,** 725–733 (1995).
9. Zuckerkandl, E. & Pauling, L. J. *Theor. Biol.* **8,** 357–366 (1965).
10. Makalowski, W. & Boguski, M. S. *Proc. Natl Acad. Sci. USA* **95,** 9407–9412 (1998).
11. Wheelan, S. J., Boguski, M. S., Duret, L. & Makalowski, W. *Gene* **238,** 163–170 (1999).
12. Xie, W. & Evans, R. M. *Drug Discovery Today* **7,** 509–515 (2002).
13. Boguski, M. S., Elshourbagy, N., Taylor, J. M. & Gordon, J. I. *Proc. Natl Acad. Sci. USA* **81,** 5021–5025 (1984).
14. Pennacchio, L. A. *et al. Science* **294,** 169–173 (2001).
15. Boguski, M. S., Birkenmeier, E. H., Elshourbagy, N. A., Taylor, J. M. & Gordon, J. I. J. *Biol. Chem.* **261,** 6398–6407 (1986).
16. Talmud, P. J. *et al. Hum. Mol. Genet.* **11,** 3039–3046 (2002).
17. www.afcs.org
18. www.hgsc.bcm.tmc.edu/projects/rat
19. Jacob, H. J. & Kwitek, A. E. *Nature Rev. Genet.* **3,** 33–42 (2002).
20. Rubin, G. M. *et al. Science* **287,** 2204–2215 (2000).
21. http://www.genome.gov/page.cfm?pageID=10002154
22. Gould, S. J. *Full House: The Spread of Excellence from Plato To Darwin* (Harmony Books, New York, 1996).