Opinion

The mouse genome sequence - the end of the tail, or just the beginning?

Janet Rossant*† and Stephen W Scherer†‡

Addresses: *Samuel Lunenfeld Research Institute and †Department of Molecular and Medical Genetics, University of Toronto, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada. ‡Department of Genetics, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada.

Correspondence: Janet Rossant. E-mail: rossant@mshri.on.ca

Published: 1 April 2003

Genome ***Biology*** 2003, **4**:109

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/4/109

Abstract

The recent flurry of papers on the mouse genome includes the description of the full genome assembly, analysis of the mouse transcriptome, the origin of interstrain variation, initial analysis of conserved non-coding regions and high-throughput expression analysis of a subset of genes. Each illustrates how the availability of the genome sequence will change the way mouse biologists do business in future.

With the availability of many sophisticated technologies for manipulating its genome, the mouse is the favored experimental mammal for studying many basic biological processes and for developing models relevant to human disease. Its pre-eminent position has only been strengthened by the recent publication in *Nature* of papers detailing the analysis of the draft mouse genome sequence [1] and the annotation of a major set of full-length mouse cDNAs [2]. Using experience and data gained from the human genome annotation, as well as mouse cDNA data and the important RIKEN cDNA resource [2], an estimate of 29,000 protein-coding genes has been made. The continued refinement of gene-prediction algorithms, combined with ongoing efforts to develop increasing numbers of sequenced full-length cDNA libraries, should lead to better and better understanding of the nature and number of protein-coding genes in the mouse genome. Indeed, a recent paper using a two-stage *de novo* gene prediction procedure has added over 1,000 new genes to the mouse gene list [3]. There are still major gaps to be filled in the gene-prediction area, most notably in the definition of single-exon genes and in all the different forms of RNA-encoding genes, but the availability of the list of protein-encoding genes provides an exciting compendium of the component parts of a mouse.

Availability of the sequence of so many genes immediately spawns a wish-list of all the additional information needed to truly understand their function. Here are just some of the issues that now need to be addressed in the same kind of genome-wide detail as the sequence acquisition. How many splicing variants are there and how functionally important are they? How are all the transcripts expressed in time and space, from egg to adult? And where are all the proteins expressed - including intracellular location - from egg to adult? What are the post-translational protein modifications, their distribution and functional importance? What are the dynamics of protein-protein interactions in time and space? What are the effects of genetic gain or loss of function - for all kinds of alleles, from null or conditional to point mutations. What are the gene-gene and gene-environment interactions, and what are their effects? How is the temporal and spatial regulation of transcription and translation achieved? How do the protein-coding sequences and gene regulation differ across species? And can we identify the genes important for susceptibility and resistance to disease? The set of genome papers does not provide answers to all these questions, but it does point to the road ahead in various ways.

How many splicing variants are there?

Although the analysis of both the human and the mouse genome sequences has produced a smaller than expected estimate of the number of protein-coding genes, it is clear

that complexity can be increased by alternative splicing, generating different protein-coding forms. Genome sequence analysis and exon prediction *per se* cannot distinguish which exons are included in the final transcripts, but comparison with cDNAs and expressed sequence tags (ESTs) can give an estimate of the extent of alternate splicing. When 60,000 RIKEN clone sequences and 44,000 mRNAs in the public databases were aligned with the genome assembly [2], 41% of the resulting transcript clusters showed evidence for alternative spliced forms; and 79% of putative splice variants would produce altered protein sequence. These numbers provide a lower limit on the extent and effect of alternative splicing in the mouse: not all transcripts are represented in the current cDNA libraries and not all functional splice variants lead to protein-coding changes. But these numbers already indicate that alternative splicing is an important contributor to the functional complexity of the mammalian transcriptome. In humans it is often difficult to fully validate whether a predicted exon represents an alternate splice form, because rare transcripts may occur in tissues that are not readily accessible, such as the embryo. The mouse provides a much more accessible model for pursuing the extent of alternative splicing. We will need more cDNA libraries from unusual tissue and cell types, and the development of oligonucleotide-based exon arrays to validate expression of spliced forms. Determining the functional consequences of alternative transcripts, whether coding or non-coding, will require genetic disruption of the different forms, a procedure readily achievable in the mouse by gene targeting.

Gene-expression profiling

Once all the transcriptional units, whether coding or non-coding, and all their spliced forms, are fully annotated, some aspects of function may be predicted on the basis of sequence analysis alone. But much more information is needed if we are to understand how individual genes act together to produce the complex biology of the organism. Detailed patterns of transcript expression at all stages of development from egg to adult provide the basis for further investigations in individual tissues and organs. Two papers in the 'mouse genome issue' of *Nature* provide an initial analysis of the expression of all identifiable mouse orthologs of the genes on human chromosome 21 [4,5]. Both groups used *in situ* hybridization to different embryonic stages and, in the case of Gitton *et al.* [5], sections of the neonatal brain, to start building an expression database. Although some interesting expression patterns were observed for novel genes, the papers largely serve to illustrate the limitations of current approaches and the scale of the task ahead. Levels of expression are difficult to quantitate by *in situ* hybridization, making cross-comparison between datasets difficult. Use of more quantitative measures, such as RT-PCR and microarray analysis, can help, but such studies lose the important spatial expression details available from *in situ* approaches. Use of reporter genes, such as green fluorescent protein (GFP) or β-galactosidase (lacZ), simplifies the detection of expression and can potentially be quantifiable. Efforts to generate large numbers of reporter transgenic strains of mouse [6] or targeted knock-in strains are under way. Embryonic stem-cell lines carrying insertions of reporter genes such as GFP and lacZ within every gene in the genome are currently being generated in large-scale gene-trap screens [7] and will provide the most extensive resource of reporter lines for *in situ* gene-expression analysis in the mouse.

The number of stages, tissues and conditions under which one would like to assess gene expression is mind-boggling: how can one deal with the huge amounts of data that would be generated? Static, non-searchable presentation of expression data in the form of images of whole embryos or tissue sections is woefully inadequate for in-depth mining of the functional significance of expression patterns. Ideally, gene expression data would be presented in a three-dimensional, searchable format, a task currently being undertaken by the Emage group at the University of Edinburgh [8] (Table 1 lists useful websites relevant to this article). It is critical that expression data of the sort presented in the recent papers [4,5] is placed in the searchable fields of the text-based Gene Expression Database (GXD) [9] and the image-based Emage database [8]. Methodologies to acquire gene-expression patterns in three dimensions, such as optical projection tomography [10], will be important tools in aiding the construction of large-scale expression databases.

Determination of transcript localization is, of course, only the first step in a comprehensive expression analysis. Protein localization and protein-protein interactions need to be analyzed at the tissue level, and, importantly, at the subcellular level. This will require a major effort aimed at generating a collection of isoform-specific antibodies for the products of all the genes in the genome. Among its other uses, the RIKEN full-length cDNA resource is an ideal substrate for such an undertaking, emphasizing again the importance of the rapid and unfettered distribution of this resource to the world-wide community.

Genome-wide functional analysis

The availability of the draft mouse genome sequence, combined with the excellent resources of genomic DNA cloned within manageable bacterial artificial chromosomes (BACs) [11], have already made much simpler the process of generating targeting vectors for gene mutation in embryonic stem (ES) cells. Further enhancement could be provided by an effort to produce sequence-annotated arrayed libraries of the genome in ready-to-use targeting vectors. Indeed, it is now not entirely unrealistic to consider the goal of generating a targeted mutation in every gene in the genome, at least within ES cells. The problem is that one mutation in a gene is not likely to be enough for full functional analysis. It may be better, therefore, to use less expensive means of generating

Table 1

Useful websites	
Mouse genome sequence browsers	
Ensembl mouse genome server	http://www.ensembl.org/Mus_musculus/
NCBI mouse genome resource	http://www.ncbi.nih.gov/genome/guide/mouse/
UCSC mouse genome browser gateway	http://genome.ucsc.edu/cgi-bin/hgGateway?org=mouse
Gene expression databases	
Mouse Genome Informatics - gene expression	http://www.informatics.jax.org/menus/expression_menu.shtml
Edinburgh Mouse Atlas Project (emap)	http://genex.hgu.mrc.ac.uk
Gene-trap databases	
German GeneTrap Consortium	http://www.genetrap.de
Mammalian Functional Genomics Centre	http://www.escells.ca
Bay Genomics gene trap resource	http://baygenomics.ucsf.edu/
Centre for Modeling Human Diseases (CHMD) - gene trap core	http://www.cmhd.ca/sub/genetrap.asp
SNP resources	
SNPview	http://www.GNF.org/SNP
Whitehead Institute mouse SNP data	http://www-genome.wi.mit.edu/snp/mouse
Roche mouse SNP database	http://mouseSNP.roche.com
The RIKEN FANTOM2 cDNA resource	http://fantom2.gsc.riken.go.jp/

generic loss-of-function mutations in every gene and reserve the more precise tools of gene targeting for those genes whose functions are revealed as critical for specific biological processes. Gene-trap mutagenesis is a random mutagenesis approach that generates large numbers of sequence-tagged insertions in the genome of ES cells, each of which interrupts the coding sequence of a gene [7]. Large-scale efforts to generate libraries of gene-trap insertions are under way worldwide, and it is now very easy to assign the sequence tag from each gene-trap insertion to its position in the genome, making gene trapping a powerful means of rapidly assessing gene function. Potentially even more rapid would be the use of short-interfering RNA (siRNA) to knock-down gene expression *in vivo* [12,13]. Again the availability of genomic and cDNA sequence makes it easy to design blocking siRNAs for any gene of interest.

Even if every gene is known and every gene can be mutated in a sequence-driven manner, this will not provide a complete analysis of the function of the mouse genome. Phenotype-driven approaches will continue to provide insight into the extent to which genetic alterations can contribute to complex traits. Mutagenesis programs using the chemical ethylnitrosourea (ENU) are producing large numbers of phenotypes relevant to human diseases that have to be mapped to a chromosomal region and then positionally cloned. The public sequencing consortium undertook full sequencing of only one inbred mouse strain, C57BL/6J, but it also performed partial sequencing of several other strains and was able to generate a resource of single-nucleotide polymorphisms (SNPs) across the genome [14]. These can be used for mapping purposes and then the draft sequence can be used to identify candidate genes for further analysis. The number of spontaneous and ENU-induced mutations positionally cloned by this hybrid mapping/candidate-gene approach is increasing rapidly and is certainly enhanced by the availability of the genome sequence and the SNP resource.

Available mouse inbred strains also show intrinsic phenotypic variations that can be mapped as quantitative trait loci (QTLs) or genetic modifiers of other mutations. The Mouse Phenome Project [15] is an initiative that is undertaking a detailed baseline analysis of as many different phenotypes as possible in a set of inbred strains, and this will help to characterize strain-specific variation. Identifying the genetic basis of the variation underlying such traits is notoriously difficult, however. How will sequence information help? High-quality sequence information from multiple strains would of course help to identify functional variation in candidate genes mapping to a QTL. But this is not likely to be a feature of the public sequencing effort and is only minimally available in the private domain. The SNP resource does, however, provide some idea of the extent of DNA sequence variation between inbred strains. What is particularly interesting about the available SNP analysis is that there are SNP-rich and SNP-poor regions of the genome when comparing any two inbred strains. Further analysis of SNP distribution among strains led Wade *et al*. [16] to conclude that

most existing mouse strains essentially represent recombinant inbred strains between the two founder species leading to the laboratory mouse, namely *Mus musculus* and *Mus domesticus*. In practice, this means that researchers undertaking a mapping backcross need to be aware of whether the region of interest is SNP-rich or SNP-poor between the two strains of interest.

It has been argued that the haplotype block structure of inbred strains, as revealed by SNP analysis, can be helpful in identifying the basis of QTLs, since the search for candidate differences between two strains can be restricted to regions where they are highly divergent [17]. But this assumes that important QTLs reside in the original species-wide variation introduced into the gene pool of the inbred strains. QTLs might instead represent recent variants specific to a given strain and might reside in SNP-poor domains. The two scenarios can be distinguished if the haplotype structure of the major inbred strains is known and can be associated with specific phenotypes. There is thus clearly a need for a 'HapMap' of the mouse genome, as has recently been proposed as a means of associating phenotype with map position in humans [18]. A recent paper by Wiltshire *et al*. [19] is a first step in this direction. It describes haplotype patterns of SNPs across eight inbred strains of mice.

Gene regulation

One of the key issues in dealing with the complexity of gene function in the living organism is to understand how gene expression is regulated in a cell-, tissue- and developmental-stage-specific manner. We know that control of gene transcription depends ultimately of the activity of transcription factors that bind to regulatory elements close to the gene of interest. We also know that such sites are likely to be conserved across evolution. The sequence analysis of the mouse genome did not make a strong effort to use the genome sequence to identify such elements, but the authors did estimate that fully 5% of the genome of the mouse is under selection [1] when compared with human sequence: this is a much higher proportion than is predicted on the basis of conserved coding sequences alone. Some of this sequence under selection may represent RNA genes, chromatin structural elements, and so on [20], but much of it is likely to represent regulatory elements. Although comparison of mouse and human sequences located near genes can help to identify conserved regulatory elements, it is not easy to find such elements based on comparisons of only two sequences. Comparative genomic analysis - sequencing whole genomes or selected regions from multiple genomes and then looking for conserved elements - can be very useful in identifying conserved regulatory elements; and a number of algorithms have been developed to support such analysis. There is a bit of a paradox, however: analysis to date has focused on identifying conserved elements in genes regulated identically across evolution; but many of the genetic changes that drive evolution are likely to be due to changes in gene regulation, and comparative sequence analysis needs to be able to identify these elements as well. Eventually there will be sufficient sequence available for many different species that some of these evolutionary questions will be resolved.

The mouse genome assembly has now been published in the formal literature, as has the draft *Fugu* genome sequence [21], and an assembly of the rat genome sequence is available online [22]. We can expect the whole-genome draft sequences of other vertebrates to follow in quick succession, making this the coming era of comparative vertebrate genomics. But the mouse still leads the way as an all-round system for in-depth analysis of genome function from proteins to disease, and is likely to continue to do so for many years to come.

References

1. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
2. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al.*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
3. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, *et al.*: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.** *Proc Natl Acad Sci USA* 2003, **100:**1140-1145.
4. Reymond A, Marigo V, Yaylaoglu MB, Leoni A, Ucla C, Scamuffa N, Caccioppoli C, Dermitzakis ET, Lyle R, Banfi S, *et al.*: **Human chromosome 21 gene expression atlas in the mouse.** *Nature* 2002, **420:**582-586.
5. Gitton Y, Dahmane N, Baik S, Ruiz i Altaba A, Neidhardt L, Scholze M, Herrmann BG, Kahlem P, Benkahla A, Schrinner S, *et al.*: **A gene expression map of human chromosome 21 orthologues in the mouse.** *Nature* 2002, **420:**586-590.
6. Heintz N: **BAC to the future: the use of BAC transgenic mice for neuroscience research.** *Nat Rev Neurosci* 2001, **2:**861-870.
7. Stanford WL, Cohn JB, Cordes SP: **Gene-trap mutagenesis: past, present and beyond.** *Nat Rev Genet* 2001, **2:**756-768.
8. **The Emage database** [http://genex.hgu.mrc.ac.uk/Emage/database/intro.html]
9. Ringwald M, Eppig JT, Begley DA, Corradi JP, McCright IJ, Hayamizu TF, Hill DP, Kadin JA, Richardson JE: **The Mouse Gene Expression Database (GXD).** *Nucleic Acids Res* 2001, **29:**98-101.
10. Sharpe J, Ahlgren U, Perry P, Hill B, Ross A, Hecksher-Sorensen J, Baldock R, Davidson D: **Optical projection tomography as a tool for 3D microscopy and gene expression studies.** *Science* 2002, **296:**541-545.
11. Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, *et al.*: **A physical map of the mouse genome.** *Nature* 2002, **418:**743-750.
12. Elbashir SM, Lendeckel W, Tuschl T: **RNA interference is mediated by 21- and 22-nucleotide RNAs**. *Genes Dev* 2001, **15:**188-200.
13. Hasuwa H, Kaseda K, Einarsdottir T, Okabe M: **Small interfering RNA and gene silencing in transgenic mice and rats.** *FEBS Lett* 2002, **532:**227-230.
14. Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, *et al.*: **Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse.** *Nat Genet* 2000, **24:**381-386.
15. **The Mouse Phenome Database** [http://www.jax.org/phenome]
16. Wade CM, Kulbokas EJ, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ: **The mosaic structure of variation in the laboratory mouse genome.** *Nature* 2002, **420:**574-578.

17. Glazier AM, Nadeau JH, Aitman TJ: **Finding genes that underlie complex traits.** *Science* 2002, **298:**2345-2349.
18. Couzin J: **Human genome. HapMap launched with pledges of $100 million.** *Science* 2002, **298:**941-942.
19. Wiltshire T, Pletcher MT, Batalov S, Barnes SW, Tarantino LM, Cooke MP, Wu H, Smylie K, Santrosyan A, Copeland NG, *et al.*: **Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse.** *Proc Natl Acad Sci USA* 2003, **100:**3380-3385.
20. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE: **Numerous potentially functional but non-genic conserved sequences on human chromosome 21.** *Nature* 2002, **420:**578-582.
21. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.*: **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.** *Science* 2002, **297:**1301-1310.
22. **The Rat Genome Page** [www.ncbi.nlm.nih.gov/genome/guide/rat/]