TITLE TO BE FOUND...

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

[*]These authors contribute 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 School of Medicine, Leuven (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 disease genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of inferring several models (based on various genomic data sources), applying each model to the candidate genes to rank those candidates against the profile of the known genes, and merging the several rankings into a global final ranking of the candidate genes. In the present article, we describe the latest developments of ENDEAVOUR. These occurred at three levels. First, we provide a web user interface (instead of our previous Java client) to make ENDEAVOUR more universally accessible. Second, we support multiple species. In addition to *Homo sapiens*, we now provide candidate gene prioritization for three major model organisms: *Mus musculus*, *Rattus norvegicus*, and *Caenorhabditis elegans*. Third, ENDEAVOUR makes use of additional data sources and now includes numerous databases: ontologies and annotations, protein-protein interactions, *cis*-regulatory information, gene expression data sets, sequence information and text-mining data. For validation, we performed leave-one-out cross-validation on known data and applied the algorithm to the DiGeorge syndrome to discover and validate a novel association with YPEL1. Additionally, we present the 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 of high-throughput technologies,many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of large amount of genomic data and the creation and maintenance of the corresponding databases. However, converting genomic data into biological knowledge is not an easy task, and, particularly, the identification of genes involved in health and disease remains a challenge. Nevertheless, there is a lot of evidence to suggest that similar phenotypes are often caused by functionally related genes. To identify which genes are responsible for which phenotype, gene-association studies and linkage

analyses are often successfully used, which results in long lists of candidates. The development of more and more precise technologies allows geneticists to narrow down the candidate list to a few dozens of genes. However, it is still too expensive and time consuming to validate them all experimentally. Often, the candidates are prioritized to validate the best candidates first. Given the amount of genomic data publicly available, it is extremely challenging to perform the prioritization manually and therefore there is a need for a computational approach.

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 evidence that suggests that similar phenotypes are caused by genes with similar 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 training and candidate 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.

[To complete]
Most of the existing approaches mainly focus on the combination of two data sources like for instance x et al [XX] who combine text-mining data and expression data or xx et al [XX] who combine interactions data with expression data to predict ...

For a more detailed description of the available tools, see the reviews by Oti and Brunner [2] or by Zhu and Zhao [3].

Aerts et al [1] presented the concept of gene prioritization through genomic data fusion and its implementation called ENDEAVOUR. It is fed with two inputs: the training genes (already known to be involved in the process under study) and the candidate genes (the list to prioritize). In return, it produces one output: the prioritized list of candidate genes along with the rankings per data source. The algorithm itself is made up of three steps: the training, scoring, and fusion steps. In the training step, we use the training genes provided by the user to infer several models, one per data source. For example, with ontology based data sources, the genes are annotated with several terms and reciprocally one term can be associated with several genes. The algorithm selects only the significant terms, being the ones that are over-represented in the training sets compared to the genome. At the end, the model consists of the significant terms together with their corresponding *p-values* that reflect the significance of the enrichment. More details about the methods for the other data sources can be found in the supplementary material (X). In the scoring step, the model is used to score and rank the candidate genes according to their score. Following the same example, 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. As in Step 1, all details about the methods for the other data sources can be found in the supplementary material (X). In the last step, the rankings per data source are fused into one global final ranking using order statistics. Among the existing methods, we have chosen the order statistics to avoid the penalization of the 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 Gene Ontology annotations, while other genes do not have their corresponding probes spotted on the microarray platform for which data is available. The order statistics allow us to combine the rankings

per data source which differ by their length. Thus, the use of unbiased data sources (gene expression data, cis-regulatory motifs, protein sequences for example) together with the use of the order statistics allow us to get results that are not overly biased towards the most studied genes. The statistics involved in this last step are fully described in the supplementary material (X). The use of several data sources is one strength of our approach. Indeed, when combined, two data sources (even incomplete) can be more powerful than the best of the two alone. This can be seen in the technical validation when the overall performance is greater than the performance of the best per data source. The fact that our approach does not rely only on a single data source also reinforces its robustness to noise and to noisy data sources (like microarray data). Thus, the number of false positive in the final ranking is expected to be lower than in the rankings per data source.

In this article, we describe three main improvements made to the original application. First, we now provide a web interface (instead of the Java client) to make ENDEAVOUR more universally accessible. Second, we now support multiple species. In addition to *Homo sapiens*, we now provide candidate gene prioritization for three major model organisms: *Mus musculus*, *Rattus norvegicus* and *Caenorhabditis elegans –Danio rerio* and *Drosophila melanogaster* are scheduled for release mid 2008. Third, ENDEAVOUR makes use of additional data sources and now includes numerous databases: (1) ontologies that describe the function of the genes and proteins (Gene Ontology, SwissProt, InterPro, Kegg), (2) physical or genetic protein-protein interactions (BIND, BioGrid, IntAct, HPRD, String), (3) *cis*-regulatory information (binding motifs and *cis*-regulatory modules), large species specific expression data sets (GEO), sequence data (BLAST, ProspectR) and text-mining data.

Besides this web application, we used ENDEAVOUR to propose YPEL1 as a novel DiGeorge syndrome gene that was then validated in zebrafish. ENDEAVOUR was also used by several teams around the world to propose novel genes involved in Type II diabetes [4], novel genes associated with cleft lip and cleft palate [5], and novel genes involved in pulmonary fibrosis [6] or to analyze the adipocyte proteome [7].

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 web server 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 in preparing the prioritization. The first step is to choose the organism to work with: human, rat, mouse or worm. The second step is to define the training set. The user can input a chromosomal band, a chromosomal interval, a list of gene symbols, a list of EnsEMBL gene identifiers, a Kegg identifier, a Gene Ontology identifier, or an OMIM disease name. 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 choose the data source to be used. The data sources available depend on the organism chosen in the first step. Some of them are species specific (gene expression data sets, or OMIM, for example) while others are more generic (Gene Ontology annotations, for example). In the last step, the user must input the candidate genes, as in Step 2, the same rules apply. The user launches the prioritization by using the dedicated button. The computation time is related to the number of data sources used, the size of the candidate set and the load on our servers, but 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 that are produced by the events described above are displayed in the console located in the middle of the main windows. Eventually, the results are produced and 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 and for each data source, the sprint plot is the graphical version of this panel. The user can sort the list 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). Further developments will include the saving and loading of this XML file.

New model organisms and more data sources

Our knowledge of the human biology benefits a lot from experiments originally performed in model organisms, such as mouse or rat. This is also true when considering the knowledge on health and disease where a lot of discoveries are mainly useful to human but were first discovered in model organisms. In our previous publication [1], we presented the *Homo sapiens* version of ENDEAVOUR. To our knowledge, none of the existing prioritization tools is available for other organisms than human. For the present study, we have enriched the application with three major model organisms: *Mus musculus*, *Rattus norvegicus,* and *Caenorhabditis elegans*. Furthermore, we are currently developing the versions for two other model organisms – *Danio rerio* and *Drosophila melanogaster* – and the release is scheduled for mid 2008. We have designed the web server so that the organism specific versions use the same algorithm and furthermore the same method for each generic data source (Gene Ontology annotations for instance).

Another strength of ENDEAVOUR is the fact that a lot of data sources are available and the user can pick up the ones that best correspond to the biological question under study.

In total, 26 data sources are available, (repartition 8,11,12,20 for rat, cel, mouse and human)

The 26 data sources can be classified into 6 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, mainly the proteins. Ontologies give more insight on the molecular functions performed (Gene Ontology [10] and SwissProt [14]), on the biological processes involved in (Gene Ontology [10] and Kegg [12]), on the cellular components in which the gene products are active (Gene Ontology [10] and EnsEMBL EST [13]), and on the active sites of the proteins (InterPro [11]).

Interaction data come from databases that collect protein pairs that interact either physically or genetically. (more details on the db we use)

The expression data sets we use are always coming from very large study in which a large panel of tissues are analyzed given thus an idea of the behaviour of the proteome.
For *homo sapiens*, the Su et al [24] and Son et al [25] studies provide gene expression for approximately 60% of the genome and for158 tissues. Su et al [24] have also produced a global analysis of the *mus musculus* proteome (35% of the genome and 90 tissues). Additionally, we are also

using data from the Hovatta et al [27] and Lindley et al [28] studies. These two data sets are complementary since Hovatta et al [27] have analyzed 38% of the genome for 87 tissues (weak genome coverage and large tissue panel) while Lindley et al [28] have analyzed 79% of the genome for 17 tissues (good genome coverage but small tissue panel).

For *rattus norvegicus*, the only dataset available is coming from the Walker et al paper [29]. The genome coverage is rather small (30%) but the tissue panel is large (122 tissues).
For *Caenorhabditis elegans*, we use the time series by Baugh et al [26], they
80% 123 tissues

Until now, we only have *cis*-regulatory information for *homo sapiens*. Using the TOUCAN toolbox [32] and the upstream sequence of the genes, the algorithm looks for putative motifs and modules (combination of motifs).

Seq
Ouzounis et al [30] and Adie et al [31] were able to discriminate the disease causing genes from the genome based on sequence features (length of the sequence, length of the UTRs, number of introns, length of the introns, for instance). Then they associate to every gene a probability of being a disease causing gene a-priori. Additionally, and for all organisms, an all-against-all Blast similarity search is performed using the NCBI Blast [33]. Oppositely to the previous a-priori data sources, It uses only the raw protein sequence.

Text [34]

Except for the regulatory information category, each organism has access to, at least, one data source per category.

Alternatively, one can use the Java client, which is also extended to include the other model organisms. This client includes more features like 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. The SOAP service itself is also available for integration in workflows, for instance, when using Taverna or Kepler.

(find a title)

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

Three use cases taken from the literature are preloaded into the application. The user can load the example by clicking on the corresponding button situated on top of the wizard. The training genes, the data sources, and the candidate genes are then loaded into the application. Then, the user can quickly go through the 4 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 [1]. 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 [4]. They have prioritized five susceptibility loci to reveal a molecular link between the two disorders. We picked up the susceptibility loci located on chromosome 11 for this example. It contains KCNJ5, an homologous gene of KCNJ11 that is known to contribute to the risk of Type II diabetes. We have built the last example right after Ebermann et al published their discovery of a novel Usher gene: DFNB31 encoding whirlin [8]. By using data 6 month prior to the discovery, we wanted to show that ENDEAVOUR was able to tag DFNB31 as the best candidate of the region that contains 32 genes.

Technical validation

As statistical validation, we performed a standard leave-one-out cross-validation using known gene sets (similarly to our previous work (Aerts et al [1])). We also produced the corresponding ROC curves and measured the performance by estimating the AUC. We applied this scheme to 29 gene sets containing the 625 genes involved in 29 diseases from the OMIM database, to 3 gene sets containing the 75 genes involved in 3 signalling pathways from the Gene Ontology database and, as a control, to 5 gene sets containing 100 genes randomly selected from the genome. The disease list contains homogeneous disorders (*e.g.*, Zellweger and Usher syndromes) as well as more complex disorders (*e.g.*, Alzheimer and diabetes). The resulting ROC curves are shown in X. The first observation is that the performance of the control validation is 49%, which is close to the random expectation of 50%. It means that when using randomly generated gene sets, we obtain random results (*i.e.,* our algorithm is not biased). By contrast, the performance of biologically meaningful sets is much higher (89% and 92% respectively for the 29 diseases and the 3 biological pathways). For a fair validation, we excluded the data sources that might contain explicitly the gene-disease or the gene-pathway association, which left us with 14 and 11 data sources respectively for the disease and the pathway validation.

The overall performance of the disease validation is 89%, but there are discrepancies between the different diseases. The performance is higher for the homogeneous disorders (Usher and Zellweger syndromes, for instance) than for the more complex disorders (diabetes and leukemia, for instance). It is no surprise since the homogeneous disorders are the perturbation of a single biological pathway while the more complex disorders can involve several perturbations. The per data source analysis reveals that the best performing data sources are x (x) and y (y) respectively for the disease and the pathway validations. The fact that the overall performance is higher than the best individual performance shows that our concept of data fusion is meaningful. Similar results were obtained with the different organisms we support (data not shown). Altogether, it indicates that the concept of similarity can be applied successfully.

Experimental validation

One major drawback of the leave-one-out cross-validation is the fact that we work with known data: for every disease gene, the link between the disease and the gene is described in the literature and sometimes we can also find evidences in the ontologies or in the interaction information. To assess the ability of ENDEAVOUR to solve real biological cases, we have described, in our previous publication, how ENDEAVOUR was used to discover a novel gene involved in DiGeorge-like syndrome. The DiGeorge syndrome is a rare congenital disease whose common symptoms are heart defects and characteristic facial features. In 1981, de la Chapelle *et al* shown that this anomaly is caused by a partial monosomy

of chromosome 22 and that the key gene probably lies in band 22q11 [9]. It was furthermore proved that TBX1 was the critical disease causing gene. For several patients with a DiGeorge-like syndrome, a region downstream from 22q11 was found to be deleted and thus related to the disorder. Using ENDEAVOUR, this region (containing 68 genes), was then prioritized with 4 different training sets – each one related to one phenotype of the DiGeorge syndrome. As an extra step, we merged together the 4 global rankings to obtain the final ranking in which YPEL1 ranked second among the 68 candidates. We further confirmed the association in zebrafish where YPEL1 knockdown embryos show malformation of the face and the heart.

Elbers *et al* have used ENDEAVOUR and other prioritization tools to define the best strategy to search for common obesity and Type 2 diabetes genes [4]. They suggest a list of genes indicated as potential candidates by at least 2 of the 6 tools. To our knowledge, none of them is yet validated. Tzouvelekis *et al* have used ENDEAVOUR to prioritize a list of genes differentially expressed in idiopathic pulmonary fibrosis [6]. They consistently find that among the top candidates, 5 and 7 genes are targets of, respectively, tumor necrosis factor (TNF) and transforming growth factor (TGF). Osoegawa *et al* have used ENDEAVOUR to propose novel genes associated with cleft lip and cleft palate phenotypes [5]. They analyse 83 syndromic cases and 104 nonsyndromic cases and conclude 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 have explored the proteome of the adipocyte, a central player in energy metabolism [7]. With 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

[...incomplete...]

Funding & acknowledgements

[...incomplete...]

References

1. Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., De Smet F., Tranchevent L-C., De Moor B., Marynen P., Hassan B., Carmeliet P. and Moreau Y. (2006) Gene prioritization through genomic data fusion. *Nature Biotechnology.* 2006 May;24(5):537-544.
2. Oti M. and Brunner HG. (2007) The modular nature of genetic diseases. *Clin Genet.* 2007 Jan; 71(1):1-11.
3. Zhu M. and Zhao S. (2007) Candidate gene identification approach: progress and challenges. *Int J Biol Sci.* 2007 Oct 25;3(7):420-7.
4. Elbers CC., Onland-Moret NC., Franke L., Niehoff AG., van der Schouw YT. and Wijmenga C. (2007) A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab.* 2007 Jan-Feb;18(1):19-26.
5. Osoegawa K., Vessere GM., Utami KH., Mansilla MA., Johnson MK., Riley BM., L'heureux J., Pfundt R., Staaf J., van der Vliet WA., Lidral AC., Schoenmakers EF., Borg A., Schutte BC., Lammer EJ., Murray JC. and De Jong PJ. (2007) Identification of novel candidate genes

associated with cleft lip and palate using array comparative genomic hybridization. *J Med Genet.* 2007 Sep 14.

6. Tzouvelekis A., Harokopos V., Paparountas T., Oikonomou N., Chatziioannou A., Vilaras G., Tsiambas E., Karameris A. and Bouros D., 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.* 2007 Dec 1;176(11):1108-19.
7. 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.* 2007 Jul;6(7): 1257-73.
8. Ebermann I., Scholl HP., Charbel Issa P., Becirovic E., Lamprecht J., Jurklies B., Millán JM., 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.* 2007 Apr;121(2):203-11.
9. De la Chapelle A., Herva R., Koivisto M. and Aula P. (1981) A deletion in chromosome 22 can cause DiGeorge syndrome. *Hum Genet.* 1981;57(3):253-6.
10. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.* 25: 25-29.
11. Mulder N.J. and al (2007) New developments in the InterPro database. *Nucleic Acids Res.* 35 (Database Issue):D224-D228.
12. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480-D484 (2008).
13. Hubbard T.J.P. and al. (2007) Ensembl 2007. *Nucleic Acids Res*. 2007 Vol. 35, Database issue:D610-D617.
14. 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(2003).
15. Bader GD., Donaldson I., Wolting C., Ouellette B.F., Pawson T. and Hogue C.W. (2001) BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res.* 2001 Jan 1;29(1): 242-5.
16. 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 Database issue:D449-51
17. 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.* Jan 1;34(Database issue):D535-9.
18. 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.* 2006 Nov 18;7:508.
19. 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.* 2007 Jan; 35(Database issue):D572-4.
20. Pagel P., Kovac S., Oesterheld M., Brauner B., Dunger-Kaltenbach I., Frishman G., Montrone C., Mark P., Stümpflen V., Mewes H.W., Ruepp A. and Frishman D. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics.* 2005; 21(6):832-834.
21. Peri S. and al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 2003 Oct;13(10):2363-71.
22. S. Kerrien S. and al. (2007) IntAct – Open Source Resource for Molecular Interaction Data. *Nucleic Acids Res.* 2007 Jan;35(Database issue):D561-5.

23. von Mering C., Jensen L.J., Kuhn M., Chaffron S., Doerks T., Krüger B., Snel B. and Bork P. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* 2007 Jan;35(Database issue):D358-62.
24. 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., Patapoutian A., Hampton G.M., Schultz P.G. and Hogenesch J.B. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA.* 2002 Apr 2;99(7):4465-70.
25. 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.* 2005 Mar;15(3):443-50.
26. 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.* 2005 Apr;132(8):1843-54.
27. 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. and Barlow C. (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature.* 2005 Dec 1;438(7068):662-6.
28. Lindsley R.C., Gill J.G., Kyba M., Murphy T.L., Murphy K.M. (2006) Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. *Development.* 2006 Oct; 133(19):3787-96.
29. Walker J.R., Su A.I., Self D.W., Hogenesch J.B., Lapp H., Maier R., Hoyer D. and Bilbe G. Applications of a rat multiple tissue gene expression dataset. *Genome Res.* 2004 Apr;14(4): 742-9.
30. López-Bigas N. and Ouzounis C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* 2004 Jun 4;32(10):3108-14.
31. Adie E.A., Adams R.R., Evans K.L., Porteous D.J. and Pickard B.S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics.* 2005 Mar 14;6:55.
32. 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.* 2005 Jul 1;33(Web Server issue):W393-6.
33. Ye J., McGinnis S. and Madden T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W6-9.
34. 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 Biology*, 5 (6):R43:1-12.