A Guide to the Mammalian Genome

Yasushi Okazaki and David A. Hume

Genome Res. 2003 13: 1267-1272
Access the most recent version at doi:10.1101/gr.1445603

Commentary

A Guide to the Mammalian Genome

Yasushi Okazaki[1,3,4] and David A. Hume[2]

[1]*Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan;* [2]*Institute for Molecular Bioscience, University of Queensland, Brisbane, Q4072, Australia*

Sequencing of a Transcriptome

The rapid completion and public release of the genome sequences of mouse and human has led to a downgrading of the number of "genes" predicted in the mammalian genome to the region of 30,000 (Mouse Genome Sequencing Consortium, Waterston et al. 2002). In simpler organisms such as yeast, the estimate of gene number is comparatively straightforward, because the majority of the genome clearly encodes proteins, and individual genes generally have a well-defined start and finish and a single mRNA output. In mammals, the task is much more complex. Only a small proportion of the genome encodes mRNAs that in turn encode protein, and protein-coding sequence is interspersed with large introns or intergenic regions. Even protein coding genes have proven difficult to annotate reliably (Kawai et al. 2001), and non–protein coding genes are essentially impossible to annotate a priori.

The key to reliable annotation of a mammalian genome is the comprehensive characterization of the transcriptional output, the transcriptome. There are two approaches to this problem. The most common is high-throughput sequencing of cDNA ends (ESTs). In mouse and human, and to a lesser extent in many other mammals, there are millions of EST sequences in various repositories. EST sequences can be computationally assembled into clusters, as in the UniGene projects (http://www.ncbi.nlm.nih.gov/UniGene). There are many drawbacks with this approach, both from the cDNA cloning and sequence quality and from computational perspectives, but the most compelling is that the sequences are generated in silico and are not necessarily supported by a physical clone. It is also rather inefficient, because even with the best subtraction and normalization, abundant transcripts have been sequenced thousands of times, whereas many rare transcripts are absent from EST databases. EST assemblies are particularly difficult to interpret when there are multigene families or complex alternative splicing.

The alternative approach is to systematically isolate and sequence full-length cDNAs. The logistics of this approach are daunting, and it is actually far more challenging than is genomic sequencing, especially using shotgun approaches because of the difficulties in the collection of the samples. Nevertheless, the RIKEN Mouse Gene Encyclopedia Project has taken this approach. In the process, the RIKEN team has provided a model for eukaryotic transcriptome projects. The task required a range of new technologies and approaches. In outline, the RIKEN team developed new approaches to production of full-length cDNAs (Carninci et al. 2003) that required (1) a novel reverse transcriptase reaction (to enable effective complete first-strand synthesis), (2) novel 5′ end capture technology, and (3) novel approaches to normalization and subtraction of cDNA libraries.

Starting with their first libraries, the RIKEN team sequenced 3′ ends (and later 5′ ends) in a Phase 1 sequencing pipeline and, for each individual clone, determined whether the sequence had been sequenced previously or could be ascribed to a new cluster. In the second phase, individual representatives of EST clusters were selected and fully sequenced to produce a full-length cDNA sequence representing the sequence of an individual physical clone. At a number of stages in the project, the RIKEN team assembled a set of cDNAs that had previously been sequenced and used them to subtract successive libraries. The success of the approach is outlined in detail in Carninci et al. (2003). The output of this pipeline was analyzed in the FANTOM2 meeting (April 29 to May 5, 2002, Yokohama, Japan), which is the basis of this special issue of *Genome Research*.

The History of FANTOM

The FANTOM (*F*unctional *An*not*o*tation *o*f *M*ouse) Consortium is a group of molecular biologists from the RIKEN Genomic Sciences Center in Yokohama, elsewhere in Japan and from around the world in Europe, North America, and Australia. Among this group are representatives of the major mouse genome resource centers and sequence databases: the Mouse Genome Informatics (MGI) at the Jackson Laboratories, the TIGR mouse gene index, the National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute (EBI). Additional participants bring a range of expertise in sequence analysis and functional annotation, largely invited as a result of the many collaborations that have come out of the RIKEN project. The first meeting of the FANTOM Consortium (FANTOM1) was assembled to annotate the initial output of the RIKEN pipeline, 21,076 sequences (Kawai et al. 2001), which at the time was the largest assembly of completed full-length cDNA sequences for any organism. In some measure inspired by earlier genome annotation jamborees (Pennisi 2000), around 100 participants were cloistered for two weeks in front of computer screens in Tsukuba, Japan, and emerged with a first-pass annotation of the individual cDNAs and a picture of the scope of the task for the future. A number of things

[3]**Corresponding author.**
E-MAIL okazaki@gsc.riken.go.jp; FAX 81-45-503-9216.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1445603.

were clear. First, despite the systematic pipeline, there was considerable redundancy in the sequenced cDNA set. One cause of this redundancy was the very high level of 3′ end variation (alternative polyadenylation/termination) in mammalian mRNAs. Second, a significant proportion of the cDNAs could not be ascribed any useful or definitive annotation and remained "unclassified." Finally, the logistics and failings of human curation were confronted, and it became clear that a much more efficient and rigorous precomputational approach would be needed for annotation of a larger cDNA set in the future.

FANTOM2: The Bigger Picture

The FANTOM2 Consortium confronted the task of annotating 60,770 completed full-length cDNAs. To plan for this task, a small group of senior international scientists joined their Japanese colleagues for the FANTOM2 Typhoon meeting in October 2001 in Yokohama. The group recognized that the key to effective use of human time was the assembly and presentation of all available information pertaining to an individual clone. If such information could be presented on a Web-based annotation interface, then the need to assemble an annotation team in one place could be avoided. The concept of the Mouse Annotation Teleconference for RIKEN cDNA Sequences (MATRICS) was born. The group also recognized the possibility that an appropriate computational pipeline could obviate the need for human curation in the large majority of cases, and such an innovation would permit continued upgrade of annotations in the future. The key to automated annotation is the establishment of a controlled vocabulary and decision tree, which forms the basis for the ordered presentation of information on an annotation interface. The design and operation of the annotation pipeline and tools are described in detail in papers by Kasukawa et al. (2003). The purpose of annotation is to choose a name that provides the maximum possible information, and to link that name to a range of relevant functional and structural information including gene ontology predictions. Ultimately, the most important applications of annotation occur in large-scale gene expression analysis and systems biology. In the pipeline, priority was given to identification of homology and orthology relationships, but this sometimes yields a very uninformative name. For example, the pathway may give the name "homolog to" human clone K1234567. A key part of the automated annotation was the successful generation of a screen for such uninformative names (Kasukawa et al. 2003).

The MATRICS activity was carried out over a period of around two to three months, involving 90 annotators who each had the opportunity to choose sets of transcripts of special personal interest and expertise. At the end of this period, the outcomes were collated. Curators changed the names in around 20% of cases, not always rationally (Kasukawa et al. 2003). A comprehensive review of the reasons will certainly permit fully automated annotation in the future.

The Cherry Blossom Meeting and Definition of a Transcriptional Unit

The final phase of the annotation of the RIKEN cDNA set was at the Cherry Blossom Meeting (April 29 to May 7, 2002) in Yokohama. Participants submitted abstracts to the meeting based upon their analyses of the set of RIKEN sequences to which they had privileged access during the MATRICS period. During this meeting, there was an ongoing effort to complete the annotation and discussion about annotation criteria and the minor deficiencies of the pipeline. The minds of participants were focused with a session of Zen meditation, and the occasion was celebrated with the ceremonial planting of a cherry blossom tree at the front of the new RIKEN Genomic Sciences Center in Yokohama (Fig. 1).

Many of the abstracts formed the basis for the full papers contained in this special issue of *Genome Research* and provided the outlines for the focused summary of the whole project in *Nature* (Okazaki et al. 2002). Early in the meeting, there was a discussion about the global synthesis of the transcriptome information, and a realization that with the combined information from RIKEN and the public domain, and the resources and skills of RIKEN, MGI, TIGR, NCBI, EBI, and other major computational biology centers, it would be possible to obtain a global overview of the transcriptional output of a mammalian genome for the first time.

Figure 1 A shot from the Zen meditation ceremony held as an excursion during the FANTOM2 Cherry Blossom Meeting. The Zen meditation was a good break and provided the participants with novel inspiration.

The determination to undertake such an analysis lead to a spirited debate as to how one would describe the transcriptome. The de facto definition of a "gene" in mammalian genome annotation is a segment of DNA from which an mRNA is transcribed. The correct functional genetic definition of a gene as a segment of DNA that is able to complement a mutant phenotype is lost, because a genome sequence alone cannot be used to infer function. The transcriptome is the set of transcripts that is derived from a genome. The problem in defining a transcriptome is that individual genomic loci can produce multiple overlapping transcripts, identified computationally as a cluster of overlapping sequences. Especially if the members of a cluster encode a protein, they would generally all have the same name, regardless of alternative 5′ or 3′ UTR or alternative splicing. Neither gene nor locus is an appropriate term to describe the DNA from which such a cluster of mRNA originates. A cluster of transcripts could arise from a so-called expressed pseudogene, and an individual locus can encode clusters from either strand, partly or completely overlapping. The term transcriptional unit (TU) was therefore adopted to describe the genomic DNA region spanned by a cluster of transcripts. The RIKEN sequences were clustered with known mouse cDNA sequences. Then a single sequence was chosen from among the individual clusters to represent each TU; the set of representative transcript sequences (RTSs) was used in global analysis of the complexity and diversity of the mouse transcriptome.

One area in which the RTS was not applied was in the analysis of alternative splicing, in which all of the transcripts were used. The RIKEN clone set combined with public sequences provided the most global analysis ever available of the impact of alternative splicing (Zavolan et al. 2003). Clearly, the large majority of protein-coding TUs undergo some form of alternative splicing, and many are controlled by alternative promoters producing variant 5′ UTRs. This fact has been known for some time, in part because alternative splicing can also be inferred from ETS assemblies. The difference in the RIKEN data is that the alternative splice forms exist as real physical clones, and one can start to analyze whether, for example, alternative promoter or termination is associated with internal alternative splicing. The high frequency of alternative splicing also indicates that the transcriptome is far from complete. Continued sequencing of additional representatives from Phase 1 sequence clusters would undoubtedly give a very high return in terms of variant forms derived from known TUs.

Cherry Blossom Highlights

How Many Genes?

The focus of all discussion of genomes is on the number of genes. In the transcriptome project, the products of ~37,000 TUs were identified as cDNA clusters. Of these, around half encode a predicted protein. The RIKEN project alone contributes around 90% of the total TUs. For a number of reasons, including mapping of numerous 5′ and 3′ EST clusters that remain to be sequenced fully, this remains a substantial underestimate of the total number of TUs in a mammalian genome (Okazaki et al. 2002). That view is supported by the fact that the ENSEMBL annotated mouse genome sequence (Waterston et al. 2002) contains ~10,000 protein coding gene predictions for which the full-length cDNA has not yet been sequenced. The total number of TUs could be as high as 70,000 based upon current analysis, and as pointed out by Wells et al. (2003), there are still significant sources of novel mRNAs that have not been fully sampled. Can we equate a TU with a gene? If a segment of DNA is transcribed, it is reasonable to assume that it has a function. The null hypothesis, that it does not have a function, is impossible to prove because the full range of physiological circumstances in which a transcript could contribute to "fitness" can never be sampled. Hence, if the definition of a gene as a unit of function is accepted, there are actually far more "genes" in the genome than have been annotated in genome sequencing projects, and the large number presumably contributes to the complexity of mammalian cell and developmental biology.

In this special issue of *Genome Research*, there are a number of full papers that arise from the Cherry Blossom meeting. Some of these papers are concerned directly with the FANTOM2 data set and MATRICS, and include the development and assessment computation and annotation tools, the analysis of the diversity of transcripts and of the proteins encoded by a subset of the transcripts, and the implications for human genetics and diseases. A separate group of papers deals with several applications of the RIKEN cDNA collection in functional genomics. Finally, we present a selected set of abstracts that have not yet been expanded to produce full papers, but which further highlight the advances arising from FANTOM2.

The Annotation Process

As mentioned above, a key resource for MATRICS was the development of an automated annotation pipeline to provide curators with information in a logical progression and expedite a rapid naming decision. The development and evaluation of the pipeline is described in the article by Kasukawa et al. (2003), which suggests that automated annotation is possible. There are still significant issues in manual gene name assignment. Ongoing review of the reasons for such events will allow refinement of the automated pipeline. One might have expected that the isolation of full-length cDNA would make the task of identification of protein coding transcripts, and assignment of the open-reading frame (ORF), a straightforward computational task. This is clearly not the case. Partly the problem arises because even the best of sequencing (99.99% accuracy) will have one error in a significant subset of cDNA sequences. Annotators were provided with a choice of ORF predictions, including some that correct possible frame shifts, as well as sequence quality information, and were required to judge the best alternative, sometimes taking account of alignments to related protein or DNA sequences. The paper by Furuno et al. (2003) shows that no one CDS prediction method, including the DECODER method developed specifically for the purpose, was accepted by human curators in >80% of instances. This outcome highlights the even greater problems in genome sequence annotation when predicted ORFs are interrupted by introns. The paper also highlights the particular problem of annotation of ORFs of <100 amino acids, which really require experimental validation before they can be accepted.

The MATRICS interface, which is publicly released, now provides the portal to all of the information pertaining to individual RIKEN cDNAs. Included on this viewer is additional information that was not directly used in

annotation, including the information from protein-protein interaction studies (Suzuki et al. 2003) and cDNA microarray expression profiling detailed below (Bono et al. 2003a). A new tool that has been developed is FACTS, described in the article by Nagashima et al. (2003), which extracts relevant information from the literature based upon text searches. The RIKEN cDNAs have also been integrated into other major mouse genome resources. The fruitful collaboration between the NCBI and MGI groups, who were active participants in FANTOM2, is described in the paper by Baldarelli et al. (2003).

Bioinformatics

The dominant motivation of many of the participants in the FANTOM2 consortium was to get their hands on lots of new genes that are rather more than predictions from genomic sequence. The headline finding in the *Nature* paper (Okazaki et al. 2002) is that a very substantial proportion of the transcripts do not encode a protein. The paper by Numata et al. (2003) adds considerable depth to this analysis and describes additional validation of a subset of these transcripts based upon expression information and sequence conservation. A subset of these transcripts represents antisense counterparts to coding transcripts, and the paper by Kiyosawa et al. (2003) describes >2000 examples of pairs of complementary transcripts, with profound implications for the nature of gene regulation in mammals. The report by Zavolan et al. (2003) analyzes the extent of alternative splicing in the mouse transcriptome and provides another example of an analysis that cannot be contemplated based upon genome sequence information alone.

The representative protein set (RPS) derived from RIKEN transcripts and known mouse protein sequences provided a substantial new discovery resource, and thousands of new predicted proteins were discovered. A global overview of the predicted proteome is provided by Kanapin et al. (2003), who also exploit the availability of this large data set to identify new conserved protein motifs and domain combinations. Individual expert groups have taken the opportunity to produce a comprehensive analysis of protein families and biological systems in a single mammal for the first time. Highlights include identification of the complete metabolome (Bono et al. 2003b); protein family analysis of the zinc finger family, which is the largest single domain class in the genome (Ravasi et al. 2003); identification of new chromodomain proteins (Tajul-Arifin et al. 2003); and analysis of new phosphoregulators (kinases and phosphatases) (Forrest et al. 2003a), G protein–coupled receptors (Kawasawa et al. 2003), proteins that control the cell cycle (Forrest et al. 2003b), cytokine-related genes (Brusic et al. 2003), members of the kinesin superfamily (Miki et al. 2003), and regulators of ubiquitin-dependent protein turnover (Semple and RIKEN GER Group and GSL Members 2003). A novel analytical tool was used to identify candidate secreted proteins, the secretome (Grimmond et al. 2003).

In many instances, the identification of protein family relationships based upon a comprehensive data set will direct reannotation of some of the cDNAs. We may also hope that it will eventually lead to the adoption of rational nomenclatures for large multigene families and identification of more stringent orthology relationships between mouse and human. As noted by the Mouse Genome Sequencing Consortium (Waterston et al. 2002), comparative analysis of sequences from humans, and the most experimentally tractable mammalian model organism, is a very powerful arm of functional genomics in the future. Schriml et al. (2003) focus on the representative transcripts that are the likely homologs of known human disease causing genes listed in OMIMS. Silva et al. take a different approach and extend the analysis to the identification of transcripts that are "similar to" human disease-causing genes. Finally, the comparative analysis of cell death and inflammatory signaling pathways between the two species by Reed et al. (2003) reveals likely orthologs for 219/227 human genes known to be involved in these processes.

Experimental Functional Genomic Studies

The RIKEN cDNA collection is clearly a major resource for functional genomics, and the possible function of the large pool of noncoding mRNAs is especially tantalizing. In the annotation viewer, annotators are provided with information about EST sequences and library of origin. The new cDNAs making up the FANTOM2 set have been used in large-scale gene expression array profiling, and the outcomes are summarized in the report by Bono et al. (2003a). As well-defining, very large clusters of tissue-restricted protein coding transcripts, these data strongly support the view that the large majority of the noncoding RNAs are, indeed, reproducibly expressed. At least some of these transcripts probably contribute to the phenomenon of genome imprinting. In a separate study, Nikaido et al. (2003) used microarrays to identify >2000 transcripts that differ in their expression depending upon parent chromosome of origin. In a more focused study, Holmes et al. (2003) confirm the imprinted expression of 12 separate transcripts from the *gnas* complex on distal mouse chromosome 2.

Beyond FANTOM2

The RIKEN Phase 1 collection of 5′ EST clusters currently numbers 116,660, and 3′ clusters 131,335, of which less than half have been fully sequenced. As shown in the report by Wells et al. (2003), there are still tissues and cell types that have not been fully sampled. So, we can be confident that the number of experimentally validated TUs in the mouse genome will continue to expand. Additionally, the RIKEN project has sequenced only one representative of each Phase 1 cluster, and the alternative splicing analysis carried out by Zavolan et al. (2003) is possible only because the variation in the 5′ and 3′ ends leads to the collapse of clusters. The sequencing of additional members of individual clusters will clearly be productive. Additionally, future gene expression analysis will require the generation of exon-specific probes, presumably involving long oligonucleotides. Additionally, systematic studies on the localization of mRNAs in individual cells, especially those of the developing embryo, are being undertaken by RIKEN and collaborators as well as in other sites worldwide.

Functional genomic analysis of all the noncoding mRNAs in the RIKEN collection will be a major challenge for the future. There are very few examples confirming a function (e.g., *Xist, AIR*) and few clues as to precisely how they function. We are likely to focus on approaches to selective disruption of expression by mutagenesis, RNAi, or other approaches. For protein-coding

transcripts, the physical clones produced in the RIKEN project provide us with the resources for high-throughput analysis of many aspects of biology. Some of the obvious future directions that are already underway include: systematic analysis of protein–protein interaction partner relationships, high-throughput expression of individual proteins for determination of three-dimensional structure (structural genomics) by crystalography or NMR, localization of proteins within cells by high-throughput transfection of epitope-tagged proteins, and functional analysis based upon the impact of over-expression on biologies such as proliferation, survival, and adhesion. Neither the genome nor the transcriptome project has addressed the impact of genetics. We know a great deal about a single mouse. We also know that mobile genetic elements in mice are important mutagens, and as we survey the full spectrum of around 50 inbred strains, we will undoubtedly identify many more TUs and also discover that there are many different genetic ways to end up being a mouse.

The major challenge of the future will be to integrate all of this information in diverse formats into accessible databases that are effectively presented and systematically curated. The RIKEN Mouse Gene Encyclopedia will be one such project. For those of us who participated in the RIKEN project, the most stimulating aspect has been the spirit of cooperation and camaraderie that has developed as a consequence and which we all hope will continue to be the driving force into the future.

REFERENCES

Baldarelli, R.M., Hill, D.P., Blake, J.A., Adachi, J., Furuno, M., Bradt, D., Corbani, L.E., Cousins, S., Frazer, K.S., Qi, D., et al. 2003. Connecting sequence and biology in the laboratory mouse. *Genome Res*. (this issue).

Bono, H., Yagi, K., Kasukawa, T., Nikaido, I., Tominaga, N., Miki, R., Mizuno, Y., Tomaru, Y., Goto, H., Nitanda, H., et al. 2003a. Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays. *Genome Res*. (this issue).

Bono, H., Nikaido, I., Kasukawa, T., Hayashizaki, Y., RIKEN GER Group and GSL Members, and Okazaki, Y. 2003b. Comprehensive analysis of the mouse metabolome based on the transcriptome. *Genome Res*. (this issue).

Brusic, V., Pillai, R.S., Silva, D.G., Petrovsky, N., RIKEN GER Group and GSL Members, and Schönbach, C. 2003. Cytokine-related genes identified from the RIKEN full-length mouse cDNA data set. *Genome Res*. (this issue).

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki. D., et al. 2003. Targeting a complex transcriptome: The construction of the mouse full-length cDNA encyclopedia. *Genome Res*. (this issue).

Forrest, A.R.R., Ravasi, T., Taylor, D., Huber, T., Hume, D., RIKEN GER Group and GSL Members, and Grimmond, S. 2003a. Phosphoregulators: Protein kinases and protein phosphatases of mouse. *Genome Res*. (this issue).

Forrest, A.R.R., Taylor, D., RIKEN GER Group and GSL Members, and Grimmond, S. 2003b. Exploration of the cell-cycle genes found within the RIKEN FANTOM2 data set. *Genome Res*. (this issue).

Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., Hayashizaki, Y., and Okazaki, Y. 2003. CDS annotation in full-length cDNA sequence. *Genome Res*. (this issue).

Grimmond, S.M., Miranda, K.C., Yuan, Z., Davis, M.J., Hume, D.A., Yagi, K., Tominaga, N., Bono, H., Hayashizaki, Y., Okazaki, Y., et al. 2003. The Mouse secretome: Functional classification of the proteins secreted into the extracellular environment. *Genome Res*. (this issue).

Holmes, R., Williamson, C., Peters, J., Denny, P., RIKEN GER Group and GSL Members, and Wells, C. 2003. A comprehensive transcript map of the mouse *gnas* imprinted complex. *Genome Res*. (this issue).

Kanapin, A., Batalov, S., Davis, M.J., Gough, J., Grimmond, S., Kawaji, H., Magrane, M., Matsuda, H., Schönbach, C., Teasdale, R.D. et al. 2003. Mouse proteome analysis. *Genome Res*. (this issue).

Kasukawa, T., Furuno, M., Nikaido, I., Bono, H., Hume, D.A., Bult, C., Hill, D.P., Baldarelli, R., Gough, J., Kanapin, A., et al. 2003. Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res*. (this issue).

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Kawasawa, Y., McKenzie, L.M., Hill, D.P., Bono, H., RIKEN GER Group and GSL Members, and Yanagisawa, M. 2003. G-Protein-coupled receptor genes in the FANTOM2 database. *Genome Res*. (this issue).

Kiyosawa, H., Yamanaka, I., Osato, N., RIKEN GER Group and GSL Members, and Hayashizaki, Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res*. (this issue).

Miki, H., Setou, M., RIKEN GER Group and GSL Members, and Hirokawa, N. 2003. Kinesin superfamily proteins (KIFs) in the mouse transcriptome. *Genome Res*. (this issue).

Nagashima, T., Silva, D.G., Petrovsky, N., Socha, L.A., Suzuki, H., Saito, R., Kasukawa, T., Kurochkin, I.V., Konagaya, A., and Schönbach, C. 2003. Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS. *Genome Res*. (this issue).

Nikaido, I., Saito, C., Mizuno, Y., Meguro, M., Bono, H., Kadomura, M., Kono, T., Morris, G.A., Lyons, P.A., Oshimura, M., et al. 2003. Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res*. (this issue).

Numata, K., Kanai, A., Saito, R., Kondo, S., Adachi, J., Wilming, L.G., Hume, D.A., RIKEN GER Group and GSL Members, Hayashizaki, Y., and Tomita, M. 2003. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res*. (this issue).

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Pennisi, E. 2000. Ideas fly at gene-finding Jamboree. *Science* **287:** 2182–2184.

Reed, J.C., Doctor, K., Rojas, A., Zapata, J.M., Stehlik, C., Fiorentino, L., Damiano, J., Roth, W., Matsuzawa, S., Takayama, S., et al. 2003. Comparative analysis of apoptosis and inflammation genes of mice and humans. *Genome Res*. (this issue).

Ravasi, T., Huber, T., Zavolan, M., Forrest, A., Gaasterland, T., Grimmond, S., RIKEN GER Group and GSL Members, and Hume, D.A. 2003. Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. *Genome Res*. (this issue).

Schriml, L.M., Hill, D.P., Blake, J.A., Bono, H., Wynshaw-Boris, A., Pavan, W.J., Ring, B.J., Beisel, K., Setou, M., RIKEN GER Group and GSL Members, et al. 2003. Human disease genes and their cloned mouse orthologs: Exploration of the FANTOM2 cDNA sequence data set. *Genome Res*. (this issue).

Semple, C.A.M., and RIKEN GER Group and GSL Members. 2003. The comparative proteomics of ubiquitination in mouse. *Genome Res*. (this issue).

Silva, D.G., Schönbach, C., Brusic, V., Socha, L.A., Nagashima, T., RIKEN GER Group and GSL Members, and Petrovsky, N. 2003. Identification of novel "pathologs" (human disease-related gene candidates) from the RIKEN full-length mouse cDNA data set. (Abstract) *Genome Res*. (this issue).

Suzuki, H., Saito, R., Kanamori, M., Kai, C., Schönbach, C., Nagashima, T., Hosaka, J., and Hayashizaki, Y. 2003. The mammalian protein–protein interaction database and its viewing system that is linked to the main FANTOM2 viewer. *Genome Res*. (this issue).

Tajul-Arifin, K., Teasdale, R., Ravasi, T., Hume, D.A., RIKEN GER Group and GSL Members, and Mattick, J.S. 2003. Identification and analysis of chromodomain-containing proteins encoded in the mouse transcriptome. *Genome Res*. (this issue).

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wells, C.A., Ravasi, T., Sultana, R., Yagi, K., Carninci, P., Bono, H., Faulkner, G., Okazaki, Y., Quackenbush, J., Hume, D.A., et al. 2003. Continued discovery of transcriptional units expressed in cells of the mouse mononuclear phagocyte lineage. *Genome Res*. (this issue).

Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D.A., RIKEN GER Group and GSL Members, Hayashizaki, Y., and Gaasterland, T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res*. (this issue).