Published online 28 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W377–W384***
doi:10.1093/nar/gkn325

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,*]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, [2]Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, [3]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and [4]Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

ABSTRACT

ENDEAVOUR (http://www.esat.kuleuven.be/endeavour web; this web site is free and open to all users and there is no login requirement) 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 (i) inferring several models (based on various genomic data sources), (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. In the present article, we describe the latest developments of ENDEAVOUR. First, we provide a web-based user interface, besides our Java 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 32 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.

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 (1–3). 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 5 years, the bioinformatics community has developed several strategies to address this question, and several tools are available online (4,5). 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 (1–3). 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

*To whom correspondence should be addressed. Tel: +32 16 32 17 09; Fax: +32 16 32 19 70; Email: yves.moreau@esat.kuleuven.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

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 combining gene expression and protein interaction data method proposed by Ma *et al*. (6) combines expression and interaction data. Several methods only rely on literature and ontologies: BITOLA (7), POCUS (8) Gentrepid (9), G2D (10) and the method defined by Tiffin *et al*. (11). In contrast, systems that use more data sources have recently been designed, such as CAESAR (12), GeneSeeker (13), SUSPECTS (14), TOM (15) and ENDEAVOUR (16). For a more detailed description of the available tools, see the reviews by Oti and Brunner (5) or by Zhu and Zhao (4).

We previously presented the concept of gene prioritization through genomic data fusion and its implementation called ENDEAVOUR (16). 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 (16). 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 (16). 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 (16), other recent independent publications confirm that ENDEAVOUR is efficient in identifying novel disease genes. Indeed, ENDEAVOUR was recently applied to analyze the adipocyte proteome (17) and to propose novel genes involved in Type II diabetes (18), cleft lip and cleft palate phenotypes (19), and pulmonary fibrosis (20).

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 (21) gene identifiers, KEGG (22) identifiers, Gene Ontology (23) identifiers or OMIM (24) 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 s 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

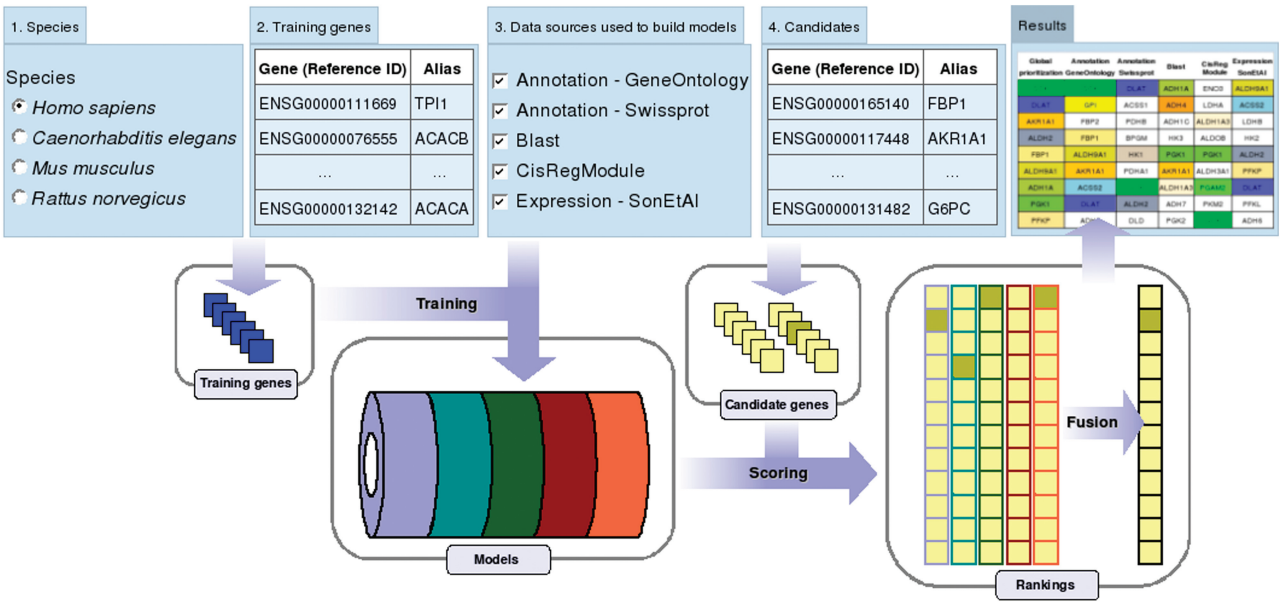

Figure 1. ENDEAVOUR: the algorithm behind the wizard. Once the organism of interest is chosen (Step 1), the user can specify the training genes (Step 2). Step 3 lets the user select the data sources that will be used to build the models. The models summarize the training gene information. The candidate 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.

represented as boxes and the top ranking boxes are coloured 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 allows one to export the results as a TSV spreadsheet or as an XML file. The user can also save the sprint plot using several picture formats (i.e. PNG, JPG and GIF).

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 (16) 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 (23) and SwissProt (25)], on the biological processes involved in [Gene Ontology and KEGG (22)], on the cellular components in which the gene products are active (Gene Ontology) and on the active domains of the proteins [InterPro (26)]. Interaction data come from databases that collect pairs of proteins that interact either physically or genetically. BIND (27) and DIP (28) 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 (29) and MIPS (30) mine the literature, either manually or automatically, to find experimentally verified protein interactions. HPRD (31) does the same with an emphasis on domain architecture, post-translational modifications, interaction networks and disease association. IntAct (32) and BioGrid (33) collect physical and genetic interactions by combining analysis of high-throughput experiments and literature curation. STRING (34) and IntNetDb (35) 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.* (36) and Son *et al.* (37)], three for *M. musculus* [Su *et al.* (36), Hovatta *et al.* (38) and Lindsley *et al.* (39)] and one for *R. norvegicus* and *C. elegans*, respectively from the Walker *et al.* paper (40) and the Baugh *et al.* study (41). Additionally, anatomical expression sequence tags (EST) expression data from EnsEMBL (21) are available for human. Regarding the *cis*-regulatory data, we only have information for *H. sapiens* currently. Using the TOUCAN toolbox (42) 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, Lopez-Bigas *et al.* (43) and Adie *et al.* (44) (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 (45). The data source based on literature mining relies on the TxtGate framework (46). The strategy is to screen the abstracts from PubMed (47) 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 (48) or Kepler (49)].

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 (16). 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.* (18) 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.* (50) 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 (16), 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 (23). 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

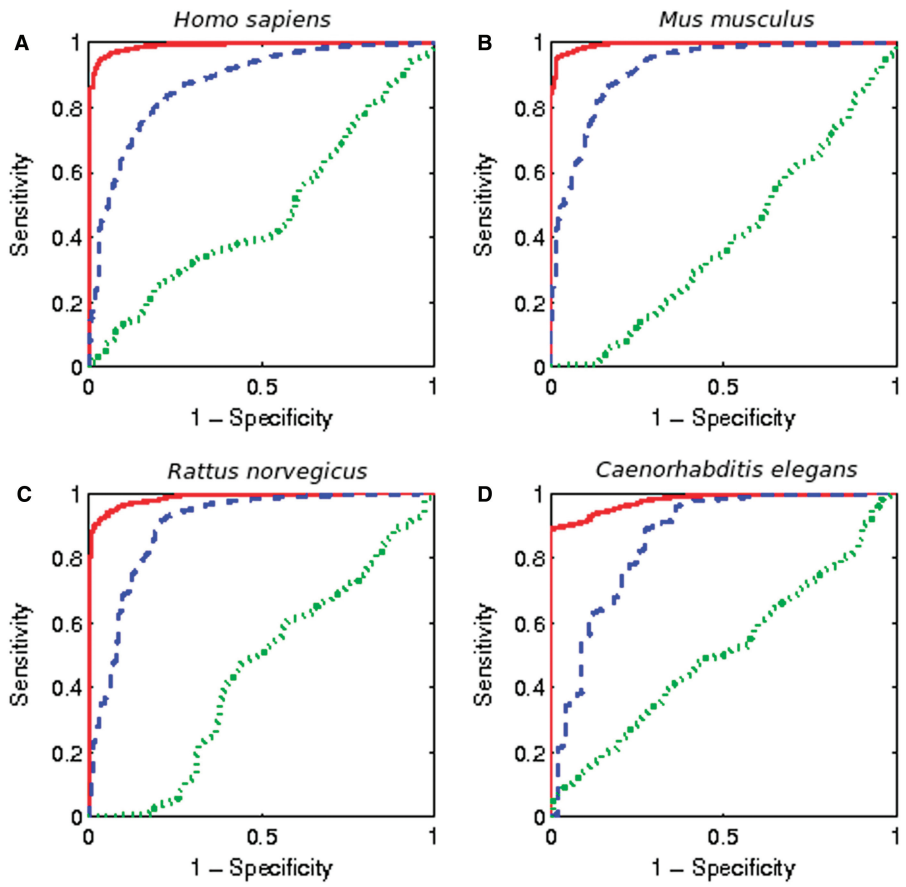

Figure 2. Results of the leave-one-out cross-validation. For each organism, the leave-one-out cross-validation was performed on three pathways sets from Gene Ontology (23), and, as a control, on five sets of 20 randomly selected genes. The ROC curves of the random (dotted green) and pathway validation (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 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.

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 gene–disease associations that were reported in *Nature Genetics* after 1 January 2008 (Table 1), 32 in total. For each disorder, we built a training set containing all the genes already known to play a role in that disorder according to the OMIM and Gene Ontology databases (both downloaded in August 2007). 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*, *CTRC* and *SORT1* rank first out of their region and *GDF5*, *RGS1* and *SH2B3* rank second. All genes but four are within the top 20% and half of them are within the top 9%.

Others have used our gene prioritization tool as well. Elbers *et al.* (18) have used ENDEAVOUR in combination with other prioritization tools to define the best strategy to search for common obesity and Type II diabetes genes. They suggest a list of genes indicated as potential candidates by at least two of the six tools. Tzouvelekis *et al.* (20) 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.* (19) applied ENDEAVOUR to propose novel genes associated with cleft lip and cleft palate phenotypes. They analysed 83 syndromic cases and 104 non-syndromic 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.* (17) explored the proteome of the adipocyte, a central player in energy metabolism.

Table 1. Results of the thirty two genetic disorder prioritizations

Gene	Disorder	Reference	Endeavour rank
BANK1	Systemic lupus erythematosus	Kozyrev *et al.* (51)	1
ITGAM	Systemic lupus erythematosus	Nath *et al.* (52)	3
TNFSF4	Systemic lupus erythematosus	Graham *et al.* (53)	16
DPP6	Amyotropic lateral sclerosis	van Es *et al.* (54)	15
CTRC	Chronic pancreatitis	Rosendahl *et al.* (55)	1
ATP6V0A2	Impaired glycosylation	Kornak *et al.* (56)	5
ATP6V0A2	Cutis laxa	Kornak *et al.* (56)	5
GALNT2[a]	LDL/HDL cholesterol	Willer *et al.* (57), Kathiresan *et al.* (58)	13
SORT1[a]	LDL/HDL cholesterol	Willer *et al.* (57), Kathiresan *et al.* (58)	1
MLXIPL[a]	LDL/HDL cholesterol	Willer *et al.* (57), Kathiresan *et al.* (58), Kooner *et al.* (59),	12
GDF5[a]	Human height	Sanna *et al.* (60)	2
C20orf44[a]	Human height	Sanna *et al.* (60)	41
MSMB[a]	Prostate cancer	Eeles *et al.* (61), Thomas *et al.* (62)	18
JAZF1[a]	Prostate cancer	Thomas *et al.* (62)	14
CTBP2[a]	Prostate cancer	Thomas *et al.* (62)	4
LMTK2[a]	Prostate cancer	Eeles *et al.* (61)	4
KLK3[a]	Prostate cancer	Eeles *et al.* (61)	9
CPNE3[a]	Prostate cancer	Thomas *et al.* (62)	42
IL16[a]	Prostate cancer	Thomas *et al.* (62)	9
CDH23[a]	Prostate cancer	Thomas *et al.* (62)	40
EHBP1[a]	Prostate cancer	Gudmundsson *et al.* (63)	19
CCR3[a]	Celiac disease	Hunt *et al.* (64)	12
RGS1[a]	Celiac disease	Hunt *et al.* (64)	2
LPP[a]	Celiac disease	Hunt *et al.* (64)	30
TAGAP[a]	Celiac disease	Hunt *et al.* (64)	3
SH2B3[a]	Celiac disease	Hunt *et al.* (64)	2
IL12A[a]	Celiac disease	Hunt *et al.* (64)	18
SCHIP1[a]	Celiac disease	Hunt *et al.* (64)	20
IL18R1[a]	Celiac disease	Hunt *et al.* (64)	3
IL18RAP[a]	Celiac disease	Hunt *et al.* (64)	4
IL2[a]	Celiac disease	Hunt *et al.* (64)	10
IL21[a]	Celiac disease	Hunt *et al.* (64)	14
		Mean (all genes)	12.25
		Mean (GWAS excluded)	6.57

[a]Associations reported with GWAS (Genome Wide SNPs Associations Studies).
The gene-disease associations were reported in Nature Genetics after 1 January 2008 to exclude the presence of explicit evidence in our data sources. The training sets were built with OMIM and Gene Ontology; 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 but four within the top 20%, and half of them within the top 9%.

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 (16), 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.* (18) and Osoegawa *et al.* (19)] or to optimize the analysis of medium-throughput experiments [Tzouvelekis *et al.* (20) and Adachi *et al.* (17)]. 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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

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 article. P.V.L. and S.A. are, respectively, supported by a PhD and a postdoctoral research fellowship of the Research Foundation—Flanders (FWO).

Conflict of interest statement. None declared.

REFERENCES

1. Smith,N.G. and Eyre-Walker,A. (2003) Human disease genes: patterns and predictions. *Gene*, **318**, 169–175.
2. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabási,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
3. Jimenez-Sanchez,G., Childs,B. and Valle,D. (2001) Human disease genes. *Nature*, **409**, 853–855.
4. Zhu,M. and Zhao,S. (2007) Candidate gene identification approach: progress and challenges. *Int. J. Biol. Sci.*, **3**, 420–427.
5. Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.
6. Ma,X., Lee,H., Wang,L. and Sun,F. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.
7. Hristovski,D., Peterlin,B., Mitchell,J.A. and Humphrey,S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.*, **74**, 289–298.
8. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
9. George,R.A., Liu,J.Y., Feng,L.L., Bryson-Richardson,R.J., Fatkin,D. and Wouters,M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.
10. Perez-Iratxeta,C., Wjst,M., Bork,P. and Andrade,M.A. (2005) G2D: a tool for mining genes associated with disease. *BMC Genet.*, **6**, 45.
11. Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, **33**, 1544–1552.
12. Gaulton,K.J., Mohlke,K.L. and Vision,T.J. (2007) A computational system to select candidate genes for complex human traits. *Bioinformatics*, **23**, 1132–1140.
13. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P., Leunissen,J.A., Brunner,H.G. and Vriend,G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, **33**, W758–W761.
14. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.
15. Rossi,S., Masotti,D., Nardini,C., Bonora,E., Romeo,G., Macii,E., Benini,L. and Volinia,S. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res.*, **34**, W285–W292.
16. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.-C., De Moor,B., Marynen,P., Hassan,B. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
17. Adachi,J., Kumar,C., Zhang,Y. and Mann,M. (2007) In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol. Cell. Proteomics*, **6**, 1257–1273.
18. Elbers,C., Onland-Moret,C., Franke,L., Niehoff,A., van der Schouw,Y. and Wijmenga,C. (2007) A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol. Metab.*, **18**, 19–26.
19. Osoegawa,K., Vessere,G., Utami,K., Mansilla,M., Johnson,M., Riley,B., L'Heureux,J., Pfundt,R., Staaf,J., van der Vliet,W. *et al.* (2008) Identification of novel candidate genes associated with cleft lip and palate using array comparative genomic hybridisation. *J. Med. Genet.*, **45**, 81–86.
20. Tzouvelekis,A., Harokopos,V., Paparountas,T., Oikonomou,N., Chatziioannou,A., Vilaras,G., Tsiambas,E., Karameris,A., Bouros,D. and Aidinis,V. (2007) Comparative expression profiling in pulmonary fibrosis suggests a role of hypoxia-inducible factor-1alpha in disease pathogenesis. *Am. J. Respir. Crit. Care Med.*, **176**, 1108–1119.
21. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
22. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
23. The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
24. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
25. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
26. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
27. Bader,G., Donaldson,I., Wolting,C., Ouellette,F., Pawson,T. and Hogue,C. (2001) BIND-The biomolecular interaction network database. *Nucleic Acids Res.*, **29**, 242–245.
28. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
29. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the molecular interaction database. *Nucleic Acids Res.*, **35**, D572–D574.
30. Mewes,H.W., Frishman,D., Mayer,K.F.X., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. and Stumpflen,V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
31. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
32. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct-open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
33. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
34. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
35. Xia,K., Dong,D. and Han,J.D. (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, **7**, 508.
36. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
37. Son,C.G., Bilke,S., Davis,S., Greer,B.T., Wei,J.S., Whiteford,C.C., Chen,Q.R., Cenacchi,N. and Khan,J. (2005) Database of mRNA

gene expression profiles of multiple human organs. *Genome Res.*, **15**, 443–450.
38. Hovatta,I., Tennant,R.S., Helton,R., Marr,R.A., Singer,O., Redwine,J.M., Ellison,J.A., Schadt,E.E., Verma,I.M., Lockhart,D.J. *et al.* (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature*, **438**, 662–666.
39. Lindsley,R.C., Gill,J.G., Kyba,M., Murphy,T.L. and Murphy,K.M. (2006) Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. *Development*, **133**, 3787–3796.
40. Walker,J.R., Su,A.I., Self,D.W., Hogenesch,J.B., Lapp,H., Maier,R., Hoyer,D. and Bilbe,G. (2004) Applications of a rat multiple tissue gene expression data set. *Genome Res.*, **14**, 742–749.
41. Baugh,L.R., Hill,A.A., Claggett,J.M., Hill-Harfe,K., Wen,J.C., Slonim,D.K., Brown,E.L. and Hunter,C.P. (2005) The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. *Development*, **132**, 1843–1854.
42. Aerts,S., Van Loo,P., Thijs,G., Mayer,H., de Martin,R., Moreau,Y. and De Moor,B. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.*, **33**, W393–W396.
43. Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.
44. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.
45. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
46. Glenisson,P., Coessens,B., Van Vooren,S., Mathys,J., Moreau,Y. and De Moor,B. (2004) TXTGate: profiling gene groups with text-based information. *Genome Biol.*, **5**, R43.
47. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
48. Oinn,T., Addis,M., Ferris,J., Marvin,D., Senger,M., Greenwood,M., Carver,T., Glover,K., Pocock,M.R., Wipat,A. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.
49. Altintas,I., Berkley,C., Jaeger,E., Jones,M., Ludäscher,B. and Mock,S. (2004) *16th International Conference on Scientific and Statistical Database Management* Santorini Island, Greece.
50. Ebermann,I., Scholl,H.P., Charbel Issa,P., Becirovic,E., Lamprecht,J., Jurklies,B., Millan,J.M., Aller,E., Mitter,D. and Bolz,H. (2007) 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.
51. Kozyrev,S.V., Abelson,A.K., Wojcik,J., Zaghlool,A., Linga Reddy,M.V., Sanchez,E., Gunnarsson,I., Svenungsson,E., Sturfelt,G., Jönsen,A. *et al.* (2008) Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat. Genet.*, **40**, 211–216.
52. Nath,S.K., Han,S., Kim-Howard,X., Kelly,J.A., Viswanathan,P., Gilkeson,G.S., Chen,W., Zhu,C., McEver,R.P., Kimberly,R.P. *et al.* (2008) A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. *Nat. Genet.*, **40**, 152–154.
53. Graham,D.S., Graham,R.R., Manku,H., Wong,A.K., Whittaker,J.C., Gaffney,P.M., Moser,K.L., Rioux,J.D., Altshuler,D., Behrens,T.W. *et al.* (2008) Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. *Nat. Genet.*, **40**, 83–89.
54. van Es,M.A., van Vught,P.W., Blauw,H.M., Franke,L., Saris,C.G., Van den Bosch,L., de Jong,S.W., de Jong,V., Baas,F., van't Slot,R. *et al.* (2008) Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, **40**, 29–31.
55. Rosendahl,J., Witt,H., Szmola,R., Bhatia,E., Ozsvári,B., Landt,O., Schulz,H.U., Gress,T.M., Pfützer,R., Löhr,M. *et al.* (2008) Chymotrypsin C (CTRC) variants that diminish activity or secretion are associated with chronic pancreatitis. *Nat. Genet.*, **40**, 78–82.
56. Kornak,U., Reynders,E., Dimopoulou,A., van Reeuwijk,J., Fischer,B., Rajab,A., Budde,B., Nürnberg,P., Foulquier,F., ARCL Debré-type Study Group. *et al.* (2008) Impaired glycosylation and cutis laxa caused by mutations in the vesicular H+-ATPase subunit ATP6V0A2. *Nat. Genet.*, **40**, 32–34.
57. Willer,C.J., Sanna,S., Jackson,A.U., Scuteri,A., Bonnycastle,L.L., Clarke,R., Heath,S.C., Timpson,N.J., Najjar,S.S., Stringham,H.M. *et al.* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.*, **40**, 161–169.
58. Kathiresan,S., Melander,O., Guiducci,C., Surti,A., Burtt,N.P., Rieder,M.J., Cooper,G.M., Roos,C., Voight,B.F., Havulinna,A.S. *et al.* (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.*, **40**, 189–197.
59. Kooner,J.S., Chambers,J.C., Aguilar-Salinas,C.A., Hinds,D.A., Hyde,C.L., Warnes,G.R., Gómez,P.J., Frazer,K.A., Elliott,P., Scott,J. *et al.* (2008) Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.*, **40**, 149–151.
60. Sanna,S., Jackson,A.U., Nagaraja,R., Willer,C.J., Chen,W.-M., Bonnycastle,L.L., Shen,H., Timpson,N., Lettre,G., Usala,G. *et al.* (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.*, **40**, 198–203.
61. Eeles,R.A., Kote-Jarai,Z., Giles,G.G., Olama,A.A.A., Guy,M., Jugurnauth,S.K., Mulholland,S., Leongamornlert,D.A., Edwards,S.M., Morrison,J. *et al.* (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.*, **40**, 316–321.
62. Thomas,G., Jacobs,K.B., Yeager,M., Kraft,P., Wacholder,S., Orr,N., Yu,K., Chatterjee,N., Welch,R., Hutchinson,A. *et al.* (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, **40**, 310–315.
63. Gudmundsson,J., Sulem,P., Rafnar,T., Bergthorsson,J.T., Manolescu,A., Gudbjartsson,D., Agnarsson,B.A., Sigurdsson,A., Benediktsdottir,K.R., Blondal,T. *et al.* (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.*, **40**, 281–283.
64. Hunt,K.A., Zhernakova,A., Turner,G., Heap,G.A.R., Franke,L., Bruinenberg,M., Romanos,J., Dinesen,L.C., Ryan,A.W., Panesar,D. *et al.* (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.*, **40**, 395–402.