1544–1552 *Nucleic Acids Research, 2005, Vol. 33, No. 5*
doi:10.1093/nar/gki296

Integration of text- and data-mining using ontologies successfully selects disease gene candidates

Nicki Tiffin*, Janet F. Kelso, Alan R. Powell, Hong Pan[1], Vladimir B. Bajic[1] and Winston A. Hide

South African National Bioinformatics Institute, University of the Western Cape, Belville 7535, South Africa and [1]Knowledge Extraction Laboratory, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

Received November 19, 2004; Revised January 26, 2005; Accepted February 22, 2005

ABSTRACT

Genome-wide techniques such as microarray analysis, Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS), linkage analysis and association studies are used extensively in the search for genes that cause diseases, and often identify many hundreds of candidate disease genes. Selection of the most probable of these candidate disease genes for further empirical analysis is a significant challenge. Additionally, identifying the genes that cause complex diseases is problematic due to low penetrance of multiple contributing genes. Here, we describe a novel bioinformatic approach that selects candidate disease genes according to their expression profiles. We use the eVOC anatomical ontology to integrate text-mining of biomedical literature and data-mining of available human gene expression data. To demonstrate that our method is successful and widely applicable, we apply it to a database of 417 candidate genes containing 17 known disease genes. We successfully select the known disease gene for 15 out of 17 diseases and reduce the candidate gene set to 63.3% (±18.8%) of its original size. This approach facilitates direct association between genomic data describing gene expression and information from biomedical texts describing disease phenotype, and successfully prioritizes candidate genes according to their expression in disease-affected tissues.

INTRODUCTION

Many diseases are thought to be caused by altered gene function, and familial studies confirm the heritability of these diseases. Understanding genetic changes that cause disease has potential diagnostic, prognostic and therapeutic benefits. Many disease phenotypes are monogenic (Mendelian) traits for which single causative genes have been identified. However, complex trait phenotypes are controlled by multiple genes, and identifying these gene loci is confounded by many factors including locus heterogeneity, epistasis and pleiotropy, low penetrance of contributing genetic variants (1), variable gene expression levels and potential environmental effects on the disease state (2). Loci are generally large, containing up to 300 genes in humans (3–5).

Two strategies are commonly used to detect candidate disease-causing genes. The 'candidate gene' approach looks for statistical correlation between genetic variants and a disease according to data derived from experimental studies, and is favoured for simple study design and greater statistical power to detect several genes of small effect (4,6). Alternatively, the entire genome is scanned for disease genes, frequently employing human linkage data from concordant and discordant sib-pairs (2,6,7), with the generic problem that such high-throughput scans analysing thousands of genes may detect several hundred candidate genes. The same problem is encountered with other genome-wide techniques such as microarray analysis, Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature Sequencing (MPSS), which similarly generate large candidate gene sets. The challenge, and the problem we address here, is analysis of the set of several hundred candidate genes selected by linkage analysis and selection of a smaller subset of most probable candidate genes before embarking on expensive and time-consuming empirical analysis.

Analysis of candidate disease genes identified by genome-wide analysis has traditionally taken place at the laboratory bench in a laborious process of experimental elimination. Bioinformatic approaches allow '*in silico*' analysis of candidate genes through integration and analysis of relevant information from many sources, including single nucleotide

*To whom correspondence should be addressed. Tel: +27219592611; Fax: 27219592512; Email: nicki@sanbi.ac.za
Present address:
Janet F.Kelso, Max-Planck-Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany

polymorphism (SNP) data; protein–protein interactions, gene regulatory networks, gene structure variation, homologs, orthologs and expression data (3,4,6,8–11). Gene expression profiles are increasingly analysed in the search for candidate disease genes, as disease gene expression is often dysregulated in affected tissues (2,12). Examples include the use of gene expression data to identify a gene causing Leigh Syndrome in humans (13), and to reduce candidate gene lists for retinopathies (14) and the rat Rf-1 disease (15). We have developed and tested a generic approach to filter candidate genes selected from disease-associated loci, according to their expression profile. We use an anatomical ontology to integrate text-mining of scientific literature and data-mining of gene expression data to identify candidate disease genes. We use text-mining of PubMed abstracts to identify association between the disease name and anatomy terms. We then use the identified anatomy terms to independently identify genes that are expressed in these tissues from the Ensembl genomic database (http://www.ensembl.org).

Ontologies define terms within a specific subject area (16,17). The Unified Medical Language System (UMLS, http://umlsks.nlm.nih.gov) is a repository of biomedical vocabularies developed by the US National Library of Medicine aiming to standardize terminology, such as naming of genes, proteins, diseases and molecular functions, and includes Medical Subject Headings (MeSH), which has a clinical bias (18). The Gene Ontology (GO) describes molecular function, process and location of action of a protein in a generic cell (19). The eVOC ontologies (20) provide simple sets of controlled terms describing human anatomical systems, cell types, diseases and developmental stages. Organized as an intuitive and simple hierarchy, the terms in the anatomical system ontology provide a human-readable description of the terms commonly used in the annotation of samples taken for expression studies. In addition, the standardized terminology and hierarchical organization of the terms make the ontology computationally parseable. Annotation of the publicly available EST and mRNA data with terms from the ontology provides a means to connect expressed sequences with terms describing the location and timing of expression. These terms can also be found in the public literature. By using the ontologies to mine the public literature, we are able to connect expression data to the public literature. eVOC annotation of genes and transcripts is available through the Ensmart database (http://www.ensembl.org/Homo_sapiens/martview), and through a central site (www.sanbi.ac.za/evoc). Here, we use the eVOC anatomical ontology to link expression phenotype and genomic sequence (20).

Text-mining of PubMed abstracts and articles is used increasingly to extract molecular information from research literature (http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html) (21–26). However, clinical articles are underutilized by molecular biologists although they offer a complex phenotype description for the disease genotype under investigation. GO has been used in data- and text-mining, e.g. in the Onto-Tools suite, an annotation database with integrated data-mining tools (27), and Dragon TF Association Miner which associates transcription factors with GO terms and diseases (28). Text-mining of biomedical literature with MeSH terms has also been used in conjunction with GO to identify candidate disease genes (29). We use the eVOC Anatomical System ontology as a bridging vocabulary that integrates clinical and molecular data through a combination of text- and data-mining, and select candidate disease genes according to their expression profiles within tissues affected by the disease of interest. We first make an association between each eVOC anatomy term and disease name according to their co-occurrence in pubmed abstracts. This step does not implicate candidate disease genes in any way. We then rank the identified anatomy terms and select candidate genes annotated with the top-ranking terms (Figure 1).

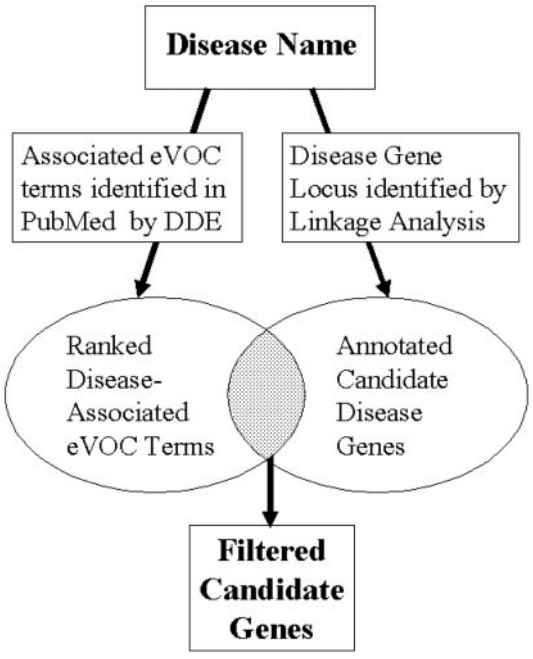

Figure 1. Schema outlining the method used to identify candidate disease genes.

METHODS

Summary

We first associate eVOC anatomy terms with disease names based on their co-occurrence in pubmed abstracts. We then rank the selected anatomy terms by calculating a *ranking score s* for each associated eVOC term, according to *frequency of association* and *frequency of annotation* of the eVOC term as defined below. *n* top-scoring eVOC terms are selected from the ranked list and these terms are compared with eVOC terms annotated to candidate disease genes selected from the Ensembl database to populate a training dataset. Our system allows *m* mismatched terms (mismatches) between the terms identified by text-mining and the terms used to annotate candidate genes. Genes selected from the training dataset as the final candidate gene list are those annotated with at least *n–m* eVOC terms that match top-scoring disease-associated eVOC terms. We apply the method with multiple values for parameters *m* and *n*, and report the results of the optimal values for *m* and *n*. We then use these optimal parameter values to run the system on a second, independent dataset (the test dataset).

Frequency of annotation

The *frequency of annotation* of RefSeq genes with terms from the eVOC Anatomical System ontology was calculated for each node (term) in the hierarchical ontology. The number of RefSeq genes at each node was the sum of all annotated RefSeq genes at the node and descendants of the node. The frequency of annotation for each term is the number of RefSeq genes at that node divided by the total number of annotated RefSeq genes.

Frequency of association

The *frequency of association* of each term as it occurs with the disease name is determined by text-mining of PubMed abstracts using Dragon Disease Explorer (DDE, http://research.i2r.a-star.edu.sg/DRAGON/DE/), and is the number of abstracts containing the term and the disease name divided by the total number of abstracts containing the disease name.

Calculation of ranking score '*s*'

Python scripts calculate a rank score for each eVOC term associated with a disease name, according to frequency of association and frequency of annotation of the term. Each associated anatomical term has a value for frequency of association and for frequency of annotation. We empirically determined the optimal weighting of these two values for the calculation of the rank score by altering the weighting of these terms and determining the effect on the ability of the system to select known disease genes from the training database. The optimal weighting of these values to determine rank of eVOC terms was empirically determined to be {rank score $s = [2*f(\text{association}) + f(\text{annotation})]/2$}. The python scripts are freely available at http://www.sanbi.ac.za/tiffin_et_al. We calculated the frequency of terms occurring in PubMed abstracts that do not contain the disease name, and found that incorporating this value in the calculation of rank score had no effect (results not shown). *n* top-scoring eVOC terms are selected from the ranked list and eVOC annotation of candidate disease genes from the Ensembl database are searched for these eVOC terms, allowing *m* mismatched terms (mismatches) between the selected top-ranking terms identified by text-mining of PubMed abstracts, and the annotated terms stored in the Ensembl database for each gene. Candidate genes from the Ensembl database that have at least *n*–*m* annotated eVOC terms that match top-ranking eVOC terms selected from PubMed abstracts in association with the disease name are listed as the selected candidate genes.

Construction of the training database

We tested our approach on a subset of genes representative of those that might be selected by a linkage analysis study. The size of the training database was chosen to approximate a set of such candidate genes. A training database was populated with data for 417 genes downloaded from Ensembl EnsMart 19.3 database (www.ensembl.org). In order to select a representative set of disease genes, we first identified various modes of gene dysregulation that may cause disease. This included point mutations in genes, mutation of the gene promoter, gene product overexpression, gene amplification, genetic translocation, loss of gene imprinting, dysregulated mRNA splicing and genes believed to predispose to disease by an unknown mechanism. For each type of dysregulation, we then selected one or more known disease gene. This generated a set of 17 known disease genes dysregulated by a large variety of genetic mechanisms. Depending on gene density at each disease gene locus, 5 to 10 additional genes were randomly selected from a region of 20 cM around each gene. To ensure sufficient representation of non-disease genes in the training database in addition to the randomly selected genes, non-disease genes were selected to fulfil the categories of metabolic housekeeping genes, rRNA genes, tRNA genes, structural and cytoskeletal genes, immune molecules (Igs), protein family members, network members and binding partners from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.ad.jp/kegg/) and the preBIND database (http://www.blueprint.org/products/prebind/prebind.html) (total = 94 genes). Again, 5 to 10 genes were selected from the same locus of each non-disease gene and added to the training database bringing the number of randomly selected genes to a total of 306 genes. The training database is created in MySQL with the same structure as the Ensembl ensMART database (www.ensembl.org/Homo_sapiens/martview), and is freely available at http://www.sanbi.ac.za/tiffin_et_al.

Selecting optimal values for *m* and *n*

The system is run with all combinations of assigned mismatched term number *m* (range: 0 to 5) and term number *n* [range: $(m + 1)$ to 12]. When using a value for *n* greater than 12, we found that few candidate genes were selected, therefore used 12 as a maximum value. For each combination of values for *n* and *m*, we determine the frequency with which the known disease gene is selected, and with which the final set size falls within specified limits for each pair of values used for *m* and *n* when the system is tested with all 17 known disease genes in the training database. In order to determine which pair of *m* and *n* values provide useful results, the user may assign cut-off values for frequency with which the known disease gene is selected when all 17 diseases are analysed *(true positive frequency)* according to the priority they wish to place on correct selection of the known disease gene, and the frequency with which the final set size falls within user-specified limits (*set size frequency*) according to the priority they wish to place on reducing the size of the selected candidate set. The system is run with all combinations of parameters *m* and *n*, and only those pairs of values that give results fulfilling the required success rates specified by *true positive frequency* and *set size frequency* are identified. For control experiments to show that selection of the known disease gene is not by chance, we assign randomly selected genes from the training database to each disease name in place of the known disease gene, and determine whether this random gene is selected by the method. We repeat this process ten times for each disease name. Genes are randomly selected from the training database using the Python pseudorandom number generator, random() (http://docs.python.org/lib/module-random.html) and the frequency with which the random gene is selected out of the ten runs per disease is calculated. We use the Fisher Exact Test to test whether the results from our experiment are significantly different from those obtained in this control experiment. As an

additional control experiment, we address the possibility that inclusion of the known disease gene in the selected subset is a random event occurring only by chance. We measure the probability of the disease gene being selected as a function only of final set size by calculating the value of final set size compared to original set size.

Validation of selected parameters

In order to verify that the system is generally applicable using the parameters determined from the training database, we constructed a second, independent database (the test dataset) containing Ensembl data for 20 known disease genes randomly selected from the OMIM database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) and all genes falling in the same cytogenetic bands as those disease genes (2191 additional genes). We apply the method to this independent dataset using optimal parameter values for *m* and *n*, as determined from running the system on the training database. The test dataset is freely available at http://www.sanbi.ac.za/tiffin_et_al/ test_dataset/).

RESULTS

Using the Anatomical System ontology, the disease gene was correctly selected from the training database with a *true positive frequency* of >80% with a user-specified set size limit of 350 genes (*set size frequency* >90%) when $n = 4$ terms and $m = 1$ mismatched term. The correct disease gene was present in the selected subset of genes for 15/17 (88.2%) of the diseases in the training database. We performed two control experiments. For the first, the average rate of selection of the assigned gene for 10 runs with random gene assignment averages only 45% (range: 10–80%, $P = 0.0006$ using the Fishers Exact Test). For the second, we measured the likelihood that the known disease gene is selected by chance alone, and find that the chance of selection of the known gene according to only the size of the selected set was an average of only 63% (range: 31–91%) (Figure 2). Size of candidate gene sets for the 15 successful cases ranged from 129 to 379 (from a total of 417 genes), an average reduction in size to 63.3% (range 30.9–90.9%) of the original candidate gene set size (Figure 3). Here, we prioritize inclusion of the known disease gene in the selected set by

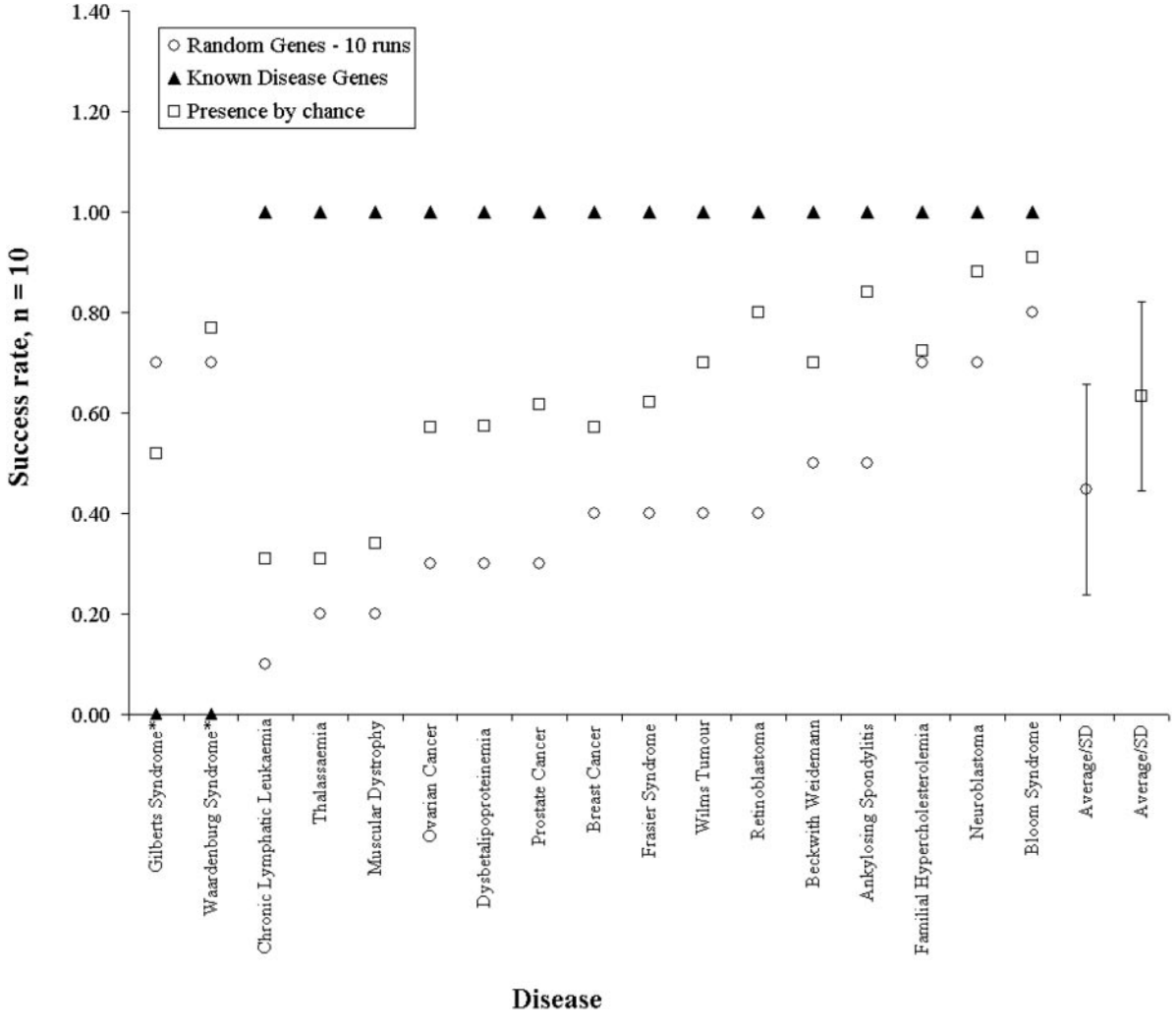

Figure 2. Success of finding the disease gene using mismatched terms $m = 1$, number of terms $n = 4$ is scored as successful = 1 and unsuccessful = 0 (closed triangle). Randomly selected genes were assigned to each disease name and their presence detected in the selected gene set. Detection rate is shown per disease, for 10 runs with random gene assignment (open circle). Likelihood of the known disease gene being selected by chance alone was calculated as the probability of the disease gene falling into the candidate gene set according to the size of the candidate disease gene set (open square).

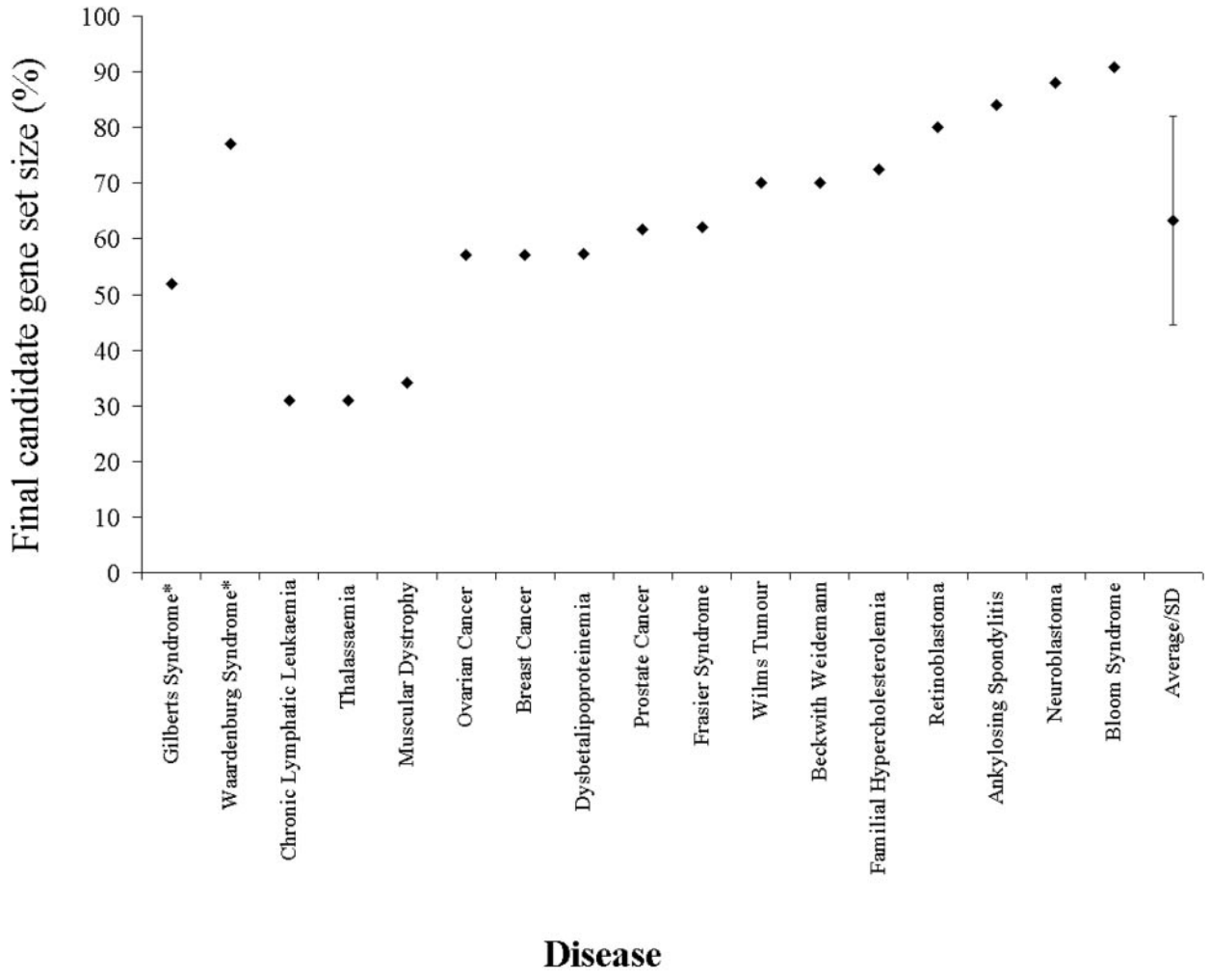

Figure 3. Reduction in size of the candidate gene set size using mismatch $m = 1$ and number of terms $n = 4$. Final set size is shown as a percentage of starting set size. Average and SD values are shown. *Known disease gene is not present in the selected candidate gene set.

specifying a *true positive frequency* value of >80%, selecting only combinations of *m* and *n* that select the correct disease gene in >80% of the 17 diseases; however, the system is flexible according to the requirements of the researcher, and *true positive frequency* and *set size frequency* limits used to determine parameters *n* and *m* could also be specified to prioritize reduction in set size over accurate selection of the known disease gene. Subsequent running of the method on the independent test dataset using the parameter values determined from the training database ($m = 1$ and $n = 4$) resulted in the presence of the known disease gene in the selected subset of candidate genes for 19/20 cases (95%), with an average reduction in size of the candidate gene set to 64.2% (±10.7%) of the original set size (data available at http://www.sanbi.ac.za/tiffin_et_al/).

The number of eVOC Anatomical System terms found associated with each disease from the training dataset ranges from 11 to 198 (Table 1), with a frequency of association range of 0.1690–0.9905. The frequency of annotation of terms to RefSeq genes ranges from 0 to 0.928, with 55% of terms having a value of 0 (Figure 4), indicating that not all eVOC terms used in PubMed abstracts are used to annotate genes, and consequently may not be informative. We found that the number of abstracts used for text-mining did not affect the success of the method (Table 1). Additional testing of the method using the eVOC Cell Type and Pathology ontologies were not successful, and this is most likely due to less frequent annotation of genes with these terms.

We tested the validity of text-mining in identifying disease-affected tissues by comparing the top 12 ranked eVOC terms associated with Wilms' tumour by text-mining to a list of 12 tissues commonly affected in Wilms' tumour as provided by specialists in this field (Dr R. D. Williams and colleagues, Department of Paediatric Oncology, Institute of Cancer Research, United Kingdom). Text-mining delivered generally disease-relevant terms, with 10 of the 12 terms selected by text-mining equivalent to 7 of the 12 terms provided by the specialists. To confirm that abstracts naming the known disease gene in conjunction with the disease name do not alter the results of text-mining, we identified the earliest paper in which the disease-causing gene WT1 was reported to be mutated in a cohort of Wilms' Tumour cases, published in 1991 (30), and analyzed only the set of Wilms' Tumour abstracts published up to 1990. The three selected top-ranking

Table 1. Information for each disease, showing terms associated with disease name (Associated terms), terms annotated to the disease gene (Annotated terms), number of terms associated with disease name and also used to annotate the disease gene (Common eVOC terms), number of common terms falling within the highest four ranking terms (Terms ranked in top 4) and total number of candidate disease genes selected

Disease	Disease gene	Number of abstracts	Candidate genes	Associated terms	Annotated terms	Common eVOC terms	Terms ranked in top 4
Gilberts syndrome[a]	UGT1A1	316	216	37	8	6	1
Waardenburg syndrome[a]	PAX3	213	321	57	10	5	1
Frasier syndrome	WT1	36	259	11	16	4	3
Dysbetalipoproteinemia	APOE	331	239	37	38	11	3
Chronic lymphatic leukemia	BCL2	407	129	65	28	20	3
Bloom syndrome	BLM	447	379	51	22	14	3
Beckwith Weidemann	IGF2	505	292	76	22	16	3
Ankylosing spondylitis	HLA-B	3451	350	137	53	38	3
Wilms tumour	WT1	4764	292	152	16	15	3
Muscular dystrophy	DMD	7210	142	166	40	33	3
Thalassemia	HBB	7510	129	143	34	31	3
Retinoblastoma	RB1	8617	334	175	35	34	3
Ovarian cancer	BRCA1	11 227	238	142	25	24	3
Neuroblastoma	MYCN	15 409	367	197	20	20	3
Familial hypercholesterolemia	LDLR	15 506	302	159	34	31	3
Prostate cancer	ERBB2	17 876	257	164	30	29	3
Breast cancer	BRCA1	51 759	238	198	25	24	3

[a]Known disease gene not present in candidate gene set

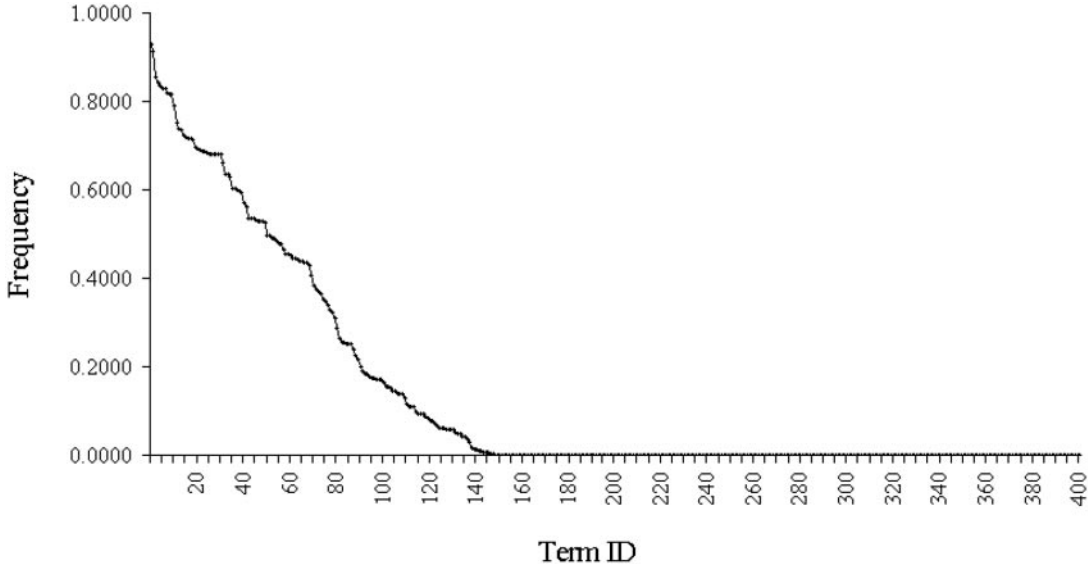

Figure 4. Frequency of annotation per 'eVOC Anatomy Term' in the EnsMart database. The frequency of RefSeq genes at each node was the sum of all annotated RefSeq genes at the node and descendants of the node, compared to total number of annotated RefSeq genes.

eVOC anatomy terms in this reduced set of abstracts were found to be the same as those selected from the total abstract set.

DISCUSSION

The system that we have designed successfully uses a controlled vocabulary of anatomical terms, the eVOC Anatomical System ontology, to match tissues associated with disease to genes expressed in those tissues, and we demonstrate this by successful selection of known candidate disease genes in subsets selected from a training database of 417 genes, and an independent test dataset of 2211 genes. Our system first selects and ranks eVOC anatomical terms that are found to be associated with disease names in PubMed abstracts. Then, candidate disease genes identified by linkage disequilibrium are selected according to their annotation in the Ensembl database with the identified eVOC terms. Our system succeeds in selecting the correct disease gene amongst other candidate genes in 15 out of 17 diseases in the training dataset (88.2% success rate). The diseases for which the candidate gene set is most reduced (to 30.9% of original size) by our method are chronic lymphatic leukemia in which *BCL2* is frequently overexpressed due to chromosomal translocation, and β-thalassemia caused by mutation in the beta-globin gene *HBB*. The disease for which the candidate gene set is least

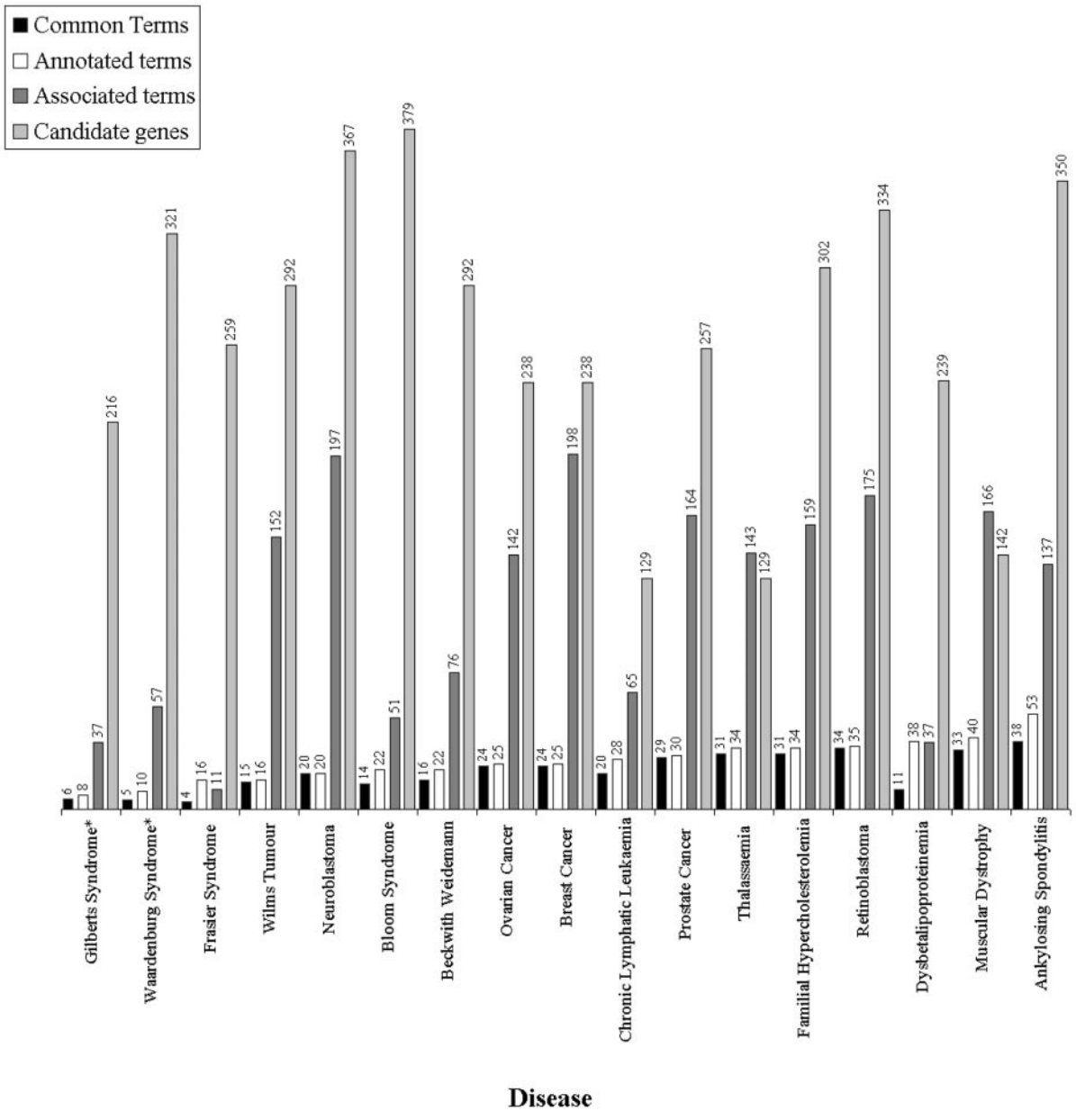

Figure 5. Schema showing the relationship between the number of candidate genes selected (Candidate genes), number of terms associated with the disease name (Associated terms), number of terms used to annotate the disease gene (Annotated terms) and number of terms associated with disease name and also used to annotate the disease gene (Common terms). *Known disease gene is not present in the candidate gene set.

reduced (to 90.9% of original size) is Bloom Syndrome, in which the *BLM* gene is mutated (see Table 1). Abstract number and type of gene mutation do not affect gene selection or final set size. For Waardenburg Syndrome and Gilberts Syndrome, the disease gene is not selected although it is present in the training database, and the terms in common between disease-associated and gene-annotated terms tend to be low ranking. In general, higher numbers of terms associated with disease name do not result in an increase in common term number, and this suggests that the number of eVOC annotations per gene, rather than associated term number, is the limiting factor in the number of common terms selected. The disease genes for Waardenburg Syndrome and Gilberts Syndrome are poorly annotated with eVOC terms, and this is likely to have contributed to these genes not being selected as candidates by our method (Table 1, Figure 5).

We confirm the wider applicability of this system by testing the parameters optimized using the training database on a second separate database, the test dataset, containing 1211 genes including 20 known disease genes. In this scenario, the system successfully selects the known disease gene in 95% of cases, and reduces the size of the candidate gene set to 64.2% (±10.7%) of the original set size (data available at www.sanbi.ac.za/tiffin_et_al/). These results are comparable to those obtained with the training database. The success rate is marginally higher for the test dataset, and the average reduction in the set size marginally lower although with a greatly reduced SD. Thus, the selected values for parameters

m and *n* appear to be applicable to datasets beyond the training database that was used to determine the optimal parameter values.

The eVOC Anatomical System ontology successfully connects clinical data in PubMed abstracts to the expression profiles of genes. It offers the advantage of small size, uncomplicated structure and accessible terminology that is applicable to both clinical and molecular biology disciplines, in contrast to more discipline-specific ontologies such as GO and MeSH. This simplicity permits the terms selected by text-mining using the eVOC ontology to be used directly for data-mining genomic databases. Several other bioinformatic methods to identify candidate disease genes have been recently described. Perez-Iratxeta *et al*. use MeSH and GO terms in a multiple-step, inference-based text- and data-mining procedure that correctly detects 55 of 100 known disease genes (29). This system uses text-mining of PubMed abstracts to make associations between MeSH pathological terms and MeSH chemical terms, and then to link the chemical terms to GO functional annotations. Genes with those GO functional annotations are selected and ranked according to the number of terms they share. Freudenberg and Propping describe a system that clusters known disease genes according to phenotypic similarity between their associated diseases, and ranks candidate disease genes by comparing their GO annotations to those of the clustered disease genes (31). Their system detects the known disease gene, within a wide range of rankings, for two-thirds of 10 672 diseases documented in OMIM. Turner *et al*. describe POCUS, a system to predict candidate disease genes according to enrichment of GO and InterPro domain annotation for genes within a set of specified disease-associated genetic loci (32). Depending on parameters and thresholds applied, this system successfully identifies two or more known disease genes for 15–65% of 29 complex diseases.

In comparison to these approaches, our method avoids the complexity of the extensive MeSH and GO vocabularies. The eVOC anatomical terminology is simple and purely descriptive, and we avoid interpretational bias that may be encountered with functional annotation systems such as GO. Also, our method does not employ indirect inference between terms, and the eVOC terms identified by text-mining are used directly in gene annotation. This combination of simplicity and direct association between biomedical texts and genetic data, without functional interpretation, allows our system a significant advantage and a high success rate (88.2%). Avoiding analysis based on gene functional annotation also allows novel candidate gene selections to be made outside current functional knowledge paradigms.

Van Driel *et al*. describe GeneSeeker, a web-based application that mines up to 9 web-based databases for candidate genes using expression terms defined by the user, and collates positional and expression/phenotypic information to provide an overview of the candidates (33). This system requires some clinical knowledge by the user, and search terms and gene data do not conform to a controlled vocabulary, which may cause appropriate information to be missed while searching. The system is demonstrated for ten human malformation syndromes, but the general success rate is not determined. In contrast, our method employs efficient and comprehensive querying of controlled vocabulary terms stored in a relational database. Also, our method provides a crucial additional step of text-mining biomedical literature to determine appropriate anatomical sites affected in a disease and thus allows the system to be widely applicable to all diseases regardless of type of disease or the domain knowledge of the user.

We are confident that text-mining of abstracts using disease names and the eVOC Anatomical System terms generates terms that are relevant to the symptoms of the disease, as illustrated by the comparison made between terms identified for Wilms' tumour by our system, and terms proposed by specialists in the field. Although there may be occasional terms identified by text-mining that are not disease-relevant, our system searches for four common eVOC terms but allows 1 mismatched term to occur, and this permitted mismatch can accommodate the effect of occasional inappropriate terms that may be selected by text-mining.

This generic method of text- and data-mining using ontologies offers a rapid and reliable approach to selection of candidate disease genes according to expression profile, giving researchers a valuable starting point to prioritize candidates for a wide range of diseases. When compared with existing systems to prioritize candidate disease genes, our method has a high success rate and an approach that is unique in its use of a simple anatomical ontology to directly associate biomedical literature describing disease phenotype and annotated gene expression data. We have avoided using functional data due to its intrinsic interpretative nature and the complexity of GO terminology. Our method is flexible, and parameter values (*m* and *n*) can be chosen to prioritize true positive frequency or candidate gene set size reduction, according to the researcher's preference. Our method relies on annotation of genes using a controlled vocabulary to integrate molecular biology expression data and clinical disease data, and these results emphasize that extensive application of appropriate controlled vocabularies to biomedical data will enhance the efficacy and productivity of text- and data-mining. Where candidate genes are poorly annotated, future research may include analysis of annotation of orthologous and paralogous genes, in order to further define expression profiles of candidates.

To date, no single method is able to accurately predict candidate disease genes in one step. Rather, a concert of methods is applied to prioritize most likely candidate disease genes from sets identified by such techniques as linkage analysis and microarray analysis. Existing methodologies mine biological and functional information about candidate genes, and we believe that our system can complement these existing approaches by using a novel method that mines expression data for candidate genes, linking this data with anatomical sites that are implicated in the disease. By using text-mining of PubMed abstracts, our method allows researchers to utilize associations between disease name and affected tissues without having a clinical understanding of the disease, and to apply this data in the selection of candidate genes. Employing the additional facet of gene expression data in this way can significantly assist in the process of focusing the search for most likely disease gene candidates.

ACKNOWLEDGEMENTS

Richard D. Williams, Paediatric Oncology, Institute of Cancer Research, United Kingdom, for generating a curated list of

tissues implicated in Wilms' tumour. Cathal Seoighe, Computational Biology Group, University of Cape Town, South Africa, for methodology advice. Charles Auffray, CNRS, Villejuif, France and Ranajit Chakraborty, Centre for Genomic Information, University of Cincinnati, USA. for critical review of the manuscript. This work was funded by the Medical Research Council South Africa, the National Bioinformatics Network South Africa and the Wellcome Trust grant number CRIG,HH7MD. Funding to pay the Open Access publication charges for this article was provided by the Medical Research Council South Africa.

Conflict of interest statement. None declared.

REFERENCES

1. Pritchard,J.K. and Cox,N.J. (2002) The allelic architecture of human disease genes: common disease-common variant... or not? *Hum. Mol. Genet.*, **11**, 2417–2423.
2. Hoh,J. and Ott,J. (2004) Genetic dissection of diseases: design and methods. *Curr. Opin. Genet. Dev.*, **14**, 229–232.
3. Glazier,A.M., Nadeau,J.H. and Aitman,T.J. (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.
4. Tabor,H.K., Risch,N.J. and Myers,R.M. (2002) Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Rev. Genet.*, **3**, 391–397.
5. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
6. McCarthy,M.I., Smedley,D. and Hide,W. (2003) New methods for finding disease-susceptibility genes: impact and potential. *Genome Biol.*, **4**, 1191–1198.
7. Ghosh,S., Reich,T. and Majumder,P.P. (2002) Linkage mapping of quantitative trait loci in humans: an overview. *Ann. Hum. Genet.*, **66**, 431–438.
8. Marnellos,G. (2003) High-throughput SNP analysis for genetic association studies. *Curr. Opin. Drug Discov. Devel.*, **6**, 317–321.
9. Bell,J.I. (2002) Single nucleotide polymorphisms and disease gene mapping. *Arthritis Res.*, **4**, S273–S278.
10. Freudenberg,J. (2003) Genome-wide prediction of disease-relevant genes and variants. *Curr. Opin. Drug Discov. Devel.*, **6**, 304–309.
11. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.*, **33**, 228–237.
12. Winter,E.E., Goodstadt,L. and Ponting,C.P. (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.*, **14**, 54–61.
13. Mootha,V.K., Lepage,P., Miller,K., Bunkenborg,J., Reich,M., Hjerrild,M., Delmonte,T., Villeneuve,A., Sladek,R., Xu,F. *et al.* (2003) Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc. Natl Acad. Sci. USA*, **100**, 605–610.
14. Katsanis,N., Worley,K.C., Gonzalez,G., Ansley,S.J. and Lupski,J.R. (2002) A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. *Proc. Natl Acad. Sci. USA*, **99**, 14326–14331.
15. Vitt,U., Gietzen,D., Stevens,K., Wingrove,J., Becha,S., Bulloch,S., Burrill,J., Chawla,N., Chien,J., Crawford,M. *et al.* (2004) Identification of candidate disease genes by EST alignments, synteny, and expression and verification of Ensembl genes on rat chromosome 1q43-54. *Genome Res.*, **14**, 640–650.
16. Schulze-Kremer,S. (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biol.*, **2**, 179–193.
17. Bard,J.B. and Rhee,S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nature Rev. Genet.*, **5**, 213–222.
18. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.
19. Camon,E., Barrell,D., Lee,V., Dimmer,E. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database—an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.*, **4**, 5–6.
20. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V., McCarthy,M.I. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.
21. Dickman,S. (2003) Tough mining: the challenges of searching the scientific literature. *PLoS Biol.*, **1**, E48.
22. Andrade,M.A. and Bork,P. (2000) Automated extraction of information in molecular biology. *FEBS Lett.*, **476**, 12–17.
23. de Bruijn,B. and Martin,J. (2002) Getting to the (c)ore of knowledge: mining biomedical literature. *Int. J. Med. Inf.*, **67**, 7–18.
24. Grivell,L. (2002) Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO Rep.*, **3**, 200–203.
25. Roberts,R.J., Varmus,H.E., Ashburner,M., Brown,P.O., Eisen,M.B., Khosla,C., Kirschner,M., Nusse,R., Scott,M. and Wold,B. (2001) Information access. Building a 'GenBank' of the published literature. *Science*, **291**, 2318–2319.
26. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, E309.
27. Khatri,P., Bhavsar,P., Bawa,G. and Draghici,S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.
28. Pan,H., Zuo,L., Choudhary,V., Zhang,Z., Leow,S.H., Chong,F.T., Huang,Y., Ong,V.W., Mohanty,B., Tan,S.L. *et al.* (2004) Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. *Nucleic Acids Res.*, **32**, W230–W234.
29. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genet.*, **31**, 316–319.
30. Cowell,J.K., Wadey,R.B., Haber,D.A., Call,K.M., Housman,D.E. and Pritchard,J. (1991) Structural rearrangements of the WT1 gene in Wilms' tumour cells. *Oncogene*, **6**, 595–599.
31. Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics.*, **18**, S110–S115.
32. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
33. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P., Leunissen,J.A. and Brunner,H.G. (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. J. Hum. Genet.*, **11**, 57–63.