Published online 2 November 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue **D975–D979**
doi:10.1093/nar/gkq1024

CCDB: a curated database of genes involved in cervix cancer

Subhash M. Agarwal[1],*, Dhwani Raghav[1], Harinder Singh[2] and G.P.S. Raghava[2]

[1]Bioinformatics Division, Institute of Cytology and Preventive Oncology, I-7, Sector-39, Noida 201301 and [2]Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India

Received August 14, 2010; Accepted October 10, 2010

ABSTRACT

The Cervical Cancer gene DataBase (CCDB, http://crdd.osdd.net/raghava/ccdb) is a manually curated catalog of experimentally validated genes that are thought, or are known to be involved in the different stages of cervical carcinogenesis. In spite of the large women population that is presently affected from this malignancy still at present, no database exists that catalogs information on genes associated with cervical cancer. Therefore, we have compiled 537 genes in CCDB that are linked with cervical cancer causation processes such as methylation, gene amplification, mutation, polymorphism and change in expression level, as evident from published literature. Each record contains details related to gene like architecture (exon–intron structure), location, function, sequences (mRNA/CDS/protein), ontology, interacting partners, homology to other eukaryotic genomes, structure and links to other public databases, thus augmenting CCDB with external data. Also, manually curated literature references have been provided to support the inclusion of the gene in the database and establish its association with cervix cancer. In addition, CCDB provides information on microRNA altered in cervical cancer as well as search facility for querying, several browse options and an online tool for sequence similarity search, thereby providing researchers with easy access to the latest information on genes involved in cervix cancer.

INTRODUCTION

Carcinoma of the uterine cervix is the most common malignancy that affects women in several parts of the world and causes high mortality (1). It is the second most common cancer affecting women worldwide and its occurrence has steadily increased in young women (2). It has been documented that approximately 500 000 new cases and 274 000 deaths occur each year due to this disease (3). This disease is more prevalent among women of low socioeconomic status and is a major health problem in developing countries. In India, it has become the leading cancer among women with an annual incidence of about 130 000 cases and 70–75 000 deaths (4). It has been established that human papillomavirus (HPV) is an etiologic agent for development of cervical cancer, yet it has been observed that an infection caused by HPV alone is not a sufficient factor for causation of cervical cancer. Rather, genetic and environmental factors that act in concordance have been suggested to contribute to the induction of cervical cancer. It is thought that other reasons including changes in tumor suppressor genes, activation of cellular proto-oncogenes, chromosomal alterations or altered expression of several genes too are required for tumorigenesis (5). As a result several investigators in recent years have, by conducting small-scale (individual gene) and high-throughput studies like microarray expression profiling (6), compared normal cervical cells with premalignant and malignant cervical cells. This has resulted in generation of enormous data and revelation of hundreds of genes that are differentially expressed, thus providing researchers important resources to potentially explore the molecular mechanisms and identify cervix-cancer-related genes. Despite presence of voluminous biomedical literature in Pubmed that provides evidence for variety of genes implicated in cervical cancer and women worldwide increasingly being affected from this malignancy, to our knowledge no resource exists that focuses on cervix cancer. Strikingly, a number of databases have been published over the years that collect information on several specialized cancers like Human Lung Cancer Database (7), Prostate Gene Database (8), Oral Cancer Gene Database (9) and Breast Cancer Gene Database (10), but none of these databases exists for genes involved in cervix cancer causation, one of the leading cancers that affect women worldwide.

Downloaded from nar.oxfordjournals.org by guest on April 28, 2011

*To whom correspondence should be addressed. Tel: +91 120 2579471 72; Fax: +91 120 2579473; Email: smagarwal@yahoo.com

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Therefore, we have collected cervix-cancer-related genes to construct an integrated database termed Cervical Cancer gene DataBase (CCDB) that catalogs the genes known to be involved in cervical carcinogenesis as evidenced from the biomedical literature. To gather and uniformly present the available information on cervix cancer genes, we have created a user-friendly interface in the form of CCDB. The database integrates heterogeneous data including basic gene and protein information, manually curated literature references to support inclusion of a gene in the database, the biological alteration, gene ontology (GO) information, homologous sequences in other eukaryotic genomes, interacting partners, microRNA (miRNA) reported to be altered in cervical carcinomas as compared with normal cervix, as well as information and links of several external resources/databases in order to help in retrieval of any other related information. In addition, the database provides search facility for querying the database and an online tool for sequence similarity search. Overall, CCDB is a specialized, first of its kind value-added database that will enable the exploration of relevant information for all experimentally determined human cervix-cancer-related genes making it a unique resource in the area of cervix cancer biology.

DATA COLLECTION AND CONTENT

For collection of genes playing role in causation of cervix cancer we have extensively searched Pubmed database and collected the relevant literature manually. After we obtained the literature, we read through the full text of each article to identify one or more genes involved in disease process. Thereafter, information regarding the type of molecular and genetic events (such as methylation, gene amplification, mutation, altered expression and polymorphism) responsible for occurrence of cervical cancer disease as documented in the references was recorded along with the tracking number (PMID). Also, we have collected information about the location from which sample was collected by their respective authors to conduct the study. This information has been introduced in the reference table. Once a non-redundant list of genes was extracted from the literature further information regarding the gene was derived and has been integrated into other tables. Overall, CCDB is divided into six tables: (i) the gene detail table, which stores information necessary for the conversion between different gene and protein identifiers, using the Gene ID as primary key, (ii) the homology table, which stores the orthology relationships derived from Homologene database (11), (iii) the reference table (detailed above), (iv) the GO table, (v) the sequence table, which stores mRNA, CDS and protein location and sequences and (vi) 3D protein structure information table (Figure 1).

miRNAs are a family of small non-coding RNA molecules that downregulate the expression of their protein-coding gene targets. The recent studies provide evidences that multiple miRNAs have altered expression in various human cancers, including cervical cancer. This provides a good indication that knowledge of differential expression of miRNA in cancer may have substantial diagnostic and prognostic value and hence these are considered as an important resource for cancer research. Therefore, we selected cervix-cancer-related miRNAs with experimental information from the literature and documented information regarding host gene that codes for miRNA, its location, the target gene and its potential role.

Once all the information was gathered we integrated the data in MySQL, an object-relational Database Management System (RDBMS), which works at the backend and the web interface, was built in PHP. Currently, we have collected 537 genes that are involved in the different stages of cervical carcinogenesis. The information thus accumulated is expected to be critical for both scientific researchers and clinicians to understand and determine the molecular mechanism that causes cervix cancer.

DATA ACCESS

The data in the CCDB can be easily accessed in a variety of ways. Users can query the database by gene name, gene ID or chromosome number, which results in display of gene-centered information in a new page (Figure 2). The main page for each gene provides the following information (i) a schematic view of intron/exon structure, (ii) general information including the gene name, gene id, contig, location, OMIM id, unigene id, Ensembl id, HPRD id, (iii) link to references that validate the presence of this gene in cervical cancer, (iv) link to homologous gene entries, (v) link to mRNA/CDS/Protein sequence details and (vi) link to other public databases, e.g. HGNC and PharmaGKB. Clicking on HPRD link leads to GO information, and homologene id link provides the details of similar sequences (mRNA and protein id) that are present in other eukaryotic genomes along with multiple sequence alignment generated using clustalX (12) in pdf format. Further, to know how a gene is related to cervix cancer, the user can click on the reference link that displays the references used as evidence for establishing relationship of the gene to cervix cancer along with the aberrant biological process responsible for causing cervix cancer. Moreover, the mRNA/CDS/Protein link provides information regarding the various isoforms encoded by the gene, their location, length as well as mRNA, CDS and protein sequences of each isoform.

Also CCDB provides three other ways to view and retrieve all cervix-cancer-related genes. First, a user can query the database for human chromosome number to display the genes present on each chromosome and then browse them individually. Second, CCDB also offers a browsing section, which allows the user to access the entire collection of genes (complete list) or search the genes by their names that are ordered in the alphabetical mode. Finally, the genes have been categorized into biological process whose alteration leads to cervix cancer disease as evident from literature. Thus, CCDB provides a gateway through which the biomedical community can

Downloaded from nar.oxfordjournals.org by guest on April 28, 2011

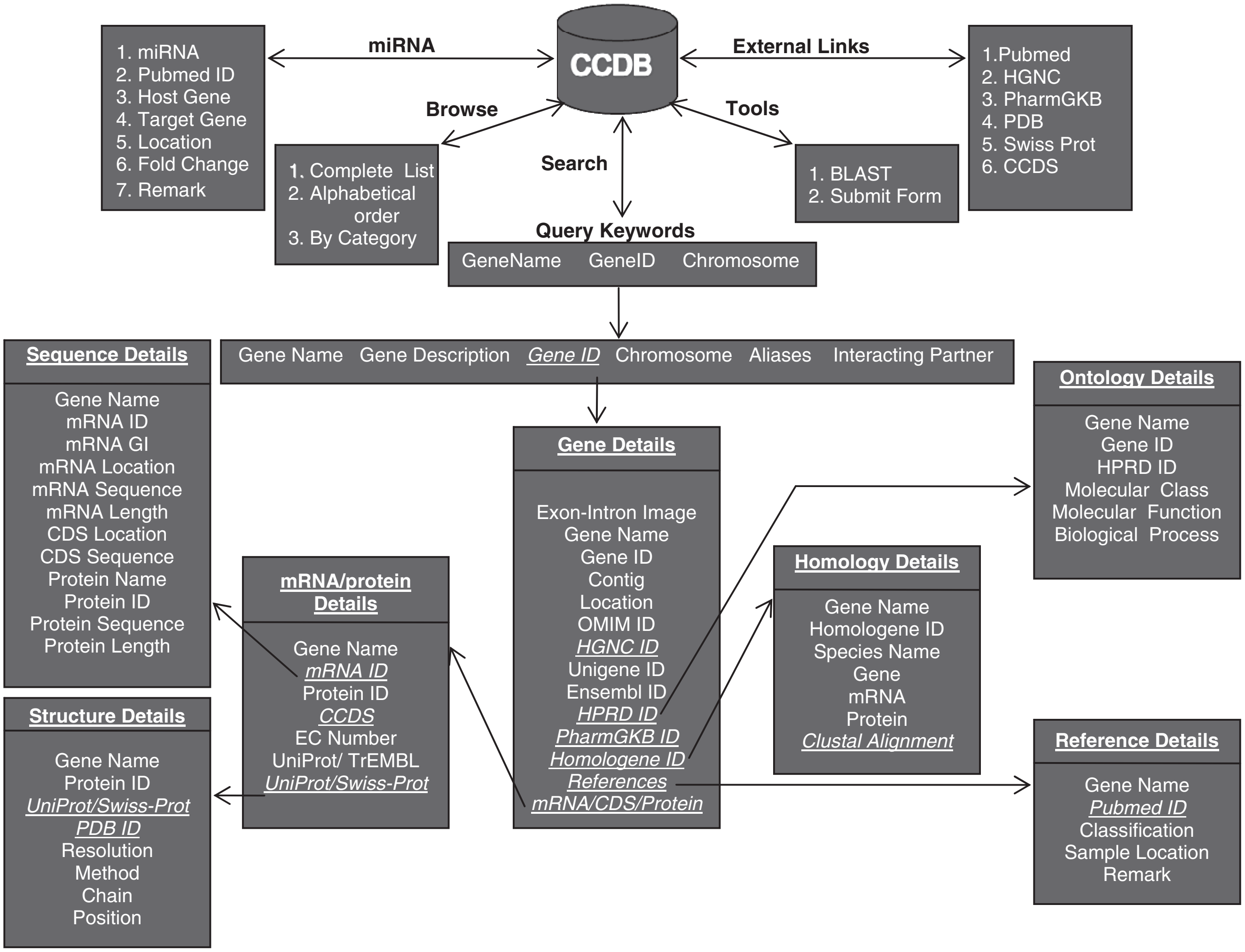


Figure 1. The database structure of CCDB.

easily access the latest information on the genes involved in cervix cancer.

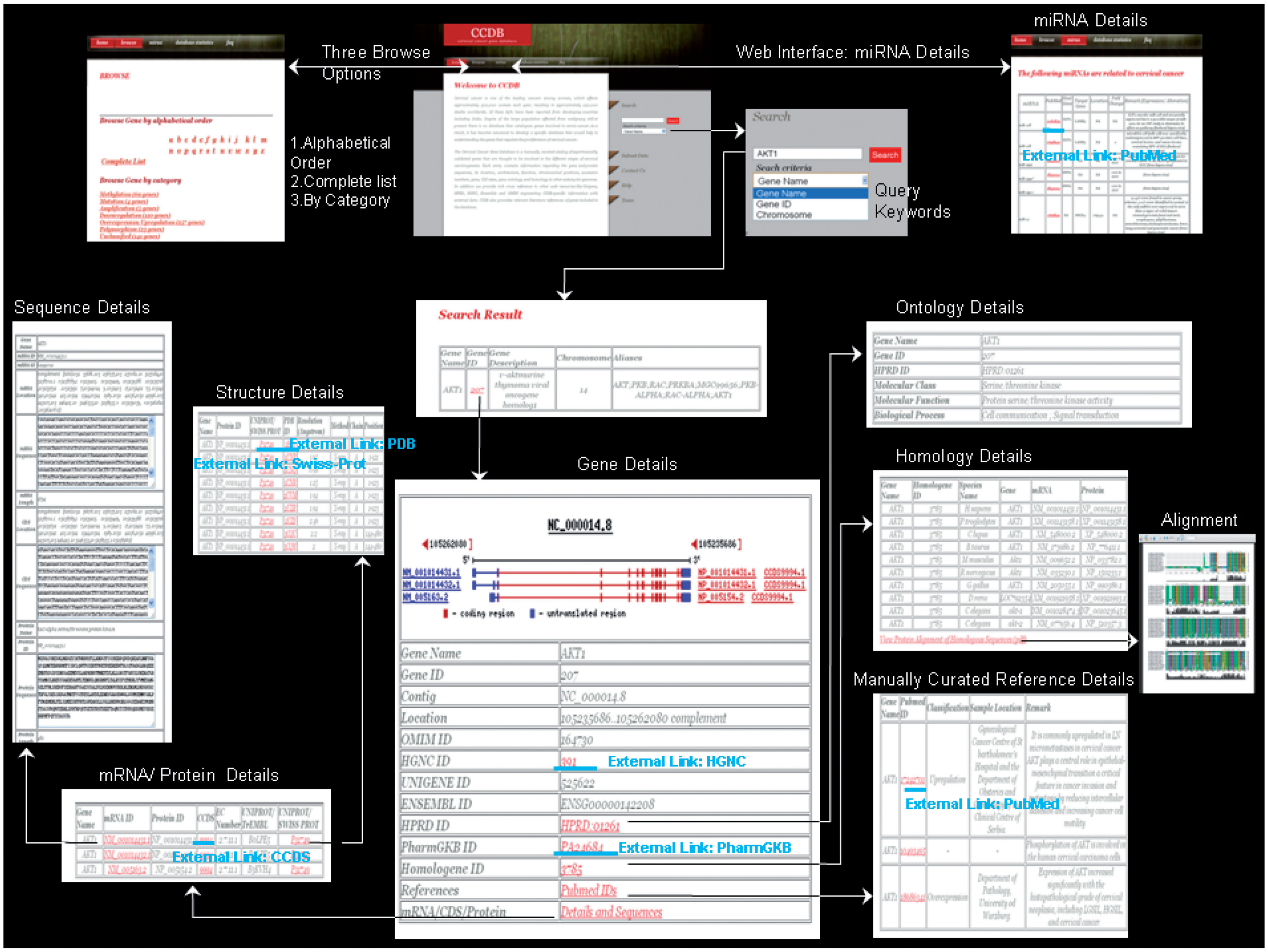
Further, a customized BLAST (13) tool has been made available that searches a user-defined query against the sequences available in the database. It may be useful in characterization of orphan sequences and fishing out of homologous sequences from the database, based on sequence similarity. Additionally, an online submission facility has been provided to add gene entries that are associated with cervix cancer (this option is yet to be functional). Once the user adds new gene information with specified fields (the fields with star are mandatory), the database would be updated after validation.

DISCUSSION

Although over the years a large number of cancer-specific databases (e.g. Human Lung Cancer Database, Prostate Gene Database, Oral Cancer Gene Database and Breast Cancer Gene Database) have been published, yet, to date to our knowledge, there is no resource available that provides detailed information about the genes known to be associated with cervix cancer. Therefore we have compiled the first CCDB, where manual curation along with information from other resources have been integrated to provide a knowledgebase that will allow researchers and clinicians to get an overview of biology of the genes involved in cervix cancer. We hope that the availability of this database would save time and effort of researchers involved in the field and thus will facilitate the biological discovery process.

GO terms have been used previously to characterize protein function and to elucidate trends in protein datasets. We too classified all cervix-cancer-related genes according to the molecular function of each protein and the biological process in which it is involved. Assignment of 537 genes to various molecular functions revealed that the top five categories represented in the dataset are: transcription factor activity, transcription regulatory activity, cell adhesion molecule activity, receptor activity and DNA binding, thereby suggesting importance of these gene products for developmental pathway of cervical cancer cells. Also, we observed that the biological processes that are enriched in CCDB are (i) cell communication and signal transduction pathway: 26%, (ii) regulation of nucleobase, nucleoside, nucleotide and nucleic acid

Downloaded from nar.oxfordjournals.org by guest on April 28, 2011



Gene Name	AKT1
Gene ID	207
HPRD ID	HPRD:01261
Molecular Class	Serine threonine kinase
Molecular Function	Protein serine threonine kinase activity
Biological Process	Cell communication, Signal transduction

Gene Name	AKT1
Gene ID	207
Contig	NC_000014.8
Location	105235686..105262080 complement
OMIM ID	164730
HGNC ID	391
UNIGENE ID	525622
ENSEMBL ID	ENSG00000142208
HPRD ID	HPRD:01261
PharmGKB ID	PA24684
Homologene ID	3785
References	Pubmed IDs
mRNA/CDS/Protein	Details and Sequences

Figure 2. Schematic workflow of CCDB.

metabolism: 17% and (iii) protein metabolism: 12%. This suggests that these biological processes may be involved in transition of cell progression to cancerous state.

Further homology information is crucial as it reveals how conserved a protein has remained through evolution, and hence the degree to which it can tolerate alteration within the sequence. Using the data of 537 genes involved in cervix cancer we find most of the proteins exhibit homology to other eukaryotic genomes except six. Although almost all the proteins exhibit homologous sequences in other genomes, majority (39%) are conserved in *Euteleostomi* and only a small fraction are conserved in eukaryotic crown group (16%), i.e. they are highly conserved across the eukaryotic domain of life. The presence of a large proportion of CCDB genes within *Euteleostomi* group suggests that these genes are present in fish species and have arisen recently on the evolutionary line and thus, it may be possible that these proteins are responsible for distinguished characteristics.

Also it is a well-established fact that hypermethylation of tumor-suppressor genes has long been associated with various cancers, including cervical cancer. In the present version of CCDB, we present evidence for 69 genes that are hypermethylated in cervical cancer, much more than the PubMeth database (14). PubMeth is an annotated and reviewed database of methylation in cancer, which catalogs information for 36 genes involved in various stages of cervical cancer. However, presently our database is not supported by as many literature references as documented in PubMeth. Therefore, we feel that our database will complement the existing database in serving scientific community.

Also, another advantage of this database will be that it allows researchers to make a thorough comparison of genes that are common in most of the cancers as well as detect the ones that are unique and demonstrate aberrant behavior in cervix cancer. For example, comparison of the present set of CCDB genes with two other databases (human lung cancer database and prostate gene database) revealed that there are 293 and 35 genes common among these cancers respectively.

CURRENT STATUS AND FUTURE DEVELOPMENTS

The current release (v. 1.0) of CCDB contains 537 unique genes that are supported by evidence from unique Pubmed

Downloaded from nar.oxfordjournals.org by guest on April 28, 2011

records. The data are presented in a systematic way and apart from search facility, several browsing options facilitate fast, efficient and user-friendly retrieval of information. We propose to update this database on a regular basis adding new data from literature as well as other data analysis tools that will help in improving our knowledge about cervix cancer and potentially contribute to the development of novel therapeutic strategies. Since not all genes published in literature are yet included in the current release of the database, our foremost objective would be to identify, collect and add those genes. Also, another limitation of the present version is that not all the related references for a given gene have been covered. Therefore, in the next update we plan to add more literature references for genes in the current version. We also propose to extend search facility and include information about transcription factor binding site. Overall, we aim to increase the quality of data and to supply additional database function.

SUMMARY

In summary, CCDB is a first attempt to provide a comprehensive non-redundant catalog of genes involved in the cervix cancer along with information about the altered biological process and supporting evidence from published literature. It is expected that availability and use of CCDB will reduce the time and effort of scientists and clinicians to survey the literature on genes and their involvement in cervix diseases. It thus provides a unique value-added resource in the field of cervix cancer biology.

FUNDING

Funding for open access charge: Institute of Cytology and Preventive Oncology, Indian Council of Medical Research.

Conflict of interest statement. None declared.

REFERENCES

1. Phongsavan,K., Phengsavanh,A., Wahlström,R. and Marions,L. (2010) Women's perception of cervical cancer and its prevention in rural laos. *Int. J. Gynecol. Cancer.*, **20**, 821–826.
2. Wang,X., Liu,R., Ma,B., Yang,K., Tian,J., Jiang,L., Bai,Z.G., Hao,X.Y., Wang,J., Li,J. *et al.* (2010) High dose rate versus low dose rate intracavity brachytherapy for locally advanced uterine cervix cancer. *Cochrane Database Syst. Rev.*, **7**, CD007563.
3. Hakim,A.A., Lin,P.S., Wilczynski,S., Nguyen,K., Lynes,B. and Wakabayashi,M.T. (2010) Indications and efficacy of the human papillomavirus vaccine. *Curr. Treat Options Oncol.*, **8**, 393–401.
4. Satyaprakash,A.K. and Tyring,S.K. (2010) Human papillomaviruses vaccine: a dermatologic perspective. *Indian J. Dermatol. Venereol. Leprol.*, **76**, 14–19.
5. Manavi,M., Hudelist,G., Fink-Retter,A., Gschwandtler-Kaulich,D., Pischinger,K. and Czerwenka,K. (2007) Gene profiling in Pap-cell smears of high-risk human papillomavirus-positive squamous cervical carcinoma. *Gynecol. Oncol.*, **105**, 418–426.
6. Hagemann,T., Bozanovic,T., Hooper,S., Ljubic,A., Slettenaar,V.I., Wilson,J.L., Singh,N., Gayther,S.A., Shepherd,J.H. and Van Trappen,P.O. (2007) Molecular profiling of cervical cancer progression. *Br. J. Cancer.*, **96**, 321–328.
7. Wang,L., Xiong,Y., Sun,Y., Fang,Z., Li,L., Ji,H. and Shi,T. (2010) HLungDB: an integrated database of human lung cancer research. *Nucleic Acids Res.*, **38**, D665–D669.
8. Li,L.C., Zhao,H., Shiina,H., Kane,C.J. and Dahiya,R. (2003) PGDB: a curated and integrated database of genes related to the prostate. *Nucleic Acids Res.*, **31**, 291–293.
9. Levine,A.E. and Steffen,D.L. (2001) OrCGDB: a database of genes involved in oral cancer. *Nucleic Acids Res.*, **29**, 300–302.
10. Baasiri,R.A., Glasser,S.R., Steffen,D.L. and Wheeler,D.A. (1999) The Breast Cancer Gene Database: a collaborative information resource. *Oncogene*, **18**, 7958–7965.
11. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.
12. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Ongenaert,M., Van Neste,L., De Meyer,T., Menschaert,G., Bekaert,S. and Van Criekinge,W. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**, D842–D846.

Downloaded from nar.oxfordjournals.org by guest on April 28, 2011