LINKING PHARMGKB TO PHENOTYPE STUDIES AND ANIMAL MODELS OF DISEASE FOR DRUG REPURPOSING

ROBERT HOEHNDORF[1,*], ANIKA OELLRICH[2], DIETRICH REBHOLZ-SCHUHMANN[2]
PAUL N. SCHOFIELD[3], GEORGIOS V. GKOUTOS[1]

[1]*Department of Genetics, University of Cambridge*
Downing Street, Cambridge, CB2 3EH, UK

[2]*European Bioinformatics Institute*
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[3]*Department of Physiology, Development and Neuroscience*
University of Cambridge
Downing Street, Cambridge CB2 3EG, UK, and
The Jackson Laboratory, 600, Main Street
Bar Harbor ME 04609-1500, USA

The investigation of phenotypes in model organisms has the potential to reveal the molecular mechanisms underlying disease. The large-scale comparative analysis of phenotypes across species can reveal novel associations between genotypes and diseases. We use the PhenomeNET network of phenotypic similarity to suggest genotype–disease association, combine them with drug–gene associations available from the PharmGKB database, and infer novel associations between drugs and diseases. We evaluate and quantify our results based on our method's capability to reproduce known drug–disease associations. We find and discuss evidence that levonorgestrel, tretinoin and estradiol are associated with cystic fibrosis ($p < 2.65 \cdot 10^{-6}$, $p < 0.002$ and $p < 0.031$, Wilcoxon signed-rank test, Bonferroni correction) and that ibuprofen may be active in chronic lymphocytic leukemia ($p < 2.63 \cdot 10^{-23}$, Wilcoxon signed-rank test, Bonferroni correction). To enable access to our results, we implement a web server and make our raw data freely available. Our results are the first steps in implementing an integrated system for the analysis and prediction of drug–disease associations for rare and orphan diseases for which the molecular basis is not known.

Keywords: phenotype, drug repurposing, animal models, ontology, pharmacogenetics

1. Background

1.1. *Drug discovery and drug repurposing*

The major challenges currently faced by pharmacological research include the high rates of attrition in the development of new compounds (mainly in Phase II), the increased cost of development itself, and increased regulatory concern about safety and efficacy.[33] Recently, it has been shown that the rate of production of novel drugs has changed little since the 1950s,[28] yet the cost of developing those drugs has soared. As a result, pharmacological research increasingly focuses on *repurposing* or *repositioning* existing drugs for new indications.

Drug repurposing aims to find new indications for existing drugs, and strategies for drug repurposing can be divided into two main types: identification of new targets for known drugs and identification of new indications for a known mechanism of action.[33] Approaches to drug-repurposing include database-driven bioinformatics approaches, *in vivo* and *ex vivo* studies as

well as high-throughput screening methods.[33]

Finding new targets for existing drugs depends on our fundamental understanding of the physiology and pathobiology that underlies diseases, their phenotypic manifestations and the molecular sites of action of small-molecule therapeutics. The manifest modularity of disease phenotypes reflects the existence of underlying physiological pathways, and a lesion in any of the components of the same pathway can produce closely overlapping disease phenotypes.[29]

The sequencing of the human genome and that of many animal models, the rapid development of high-throughput phenotyping and genotyping technologies and our ability to create specific mutations in the genomes of model organisms have provided us with a vast amount of information that supports the discovery of meaningful associations between the genotype and phenotype of an organism. This information, in turn, extends our ability to comprehensively characterize the phenotypic manifestations of diseases and generate hypotheses on which the intelligent design of drugs can be built.

With the availability of increasing amounts of information in public and private database, the development of *in silico* approaches that can analyse the data, propose potential drug indications and ultimately reduce the cost and time of drug development is required. Large scale analysis frameworks such as the *PREDICT* framework[14] are examples of such approaches. These frameworks are successfully being applied for drug repositioning and the suggestions of potentially novel drugs.[14] They adopt a variety of methods such as drug response gene expression profiles, drug–drug similarity, drug–disease similarity with respect to known drug–disease associations, text mining of known associations and several further resources. One of the areas though that still remains to be fully exploited is the wealth of phenotypic information that is increasingly becoming available from a variety of studies on animal models of human diseases.

Here, we present an approach for predicting novel associations between drugs and diseases based on the PhenomeNET method for comparing phenotypes across species.[20] We apply our method to the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB),[16] a central repository containing a wealth of relationships between genetic, genomic, drug-response related phenotype data and clinical information. Our method utilizes predictions of disease–gene associations from the PhenomeNET.[20] These predictions are based on a similarity-based comparative analysis of the phenotypic data associated with humans in the Online Mendelian Inheritance in Men (OMIM) database[1] and the phenotypes associated with animal studies in five major model organisms. We use these gene–disease associations and combine them with the drug–gene associations available from the PharmGKB to suggest new diseases in which a drug may be active.

1.2. *Animal models of human disease*

To gain an in depth understanding of specific drug actions it is often necessary to study and validate their effects as part of a whole system that involves, for example, the organism's physiology including metabolism, absorption, excretion, distribution and toxicity effects. The better we understand the pathophysiology and underlying *in vivo* biology of an organism, the more likely we are to take advantage of the development of new technologies that enable drug discovery.

As such, animal models provide a powerful mechanism for drug discovery since they determine the physiological conditions and complex interdependencies among different cell types and tissues in which the interactions of chemicals with drug targets can be studied. Based on the premise of evolutionarily conserved pathogenetic mechanisms, animal models such as mouse, zebrafish, fruitfly, yeast and worm have been harnessed to provide an in-depth understanding of the biological mechanisms that govern the effect of drug administration. These benefits assign an important role to animal models in the drug discovery process.

The advent of functional genomics allowed for the large scale exploration of gene function based on the systematic comparative analysis of gene activity. The comparison of mutant and "wild-type" phenotypes within a single organism as well as with respect to homologous genes in different organisms allows us to gain a better understanding of human disease. Despite the variations of physiology and pathobiology between species, phenotype information collected from animal models have proved extremely useful in providing new insights into disease mechanisms and etiologies. The analysis and study of phenotypes arising from the various models has direct implications for understanding mammalian physiology in the context of pharmacodynamics and pharmacokinetics studies, in understanding signalling and regulatory networks, in studies that focus on the identification of response regulators, activators and inhibitors, and in chemical genetics.

As a result, mutant strains derived from hypothesis-driven research are now being augmented for several animal models, including the mouse. For example, following the success of phenotype-driven ENU mutagenesis projects,[25] large scale gene knockout programmes have now been established with the ultimate goal of discovering the functions of all of the protein coding genes in the mouse genome.[6]

1.3. *Ontologies and phenotype information*

One of the consequences of all those efforts is the increase in the amount of phenotype information collected around the world and stored in various databases. This increase in phenotype data necessitated the development of computational frameworks that enable the retrieval, comparison and analysis of phenotypes. In response, the biomedical community has developed a plethora of species-specific phenotype ontologies[27,31,34] that are used as controlled vocabularies to annotate phenotypes in several model organism databases.

Ontologies formally specify the meaning of terms in a vocabulary[15] and express this meaning by utilising languages that provide an explicit, formal semantics. Many of the species-specific phenotype ontologies have been augmented with class definition[13,27] based on the Phenotype And Trait Ontology (PATO)[12] and other species-independent ontologies such as the Gene Ontology (GO)[2] and the Chemical Entities of Biological Interest (ChEBI) ontology.[8] Methods have been proposed to formally represent these structured definitions in OWL[19] and utilize them for automated reasoning and verification of annotations.[17,20]

1.4. *PhenomeNET*

PhenomeNET is a network in which nodes represent complex phenotypes resulting from either the phenotype annotations available in model organism databases or the disease descriptions

available from the OMIM database. To enable cross-species comparative analyses of phenotypes, PhenomeNET integrates multiple species-specific phenotype ontologies to derive a cross-species phenotype ontology.[20] This integrated ontology is based on the structured definitions of phenotypes that were developed for several species-specific phenotype ontologies,[27] the UBERON mappings between species-specific anatomy ontologies,[37] the Gene Ontology (GO),[2] the PATO ontology[12] as well as several other species-independent ontologies. Efficient automated reasoning over the PhenomeNET is enabled through ontology modularization[18] and design patterns for expressing phenotypes and their links to anatomy and physiology ontologies.[19]

Integrating phenotype ontologies allows for a *direct comparison* of phenotypes of multiple species, and PhenomeNET performs a pairwise comparison of phenotypes using a measure of semantic similarity. As a result, PhenomeNET ranks phenotypes for diseases as well as phenotype annotations from model organism databases. This ranking can predict genes that participate in the same pathway, orthologous genes as well as gene–disease associations based on comparing phenotypes alone.

2. Method

2.1. *Preparation of data*

We obtained the raw PhenomeNET dataset (version 16 September 2011) from `http://phenomeblast.googlecode.com`. The dataset consists of a similarity matrix that represents the pair-wise phenotypic similarity between 87,037 complex phenotypes. These complex phenotypes either represent a phenotype annotation available from one of the model organism databases or a disease phenotype from OMIM.

For our analysis, we filter the PhenomeNET similarity matrix for nodes that represent human genes, human diseases and mouse models. As a result, we obtain a square matrix of similarities between 2,964 OMIM diseases and 26,148 genotypes or genes. Each column and row in this matrix represents either phenotypic descriptions of genes and genotypes in the MGI database[3] or phenotypes associated with OMIM genes and diseases.[1]

We use the MGI report `MGI_PhenotypicAllele.rpt` to map mouse alleles to their corresponding gene. This report includes the MGI allele accession identifier and the corresponding MGI gene identifier. To link mouse genes to their human orthologs, we use the human-mouse orthology mapping available from the MGI (MGI report `HMD_Human4.rpt`).

To link the genes in PhenomeNET to PharmGKB, we use the Gene ID available in MGI report `HMD_Human4.rpt`. OMIM records in PhenomeNET are mapped to their corresponding Gene ID using the `mim2gene` file available from the NCBI (`ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/`).

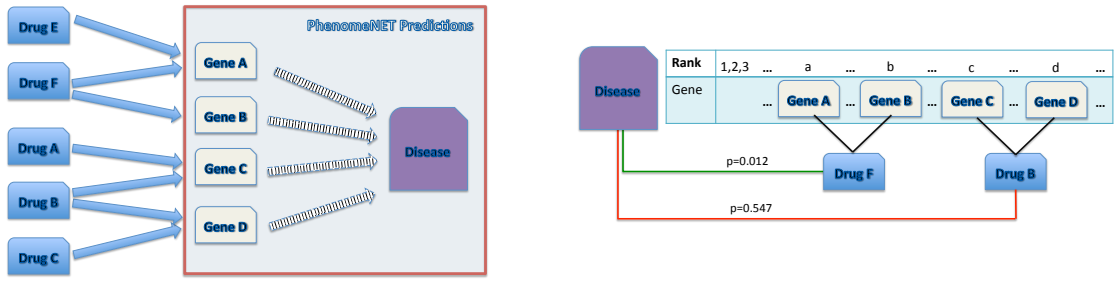
PhenomeNET contains only heritable diseases that are available in OMIM. The PharmGKB, on the other hand, contains a classification based on UMLS. We use the UMLS to identify the subset of diseases in PharmGKB that can be mapped to OMIM and restrict our analysis to these diseases. The files used for evaluation and mapping were obtained on 9 July 2011.

2.2. *Experimental setup*

PharmGKB provides a list of drugs, genes and diseases, and it associates them when they are discussed together in a published article. This co-occurrence in literature may be indicative of biologically meaningful relations between the associated entities. For example, articles that discuss a drug and a gene together may state that there is some *interaction* between the drug and the gene. Similarly, discussing a drug and a disease together may indicate that a drug can play a pharmacological role in the disease.

Our underlying hypothesis is that we can employ criteria that allow us to combine relations to infer new associations. In particular, we assume that, if a drug is associated with a gene (in PharmGKB), and this gene is involved in a disease, then the drug may play a pharmacological role in the disease. In our analysis, we combine the drug–gene associations available from PharmGKB and the predictions of gene–disease relations from PhenomeNET to infer both known and novel drug–disease associations.

2.3. *Statistical testing*



(a) (b)

Fig. 1. **Illustration of parts of our method for establishing drug–disease associations.** Our method uses known drug–gene interactions as input and links them to the gene–disease predictions generated by PhenomeNET (Fig. a). For each disease, PhenomeNET provides a ranked list of genes and genotypes based on their phenotypic similarity to the disease (Fig. b). We identify the subset of the list of genes that is known to interact with a particular drug and performs a Wilcoxon signed-rank test to determine the p-value of observing these genes at this position in the ranked list for that disease. If the genes associated with a drug are distributed uniformly across the ranks, the Wilcoxon signed-rank test will return a relatively high p-value. If the genes associated with a drug are phenotypically very similar to the disease, the test will return a relatively low p-value. Finally, we assign the p-value returned by the Wilcoxon signed-rank test as the value of the association between the drug and the disease.

As a first step in our method, we use the list of drug–gene associations in PharmGKB and the mappings we generate to create a list of PhenomeNET entries for each drug in PharmGKB. As a result, we obtain a list of pairs $(C, < G_1, ..., G_n >)$ such that C is a drug and $G_1, ..., G_n$ are genes or genotypes (in PhenomeNET) with which C interacts.

The second step of our method identifies the distribution of phenotypic similarity for each disease. For each disease that is available in PhenomeNET, we then extract a ranked list of phenotypic similarity of *all* genes and genotypes that we include in our analysis (i.e., all genes and genotypes that have a phenotypic annotation in either the OMIM or MGI database). As a result, for each disease D, we obtain a list of pairs $< G_i, \tau >$ of genes or genotypes G_i and their phenotypic similarity to τ to the disease D. This list includes the genes from the first step of our method, i.e., the genes that have known interaction to particular drugs.

The third step performs a statistical test for significant drug–disease associations. For each disease D, we perform a Wilcoxon signed-rank test for each drug and the its assigned list of gene interactions $(C, < G_1, ..., G_n >)$. The test allows us to identify whether the genes with which C interacts are phenotypically significantly more similar to the disease D than expected by chance. Formally, given a disease D, a list of pairs $< G_i, \tau >$ and a drug C with its interaction partners $< G_1, ..., G_n >$, we perform the Wilcoxon signed-rank test on the distributions of τ_k for $(G_1, \tau_1), ..., (G_n, \tau_n)$ and $< G_i, \tau >$. As a result, we obtain p-values for each pair of a drug and a disease (C, D).

Based on the list of p-values, we perform a correction for multiple testing. We apply two corrections methods to our data set. First, we use the Bonferroni correction to provide a conservative estimate of the p-value for a drug–disease association. However, for many applications it is sufficient to control the false discovery rate in a data set. Therefore, we further apply Benjamini-Hochberg's method (which controls the false discovery rate) to correct for multiple testing.

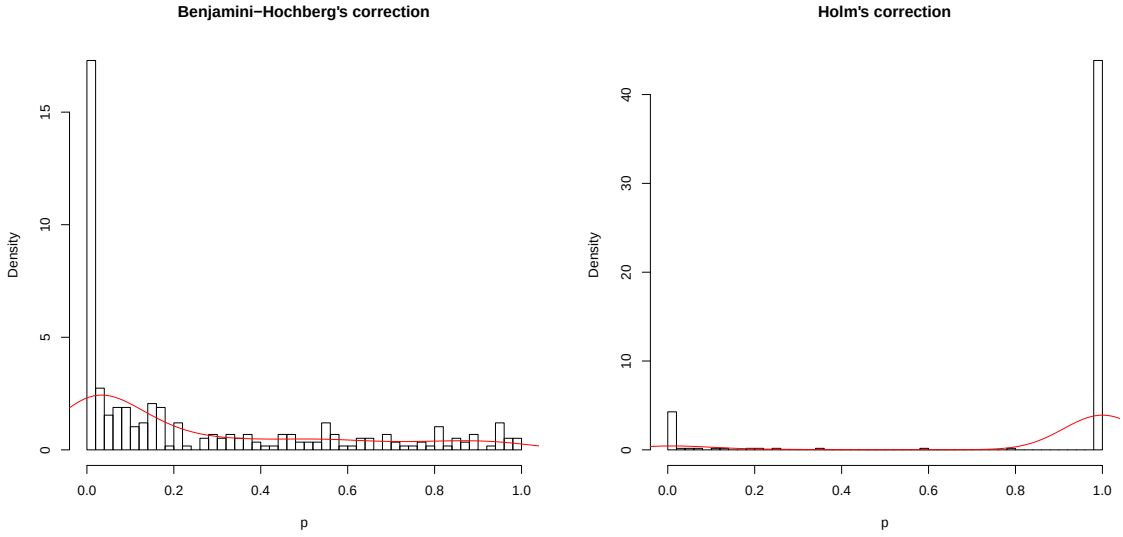


Fig. 2. The figures demonstrate the distribution of p-values for drug–disease associations available in PharmGKB using Bonferroni correction as well as Benjamini-Hochberg's correction for multiple-testing.

3. Results

3.1. *Predicting known drug-disease associations*

The result of our method are p-values for drug–disease associations, and we first identify significant drug–disease associations. In total, using Bonferroni correction, we identify 23,927 drug–disease associations as significant (with $p < 0.1$), and we identify 419,335 drug–disease associations as significant when controlling the false discovery rate using Benjamini-Hochberg's correction. Using a more conservative p-value of 0.0001 as measure of significance, we identify 6,473 significant drug–disease associations when using Bonferroni correction and 55,931 significant associations when using Benjamini-Hochberg's method. Figure 2 shows the distribution of p-values we obtain for the drug–disease associations that are included in PharmGKB.

We then use the manually curated drug–disease assocations available from PharmGKB to evaluate the performance of our method. Within PharmGKB, associations between drugs and diseases are established when both a drug and a disease are discussed together in a scientific article. Such an association does not necessarily indicate that the drug may play a pharmacological role in the disease. Additionally, PharmGKB primarily focuses on genotypes, genes and drugs, and does not provide a comprehensive repository of drug–disease associations. Therefore, we do not only use the drug–disease associations available in PharmGKB in our evaluation, but further use the drug–disease associations from FDA-approved drugs as well as high-confidence (inference score higher than 50) drug–disease associations from the Clinical Toxicogenomics Database (CTD)[7] to evaluate the performance of our method. Our quantitative evaluation is based on analyzing the receiver operating characteristic (ROC) curve for predicting drug–disease associations.

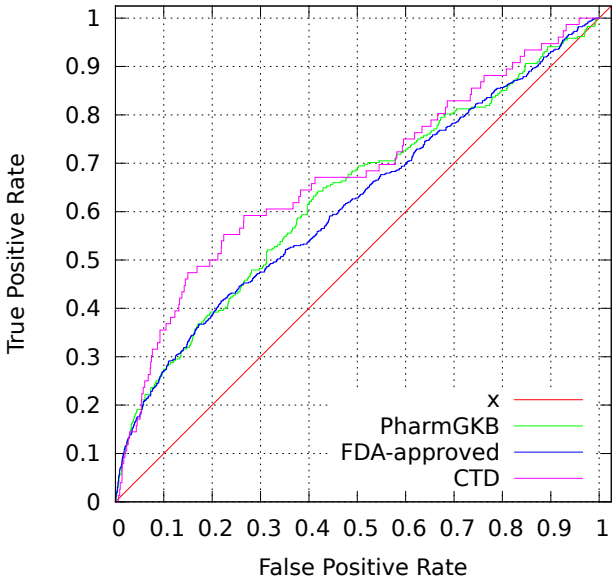


Fig. 3. The ROC curve for predicting known drug–disease associations using our method. We show the ROC curves for comparison with PharmGKB's drug–disease associations, FDA-approved drug indications, and CTD predictions of drug–disease interactions.

A ROC curve is a plot of the true positive rate as a function of the false positive rate. However, while we have a large number of *positive* drug–disease associations (available from PharmGKB), we lack *negative* examples (i.e., drug–disease pairs for which it is *known* that the drug has no effect in the disease). Therefore, we make the simplifying assumption that an *unknown* drug–disease association constitutes a *negative* example. Since our aim is to identify new indications for known drugs, this assumption does not hold for either data set we use for evaluation. Consequently, the result of the ROC analysis represents a *lower boundary* of our method's performance.

The ROC curves resulting from our method are shown in Figure 3. The area-under-curve (AUC) is a measure of the performance of the prediction and equivalent to the probability that a randomly chosen positive example (a known drug–disease association) is ranked higher than a randomly chosen negative example (an unknown drug–disease association).[11] Evaluating against PharmGKB's drug–disease associations yields an AUC of 0.629, against FDA-approved drug indication an AUC of 0.613 and against the CTD dataset an AUC of 0.672.

3.2. *Novel drug-disease associations*

Using the results of our method, we can explore drug–disease association where we predict new uses for existing drugs and potential new drugs for conditions for which they have not been approved or tested. Our criteria for novelty are that the drug-disease relationships are not in the curated knowledge of PharmGKB, not in the list of FDA-approved indications for the drug and not in the high-confidence drug–disease associations available from the CTD.

For example, cystic fibrosis (`OMIM:219700`) is caused by mutations in the *CFTR* gene which is an epithelial apical membrane anion channel regulated by nucleotides and phosphorylation.[30] The disease affects multiple tissues: pancreas, liver, reproductive tract, and the heart, with progressive lung disease accounting for most of the morbidity and mortality. It is an autosomal recessive disease with a mean prevalence of 0.737/10,000 in the EU.[10] The drugs that are significantly associated with cystic fibrosis using our method are levonorgestrel, tretinoin and drospirenone/ethinyl estradiol ($p < 2.65 \cdot 10^{-6}$, $p < 0.002$ and $p < 0.031$, Wilxocon signed-rank test, Bonferroni correction).

Tretinoin (retinoic acid) is already implicated as a potential therapy for cystic fibrosis in the scientific literature. For example, a case report in 2006 showed dramatic improvement in cystic fibrosis patients treated with isotretinoin for acne.[4] Another study shows that retinoic acid can restore the levels of nucleoside diphosphate kinase, which are reduced in CF, and reduce membrane transgutaminase activity, suggesting that retinoic acid could be a novel therapy for some aspects of CF.[36]

Inspection of the genes associated with the functionally related drugs levonorgestrel and estradiol in PharmGKB shows that mouse phenotypes associated with mutations of these genes have a high representation of reproductive phenotypes. Recent reports show that β-estradiol directly rescues F508CFTR functional expression in human cystic fibrosis airway cells[9] and that 17b-estrodiol potentiates activity of the cardiac isoform of CTFR. This may indicate that a potential role for sex steroids in cystic fibrosis therapy may be directly on CFTR itself and not only through action in the reproductive system. CTFR has been known

for some time to be up-regulated by estrogens[32] and down-regulated by progesterone.[26]

Instead of looking for potential drugs for a disease, we can also explore the drug–disease associations by aiming to identify indications for a known drug. For example, the drug *ibuprofen* is associated with four diseases in PharmGKB (hemorrhage, pain, myocardial infarction and stomach neoplasm) and FDA-approved indications of ibuprofen are rheumatoid arthritis, osteoarthritis, pain and primary dysmenorrhea. Our method strongly suggests chronic lymphocytic leukemia ($p < 2.63 \cdot 10^{-23}$, Wilcoxon signed-rank test, Bonferroni correction) as an indication in which ibuprofen may be active. Ibuprofen has been shown to inhibit proliferation and induce apoptosis in promyelocytic cells,[21] and intake of ibuprofen has been linked with significantly decreased risk of leukemia.[22]

3.3. *Interface and availability*

We make our results freely available on our project's website at `http://code.google.com/p/phenomeblast/wiki/PhenomeDrug`. In addition, we enable access to the results of our method through a web server. The webserver enables search for either drugs (from PharmGKB) or diseases (from the OMIM database) and lists the drug–disease pairs as well as their p-values.

4. Discussion

We developed and implemented a method to suggest novel drug–disease associations based on analyzing information about phenotypes available from animal studies. Our initial results demonstrate that we can identify known drug–disease associations and suggest new ones, although our method does not perform as well as comprehensive drug repurposing and discovery frameworks such as PREDICT, which achieve an AUC of over 0.9 for predicting known drug–disease associations.[14] However, these frameworks are commonly based on machine learning and rely on the availability of information about the molecular mechanisms underlying disease (i.e., known gene–disease or drug–disease interactions that can be used to identify the features based on which a classifier can be trained). Our approach does not utilize machine learning, and neither PhenomeNET nor our method for drug-repurposing require any prior knowledge about gene–disease or drug–disease associations. Since our approach relies only on the knowledge of drug targets and the comparisons of phenotypes to establish a link between drug and disease, it can be applied to suggest drugs for orphan diseases of which the molecular basis is unknown.

A major goal for future research is to refine the analysis and evaluation of our method. At the moment, PharmGKB provides associations between drugs and genes based on co-occurrence in scientific articles. Discussing a drug and a gene together may indicate that the drug targets the gene, but can also be indicative of further biological relations. Since our method crucially depends on curated information about drug–gene relations, it may be improved when more specific relations between drugs and genes (or genotypes) become available within PharmGKB. Furthermore, we intend to apply our method to other databases such as DrugBank[23] and the Comparative Toxicogenomics Database[7] which also contain drug–gene interactions. We also plan to include phenotypes of further model organism available in PhenomeNET, such as yeast, fly, worm and fish, to develop an integrated system for the analysis

and prediction of drug–disease associations for rare and orphan diseases based on comparisons of phenotypes.

Performing enrichment analyses over both the diseases that are associated with a drug, using disease-relevant ontologies such as the Human Disease Ontology,[24] and enriching the drugs that are associated with a disease using the chemical classifications of the ATC, MeSH or ChEBI, are areas of further research that may improve the performance of our method.

5. Conclusions

We use the PhenomeNET cross-species phenotype network to provide ranked lists of gene–disease associations. Since the PhenomeNET method performs a comparison of phenotypes directly, it can prioritize genes for orphan diseases of which the molecular basis is unknown. Using this information in our study, we can predict drugs for diseases when the disease mechanisms are not known and therefore suggest potentially new drugs for rare and orphan diseases. The use of PhenomeNET further provides direct links to animal models that can be used to investigate the drug and disease mechanisms as well as the drug's role in the disease.

Animal models of disease play an important role in the investigation of the pathobiology of disease[5] and of drug mechanisms.[35] Our method can identify relevant animal models and thereby improve the speed and reduce the cost required to test and investigate a novel candidate drug. PhenomeNET contains not only phenotype description associated with genes, but rather with the *genotype* of a model organism that belongs to a particular *strain* and within a certain *environment*. The link between PharmGKB and the PhenomeNET resource will therefore allow us to target particular strains and genotypes, and utilize the results from large-scale phenotyping projects such as the IMPC.[3,6] The link to the environmental conditions present in some model organism databases may further improve identification of targets for testing drugs within an *in vivo* environment.

Acknowledgments

Funding for RH was provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502. Funding for PS was provided by an NIH grant, number R01 HG004838-02. Funding for GG was provided by BBSRC grant BBG0043581. Funding for AO and DRS was provided by the European Bioinformatics Institute.

References

1. J. Amberger, C. Bocchini, and A. Hamosh. A new face and new challenges for online mendelian inheritance in man (OMIM). *Hum Mutat*, 2011.
2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, M. J. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. I. Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1), May 2000.
3. J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, J. T. Eppig, and the Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Research*, 39(suppl 1):D842–D848, 2011.

4. J. L. Buckley, M. A. Chastain, and R. L. Rietschel. Improvement of cystic fibrosis during treatment with isotretinoin. *SKINmed: Dermatology for the Clinician*, 5(5):252–225, 2006.
5. C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and J. A. a. Blake. The mouse genome database (mgd): mouse biology and model systems. *Nucleic acids research*, 36(Database issue), January 2008.
6. F. S. Collins, R. H. Finnell, J. Rossant, and W. Wurst. A new partner for the international knockout mouse consortium. *Cell*, 129(2):235, 2007.
7. A. P. Davis, B. L. King, S. Mockus, C. G. Murphy, C. Saraceni-Richards, M. Rosenstein, T. Wiegers, and C. J. Mattingly. The comparative toxicogenomics database: update 2011. *Nucleic Acids Research*, 2010.
8. K. Degtyarenko, P. Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 2007.
9. T. Fanelli, R. A. Cardone, M. Favia, L. Guerra, M. Zaccolo, S. Monterisi, T. De Santis, S. M. Riccardi, S. J. Reshkin, and V. Casavola. Beta-oestradiol rescues deltaf508cftr functional expression in human cystic fibrosis airway cfbe41o- cells through the up-regulation of nherf1. *Biol Cell*, 100(7):399–412, 2008.
10. P. M. Farrell. The prevalence of cystic fibrosis in the european union. *Journal of Cystic Fibrosis*, 7(5):450 – 453, 2008.
11. T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.
12. G. V. Gkoutos, E. C. Green, A.-M. M. Mallon, J. M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1), 2005.
13. G. V. Gkoutos, C. Mungall, S. Dolken, M. Ashburner, S. Lewis, J. Hancock, P. Schofield, S. Kohler, and P. N. Robinson. Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 1:7069–7072, 2009.
14. A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7, June 2011.
15. N. Guarino. Formal ontology and information systems. In N. Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*, pages 3–15. IOS Press, 1998.
16. T. Hernandez-Boussard, M. Whirl-Carrillo, J. M. Hebert, L. Gong, R. Owen, M. Gong, W. Gor, F. Liu, C. Truong, R. Whaley, M. Woon, T. Zhou, R. B. Altman, and T. E. Klein. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic acids research*, 36(Database issue), Jan. 2008.
17. R. Hoehndorf, M. Dumontier, A. Oellrich, D. Rebholz-Schuhmann, P. N. Schofield, and G. V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLOS ONE*, 6(7):e22006, July 2011.
18. R. Hoehndorf, M. Dumontier, A. Oellrich, S. Wimalaratne, D. Rebholz-Schuhmann, P. Schofield, and G. V. Gkoutos. A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*, 27(7):1001–1008, April 2011.
19. R. Hoehndorf, A. Oellrich, and D. Rebholz-Schuhmann. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26(24):3112 – 3118, 10 2010.
20. R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 2011.
21. J. Jakubikova, T. Duraj, X. Takacsova, L. Hunakova, B. Chorvath, and J. Sedlak. Non-steroidal anti-inflammatory agent ibuprofen-induced apoptosis, cell necrosis and cell cycle alterations in

human leukemic cells in vitro. *Neoplasma*, 48(3):208–213, 2001.
22. C. M. Kasum, C. K. Blair, A. R. Folsom, and J. A. Ross. Non-steroidal anti-inflammatory drug use and risk of adult leukemia. *Cancer Epidemiology Biomarkers & Prevention*, 12(6):534–537, June 2003.
23. C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. Drugbank 3.0: a comprehensive resource for omics research on drugs. *Nucleic Acids Research*, 2010.
24. P. LePendu, M. Musen, and N. Shah. Enabling enrichment analysis with the human disease ontology. *Journal of Biomedical Informatics*, 2011. In press.
25. H. Morgan, T. Beck, A. Blake, H. Gates, N. Adams, G. Debouzy, S. Leblanc, C. Lengger, H. Maier, D. Melvin, H. Meziane, D. Richardson, S. Wells, J. White, J. Wood, T. E. Consortium, M. H. de Angelis, S. D. M. Brown, J. M. Hancock, and A.-M. Mallon. EuroPhenome: a repository for high-throughput mouse phenotyping data. *Nucleic Acids Research*, 38(suppl 1):D577–D585, 2010.
26. A. Mularoni, L. Beck, R. Sadir, G. L. Adessi, and M. Nicollier. Down-regulation by progesterone of cftr expression in endometrial epithelial cells: A study by competitive rt-pcr. *Biochemical and Biophysical Research Communications*, 217(3):1105 – 1111, 1995.
27. C. Mungall, G. Gkoutos, C. Smith, M. Haendel, S. Lewis, and M. Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, 2010.
28. B. Munos. Lessons from 60 years of pharmaceutical innovation. *Nature reviews. Drug discovery*, 8(12):959–968, Dec. 2009.
29. M. Oti and H. G. Brunner. The modular nature of genetic diseases. *Clinical Genetics*, 71:1–11, 2007.
30. J. R. Riordan. CFTR function and prospects for therapy. *Annual review of biochemistry*, 77:701–726, 2008.
31. P. N. Robinson, S. Koehler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–615, 2008.
32. L. Rochwerger and M. Buchwald. Stimulation of the cystic fibrosis transmembrane regulator expression by estrogen in vivo. *Endocrinology*, 133(2):921–30, 1993.
33. S. H. Sleigh and C. L. Barton. Repurposing strategies for therapeutics. *Pharmaceutical Medicine*, 24(3):151–159, 2010.
34. C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7, 2004.
35. S. Tickoo and S. Russell. Drosophila melanogaster as a model system for drug discovery and pathway screening. *Current Opinion in Pharmacology*, 2(5):555 – 560, 2002.
36. K. J. Treharne, O. Giles Best, and A. Mehta. Transglutaminase 2 and nucleoside diphosphate kinase activity are correlated in epithelial membranes and are abnormal in cystic fibrosis. *FEBS Letters*, 583(17):2789–2792, 2009.
37. N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7(11):e1000247, 11 2009.