ENDEAVOUR update: a web resource for gene prioritization in multiple species

Léon-Charles Tranchevent[1,*], Roland Barriot[1,*], Shi Yu[1], Steven Van Vooren[1], Peter Van Loo[1,2,3], Bert Coessens[1], Bart De Moor[1], Stein Aerts[3,4] and Yves Moreau[1,#]

[*]These authors contributed equally to this work;[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven (Belgium); [2]Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium); [3]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine, (Belgium); [4]Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium); [#]e-mail: yves.moreau@esat.kuleuven.be

Keywords: gene prioritization, disease gene, data fusion.
This web site is free and open to all users and there is no login requirement.

Server site: http://www.esat.kuleuven.be/endeavourweb

Abstract

ENDEAVOUR is a web resource for the prioritization of candidate genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of (1) inferring several models (based on various genomic data sources), (2) applying each model to the candidate genes to rank those candidates against the profile of the known genes, and (3) merging the several rankings into a global ranking of the candidate genes. In the present article, we describe the latest developments of ENDEAVOUR. These occur at three levels. First, we provide a web-based user interface, besides our Java Web Start client, to make ENDEAVOUR more universally accessible. Second, we support multiple species: in addition to *Homo sapiens*, we now provide gene prioritization for three major model organisms: *Mus musculus*, *Rattus norvegicus*, and *Caenorhabditis elegans*. Third, ENDEAVOUR makes use of additional data sources and is now including numerous databases: ontologies and annotations, protein-protein interactions, *cis*-regulatory information, gene expression data sets, sequence information and text-mining data. We tested the novel version of ENDEAVOUR on seven recent disease gene associations from the literature. Additionally we describe a number of recent independent studies that made use of ENDEAVOUR to prioritize candidate genes for obesity and Type II diabetes, cleft lip and cleft palate, and pulmonary fibrosis. ENDEAVOUR is publicly available at http://www.esat.kuleuven.be/endeavourweb.

Background

With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes . To identify which genes are responsible for which phenotype, association studies and linkage analyses are often used, resulting in large lists of

candidate genes. In many cases, the list of candidates can be narrowed down to a few dozen. However, it is generally too expensive and time-consuming to perform experimental validation for all these candidates. Therefore, these candidates may be prioritized to first validate the best ones. Given the amount of genomic data publicly available, it is often prohibitive to perform the prioritization manually and consequently, there is a need for computational approaches.

During the past five years, the bioinformatics community has developed several strategies to address this question, and several tools are available online . To our knowledge, all the tools use the concept of similarity. It is based on the assumption that similar phenotypes are caused by genes with similar or related functions . However, the tools differ by the strategy they adopt in calculating the similarity (either between the candidate genes and the phenotypes or between the candidate genes and the training genes) and by the data sources they use. The most commonly used data sources are text-mining data, gene expression data, and sequence information. Additionally, phenotypic data, protein-protein interactions, ontologies, and *cis*-regulatory information are sometimes included. However, most of the existing approaches mainly focus on the combination of few data sources. For instance, the CGI method proposed by Ma *et al.* combines expression and interaction data. Several methods only rely on literature and ontologies: BITOLA , POCUS Gentrepid , G2D and the method defined by Tiffin *et al.* . By contrast, systems that use more data sources have recently been designed, such as CAESAR , GeneSeeker , SUSPECTS , TOM and ENDEAVOUR . For a more detailed description of the available tools, see the reviews by Oti and Brunner or by Zhu and Zhao .

We previously presented the concept of gene prioritization through genomic data fusion and its implementation called ENDEAVOUR . This tool requires two inputs: the training genes, already known to be involved in the process under study, and the candidate genes to prioritize. ENDEAVOUR produces one output: the prioritized list of candidate genes, along with the rankings per data source. The algorithm is made up of three stages, called the training, scoring, and fusion stages. In the training stage, ENDEAVOUR uses the training genes provided by the user to infer several models, one per data source. For example, with ontology based data sources, genes are annotated with several terms and reciprocally one term can be associated to several genes. The algorithm selects only the significant terms, the ones that are over-represented in the training sets compared to the complete genome. Hence, the model consists of these significant terms together with their corresponding *p-values* that reflect the significance of the enrichment. In the scoring stage, the model is used to score the candidate genes and rank them according to their score. For ontologies, the algorithm scores each candidate independently by combining the *p-values* of its associated terms that are, at the same time, present in the model. The scores are then used to rank the candidates based on this one data source. In the final stage, the rankings per data source are fused into one global ranking using order statistics. Among the existing methods, the order statistics has the advantage of avoiding penalizing genes that are absent from a given data source. Indeed, the genomic data sources are almost always incomplete. For instance, some genes do not have any ontology annotations, while other genes do not have their corresponding probes spotted on the microarray platform for which data is available. The order statistics allows us to combine the rankings per data source, taking missing values into account. Thus, the use of 'unbiased' data sources (*e.g.*, gene expression data, *cis*-regulatory motifs, and protein sequences), together with the use of the order statistics, allows us to obtain results that are not overly biased towards the most studied genes . The use of several data sources is indeed an important strength of our approach: combining two data sources, although possibly incomplete, can be more powerful than either individual data source, as shown by our validation experiments . The fact that our approach does not rely only on a single data source also reinforces its robustness to noisy data

sources like microarray data. More details about the training and scoring methods, the data sources and the order statistics can be found in Supplementary Tables 1 and 2, and in Supplementary Note 1.

In the present article, we describe a novel intuitive web interface in addition to the original Java client. Furthermore, three major model organisms have been added to the application: *M. musculus*, *R. norvegicus*, and *C. elegans* (*Danio rerio* and *Drosophila melanogaster* versions will be made available in 2008). Finally, novel data sources have been integrated including numerous protein-protein interaction databases and large species-specific expression data sets, bringing the number of available data sources to 26. Apart from our extensive validation , other recent independent publications confirm that ENDEAVOUR is efficient in identifying novel disease genes. Indeed, ENDEAVOUR was recently applied to analyze the adipocyte proteome and to propose novel genes involved in Type II diabetes , cleft lip and cleft palate phenotypes , and pulmonary fibrosis .

Outline of the ENDEAVOUR web server

ENDEAVOUR was first implemented as a Java client application interacting with a SOAP server and a MySQL database. To make it more universally accessible, we have developed a PHP web-based interface that runs with the most common web browsers, without the need for Java to be installed. It is freely accessible and there is no login requirement.

A four-step wizard guides the user through the preparation of the prioritization (Figure 1). The first step is to choose the organism: human, rat, mouse or worm. The second step is to specify the training set. The user can input a mixture of chromosomal bands, chromosomal intervals, gene symbols, EnsEMBL gene identifiers, KEGG identifiers, Gene Ontology identifiers, or OMIM disease names. Each input has to be prefixed according to its type. The rules are explained in the supplementary material and in the online manual. The genes corresponding to the input are retrieved and loaded into the application. The third step is to select the data sources to be used. The data sources available depend on the organism chosen in the first step. Some of these are species specific (*e.g.*, gene expression data sets) while others are more generic (*e.g.*, Gene Ontology annotations). The last step lets the user specify the candidate genes applying the same rules as in the second step. The user launches the prioritization by using a dedicated button. The computation time is dependent on the number of data sources used, the number of candidates and the load on our servers. The application can handle the prioritization of hundreds of genes (*e.g.*, the average computation time for 400 candidates using 10 data sources is 19.14 seconds over 100 repeats). Warnings and errors, such as unrecognized gene identifiers, are displayed in the console located in the middle of the main windows. The results are displayed at the bottom of the main page in three panels. The first panel contains the sprint plot, a graphical representation of the rankings with one column per data source plus an additional one for the global ranking. The genes are represented as boxes and the top ranking boxes are colored for better interpretation of the results. The second panel contains the raw scores and ranks for each gene in each data source. The user can sort the columns according to the global ranking or to any ranking per data source. The third panel contains the results as returned by the application (*i.e.* an XML file).

New model organisms and more data sources

ENDEAVOUR is designed as a generic prioritization tool and is equally useful for the prioritization of candidate disease genes as for candidate members of biological pathways and processes. This is illustrated in our previous publication where we used ENDEAVOUR to identify downstream genes of

myeloid differentiation. Since the fundamental study of biological processes is predominantly performed in model organisms, we decided to extend our framework to several model organisms. Currently, gene prioritization can be performed for *M. musculus*, *R. norvegicus,* and *C. elegans*, and we are also developing the versions for *D. rerio* and *D. melanogaster*. We have designed the web server so that the organism-specific versions use the same method for each generic data source (*e.g.*, Gene Ontology annotations).

The key strength of ENDEAVOUR resides in the fact that a lot of data sources are available and the user can select the ones that best correspond to the biological question under study. There are 8, 11, 12, and 20 data sources available respectively for *R. norvegicus*, *C. elegans*, *M. musculus*, and *H. sapiens*, which, in total, result in 26 distinct data sources. They can be classified into six categories: ontologies, interactions, expression, regulatory information, sequence data and text-mining data. Ontologies are structured vocabularies that are used to describe the function of the gene products. Ontologies give more insight on the molecular functions performed (Gene Ontology and SwissProt), on the biological processes involved in (Gene Ontology and KEGG), on the cellular components in which the gene products are active (Gene Ontology) and on the active domains of the proteins (InterPro). Interaction data come from databases that collect pairs of proteins that interact either physically or genetically. BIND and DIP curate the experimentally determined interactions collected from large-scale interaction and mapping experiments done using yeast two hybrid, mass spectrometry, genetic interactions and phage display. MINT and MIPS mine the literature, either manually or automatically, to find experimentally verified protein interactions. HPRD does the same with an emphasis on domain architecture, post-translational modifications, interaction networks and disease association. IntAct and BioGrid collect physical and genetic interactions by combining analysis of high-throughput experiments and literature curation. STRING and IntNetDb are large databases that contain all kinds of interactions. They rely on a statistical framework to integrate data coming from numerous experiments and databases (including several databases described above), and, additionally, the interactions are transferred across the different organisms, when applicable. Regarding the expression data, the preferred studies are the ones that include a large number of tissues and a large number of genes. Two sets are available for *H. sapiens* (Su *et al.* and Son *et al.*), three for *M. musculus* (Su *et al.* , Hovatta *et al.* , and Lindsley *et al.*) and one for *R. norvegicus* and *C. elegans,* respectively from the Walker *et al.* paper and the Baugh *et al.* study . Additionally, anatomical EST expression data from EnsEMBL are available for human. Regarding the *cis*-regulatory data, we only have information for *H. sapiens* currently. Using the TOUCAN toolbox and the upstream sequence of the genes, the algorithm looks for putative motifs and modules (combination of five motifs). There are two data sources that are based on sequences: the protein sequence similarities and the disease probabilities. For the latter, Ouzounis *et al.* and Adie *et al.* (ProspectR) used sequence features (*e.g.*, length of the sequence, length of the UTRs, number of introns, length of the introns) and a statistical framework to discriminate the human disease causing genes from the rest of the genome. Next, they associated to every gene a probability of being a disease causing gene, *a priori*. As for sequence similarity, an all-against-all similarity search is performed for all organisms using the NCBI BLAST . The data source based on literature mining relies on the TxtGate framework . The strategy is to screen the abstracts from PubMed with a manually curated vocabulary based on Gene Ontology. Similarly to the ontologies described above, it provides more information on the molecular functions and biological processes of the genes. It is important to notice that, except for the regulatory information category, each organism is provided with at least one data source per category.

As an alternative to the novel web-based application, one can use the original Java Web Start client, which is also extended to include the other model organisms. This application includes a few additional features, such as a full description of the models created, a full genome screening service in which the whole genome of the given organism can be prioritized, and the possibility for users to make use of their own microarray data sets. A SOAP service is also available to allow integration in workflows (*e.g.*, when using Taverna or Kepler).

Software documentation

ENDEAVOUR comes with an online manual. A subsection describes the concept of gene prioritization through genomic data fusion. Another subsection contains the answers to frequently asked questions and gives more details on how to perform a prioritization and how to interpret the results. Finally, a step-by-step example is given together with the corresponding screenshots.

The application is provided with three use cases taken from the literature. The user can run the examples by clicking on the corresponding buttons situated above the wizard that cause the training genes, the data sources, and the candidate genes to be loaded automatically into the application. Then, the user can quickly go through the four steps and launch the prioritization process. The three use cases can be used as a first step to understand the mechanisms of ENDEAVOUR. The first example is derived from our previous publication in which we studied the DiGeorge syndrome . This example shows why YPEL1 was first selected for wet lab experiments that eventually confirmed the phenotypic association in zebrafish. The second example is taken from the Elbers *et al.* review on obesity and Type II diabetes . They have prioritized five susceptibility loci to reveal a molecular link between the two disorders. ENDEAVOUR uncovered the susceptibility loci located on chromosome 11 for this example. It contains KCNJ5, a homolog of KCNJ11 that is known to contribute to the risk of Type II diabetes. We have built the last example after Ebermann *et al.* published their discovery of a novel Usher gene, DFNB31, that encodes the whirlin protein . By using data six months prior to the publication, we made sure that the association was not yet present in the databases. Among the 32 candidates of the chromosomal band 9q32, DFNB31 ranked first, showing that, retrospectively, it was indeed a good candidate.

Validation

Similarly to our previous work , we statistically validate the approach with a standard leave-one-out cross-validation using known gene sets. We produced the corresponding receiver operating characteristic (ROC) curves and measured the performance by calculating the area under the curve (AUC) (Figure 2). Here, we focused on the pathway gene prioritization for the newly added species by applying this scheme to three signalling pathways taken from the Gene Ontology database . These pathways are common to the four organisms and involve respectively 193, 170, 126, and 44 genes for *H. sapiens*, *M. musculus*, *R. norvegicus,* and *C. elegans*. We performed both a fair validation and a complete validation. For the fair validation, we excluded the data sources that might contain explicitly the gene-pathway association (*i.e.*, Gene Ontology, Kegg, String, and Text) while all data sources were used for the complete validation. The first observation is that the performance of the four control validations stays close to the theoretical expectation of 50% (respectively 48%, 39%, 45% and 51%). This means that when using randomly generated gene sets for training, we obtain random results. In contrast, the performance of biologically meaningful sets is much higher (respectively 88%, 92%, 90%,

and 86% for the fair validation and 99%, 99%, 99%, and 98% for the complete validation). An analysis per data source of the fair validation reveals that the global performance (*e.g.*, 88% for human) is always higher than the best performing data source performance (*e.g.*, 78% for human InterPro). It shows that our data fusion approach is scientifically sound and that it is crucial to make use of complementary data sources. Altogether, this indicates that our approach based on the assumption that functionally related genes often cause similar phenotypes can be applied successfully.

A difficulty of validating gene prioritization methods is the fact that known data are used for the ranking. In other words, for every disease or pathway gene, the link between the disease and the gene is described in the literature and sometimes evidence is also present in the ontologies or in the interaction information. Therefore, we excluded in the above analysis the data sources that contain explicit information about the similarity of the true positive to the training set. To assess the full performance of ENDEAVOUR to solve real biological cases, using all data sources, we therefore focused on genetic disorders for which associations were reported very recently in the literature, so that the explicit information is not yet present in our data. Particularly, we used all gene-disease associations that were reported in *Nature Genetics* after January 1st 2008 (Table 1), eight in total. For each disorder, we built a training set containing all the genes already know to play a role in that disorder according to the OMIM database (downloaded in august 2007). There was no gene yet known to cause the Kawasaki disorder in OMIM so we were unable to build a training set for this disease, leaving us with seven cases to study. As candidate genes to be ranked we used the true positive gene together with 99 genes that flank the true positive in the genome. These regions were then prioritized with ENDEAVOUR using all data sources and their specific training sets. The results are presented in Table 1. Interestingly, BANK1 and CTRC rank first out of their region and ITGAM ranks third. All the genes are within the top 20%, and five out of the seven genes are within the top 5%.

Others have used our gene prioritization tool as well. Elbers *et al.* have used ENDEAVOUR in combination with other prioritization tools to define the best strategy to search for common obesity and Type 2 diabetes genes . They suggest a list of genes indicated as potential candidates by at least two of the six tools. Tzouvelekis *et al.* have used ENDEAVOUR to prioritize a list of genes differentially expressed in idiopathic pulmonary fibrosis . They consistently find that among the top candidates, five and seven genes are targets of, respectively, tumor necrosis factor (TNF) and transforming growth factor (TGF). Osoegawa *et al.* applied ENDEAVOUR to propose novel genes associated with cleft lip and cleft palate phenotypes . They analysed 83 syndromic cases and 104 nonsyndromic cases and concluded that estrogen receptor 1 (ESR1) and fibroblast growth factor receptor 2 (FGFR2) were the most likely candidates respectively from region 6q25.1-25.2 and region 10q26.11-26.13. Using mass spectrometry and bioinformatics, Adachi *et al.* explored the proteome of the adipocyte, a central player in energy metabolism . Using ENDEAVOUR, they were able to associate a number of factors with vesicle transport in response to insulin stimulation, which is a key function of adipocytes.

Conclusion

ENDEAVOUR is a web server that allows users to prioritize candidate genes with respect to their biological processes or diseases of interest. It is provided with an intuitive four step wizard and an online manual. It is available for four organisms (*H. sapiens*, *M. musculus*, *R. norvegicus*, and *C. elegans*). ENDEAVOUR relies on the similarity between the candidates and the models built with the training genes. The approach has been validated experimentally , by extensive leave-one-out cross-validations, and by analysis of recently reported cases from the literature. Additionally, several independent

laboratories have used ENDEAVOUR to propose novel disease genes (Elbers *et al.* and Osoegawa *et al.*) or to optimize the analysis of medium-throughput experiments (Tzouvelekis *et al.* and Adachi *et al.*). Importantly, the cross-validation revealed the added value of combining several complementary data sources. With 26 distinct data sources (51 in total) covering most aspects of the knowledge available on genes and gene products (functional annotations, protein interactions, expression profiles, regulatory information, sequence based data and literature mining), ENDEAVOUR exploits the most comprehensive collection of publicly available knowledge.

Funding & acknowledgements

This research was supported by the Research Council KUL (GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc & fellow grants), FWO (PhD/postdoc grants, projects G. 0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM)), IWT (PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM Endometriosis), the Belgian Federal Science Policy Office (IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011), and the EU-RTD (ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap). The authors thank Sonia Leach for critical comments and helpful suggestions on the manuscript. PVL and SA are respectively supported by a PhD and a postdoctoral research fellowship of the Research Foundation – Flanders (FWO).

Tables

Gene	Disorder	Reference	Date	Endeavour rank
BANK1	Systemic lupus erythematosus	Kozyrev *et al.*	February 1st, 2008	1
ITGAM	Systemic lupus erythematosus	Nath *et al.*	February 1st, 2008	3
TNFSF4	Systemic lupus erythematosus	Graham *et al.*	January 1st, 2008	16
DPP6	Amyotropic lateral sclerosis	van Es *et al.*	January 1st, 2008	15
CTRC	Chronic pancreatitis	Rosendahl *et al.*	January 1st, 2008	1
ATP6V0A2	Impaired glycosylation	Kornak *et al.*	January 1st, 2008	5
ATP6V0A2	Cutis laxa	Kornak *et al.*	January 1st, 2008	5
			Mean	**6.57**
			Standard deviation	**6.32**

Table 1. Results of seven genetic disorder prioritizations. The gene-disease associations were reported in Nature Genetics after January 1st 2008 to exclude the presence of explicit evidence in our data sources. The training sets were built with OMIM and the candidate regions contain the novel gene and its 99 nearest neighbours. The 20 human data sources were used to perform the prioritizations. The results show that ENDEAVOUR ranked all the novel genes within the top 20%, and in five out of the seven cases, within the top 5%.

Figures

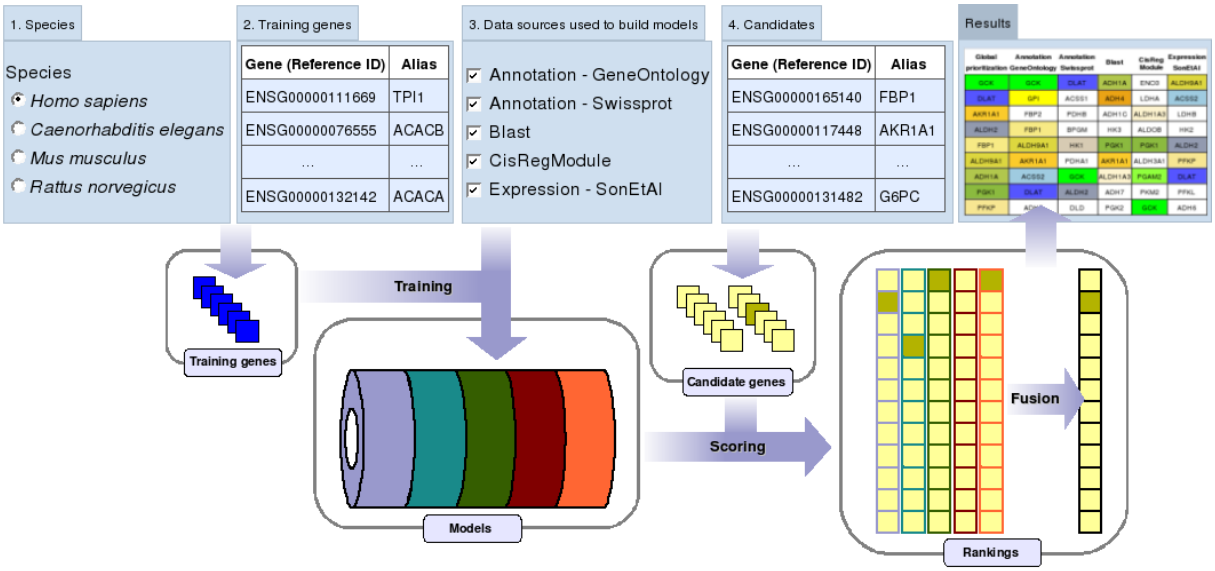

Figure 1. ENDEAVOUR: the algorithm behind the wizard. Once the organism of interest is chosen (Step 1), the user can specify the training genes (Step 2). Step 3 lets the user select the data sources that will be used to build the models. The models summarize the training gene information. The candidate genes specified by the user in Step 4 are then scored against the model. This produces one ranking per data source plus one global ranking obtained by fusion of the rankings per data source. The global ranking together with the rankings per data source are returned to the application and can be viewed in the 'Results' panel.

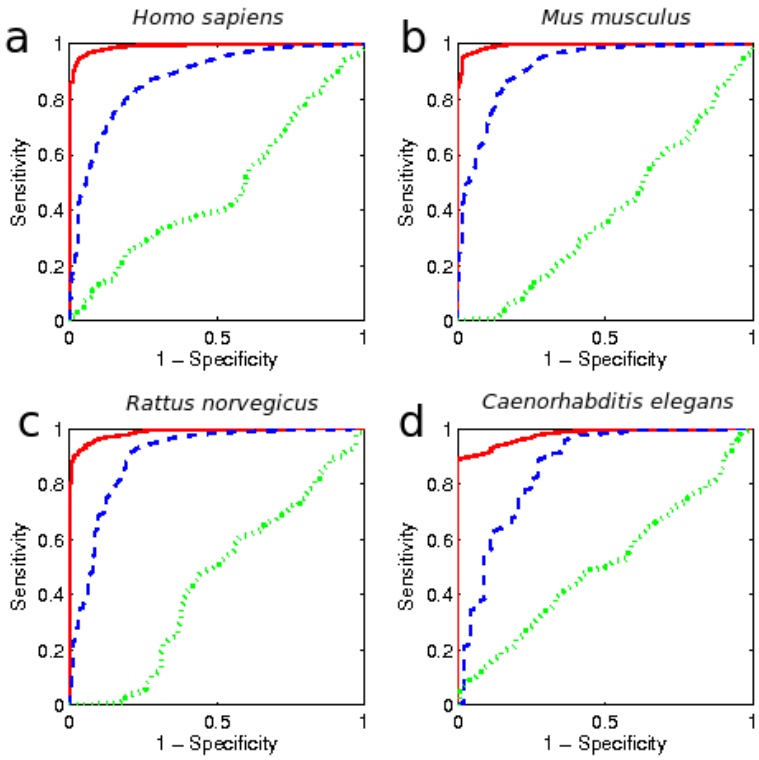

Figure 2. Results of the leave-one-out cross-validation. For each organism, the leave-one-out cross-validation was performed on 3 pathways sets from Gene Ontology , and, as a control, on 5 sets of 20 randomly selected genes. The ROC curves of the random (dotted green) and pathway validation (solid red and dashed blue) are plotted for (a) H. sapiens, (b) M. musculus, (c) R. norvegicus, and (d) C. elegans. Notice that for the fair validation (dashed blue), Gene Ontology, KEGG, Text and String were excluded while all data sources were used for the complete validation (solid red).The area under the curve (AUC) of the control validations are respectively 48%, 39%, 45% and 51% indicating a random performance. On the opposite, the AUC of the pathway validations are respectively 88%, 92%, 90%, and 86% for the fair validation and 99%, 99%, 99%, and 98% for the complete validation showing the validity of our approach.

Supplementary material

Type	Prefix (*)	Examples				Gene(s) loaded
		Human	***Mouse***	***Rat***	***Worm***	
Gene identifier		ENSG00000184895	ENSMUSG00000071964	ENSRNOG00000012772	WBGene00000966	The gene whose main identifier matches exactly the input
Gene symbol		OPTN	Tmem58	TAGL	T09E11.8	The gene whose symbol matches exactly the input
Chromosomal band	chr:	chr:1p36	chr:11A4	chr:3q22	*Not supported*	All genes located in the chromosomal band
KEGG	kegg:	kegg:05211	kegg:04540	kegg:00230	kegg:00624	All genes involved in the given KEGG pathway
Gene Ontology	go:	go:0019321	go:0005747	go:0004114	go:0006421	All genes annotated with the given GO term
OMIM	omim:	omim:parkin	*Not supported*	*Not supported*	*Not supported*	All genes involved in a disease that matches partially the input

Supplementary Table 1. Syntax for adding genes (Step 2 and Step 4). (*) The prefixes are case insensitive.

Supplementary Note 1: order statistics

A *Q* statistic is calculated from *n* rank ratios using the joint cumulative distribution of an *n*-dimensional order statistic as previously done by Stuart *et al* . The original recursive formula turned out to be highly inefficient because its complexity is O(*n*!). We use the recursive formula proposed by Aerts *et al* which presents the advantage of having a tractable complexity of O(n^2). An additional fitting step is performed to make the *Q* statistics uniformly distributed. For a number of data sources smaller or equal to 5, the distribution of the *Q* statistics is best fitted with a beta distribution. For a greater number of data sources, a gamma distribution is found to be the best approximation. The cumulative distribution function of these distributions provides us with uniformly distributed *p-values* that are then used to build the global ranking.

ENDEAVOUR data sources			
Data type	**Data sources**	**Training**	**Scoring**
Ontologies	Gene Ontology , InterPro , KEGG , EnsEMBLEst , and SwissProt/Uniprot	The enrichment of the terms is assessed using the binomial law. The model consists of the over-represented terms and their *p-values* that reflect the significance of the enrichment.	For a candidate, the score is obtained by combining the *p-values* of its annotations that are, at the same time, present in the model using the Fisher's omnibus.
Interactions	BIND , DIP , BioGrid , IntNetDb , MINT , MIPS , HPRD , IntAct , and STRING	The training genes and all their interactors are grouped into the model.	The score is the relative size of the overlap between the candidate gene plus its interactors and the model.
Expression	Su *et al.* , Son *et al.* , Baugh *et al.* , Hovatta *et al.* , Lindley *et al.* , and Walker *et al.*	All the expression profiles of the training genes are grouped into the model.	Pearson correlation between the candidate profile and each of the model profiles are computed. The score is the average of the 50% best correlations.
Regulatory information	Motifs	The model is the average profile, computed with all the training profiles.	The score is the Pearson correlation between the candidate profile and the model.
	cis-Regulatory module	The model is the best module of 5 motifs reported by ModuleSearcher.	The score, computed using ModuleScanner, represents the probability for a candidate gene to be regulated by the model module.
Sequence based data	Ouzounis *et al.* and ProspectR	*No training is needed.*	The score is one minus the probability of being involved in a disease.
	BLAST		The *e-values* of the protein alignments between all training and all candidate genes are collected. For a candidate, the score is the best *e-value* obtained.
Literature data	TxtGate using PubMed abtracts .	The model is the average term weight vector of the training genes.	The score is the cosine similarity measure between the candidate profile and the model.

Supplementary Table 2. The ENDEAVOUR data sources and details about the methods used for training and scoring.