GENOME SCIENCE OF VERTEBRATE

Naoki Osato, Hidenori Kiyosawa, Jun Kawai, and **Yoshihide Hayashizaki**
RIKEN Yokohama Institute, Japan

Keywords: inbred strains, linkage analysis, gene-knockout mice, transgenic mice, genome sequencing project, single-read expressed sequence tags (EST), full-length cDNA project, positional cloning, mouse cDNA encyclopedia, positional candidate

Contents

Summary

Mice began to be used for biological experiments for many reasons outlined in Section 1 of this article. Mouse genome sequences have been determined by two organizations. In the first, the single-read expressed sequence tags (EST) project began with the use of ESTs as markers on chromosomes. In the second, Japan originally started a full-length cDNA project. The Mouse Encyclopedia Project is mainly conducted by the Laboratory for Genome Exploration Research Group of the RIKEN Yokohama Institute, which leads the field in the collection and sequencing of full-length cDNA clones. The United States in turn started the Mammalian Gene Collection project. A Mouse cDNA annotation meeting was held at RIKEN in Japan. Researchers from all over the world predicted 13 000 novel mouse genes and found 8000 genes to be novel for mammalians by a computational approach. A positional cloning method is a powerful tool to identify causative genes for phenotypes such as diseases, especially in the mouse, because it is possible to mate any combination of individual mice and construct large pedigrees to facilitate genetic analyses. Once the mouse cDNA encyclopedia is established, it will facilitate many ways to find genes based on the new positional candidate approach. After delineating the chromosomal region by linkage, all genes in the region can be

easily input into a computer. Isolating the cDNA clones *de novo* is no longer necessary. Instead, full-length cDNA clones are chosen from the cDNA Encyclopedia.

1. The Importance of the Mouse Genome and Transcriptome Study

In the era of the genome sequencing project, there must be a reason for any organism to be chosen for the project. Generally, each organism must be a representative model for various biological phenomena. For example, in the case of budding yeast (*Saccharomyces cerevisiae*), its genotypes are directly linked to its phenotypes and the screening of mutants is easy because yeast is unicellular and spends almost all periods of its life cycle as a haploid. The rescue of mutated phenotypes is relatively simple through the complementation and identification of the genes responsible for the mutated phenotypes. Though fruit flies (*Drosophila melanogaster*) are multicellular organisms, their life cycle is relatively short and observation during embryo genesis is simple. The genetic map of *D. melanogaster* was established in the early twentieth century and enormous numbers of mutants have been isolated and characterized. The nematodes (*Caenorhabditis elegans*) have similar advantages, such as a short life cycle and easy observation during embryo genesis. In addition, the cell numbers constituting their bodies are so limited that individual cell lineage can be traced through embryo genesis. In this way, nematodes have made a great contribution to developmental biology. Zebra fish are also one of the greatest models for developmental biology, in that they are vertebrates and the developmental process can be seen in their transparent eggs.

There are also many reasons for using mice. They were first used for biological experiments in the early twentieth century when inbred strains were produced. Although the life cycle of mice is much shorter than that of the human, its physiology is almost the same and mice, as mammals, have been models for various aspects of the life activities of humans. The establishment of many inbred strains and cross-breeding with other species made it possible to perform genetic linkage analysis. Genome-wide mutant production became available with the use of ethylnitrosourea (ENU). Isolation of embryonic stem (ES) cells and the development of gene technology brought about the production of knockout mice and transgenic mice. Under these circumstances, the mouse was chosen as the second most important mammal used in the genome sequencing project after the human.

Compared with the human, the mouse is superior as a model for experiment for the following reasons:

- The mouse's biochemical pathways are similar to those of the human. Hence, the mouse can be used to study the effects of environmental factors that cannot be controlled in human diseases.
- Inbred strains that have the same genetic background are available. These strains enable the operation of genes by mating experiments. Embryological techniques such as gene induction, gene knockout, and site-directed and random mutagenesis are also possible, so that the mechanisms of polygenic diseases can be investigated.

- A significant number of studies have accumulated information about mice mutagenesis and their phenotypes in comparison with human Mendelian genetic diseases.
- The human and mouse genomes are almost the same length (3 billion base-pairs). Therefore, human genes are expected to exist in the mouse genome.
- Human diseases are expected to occur in the mouse.
- The mouse can be employed to extract genes that are difficult to extract from the human, such as genes derived from embryo genesis, so that lethal genes can be approached.

Hence, the mouse is an important model animal for the analysis of human biological processes in normal and disease conditions.

2. The History of Mouse Genetics

2.1. The Beginning of Mouse Genetics and the Establishment of Inbred Strains

Mouse genetics began with a confirmation of Mendel's laws using mice just after the rediscovery of his laws in 1900. Many breeding experiments initially analyzed the genetic basis for body hair color. As a result, the famous Ay gene locus (dominant and lethal), which coded yellow hair color, was identified. The production of inbred strains also began, and the first inbred strain, DBA (dilute, brown, non-agouti mice), was established. An inbred strain can be produced by breeding between siblings for more than 20 generations. In due course, large numbers of inbred strains were established, often in relation to cancer research. This research has shown that multiple loci are involved in cancers and each type of cancer may have a specific set of loci responsible for its phenotype.

Another major field in the early days of mouse genetics was targeted toward studies of the complex T locus, which is responsible for notochord formation during embryo genesis. Because mutations in this locus led to embryonal lethality, studies of the T locus developed a close relationship to developmental biology and brought attention to pleiotropism, a phenomenon where a mutation in a single gene affects various phenotypes. Nowadays, this phenomenon can be understood as being caused by a mutation in transcription factors or by a dysfunction in the signal transduction and cascade systems among genes. In the past, however, genes were not known to be DNA, and much research focused on this matter. Threshold characters during embryo genesis, histocompatibility (H2 locus), sex determination, and X chromosome inactivation have also been major subjects of mouse genetics since the early days.

In the 1940s a new type of inbred strain, a congenic strain, was developed. This development was achieved by transferring a small segment of a chromosome from one strain to the chromosome of another. More than 10 generations of backcrosses were needed. Two other types of inbred strains were also produced: recombinant inbred strains were obtained by crossing two inbred strains, each of which possessed 50% of the genetic material from the two parental strains, and the recombinant congenic strains, which have more than half the chromosomes from one of the parental strains. These

strains were initially produced to promote studies of diseases caused by multiple factors such as cancers.

Though genetic mapping based on the phenotypes of mice had been performed between inbred strains, limited numbers of polymorphisms presented a weak point. In the 1980s, inter-species breeding (for example, *Mus spretus*) became available and the development of microsatellite markers that enabled typing with PCR rapidly accelerated genetic mapping.

A method to map genes on chromosomes without directly mating mice, called radiation hybrid (RH) mapping was also developed. This method utilizes hamster cells containing various portions of the fragmented mouse genome. The markers on the mouse genome, such as microsatellite markers, single-read expressed sequence tags (EST), and cDNA can be mapped by PCR with hamster genomic DNA.

As many genes were isolated and mapped on the chromosomes in various species, the mapping of genes using information from the conservation of the genome sequences among species was performed. Continuous portions on the chromosomes were often conserved in both mice and humans. When one gene is mapped on the human chromosome, this gene, even if not yet discovered in the mouse, can be mapped on the mouse chromosome, assuming the region around this gene is conserved between humans and mice.

With the advance in gene engineering technologies, a new field for studying the phenotypes of individuals containing artificially modified genes reversed approaches to classical genetic mapping. It entailed both the production of transgenic mice and the production of gene-knockout mice. Genes associated with human diseases are introduced to fertilize mouse eggs and the resultant transgenic mice can be used to analyze various aspects of the diseases. Many inbred strains with diverse genetic backgrounds exist in mice and this diversity makes it possible to study the difference in occurrence of specific diseases associated with specific genetic backgrounds. The genes may be introduced as cDNA or BAC (bacterial artificial chromosome) fragments. When a BAC fragment is introduced, expression close to that of a normal condition can be expected. The knockout mice are produced by introducing a mutation to a specific gene on the chromosomes, rather than by introducing new copies of genes in transgenic mice. Next, the function of the gene is analyzed by studying the phenotypes of the knockout mice. This analysis is also a powerful technique to further prove the candidacy of the genes positionally cloned in human diseases. It is thus becoming possible to time the expression and tissue specificity of gene knockout conditions.

In the post-genome era where the human genome draft sequence is complete and the mouse genome draft sequence is in the assembly stage, approaches to genetics are genome-wide. Systematic high throughput production of mutants with ENU is performed all over the world. ENU, applied to mature male mice, causes point mutations in their sperm DNA. The phenotypes of the offspring of these male mice are screened and interesting phenotypes are chosen for further analysis.

In contrast to mutagenesis with ENU, which is driven by phenotype screening, a gene trap represents an approach from genotypes. A neomycin resistant gene without a promoter is introduced to the chromosomes of ES cells to randomly disrupt the genes on the chromosomes. Next, the chromosomal DNA of the promoter region of the "trapped gene" is recovered from the surviving ES cells with the presence of neomycin. About 10–15% of the trapped genes have homologies to the sequences found in the public databases. Knockout mice with mutations of interest for future studies will be produced with ES cells.

3. The Determination of Mouse Genome Sequences

Mouse genome sequences have been determined by two organizations, the Mouse Sequencing Consortium (MSC) and Celera. The MSC comprises 10 public sequencing centers who make their raw data freely available to assist in deciphering the genetic code of the mouse and producing a high quality genome sequence. Deciphering the genetic code of the mouse helps us to understand both mouse and human genes by comparing them with human genome sequences. The first MSC effort was completed in 2000, but the Consortium did not assemble their sequencing data. MGC chopped genomic DNA from B57BL/6 mice and sequenced this DNA in threefold coverage of the entire genome.

The sequences are typically short: no more than 500 bases long, which are shorter than most single genes. MSC is planning further genome sequencing of the mouse in detail. The sequence will appear in 2003 as a high-quality draft of the mouse genome sequence, and the final version will be finished by 2005.

Celera, a private sector company, claimed in 2000 to have determined a mouse genome sequence of about 2.6 billion bases, which covers 99% of the entire mouse genome. Celera sequenced DNAs from three strains (129X1/SvJ, DBA/2J and A/J) and provided sixfold coverage of the genome. This sequence is redundant enough to assemble and order the sequence fragments into a rough draft of the genome. Currently, however, a fee is charged for access to the data.

-
-
-

Bibliography

Blizard D.A. (1992). Recombinant-inbred strains: general methodological considerations relevant to the study of complex characters. *Behav Genet* **22**, 621–633. [Reviews application of recombinant-inbred strains to gene mapping, investigates of pathways between genes and complex characters, and so on, stressing methodological considerations.]

Boyse E.A. (1977). The increasing value of congenic mice in biomedical research. *Lab Anim Sci* **27**, 771–781. [Explains the utilities of congenic strains and gives examples of important findings made possible by the use of such strains, e.g., influences of histocompatibility complex to leukemia.]

Bradley A., Zheng B., and Liu P. (1998). Thirteen years of manipulating the mouse genome: a personal history. *Int J Dev Biol* **42**, 943–950. [The authors, who established the production of germline chimaeras from cultured embryonic stem cells, describe the development and testing of concepts and techniques, and the reflection of this field.]

Carninci P., Shibata Y., Hayatsu N., Sugahara Y., Shibata K., Itoh M., Konno H., Okazaki Y., Muramatsu M., and Hayashizaki Y. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630. [This presents a method of preparing full-length cDNA libraries.]

de Angelis M.H. and Balling R. (1998). Large scale ENU screening in the mouse: genetics meets genomics. *Mut Res* **400**, 25–32. [Explains the necessity of large scale mouse mutant production by ethylnitrosourea (ENU) in the era of human and mouse genome sequence projects. Explains also the reason for choosing mice and ENU, the logistics of ENU mutagenesis, and currently known projects.]

Griffiths A.J., Gelbart W.M., Miller J.H., and Lewontin R.C. (1999). *Modern Genetic Analysis*. New York: W. H. Freeman. [A genetics textbook that covers genes, chromosomes, recombinant DNA technology, genomics, gene regulation, population genetics, evolutionary genetics, and quantitative genetics.]

Marshall E. (2000). Genomics: Public–private project to deliver mouse genome in six months. *Science* **290**, 242–243. [This reports the progress of mouse genome sequencing in several groups.]

McGhee J.D. (1995). Cell fate decisions in the early embryo of the nematode *Caenorhabditis elegans*. *Dev Genet* **17**, 155–166. [This is a summary of workings of the *Caernorhabditis elegans* early embryo and their cell fate decisions.]

Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J.L., Eisen M.B., Iyer V.R., Brown P.O., Muramatsu M., Shimada H., Okazaki Y., Hayashizaki Y. (2001). Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18 816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci.* **98**, 2199–2204. [This presents the analysis of full-length mouse cDNA arrays that reveal coordination of mRNAs expression within known biochemical pathways among different tissues.]

Muller U. (1999). Ten years of gene targeting: targeted mouse mutants, from vector design to phenotype analysis. *Mech Dev* **82**, 3–21. [This review summarizes various aspects of gene targeting in 1990s, including such techniques as basic construct design and analysis of complex phenotypes. Many examples from different areas of biomedical research are given to illustrate the purpose and limitations of the employed experimental approaches employed.]

Rubin G.M. and Lewis E.B. (2000). A brief history of Drosophila's contributions to genome research. *Science* **287**, 2216–2218. [This is a very short review of Drosophila biology, describing how it has led to major conceptual or technical breakthroughs in understanding of animal genomes.]

Rubin G.M., Yandell M.D., Wortman J.R., Gabor Miklos G.L., Nelson C.R., Hariharan I.K., Fortini M.E., Li P.W., Apweiler R., Fleischmann W., Cherry J.M., Henikoff S., Skupski M.P., Misra S., Ashburner M., Birney E., Boguski M.S., Brody T., Brokstein P., Celniker S.E., Chervitz S.A., Coates D., Cravchik A., Gabrielian A., Galle R.F., Gelbart W.M., George R.A., Goldstein L.S., Gong F., Guan P., Harris N.L., Hay B.A., Hoskins R.A., Li J., Li Z., Hynes R.O., Jones S.J., Kuehl P.M., Lemaitre B., Littleton J.T., Morrison D.K., Mungall C., O''Farrell P.H., Pickeral O.K., Shue C., Vosshall L.B., Zhang J., Zhao Q., Zheng X.H., and Lewis S. (2000). Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215. [A comparative analysis of the genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*, and the proteins they are predicted to encode was presented in the context of cellular, developmental, and evolutionary processes.]

Russell E.S. (1985). A history of mouse genetics. *Ann Rev Genet* **19**, 1–28. [This review starts from the very first stages of mouse genetic and explains major fields of mouse genetics, including establishing

strains, cancer, developmental biology, pleiotropism, threshold characters, histocompatibility, and sex determination.]

Russo S., Berkovitz Siman-Tov R., and Poli G. (1995). Yeasts: from genetics to biotechnology. *J Environ Pathol Toxicol Oncol* **14**, 133–157. [This review describes how yeast (*Saccharomyces cerevisiae*) has become useful organism for various aspects of sciences and industrial applications with the advances of biotechnologies.]

Silver L.M. (1995). *Mouse Genetics*. Oxford, UK: Oxford University Press. [A comprehensive textbook of mouse genetics, covering mouse strain origin, reproduction, breeding, mutagenesis, genomics, classical and modern mapping.]

The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. (2001). *Nature* **409**, 685–690. [This mainly describes mainly the results of a mouse full-length cDNA annotation meeting held at RIKEN.]

Zambrowicz B.P. and Friedrich G.A. (1998).: Comprehensive mammalian genetics: history and future prospects of gene trapping in the mouse. *Int J Dev Biol* **42**, 1025–1036. [The characteristics of the gene-trap method are described in order to help determine the role of the gene product in mammalian physiology and re relevance of gene product to human disease.]

Biographical Sketches

Naoki Osato graduated from the Department of Applied Biochemistry, Tsukuba University in 2000 with an M.Sc., and is currently a member of the Technical Staff at Genome Exploration Research Group, Genomic Sciences Center (GSC), RIKEN (The Institute of Physical and Chemical Research).

Hidenori Kiyosawa graduated from Department of Biology, University of Tsukuba, in 1984 with a B.Sc. (major in basic biology) and took his Ph.D. (in biology) at the Department of Biology, Utah State University, in 1994. He served with the Muscular Dystrophy Association (USA) as a postdoctoral fellow at the University of Pennsylvania/The Children's Hospital of Philadelphia, from 1994 till 1996; was a postdoctoral fellow at the National Institute of Neuroscience, from 1996 till 1997, and Assistant Professor at Department of Cell Science, Fukushima Medical University from 1997 till 1999. Currently, he is Senior Scientist at the Genome Exploration Research Group, RIKEN Genomic Sciences Center, and also works at the Graduate School Research Associate at Division of Genomic Sciences, Yokohama City University Graduate School of Integrated Science.

Jun Kawai graduated from Faculty of Science, Kyoto University (1986), was Research Scientist in Shionogi & Co., Ltd. (1986–1996), Senior Research Scientist (1997–2001) in RIKEN, and has been a Team Leader, at the Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC) since 1998.

Yoshihide Hayashizaki graduated from the Faculty of Medicine, Osaka University in 1986 as a Ph.D. He was Senior Research Scientist at the Institute for Molecular and Cellular Biology, Osaka University from 1986 to 1988, Research scientist at Department of Bioscience, National Cardiovascular Center Research Institute until 1992, and Research Scientist in Gene Bank, RIKEN Tsukuba Life Science Center until 1994. Currently, he is Chief Scientist, Department Director, and Chairman at the Genome Science Laboratory, RIKEN Tsukuba Life Science Center, and Project director of the Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC). He also serves as Professor at the Division of Genomic Sciences, Yokohama City University Graduate School of Integrated Science.