BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 10. NO 5. 258–265 doi:10.1093/bfgp/elr031

New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models

Paul N. Schofield, John P. Sundberg, Robert Hoehndorf and Georgios V. Gkoutos

Abstract

The systematic investigation of the phenotypes associated with genotypes in model organisms holds the promise of revealing genotype–phenotype relations directly and without additional, intermediate inferences. Large-scale projects are now underway to catalog the complete phenome of a species, notably the mouse. With the increasing amount of phenotype information becoming available, a major challenge that biology faces today is the systematic analysis of this information and the translation of research results across species and into an improved understanding of human disease. The challenge is to integrate and combine phenotype descriptions within a species and to systematically relate them to phenotype descriptions in other species, in order to form a comprehensive understanding of the relations between those phenotypes and the genotypes involved in human disease. We distinguish between two major approaches for comparative phenotype analyses: the first relies on evolutionary relations to bridge the species gap, while the other approach compares phenotypes directly. In particular, the direct comparison of phenotypes relies heavily on the quality and coherence of phenotype and disease databases. We discuss major achievements and future challenges for these databases in light of their potential to contribute to the understanding of the molecular mechanisms underlying human disease. In particular, we discuss how the use of ontologies and automated reasoning can significantly contribute to the analysis of phenotypes and demonstrate their potential for enabling translational research.

Keywords: *phenotype; animal model; disease; database; comparative phenomics; ontology*

The use of animals as models for human disease is one of the most important paradigms of biomedical research. Historically, these have been of central importance to the study of human disease and treatments, and are based on the now well-established premise of evolutionarily conserved pathogenetic mechanisms. The comparative approach has also long been an important method for the study of problems in fundamental biology, and has been made even more powerful by the resources of modern genomics, where it can be combined with phylogenetic analysis. Our understanding of gene function can be informed by the comparison of mutant and 'wild-type' phenotypes in a single organism as well as by the comparison of the

Corresponding author. Georgios V. Gkoutos, University of Cambridge, Department of Genetics, Downing Street, CB2 3EH, UK. Tel: +44 0 1223 766336; Fax: +44 0 1223 333992; E-mail: gg295@cam.ac.uk

Paul N. Schofield is Senior Lecturer in Anatomy. His expertise is in mammalian genetics and bioinformatics. He works on the control of mammalian pre-natal growth and the development of biomedical ontologies to relate human disease to mouse models.

John P. Sundberg's research focuses on the biology and pathobiology of hair and skin using the mouse as a model system, and the role of papillomaviruses in the genesis of skin cancers and related diseases. In addition, he integrates large-scale pathological phenotyping with public access databases.

Robert Hoehndorf received his PhD at Leipzig University and the Max Planck Institute for Evolutionary Anthropology. The focus of his research is the application of ontologies for large-scale data analysis in biomedicine and their use for translational research.

Georgios V. Gkoutos's expertise lies in molecular informatics and bioinformatics. He has developed a number of biomedical ontologies and his research focuses on the association of genotype to phenotype, translational research and the discovery of novel disease genes and drugs.

phenotype consequences of mutation of a particular gene in species *A* with those of a mutation of the evolutionarily homologous gene in species *B*. The increasing importance of these cross-species comparisons is becoming ever more apparent as large-scale mutation screens (or their epigenetic analogs using interfering RNAs) are conducted in model organisms, particularly mice, zebrafish, *Drosophila* and *Caenorhabditis elegans*. Conservation of gene function across species is strongly supported by similar phenotype consequences of loss-of-function mutations in orthologous genes in both species [1] and functional replacement of mouse genes by their human counterparts [2, 3] as well as the remarkable phylogenetic conservation of patterns of gene expression [4].

There are profound challenges in relating disease processes in humans and other animals, partly as a consequence of intrinsic variation in normal and pathobiology between species, and partly through historical, conceptual and pragmatic differences between clinical and lab approaches to describing diseases and phenotypes. Many such challenges are being faced head-on by those working with animal models of human disease, specifically with regard to inbred strains of laboratory mice. Robert Koch's exhortation, 'Gentlemen, never forget that mice are not human beings' [5], reminds us that although use animal models of human disease can be extremely valuable, knowledge of comparative anatomy, pathology and pharmacology are needed to bridge the species gap. While genetic models often faithfully reflect the major features of human diseases, differences in phenotypes between mice and humans with mutations in orthologous genes can be as informative on the biological processes as those where there is a good match. Several diseases were first defined in mice and later in humans [6, 7] and more recently we have seen examples where apparent differences between mouse and human phenotypes were resolved with the realization that the clinical phenotype description in humans is incomplete and the mouse phenotypes are actually seen in the human disease [8]. Here, the mouse phenotype informs and expands the human clinical picture and provides new insights into the pathogenetic mechanism.

PREDICTING GENE FUNCTION FROM MUTANT PHENOTYPES AND DISEASES

We have been remarkably unsuccessful at being able to predict gene function at a physiological or whole organism level from gene sequence alone, and our understanding of gene function must ultimately depend on experimental manipulation of the genome and assessment of its consequences. The mouse has been a particularly successful organism in this regard as a consequence not only of well-characterized natural genetic variation, but also the ready ability to make loss-of-function, gain-of-function and conditional mutations in the mouse genome. Consequently, the mouse has proved to be one of the most powerful model organisms in the modern approach to understanding human disease [9–11], providing insights into normal and pathobiology and gene function as well as serving as preclinical tools for drug discovery and efficacy testing [12–16].

Data from the mouse genome database (MGD) [17] currently (June 2011) lists 14 820 genes with mutant alleles in mice and 11 210 in mouse embryonic stem (ES) cell lines alone. Despite this rich genetic resource, we have at best partial phenotype information on only 8200 of its approximately 22 000 protein-coding genes. Even at the level of protein class or Gene Ontology (GO) [18] molecular function annotation, we only have 13 000 genes with any experimentally based functional annotations. Of the 8600 GO 'molecular function' terms most have very few associated gene products and only 7000 genes have 'biological process' annotations in the mouse.

In order to maximize leverage experimental data from model organisms to understand gene function, we ideally need to have rich phenotype–genotype annotations for as many genes as possible, and to represent that information in a way that enables the biologically meaningful integration and comparison of data between species. Investigators wishing to make maximum use of phenotype data from humans and model organisms currently face two problems which we address in this commentary.

First, the incomplete and incoherent description of phenotypes in different organisms impairs systematic analyses of phenotype information. In particular, partial descriptions of different aspects of the phenotype for different alleles within a species, or for orthologs from different species, need to be combined to form the complete picture of the phenotypes associated with mutations in a gene.

Secondly, the lack of computationally tractable methods for describing and analyzing phenotypes across multiple species prevents large-scale automated analyses.

PHENOTYPES FOR EVERY GENE; HUMAN AND MOUSE PHENOTYPE DATA SETS

A limitation of the phenotype information that we can gather from the published literature is that it is based on a 'you get what you look for' approach. This, together with background strain effects and environmental factors, may contribute to the frequently reported absence of discernable phenotypes in mutant mice, in addition to the often-suggested 'redundancy' in the genome [19, 20]. However, the degree of pleiotropy which we intuitively expect suggests that if one looks in the right place at the right time one will more often than not find a phenotype consequence of genetic mutation. To quote Mario Capecchi on missing phenotypes: '. . . you have to look in the right place because every gene has to have a function' [21].

It now seems highly likely that comprehensive physiological, behavioral and structural phenotyping will reveal discernable phenotype change for many, if not all, null mutations. This is borne out by the phenotype effects discovered in a recent targeted study of mutants in 472 transmembrane and secreted proteins from Tang *et al.* [22], where 89% of mice showed phenotypes in at least one organ system and 57% in two or more based on measurement of 85 assays spanning immunology, metabolism, cardiology, oncology, growth, ophthalmology, neurobiology, pathology, reproduction, viability and embryonic lethality. While this suffers from some inevitable systematic biases [23], nevertheless, it suggests that the phenotypes are out there as long as you look carefully.

None of the existing human or mouse phenotype data sets is currently optimal for cross-species analysis. Human data are not coded systematically between different resources, and databases are not designed for interoperability [24]; however, the range of phenotype information available is extremely broad and deep. The recent commitment of online Mendelian inheritance in man (OMIM) to cross-reference to the human phenotype ontology (HPO [25]; see below), elements of morphology terms [26], International Classification of Diseases (ICD) and the mammalian phenotype ontology (MPO; [27]) is a welcome move to standardize the structure of the clinical synopses and make them available to searching and analysis [28]. This will hopefully be the first of similar moves to standardization of human phenotype description in phenotype–genotype databases. Recent examination of the databases containing human phenotype data by Oti *et al.* [29], prompted a strong argument that integration of OMIM [28], Possum [30] and Orphanet [31] would generate a much more powerful resource. This depends on the implementation of a method of standardization for human phenotype data and development of a formal model to describe elements of disease such as phenotype frequency, desirables that are discussed in the next section.

In contrast, for the mouse we have a very heterogeneous data set in MGD with various degrees of depth and annotation density, though it is all coded and structured in ways which allow for powerful analysis. The incompleteness of the mouse phenotype data sets also manifest in a way which underlies the difference between clinical practice and laboratory phenotyping. In clinical medicine, a complete picture is built up with a summative, often etiological diagnosis, whereas information on the pathobiological context of individual findings is often lacking in the mouse due to incomplete data.

In addition to the established uses of mouse phenotype data discussed above, model organism phenotype data are now finding important uses in the exploration and validation of human genome-wide association studies (GWAS) and in the dissection of human copy number variation (CNV) data [32, 33]. These applications have depended on the development of appropriate tools and informatics for capturing and analyzing phenotype data discussed below. Currently, the data used for the mouse are the very high quality manually curated genotype–phenotype data set from the mouse genome informatics (MGD) database, but for reasons discussed by Kitsios *et al.* [32] the utility of these data are limited by its derivation from hypothesis-driven experiments. Kitsios *et al.* analyzed human GWAS data from the National Human Genome Research Institute (NHGRI) catalog [34] comparing human against mouse phenotype data and concluded that while there is an excellent concordance between human and mouse phenotypes associated with orthologous loci, a significant number of GWAS associations do not have mouse mutants annotated to equivalent phenotype descriptors—in this case MPO terms—and suggest that the problem is likely to get worse as the depth of human phenotype determination increases with time. This is almost certainly due to the fact that curated mouse data are derived from reports of hypothesis-driven science, often

incomplete, and driven by the interests and competence of the investigators, and underlines the need for systematic, agnostic phenotyping of the mouse genome. Such systematic high-throughput phenotyping is now underway as part of the International Mouse Phenotyping Consortium (IMPC) [35], which aims to generate system-wide phenotypes for null mutants in all the approximately 22 000 protein sequence-associated genes in the mouse by 2021 (http://www.mousephenotype.org/).

DESCRIBING PHENOTYPE AND DISEASE DATA

The lack of a computationally tractable method for comparing phenotypes in different species is a current fundamental weakness. Not only does each of the model organisms have its own vocabulary for describing the phenotype consequences of mutation, vocabularies often tied to the particular conceptualizations of the anatomy or physiology of the organism, but worse, these descriptions are usually recorded in the literature as free text. Free text is highly expressive, but it does not help a computer program recognize the fact that there is likely to be a significant connection between the mutations that result in 'small' mice, 'dwarf' *Drosophila*, 'short' fish and 'reduced stature' humans. Two types of approaches have been made to solving the cross-species problem.

The first makes use of a 'translation layer' between phenotype descriptors. This translation layer may be either a direct mapping, based on automated, manual or human-assisted automatic lexical mapping, or indirect mapping using an intermediate framework. The intermediate framework may be a network or terminology such as Unified Medical Language System (UMLS) [36] or Medical Subject Headings (MeSH), or in recent approaches may make the use of the semantic information in a phenotype ontology to bridge species-specific ontologies using computable logical definitions of phenotype classes.

The second approach makes use of gene orthology and is predicated on the assumption that orthologous genes are involved in orthologous pathways. These genes give rise to orthologous phenotypes—'phenologs', as coined by McGary *et al*. [37]. In their study McGary *et al*. [37] use orthology to identify overlapping gene sets between species and infer that the associated phenotypes are in some way homologous. While there is considerable evidence to suggest that at least in closely related genera, such as mouse and human this is true, this assumption is not needed for the phenotype-alone methods of approach 1. Orthology has also been exploited by Espinosa and Hancock [38] in order to map mouse model phenotype annotations onto OMIM diseases to develop a phenotype–genotype network amenable to graphical analysis and is used for integration of phenotype data in Phenomic DB [39], where native text phenome data are captured into a cross-species database from OMIM and model organism databases.

PHENOTYPE ONTOLOGIES AND TERMINOLOGIES

Automated text mining from the literature and lexical methods for phenotype term matching has been used with some considerable success in establishing standardized gene–phenotype correlation data sets for humans. For example, van Driel *et al*. used MeSH to mine OMIM and established that similarity between phenotypes reflected underlying biological modules of interacting functionally related genes, proposing the modularity of phenotypes [40, 41]. Butte and Kohane have used UMLS and MeSH to extract information from text annotations of Gene Expression Omnibus (GEO) to establish human disease–genotype relations and it is clear that UMLS, which is a unifying resource for medical terminologies, is a powerful tool in integrating vocabularies and extracting information from unstructured text [42].

The development of the MPO in 2004 provided the first formal ontology for the description of mammalian phenotypes. The combination of a well-formed ontology and the manual curation to the ontology carried out by Mouse Genome Informatics (MGI) curators has meant that the mouse is still probably the best phenotypically annotated vertebrate. Several strategies for mapping of MPO to human phenotype terms have been developed. Kitsios *et al*. [32] employed a manually guided automated lexical mapping of the free-text phenotype terms used in the NHGRI GWAS browser and supplemented with MeSH used mouse phenotypes to assist in the prioritization of GWAS candidate genes. Burgun *et al*. [43] used an approach involving lexical matching to map MPO to UMLS, which then permits interrogation of OMIM and other clinical data sets using one of the more

than 100 terminologies integrated into UMLS. Problems with disambiguation of gene names, conceptual dissonance between the terminologies and semantic inconsistency within MPO show some of the drawbacks with this approach. Nevertheless, the possibility of mapping a large number of terminologies via UMLS in a single mapping is very powerful. This approach can be used as a way of identifying equivalent terms in different terminologies for both intra- and cross-species data integration.

To a great extent the almost exclusive use of lexical matching for knowledge extraction and cross-species phenotype bridging in phenotype analysis was determined by the lack, until 2008, of a widely applicable HPO [25]. Such an ontology could be used in direct annotation or computational markup and data extraction from documents and on-line resources such as OMIM. The existence of controlled vocabularies for human phenotypes (for example, those used in the London Dysmorphology Database (LDDB), Orphanet and POSSUM), has been extremely helpful. However, the harmonization of these vocabularies is an ongoing task and until recently it has not been possible to integrate or co-analyze annotation date from these sources. Additional formal nomenclatures are being developed, notably for dysmorphology [44].

CROSSING THE SPECIES GAP

The problems associated with crossing the phenotype 'gap' between different ways of describing phenotypes in different genera are discussed in Schofield *et al.* [45]. The issues are not simply those of establishing straight mappings, though that is difficult enough, but where there are large evolutionary distances between species, bridging to the most closely related phenotypes. For example, cardiac defects such as the tetralogy of Fallot are closely related in human and mouse, but while the syndrome is impossible in fish due to the different anatomical structure of the fish heart, other cardiac morphological defects may well be the consequence of dysregulation of the same morphogenetic processes in all three organisms. It is useful to include such related defects in the analysis. This requires a way to provide rich, explicit and consistent descriptions, so that automated systems are able to process and distinguish the meaning of their terms and use them to infer new information.

The problems with the use of lexical matching and lack of a formal ontology for human phenotypes have now been resolved with the much needed development of the HPO [25]. We now have phenotype ontologies for the mouse, yeast, worm and fly, all of which are available from the Open Biological and Biomedical Ontology (OBO) foundry (http://www.obofoundry.org/) [46]. Lexical cross-mapping of MPO and HPO using UMLS as a translation layer has been reported [47], but this suffers from the same problems reported by Burgun and co-workers. In recent years an approach has been developed to circumvent the species specificity of the phenotype ontologies using matching of logical definitions for classes within each ontology rather than text. The definitions utilize species-agnostic ontologies such as GO to provide a common semantic level at which they can be integrated into a single framework [48, 49] and a post-composition strategy termed Entity–Quality (EQ).

Rather than using a precomposed ontology such as HPO, phenotypes may be described using the EQ formalism [49]. In the EQ method, a phenotype is characterized by an affected Entity (from an anatomy or process ontology) and a Quality [from the Phenotype and Trait Ontology (PATO)] that specifies how the entity is affected [48]. The affected entity can either be a biological function or process as specified in GO, or an anatomical entity. The Zfin database uses the EQ method exclusively to capture phenodeviance [50], but these statements may be used as logical definitions for classes within a phenotype ontology. While the ontologies used to write the definitions (cross-products) are largely species agnostic, such as GO, CheBI, MPATH [51], anatomical entities are almost exclusively specified using a species-specific anatomy ontology, for example, the Foundational Model of Anatomy (FMA), the Mouse Adult Anatomy (MA) or the Zebrafish Anatomy (ZFA), and to make mappings between these vertebrate anatomies the metazoan, species-independent UBERON ontology is used in constructing anatomically based cross-products [52]. Once their classes are formally defined, phenotype ontologies may be linked through common or related cross-product definitions and striking concordances may be discovered between the phenotypes of different species. This method was successfully exploited by Washington *et al.* [53] who annotated the phenotypes of 11 gene-linked human diseases from OMIM and computationally compared these with

other ontology-based phenotype descriptions from model organisms. They showed that, based on the subsumption of classes in the ontologies and the frequency of annotation, they could detect other alleles of the same gene, other members of a signaling pathway, and orthologous genes and pathway members across species through the similarity of the phenotypes, demonstrating a proof of principle for the EQ approach.

More recently, a whole-phenome approach to comparative phenomics was developed exploiting the full semantic content of ontologies [54]. The method requires the formalization of anatomy and phenotype ontologies so that they can be integrated using the parthood relation followed by the generation of a single, unified and logically consistent representation of phenotype data for multiple species annotated to the species-specific phenotype ontologies within a single semantically coherent framework, amenable to automated reasoning [55]. The unified framework, PhenomeNET, permits the use of phenotype information alone to query ('phenomeblast') the gathered phenotype annotations from OMIM and the mouse, zebrafish, fly, yeast and worm model organism databases. The ontology contains more than 275 000 classes and more than a million axioms, including classes for 86 203 complex phenotype annotations drawn from the model organism databases and OMIM. A great advantage is that the ontology can be regenerated to include new phenotype annotations and future developments of all of the constituent ontologies, and a tool to query the data has been made available on http://www.phenomebrowser.net/. This method is the first to be able to allow automated reasoning over all of the phenotype ontologies and the gathered ontologies involved in the logical class definitions. It permits a simultaneous survey and computation over all of the phenotype data available from the main model organism and human databases. The network can be used to successfully identify orthologous genes through related phenotypes, and genes involved in the same pathway as well as genes giving rise to the same disease.

Development of PhenomeNET has allowed the comparison of usefulness of the phenotype annotations in the model organism databases and OMIM. It is clear that the manual annotation of MGD represents a gold standard for literature annotation to a precomposed phenotype ontology. In contrast, the heterogeneity and in some case sparseness of annotation in OMIM is problematical. Oti *et al*. found that the under-annotation of diseases in OMIM is a weakness in its ability to provide a resource for identifying animal models of OMIM diseases, and this observation could be confirmed with PhenomeNET. The use of annotations in Orphanet, and particularly the possibility of incorporating frequency data into the 'phenomeblast', are important areas of research to extend automated cross-species analyses of phenotype information.

REQUIREMENTS FOR INTEROPERABILITY AND LARGE-SCALE ANALYSIS OF PHENOTYPES

To further facilitate and improve the automated analysis of the growing information about phenotypes, three areas of research need to be addressed. First, the documentation of phenotype information in scientific databases and publications needs to be further standardized. While species-specific phenotype ontologies are being applied in several model organism databases, it is a major challenge to unify these phenotype ontologies across species. The EQ approach based on the PATO ontology has been successfully applied to several model organism databases as well as to formally define classes in species-specific phenotype ontologies. Moreover, PATO has been demonstrated to support cross-species integration and comparison. It would be desirable to further align phenotype ontologies across species based on the PATO framework, to improve mappings between species-specific anatomy ontologies and to develop new phenotype ontologies based on the PATO framework so that ontologies immediately connect to the existing web of cross-species phenotype knowledge.

The second area requiring further development is the establishment of an infrastructure for the representation, processing and analysis of phenotype information. Large-scale reasoning over phenotypic information enables highly expressive queries across multiple domains, but is limited by the complexity of phenotype descriptions. Modularization approaches, efficient reasoning and reasoning servers may help to allow access to phenotype information [55]. Further extensions include semantic web service frameworks [56] and general tools to create, edit and analyze information about phenotypes without requiring knowledge about the structure of the underlying ontologies.

Finally, reference databases and resources in biomedicine need to link to and contribute to the unifying web of knowledge that is enabled by formal, ontology-based phenotype descriptions. In particular, ontology-based phenotype descriptions for diseases, as represented in databases such as OMIM or Orphanet, need to be consistently linked to ontologies, and the phenotypes associated with diseases completed and corrected where errors are detected.

FUTURE PROSPECTS

Currently, diseases in humans and model organisms are described as the product of a set of constituent phenotypes, or 'endophenotypes', ignoring the frequency and co-occurrence of specific phenotype aspects of the disease: time, prognosis, molecular signatures and therapeutic responsiveness. These are important aspects of human disease and are in some cases available for model organisms but not captured in current curation practices as we lack a formal model for disease. Approaches toward developing such a model are being proposed currently [57, 58] and it is clear that such developments will improve the richness and accuracy of phenotype–disease descriptions with concomitant improvement in the power of informatics to detect similarities between related diseases and subtypes within apparently uniform conditions.

Key Points

- Phenotype studies in model organisms are successful in revealing genotype–phenotype relations and the molecular mechanisms underlying human disease.
- Challenges for analyzing phenotype data include incomplete and noisy information in databases describing model organisms and human diseases.
- Approaches toward integrating phenotypes across species include ontology-based approaches and the direct comparison of phenotypes.
- Biomedical ontologies and automated reasoning are a means to integrate phenotypes across species.
- The combination of ontology- and similarity-based approaches has been shown to successfully suggest novel genes for rare and orphan diseases.

FUNDING

Related work in the authors' laboratories is supported by grants from the Ellison Medical Foundation (to J.P.S.), the National Institutes of Health (AG25707, for the Shock Aging Center, CA89713, and AR056635) to J.P.S., P.N.S. acknowledges funding from NIH (R01 HG004838-02) and the Commission of the European Union (EUMODIC contract number LSHG-CT-2006-037188), funding for RH was provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502 and funding for G.V.G. was provided by BBSRC grant BBG0043581.

References

1. Al-Hasani K, Vadolas J, Knaupp AS, *et al*. A 191-kb genomic fragment containing the human alpha-globin locus can rescue alpha-thalassemic mice. *Mamm Genome* 2005;**16**: 847–53.
2. Nakatani J, Tamada K, Hatanaka F, *et al*. Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism. *Cell* 2009;**137**: 1235–46.
3. Wallace HA, Marques-Kranc F, Richardson M, *et al*. Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* 2007;**128**:197–209.
4. Zheng-Bradley X, Rung J, Parkinson H, *et al*. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol* 2010;**11**:R124.
5. Greep RO. Animal model in biomedical research. *J Anim Sci* 1970;**31**:1235–46.
6. Kljuic A, Bazzi H, Sundberg JP, *et al*. Desmoglein 4 in hair follicle differentiation and epidermal adhesion: evidence from inherited hypotrichosis and acquired pemphigus vulgaris. *Cell* 2003;**113**:249–60.
7. Sundberg JP, Price VH, King LE. The ''hairless'' gene in mouse and man. *Arch Dermatol* 1999;**135**:718–20.
8. Lisse TS, Thiele F, Fuchs H, *et al*. ER stress-mediated apoptosis in a new mouse model of osteogenesis imperfecta. *PLoS Genet* 2008;**4**:e7.
9. Rosenthal N, Brown S. The mouse ascending: perspectives for human-disease models. *Nat Cell Biol* 2007;**9**: 993–99.
10. Peters LL, Robledo RF, Bult CJ, *et al*. The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat Rev Genet* 2007;**8**:58–69.
11. Sundberg JP. Mouse mutations: animal models and biomedical tools. *Lab Anim* 1991;**20**:40–9.
12. Zambrowicz BP, Sands AT. Knockouts model the 100 best-selling drugs–will they model the next 100? *Nat Rev Drug Discov* 2003;**2**:38–51.
13. Zambrowicz BP, Turner CA, Sands AT. Predicting drug efficacy: knockouts model pipeline drugs of the pharmaceutical industry. *Curr Opin Pharmacol* 2003;**3**:563–70.
14. Sutherland KD, Berns A. Cell of origin of lung cancer. *Mol Oncol* 2011;**4**:397–403.
15. Antony PM, Diederich NJ, Balling R. Parkinson's disease mouse models in translational research. *Mamm Genome* 2011;**7–8**:401–19.
16. Van Dam D, De Deyn PP. Animal models in the drug discovery pipeline for Alzheimer's disease. *Br J Pharmacol* 2011, doi: 10.1111/j.1476-5381.2011.001299.x.

17. Blake JA, Bult CJ, Kadin JA, *et al*. The mouse genome database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res* 2011; **39**:D842–8.
18. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
19. Barbaric I, Miller G, Dear TN. Appearances can be deceiving: phenotypes of knockout mice. *Brief Funct Genomic Proteomic* 2007;**6**:91–103.
20. Barthold SW. Genetically altered mice: phenotypes, no phenotypes, and Faux phenotypes. *Genetica* 2004;**122**:75–88.
21. Kain K. The first transgenic mice: an interview with Mario Capecchi. *Dis Model Mech* 2008;**1**:197–201.
22. Tang T, Li L, Tang J, *et al*. A mouse knockout library for secreted and transmembrane proteins. *Nat Biotechnol* 2010; **28**:749–55.
23. Wurst W, de Angelis MH. Systematic phenotyping of mouse mutants. *Nat Biotechnol* 2010;**28**:684–5.
24. Patrinos GP, Brookes AJ. DNA, diseases and databases: disastrously deficient. *Trends Genet* 2005;**21**:333–8.
25. Robinson PN, Kohler S, Bauer S, *et al*. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 2008;**83**:610–5.
26. Allanson JE, Cunniff C, Hoyme HE, *et al*. Elements of morphology: standard terminology for the head and face. *Am J Med Genet A* 2009;**149A**:6–28.
27. Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* 2005;**6**:R7.
28. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM((R))). *Hum Mutat* 2011;**32**:564–7.
29. Oti M, Huynen MA, Brunner HG. The biological coherence of human phenome databases. *Am J Hum Genet* 2009; **85**:801–8.
30. Bankier A, Keith CG. POSSUM: the microcomputer laser-videodisk syndrome information system. *Ophthalmic Paediatr Genet* 1989;**10**:51–2.
31. Weinreich SS, Mangon R, Sikkens JJ, *et al*. [Orphanet: a European database for rare diseases]. *Ned Tijdschr Geneeskd* 2008;**152**:518–9.
32. Kitsios GD, Tangri N, Castaldi PJ, *et al*. Laboratory mouse models for the human genome-wide associations. *PLoS One* 2010;**5**:e13782.
33. Chen J, Xu H, Aronow BJ, *et al*. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* 2007;**8**:392.
34. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.
35. Abbott A. Mouse project to find each gene's role. *Nature* 2010;**465**:410.
36. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
37. McGary KL, Park TJ, Woods JO, *et al*. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci USA* 2010;**107**:6544–9.
38. Espinosa O, Hancock JM. A gene-phenotype network for the laboratory mouse and its implications for systematic phenotyping. *PLoS One*;**6**:e19693.
39. Groth P, Pavlova N, Kalev I, *et al*. PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res* 2007;**35**:D696–9.
40. van Driel MA, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**: 535–42.
41. Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet* 2007;**71**:1–11.
42. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol* 2006;**24**:55–62.
43. Burgun A, Mougin F, Bodenreider O. Two approaches to integrating phenotype and clinical information. *AMIA Annu Symp Proc* 2009;**2009**:75–9.
44. Allanson JE, Biesecker LG, Carey JC, *et al*. Elements of morphology: introduction. *Am J Med Genet A* 2009;**149A**:2–5.
45. Schofield PN, Gkoutos GV, Gruenberger M, *et al*. Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis Model Mech* 2010;**3**:281–9.
46. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.
47. Sardana D, Vasa S, Vepachedu N, *et al*. PhenoHM: human-mouse comparative phenome-genome server. *Nucleic Acids Res* 2010;**38**:W165–74.
48. Gkoutos GV, Green ECJ, Mallon A-M, *et al*. Building mouse phenotype ontologies. *Pac Symp Biocomput* 2004;**9**: 178–89.
49. Mungall CJ, Gkoutos GV, Smith CL, *et al*. Integrating phenotype ontologies across multiple species. *Genome Biol* 2010;**11**:R2.
50. Bradford Y, Conlin T, Dunn N, *et al*. ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res* 2011;**39**:D822–9.
51. Schofield PN, Gruenberger M, Sundberg JP. Pathbase and the MPATH ontology: community resources for mouse histopathology. *Vet Pathol* 2010;**47**:1016–20.
52. Haendel M, Gkoutos GV, Lewis SE, *et al*. *Uberon: Towards a Comprehensive Multi-species Anatomy Ontology*. Buffalo, NY: International Consortium of Biomedical Ontology, 2009.
53. Washington NL, Haendel MA, Mungall CJ, *et al*. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol* 2009;**7**:e1000247.
54. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res* 2011; 1–12.
55. Hoehndorf R, Dumontier M, Oellrich A, *et al*. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS One* 2011;**6**:7.
56. Wilkinson M, McCarthy L, Vandervalk B, *et al*. SADI, SHARE, and the in silico scientific method. *BMC Bioinformatics* 2011(Suppl. 12):S7.
57. Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol* 2007;**3**:124.
58. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. *Summit Translat Bioinforma* 2009;116–20.