Computational tools for prioritizing candidate genes: boosting disease gene discovery

Yves Moreau and Léon-Charles Tranchevent

Abstract | At different stages of any research project, molecular biologists need to choose — often somewhat arbitrarily, even after careful statistical data analysis — which genes or proteins to investigate further experimentally and which to leave out because of limited resources. Computational methods that integrate complex, heterogeneous data sets — such as expression data, sequence information, functional annotation and the biomedical literature — allow prioritizing genes for future study in a more informed way. Such methods can substantially increase the yield of downstream studies and are becoming invaluable to researchers.

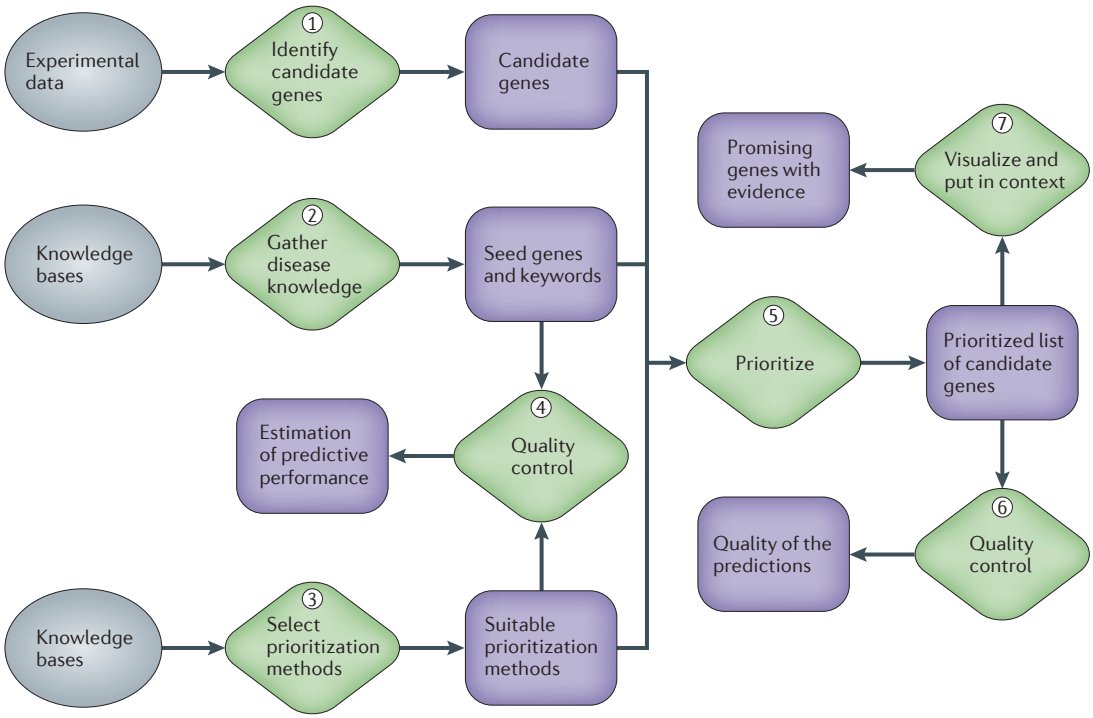
Homozygosity mapping
A form of recombination mapping that allows the localization of rare recessive traits by identifying unusually long stretches of homozygosity at consecutive markers.

Department of Electrical Engineering ESAT-SCD and IBBT–KU Leuven Future Health Department, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.
e-mails: yves.moreau@esat.kuleuven.be; leon-charles.tranchevent@esat.kuleuven.be
doi:10.1038/nrg3253
Published online 3 July 2012

Gene prioritization aims to identify the most promising genes (or proteins) among a larger pool of candidates through integrative computational analysis of public and private genomic data. Its goal is to maximize the yield and biological relevance of further downstream screens, validation experiments or functional studies by focusing on the most promising candidates. Bioinformatics techniques for prioritization are useful at several stages of any gene-hunting process. These bioinformatics tools were initially developed to help to identify the disease-causing gene within a multigene locus that has been identified by a positional genetic study, as they allowed focusing the resequencing of case and control samples on a few of the most likely candidate genes[1–3]. For instance, a linkage analysis on patients with anauxetic dysplasia identified a locus on 9p13–p21 (REF. 4). Prioritization of the 77 genes from this locus using GeneSeeker[5] pinpointed RNA component of mitochondrial RNA-processing endoribonuclease (*RMRP*) as a promising candidate, for which mutation in disease cases was then confirmed by sequencing[4]. Homozygosity mapping followed by mutation screening of the most promising candidates[6–9] is another typical scenario for gene prioritization. For instance, GeneDistiller[10] was used to prioritize 74 genes from a 2 Mb region on chromosome 17 that is associated with cardiac arrythmias, and a mutation in the top-ranking gene *PTRF* (also known as *CAVIN*) was found[7]. Similarly, Gentreprid[11] was used to prioritize the 200 genes from a 10 Mb locus on chromosome 17 that is associated with spondylocostal dysostosis; a disease-specific variant within hairy and enhancer of split 7 (*HES7*) was then identified through sequencing[6]. Even in such simple scenarios, the task of identifying which genes from a given locus potentially underlie a monogenic disease would be laborious without the automation provided by gene prioritization tools. Manually reviewing the literature and perusing public databases of functional annotation (such as Gene Ontology[12] and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[13]), sequence data (such as Ensembl[14] or the UCSC Genome Browser[15]) or expression data (such as ArrayExpress[16] or Gene Expression Omnibus[17]) is a daunting task. Furthermore, prioritization methods have since proved to be applicable in many other situations, such as in more complex genetic studies of contiguous gene syndromes, genetic modifiers, acquired somatic mutations at multiple loci or genome-wide association studies (GWASs)[18–21]. For instance, using G2D[22] identified 10 potential candidate genes for asthma, and a subsequent association study of 91 SNPs in these genes found a variant in protein tyrosine phosphatase, receptor type E (*PTPRE*) that is associated with early-onset asthma[23].

Beyond positional disease gene identification, gene prioritization can be used to identify promising candidates from many studies that generate gene lists, such as differentially expressed genes from microarray or proteomics experiments or hits from RNAi screens or proteomics pull-down experiments. This broadening of applications is beginning to be reflected in the tools themselves: although the tools have a historical bias towards prioritization of human disease genes, methods are emerging that are tailored towards other applications, such as to select genes for a genetic screen in a model organism[24].

Box 1 | **Gene prioritization workflow**



The first step in gene prioritization consists of building the list of candidate genes to prioritize. Typical lists come from linkage regions, chromosomal aberrations, association study loci, differentially expressed gene lists or genes identified by sequencing variants. Alternatively, the complete genome can be prioritized, but substantially more false positives would then be expected. Step two consists of collecting prior knowledge about the disease, in the form of seed genes (known disease genes) or disease-relevant keywords, through knowledge bases or text-mining tools that collect data about diseases or biological processes. For seed genes, it is essential to review each gene across such databases or to use expert knowledge to make sure that it is truly relevant. Also, if the set contains too few genes, the pattern will be insufficiently informative, whereas if the set is too large, the pattern will often be molecularly too heterogeneous to be useful. In our experience, good sets of seed genes contain between 5 and 30 genes. Step three consists of selecting prioritization methods that best match the specific task (BOX 3). In some cases, little or no prior knowledge is available, and in these cases seed genes cannot be readily collected, and only some methods remain applicable (see the main text). Step four is the crucial step of assessing whether the selected seed genes, keywords and tools are suitable and whether reliable predictions can be expected. Cross-validation makes it possible to assess whether a set of seed genes provides a coherent pattern (see the 'Statistical benchmarking by cross-validation' section of the main text). It is also advisable to create multiple sets of seed genes or keywords covering complementary phenotypic aspects of the disease and to assess their performance separately. In step five, the actual prioritization takes place, possibly using multiple tools or multiple sets of seed gene or keywords. These results can also be combined hierarchically to obtain a consensus result (see 'Carrying out complex strategies' in the main text). At this stage, an optional step is to perform a quality assessment of the global prioritization results to make sure that they are relevant (step six): for example, using functional enrichment (see 'Other quality-control methods' in the main text). Finally, step seven consists of interpreting the results using the prioritization tools themselves or other third-party tools to identify relations between candidate genes and known disease genes to guide the final the selection of genes for experimental validation. For instance, if a top-ranking gene contains variants that are associated with phenotypically related disorders or to relevant traits in animal models, this provides strong support for a candidate. Also, confirmed or predicted physical binding between the products of a seed gene and a top-ranking candidate will immediately direct the validation experiment.

Guilt by association
A statistical rule of thumb that asserts that reliable predictions about the function or disease involvement ('guilt') of a gene or protein can generally be made if several of its partners (for example, genes with correlated expression profiles or protein–protein interaction partners) share a corresponding 'guilty' status ('association').

Gene prioritization methods (BOX 1) typically involve two inputs: a list of candidate genes for prioritization and the criteria for prioritization, such as for the involvement in a particular disease or cellular process. These prioritization criteria are typically in the form of biological keywords or a set of 'seed' genes (also known as training genes) that are already linked to that disease or process. The methods are based on the well-established concept of guilt by association[25,26], (see REF. 27 for a review on the use of guilt by association in the context of disease gene discovery). They query databases that contain webs of simple relations between genes or proteins (such as protein–protein interaction (PPI) data[28]) to discover unexplored relations between those entities. Thus, genes can be prioritized on the basis of putative links to other genes that have more established roles in the disease or process of interest. For example, a gene could be prioritized for a role in a disease if PPI data show that its protein product is found in a multiprotein complex with other proteins in which some mutations are known to cause the disease or

a phenotypically related disease[29]. For instance, receptor-interacting serine/threonine protein kinase 1 (*RIPK1*) was proposed as a novel candidate for inflammatory diseases through the identification of a protein complex that links RIPK1 with genes that are involved in inflammatory diseases[29].

The explosion in large-scale 'omics' data, such as high-throughput sequencing data, has created a pressing need for effective gene prioritization tools[30]. In turn, the tools have been developing quickly owing to innovative advances in machine learning methods for the integration of complex heterogeneous data[31–34] and broad public availability of omics data. This Review primarily aims at helping molecular biologists and geneticists to incorporate gene prioritization into their gene discovery projects. Because gene prioritization tools have become easy to use, this article is targeted at biologists rather than bioinformaticians, in contrast to more technical reviews on gene prioritization[35–38]. To this end, this article provides a novel tutorial component that bridges the gap for biologists towards adopting prioritization methods. In this Review, we discuss key principles of computational methods for prioritization, guidelines for assessing the results of prioritization and finally some future perspectives for improving gene prioritization and extending its scope. Our discussion of how to carry out complex prioritization strategies and of how to assess prioritization results addresses two crucial issues for biologists, which were covered only marginally in previous reviews. The goals of this Review are to allow readers to distinguish between the key features of different gene prioritization tools, so as to allow selection of a suitable tool for their specific purpose, to avoid some pitfalls of such methods and to carry out a simple prioritization task in practice using some of the available Web applications.

Gene prioritization tools and data sources

Many tools are now available for candidate gene prioritization. However, different tools use different data sources (BOX 2) and also compile different relations between genes and then combine this information in different ways[39]. Data are highly heterogeneous (for example, sequence, expression, PPIs, annotation and literature) and lead to various relevant relations that can be detected between genes: sequence homology, co-expression[40], PPIs[20,41], shared functional annotations or co-occurrence in literature abstracts[42]. No single source of data can be expected to capture all relevant relations. For example, PPI data cannot capture transcriptional regulation, whereas expression data will fail to detect many effects of post-transcriptional modifications. Thus different data types are complementary and need to be merged to provide broader coverage than any single data source and to infer stronger relationships through the accumulation of evidence. Several general strategies are available for this integration, such as creating information profiles across different sources and matching candidates against those profiles[1,2] or using network algorithms to capture putative relationships[43].

There is now a wealth of gene prioritization tools, and their technical details (such as inputs, outputs and computational methods) have been reviewed in several articles[35–38]. To help the reader to get acquainted with the different tools quickly, in BOX 3 we describe the Gene Prioritization Portal[35], which hosts links to most of the prioritization tools made available over the Web by various research groups and which helps users to select the right tool for their needs. In the present article, we focus on key principles and potential pitfalls for biological users rather than on exhaustive technical details.

Approaches for gene prioritization

Selecting a prioritization strategy. The precise prioritization strategy is influenced by the set of candidate genes and is tailored to the type of biological question that is being answered. As an example, from the same list of candidates from a cancer genome project on metastasis, different researchers might prefer to look for genes that are related to vasculogenesis or alternatively to cell–cell adhesion processes.

The capacity for downstream experiments is also a major consideration. For low-throughput validation and functional characterization (for example, *in vivo* studies), prioritization would be stringent so as to result in an output of only a few genes. However, to elucidate a large portion of a pathway or to perform a medium-throughput RNAi or genetic interaction screen, tens or hundreds of output genes would be more appropriate. The type of biology being studied will also influence the number of genes, both in the input candidate list and the number of prioritized genes undergoing downstream analysis. The input candidate list could comprise genes from a single locus, multiple loci or lists from omics experiments, or it could even comprise an agnostic approach towards candidates by prioritizing the whole genome. The number of output genes characterized will depend on whether a single gene for a monogenic disease is sought or rather whether multiple genes could be relevant, such as among a set of differentially expressed genes that might underlie a particular disease state. Finally, the level of prior knowledge (BOX 2) about the biological process is an important consideration. Prioritization strategies for adding a novel gene to a well-characterized disease or pathway differ from those for which limited or no prior knowledge is available about the molecular basis of the disease[44], because it is difficult to identify enough relevant seed genes or keywords. All factors mentioned above influence the choice of a suitable prioritization tool.

Gathering candidate genes. Carefully selecting a set of genes among which to search for promising candidates greatly influences the quality of the prioritization. Candidate genes can be obtained from primary or secondary data sources (BOX 2). Researchers still tend to carry out research by first designing an experiment and generating primary data. However, so many secondary data are now available that it is often worthwhile first to analyse secondary data and to prioritize them, as a pilot study for evaluating feasibility, refining the original biological question and informing the experiment design — or possibly as a way of skipping primary data generation entirely.

Machine learning methods
The design and development of algorithms that allow computers automatically to learn to recognize complex patterns in data and to make intelligent decisions on the basis of such data.

Box 2 | **Biological data sources**

There is a plethora of databases that contain large amounts of relevant gene and protein data, such as sequences, molecular functions, roles in pathways and biological processes, expression profiles, regulatory mechanisms, interactions with other biomolecules and biomedical literature. Such biological data sources are at the core of gene prioritization methods, because prioritization algorithms sift through these data to create a computational model of promising candidates. The integration of high-quality biological data sources is necessary, but not sufficient, to obtain accurate predictions.

Data standardization and interoperability

Acquiring and merging numerous sources of heterogeneous data present severe technical challenges. First, multiple identifiers are available for genes, transcripts and proteins (such as HUGO Gene Nomenclature Committee names, Ensembl gene identifiers, Affymetrix probe identifiers or SwissProt identifiers), and there is not necessarily a one-to-one relationship between them. Thus, data from different sources will need to be appropriately mapped and merged[121,122]. Moreover, information about diseases, phenotypes and biological processes is far from being fully standardized. Ontologies, which can be seen as logically structured computer-processable vocabularies, are of great help for computers to retrieve and process complex data sets. Relevant examples here include Gene Ontology[12], Human Phenotype Ontology[123] or Disease Ontology[124]. Some data sets are easily retrievable over the Web in a well-structured format — for example, data that are retrievable through the Ensembl BioMart[125] — whereas in other cases format might be subject to change over time, or identifiers might become obsolete. Furthermore, data sets are not static, and the data underlying gene prioritization tools need to be updated regularly. However, because it is difficult for all mapping and merging steps to be carried out automatically across numerous data sources, it is still a major challenge for developers of gene prioritization tools to update the data underlying their tools frequently. The gradual adoption of semantic Web technology[114], which aims to improve the interoperability of Web resources, will alleviate such problems over time.

Data representation

Different data sources are represented in multiple heterogeneous ways. Indeed, whether the data are presented as a matrix of numbers (for example, expression data), as a graph (for example, protein–protein interactions) or as lists of terms (for example, keywords extracted from MEDLINE abstracts) will influence the way in which these data will be analysed and used for prediction. For instance, sequence data are best analysed using dedicated tools, such as BLAST for sequence alignment. Expression data and other vector data can be analysed using basic techniques (for example, correlation), as well as more advanced techniques (for example, principal components analysis or clustering). For gene and protein networks, which are popular because of their seemingly easy interpretation, different and specific strategies have been developed (for example, shortest paths or random walks). Last, annotation data are a particular case of vector data and are characterized by the use of ontologies (that is, hierarchical relations between concepts). Methods that take into account the structure of ontologies are therefore preferred for analysing such data.

Primary and secondary data

An important distinction regarding data sources should be made between primary and secondary data. Primary data are data that are specifically generated (typically in-house) to answer a biological question. Such an example would be a microarray experiment in which the experimental design is dedicated to answering your question. Secondary data are data that are available through public repositories (such as ArrayExpress, Gene Expression Ominbus, Ensembl or the UCSC Genome Browser) or through large in-house facilities independently of the biological question being asked (and are usually made available by third parties).

Metagenes

The definition of a gene is typically a rather vague concept in gene prioritization methods. Usually, no distinction is made between genes and their corresponding proteins or between alternative transcripts or protein isoforms. Furthermore, information might be transferred across species through homology (especially orthology) relations. So, the genes we refer to are in fact 'metagenes', collapsing together the notions of genes and proteins, possibly across species. This makes it challenging to collect species- or isoform-specific information in an automated fashion or to use such information in prioritization tools. In particular, cross-species data integration raises the classical problems of identifying orthologous genes[126] and interologue protein–protein interaction gene pairs[127] and of how to transfer functional information accurately across species.

Data and knowledge

The terms data and knowledge are often used indiscriminately, even though they provide useful semantic distinctions in terms of levels of abstraction and relevance. As an example, gene expression profiles generate raw and normalized data, whereas the fact that gene A is a transcription factor that regulates gene B is a form of knowledge. Data are detailed but their meaning is loosely organized, whereas knowledge is highly structured and has a clear and usable meaning. Dedicated algorithms must be used on data to detect relevant biological signals and thus to extract information. Gene prioritization relies both on those data sources that contain knowledge and those that contain data. By doing so, it can make predictions that are accurate (by relying on knowledge to suggest potential relationships among well-characterized objects) as well as novel (by relying on data to detect unexpected or previously uncharacterized relationships). Note that the overrepresentation of well-characterized genes in relationship databases creates a 'knowledge bias' because those well-characterized genes tend to be favoured over potential novel discoveries (see the main text).

Principal components analysis
A statistical method that is used to simplify a complex data set by transforming a series of correlated variables into a smaller number of uncorrelated variables called principal components.

Interologue
A protein–protein interaction that is conserved between orthologous proteins in different species.

Alternatively, the entire genome can be prioritized, but this can generate large, unmanageable lists of prioritized genes. It is also challenging to assess how strong the results of the prioritization are (see below). Indeed, if genes that are already known to be involved in the biological process are not among the top results of the genome-wide ranking, it is then difficult to assess whether high-ranking genes are false positives or not. There has been, however, at least one success for Parkinson's disease. CAESAR was used to prioritize the whole human genome; from a mutation screen across two of the top ten genes, five variants associated with Parkinson's disease were identified in the South African population[45].

Prioritization criteria based on keywords or known seed genes. The criteria that are used to prioritize a set of candidate genes are typically in the form of keywords or seed genes. The advantage of keywords is that they are easy to formulate and to gather. However, their expressive power is actually lower than would intuitively be believed, and if expression of more complex relations is needed, keywords quickly result in complex queries or long lists of largely irrelevant output genes. Also, keywords capture only explicit relations, and if an important biological aspect is missing (for example, the involvement of some key pathways), this knowledge will not be captured by the gene prioritization.

The collection of seed genes is more time consuming, but it is a flexible way to formulate complex queries implicitly, and it can capture aspects of the process of which we may not be aware (through the shared characteristics of the seed genes).

Selecting keywords or seed genes. Choosing appropriate keywords or seed gene lists are not trivial exercises. Poorly informative genes or keywords should be avoided. For example, disease biomarkers can be bad choices because they are often only indirectly linked to the disease and will weaken the homogeneity of the gene set. Similarly, general keywords that are weakly associated to the disease are likely to introduce noise in the analysis. The key is to focus on relevant information to obtain a consistent functional pattern that will be recognizable in good candidates. For example, in a study of genes that are involved in cancer progression in squamous cell carcinoma, the more specific term 'squamous cell carcinoma' would be preferable as a keyword to the overly broad term 'cancer'.

Several databases collect phenotypic information both about diseases and about their associated genetic factors and are thus useful sources of keywords and seed genes (reviewed in REF. 46). For instance, Online Mendelian Inheritance in Man (OMIM) is a manually curated knowledge base for genetic disorders with Mendelian inheritance[47,48]. Each OMIM disease entry contains a gene–phenotype relationship table that can be used to identify the known disease genes and a general description that can be used to identify relevant keywords. The Genetic Association Database (GAD) focuses on association studies of complex disorders[49] and can therefore be used to identify causative variants. Because they are based on manual curation, knowledge bases are sometimes incomplete, and additional strategies are required to get the latest data. For instance, GoPubMed[50] mines MEDLINE using biomedical ontologies to associate ontological terms and genes to the biological process of interest and therefore can be used to retrieve both genes and keywords from the scientific literature. Also, commercial systems, such as Ingenuity Pathway Analysis[51,52], MetaCore from GeneGo[53,54] and the Human Gene Mutation Database[55,56], contain manually curated disease–gene associations that might not be available through public databases.

Computational strategies for gene prioritization. Prioritization tools typically produce their outputs either by filtering the candidate genes into smaller subsets or by ranking the candidate genes (FIG. 1; see also reviews in REFS 36,37). In light of the properties that an ideal gene should fulfil, filtering reduces the list of candidates into a smaller list of output genes by assessing those criteria using the available data (FIG. 1a). For example, TEAM filters genes on the basis of their function (from Gene Ontology) as well as their association status (from GWASs)[57]. Furthermore, Biofilter integrates several more databases and includes pathway annotations and PPIs[58]. The main limitation of such methods is that the strict filtering process does not allow for a fine analysis of the candidate set. If a relevant gene fails to meet just one of the criteria, it is simply filtered out and thus becomes a false negative.

By contrast, ranking methods tackle this limitation by ranking candidates from most promising to least promising. They can combine multiple viewpoints or criteria but avoid the hard thresholding of filtering methods. Ranking methods can roughly be classified into three categories: text mining[59,60], similarity profiling and network analysis[43,61–63] (FIG. 1b–d). Text mining gathers all methods that only rely on the use of text data (FIG. 1b). First, a set of keywords or knowledge fragments is used to retrieve a set of documents (for example, abstracts) that are relevant to the disease

Box 3 | **Gene Prioritization Portal**

The Gene Prioritization Portal is an online resource that is designed to help biologists and geneticists to select the prioritization methods that best correspond to their needs. It is frequently updated and currently describes 33 publicly available prioritization tools by the inputs they require (such as genes or keywords), the outputs they produce (such as a prioritized list or a gene selection through filtering) and the data they use (for example, text-mining data, expression data — see BOX 2 for more details)[35]. A search page can be used to identify the best tools for use in different situations, such as prioritizing genes in a chromosomal locus from a linkage analysis, prioritizing genes in the absence of known disease genes or incorporating user-specific genomic data sets in the prioritization. This portal is also a repository of experimental validation studies that demonstrate the ability of prioritization methods to identify promising candidate genes and therefore to speed up disease gene discovery (see FIGS 2,3 for two illustrative examples). In addition, recent reviews can be used to determine which methods are most suitable[36–38].

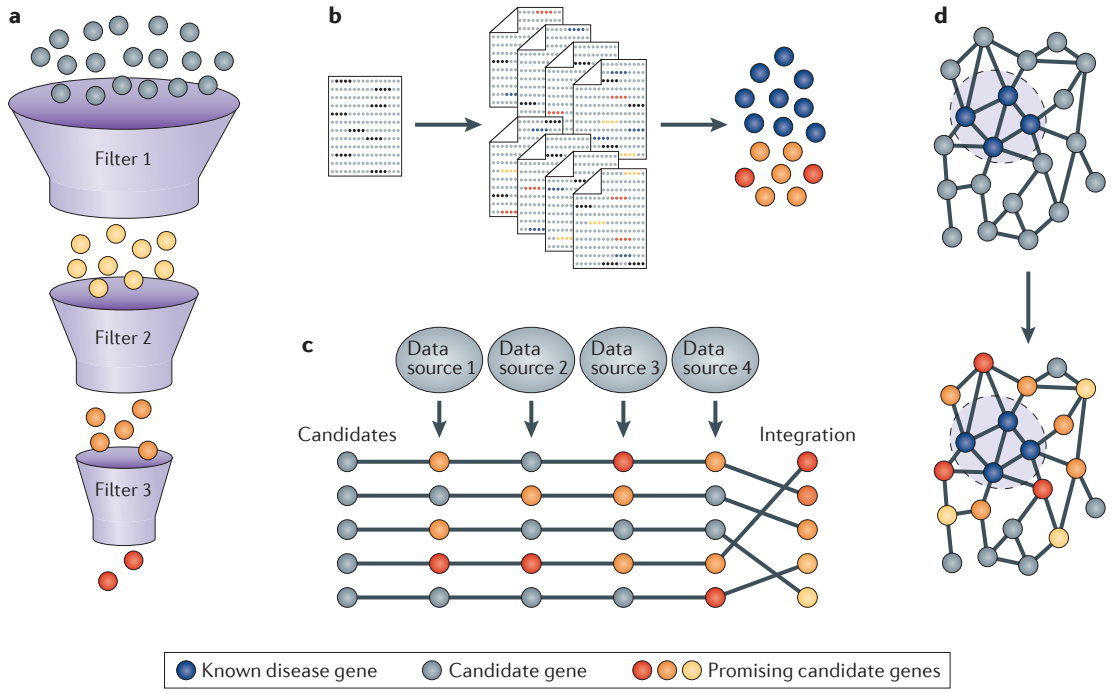


Figure 1 | **Computational strategies for prioritization.** Prioritization methods can roughly be classified into a filtering strategy (**a**) and three ranking strategies (**b**–**d**). **a** | Filtering strategy. First, the properties of the ideal candidate gene are defined, and filters are created accordingly. These filters are then used to select the most promising genes from the pool of candidate genes. **b** | Text-mining strategy. In the first step, a set of disease-relevant keywords is used to retrieve a corpus of disease-relevant documents. This corpus is then mined to identify both already known genes and promising candidate genes. **c** | Similarity profiling and data fusion strategy. Several complementary data sources are considered to define the most promising candidates. The similarities between the candidate genes and the known seed genes are computed for each data source and are then integrated over all data sources to obtain the final prioritization. **d** | Network-based strategy. The known disease genes are identified in a gene network. Candidate genes are then selected on the basis of their distance from the known genes.

under study. Second, the genes mentioned in these documents are extracted through information-retrieval methods. Third, a statistical assessment of the strength of the extracted information is used to score each gene. The result is then a combination of already known disease genes and promising candidate genes for which some evidence from the literature already points to a link to the biological process or to the disease of interest. Systems such as GeneProspector[64] and AGeneApart[65] mine MEDLINE to discover known and potentially new disease–gene relations. For example, AGeneApart has been integrated into the DECIPHER database of chromosomal aberrations to support the interpretation of disease loci in terms of genes that are known to be linked to a phenotype on the basis of MEDLINE abstracts[66].

Although mining the literature is a powerful way of identifying promising candidates, it tends to identify straightforward candidates for which abundant knowledge is already available[67]. By contrast, similarity-profiling methods integrate both knowledge bases (for reliable predictions) and raw data (for novel predictions)[1,5] (FIG. 1c). Most of these methods identify the most promising candidate genes according to their similarity to the already known seed genes for that disease or biological process. For example, they can assess which Gene Ontology categories tend to be overrepresented among the known genes and can favour candidates that belong to these Gene Ontology categories. Likewise, they can assess the BLAST scores of candidates against the seed genes and can favour candidates that are homologous to some of the seed genes. Next, the procedure of data fusion aggregates the similarity profile scores from multiple data sources into a global ranking. Tools such as Endeavour[1,68] and GeneDistiller[10] carry out such strategies and integrate more than six types of genomic data from over a dozen data sources. Additionally, data from model organisms have become a particularly rich source of information for human gene prioritization[30], although it presents specific challenges of transferring data across species (BOX 2). For example, GeneSeeker[5] incorporates mouse expression data to help prioritize human genes, whereas ToppGene[69] incorporates information about phenotypes from mouse mutants. Alternatively, Genie provides large-scale cross-species text mining[70]. Recently, GPsy[71] proposed a prioritization scheme that extends Endeavour to integrate data across species and with a flexible weighting scheme, although it is specifically tailored to a precompiled lists of developmental processes.

Recently, prioritization methods based on network analysis have also become popular[25,43,72] (FIG. 1d). Network analysis uses strategies that are similar to data fusion methods by determining the similarity between candidates and known genes, except the data are represented as networks. Known seed genes are identified, and candidate genes are scored according to their network distance to the known disease genes. Such approaches are reviewed in REFS 73,74. For instance, GeneWanderer uses random walks or a diffusion kernel on a PPI network[75], and ToppNet (from the ToppGene suite) uses Web and social network methods on a PPI network[69,76]. The network can either be a true PPI network (such as BioGrid[77]) or an integrative (functional linkage) network[78] (such as STRING[79]). Network-based prioritization differs from network inference in that the goal of the data integration is to identify nodes of the network that are relevant to the disease or biological process of interest rather than to infer the edges (that is, the connections) of the network. It also differs from similarity profiling in that it relies on a pre-established network across which information is propagated. This gives it the advantage of easier interpretability (relationships can be expressed in terms of links in the network) but the disadvantage of being limited to those genes that belong to the network (for example, BioGrid v3.1.88 covers only 14,528 unique human proteins).

In BOX 4, we provide a tutorial for using Endeavour and GeneWanderer to 'rediscover' a known disease–gene association. In this example, candidate genes from a single chromosomal region are prioritized using seed genes as prioritization criteria.

Finally, a delicate problem arises when little or no prior knowledge is available, which is an interesting situation because the potential for discoveries is the greatest. In this case, seed genes will be difficult to collect. A first possibility is to rely on methods that do not use any prior knowledge about disease phenotype and that perform a priori prioritization using sequence features[80,81] or topological network features only[69]. Another approach is to collect sets of seed genes for closely related biological processes or phenotypes and to use those for prioritization. Collecting keywords is usually easier, but in this situation text-mining strategies will fail owing to the lack of published information. Network-based methods offer several interesting possibilities. For example, relevant protein complexes in a PPI network have been identified on the basis of similarities between phenotypic descriptions of known disease genes and a target phenotype[29], and pairs or triplets of interacting proteins have been found across multiple disease loci[2]. Furthermore, ranking of candidates can also be carried out if signals other than seed genes are available. For example, PINTA[44] uses differential expression data to prioritize candidates. Promising candidates are the genes for which strong differential expression signals — for example, between affected versus healthy individuals — are observed in the neighbourhood of the candidate. Other signals, such as GWAS association scores, could also be propagated in this way across a network to prioritize candidates[20].

Carrying out complex strategies. Despite most prioritization tools relying on similar concepts, using different data sources, different prioritization strategies and different representations of prior knowledge means that currently no method universally dominates[36,82]. Some methods are better suited for the analysis of multiple loci from GWASs (for instance, G2D[83] and Prioritizer[2]), whereas others are more suitable when no disease genes are known (for instance, Candid[84] and PolySearch[85]). It can therefore be useful to perform the analysis using multiple tools concurrently to maximize the chances of identifying the relevant genes (FIG. 2). In that case, each tool generates its own prioritization, representing one line of evidence that is then combined with the other prioritizations. For example, candidate genes for a complex disease are typically harder to prioritize than for a monogenic disorder, but using multiple methods in conjunction can improve the quality of the predictions, as shown by several studies on type 2 diabetes and obesity[86–88].

As a tutorial for using and comparing multiple gene prioritization tools, Supplementary information S1 (table) contains the candidate gene lists and prioritization criteria for 42 disease–gene associations, which can be used to compare the working of prioritization tools. Researchers can simply cut and paste the input data into any of the available gene prioritization tools, such as those linked through the Gene Prioritization Portal (BOX 3), to compare the abilities of these tools to 'rediscover' recently discovered disease–gene associations.

Using a single set of seed genes can be enough to study simple monogenic conditions. However, more advanced strategies are often required to model disorders that encompass effects across multiple biological processes, multiple phenotypes or multiple and distinct disease subtypes. For example, if distinct phenotypes are linked to the disease (such as a heart anomaly and intellectual disability), a single set of seed genes will probably be too heterogeneous at the molecular level, and therefore predictions will be less accurate. In such a case, it is preferable to model each phenotypic aspect separately and then to merge the resulting predictions[37,89]. An example is the analysis of a locus on chromosome 6 that was associated with congenital heart defects using seven models corresponding to seven phenotypes or biological processes that are linked to heart development (FIG. 3).

Prioritization tools are increasingly applied to study monogenic diseases with locus heterogeneity and oligogenic or complex diseases by prioritizing many candidates across several loci for downstream characterization (instead of focusing on one locus at a time). For example, five prioritization tools were used to analyse 47 non-overlapping rare copy number variants (CNVs) from 255 patients with intellectual disability, resulting in 28 novel promising candidate genes[90]. Because of the rapid decrease in sequencing costs, such strategies are becoming particularly attractive. Indeed, instead of focusing on resolving the disease gene for one disease locus at a time, it is becoming feasible to sequence multiple candidate genes from multiple disease loci simultaneously in a panel of patients. This strategy increases the likelihood of confirming disease genes and makes it

Random walk
A mathematical formalization of the path resulting from taking successive random steps. Classical examples of random walks are Brownian motion, the fortune of a gambler flipping a coin or fluctuations of the stock market. In the context of graphs, a random walk typically describes a process in which a 'walker' moves from one node of the graph into another with a probability proportional to the weight of the edge connecting them.

Diffusion kernel
A type of kernel similarity matrix that is derived from the notion of a random walk on a graph. Diffusion kernels measure similarity between nodes of a graph (in this case, between genes) — for example, by estimating the average length of a random walk from one node to the other.

Locus heterogeneity
The appearance of phenotypically similar characteristics that result from mutations at different genetic loci. Differences in effect size or in replication between studies and samples are often ascribed to different loci leading to the same disease.

Box 4 | **A single-locus, monogenic gene prioritization tutorial**

This step-by-step tutorial is based on a study by Ebermann and colleagues[128], who reported a novel Usher syndrome gene, deafness, autosomal recessive 31 (*DFNB31*). Usher syndrome combines hearing loss and retinitis pigmentosa (which is a disorder of the retina leading to blindness). We mimic the situation in which this disease–gene association is still unknown and describe how using Endeavour and GeneWanderer we can rediscover this association. This example is purely illustrative because *DFBN31* is now an established Usher syndrome gene. Note that information pertaining to the role of *DFBN31* in Usher syndrome will be contained in some of our data sources; this concept of 'knowledge contamination' is discussed in the main text and makes our prioritization task easier than in the case of a novel discovery.

Identifying candidate genes

In this example, we consider all genes located on chromosome 9q (where *DFBN31* is located) as candidate genes. With Endeavour and GeneWanderer, candidates can be defined using chromosome arms, coordinates or cytogenetic bands, so there is no need to retrieve the complete list of genes.

Gathering seed genes

A useful starting point is to browse Online Mendelian Inheritance in Man (OMIM) to identify the genes that are already associated with Usher syndrome. The query 'Usher syndrome' matches 10 OMIM pages that describe what is known about the different types of Usher syndrome (those pages are #276900, #605472, #276904, #601067, #276901, #276902, #602083, %612632, #606943 and %602097). Each page starts with a table that contains phenotype–gene relationships. In total, the 10 tables corresponding to the 10 OMIM pages contain 9 genes (see the table). To mimic searching for unknown disease–gene associations, we have excluded *DFNB31* (page #611383) from the seed gene list.

The seed gene list can be expanded through a literature search to identify genes with putative links to the disease that might not yet be included in OMIM. In PubMed, an advanced query can be built by selecting all publications that contain 'Usher syndrome' in their title and that are also review articles; here, the search input would be: "Usher syndrome" [title] review [publication type]. In this case, no extra seed genes are identified in the abstracts of the retrieved articles.

Prioritizing the candidates with Endeavour

Running Endeavour is a four-step process. First, the species has to be selected. In this example, 'human' is the appropriate selection because the candidates are human genes. Second, the seed genes are provided (see the table in this box) one gene at a time. For *Homo sapiens* genes, Endeavour recognizes official HUGO gene names, so care should be taken to avoid unofficial gene name synonyms. Third, the suitable data sources — that differ in the types of relationship data they contain — must be selected from the displayed list. For simplicity, all of them can be selected for this example. Fourth, the candidate genes are entered using the term 'chr:9q'; the program then automatically loads the 593 genes from that region. The prioritization can then be launched. When the prioritization is complete, the results are presented in a coloured ranked table with the most promising genes at the top. The output table includes separate columns of rankings according to each of the chosen data sources that were interrogated, in addition to a combined ranking that encompasses results from all of the chosen data sources.

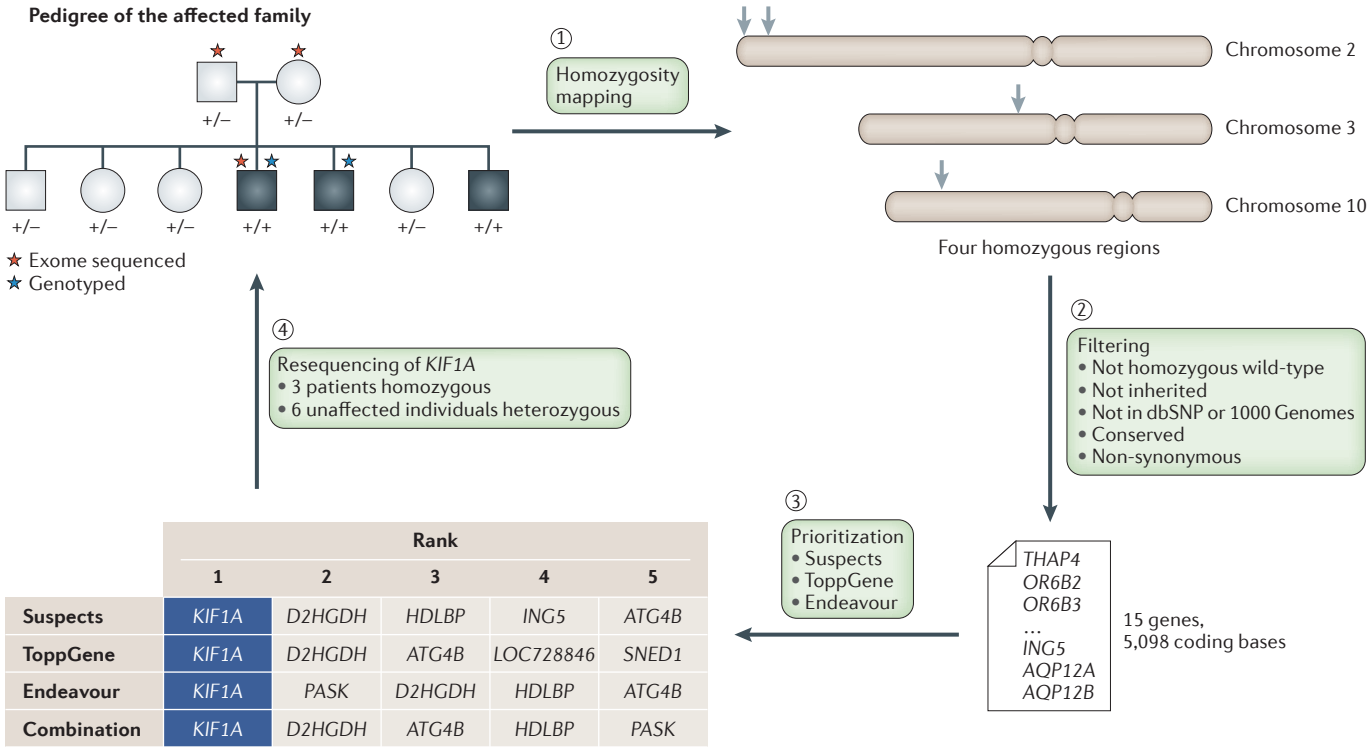
Prioritizing the candidates with GeneWanderer

There are four inputs that are required to run GeneWanderer. First, the candidate genes are defined through chromosomal coordinates. In this case, the coordinates of 9q can be used (9, 51274031 and 140273252). Then the ranking algorithm needs to be selected; the default option 'Random Walk' can be used as it usually returns the best results[75]. Third, the seed genes need to be provided (see the table). Alternatively, users can select the disease name from a predefined list. However, in our case, 'Usher syndrome' is not the list, so we input the genes manually. The last option is the network to be used. Once again, the default option can be used for this example. Similarly to Endeavour, the output table contains the most promising genes on top, together with their final scores.

Gene name	Gene ID	Location
MYO7A	4647	11q13.5
GPR98 (also known as *VLGR1*)	84059	5q14.3
PDZD7	79955	10q24.31
USH1C	10083	11p15.1
PCDH15	65217	10q21.1
CDH23	64072	10q22.1
USH2A	7399	1q41
CLRN1	7401	3q25.1
USH1G (also known as *SANS*)	124590	17q25.1

possible to identify entire molecular networks in which mutations lead to the disease.

The tools can also be tightly integrated with medium-throughput screens so that researchers can rapidly cycle between experiments and computational analysis. An example is the integration of gene prioritization in a screen for genetic interactors of the *Atonal* proneural gene in *Drosophila melanogaster*[24]. Initially, screening of deficiency lines identified 12 loci (containing a total of 1,100 candidate genes) that were positively associated with the phenotype of interest. Prioritization using a fly-specific version of Endeavour then selected the top 30% of candidate genes for genetic screening, from which all 12 causal genes were identified through functional analysis *in vivo*. In fact, 11 of these 12 genes were found in the top 6% of the prioritized candidate gene list. Subsequently, analysis of the STRING network for the newly identified genes and the seed genes identified a dense subnetwork containing most of those genes and an additional 66 promising candidates across the whole genome. Those candidates could then be used directly to plan a second medium-throughput screen. Such strategies can substantially speed up experimental work and reduce associated costs.



	Rank				
	1	2	3	4	5
Suspects	*KIF1A*	*D2HGDH*	*HDLBP*	*ING5*	*ATG4B*
ToppGene	*KIF1A*	*D2HGDH*	*ATG4B*	*LOC728846*	*SNED1*
Endeavour	*KIF1A*	*PASK*	*D2HGDH*	*HDLBP*	*ATG4B*
Combination	*KIF1A*	*D2HGDH*	*ATG4B*	*HDLBP*	*PASK*

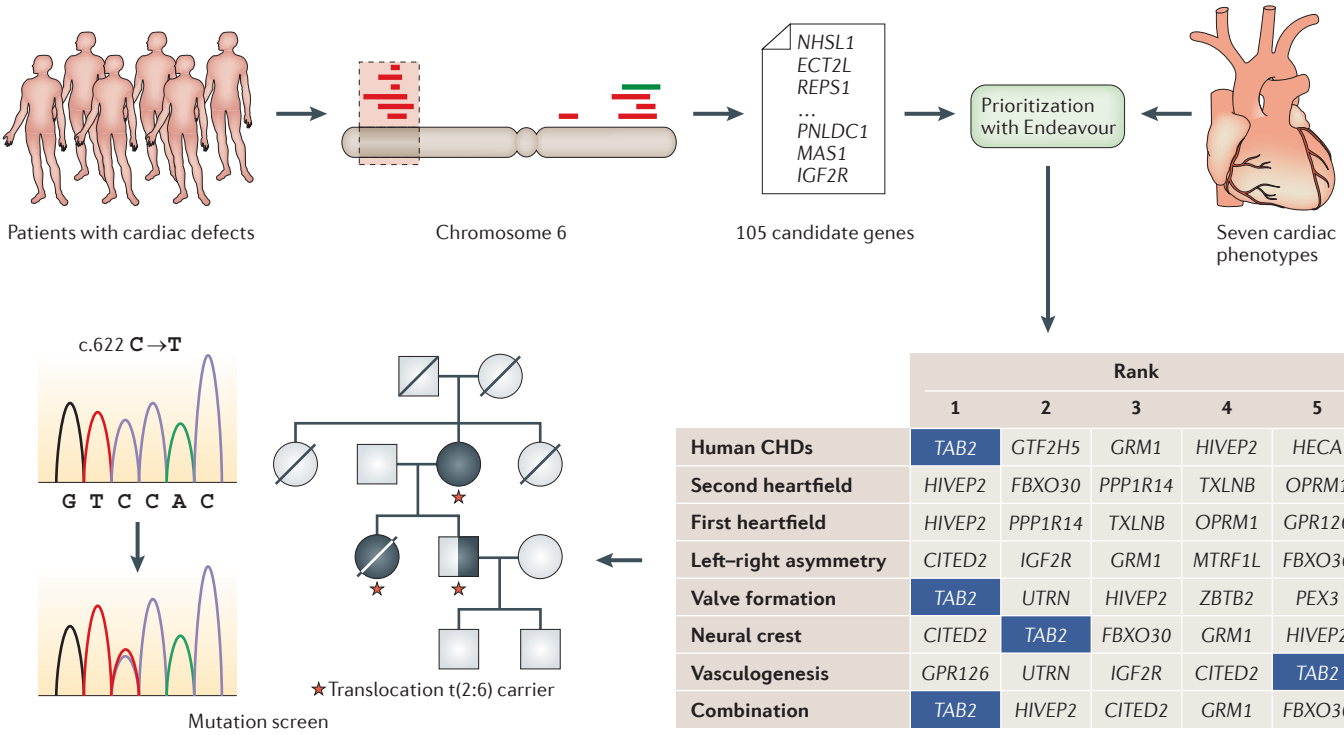
Figure 2 | **Exome sequencing and disease network analysis of a single family implicate a mutation in *KIF1A* in hereditary spastic paraparesis.** A familial case of hereditary spastic paraparesis (HSP) was analysed through whole-exome sequencing and homozygosity mapping[99]. The four largest homozygous regions between two of the three affected brothers were considered to be potential disease loci, containing a total of 44 genes. Because the exome-sequencing data provided detailed information on the genetic variants in these genes, the genes were considered to be potentially causative if they contained at least one variant meeting each of the following characteristics: non-wild-type and homozygous; under purifying selection; not inherited from the parents; not present in dbSNP or the 1000 Genomes Project data; and non-synonymous. After this filtering step, 15 candidate genes remained. The list was then prioritized using three computational methods (namely, Suspects, ToppGene and Endeavour) to assess the robustness of the prioritization results and because those tools use different data sources. The prioritization criteria were a list of 11 seed genes that were obtained through a review of the literature and are known to be associated with forms of HSP in which mutations lead to the core HSP phenotypic traits (that is, progressive lower-extremity spastic weakness, hypertonic urinary bladder disturbance and mild diminution of lower-extremity vibration sensation) but not to unrelated traits. The top-ranking gene from the prioritization was kinesin family member 1A (*KIF1A*). Sanger sequencing confirmed that *KIF1A* is the causative variant: the third affected brother was also homozygous at the *KIF1A* locus (whereas the parents and four unaffected siblings were heterozygous), and a homozygous Ala255Val variant was identified in the protein motor region of the encoded KIF1A protein.

Assessment of the prioritization

Experimental benchmarking. Even though some prioritization methods return a *P* value estimate with each output gene, these values can be unreliable owing to the complexity of the underlying statistical models and some multiple testing issues. Evaluating the actual performance of gene prioritization methods is challenging. In an ideal setting, a large set of prioritizations would be carried out using a given tool and then those hypotheses would be tested experimentally to determine the proportion of false positives and ideally of false negatives as well. So far, only a few such studies have been carried out (an example is the *D. melanogaster* screen mentioned in the previous section)[18,24,91–94]. Although such studies clearly show the value of gene prioritization, they are aimed at a single biological question and thus provide little guidance about how the method will perform on a different problem.

Statistical benchmarking by cross-validation. In contrast to experimental benchmarking, statistical benchmarks collect extensive sets of known disease–gene associations and evaluate how well a method recovers those known associations. An easy and common statistical benchmarking method is called cross-validation[95]. In a cross-validation setup, a proportion of the data is used to build a model, whereas the remaining part of the data is set aside to evaluate the model. This split is repeated multiple times. Cross-validating gene prioritization tools involves removing a known disease-related gene from the seed gene list and instead including it in the longer list of random candidate genes for prioritization. This procedure is repeated for each seed gene, and the average rank of the seed gene among the random genes is computed across all of the runs. If the prioritizations rank these genes within the top

Multiple testing
A statistical problem that arises from carrying out multiple hypothesis tests together. *P* values obtained from hypothesis tests under the assumption of a single test must be appropriately corrected to reflect multiple testing.



	Rank				
	1	2	3	4	5
Human CHDs	*TAB2*	*GTF2H5*	*GRM1*	*HIVEP2*	*HECA*
Second heartfield	*HIVEP2*	*FBXO30*	*PPP1R14*	*TXLNB*	*OPRM1*
First heartfield	*HIVEP2*	*PPP1R14*	*TXLNB*	*OPRM1*	*GPR126*
Left–right asymmetry	*CITED2*	*IGF2R*	*GRM1*	*MTRF1L*	*FBXO30*
Valve formation	*TAB2*	*UTRN*	*HIVEP2*	*ZBTB2*	*PEX3*
Neural crest	*CITED2*	*TAB2*	*FBXO30*	*GRM1*	*HIVEP2*
Vasculogenesis	*GPR126*	*UTRN*	*IGF2R*	*CITED2*	*TAB2*
Combination	*TAB2*	*HIVEP2*	*CITED2*	*GRM1*	*FBXO30*

Figure 3 | **Haploinsufficiency of *TAB2* causes congenital heart defects in humans.** A locus for congenital heart defects (CHDs) is identified on 6q24–q25 through a genotype–phenotype correlation in 12 patients[89]. The locus was prioritized using Endeavour with seven sets of seed genes corresponding to seven relevant aspects of the cardiac phenotypes (as defined by experts and with the use of CHDWiki[129]). The main motivation behind using seven disease models is the improvement of the benchmarking performance when compared with using a single large gene set. When combined, the seven rankings reveal that TGFβ-activated kinase 1/MAP3K7-binding protein 2 (*TAB2*) is the most promising candidate gene among the 105 candidate genes from this locus. Its role in cardiac development is supported by its conserved expression in the developing human and zebrafish heart. Moreover, a family is identified in which a balanced translocation that disrupts *TAB2* segregates with CHDs. Finally, mutation analysis in 402 patients with CHDs reveals two evolutionarily conserved missense mutations. Taken together, the experimental primary data and the results of the prioritization firmly establish the role of *TAB2* as a disease gene for CHDs.

5–15% of random genes, the prioritization has been able to capture useful information.

Various cross-validation tests have been performed for many hundreds of disease–gene associations for over 100 disease families, as prioritized by various tools[75,96,97]. Each benchmarking study showed that disease genes rank on average within the top 10% of the prioritized list, although this value varies according to the settings. However, the primary disadvantage of cross-validation is that it measures the ability of an algorithm to capture what is already known by falsely pretending that it is not known. After publication, information on disease–gene associations becomes rapidly integrated into resources such as MEDLINE, Gene Ontology and KEGG. Because such data sources are at the core of the prioritization tools and already contain this disease–gene association information (so-called 'knowledge contamination'), the retrieval of the test genes is facilitated and hence cross-validation provides optimistic estimates of the predictive power of gene prioritization tools[98]. However, cross-validation remains an assessment of choice because good cross-validation performance is a requirement for good prioritization, albeit it is not a guarantee.

Other quality-control methods. An alternative assessment for prioritization tool performance is to rerun the prioritization using a set of negative control seed genes (for example, genes for other unrelated diseases)[89,99]. If top-ranking candidates that are identified using the relevant seed genes also rank highly when using the negative control seed genes, this indicates that some systematic bias is present and that the results are unreliable.

If the set of candidates is a small subset of the genome, another simple technique is to perform prioritizations both on the actual set of candidates and on the whole genome. When comparing prioritization outputs, if the top-ranking candidates from the small subset do not rank within the top 5–15% of the whole genome, this variability suggests that the prioritization might simply not have been able to capture enough information to identify any good candidates.

Finally, another option when prioritizing large sets of candidates is to check for functional enrichment (for example, in Gene Ontology categories) among the top candidates in the prioritized list[100]. The enriched terms should match expectations for the biological process or phenotype of interest. Because prioritization methods

involve capturing Gene Ontology information or related information, a Gene Ontology enrichment matching expectation is a necessary — but not sufficient — indication that the prioritization was successful.

Contextualization and visualization

Because of the complexity of the retrieval, analysis and aggregation of heterogeneous data sources, it is difficult to dissect the contribution of each nugget of the underlying relationship data to the final ranking of a candidate. Prioritization tools rarely provide the data that underlie the ranking of a candidate, making these tools somewhat 'black box' in nature. This hinders the interpretation of the prioritization results and the design of downstream functional analysis of promising candidates. Until improved 'explanation support' becomes available in prioritization tools, third-party tools can alleviate this difficulty by providing the functional context of the top candidate genes, most often in a graphical manner. For instance, by querying the STRING protein network[79,101] with the seed genes and the top candidate genes, it is possible to visualize a global functional network and therefore to understand why these candidate genes are considered to be promising. In addition, an enrichment analysis of the top candidates can be performed using DAVID[102] or GSEA[103] to detect overrepresented pathways and to check whether they make sense with respect to the biological process of interest.

Conclusions and future directions

Computational methods for gene prioritization have progressed quickly. They now demonstrably contribute to biological discovery. Their ability to gather and to integrate data from multiple sources provides a more thorough and less biased global assessment of candidate genes than can be manually achieved. Such methods are not confined to guiding the discovery of disease genes in monogenic Mendelian disorders but are useful whenever genes or proteins are to be selected on the basis of heterogeneous functional data (for example, selecting genes for a genetic interaction screen). The fact that such analyses can be carried out quickly using simple tools without the need for the direct support of a bioinformatics expert makes them particularly attractive. However, many tools are available, and different biological questions may require using different prioritization tools, depending on which data sources are required by the user. Rather than being an 'oracle' that provides predictions — which a researcher would then simply be left to validate experimentally — gene prioritization is increasingly used as a line of evidence that is complementary to primary experimental data when showing the association of a gene to a disease or a biological process[99,104].

Although prioritization methods have greatly improved in the past few years, some methodological improvements are still necessary. First, our understanding of how to perform useful predictions using multiple data sources or across biological networks is still rudimentary. For example, the principle of guilt by association has been called into question as presenting important statistical artefacts (such as node degree effects or exceptional edges that bias the performance assessment)[105,106]. Methodological work is needed to improve data and network quality towards integrative predictions and to remove biases in predictive methods. Second, the field needs to consolidate through improved benchmarking efforts. Benchmarks do not provide a gold standard in evaluating the performance of prioritization methods, thus their quality could be considerably improved. There is a need for a large-scale community effort — similar in spirit to the CASP[107,108], BioCreative[109,110], CAMDA[111,112] or DREAM[113] competitions — in which multiple tools can be compared across common prospective benchmarks that have been designed by the community. These efforts can serve as a guide for methodological developments in the field by allowing a reasonably objective comparison of tools. Also, prioritization methods integrate data from numerous sources with all the resulting challenges of data standardization and updates. As such, they will greatly benefit from all efforts related to the semantic Web[114] (standardized use of ontologies across databases and of automated queries over the Web).

There is also a need for improved reporting of the underlying relationships so that all tools can move beyond the black-box stage to have greater explanatory power. Currently, only prioritization methods based on text mining provide easy access to the evidence for the prioritization through links to the relevant literature[85,115]; however, the ToppNet tool does provide a network view of candidate genes and seed genes, which is a first step in this direction. Additionally, methods need to supplement their output rankings with meaningful and reliable *P* values to improve confidence in the results.

Future research directions for prioritization mostly focus on broadening its scope beyond the ranking of individual genes. A key opportunity is the prioritization of genomic variants from next-generation sequencing data. Full-genome sequencing of any individual will identify on the order of 4,000,000 variants, ~10,000 of which are in coding regions. Sequencing projects for cancer and other diseases deliver huge lists of genomic variants[116,117] (such as single-nucleotide variants, insertions and deletions, and rearrangements), but it is extremely challenging to assess which variants are causative for or associated with the phenotype. Although there has been considerable progress in filtering variants (see the recent Review in this journal[118]), current methods mainly focus on how variants affect sequence properties (in particular, evolutionary conservation) and protein structure, rather than being based on phenotypic information. However, existing gene prioritization tools cannot handle information at the level of individual variants and are thus not directly suitable for this purpose either. Nevertheless, many relevant types of biological information on genetic variants are available, such as disease-association scores, whether a variant falls in a locus that has been associated to the phenotype in linkage or copy number studies or whether a variant affects a gene that is potentially implicated with the phenotype. Therefore, such integration tasks would be well suited for novel prioritization strategies.

Another important extension of gene prioritization methods is termed 'edge prioritization'[119]. Instead of only prioritizing genes in isolation, methods can be developed to generate hypotheses about potential interactions among top candidates and seed genes (for example, transcription factor–target interactions, ligand–receptor interactions and post-transcriptional modifications). This would greatly increase the value of prioritization methods because it would generate hypotheses that directly suggest relevant experimental validation. For example, if a gene among the top-ranking candidates is a transcription factor and if the regulatory region of a corresponding seed disease gene carries potential transcription factor binding sites, the seed gene might be a direct target of this candidate.

Finally, although we focused this Review on prioritizing genes and their protein products, the computational methods could potentially be applied to other biomolecules as long as multiple relevant data sources are available. Thus, prioritization schemes can potentially be developed for alternative transcripts and protein isoforms, peptides, non-coding RNAs, metabolites and drugs[120]. This last option could be highly useful for the pharmaceutical industry and could help it to handle the large amounts of genomic data that it has at its disposal.

1. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nature Biotech.* **24**, 537–544 (2006). **This is the original description of the prioritization tool Endeavour, which uses a similarity profiling strategy.**
2. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006). **This is the original description of the prioritization tool Prioritizer, which relies on a human functional network.**
3. Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Association of genes to genetically inherited diseases using data mining. *Nature Genet.* **31**, 316–319 (2002).
4. Thiel, C. T. *et al.* Severely incapacitating mutations in patients with extreme short stature identify RNA-processing endoribonuclease RMRP as an essential cell growth regulator. *Am. J. Hum. Genet.* **77**, 795–806 (2005).
5. van Driel, M. A., Cuelenaere, K., Kemmeren, P. P.C. W., Leunissen, J. A. M. & Brunner, H. G. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. J. Hum. Genet.* **11**, 57–63 (2003).
6. Sparrow, D. B., Guillén-Navarro, E., Fatkin, D. & Dunwoodie, S. L. Mutation of hairy-and-enhancer-of-split-7 in humans causes spondylocostal dysostosis. *Hum. Mol. Genet.* **17**, 3761–3766 (2008).
7. Rajab, A. *et al.* Fatal cardiac arrhythmia and long-QT syndrome in a new form of congenital generalized lipodystrophy with muscle rippling (CGL4) due to *PTRF-CAVIN* mutations. *PLoS Genet.* **6**, e1000874 (2010).
8. Kaufmann, R. *et al.* Infantile cerebral and cerebellar atrophy is associated with a mutation in the *MED17* subunit of the transcription preinitiation mediator complex. *Am. J. Hum. Genet.* **87**, 667–670 (2010). **This study shows that *MED17* mutations are associated with infantile cerebral and cerebellar atrophy using GeneDistiller.**
9. Spinazzola, A. *et al. MPV17* encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nature Genet.* **38**, 570–575 (2006).
10. Seelow, D., Schwarz, J. M. & Schuelke, M. GeneDistiller—distilling candidate genes from linkage intervals. *PLoS ONE* **3**, e3874 (2008).
11. George, R. A. *et al.* Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.* **34**, e130 (2006).
12. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
13. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
14. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012).
15. Dreszer, T. R. *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* **40**, D918–D923 (2012).
16. Parkinson, H. *et al.* ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* **39**, D1002–D1004 (2011).
17. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
18. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
19. van Vliet-Ostaptchouk, J. V. *et al. HHEX* gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort. *Eur. J. Hum. Genet.* **16**, 652–656 (2008). **This is a biological validation of Prioritizer, showing that variants near the *HHEX* gene contribute to the risk of T2D in a Dutch population.**
20. Pers, T. H. *et al.* Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. *Genet. Epidemiol.* **35**, 318–332 (2011).
21. Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
22. Perez-Iratxeta, C., Bork, P. & Andrade-Navarro, M. A. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res.* **35**, W212–W216 (2007).
23. Tremblay, K. *et al.* Genes to diseases (G2D) computational method to identify asthma candidate genes. *PLoS ONE* **3**, e2907 (2008).
24. Aerts, S. *et al.* Integrating computational biology and forward genetics in *Drosophila*. *PLoS Genet.* **5**, e1000351 (2009).
25. Goh, K.-I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
26. Smith, N. G. C. & Eyre-Walker, A. Human disease genes: patterns and predictions. *Gene* **318**, 169–175 (2003).
27. Oti, M. & Brunner, H. G. The modular nature of genetic diseases. *Clin. Genet.* **71**, 1–11 (2007). **This paper provides a motivation to use the guilt by association principle to identify novel disease causing genes.**
28. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
29. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotech.* **25**, 309–316 (2007).
30. Tiffin, N., Andrade-Navarro, M. A. & Perez-Iratxeta, C. Linking genes to diseases: it's all in the data. *Genome Med.* **1**, 77 (2009). **In this paper, a discussion is presented of how disease gene discovery will be facilitated by improved data integration and the use of clinical data.**
31. Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635 (2004).
32. De Bie, T., Tranchevent, L.-C., van Oeffelen, L. M. M. & Moreau, Y. Kernel-based data fusion for gene prioritization. *Bioinformatics* **23**, i125–i132 (2007).
33. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A. Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA* **100**, 8348–8353 (2003).
34. Kondor, R. I. & Lafferty, J. Diffusion kernels on graphs and other discrete input spaces. *Proc. 19th Int. Conf. Machine Learning* **2002**, 315–322 (2002).
35. Tranchevent, L.-C. *et al.* A guide to web tools to prioritize candidate genes. *Brief. Bioinformat.* **12**, 22–32 (2011). **This paper discusses a Web portal describing multiple prioritization tools and supporting the selection of appropriate tools for given requirements.**
36. Oti, M., Ballouz, S. & Wouters, M. A. Web tools for the prioritization of candidate disease genes. *Methods Mol. Biol.* **760**, 189–206 (2011). **This paper provides a detailed description of several Web-based prioritization methods together with their specificities.**
37. Tiffin, N. Conceptual thinking for *in silico* prioritization of candidate disease genes. *Methods Mol. Biol.* **760**, 175–187 (2011). **This is a review on gene prioritization that also describes the development of your own data integration method.**
38. Piro, R. M. & Di Cunto, F. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.* **279**, 678–696 (2012). **This review focuses on the different data sources and the algorithms underlying the prioritization methods.**
39. Kann, M. G. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinformat.* **11**, 96–110 (2010).
40. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
41. Ma, X., Lee, H., Wang, L. & Sun, F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* **23**, 215–221 (2007).
42. Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* **28**, 21–28 (2001).
43. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Rev. Genet.* **12**, 56–68 (2011). **This is a review of network-based methods to unravel the molecular mechanisms underlying diseases.**
44. Nitsch, D. *et al.* PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res.* **39**, W334–W338 (2011).
45. Keyser, R. J., Oppon, E., Carr, J. A. & Bardien, S. Identification of Parkinson's disease candidate genes using CAESAR and screening of MAPT and SNCAIP in South African Parkinson's disease patients. *J. Neural Transm.* **118**, 889–897 (2011).
46. Oti, M., Huynen, M. A. & Brunner, H. G. The biological coherence of human phenome databases. *Am. J. Hum. Genet.* **85**, 801–808 (2009).
47. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).
48. Antonarakis, S. E. & McKusick, V. A. OMIM passes the 1,000-disease-gene mark. *Nature Genet.* **25**, 11 (2000).
49. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nature Genet.* **36**, 431–432 (2004).
50. Doms, A. & Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.* **33**, W783–W786 (2005).

51. Racine, J. *et al.* Comparison of genomic and proteomic data in recurrent airway obstruction affected horses using ingenuity pathway analysis®. *BMC Vet. Res.* **7**, 48 (2011).
52. Thomas, S. & Bonchev, D. A survey of current software for network analysis in molecular biology. *Hum. Genom.* **4**, 353–360 (2010).
53. Wickramasinghe, S., Rincon, G., Islas-Trejo, A. & Medrano, J. F. Transcriptional profiling of bovine milk using RNA sequencing. *BMC Genom.* **13**, 45 (2012).
54. Ekins, S., Nikolsky, Y., Bugrim, A., Kirillov, E. & Nikolskaya, T. Pathway mapping tools for analysis of high content data. *Methods Mol. Biol.* **356**, 319–350 (2007).
55. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
56. Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
57. Franke, L. *et al.* TEAM: a tool for the integration of expression, and linkage and association maps. *Eur. J. Hum. Genet.* **12**, 633–638 (2004).
58. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput.* **14**, 368–379 (2009).
59. Krallinger, M., Valencia, A. & Hirschman, L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* **9** (Suppl. 2), S8 (2008).
60. Winnenburg, R., Wächter, T., Plake, C., Doms, A. & Schroeder, M. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief. Bioinformat.* **9**, 466–478 (2008).
61. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218–223 (2009).
62. Baudot, A., Gómez-López, G. & Valencia, A. Translational disease interpretation with molecular networks. *Genome Biol.* **10**, 221 (2009).
63. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
64. Yu, W., Wulf, A., Liu, T., Khoury, M. J. & Gwinn, M. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformat.* **9**, 528 (2008).
65. Van Vooren, S. *et al.* Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res.* **35**, 2533–2543 (2007).
66. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
67. Kowald, A. & Schmeier, S. *Data Mining in Proteomics. Inform. Retrieval* **696**, 305–318 (Humana Press, 2011).
68. Tranchevent, L.-C. *et al.* ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.* **36**, W377–W384 (2008).
69. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
70. Fontaine, J.-F., Priller, F., Barbosa-Silva, A. & Andrade-Navarro, M. A. Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.* **39**, W455–W461 (2011).
71. Britto, R. *et al.* GPSy: a cross-species gene prioritization system for conserved biological processes—application in male gamete development. *Nucleic Acids Res.* 8 May 2012 (doi:10.1093/nar/gks380).
72. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
73. Kann, M. G. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief. Bioinformat.* **8**, 333–346 (2007).
74. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**, 1057–1063 (2010).
This is a recent review about predicting disease–gene associations using gene–protein networks and network-based algorithms.
75. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
76. Chen, J., Xu, H., Aronow, B. J. & Jegga, A. G. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformat.* **8**, 392 (2007).
77. Breitkreutz, B.-J., Stark, C. & Tyers, M. The GRID: the General Repository for Interaction Datasets. *Genome Biol.* **4**, R23 (2003).
78. Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y. & Delisi, C. Genome-wide prioritization of disease genes and identification of disease–disease associations from an integrated human functional linkage network. *Genome Biol.* **10**, R91 (2009).
79. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444 (2000).
80. Lõpez-Bigas, N. & Ouzounis, C. A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* **32**, 3108–3114 (2004).
81. Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformat.* **6**, 55 (2005).
82. Thornblad, T. A., Elliott, K. S., Jowett, J. & Visscher, P. M. Prioritization of positional candidate genes using multiple web-based software tools. *Twin Res. Hum. Genet.* **10**, 861–870 (2007).
83. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. G2D: a tool for mining genes associated with disease. *BMC Genet.* **6**, 45 (2005).
84. Hutz, J. E., Kraja, A. T., McLeod, H. L. & Province, M. A. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet. Epidemiol.* **32**, 779–790 (2008).
85. Cheng, D. *et al.* PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **36**, W399–W405 (2008).
86. Tiffin, N. *et al.* Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.* **34**, 3067–3081 (2006).
This is an example of the application of prioritization to a complex disorder using multiple prediction algorithms to create a consensus.
87. Teber, E. T., Liu, J. Y., Ballouz, S., Fatkin, D. & Wouters, M. A. Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. *BMC Bioinformatics* **10** (Suppl. 1), S69 (2009).
88. Elbers, C. C. *et al.* A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol. Metab.* **18**, 19–26 (2007).
89. Thienpont, B. *et al.* Haploinsufficiency of TAB2 causes congenital heart defects in humans. *Am. J. Hum. Genet.* **86**, 839–849 (2010).
This is a biological validation of Endeavour that shows a role for TAB2 in human cardiac development.
90. Qiao, Y. *et al.* Outcome of array CGH analysis for 255 subjects with intellectual disability and search for candidate genes using bioinformatics. *Hum. Genet.* **128**, 179–194 (2010).
91. Hwang, S., Rhee, S. Y., Marcotte, E. M. & Lee, I. Systematic prediction of gene function in *Arabidopsis thaliana* using a probabilistic functional gene network. *Nature Protoc.* **6**, 1429–1442 (2011).
92. Hess, D. C. *et al.* Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. *PLoS Genet.* **5**, e1000407 (2009).
93. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
94. Lee, I. *et al.* Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc. Natl Acad. Sci. USA* **108**, 18548–18553 (2011).
95. Kohavi, R. A. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. 15th Int. Joint Comp. Artificial Intelligence* **2**, 1137–1143 (1995).
96. Chen, Y. *et al. In silico* gene prioritization by integrating multiple data sources. *PLoS ONE* **6**, e21137 (2011).
97. Schuierer, S., Tranchevent, L.-C., Dengler, U. & Moreau, Y. Large-scale benchmark of Endeavour using MetaCore maps. *Bioinformatics* **26**, 1922–1923 (2010).
98. Huttenhower, C. *et al.* The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. *Bioinformatics* **25**, 2404–2410 (2009).
99. Erlich, Y. *et al.* Exome sequencing and disease-network analysis of a single family implicate a mutation in *KIF1A* in hereditary spastic paraparesis. *Genome Res.* **21**, 658–664 (2011).
This is a study in which traditional mapping methods, new sequencing tools and network analysis are combined to identify the causal mutation for a rare monogenic disease.
100. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
101. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
102. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.* **4**, 44–57 (2009).
103. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
104. Casci, T. Human disease: something old, something new. *Nature Rev. Genet.* **12**, 382–383 (2011).
105. Gillis, J. & Pavlidis, P. The impact of multifunctional genes on "guilt by association" analysis. *PLoS ONE* **6**, e17258 (2011).
106. Gillis, J. & Pavlidis, P. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* **8**, e1002444 (2012).
107. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. & Pedersen, J. T. Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* **29** (Suppl. 1), 2–6 (1997).
108. Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins* **79** (Suppl. 1), 1–5 (2011).
109. Arighi, C. N. *et al.* BioCreative III interactive task: an overview. *BMC Bioinformatics* **12** (Suppl. 8), S4 (2011).
110. Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6** (Suppl. 1), S1 (2005).
111. Tilstone, C. DNA microarrays: vital statistics. *Nature* **424**, 610–612 (2003).
112. Johnson, K. & Lin, S. Call to work together on microarray data analysis. *Nature* **411**, 885 (2001).
113. Prill, R. J., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K. & Stolovitzky, G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.* **4**, mr7 (2011).
114. Stein, L. D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Rev. Genet.* **9**, 678–688 (2008).
115. Yoshida, Y. *et al.* PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.* **37**, W147–W152 (2009).
116. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
117. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
118. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Rev. Genet.* **12**, 628–640 (2011).
119. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
120. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* **36**, D684–D688 (2008).
121. Baron, D. *et al.* MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. *Bioinformatics* **27**, 725–726 (2011).
122. Chen, R., Li, L. & Butte, A. J. AILUN: reannotating gene expression data automatically. *Nature Methods* **4**, 879 (2007).
123. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).

124. Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10** (Suppl. 1), S6 (2009).
125. Smedley, D. *et al.* BioMart—biological queries made easy. *BMC Genom.* **10**, 22 (2009).
126. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
127. Yu, H. *et al.* Annotation transfer between genomes: protein–protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
128. Ebermann, I. *et al.* A novel gene for Usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss. *Hum. Genet.* **121**, 203–211 (2007).
129. Barriot, R. *et al.* Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Med.* **2**, 16 (2010).
This study describes CHDWiki, the first knowledge portal to annotate and analyse gene–phenotype networks collaboratively.

Acknowledgements
This work was supported in part by the following grants: KUL PFV/10/016 SymBioSys, KUL GOA MaNet, Hercules III PacBio RS and FP7-HEALTH CHeartED.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
ArrayExpress: http://www.ebi.ac.uk/arrayexpress
Ensembl Genome Browser: http://www.ensembl.org
Gene Expression Omnibus: http://www.ncbi.nlm.nih.gov/geo
Gene Prioritization Portal: http://www.esat.kuleuven.be/gpp
Gene Ontology: http://www.geneontology.org
Genetic Association Database: http://geneticassociationdb.nih.gov
GoPubmed: http://www.gopubmed.org
HUGO Gene Nomenclature Committee: http://www.genenames.org
Human Gene Mutation Database: http://www.hgmd.cf.ac.uk
Ingenuity Pathway Analysis: http://www.ingenuity.com/products/pathways_analysis.html
KU Leuven Bioinformatics Laboratory: http://www.kuleuven.be/bioinformatics
KU Leuven *SymBioSys* Center for Computational Systems Biology: http://www.kuleuven.be/symbiosys
Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.jp/kegg
MetaCore (from GeneGO): http://www.genego.com/metacore.php
Online Mendelian Inheritance in Man (OMIM): http://omim.org
STRING: http://string-db.org
UCSC Genome Browser: http://genome.ucsc.edu

SUPPLEMENTARY INFORMATION
See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF