

# Taking a functional genomics approach in molecular medicine

Marie-Laure Yaspo

The elucidation of genetic components of human diseases at the molecular level provides crucial information for developing future causal therapeutic intervention. High-throughput genome sequencing and systematic experimental approaches are fuelling strategic programs designed to investigate gene function at the biochemical, cellular and organism levels. Bioinformatics is one important tool in functional genomics, although showing clear limitations in predicting *ab initio* gene structures, gene function and protein folds from raw sequence data. Systematic large-scale data-set generation, using the same type of experiments that are used to decipher the function of single genes, are being applied on entire genomes. Comparative genomics, establishment of gene catalogues, and investigation of cellular and tissue molecular profiles are providing essential tools for understanding gene function in complex biological networks.

The explosion of genomic information holds great promise for future developments in the field of molecular medicine. Many diseases have a genetic component, and understanding gene function provides insight into disease pathogenesis, the knowledge of which could be applied to the development of causal therapies. The human genome is estimated to contain ~35 000 protein-coding genes among which 60% have not been ascribed any functional attribute<sup>1,2</sup>; either a biochemical function (e.g. kinase), a cellular function (e.g. a specific signaling pathway), or a function at the organism level (e.g. brain development, immune response, etc.). Characterizing the function of a gene is not straightforward since it will depend on the molecular context within a given cell type and in a particular cellular microenvironment. Therefore, different and complementary experimental approaches should address the elucidation of gene function at those various levels. Functional genomics is a systematic effort to understand the function of genes and gene products by high-throughput analysis of gene products (transcripts, proteins) and biological systems (cell, tissue or organism) using automated procedures allowing to scale up experiments classically performed for single genes (e.g. generation of mutants, analysis of transcript and protein expression on a genome-wide basis, analysis of protein structure and protein-protein interactions, etc.).

Functional genomics can be conceptually divided in gene-driven and phenotype-driven approaches (Fig. 1). Gene-driven approaches use genomic information for identifying, cloning, expressing and characterizing genes at the molecular level. Phenotype-driven approaches analyze phenotypes from random mutation screens or naturally occurring variants (mouse mutants, human diseases) to identify and clone

the gene(s) responsible for the phenotype, without prior knowledge of the underlying molecular mechanisms. These two strategies are highly complementary at virtually all levels of analysis and lead collectively to the correlation of phenotypes with genotypes (Fig. 1). The novel dimension brought by systematic and global approaches is the unique potential for integrating and mining large data sets for obtaining insight into molecular networks associated with specific cellular processes. This review will focus on comparative genomics as an especially important component of functional genomics, because model organisms play a pivotal role in the functional characterization of genes and in the dissection of basic biochemical mechanisms *in vivo*. I will present which type of explicit and implicit functional information can be extracted from genome sequences, as well as from experiments carried out systematically in a high-throughput fashion, how this information can be linked to understanding the function of human genes by exploiting the respective features of model systems, and how monitoring of cellular and tissue phenotypes contribute to exploring biological networks at the molecular level. I will discuss the relevance of genome research for disease gene identification and the future progress in molecular medicine.

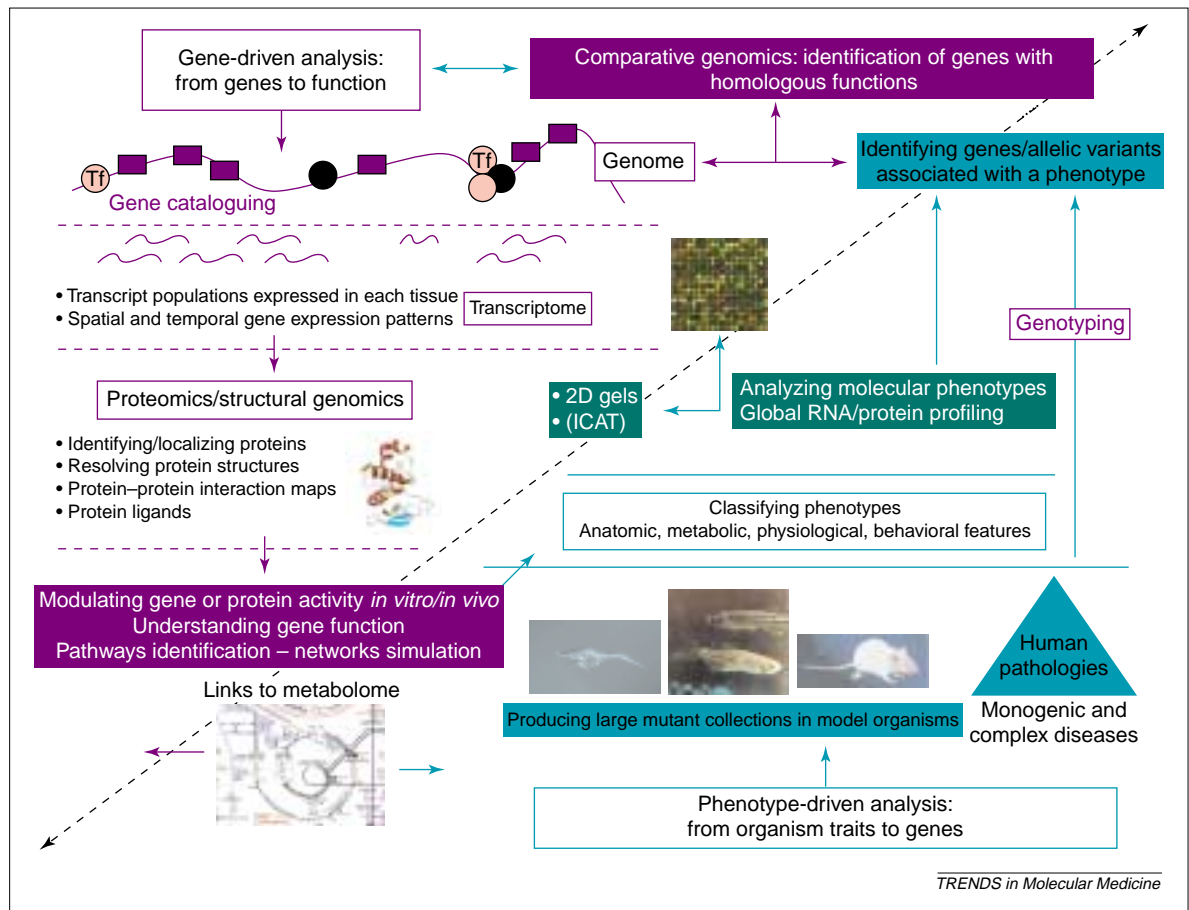
## Comparative genomics for understanding gene function

### *Genome analysis and sequence comparisons*

With >50 microbial genomes sequenced, the microbiology community has pioneered comparative genomics as a lead to deduce molecular mechanisms underlying physiological characteristics of different bacterial strains, and to identify essential genes involved in infection. Sequencing of leading pathogens, including that of *Streptococcus pneumoniae*<sup>3</sup>, enables the study of the biology of pathogens and host-pathogen relationships, knowledge that might contribute to future disease control. The comparative genomics concept applies just as well to eukaryotes, where it becomes particularly useful for studying reference organisms such as chicken, frog, fruit fly and also rodents that have been routinely used in developmental biology and classical genetics. Model organisms can be classified in two categories: (1) embryologically and/or genetically tractable organisms used for engineering and analyzing phenotypes and (2) genome models that are not necessarily amenable to experiments but which exhibit a compact genome and/or occupy a pivotal

Marie-Laure Yaspo  
Max Planck Institute for  
Molecular Genetics,  
Innestrasse 73, D-14195  
Berlin, Germany.  
e-mail:  
yaspo@molgen.mpg.de

Fig. 1. Overview of the two complementary strategies in functional genomics: the gene-driven and the phenotype-driven approaches, artificially separated along a diagonal axis. The main analysis tools and routes of investigation are represented together with the links between different levels of information.



position in the phylogenetic tree, hence contributing essential cues for understanding genome structure, function and evolution.

The 'Rosetta stone' analogy is a leading rationale in comparative genomics, enabled by completely sequenced genomes and EST collections that are now available for several metazoans (Fig. 2). Systematic sequence analysis allows the identification of HOMOLOGOUS and ORTHOLOGOUS GENES for establishing molecular links between those different model systems. Emerging gene catalogues established for different organisms, stating the number and nature of protein-coding sequences (Fig. 2), are precious tools for the global analysis of molecular events. The remarkable conservation of basic developmental programs during evolution allows us to infer gene function between distantly related species, based on sequence homologies and experimental cues. Whole-genome comparisons of yeast, fruit fly, worm and human helped to classify phylogenetically related proteins into functional categories based on groups of orthologous genes<sup>4,5</sup>. Yet, it has to be noted that complete structural and functional correspondence of orthologs does not necessarily apply to genes encoding multi-domain proteins<sup>6</sup>. *Drosophila* and principally vertebrates have evolved more complex domain organization than worm or fungi<sup>7</sup>. There are close to 2030 human-worm orthologs, 2700 human-fly orthologs, and 1500 human proteins with strict orthology in both fruitfly and worm<sup>1,2,7</sup>. Conserved

function across phyla is validated by functionally interchangeable proteins (e.g. human *OTX* genes can rescue cephalic defects in *Drosophila otd* mutant<sup>8</sup>).

*Comparative genomics to study human disease genes*  
 Nearly 62% of known human disease proteins and 68% of cancer genes surveyed have a counterpart in the fruit fly and/or worm<sup>7,9</sup>. These mostly represent classes of disease genes for cancers, neurological and metabolic conditions. This is illustrated, for example, by the sonic hedgehog signaling pathway, which is essential to embryo patterning and neural development from fly to vertebrates, but is also involved in holoprosencephaly, a genetic defect affecting craniofacial development<sup>10</sup>, and found active in several types of cancers<sup>11</sup>. There is a core of 1400 gene clusters with homologs found across yeast, worm and fly and mammals<sup>1,5</sup>. Those 'core proteins' are mainly enzymes from the transcriptional machinery, intermediary metabolism, cell cycle regulators, protein transport and trafficking, pointing to essential cellular processes. Human diseases caused by defects in basic cellular processes are likely to involve enzymatic or morphogenetic pathways carried out by several of these 'core proteins'. By contrast, cardiac, immunological and endocrine disease genes are underrepresented among genes conserved in invertebrates, as a reflection of more specialized processes typical for vertebrates. Protein repertoires reflect physiological functions taking

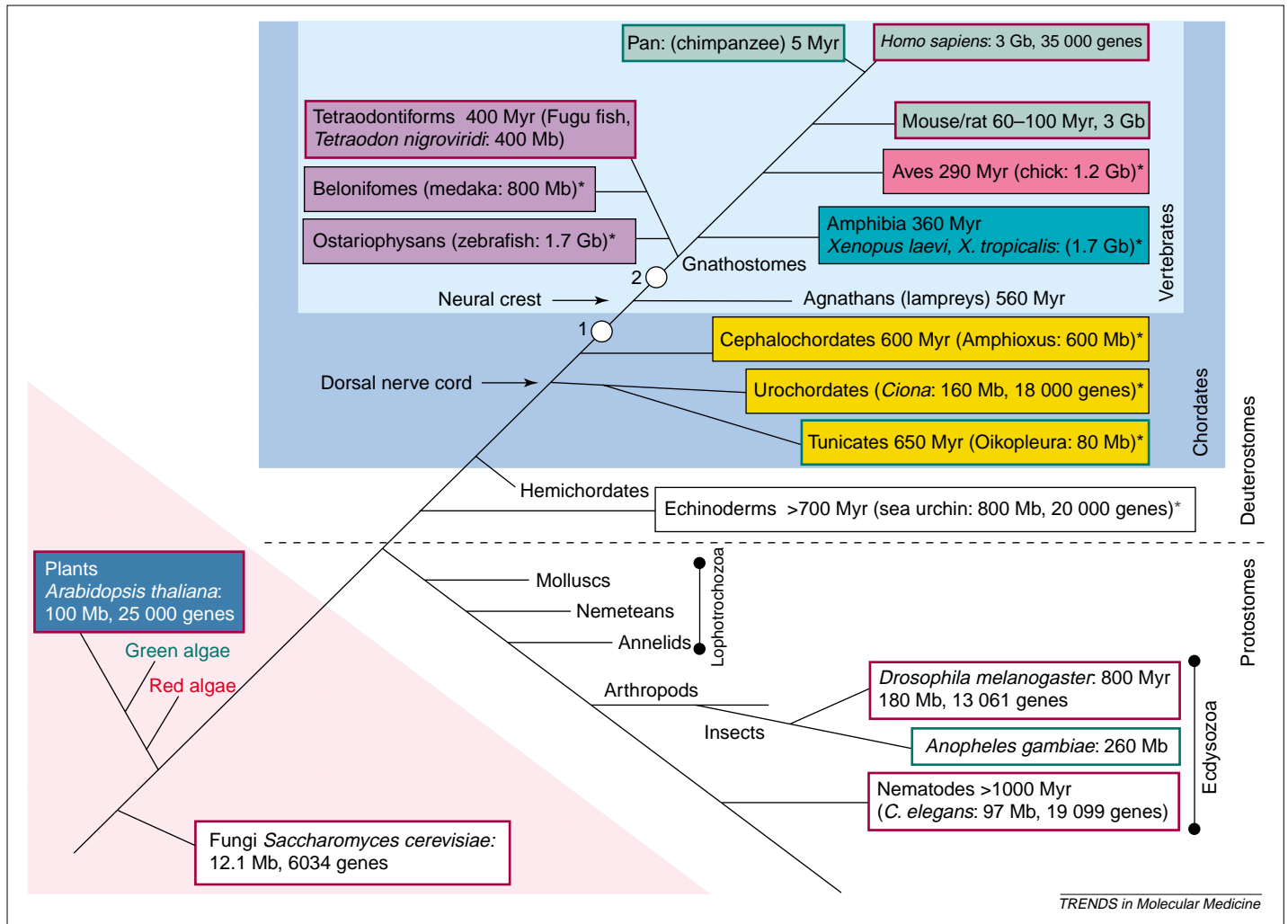


Fig. 2. Status of sequence resources for metazoans represented on an evolutionary tree drawn from compiled information sources. Estimated time indicating the separation between a given species and humans is in million years (Myr). White dots along the tree axis indicate the presumed first and second genome duplication, respectively. Genome sizes are given in megabases (Mb) or gigabases (Gb). Red frames indicate species for which complete genome sequence and ESTs are available. Green frames indicate genomes currently being sequenced. Asterisks indicate species with ESTs resources, but no complete genome sequence. A list of sequenced genomes is available at the Kyoto Encyclopedia of Genes and Genomes (KEGG) (<http://www.genome.ad.jp/kegg/kegg2.html>) and at National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov>).

place in a given living system. Strikingly, humans have less than twice as many genes than nematodes demonstrating that organism complexity does not reside merely within the gene number<sup>12</sup>. Vertebrates display an enlarged repertoire of cellular functions, achieved in part by protein domain accretion but also by a combination of molecules and interactions generated by mechanisms controlling gene expression, RNA splicing and protein processing.

#### Genome models for comparative genomics

The study of 'genome model' sequences should contribute simultaneously to the understanding of gene function in complex chordates such as vertebrates as well as in protostomes (Fig. 2). Chordates are characterized by a notochord, pharyngeal clefts, and a dorsal nerve tube developing into brain and spinal cord in higher chordates. *Oikopleura*, a small marine

plankton with the smallest known genome of the chordate phylum which has a significant biological value for identifying markers specifying the chordate lineage<sup>13</sup>, provides a missing molecular link at the hemichordate–chordate transition. *Amphioxus*, a cephalochordate with a body plan arranged in segmental blocks, will give clues to vertebrate-specific molecular events that occurred before genome duplications (Fig. 2). The pufferfish *Fugu rubripes* has become a popular model because of its small, compact genome and a gene repertoire similar to that of human<sup>14</sup>. The sequence of a closely related fish, *Tetraodon nigroviridis*, was used to establish a novel method for predicting the number of genes in the human genome<sup>15</sup>. Uniquely human traits in terms of anatomy, cognition, behavior and disease susceptibility can partly be understood by molecular analysis of its closest kin, the chimpanzee. For example, chimps are not prone to developing neurodegenerative diseases linked to triplet expansion<sup>16</sup>. Genome models placed at key evolutionary transitions will help to identify noncoding regulatory elements in each lineage, and provide essential information for a functional interpretation of the human genome. Hypotheses formulated by genomic studies could be then tested in experimentally more amenable model organisms.

### Relating protein repertoires and function in invertebrates

One important question is how genome information and knowledge of protein repertoires can be exploited for inferring gene function from model organisms to humans. For example, neurological conditions are of particular interest to molecular medicine and can be modelled in different systems. With over 250 000 neurons and a fairly elaborate brain structure, *Drosophila* can be used as a 'test tube' for genetic manipulation of neurological phenotypes and for identifying protein partners and biochemical pathways participating in neurodegeneration processes. Knock-in mutant forms of human  $\alpha$  synuclein (SNCA) expressed in *Drosophila* trigger adult-onset loss of dopaminergic neurons, recapitulating anatomical and functional deficits of familial Parkinson's disease<sup>17</sup>. The fruit fly is a flexible model with an interesting potential for identifying drug targets and testing compounds acting on neurotoxic mutations. *Drosophila* and *C. elegans* share a conserved repertoire of neuronal signaling molecules also present in mammals<sup>7</sup>. With close to 19 000 genes, *C. elegans* has only 302 neurons among 800 cells. The expansion of specific classes of protein families found in worm but not in the fly includes voltage-gated potassium channels, serine/threonine protein kinases, G-protein-coupled receptors, histones H3/H4, extracellular matrix adhesion molecules, nicotinic acetylcholine receptors and GABA receptors, and olfactory receptors<sup>2</sup>. In nematodes, sensory and neuronal connections are thought to be established according to a different mode of complexity than the intricate wiring seen in fly. Those differences argue for using more than one model organism for studying gene function, taking advantage of different cellular and biochemical contexts *in vivo*. As a simple organism, *C. elegans* is also used for studying pathways associated with longevity and ageing<sup>18</sup>. An increasingly popular model for studying the central nervous system is *Ciona intestinalis* (Fig. 2), a simple marine invertebrate chordate for studying embryonic patterning and gene ancestors of the vertebrate lineage<sup>19</sup>. With roughly 15 000 genes, *Ciona* has 80 neurons among 2500 cells, and the simplicity of this system allows precise tracing of certain markers of neuronal development despite the lack of elaborate anatomical structures. Moreover, simple organisms with a short life cycle are economically more feasible in high-throughput mutagenesis, drug screening and therapeutic target discovery<sup>20</sup>.

### Considerations and limitations

Bottlenecks of comparative genomics are caused by bioinformatics and biological considerations. First, the concept of comparative genomics relies upon the knowledge of predicted protein complements of sequenced genomes and the outcome depends heavily on accurate genome annotation. Bioinformatics tools such as exon-prediction programs show pitfalls in the *ab initio* identification of gene structures. Gene extremities and exon-intron boundaries are often

incorrectly defined, and alternatively spliced isoforms can currently not be predicted correctly. Large genes consisting of small exons spread over large DNA segments are also difficult to tackle. These shortcomings are evident for large vertebrate genomes. As yet, the human genome is only partially annotated and the mammalian gene catalogue is incomplete (24 000 proteins in the ENSEMBL human set). The identification of conserved regions between complete genomes obtained by comparing whole translated sequences could identify genes omitted by these programs, as shown in a recent re-annotation of the *Drosophila* genome<sup>21</sup>.

Second, the notion of gene orthology from simple organisms to vertebrates is not always straightforward (e.g. after two potential genome duplications; see Fig. 2) and 'function' can be erroneously inferred from genes that are structurally related but functionally divergent. Conversely, sequence divergence during evolution might obscure the detection of structural homologies between proteins that are indeed functionally related. Despite spectacular advances in our knowledge of the gene catalogue, bioinformatics alone cannot predict a phenotype solely on the knowledge of an ancestral gene function. Experimentally based gene-driven and phenotype-driven approaches need to be applied simultaneously for filling gaps in our current knowledge (Fig. 1).

Other limitations of data transferability from invertebrates to humans become apparent for modelling complex human phenotypes, where mouse models will be more appropriate. Bottlenecks can also occur for studying major functional classes of proteins that did undergo creation or large expansion in humans. Those include mostly proteins involved in immunity [e.g. intercalins (a family of genes acting as stimulating factors), interleukins, T-cell receptors], specialized neurological function (e.g. myelin, nerve growth factor) and homeostasis (e.g. glucagon, calcitonin). Yet, invertebrates are helpful for expressing these genes ectopically for unmasking possible secondary functions or for studying dominant-negative effects with knock-in experiments. The comparative genomics concept has its strengths and limitations, but it is nonetheless a unique systematic route for deciphering gene function and dissecting biochemical mechanisms *in vivo*. Gene function can be addressed by gene inactivation, mutation, overexpression or ectopic expression methodologies applied in different model species.

### Systematic generation of phenotypes in invertebrate and vertebrate models

In many respects the most interesting aspect of the function of a gene is its function in the organism as a whole, requiring either the mutations of specific genes, followed by the analysis of the phenotype of the organisms carrying this mutation (gene-driven paradigm), or the identification of organisms with an interesting phenotype followed by the mapping of mutated genes (phenotype-driven paradigm). It was

previously estimated that about one third of genes in yeast, worm, fly and mice is essential for viability<sup>22</sup>. It is likely that lethal mutants hit a significant fraction of core conserved genes. Generating phenotypes in experimentally tractable organisms and linking these phenotypes to genes is crucial for ascribing a biological role to genes *in vivo*. In *Drosophila*, 20% of the genes generate observable phenotypes whereas in yeast, 50% of disruptions lead to phenotypes<sup>22</sup>. In mouse, gene targeting methods, despite their exquisite potential for addressing gene function in defined contexts and for rescuing lethal phenotypes, have hit at most a few thousand loci, representing only a minute fraction of the 35 000 genes to be analyzed. Systematic mutagenesis strategies by gene-driven and phenotype-driven approaches aim at saturating genomes for creating extensive collections of phenotypes, as discussed by Janet Rossant and Colin McKerlie (see page 502).

#### *Gene-driven approaches*

Gene-driven knockout strategies generally require prior knowledge of genomic information of the targeted loci and are well suited to functional testing of novel predicted genes. RNA INTERFERENCE (see Glossary) (RNAi) induces specific gene silencing at the post-transcriptional level upon introduction of double-strand RNA, and was reported to be efficient in worm and fly<sup>23</sup>, and early mouse embryogenesis<sup>24</sup>, but not in zebrafish<sup>25</sup>. Systematic RNAi in *C. elegans* indicated that close to 14% of tested loci produced phenotypes, whereas 60% of these were embryonic lethal<sup>26</sup>. Interestingly, this data set reveals that genes that are conserved and expressed at higher levels are more prone to exhibiting a phenotypic effect. A large fraction of genes of unknown function produced phenotypes in late developmental processes, suggesting that they might play a more specialized role, in contrast to genes involved in basic metabolic machinery often required for embryonic viability. Antisense morpholino-based gene targeting<sup>27</sup> is a powerful tool for generating phenotypes, by specific inhibition of gene translation, and has been shown to be effective in both invertebrates and vertebrates. This method is expensive but is flexible, and might overcome tissue-restricted limitations encountered with RNAi, particularly apparent with neurons. In zebrafish, compelling parallels with human phenotypes for uroporphyrinogen decarboxylase and holoprosencephaly were generated with morpholinos<sup>28</sup>. The forthcoming zebrafish genomic sequence and EST resources developed for the diploid frog *Xenopus tropicalis* will pave the way for systematic knockout schemes in non-mammalian vertebrates. Taken together with classical gene manipulation technologies, an arsenal of tools is available for modulating gene activity *in vivo*. By using these technologies and the existing knowledge of protein sets, gene inactivation strategies could be combined to target entire generic classes of proteins, or to inactivate several genes in concert to trace multiple-hit effects or

compensation mechanisms. In the mouse, large-scale gene-trap insertion enterprises have already generated collections of archived mutant embryonic stem (ES) cells<sup>29</sup>. In this approach, the identity of the locus of interest can be directly assessed by reverse transcription (RT)-PCR and sequence analysis, a capital advantage over chemical mutagenesis. It is as yet unknown how many and which types of phenotypes can be generated from gene-trap strategies.

#### *Phenotype-driven approaches*

So far, yeast, zebrafish and mouse have been used for phenotype driven studies. An impressive screen performed in zebrafish using N-ethyl-N-nitrosourea (ENU) mutagenesis has revealed a large panel of recognizable mutants that have been classified and complemented, revealing several functional pathways<sup>30</sup>. However, the hurdle awaiting chemical mutagenesis screens is the very laborious mapping of mutated genome loci requiring high-throughput mapping technologies. For the zebrafish screen, only 10% of the phenotypes have been mapped to date<sup>31</sup>. In mammals, large-scale genome-wide ENU mutation programs in mouse complement gene-traps strategies<sup>32</sup> (see Janet Rossant and Colin McKerlie, page 502), where ENU screens recover typically 1.5–2 % of inherited dominant mutations<sup>33,34</sup>, revealing dysmorphic defects, behaviour abnormalities, hematologic and immunoglobulin defects<sup>34</sup>. Extrapolating to the number of known dominant mutations reported in humans (close to 3000, OMIM), screening of more than 200 000 mice would be necessary to generate a comparable phenotype collection, although recessive effects will be overlooked, despite efforts at crossing ENU mouse lines.

#### *Comparative mapping and chromosome engineering*

Strategies based on comparative mapping are restricted to vertebrates and are widely used for human-mouse maps. Comparative maps are built upon long-range chromosomal SYNTENY, reflected by a near perfect conserved gene content and order within paralleled genomic regions<sup>35</sup>. Positional cloning by homology is instrumental for disease gene identification. A human gene can be inferred as a candidate for a human disease when an explicit mouse mutant homolog associated with a comparable phenotype could be mapped. Homology mapping was used successfully for isolating factors involved in diabetes, hypertension and obesity<sup>35</sup>. The mouse sequence will expedite positional cloning and will delineate precise DNA boundaries of syntenic regions. Comparative maps support chromosome engineering in the modelling of human aneuploidy phenotypes, mainly developmental defects and cancer, using the Cre-loxP technology for creating chromosome deletions, duplications and inversions<sup>36</sup>. Models of DiGeorge syndrome in mice<sup>37</sup> are preludes to generating random deletions in the mouse genome. One could envisage creating

duplications recapitulating gene dosage effects in trisomy 21 that would complement current mouse models of Down syndrome<sup>38</sup>.

#### Monitoring of cellular and tissue phenotypes

*Gene expression patterns by in situ hybridization*  
Systematic functional screens by *in situ* hybridization (ISH) reveal temporal and spatial gene expression patterns, which can contribute to the discovery of novel gene functions and *SYNEXPRESSION* groups leading to pathway identification. In gene-driven approaches, study of embryonic patterning using collections of random cDNA provides a direct link between DNA sequences and endogenous gene expression patterns. High-throughput whole-mount ISH developed for *X. laevis* and mouse embryos showed that roughly 18–30% of randomly chosen genes revealed a regionalized pattern<sup>39,40</sup>. Systematic efforts are currently ongoing in mouse with the aim of establishing an atlas of gene expression patterns for all 35 000 mammalian genes. This resource will be extremely valuable for attributing topological information to microarray data or protein–protein interaction maps, and for identifying markers for specific cell types.

#### *Gene expression profiles for revealing molecular basis of phenotypes*

The analysis of gene expression patterns can contribute to the understanding of complex molecular pathways. Arrays are leading technologies in this field, using target molecules (oligonucleotides or cDNA fragments) immobilized on a surface (e.g. nylon, treated glass)<sup>41–43</sup>. Typically, hybridization of labeled complex probes onto arrays generates a global qualitative snapshot of the corresponding transcript population. Quantitative measurements of expression levels are only relative to that of the chosen reference *TRANSCRIPTOME*. The selection of immobilized target genes and the choice of controls are crucial in the experimental design. Computational data analysis methods and clustering algorithms need to be optimized for each project to extract a maximum of relevant information<sup>44,45</sup>, however mining algorithms are still in development. Cellular programs as well as subverted mechanisms associated to diseases can be partly revealed and predicted by analyzing global gene expression profiles surveying the whole genome. The concept relies upon the monitoring of gene expression levels associated with a given phenotype compared to that of a reference state (e.g. cell type, developmental stage, healthy control individuals)<sup>41</sup>. The most immediate application of transcriptome surveys in molecular medicine is the establishment of cancer-specific expression profiles. Molecular classification of phenotypes can distinguish cancer subtypes that are histologically similar and indicate potential markers of disease progression<sup>46</sup>. Molecular phenotyping is therefore valuable for stratifying patient populations. Classification and diagnostic prediction of several childhood cancer types was achieved by gene

expression profiling and artificial neural networks<sup>47</sup>. Tracing the histological origins of cancer cell lines from a chip-based transcriptome survey of 8000 genes linked coherent gene clusters to biological properties shared by different cell types<sup>48</sup>, whereas the analysis of drug responses built a molecular pharmacology catalogue<sup>49</sup>. This type of study could be seminal for simulating and predicting drug response in clinical assays, raising hopes for adopting personalized treatments to defined patient populations. In the search of human disease genes, molecular profiling of patient samples contribute to the identification of candidate and modifier genes. Arrays are now being applied to stem cell research, as well as being used to study the host response to microbial pathogens. The response of human respiratory epithelial cells to *Bordetella pertussis* indicated key strategic defense molecules including cytokines and mucin<sup>50</sup>.

Ideally, arrays should contain the full gene complement of a given genome when the point of array-based approached is to survey global transcriptomes. The yeast *S. cerevisiae*<sup>51</sup> and *C. elegans*<sup>52</sup> are the only species for which diverse cellular processes have been analyzed on nearly complete gene arrays. Gene catalogs are the basis for designing comprehensive arrays. Most arrays are designed with partial gene collections, because gene catalogues and EST collections are incomplete. Comprehensive arrays should be soon available for different organisms, including human and mouse.

#### Proteome maps and phenotype analysis

Protein processing and modification can not be predicted from genomic information. In a given cell at a certain time point, up to 10 000–20 000 transcripts and 50 000–80 000 proteins could be expressed simultaneously, given estimates of about four isoforms per gene. Transcription levels do not necessarily correlate with those of proteins, particularly for rarely expressed genes<sup>53</sup>. In theory, array experiments should be mirrored with proteomics data for a full understanding of physiological networks. Complex protein mixtures can be resolved as a constellation of up to 10 000 spots by 2D-gel electrophoresis technology, but will survey only the most abundantly expressed and hydrophilic proteins<sup>54</sup>. Using mass spectrometry analysis, peptide spots can be related with gene identity by matching their spectra to database entries, permitting the building of protein maps<sup>55,56</sup>. The technique can be used to analyze and compare cell and disease phenotypes<sup>57,58</sup>. This sophisticated approach remains the unique way for generating qualitative protein profiles but is not in routine use, because it is technically demanding and associated with problems of reproducibility. Analogous to array profiling, novel mass spectrometry-based technologies using isotope-coded affinity tags (ICAT) should allow the comparison of relative protein expression levels between two complex samples, thus bypassing 2D polyacrylamide gel electrophoresis

(PAGE)<sup>59</sup>. Initially tested in yeast, this technique is currently in development. When adapted to high-throughput analysis of complex samples, global proteomics methods, including the development of protein chips, will offer a breakthrough for analyzing molecular profiles of human and animal phenotypes. Databases for 2D-PAGE are collected in databases (<http://www.expasy.ch/ch2d/>) providing also tissue proteome map information.

#### Proteomics and structural genomics

Structural information can provide a lead for protein function and drug action at the molecular level<sup>60</sup>. The Protein Data Bank<sup>61</sup> currently lists 14 150 protein structures, among which close to 3000 match the keyword 'human'. World-wide efforts developing high-throughput structural genomics operations are setting up remarkably quickly (<http://www.rcsb.org/pdb/strucgen.html>). Structural genomics initiatives aim at resolving experimentally 10 000 protein structures within the next ten years, which would represent all existing domain folds, by using nuclear magnetic resonance (NMR) and X-ray diffraction. Automation of expression and crystallization procedures and development of NMR<sup>62</sup> are required to make this possible (currently ~40–50 structures can be obtained from 1000 random cDNAs). The fraction of proteins that are refractory to this procedure could be minimized by using sequence information (e.g. sorting hydrophobic segments, domain classification). In theoretical approaches using genomic data, *ab initio* protein modeling can predict potential structures of ~20% of proteins by building a model based upon recognizable folds<sup>63</sup>.

#### Protein–protein interaction maps

The exploration of biological pathways and cellular functions builds upon knowledge of protein–protein interactions. Vertebrates made wide use of protein domain shuffling, increasing combinatorial diversity and the capacity to mediate protein–protein interactions, by adapting ancestral function or creating novel ones. The current gold standard for systematic search of protein–interaction partners is the yeast two-hybrid system<sup>64</sup>, whereby unknown partners (preys) can be screened against a bait library. Despite the occurrence of false positive (illegitimate interactors) seen with the method, the yeast two-hybrid method has been carried out on a large scale for *S. cerevisiae*, *C. elegans*, vaccinia virus and *H. pylori* (reviewed in Ref. 65). A database of protein–protein interactions is publicly available<sup>66</sup>. In yeast, several major large-scale efforts have reported more than 2500 interactions, providing a rich data resource that was mined for predicting functional networks<sup>67</sup>. This approach led to a proposed functional category for more than 350 uncharacterized proteins and revealed cross-talks between functional groups. Interestingly, only one third of assigned function were similar to those suggested based upon RNA expression profiles

and genome properties in an independent study<sup>67,68</sup>. Initially pioneered in yeast, protein–interaction maps are still in their infancy for metazoans.

#### Lessons from model organisms and remaining challenges

Non-mammalian models will remain essential for tackling gene function in a systematic fashion, providing different contexts to assign function to many predicted genes that might have specialized or pleiotropic roles. Despite their value, mouse models are somewhat limiting owing to the relative paucity of mouse knockouts that give tractable phenotypes. It will be technically difficult and costly to generate most genes in the mouse genome and to study phenotypic outcomes even with standardized assessment protocols and archiving. In gene-driven approaches, stable inheritance of phenotypes does not occur in RNAi and morpholinos methodologies, and further developments are necessary in that direction. Considering the number of human orthologs in invertebrates and the percentage of lethal loci and overt phenotypes obtained by systematic mutagenesis in mouse and other models, current data sets led to speculate that at most 40% of the human genes can be analyzed from mutant collections. Functional information for the other 60% of the human genome will originate from molecular profiling at the transcriptome and PROTEOME levels, and from knowledge of protein structures and protein–interaction maps. A large part of those 60% are as yet anonymous genes for which identification of functional attributes at any level could help to decipher their biological role, justifying the necessity of integrating data from various systematic approaches.

#### Concluding remarks

This review explores highlights of ongoing efforts in functional genomics and illustrates how these approaches can be applied in biomedical research. Collectively, functional genomics approaches provide a matrix of information, allowing the 'tagging' of gene products with functional attributes and to reveal physiological pathways. Large data sets need to be archived, annotated, analyzed and made publicly available in a standardized and usable form to be best exploited. Integration with SNP information and allelic variants in patient populations will be particularly useful. Functional genomics research forces scientists to formulate new questions that are gradually changing the way experiments are being designed. This approach will lead to novel discoveries and avoids bias towards *a priori* interesting genes, and thus will be of particular importance in novel therapeutic target discovery. The impact of genome research for therapeutic target discovery is hard to evaluate at present but might be in the range of several thousands, based on sequence homologies with known targets and considering the number of genes encoding kinases or secreted molecules. Current drug therapy addresses about 500 drug targets, mostly

represented by cell membrane receptors (e.g. G-protein-coupled receptors) and enzymes<sup>69</sup>. Therapeutic utility of pharmaceuticals targeting nuclear receptors (2% of targets) such as glucocorticoid, retinoic acid, estrogen and progesterone receptors suggests possibilities for targeting other classes of transcription factors whose inappropriate activity leads to a wide range of human pathologies<sup>70</sup>.

Understanding disease mechanisms at the molecular level is an informative source of potential targets that might indicate novel entry points for pharmaceuticals. Prospects of functional genomics raises enthusiasm and skepticism, but there is little doubt that we are marching towards a new era of biology that will revolutionize our understanding of disease processes, and create openings towards causal therapies.

#### References

- Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
- Martin, A.C. *et al.* (1996) Analysis of the complete nucleotide sequence and functional organization of the genome of *Streptococcus pneumoniae* bacteriophage Cp-1. *J. Virol.* 70, 3678–3687
- Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
- COPSE: A platform for reconstructing vertebrate phylogeny ([www.dkfz.de/tbi/services/copse/form](http://www.dkfz.de/tbi/services/copse/form))
- Koonin, E.V. *et al.* (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101, 573–576
- Rubin, G.M. *et al.* (2000) Comparative genomics of the eukaryotes. *Science* 287, 2204–2215
- Nagao, T. *et al.* (1998) Developmental rescue of *Drosophila* cephalic defects by the human *Otx* genes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3737–3742
- Banfi, S. *et al.* (1996) Identification and mapping of human cDNAs homologous to *Drosophila* mutant genes through EST database searching. *Nat. Genet.* 13, 167–174
- Ming, J.E. *et al.* (1998) Human developmental disorders and the Sonic hedgehog pathway. *Mol. Med. Today* 4, 343–349
- Dahmane, N. *et al.* (1997) Activation of the transcription factor Gli1 and the Sonic hedgehog signalling pathway in skin tumours. *Nature* 389, 876–881
- Claverie, J.M. (2001) Gene number. What if there are only 30 000 human genes? *Science* 291, 1255–1257
- Spada, F. *et al.* (2001) Molecular patterning of the oikoplasmic epithelium of the larvacean tunicate *Oikopleura dioica*. *J. Biol. Chem.* 276, 20624–20632
- Brenner, S. *et al.* (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 366, 265–268
- Roest Crollius, H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238
- Djian, P. *et al.* (1996) Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. *Proc. Natl. Acad. Sci. U. S. A.* 93, 417–421
- Feany, M.B. and Bender, W.W. (2000) A *Drosophila* model of Parkinson's disease. *Nature* 404, 394–398
- Antebi, A. *et al.* (1998) daf-12 regulates developmental age and the dauer alternative in *Caenorhabditis elegans*. *Development* 125, 1191–1205
- Hudson, C. and Lemaire, P. (2001) Induction of anterior neural fates in the ascidian *Ciona intestinalis*. *Mech. Dev.* 100, 189–203
- Link, E.M. *et al.* (2000) Therapeutic target discovery using *Caenorhabditis elegans*. *Pharmacogenomics* 1, 203–217
- Gopal, S. *et al.* (2001) Homology-based annotation yields 1042 new candidate genes in the *Drosophila melanogaster* genome. *Nat. Genet.* 27, 337–340
- Miklos, G.L. and Rubin, G.M. (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell* 86, 521–529
- Kennerdell, J.R. and Carthew, R.W. (2000) Heritable gene silencing in *Drosophila* using double-stranded RNA. *Nat. Biotechnol.* 18, 896–898
- Wianny, F. and Zernicka-Goetz, M. (2000) Specific interference with gene function by double-stranded RNA in early mouse development. *Nat. Cell Biol.* 2, 70–75
- Zhao, Z. *et al.* (2001) Double-stranded RNA injection produces nonspecific defects in zebrafish. *Dev. Biol.* 229, 215–223
- Fraser, A.G. *et al.* (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* 408, 325–330
- Summerton, J. and Weller, D. (1997) Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev.* 7, 187–195
- Nasevicius, A. and Ekker, S.C. (2000) Effective targeted gene 'knockdown' in zebrafish. *Nat. Genet.* 26, 216–220
- Wiles, M.V. *et al.* (2000) Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nat. Genet.* 24, 13–14
- Haffter, P. *et al.* (1996) The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development* 123, 1–36
- Talbot, W.S. and Hopkins, N. (2000) Zebrafish mutations and functional analysis of the vertebrate genome. *Genes Dev.* 14, 755–762
- Brown, S.D. and Nolan, P.M. (1998) Mouse mutagenesis-systematic studies of mammalian gene function. *Hum. Mol. Genet.* 7, 1627–1633
- Nolan, P.M. *et al.* (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat. Genet.* 25, 440–443
- Hrabe de Angelis, M.H. *et al.* (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat. Genet.* 25, 444–447
- O'Brien, S.J. *et al.* (1999) The promise of comparative genomics in mammals. *Science* 286, 458–462, 479–481
- Mills, A.A. and Bradley, A. (2001) From mouse to man: generating megabase chromosome rearrangements. *Trends Genet.* 17, 331–339
- Lindsay, E.A. *et al.* (1999) Congenital heart disease in mice deficient for the DiGeorge syndrome region. *Nature* 401, 379–383
- Reeves, R.H. *et al.* (2001) Too much of a good thing: mechanisms of gene action in Down syndrome. *Trends Genet.* 17, 83–88
- Neidhardt, L. *et al.* (2000) Large-scale screen for genes controlling mammalian embryogenesis, using high-throughput gene expression analysis in mouse embryos. *Mech. Dev.* 98, 77–94
- Gawantka, V. *et al.* (1998) Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mech. Dev.* 77, 95–141
- Gress, T.M. *et al.* (1996) A pancreatic cancer-specific expression profile. *Oncogene* 13, 1819–1830
- Duggan, D.J. *et al.* (1999) Expression profiling using cDNA microarrays. *Nat. Genet.* 21, 10–14
- Eickhoff, H. *et al.* (2000) Tissue gene expression analysis using arrayed normalized cDNA libraries. *Genome Res.* 10, 1230–1240
- Steinfath, M. *et al.* (2001) Automated image analysis for array hybridization experiments. *Bioinformatics* 17, 634–641
- Claverie, J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* 8, 1821–1832
- Welsh, J.B. *et al.* (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* 98, 1176–1181
- Khan, J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673–679
- Ross, D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24, 227–235
- Scherf, U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244
- Belcher, C.E. *et al.* (2000) From the cover: the transcriptional responses of respiratory epithelial cells to *Bordetella pertussis* reveal host defensive and pathogen counter-defensive strategies. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13847–13852
- DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
- Jiang, M. *et al.* (2001) Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 218–223
- Anderson, L. and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537
- Klose, J. (1999) Large-gel 2-D electrophoresis. *Methods Mol. Biol.* 112, 147–172
- Klose, J. and Kobalz, U. (1995) Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis* 16, 1034–1059

- 56 Mollenkopf, H.J. *et al.* (1999) A dynamic two-dimensional polyacrylamide gel electrophoresis database: the mycobacterial proteome via Internet. *Electrophoresis* 20, 2172–2180
- 57 Klose, J. (1999) Genotypes and phenotypes. *Electrophoresis* 20, 643–652
- 58 Chambers, G. *et al.* (2000) Proteomics: a new approach to the study of disease. *J. Pathol.* 192, 280–288
- 59 Gygi, S.P. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999
- 60 Skolnick, J. *et al.* (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.* 18, 283–287
- 61 Berman, H.M. *et al.* (2000) The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* 7 (Suppl.) 957–959
- 62 Heinemann, U. (2001) The Berlin 'protein structure factory' initiative: a technology-oriented approach to structural genomics. *Ernst Schering Res. Found Workshop* 34, 101–121
- 63 Kolinski, A. *et al.* (2001) Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 44, 133–149
- 64 Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246
- 65 Legrain, P. *et al.* (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.* 17, 346–352
- 66 Xenarios, I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291
- 67 Schwikowski, B. *et al.* (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261
- 68 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
- 69 Drews, J. (2000) Drug discovery: a historical perspective *Science* 287, 1960–1964
- 70 Emery, J.G. *et al.* (2001) Therapeutic modulation of transcription factor activity. *Trends Pharmacol. Sci.* 22, 233–240

# Mouse-based phenogenomics for modelling human disease

Janet Rossant and Colin McKerlie

The powerful and wide-ranging genetic tools available in the laboratory mouse make it the major experimental model for studying mammalian gene function *in vivo* and modelling human disease traits. Large-scale random mutagenesis approaches, either gene-driven or phenotype-driven, promise to identify new clinically relevant phenotypes and their associated genes. Development of appropriate tools for assessing clinical phenotypes in mice is a crucial component of these endeavours, as is the establishment of the infrastructure for archiving and distribution of the growing mutant resource to the community. Integrated, multidisciplinary programs will be needed to fully exploit the power of the mouse in molecular medicine.

The completion of the sequence of the human genome emphasizes the gap between the accumulation of sequence information and the application of that information to the understanding of human disease. POSITIONAL CLONING (see Glossary) of diseases caused by single gene mutations is proceeding apace, but the genetic underpinnings of the more common, complex diseases are still poorly understood. The mouse is a useful experimental system to study the genetics and pathobiology of human disease, because its genes, as well as its biochemical pathways, physiological and organ functions are closely related to humans, and because it offers unparalleled opportunities for genome manipulation. Its genome sequence is currently available in draft form in the public domain (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>) and in an assembled version in the private domain (<http://www.celera.com>). Mutagenesis in mice provides powerful tools for interrogating the genome for information relevant to human disease mechanisms, prognosis and therapy. Many genetic alterations in mice give remarkably similar clinical manifestations to their human counterparts and can

provide direct models for studying disease mechanisms and testing potential therapies. In other cases, the mouse and human mutant phenotypes are different; those differences themselves can be informative in understanding disease mechanisms. There is thus an urgent need to develop tools to scan the entire mouse genome for mutations related to human disease. An integrated approach is required, combining high-throughput, genome-wide mutagenesis with detailed phenotypic analysis from the whole organism to the molecular level (Fig. 1). Here we briefly discuss several approaches to large-scale mutagenesis in mice, some of the phenotyping screens and tools being developed and the issues around archiving and distributing this rapidly growing resource to the community.

## Interrogating the mouse genome for information relevant to human disease

### *Spontaneous mutants and inbred strains*

Long before the era of genomics, laboratory mouse stocks were a fertile source of genetic variation and spontaneous mutations relevant to human biology and disease (<http://www.informatics.jax.org/>). Over 1000 spontaneous or radiation-induced mutations have been documented and many have served as critical entry points into human physiology. For example, the cloning of the *ob* and *db* mouse mutants identified leptin and its receptor, respectively<sup>1</sup> and opened up an entirely new area of research into the control of body weight and other physiological processes. However, this source of genetic variation is usually confined to mutations detectable in the heterozygous state by their visible, semi-dominant

### Janet Rossant\*

Dept of Molecular and Medical Genetics, University of Toronto, and Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada.  
\*e-mail: rossant@mshri.on.ca

### Colin McKerlie

Dept of Pathology and Laboratory Medicine, University of Toronto, and Sunnybrook and Women's Health Sciences Centre, Toronto, Ontario, Canada.