Gene prioritization through genomic data fusion

Stein Aerts[1,4,5], Diether Lambrechts[2,5], Sunit Maity[2,5], Peter Van Loo[3–5], Bert Coessens[4,5], Frederik De Smet[2], Leon-Charles Tranchevent[4], Bart De Moor[4], Peter Marynen[3], Bassem Hassan[1], Peter Carmeliet[2] & Yves Moreau[4]

The identification of genes involved in health and disease remains a challenge. We describe a bioinformatics approach, together with a freely accessible, interactive and flexible software termed Endeavour, to prioritize candidate genes underlying biological processes or diseases, based on their similarity to known genes involved in these phenomena. Unlike previous approaches, ours generates distinct prioritizations for multiple heterogeneous data sources, which are then integrated, or fused, into a global ranking using order statistics. In addition, it offers the flexibility of including additional data sources. Validation of our approach revealed it was able to efficiently prioritize 627 genes in disease data sets and 76 genes in biological pathway sets, identify candidates of 16 mono- or polygenic diseases, and discover regulatory genes of myeloid differentiation. Furthermore, the approach identified a novel gene involved in craniofacial development from a 2-Mb chromosomal region, deleted in some patients with DiGeorge-like birth defects. The approach described here offers an alternative integrative method for gene discovery.

With the advent of 'omics, identifying key candidates among the thousands of genes in a genome that play a role in a disease phenotype or a complex biological process has paradoxically become one of the main hurdles in the field. Indeed, contrary to some early concerns in the community that a lack of sufficient global data would still be a limiting factor[1], it is precisely the opposite, a bounty of information that now poses a challenge to scientists. This has translated into a need for sophisticated tools to mine, integrate and prioritize massive amounts of information[2,3].

Several gene prioritization methods have been developed[4–10]. Most of them determine, either directly or indirectly, the similarity between candidate genes and genes known to play a role in defined biological processes or diseases. These methods offer several advantages but also pose a number of challenges. Indeed, even though multiple data sources are available, such as Gene Ontology (GO) annotations[4–6,10], protein domain databases[6,10], the published literature[5,7], gene expression data[5,7,10] and sequence information[8–10], most of the available programs access only one or two of these databases, which each have their limitations. For instance, functional data sources (GO and literature) are incompletely annotated and biased toward better-studied genes[8], whereas sequence databases have thus far been used only to produce general disease probabilities[8,9]. Some of the existing tools access more than two databases, but do not provide an overall ranking that integrates the separate searches[5,10]. Several tools rank disease genes but only one of them prioritizes genes involved in biological pathways[10], and none offers the combination of both. Thus far, only two prioritization tools[5,10] are publicly available. Thus, there is still a need for improvement of gene prioritization.

Here, we report the development and characterization of a new gene prioritization method, and offer its freely accessible, interactive and flexible software[1]. Compared to existing methods, ours provides additional opportunities for candidate gene prioritization: it accesses substantially more data sources and offers the flexibility to include new databases; it provides the user control over the selection of training genes and thereby takes advantage of the expertise of the user; it prioritizes both known and unknown genes, ranks genes involved in human diseases and biological processes, and it uses rigorous statistical methods to fuse all the individual rankings into an overall rank and probability.

RESULTS

Principles of prioritization used by Endeavour

Genes involved in the same disease or pathway often share annotations and other characteristics in multiple databases. Indeed, genes involved in the same disease share up to 80% of their annotations in the GO and InterPro databases[6], whereas genes involved in a similar biological pathway often share a high degree of sequence similarity with other pathway members[11]. It is therefore reasonable to assume that this similarity among genes is not restricted to their annotation or sequence alone, but is also true for their regulation and expression. We reasoned that a bioinformatics framework capable of comparing and integrating all available gene characteristics might be a powerful tool to rank unknown candidate 'test' genes according to their similarity with known 'training' genes, and based on this notion, we developed Endeavour. Prioritization of genes using this algorithm involves three steps (**Fig. 1**). To validate its performance, we used several complementary strategies discussed below.

[1]Laboratory of Neurogenetics, Department of Human Genetics, [2]The Center for Transgene Technology and Gene Therapy, [3]Human Genome Laboratory, Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology (VIB), University of Leuven, Herestraat 49, bus 602, 3000 Leuven, Belgium. [4]Bioinformatics Group, Department of Electrical Engineering (ESAT-SCD), University of Leuven, Belgium. [5]These authors contributed equally to this work. Correspondence should be addressed to S.A. (stein.aerts@med.kuleuven.be).

Published online 5 May 2006; doi:10.1038/nbt1203

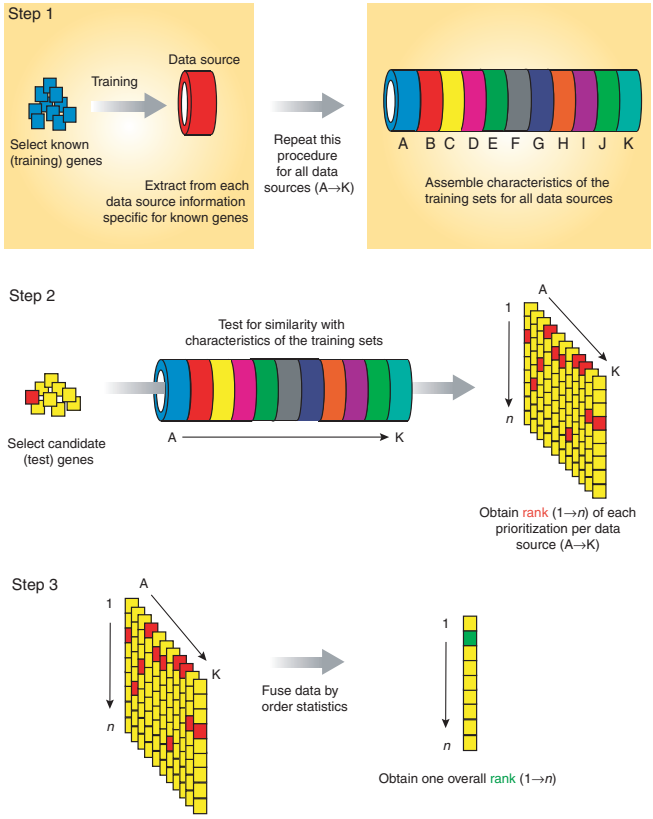

Figure 1 Concept of prioritization by Endeavour. Candidate test genes are ranked with Endeavour based on their similarity with a set of known training genes in a three-step analysis. In the first step (upper panel), information about a disease or pathway is gathered from a set of known training genes by consulting various data sources. Training genes can be loaded automatically (based on a Gene Ontology term, a KEGG pathway ID or an OMIM disease name) or manually. The latter allows the incorporation of expert knowledge. The following data sources are used: A, literature (abstracts in EntrezGene); B, functional annotation (Gene Ontology); C, microarray expression (Atlas gene expression); D, EST expression (EST data from Ensembl); E, protein domains (InterPro); F, protein-protein interactions (Biomolecular Interaction Network Database or BIND); G, pathway membership (Kyoto Encyclopedia of Genes and Genomes or KEGG); H, *cis*-regulatory modules (TOUCAN); I, transcriptional motifs (TRANSFAC); J, sequence similarity (BLAST); K, additional data sources, which can be added (e.g., disease probabilities). In the second step (middle panel), a set of test genes is loaded (again, either manually or automatically by querying for a chromosomal region or for markers). These test genes are then ranked based on their similarity with the training properties obtained in the first step, which results in one prioritized list for each data source. Vector-based data are scored by the Pearson correlation between a test profile and the training average, whereas attribute-based data are scored by a Fisher's omnibus analysis on statistically overrepresented training attributes. Finally, in the third step (lower panel), Endeavour fuses each of these rankings from the separate data sources into a single ranking and provides an overall prioritization for each test gene. As such, Endeavour prioritizes genes through genomic data fusion—a term, borrowed from engineering to reflect the merging of distinct heterogeneous data sources, even when they differ in their conceptual, contextual and typographical representations.

Validation of Endeavour when accessing individual data sources

For each individual data source, we assessed whether our approach is capable of prioritizing genes known to be involved in specific diseases or receptor signaling pathways. To this end, we performed a large-scale leave-one-out cross-validation. In each validation run, one gene, termed the 'defector' gene, was deleted from a set of training genes and added to 99 randomly selected test genes. The software then determined the ranking of this defector gene for every data source separately. We used 627 training genes, ordered in 29 training sets of particular diseases automatically selected from the Online Mendelian Inheritance In Man (OMIM) database (see **Supplementary Notes** online for selection procedure). For pathway genes, we compiled three sets of training genes involved in the WNT (43 genes), NOTCH (18 genes) and epidermal growth factor (15 genes) pathways. As a negative control for training genes, we assembled 10 sets of 20 randomly selected genes.

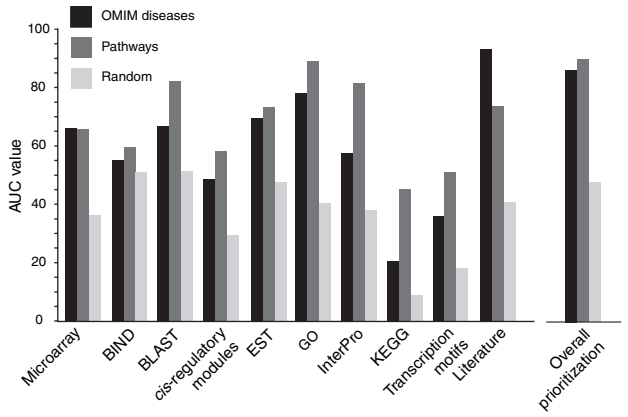

Figure 2 Cross-validation results. The AUC values obtained for all individual data sources are shown for disease prioritizations (black), pathway prioritizations (dark gray) and random prioritizations (light gray). The AUC values from the overall prioritization obtained after fusing all individual prioritizations are also shown.

Thus, a total of 903 prioritizations (627 for the disease genes, 76 for the pathway genes and 200 for the random sets) were performed for each data source. From these, we calculated sensitivity and specificity values. Sensitivity refers to the frequency (% of all prioritizations) of defector genes that are ranked above a particular threshold position. Specificity refers to the percentage of genes ranked below this threshold. For instance, a sensitivity/specificity value of 70/90 would indicate that the correct disease gene was ranked among the best-scoring 10% of genes in 70% of the prioritizations. To allow comparison between data sources we plotted rank receiver operating characteristic (ROC) curves, from which sensitivity/specificity values can be easily deduced. The area under this curve (AUC) is a standard measure of the performance of this algorithm. For instance, an AUC-value of 100% indicates that every defector gene ranked first, whereas a value of 50% means that the defector genes ranked randomly.

For every single data source, Endeavour reached a higher AUC score for disease and pathway genes than for randomly selected genes, indicating that it was sensitive and specific in ranking the defector gene, regardless of the type of data source consulted (**Fig. 2**). Not surprisingly, the data sources differed in their usefulness and suitability to rank genes (**Supplementary Notes**).

Overall prioritization by fusing multiple data sources

Although in most cases the defector gene ranked high in the prioritization list, this was not always the case (**Supplementary Fig. 1** online).

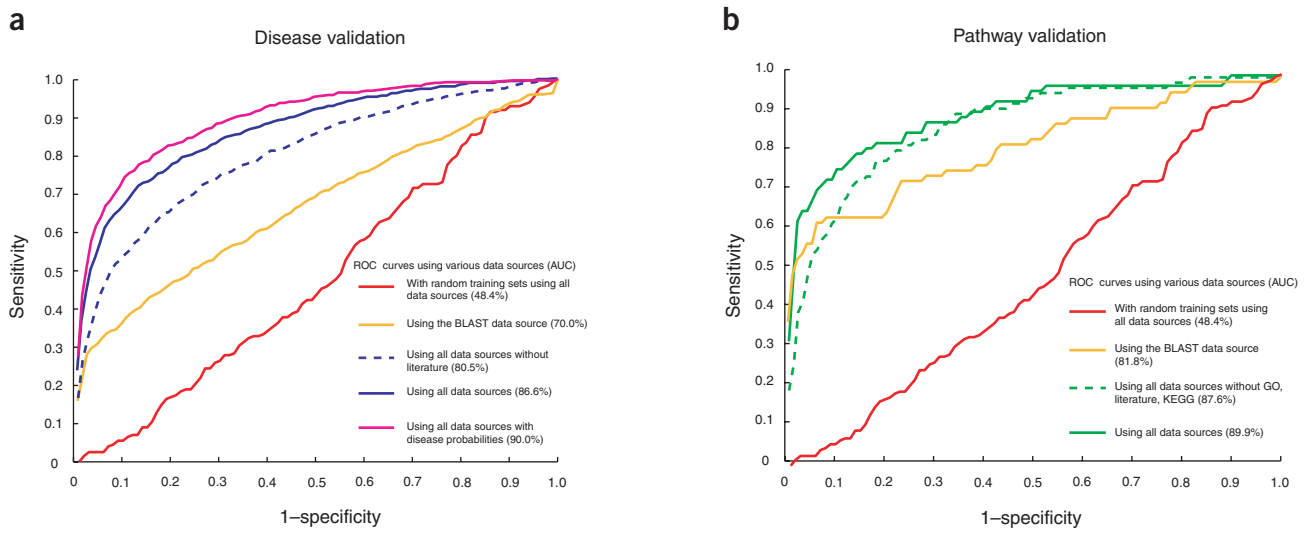

Figure 3 Cross-validation results. (**a**) Rank ROC curves obtained for the disease validation. (**b**) Rank ROC curves obtained for the pathway validation. In both figures, the control ROC curve (red line) was obtained after prioritization with randomly constructed training sets and by using all data sources. For all other ROC curves, disease or pathway-specific training sets were generated. The data sources used to construct every ROC curve are indicated on the figure.

To minimize this variability and to increase the performance of ranking, we integrated all individual prioritizations into a single overall rank by implementing an algorithm based on order statistics. With this algorithm, the probability of finding a gene at all the observed positions is calculated and a single overall rank is obtained by ranking genes according to these probabilities. To evaluate the performance of this overall ranking, we calculated its AUC values, as described above for the individual data sources. The AUC scores were 86.6% and 89.9% for disease and pathway genes compared to 48.4% for randomly selected genes (**Fig. 3a**,**b**). The correct pathway gene ranked among the top 50% of test genes in 95% of the cases, or among the top 10% in 74% of the cases. The variability of the overall prioritization was substantially smaller than that of individual data sources (**Supplementary Fig. 1**), and each of the data sources contributed to the overall ranking (**Supplementary Fig. 2** online). Our validation experiment thus results in biologically meaningful prioritizations.

Almost every data source but especially functionally annotated databases are incompletely annotated. For instance, only 63% of the genes are currently annotated in the GO database. Consequently, existing methods using these data sources introduce an undesired bias toward better-studied genes. Our approach should suffer less from these shortcomings as it also uses sequence-based sources containing information about known and unknown genes. In support of this, we found that the overall ranking of defector genes was not substantially influenced by the number of data sources if at least three sources with data annotations were available (**Supplementary Fig. 3a** online). In fact, even unknown genes lacking a

Table 1 Prioritizations of recently identified monogenic disease genes

				Rank position using the indicated data sources	
Disease	Gene	Ensembl ID	Publication date	All	Literature
Arrhythmia	*CACNA1C*	ENSG00000151067	October 2004 (ref. 34)	4	3
Congenital heart disease	*CRELD1*	ENSG00000163703	April 2003 (ref. 35)	3	1
Cardiomyopathy 1	*CAV3*	ENSG00000182533	January 2004 (ref. 36)	2	1
Parkinson disease	*LRRK2*	ENSG00000188906	November 2004 (ref. 37)	50	*
Charcot-Marie-Tooth disease	*DNM2*	ENSG00000079805	March 2005 (ref. 38)	14	100
Amyotrophic lateral sclerosis	*DCTN1*	ENSG00000135406	August 2004 (ref. 39)	27	97
Klippel-Trenaunay disease	*AGGF1* (also known as *VG5Q*)	ENSG00000164252	February 2004 (ref. 40)	3	39
Cardiomyopathy 2	*ABCC9*	ENSG00000069431	April 2004 (ref. 41)	1	51
Distal hereditary motor neuropathy	*BSCL2*	ENSG00000168000	March 2004 (ref. 42)	15	62
Cornelia de Lange syndrome	*NIPBL*	ENSG00000164190	June 2004 (refs. 43,44)	9	75
			Average rank	13 ± 5	48 ± 13

For all genes, a mutation was inherited in a mendelian fashion (or was shown to cause the disease phenotype). The name of the disease and disease-causing gene, the Ensembl ID and the publication date of the article reporting the gene mutation (month-year) are shown, together with the rank (out of 200 test genes) at which they were prioritized by Endeavour, using all data sources or using the pre-publication date literature source alone. The average rank (mean ± s.e.m.) for each prioritization is indicated. For *LRRK2*, no literature information was available. This has been indicated in the table by an asterisk (*).

Table 2 Prioritizations of recently identified polygenic disease genes

Disease	Gene	Ensembl ID	Publication date	Rank
Atherosclerosis 1	*TNFSF4*	ENSG00000117586	April 2005 (ref. 45)	54
Crohn disease	*SLC22A4, SLC22A5*	ENSG00000197208	May 2004 (ref. 46)	71
Parkinson disease	*GBA*	ENSG00000188906	November 2004 (47)	23
Rheumatoid arthritis	*PTPN22*	ENSG00000134242	August 2004 (ref. 48)	11
Atherosclerosis 2	*ALOX5AP*	ENSG00000132965	February 2004 (ref. 49)	29
Alzheimer disease	*UBQLN1*	ENSG00000135018	March 2005 (ref. 50)	54
			Average rank	40 ± 10

The nature of the genetic variation in these genes was in each case a polymorphism, which typically was inherited as a risk factor for the respective disease. The name of the complex disease in which these genes were identified, their gene name, Ensembl ID and the publication date when the disease gene was reported as a susceptibility gene are given, together with the rank (out of 200 test genes) at which they have been prioritized by all data sources with rolled-back literature. The relative contribution of these genetic variations as risk factors for disease susceptibility will become clearer once replication studies are performed. The average rank (mean ± s.e.m.) for each prioritization is indicated.

HUGO name and with very little information available could be ranked highly (**Supplementary Fig. 3b**). Thus, our method takes into account data sources with relevant information, while disregarding noninformative ones. This may be particularly advantageous for the prioritization of disease genes, as unknown genes are not readily considered as disease candidates when selected manually.

Endeavour does not rely on literature-derived data alone

For each OMIM gene used in the disease validation, a mutation causing the disease had previously been reported in a landmark study. Because the inclusion of these publications may artificially increase the relative contribution of the literature data source in the overall performance of this algorithm, we excluded, as a test, the entire literature database from the disease validation protocol. For the same reason, the GO, KEGG and literature data sources were excluded from the pathway validation. Even under such unrealistic conditions where entire data sources were not used, the overall performance of the algorithm was only negligibly affected: the performance dropped by only 6.1% for disease genes (from 86.6% to 80.5%; **Fig. 3a**) and by only 2.3% for pathway genes (from 89.9% to 87.6%; **Fig. 3b**). Thus, the diversity of data sources used in our approach enables meaningful prioritizations, even without the use of literature information.

Clearly, this caution is only of importance in the context of a validation. In a more realistic situation, when the precise function of a disease gene is not known yet, the literature could still provide valuable indirect information about other properties of a gene. In a study of ten monogenic diseases (see below), we mimicked this situation by using only 'rolled-back' literature information, available one year before the landmark publication. Even then Endeavour provided a high rank for three genes (position 1, 1 and 3 out of 200 test genes, **Table 1**), illustrating that the literature contributes to the prioritization of yet undiscovered disease genes. For the seven other genes, use of the literature as the only data source was not very efficient, but inclusion of all the other data sources yielded a high rank (**Table 1**). Overall, even though the literature may provide valuable information, our method does not rely on literature as the only critical data source. But also, its performance is not restricted by the lack of available literature data, because of its ability to access and integrate multiple other data sources.

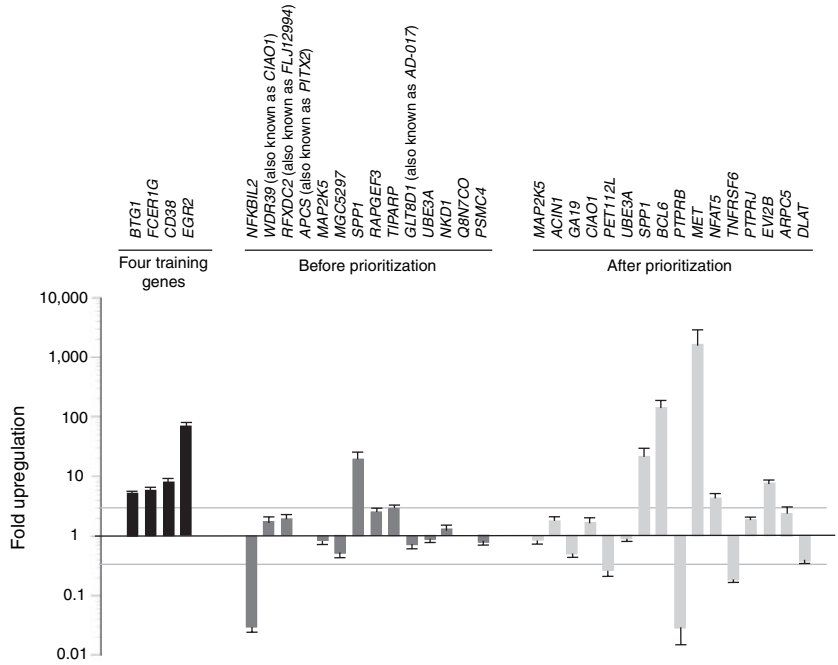

Figure 4 *In vitro* functional validation of Endeavour. Results of real-time quantitative PCR measurements in differentiated versus undifferentiated HL-60 cells. Expression profiles of 4 out of 18 training genes (left), which were tested as a positive control, and 20 target genes predicted by the *cis*-regulatory module model (center) are shown. Expression levels of *SPP1* and *NGKBIL2* differed more than threefold between differentiated and undifferentiated cells; expression levels for six genes could not be measured. The expression profiles of the 20 highest-ranking target genes after prioritization by Endeavour (right) are also shown. Expression levels of eight genes (*SPP1, BCL6, PTPRB, MET, TNFRSF6, NFAT5, PET112L* and *EVI2B*) differed more than threefold between differentiated and undifferentiated cells; four genes could not be measured. The fold difference is depicted on a logarithmic scale; error bars represent the s.e.m. The line indicates the threshold (threefold up- or downregulation).

Use of disease-specific data sources

An important asset of Endeavour is that its framework was designed to allow the inclusion of additional data sources, such as disease-related features, in the prioritization strategy. We illustrate this for the prioritization of disease genes. On the basis of a number of selection criteria (e.g., protein length, phylogenetic conservation), Lopez-Bigas and Adie determined for every gene a 'general' disease probability, or its probability as a disease candidate gene[8,9]. When integrating the Lopez-Bigas or Adie criteria in Endeavour as an additional data source, we found that its performance improved further (AUC scores increased by up to 5% regardless of the inclusion of literature sources). Likewise, microarray data specific for the process or disease under study can be included. Our approach thus allows the user to add, in a flexible and modular manner, additional data sources, such as appropriate disease-specific data sources, to enhance its overall performance.

Prioritization of genes causing monogenic diseases

In the large-scale validation, 627 genes were automatically selected from the OMIM database, without taking their mono- or polygenic nature into account. We therefore assessed whether our approach could be used to prioritize genes that cause monogenic diseases. As experimentalists often prefer to select their own sets of training genes, instead of relying on automatically derived genes or characteristics, we selected ten monogenic diseases and constructed sets of training genes together with a biological expert (**Table 1** and **Supplementary Table 1**). To simulate the real life situation, we deliberately chose recently identified disease-causing genes, and used rolled-back literature together with all other data sources. The set of test genes included the gene causing the monogenic disease, and 199 genes flanking its immediate chromosomal surroundings. The algorithm gave the ten monogenic disease–causing genes an average rank of 13 ± 5 out of 200 test genes (**Table 1**). When using a training set not related to the disease under study to prioritize the test sets as a negative control, the disease genes ranked randomly (position 96 on average). As a further validation the algorithm was applied to a very large set of test genes (that is, all 1,048 genes from chromosome 3; **Supplementary Notes** and **Supplementary Table 2** online).

This pseudo-prospective analysis, using rolled-back literature, reveals that expert-based construction of training sets may lead to high discovery rates when hunting for monogenic disease genes in both small and large test sets.

Prioritization of genes underlying polygenic diseases

In many cases, human disease is not monogenic, but polygenic in nature. We therefore prioritized six genes, recently identified as polygenic disease genes, together with 199 chromosomal flanking genes (**Table 2**). The sets of training genes used for these prioritizations are explained in **Supplementary Table 1**. On average, the susceptibility genes ranked at position 40 ± 10, when using the rolled-back literature together with all the other data sources. As expected, the prioritization of polygenic disease candidate genes is a greater challenge than ranking monogenic disease genes. Nonetheless, the ranking was still specific, as the susceptibility genes ranked at position 96 ± 10, when training sets for these disorders were randomly assigned to other test sets as a negative control. Thus, although the performance is lower than for monogenic diseases (as anticipated), susceptibility genes to polygenic diseases can be enriched by Endeavour's prioritization.

Prioritization of regulatory pathway genes

To analyze whether Endeavour could also rank genes involved in a particular biological process, we combined computation with functional validation *in vitro*. First, using the previously characterized ModuleSearcher

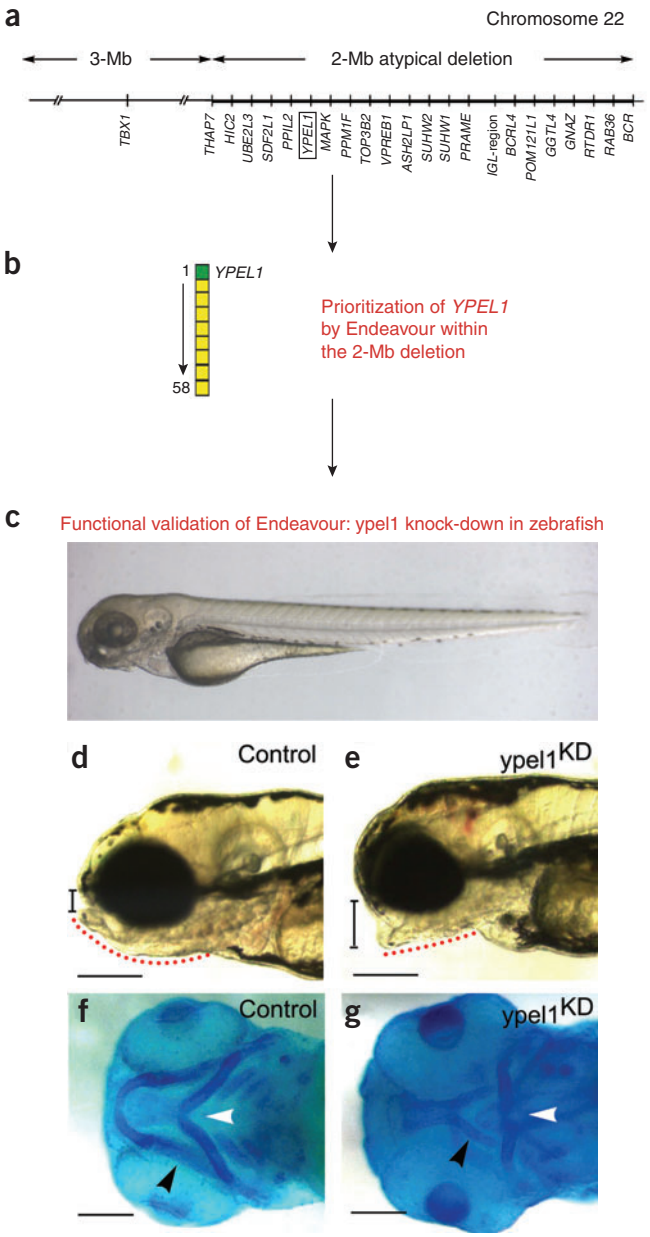

Figure 5 Functional validation of Endeavour in zebrafish. (**a**) Part of chromosome 22, illustrating the hemizygous 3-Mb region deleted in many DGS patients and the atypical 2-Mb region, which is deleted in some (atypical) DGS patients. For clarity, only some of the 58 Ensembl-annotated genes within the 2-Mb region, and only *TBX1* in the 3-Mb deleted region, are shown. It remains unknown whether any of the genes in the 2-Mb region play a role in pharyngeal arch development defects seen in DGS. (**b**) *YPEL1* was prioritized among the 58 genes of the 2-Mb deleted region by Endeavour as the most likely candidate involved in pharyngeal arch development. (**c**) Photo of a zebrafish, which has been used as a suitable model to study the role of *YPEL1* in pharyngeal arch development. (**d**,**e**) Lateral view of the head in live embryos at 4 d after fertilization. The lower jaw is clearly visible in the control, whereas ypel1^{KD} embryos show an underdeveloped lower jaw (mandibular arch; indicated by the red dotted line) and open mouth (indicated by the vertical line). (**f**,**g**) Ventral view of the pharyngeal arch cartilage using alcian blue stain at 3 d after fertilization. Black arrow depicts the mandibular arch; white arrow depicts hyoid arch. In ypel1^{KD} embryos, the jaw arches were severely malformed with the mandibular arch often reduced in size. The pharyngeal arch cartilage also showed reduced or no staining.

Table 3 Prioritization of *YPEL1* by Endeavour

Training sets used to prioritize *TBX1* or *YPEL1*	Rank assigned to *YPEL1*	Rank assigned to *TBX1*
DGS-related		
DGS (14)	1*	1*
Cardiovascular birth defects (14)	3*	1*
Cleft palate birth defects (9)	2*	1*
Neural crest genes (14)	1*	2*
Average rank	1.75 ± 0.48	1.25 ± 0.25
DGS-unrelated		
Atherosclerosis (24)	12	24
Parkinson disease (9)	31	15
Distal hereditary motoneuropathy (8)	13	41
Charcot-Marie-Tooth disease (17)	9	16
Alzheimer's disease (5)	21	14
Rheumatoid arthritis (8)	20	7
Inflammatory bowel disease (7)	7	24
Average rank	16 ± 3	20 ± 4

The set of test genes contained the 58 genes present in the 2-Mb atypical deletion region on chromosome 22q11 (middle column) or, in addition, the *TBX1* gene (right column). These test genes were prioritized by Endeavour for their similarity to the indicated set of training genes, which were related or unrelated to DGS. As shown, *TBX1* and *YPEL1* ranked among the first three test genes, indicating their high degree of similarity with the set of training genes (*, probability of $P < 0.05$ that the test and training genes had a similar profile). The number of training genes is indicated between brackets.

algorithm within TOUCAN[12,13], we predicted a *cis*-regulatory module (CRM) in the regulatory regions of 18 genes, known to be upregulated during myeloid differentiation[14]. We then selected 100 putative target genes containing this CRM from the genome, and ordered them according to their CRM score (see **Supplementary Notes**). These 100 genes were then prioritized with the algorithm, using the 18 genes involved in myelopoiesis as a training set. To investigate whether it enriched the number of true-positive target genes involved in myeloid differentiation, we induced differentiation of HL-60 cells *in vitro* and analyzed which of the 20 best ranking genes, before and after prioritization by Endeavour, were more than threefold up- or downregulated. Before prioritization, the expression of two genes (**Fig. 4**) was differentially regulated, whereas after prioritization up to eight genes were differentially regulated ($P < 0.05$; **Fig. 4**). Importantly, several of these differentially regulated genes are implicated in myeloid function: *SPP1*, *BCL6* and *MET* are known to be involved in myeloid differentiation[15–17], whereas *FRSF6*, better known as the *FAS* inducer of apoptosis, is a suppressor of macrophage activation[18]. The possible involvement of *PTPRB*, *NFAT5*, *PET112L* and *EVI2B* in myeloid differentiation was, however, unknown. Our prioritization protocol can thus be used for gene discovery as well.

Functional validation of Endeavour in zebrafish

As a final and most stringent test, we validated our approach in an animal model *in vivo*. The DiGeorge syndrome (DGS) is a common congenital disorder, in which craniofacial dysmorphism and other defects result from abnormal development of the pharyngeal arches[19,20]. Many DGS patients typically have a 3-Mb hemizygous deletion in chromosome 22 (*del22q11*)[19,20]. Genetic studies in mice and zebrafish have established *Tbx1* as a key DGS disease candidate gene in this region[21–24] (**Fig. 5a**). In atypical DGS cases, a 2-Mb region, downstream of *del22q11* is deleted[25], but it remains unknown which of the 58 Ensembl-annotated genes in this region plays a role in pharyngeal arch development. In this experiment, we first assessed whether the algorithm would prioritize any of these genes as a possible regulator of pharyngeal arch development, and then analyzed whether this gene indeed affected this process *in vivo*.

We first tested, as a positive control, whether Endeavour would identify *TBX1* as a DGS candidate when added to the list of 58 test genes. To avoid possible selection bias due to an overly restricted choice of training genes, we used various training sets according to their relationship with DGS, cardiovascular or cleft palate birth defects (typical DGS symptoms), or neural crest biology (neural crest cell anomalies cause DGS-like symptoms; **Supplementary Notes**). When using these training sets, *TBX1* ranked first or second (**Table 3**). This prioritization was specific, as *TBX1* was not identified as a DGS candidate gene when using training genes unrelated to DGS. We then used our approach to prioritize the 58 genes of the 2-Mb deleted region. When using various sets of DGS-related training genes, the top-ranking gene was always *YPEL1* (**Table 3** and **Fig. 5b**). Similar to the *TBX1* simulation, use of a set of training genes, unrelated to DGS, confirmed that the prioritization was specific for DGS.

To assess the functional role of *YPEL1 in vivo*, we used the zebrafish model, which has been previously used as a suitable model to study pharyngeal arch development[26] (**Fig. 5c**). Ypel1 protein levels in zebrafish embryos were knocked down using a set of antisense morpholino oligonucleotides (morpholinos), each targeting different sequences of the *ypel1* transcript and dose-dependently and specifically inhibiting *ypel1* translation (not shown). The role of *ypel1* in pharyngeal arch morphogenesis was evaluated by phenotyping the development of its derivatives, that is, the jaws and other skeletal structures of the skull[27]. Ypel1 knockdown (ypel1^{KD}) embryos displayed various craniofacial defects. In particular, they exhibited an underdeveloped jaw, with the most severely affected embryos displaying an open-mouth phenotype suggestive of craniofacial dysmorphism (**Fig. 5d**,**e**). Ypel1^{KD} embryos also displayed defects in pharyngeal arch cartilage formation, ranging from an overall disorganization to a complete loss of the jaw and pharyngeal arch cartilage. In some ypel1^{KD} embryos, the mandibular arch was strongly reduced in size. Occasionally, no staining of cartilage could be detected at all (**Fig. 5f**,**g**). Ypel1^{KD} embryos exhibited additional pharyngeal arch defects, which will be described in more detail elsewhere.

In summary, our method identified *YPEL1* as a candidate DGS gene and *in vivo* experiments confirmed its role in pharyngeal arch development. These data raise the intriguing question whether *YPEL1* might be a novel disease candidate gene of atypical DGS in humans.

DISCUSSION

The number of publicly available databases containing information about human genes and proteins continues to grow. Here, we developed a method to integrate all this information and prioritize any set of genes based on their similarity to a set of reference genes. Such a prioritization is not only useful for gene hunting in human diseases, but also for identifying members of biological processes.

Our approach is useful in several respects. First, it uses genes to retrieve information about a disease or biological pathway, instead of disease characteristics. Existing methods that use disease characteristics can only retrieve information from databases that use the same disease vocabulary[4,5,7]. By using genes, Endeavour accesses not only these vocabulary-based data sources, but also other data sources, storing

information about a gene (e.g., derived from a microarray experiment) or a gene sequence (e.g., BLAST sequence similarity). Moreover, by using genes, the method is also suitable for gene prioritization in biological processes as well.

Second, compared to existing methods, which access only one or two data sources[4–7], our method accesses many more data sources (currently up to 12). Importantly, consultation of each of the individual sources by Endeavour generates biologically relevant prioritizations. We developed an algorithm based on order statistics to fuse all these separate prioritizations into a single overall rank. This algorithm is able to handle genes with missing values, thereby minimizing the bias for known or well-characterized genes. This bias will decrease even further in the future, when new and better high-throughput data become available, and when the genome annotation and curation processes reach their finalization.

Third, the algorithm is publicly available as a software tool, built by bioinformaticians, but designed for experimentalists, helping them to focus readily on key biological questions. The only other available prioritization tool for diseases, G2D, uses GO and literature data sources and is therefore restricted in making predictions about annotated or known genes[5].

Fourth, the approach gives the user maximal control over the set of training and test genes. Biologists prefer the flexibility of interactively selecting their own set of genes over an automatic and noninteractive data-mining selection procedure.

We validated the method extensively, in a large-scale validation study of 703 disease and pathway genes, and in a number of case-specific analyses. The validation results were remarkably good: on average, the correct gene was ranked 10th out of 100 test genes—for monogenic diseases, the performance was even better. The algorithm was capable of prioritizing large test sets (up to 1,000 genes)—the upgrade of Endeavour into a package capable of prioritizing the entire genome would be an interesting perspective for the future. Functional validation studies *in vitro* further demonstrated that the method worked equally well for prioritization of pathway genes. Furthermore, *in vivo* studies in zebrafish revealed that *YPEL1*, a gene prioritized by Endeavour in a 2-Mb chromosomal region deleted in patients with craniofacial defects, indeed regulates morphogenesis of the pharyngeal arches and their craniofacial-derivative structures.

Lastly, the Endeavour software design is modular and allows the inclusion of publicly available or proprietary data sources (e.g., disease-specific microarray experiments). We have illustrated and validated this possibility by including the general disease probability criteria of Lopez-Bigas[9] and Adie[8].

In summary, we present a computational method for fast and interactive gene prioritization that fuses genomic data regardless of its origin.

METHODS

Data sources. A more detailed description of the data sources is available as **Supplementary Methods** online. Briefly, for information retrieved from attribute-based data sources (that is, Gene Ontology, EST expression, InterPro and KEGG), the algorithm uses a binomial statistic to select those attributes that are statistically overrepresented among the training genes, relative to their genome-wide occurrence. Each overrepresented attribute receives a *P*-value p_i that is corrected for multiple testing. For information retrieved from vector-based data sources (that is, literature, microarray expression data or *cis*-regulatory motif predictions), the algorithm constructs an average vector profile of the training set. The literature profile is based on indexed abstracts and contains inverse document frequencies for each term of a GO-based vocabulary[28]; the expression profile contains expression ratios; the motif profile contains scores of TRANSFAC position weight matrices, obtained by scanning promoter sequences of the training genes that are conserved with their respective mouse orthologous sequences. To rank a set of test genes, attribute-based data are scored by Fisher's omnibus meta-analysis (Σ-2logp_i), generating a new *P*-value from a χ^2-distribution. Vector-based data are scored by Pearson correlation between the test vector and the training average. The data in the BLAST, BIND and *cis*-regulatory module (CRM) databases are neither vector- nor attribute-based. For BLAST, the similarity score between a test gene and the training set is the lowest *e*-value obtained from a BLAST against an *ad hoc* indexed database consisting of the protein sequences of the training genes. For BIND (Biomolecular Interaction Network Database)[29], the similarity score is calculated as the overlap between all protein-protein interaction partners of the training set and those of the test gene. Lastly, for CRM data, the best combination of five clustered transcription factor binding sites—in all human-mouse conserved noncoding sequences (up to 10 kb upstream of transcription start site) of the training genes—is determined using a genetic algorithm[12,30]. The similarity of this trained model to a test gene is determined by scoring this motif combination on the conserved noncoding sequences of the test gene.

Order statistics. The rankings from the separate data sources are combined using order statistics. A *Q* statistic is calculated from all rank ratios using the joint cumulative distribution of an *N*-dimensional order statistic as previously done by Stuart *et al.*[31]

$$Q(r_1,r_2,\ldots,r_N) = N!\int_0^{r_1}\int_{s_1}^{r_2}\ldots\int_{s_{N-1}}^{r_N} ds_N ds_{N-1}\ldots ds_1 \qquad (1)$$

They propose the following recursive formula to compute the above integral:

$$Q(r_1,r_2,\ldots,r_N) = N!\sum_{i=1}^{N}(r_{N-i+1}-r_{N-i})Q(r_1,r_2,\ldots,r_{N-i},r_{N-i+2},\ldots,r_N) \qquad (2)$$

where r_i is the rank ratio for data source *i*, *N* is the number of data sources used, and $r_0 = 0$. However, two problems arose when we tried to use this formula. First, we noticed that this formula is highly inefficient for moderate values of *N*, and even intractable for $N > 12$ because its complexity is $O(N!)$. We therefore implemented a much faster alternative formula with complexity $O(N^2)$:

$$V_k = \sum_{i=1}^{k}(-1)^{i-1}\frac{V_{k-i}}{i!}r_{N-k+1}^{i} \qquad (3)$$

with $Q(r_1,r_2,\ldots,r_N) = N!V_N$, $V_0 = 1$, and r_i is the rank ratio for data source *i*.

Second, we noticed that the *Q* statistics calculated by (1) are not uniformly distributed under the null hypothesis and can thus not directly be used as *P*-values. Therefore, we fitted a distribution for every possible number of ranks and used this distribution to calculate an approximate *P*-value. We found that the *Q* statistics for $N \leq 5$ randomly and uniformly drawn rank-ratios are approximately distributed according to a beta distribution. For $N > 5$ the distributions can be modeled by a gamma distribution. The cumulative distribution function of these distributions provides us with a *P*-value for every *Q* statistic from the order statistics. Next to the original *N* rankings, we now have an $(N + 1)^{th}$ that is the combined rank of all separate ranks.

Cell culture, RNA isolation and RT-PCR. HL-60 cells were grown in RPMI 1640, supplemented with 10% FCS. Differentiation was induced by 10 nM phorbol 12-myristate 13-acetate (PMA), when cells were grown to a density of 7×10^5/ml. Before induction and 24 h after induction, cells were harvested by centrifugation and RNA was isolated using the trizol reagent (Invitrogen), and subsequently treated with Turbo DNA-free DNase (Ambion). First-strand cDNA was synthesized using Superscript II reverse transcriptase (Invitrogen). Real-time quantitative PCR was performed using the qPCR core kit for SYBR green (Eurogentec), on an ABI PRISM 7700 SDS (Applied BioSystems). The mRNA levels were normalized to the geometric average of four different housekeeping genes: *ACTB*, *GAPDH*, *UBC* and *HPRT1*. Numbers of differentially expressed genes before and after prioritization were compared with a chi-square test.

Zebrafish care and embryo manipulations. Wild-type zebrafish (*Danio rerio*) of the AB strain were maintained under standard laboratory conditions[32]. Morpholino oligonucleotides (Gene Tools) were injected into one- to four-cell-stage embryos[27]. Alcian blue cartilage staining was carried out as previously described[33]. All animal studies were reviewed and approved by the institutional animal care and use committee for Medical Ethics and Clinical Research of the University of Leuven.

Software availability. Endeavour is freely available for academic use as a Java application at http://www.esat.kuleuven.be/endeavour.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We wish to thank all groups and consortia that made their data freely available: Ensembl, NCBI (EntrezGene and Medline), Gene Ontology, BIND, KEGG, Atlas, InterPro, BioBase, the Disease Probabilities from Lopez-Bigas and Ouzounis[9] and the Prospectr scores from Euan Adie[8]. Ouzounis[8] and the Prospectr scores from Euan Adie[9]. We also thank the following people for their help in particular areas: Robert Vlietinck with the manuscript, Patrick Glenisson with text mining, Joke Allemeersch and Gert Thijs with the order statistics and Camilla Esguerra with the zebrafish experiments. S.A., D.L. and P.V.L. are sponsored by the Research Foundation Flanders (FWO). This work is supported by Flanders Institute for Biotechnology (VIB), Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT) (STWW-00162), Research Council KULeuven (GOA-Ambiorics, IDO genetic networks), FWO (G.0229.03 and G.0413.03), IUAP V-22, K.U.L. Excellentiefinanciering CoE SymBioSys (EF/05/007), EU NoE Biopattern and EU EST BIOPTRAIN to Y.M., and by the FWO (G.0405.06), GOA/2006/11 and GOA/2001/09, Squibb and EULSHB-CT-2004-503573 to P.C.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Quackenbush, J. Genomics. Microarrays—guilt by association. *Science* **302**, 240–241 (2004).
2 Kanehisa, M. & Bork, P. Bioinformatics in the post-sequence era. *Nat. Genet.* **33** Suppl. 305–310 (2003).
3. Ball, C.A., Sherlock, G. & Brazma, A. Funding high-throughput data sharing. *Nat. Biotechnol.* **22**, 1179–1183 (2004).
4. Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18** Suppl. 2, S110–S115 (2002).
5. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316–319 (2002).
6. Turner, F.S., Clutterbuck, D.R. & Semple, C.A. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* **4**, R75 (2003).
7. Tiffin, N. *et al.* Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* **33**, 1544–1552 (2005).
8. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**, 55 (2005).
9. Lopez-Bigas, N. & Ouzounis, C.A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* **32**, 3108–3114 (2004).
10. Kent, W.J. *et al.* Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737–741 (2005).
11. Altermann, E. & Klaenhammer, T.R. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics* **6**, 60 (2005).
12. Aerts, S. *et al.* TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.* **33**, W393–W396 (2005).
13. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. & De Moor, B. Computational detection of cis-regulatory modules. *Bioinformatics* **19** (Suppl 2), II5–II14 (2003).
14. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
15. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet.* **36**, 257–263 (2004).
16. Pixley, F.J. *et al.* BCL6 suppresses RhoA activity to alter macrophage morphology and motility. *J. Cell Sci.* **118**, 1873–1883 (2005).
17. Galimi, F. *et al.* Hepatocyte growth factor is a regulator of monocyte-macrophage function. *J. Immunol.* **166**, 1241–1247 (2001).
18. Brown, N.J. *et al.* Fas death receptor signaling represses monocyte numbers and macrophage activation in vivo. *J. Immunol.* **173**, 7584–7593 (2004).
19. Scambler, P.J. The 22q11 deletion syndromes. *Hum. Mol. Genet.* **9**, 2421–2426 (2000).
20. Baldini, A. Dissecting contiguous gene defects: TBX1. *Curr. Opin. Genet. Dev.* **15**, 279–284 (2005).
21. Jerome, L.A. & Papaioannou, V.E. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat. Genet.* **27**, 286–291 (2001).
22. Merscher, S. *et al.* TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell* **104**, 619–629 (2001).
23. Lindsay, E.A. *et al.* Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410**, 97–101 (2001).
24. Piotrowski, T. *et al.* The zebrafish van gogh mutation disrupts tbx1, which is involved in the DiGeorge deletion syndrome in humans. *Development* **130**, 5043–5052 (2003).
25. Rauch, A. *et al.* A novel 22q11.2 microdeletion in DiGeorge syndrome. *Am. J. Hum. Genet.* **64**, 659–666 (1999).
26. Graham, A. The development and evolution of the pharyngeal arches. *J. Anat.* **199**, 133–141 (2001).
27. Stalmans, I. *et al.* VEGF: a modifier of the del22q11 (DiGeorge) syndrome? *Nat. Med.* **9**, 173–182 (2003).
28. Glenisson, P. *et al.* TXTGate: profiling gene groups with text-based information. *Genome Biol.* **5**, R43 (2004).
29. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
30. Aerts, S., Van Loo, P., Moreau, Y. & De Moor, B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics* **20**, 1974–1976 (2004).
31. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
32. Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish*, (University of Oregon Press, Eugene, Oregon, 1994).
33. Kimmel, C.B. *et al.* The shaping of pharyngeal cartilages during early development of the zebrafish. *Dev. Biol.* **203**, 245–263 (1998).
34. Splawski, I. *et al.* Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
35. Robinson, S.W. *et al.* Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects. *Am. J. Hum. Genet.* **72**, 1047–1052 (2003).
36. Hayashi, T. *et al.* Identification and functional analysis of a caveolin-3 mutation associated with familial hypertrophic cardiomyopathy. *Biochem. Biophys. Res. Commun.* **313**, 178–184 (2004).
37. Zimprich, A. *et al.* Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
38. Zuchner, S. *et al.* Mutations in the pleckstrin homology domain of dynamin 2 cause dominant intermediate Charcot-Marie-Tooth disease. *Nat. Genet.* **37**, 289–294 (2005).
39. Munch, C. *et al.* Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* **63**, 724–726 (2004).
40. Tian, X.L. *et al.* Identification of an angiogenic factor that when mutated causes susceptibility to Klippel-Trenaunay syndrome. *Nature* **427**, 640–645 (2004).
41. Bienengraeber, M. *et al.* ABCC9 mutations identified in human dilated cardiomyopathy disrupt catalytic KATP channel gating. *Nat. Genet.* **36**, 382–387 (2004).
42. Windpassinger, C. *et al.* Heterozygous missense mutations in BSCL2 are associated with distal hereditary motor neuropathy and Silver syndrome. *Nat. Genet.* **36**, 271–276 (2004).
43. Tonkin, E.T., Wang, T.J., Lisgo, S., Bamshad, M.J. & Strachan, T. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nat. Genet.* **36**, 636–641 (2004).
44. Krantz, I.D. *et al.* Exclusion of linkage to the CDL1 gene region on chromosome 3q26.3 in some familial cases of Cornelia de Lange syndrome. *Am. J. Med. Genet.* **101**, 120–129 (2001).
45. Wang, X. *et al.* Positional identification of TNFSF4, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nat. Genet.* **37**, 365–372 (2005).
46. Peltekova, V.D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet.* **36**, 471–475 (2004).
47. Aharon-Peretz, J., Rosenbaum, H. & Gershoni-Baruch, R. Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. *N. Engl. J. Med.* **351**, 1972–1977 (2004).
48. Begovich, A.B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**, 330–337 (2004).
49. Helgadottir, A. *et al.* The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat. Genet.* **36**, 233–239 (2004).
50. Bertram, L. *et al.* Family-based association between Alzheimer's disease and variants in UBQLN1. *N. Engl. J. Med.* **352**, 884–894 (2005).