D696–D699 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 28 September 2006*
doi:10.1093/nar/gkl662

PhenomicDB: a new cross-species genotype/phenotype resource

Philip Groth[1], Nadia Pavlova[2], Ivan Kalev[2], Spas Tonov[2], Georgi Georgiev[2], Hans-Dieter Pohlenz[1] and Bertram Weiss[1,*]

[1]Research Laboratories of Schering AG, Muellerstrasse 178, 13442 Berlin, Germany and [2]MetaLife AG, Metapark 1, 79297 Winden, Germany

Received July 14, 2006; Accepted August 24, 2006

ABSTRACT

Phenotypes are an important subject of biomedical research for which many repositories have already been created. Most of these databases are either dedicated to a single species or to a single disease of interest. With the advent of technologies to generate phenotypes in a high-throughput manner, not only is the volume of phenotype data growing fast but also the need to organize these data in more useful ways. We have created PhenomicDB (freely available at http://www.phenomicdb.de), a multi-species genotype/phenotype database, which shows phenotypes associated with their corresponding genes and grouped by gene orthologies across a variety of species. We have enhanced PhenomicDB recently by additionally incorporating quantitative and descriptive RNA interference (RNAi) screening data, by enabling the usage of phenotype ontology terms and by providing information on assays and cell lines. We envision that integration of classical phenotypes with high-throughput data will bring new momentum and insights to our understanding. Modern analysis tools under development may help exploiting this wealth of information to transform it into knowledge and, eventually, into novel therapeutic approaches.

INTRODUCTION

Phenotypes, especially those concerning health, have been an intensive subject of research in humans and in many model organisms. New technologies to generate phenotypes in a high-throughput manner, such as RNA interference in higher organisms, have further advanced the field (1). In the past years, an increasing number of phenotypes associated with genotypes have been gathered in online repositories dedicated to specific model organisms or diseases, many of which are listed in (2). The impact of phenotype data on biomedical research, an overview of repositories and useful analysis methods have been presented in detail (2).

However, until recently, little effort has been dedicated to connecting genotype/phenotype information across species. To advance this effort, we have created PhenomicDB, a multi-species genotype/phenotype resource freely available at http://www.phenomicdb.de. It enables easy cross-species mining of phenotypes and their associated genotypes by taking advantage of orthology relationships (3). Here, phenotype information for many organisms become condensed into a single view where all known genes are grouped by orthologies and, if available, associated with phenotypes obtained from studies as diverse as mutant screens, k.o. mice and RNA interference. In addition, clinical descriptions and naturally occurring mutants are shown.

Besides the Online Mendelian Inheritance in Animals (OMIA) (4), a small-scale equivalent of the Online Mendelian Inheritance in Man (OMIM) (5), and the beginning efforts of the Ensembl group to gather phenotypic information (6), PhenomicDB continues to be the only database containing in-depth phenotypic information for more than one species. In a recent effort, PhenomicDB has been updated to its current version 2.1, now offering the capability to include data from whole-genome RNAi screens with detailed information on experimental design, ontology terms from the MGI's Mammalian Phenotype Ontology (7) and keywords for cell lines and experimental assays. Also, direct linking from external sources by search term or identifier is now possible.

THE DATABASE

Data content

PhenomicDB hosts classical phenotype data from a variety of sources, namely OMIM, the Mouse Genome Database (MGD) (8), WormBase (9), FlyBase (10), the Comprehensive Yeast Genome Database (CYGD) (11), the Zebrafish Information Network (ZFIN) (12), and the MIPS *Arabidopsis thaliana* database (MAtDB) (13). The vast majority of these

*To whom correspondence should be addressed. Tel: +49 30 468 14 42 4; Fax: +49 30 468 94 42 4; Email: bertram.weiss@schering.de

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

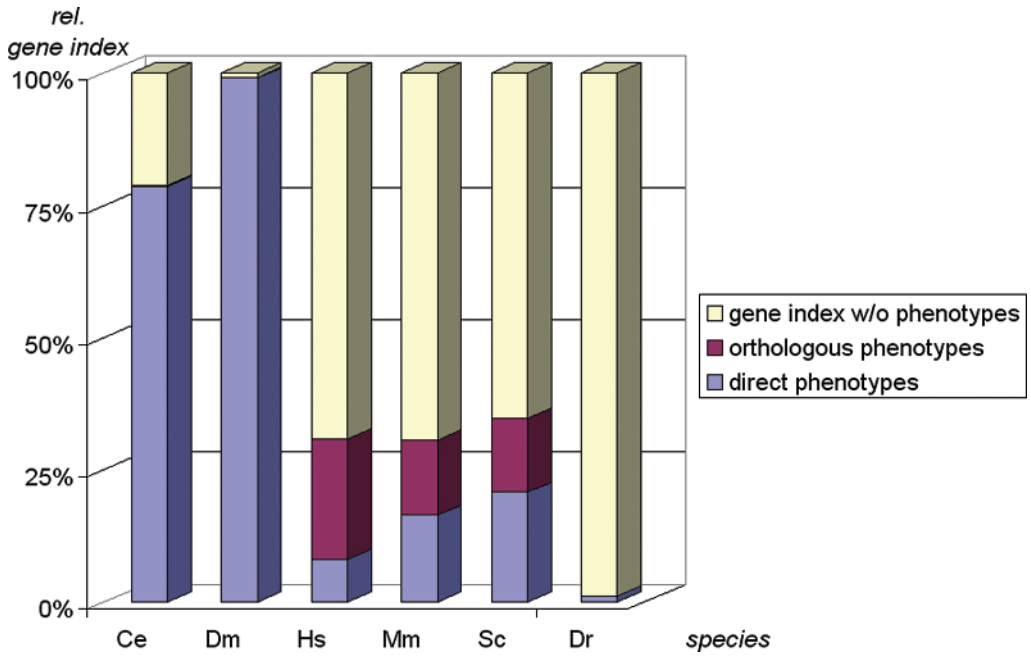

Figure 1. Percentage of NCBI Entrez Gene indices with phenotypic information in PhenomicDB for 5 model organisms and human. (Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Sc, *Saccharomyces cerevisiae*; Dr, *Danio rerio*). The percentages of genes with one or more phenotype from the given species is shown in blue ('direct phenotypes'), of those with one or more phenotype associated by orthology are shown in red ('orthologous phenotypes'), and of those genes that have no phenotype associated are shown in yellow. The red bars thus indicate the direct benefit from cross-species integration in PhenomicDB. The high coverage of *C.elegans* and *D.melanogaster* gene indices with phenotypic information is mainly owed to recently integrated RNA interference data.

phenotypes is associated with genes mapped to a common index, the Entrez Gene Index of the National Center for Biotechnology Information (NCBI) (14). Functionally equivalent (i.e. orthologous) genes from different species are grouped by taking advantage of the NCBI's HomoloGene database (15). Full annotations including the Gene Ontology (16) are provided with the genotype information as taken from Entrez Gene.

In its last major update, PhenomicDB has been redesigned to accept large datasets from whole-genome RNAi screens and thus has become a central home of data spread over dedicated smaller databases, e.g. PhenoBank (17) which has been created for a single screen, or FlyRNAi (18) for fly-specific screens, or supplementary information of journals. RNAi screens in *Caenorhabditis elegans* (17) and in *Drosophila melanogaster* (19–25) have been added as well as data from other species, subject to open access publication and availability of the data. All data in PhenomicDB are referenced and links to the original data sources are provided. PhenomicDB is kept up-to-date on a quarterly schedule and is freely accessible without restrictions.

In total, PhenomicDB hosts 399 772 phenotypes, connected to 77 400 eukaryotic genes. The percentage of the Entrez Gene index with a phenotype varies between species: It is ~99% for *D.melanogaster*, 79% for *C.elegans*, 21% for *Saccharomyces cerevisiae*, ~16% for *Mus musculus* (this number is estimated on the basis of the human Entrez Gene number, as Entrez Gene index for mouse (62 907 Gene IDs) is still in progress and therefore has not collapsed yet) and 8% for *Homo sapiens*. 84% of all available phenotypes in PhenomicDB come from *D.melanogaster* and *C.elegans*. 16.2% of phenotypes are associated with a gene having no orthologs, and <1.5% have no gene associated at all. 40 299 eukaryotic orthology groups are registered and a third of them (13 695) have at least one phenotype in any of the species. For *H.sapiens*, 2850 genes are linked to 4009 human phenotypes and for another 7592 human genes there is at least one 'orthologous phenotype' available, thus raising the percentage of human genes with phenotypic information from 8% to 31% of the Entrez Gene index. For *M.musculus*, 'orthologous phenotypes' increase available phenotypic information for mouse genes to over 30% of the gene index (see Figure 1 for more details, also on other species). These figures clearly show how integrating disparate phenotype data from different species can generate unexpected contexts for this wealth of information.

Data presentation

In PhenomicDB, genotype and phenotype data have been organised in a single database schema. Having all genes annotated and also indexed over orthology groups, this data organization allows to present orthologous genotype and phenotype data with a single database query. The advent of RNAi data required the schema to be extended in order to cope with a 'qualitative' phenotype, e.g. the description of a visual inspection via microscopy, but also with a 'quantitative' phenotype, i.e. a floating point number expressing an absolute or relative deviation from an expected 'normal' or average phenotype. Also, important aspects of RNAi study design, e.g. assay, cell line, time point, mRNA knockdown efficiency, phenotype penetrance, etc. have been addressed

Figure 2. Result list for the frataxin orthology group (some entries omitted for simplicity). In marble the frataxin genes from different species are shown; indented and in green the corresponding phenotypes. Hyperlinks lead to the source database, the 'Show Entry' button displays the full genotype/phenotype information. For *Gallus gallus*, no phenotype (in red) is available.

adequately. Furthermore, we enriched PhenomicDB with tables holding MGI's Mammalian Phenotype Ontology and controlled vocabulary for cell lines and RNAi assays.

PhenomicDB's graphical user interface has been designed to be as simple and as effective as possible. A basic query can be started intuitively by entering any search term (e.g. apoptosis, BUB1) or identifier (e.g. NM_001211). Users can configure the output data fields to be shown individually, e.g. gene symbol, phenotype name, ontology, chromosomal localization, etc. Queries allow wildcards and logical operators ('AND', 'NOT' and 'OR') and can further be refined by limiting to data fields, data domains or organisms.

The customizable results interface (Figure 2) lists all hits organised by genes with their associated phenotypes indented and provides further links to more detailed views. Two buttons, 'Orthologies' for each gene and 'Show entry' for each hit, enable the user to show all orthologous genes with their associated phenotypes or to show the full genotype and phenotype entry for a gene of interest, respectively. Also, the entire hit list can be expanded to show the orthologs of all or selected genes as well as their corresponding phenotypes. All entries consistently link back to their original sources (e.g. entries derived from OMIM link back to OMIM) to make sure data will be properly referenced by users.

For convenient external access to PhenomicDB, static hyperlinks can be created to direct to any genotype or phenotype using e.g. the Entrez Gene ID. Dynamic URLs using any

query term behave as if the term was entered into the search mask of the homepage. A manual is available on the homepage. External linking to PhenomicDB is also featured in the browser task bar *BioBar* (http://biobar.mozdev.org/).

Future direction

During its 2 years of existence, PhenomicDB has seen important functional improvements as well as large increases in data content and more data, especially from whole-genome RNAi screens, are expected to be included in the very near future. We therefore expect the percentage of human genes associated with phenotypic data to steadily rise, making it an increasingly valuable resource in biomedical research. In the past quarter, PhenomicDB's content has been requested ~6000 times per month on average.

The wealth of steadily growing information raises the question on how to benefit beyond the mere rearrangement of views and data. We are working on data mining tools taking advantage of consistent phenotype ontologies with the aim to improve further the usefulness of PhenomicDB's data content, thus helping to transform it into knowledge and eventually into novel therapeutic approaches.

ACKNOWLEDGEMENTS

The authors are grateful to Bernard Haendler (Schering AG) for useful discussions of the manuscript. Funding to pay the Open Access publication charges for this article was provided by Schering AG, Berlin, Germany.

Conflict of interest statement. None declared.

REFERENCES

1. Shi,Y. (2003) Mammalian RNAi for the masses. *Trends Genet*., **19**, 9–12.
2. Groth,P. and Weiss,B. (2006) Phenotype data: a neglected resource in biomedical research? *Curr. Bioinformatics*, **1**, 347–358.
3. Kahraman,A., Avramov,A., Nashev,L.G., Popov,D., Ternes,R., Pohlenz,H.D. and Weiss,B. (2005) PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, **21**, 418–420.
4. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res*, **34**, D599–D601.
5. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.
6. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
7. Smith,C.L., Goldsmith,C.A. and Eppig,J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*., **6**, R7.
8. Blake,J.A., Eppig,J.T., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res*., **34**, D562–D567.
9. Schwarz,E.M., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Canaran,P., Chan,J., Chen,N., Chen,W.J., Davis,P. *et al*. (2006) WormBase: better software, richer content. *Nucleic Acids Res*., **34**, D475–D478.
10. Grumbling,G. and Strelets,V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res*., **34**, D484–D488.
11. Guldener,U., Munsterkotter,M., Kastenmuller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al*. (2005) CYGD: the comprehensive yeast genome database. *Nucleic Acids Res*., **33**, D364–D368.
12. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al*. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res*., **34**, D581–D585.
13. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res*., **32**, D373–D376.
14. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
15. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, D173–D180.
16. GeneOntologyConsortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*., **34**, D322–D326.
17. Sonnichsen,B., Koski,L.B., Walsh,A., Marschall,P., Neumann,B., Brehm,M., Alleaume,A.M., Artelt,J., Bettencourt,P., Cassin,E. *et al*. (2005) Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature*, **434**, 462–469.
18. Flockhart,I., Booker,M., Kiger,A., Boutros,M., Armknecht,S., Ramadan,N., Richardson,K., Xu,A., Perrimon,N. and Mathey-Prevot,B. (2006) FlyRNAi: the Drosophila RNAi screening center database. *Nucleic Acids Res*., **34**, D489–D494.
19. Agaisse,H., Burrack,L.S., Philips,J.A., Rubin,E.J., Perrimon,N. and Higgins,D.E. (2005) Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science*, **309**, 1248–1251.
20. Baeg,G.H., Zhou,R. and Perrimon,N. (2005) Genome-wide RNAi analysis of JAK/STAT signaling components in Drosophila. *Genes Dev*., **19**, 1861–1870.
21. Boutros,M., Kiger,A.A., Armknecht,S., Kerr,K., Hild,M., Koch,B., Haas,S.A., Paro,R. and Perrimon,N. (2004) Genome-Wide RNAi analysis of growth and viability in Drosophila cells. *Science*, **303**, 832–835.
22. Cherry,S., Doukas,T., Armknecht,S., Whelan,S., Wang,H., Sarnow,P. and Perrimon,N. (2005) Genome-Wide RNAi screen reveals a specific sensitivity of IRES-containing RNA viruses to host translation inhibition. *Genes Dev*., **19**, 445–452.
23. Eggert,U.S., Kiger,A.A., Richter,C., Perlman,Z.E., Perrimon,N., Mitchison,T.J. and Field,C.M. (2004) Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol*., **2**, e379.
24. Philips,J.A., Rubin,E.J. and Perrimon,N. (2005) Drosophila RNAi screen reveals CD36 family member required for mycobacterial infection. *Science*, **309**, 1251–1253.
25. Zhang,S.L., Yeromin,A.V., Zhang,X.H., Yu,Y., Safrina,O., Penna,A., Roos,J., Stauderman,K.A. and Cahalan,M.D. (2006) Genome-wide RNAi screen of Ca2+ influx identifies genes that regulate Ca2+ release-activated Ca2+ channel activity. *Proc. Natl Acad. Sci. USA*, **103**, 9357–9362.