Applications of comparative evolution to human disease genetics
Claire D McWhite\textsuperscript{1}, Benjamin J Liebeskind\textsuperscript{1} and Edward M Marcotte

Direct comparison of human diseases with model phenotypes allows exploration of key areas of human biology which are often inaccessible for practical or ethical reasons. We review recent developments in comparative evolutionary approaches for finding models for genetic disease, including high-throughput generation of gene/phenotype relationship data, the linking of orthologous genes and phenotypes across species, and statistical methods for linking human diseases to model phenotypes.

Address
Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, & Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, United States

Corresponding author: Marcotte, Edward M (marcotte@cimb.utexas.edu)

\textsuperscript{1}These authors contributed equally.

Current Opinion in Genetics & Development 2015, 35:16–24
This review comes from a themed issue on Genomes and evolution
Edited by Antonis Rokas and Pamela S Soltis
For a complete overview see the Issue and the Editorial

Available online 4th September 2015
http://dx.doi.org/10.1016/j.gde.2015.08.004
0959-437X/\textsuperscript{c} 2015 Elsevier Ltd. All rights reserved.

Introduction
In a natural extension of the traditional model organism approach, new data sources and techniques are allowing connections to be drawn between human and model systems, even when phenotypes do not obviously match. As organisms diverge over evolutionary time, the relationship between genes and the phenotypes they encode often also diverge. Many novel phenotypes arise from repurposed gene networks, rather than novel genes \cite{1,2}, while, conversely, molecular networks can lose their associations with conserved phenotypes \cite{3}. Such complexity gives rise to a wealth of potential model systems, each capable of providing useful insights into human disease.

Comparative evolutionary approaches to study human disease are rooted in the traditional use of experimental and genetic data from diverse organisms to explore mechanisms of human genetics. However, new methods for discovering relevant organismal models for human disease are being developed, most notably methods drawing on computer science and evolutionary analyses to incorporate the growing wealth of genetic and phenotypic data in increasingly diverse species.

Here, we review recent advances in using phenotypic, genetic, and evolutionary information in both model and non-model organisms to rationally identify the genetic underpinnings of human disease. Figure 1 introduces a general framework that categorizes elements of comparative approaches. Typical approaches include identifying relationships between genes and phenotypes in a model organism via forward and/or reverse genetics, followed by comparison to human disease via gene orthology, phenotype similarity, or a combination of both. The different components have all been touched by an ever-increasing emphasis on high-throughput methods. Using the framework in Figure 1, we categorize the major approaches taken by researchers, and illustrate these approaches with several examples of how the new methods are applied to the study of human diseases.

Generating data with forward and reverse genetics
High-throughput identification of mutant phenotypes and their underlying genetics often provide the raw material for human disease model identification. The comparative approach often begins with a genetic screen in a model system to identify relevant pathways and genes. Genetic screens are traditionally divided into reverse genetics, which perturbs specific genes and looks for phenotypic effects, and forward genetics, which identifies phenotypes of interest and then uncovers their genetic basis. Generation of mutants by both approaches, and the corresponding large-scale identification of phenotypes, has been revolutionized by use of high-throughput sequencing, synthetic biology techniques, and image analysis.

In particular, over the last decade, reverse genetics has been scaled up by high-throughput knockdowns, especially RNAi screens and comprehensive gene knockout collections in model species that span the tree of life \cite{4–8}. More recently, CRISPR screens, which are potentially amenable to any organism of interest, have been used to create libraries of human cell line knockouts \cite{9,10}, allowing reverse genetics to be applied to human systems;
phenotypic analysis of the resulting CRISPR libraries now seems to be the bottleneck. A group of methods termed ‘Deep Mutational Scanning’ allows the systematic switching of, for instance, every codon in an open reading frame to every other codon in search of phenotype-causing substitutions [11**,12,13].

Similar gains in throughput are now being seen in forward genetics screens as well. Often, forward genetic screens introduce untargeted mutations into a genome, and then identify lines with phenotypes of interest before screening for mutations. Recent efforts have utilized high-throughput sequencing to make this approach feasible in mammals [14] and on much larger scales than in previous generations [15]. Other methods utilize the diversity of natural populations as a basis for phenotypic screening [16*,17]. Platforms in yeast, plants, worms, and fruit flies now exist to record quantitative data on multiple levels, including morphology, metabolism, transcription, and translation [16*,18–21]. Combining these measurements make it possible not only to uncover the biology of model systems, but also to screen human genes and candidate disease alleles in a model system background, a method that Jasper Rine and colleagues termed ‘surrogate genetics’ [22,23*]. Although model organisms will remain a mainstay of human disease genetics, the advent of these novel molecular tools has raised the possibility of high-throughput screens of genotype/phenotype relations in any organism of interest, blurring the lines between model and non-model organism. Importantly, the advances in both forward and reverse genetics have produced hundreds of thousands of gene — phenotype associations across multiple organisms [24–27], providing deep datasets that now make computational analyses of new disease genes increasingly possible.

**Finding models through phenotype comparison**

Traditionally, non-human models of disease are often identified by the direct comparison of a model organism phenotype with traits of a human disease. However, such comparisons have historically relied only on the expertise of researchers, and tended to make use of organism-specific language to describe phenotypes. The development of ontologies, formal hierarchies of descriptive annotations [28–30], now allows researchers to find new human disease models by directly searching for homologous phenotypes, an approach easily scalable to large phenotypic datasets. Multiple ontologies [31–33] have been developed, enabling systematic analyses of phenotypes in a way that is descriptive, robust, programmatically accessible, and extensible across species. Notably, major organismal databases now use ontologies to describe phenotypes, for example as for the Worm [34], Human [35], and Mammalian [36] Phenotype Ontologies.

Formal ontologies allow phenotype databases to be cross referenced, much as researchers might search for homologous sequences across organisms. This functionality can significantly improve the throughput and sensitivity of the comparative approach, and can be used to identify
disease models on the basis of phenotypic qualities alone. As one such example, PhenomeNET [37], for instance, employs the Phenotype and Trait Ontology (PATO) developed by Gkoutos et al., and was used to suggest novel genes involved in the Tetralogy of Fallot, a congenital heart defect. Other algorithms have also made use of phenotype annotations to facilitate the discovery of candidate disease genes [38,39,40,41,42].

Finding models through genetic comparison
Just as phenotype ontologies can be used to identify disease models on the basis of phenotypes alone, appropriate models can also be selected using orthologs [43] of human disease genes identified in model organisms (Figure 1). Although this approach is limited by its reliance on a priori knowledge of the genetic basis of a disease, the basic step of determining gene orthology between organisms of interest still forms the core of most comparative studies. This is because orthologs, which are separated historically only by speciation events, tend to be more closely related in function than paralogs [44–46], which result from shared ancestral gene duplications and can often partition an ancestral function, or take on whole new functions [47–49].

New methods for inferring orthologs, and new databases for storing this information are proliferating. As of 2015, there are at least 37 different orthology databases using a variety of algorithms, reviewed in detail by Sonnhammer et al. [50]. Benchmarking suggests that no one method outperforms all others, with methods differing in their precise definition of ‘orthology’ as well as in their tendency to favor either precision or recall for discovery of correct orthologs [51,52]. Meta-analyses that compare and compile information from different algorithms and databases are therefore expected to significantly improve performance [53,54]. Another promising direction is to use information about the species phylogeny to probabilistically infer gene tree inference [55–57]. When the gene of interest for a disease has not yet been identified, computational strategies now exist for prioritizing potential candidate genes, as we discuss next, but even these methods usually require knowledge of orthologs as a starting point.

Statistically associating genes and diseases
The methods described above either rely on prior information about the genetic nature of a human disease, or on a clear phenotypic similarity between organisms. However, phenotypes arising from conserved genetic pathways may have diverged so far that their homology is unrecognizable. A number of methods have been developed to derive useful information from such cases, as well as to facilitate the synthesis of data from multiple species (Figure 2). These methods have the added benefits of identifying both novel human disease genes and appropriate model systems for studying those genes. A common principle of these methods is to group genes together by some criterion that reports on function (‘statistical association’). These groupings then enable the statistical inference of novel disease-associated genes. Methods vary in way that they group genes and associate them with phenotypes (Figure 2). Below, we highlight some of the most common methods, and some recent innovations in this area.

Phylogenetic profiles
Genes with linked function tend to have similar patterns of presence/absence across species, and this presence/absence vector is termed a ‘phylogenetic profile’ (Figure 2a). Phylogenetic profiling allows the search for candidate genes which may have co-evolved with the disease-linked gene, and be involved in the same disease-causing process. Inferring gene functions by their phylogenetic profile, as proposed by Pellegrini et al., is not new [58–60], however, these methods are increasingly being applied to the genetics of human disease [61]. Recent improvements to this technique, such as using orthologous groups of genes, weighting by species divergence, and ancestral state inference, have increased the power of phylogenetic profiling methods, and by extension their application to candidate gene discovery [62,63,64,65]. As one recent example, Dey et al. applied phylogenetic profiling to discover genes involved in ciliary and centrosomal defects [62].

Gene set overlap approaches
Gene set overlap methods determine if two groups of phenotype-associated genes in different species significantly share (orthologous) members (Figure 2b). These methods employ a statistical model to test the significance (commonly, the hypergeometric probability) of two phenotype-associated groupings from two species sharing a set of orthologous genes. This has proven to be an effective approach for identifying extremely divergent model phenotypes which employ the same genetic pathways involved in human diseases [27,66], such as, for example, the identification of a plant model of human Waardenburg syndrome [66]. Because these phenotypes employ orthologous genetic mechanisms, they have been termed ‘phenologs’ [66]. The original approach inferred pairwise phenologs [66], but has been subsequently improved upon by extension to multiple species [27]. The success of this extension indicates an important fact: more comparative data means more inferential power for discovering novel genes associated with human disease.

Involvement in a particular phenotype is not the only way to classify genes in overlap-based studies. Korcsma et al. used pathway annotations in model species to predict more complete, integrated human signaling pathways [67]. Increasingly, multiple sources of data are being
combined in databases [68,69] that allow comprehensive comparison of overlap between gene sets.

**Network approaches**

Gene network-based approaches use networks to provide statistical frameworks for inferring new gene functions or disease associations. A network is first built from interactions between genes (or their encoded proteins); in a comparative gene-discovery framework, these interactions may be experimentally derived from multiple species. The resulting network can then be used to propagate information from genes (network nodes) whose function is known, to genes of unknown function (Figure 2c [70]), using the interactions (network edges) to functionally annotate new genes [71,72]. In principle, edges in the network can incorporate interaction data gleaned from any data source or species, and are therefore often a preferred method of generating consensus annotations. Many kinds of functional annotations can be propagated and therefore predicted, including new gene annotations [73] and disease gene associations, such as might arise from genome-wide association scans, for example as shown for Crohn’s disease using the human gene network HumanNet [71]. There are many methods used for information propagation (reviewed in depth in Wang and Marcotte [70]). Networks have also been constructed from genetic interactions, as might be gleaned from, for instance, double deletion screens [74], and used to find...
Recent applications of statistical linking methods to identify and treat human disease. (a) Cha et al. [84**] used overlap between vasculature genes and genes linked to antifungal sensitivity in yeast to identify TBZ as a novel vascular disrupting agent. TBZ disassembled vasculature in Xenopus embryos (top panels) and slowed human fibrosarcoma tumor growth in mice (bottom panels). Adapted from Cha et al. [84**]. (b) Tardiff et al. [88] overexpressed α-synuclein and screened for phenotype rescuing compounds. One such compound, NAB, in turn ameliorated neuronal proteinopathies in worm neurons. Adapted from Tardiff et al. [88]. (c) Chan et al. [90**] identified divergent phenotypes in planarians and schistosomes in response to the small molecule PZQ. Ca<sub>v</sub>, voltage-gated calcium channel; Na<sub>v</sub>, voltage-gated sodium channel; HVA, high-voltage activated; LVA, low-voltage activated. Adapted from Chan et al. [90**].
new genetic modifiers [75]. Many gene networks are now available online that combine evidence from different interaction types, enabling the use of network-based inferences in most major model organisms [76,77,78*, 79,80**].

In one particularly interesting recent application of gene networks, the networks are not just used for candidate gene prioritization or annotation. Vidal et al. have suggested that disease phenotypes can be viewed as disruptions between interacting genes within the network structure (‘edgetics’) [81,82,83**], suggesting a wider use for gene networks in human disease research in guiding the disruption of only some, but not all, of a given gene’s interactions to affect a specific biological outcome.

Recent applications
In reviewing the methods above, we have focused on the discovery of novel genes associated with human disease. One ultimate goal of these studies is to identify novel therapeutic agents that ameliorate the disease, which can be a challenge even when the target is known. We briefly highlight three recent studies that use the methods outlined above to identify novel drugs that target cancers, neurodegenerative diseases, and parasites (Figure 3).

One of the predictions of the original phenolog study [66] was a yeast gene set that models vertebrate angiogenesis, a key dependency of tumor growth. Cha et al. [84**] used prior information about gene-drug genetic interactions between the yeast pathway and a variety of small compounds [85] to prioritize drugs that might block blood vessel growth. They identified thiabendazole (TBZ) as a candidate angiogenesis inhibitor, and found that not only did it indeed prevent vascularization in Xenopus embryos, but it disrupted pre-existing immature vasculature and slowed fibrosarcoma tumor growth in a mouse model, making it the first such vascular disrupting agent with FDA approval for human use (here, for its antifungal activity). TBZ illustrates that guilt-by-association approaches can be predictive, even between yeast and vertebrates.

Yeast is especially useful for high-throughput drug and genetic screens, as shown by recent work on the protein α-synuclein. This protein is associated with Parkinson’s disease and related neurodegenerative diseases, termed synucleinopathies [86]. Susan Lindquist et al. used drug screens to identify compounds that inhibited aggregation of the protein α-synuclein exogenously expressed in yeast [87]. They identified a class of compounds called N-Aryl Benzimidazoles (NABs) that inhibit α-synuclein aggregation in yeast cells and animal neurons [88]. They further utilized the genetic tractability of yeast to identify pathways affected by α-synuclein [89**], suggesting potential new therapeutic avenues for synucleinopathies.

Whereas these two comparative approaches were applied to understand human genetic disease, the same approaches can also inform on infectious disease. Chan et al. [90**] used planarians, a model platyhelminth worm, to identify the target of the drug praziquantel (PZQ), which kills schistosomes, the pathogenic platyhelminths that cause schistosomiasis, but whose molecular target is unknown. Remarkably, while PZQ kills schistosomes, it causes an axis-duplication phenotype in planarians. This axis duplication phenotype, unlike death, can be explored using RNAi screens, readily applied in planarians. Using this method, Chan et al. identified the cellular basis of the axis duplication phenotype and also identified novel gene targets, as well as new compounds that phenocopy the effect of PZQ and are therefore candidate anti-schistosomal agents.

Outlook for the future
We have touched on recent work using comparative approaches to relate human phenotypes to those in model organisms. We expect opportunities for connecting human genetics with the genetics of non-model organisms to increase considerably over the near time. In particular, data are increasingly available on human genetic variation, including familial inheritance, most recently across the Icelandic population [91], and increasingly provide a reference of human genetic variation for comparative approaches. It also seems a safe bet that the capacity for high-throughput CRISPR screens will dramatically increase known gene-phenotype associations from ever more diverse organisms, including non-traditional models. These developments will only increase in power and accuracy of comparative genomic approaches, and going forward, such methods will serve as a foundation to discover trends across life that point to the cause and treatment of human disease.

Acknowledgements
This work was supported by a fellowship from the NIH to B. J. L. and grants from the NIH, NSF, CPRIT, and the Welch Foundation (F-1515) to E.M.M.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest


42. Robinson PN, Webber C: Phenotype ontologies and cross-species analysis for translational research. PLoS Genet 2014, 10:e1004268.


45. Altenhoff AM, Studer RA, Robinson-Rechavi M, Desimoz C: Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 2012, 8:e1002514.


Written by the Quest for Orthologs Consortium, this review highlights the state of the art for the accurate characterization of orthologs across species, which forms the foundation for many comparative analyses. They also introduce new community standards, such as benchmarking datasets and formats.


The authors examine phylogenetic profiles of human disease genes to characterize the evolutionary trait of genes involved in disease. They suggest that model organism selection should be done on a disease-by-disease basis, and informed by evolutionary history of the target disease genes.


The authors apply phylogenetic profiling to groups of orthologous genes in order to predict involvement of uncharacterized genes in functional modules. They predict and experimentally validate novel WASH complex interactors and cilia/basal components.


The authors determine that genes associated with a disease will have similar phylogenetic profiles. They used this approach to identify cofactors of disease associated transcription factors.


Integrating gene set overlap and network analysis, MORPHIN takes a set of model organism genes as input, and uses an orthology-based projection of those genes onto a human genome network to suggest relevant human diseases.


The authors describe an approach which uses yeast genetics to identify gene networks relevant to a particular human disease. They demonstrate their approach by identifying and validating potential therapeutic targets for the treatment of muscular atrophy disorder.
86. Cha HJ, Byrom M, Mead PE, Ellington AD, Wallingford JB.


90. Chan JD, Agbedanu PN, Zamanian M, Gruba SM, Haynes CL.