Dragon Database for Exploration of Ovarian Cancer Genes

Database Overview

Introduction

Ovarian cancer (OC) is the leading cause of death among gynecological malignancies and represents the fifth leading cause of cancer-related death in women. Statistics indicate that at the time of diagnosis, the cancer has already metastasized beyond the ovary in approximately 70% of patients and only 30% of patients with this advanced-stage OC survive five years after the initial diagnosis (1). The lethality of ovarian carcinoma primarily stems from the inability to detect the disease during the early organ-confined stage, combined with the lack of effective therapies for advanced-stage disease. In patients with metastasized OC, most relapse and ultimately die due to the development of drug resistance (2). Efforts made by the scientific community to improve the survival rate associated with OC have resulted in a wealth of scattered research data that makes it difficult for researchers to select the relevant information.

To support current research endeavors in OC, we have developed an integrated knowledge database called **D**ragon **D**atabase for Exploration of **O**varian **C**ancer Genes (DDOC). DDOC provides a comprehensive compilation of the published research related to the genes associated with OC. Many aspects of the information provided in the DDOC were curated by biologists, which emphasize its accuracy compared to databases that are populated in an automated fashion. Taking into account that experimental conditions influence gene expression, DDOC provides details of the cell line, tissue or cell type, expression status, disease stage, tumor grade, OC type and laboratory method provided in the literature. We have provided links to the relevant sources of data used to extract information related to genes included in DDOC. DDOC is freely accessible for academic and non-profit users at http://apps.sanbi.ac.za/ddoc/index.php.

Methods and Implementation

The methods used to collect and annotate data, as well as the structure of the database, are summarized below. The data was populated on 18th May, 2008.

1) Data Collection

The inclusion of a gene into DDOC is based on results of curation process. To be included in DDOC, the gene must satisfy the following criteria:

(i) Gene expression must be experimentally confirmed in ovarian cancer tissue using techniques such as; RT-PCR, immunohistochemistry, western blotting or FISH (Fluorescent In Situ Hybridization). Gene documented as having OC linked SNP were also included in the database.

(ii) Genes identified using microarray technology were not included at this stage since microarray provides initial evidence about expression of gene in certain tissue or cell type, but there are limitations associated with analysis of microarray data. The results obtained using high throughput technologies are debatable in terms of deciding about a meaningful level of differential expression and statistical methods used for analysis and interpretation of data (3;4). Therefore, it was decided to include only the genes experimentally confirmed as described in (i). We plan to incorporate at later stage the genes showing differential expression in OC specific microarray analysis.

The gene related information provided in the database was collected from various repositories. Initially a list of 900 genes was collected from sources like Cancer Gene Census (5) (http://www.sanger.ac.uk/genetics/CGP/Census/), GeneCards (6) (http://www.genecards.org/index.shtml), SymAtlas (7), OMIM (8) (Online Mendelian Inheritance in Man, 2007) (http://www.ncbi.nlm.nih.gov/), Ovarian Kaleidoscope Database (9) (http://ovary.stanford.edu/), Entrez Gene (10) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) and GenAtlas (11) (http://www.genatlas.org/). A literature search was conducted by the biologists in our team and this process resulted in a final list of 379 genes which was used to populate the database.

Table 1- Summary of statistics of various features linked with genes included in DDOC

Features	Number of genes
Reactome pathways	50
PDB IDs	184
Previous Symbols	193
Previous Name	199
KEGG pathways	211
OMIM IDs	324
Aliases (for gene symbols)	346
eVOC	353
Gene ontology annotations (GO)	367
Unigene IDs	369
Ensembl IDs	370
RefSeq peptide	370
EMBL IDs	372
RefSeq mRNA	372
UniProt IDs	372
HGNC IDs	374
GenBank IDs	374
Total number of genes in DDOC	379
Gene symbols, gene names and Entrez Gene IDs	379
TFs used for prediction of TFBS(according to Transfac IDs)	1449
PubMed abstracts for text-mining	588,727

Gene Identifiers

The general information about the genes which include, HGNC ID, approved symbol, approved name, entrez ID, previous symbol, previous name, Aliases, OMIM and chromosome location were extracted from sources such as HUGO (12) (http://www.genenames.org/) and GeneCards (6) (http://www.genecards.org/index.shtml). Other gene related identifiers were also compiled. Ones included in the database are EMBL (13) (http://www.ebi.ac.uk/embl/), Ensembl (14) (http://www.ensembl.org/index.html), Refseq (15), Genbank (16) (http://www.ncbi.nlm.nih.gov/), Unigene (17) http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene&orig_db), Uniprot (18) http://www.ebi.ac.uk/uniprot/), Swiss-Prot (19) (http://www.expasy.ch/sprot/) and PDB (20) (http://www.rcsb.org/pdb/home/home.do). ID conversion tools like IDconvertor (21) (http://idconverter.bioinfo.cnio.es/) and Onto-tools (22) (http://vortex.cs.wayne.edu/ontoexpress/servlet/UserInfo) were used to convert among different types of identifiers.

2) Annotation of incorporated genes

In order to support uniformity in research, genes in DDOC were annotated using classification systems adopted from controlled vocabularies like GO (23) (http://www.geneontology.org/) and eVOC ontologies (24;25). eVOC provides a platform that permits the evaluation of differentially expressed genes in different tissues based on EST, SAGE and microarray data. The public version of eVOC is available at (http://www.evocontology.org/). We annotated the genes using following eVOC categories: anatomical system, associated with, cell type, development stage, experimental technique, microarray platform, pathology, taxonomy, tissue preparation and treatment. Information about the genes being implicated in other diseases has also been included in the database.

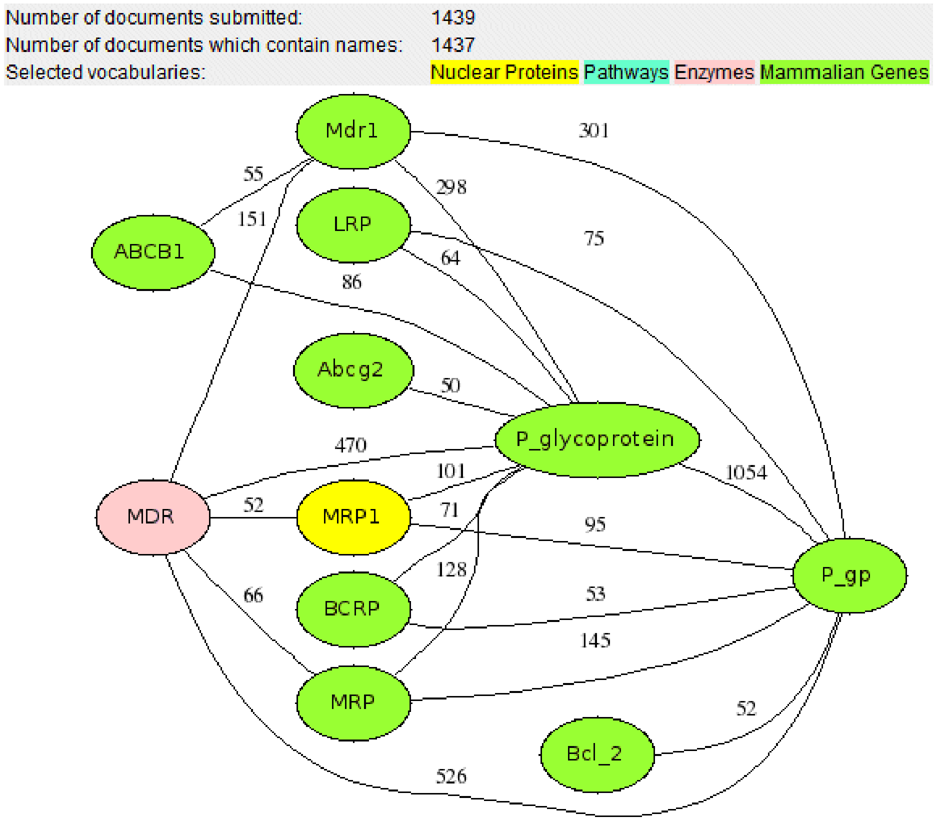
In addition to the ontology classification, we also collected information about the pathways in which the genes are involved in order to give the user a deeper insight into the mechanisms associated with the genes expression. The genes were mapped to pathways via KEGG (26) (http://www.genome.jp/kegg/), REACTOME (27) (http://www.reactome.org/) and TRANSPATH (28) (http://www.transpath.com). Live links are provided for each pathway. TRANSPATH is a commercial database provided by BIOBASE and therefore access is limited to users having subscription to TRANSPATH.

Another type of information included is about the orthology of genes. The sub-menu 'ortholog genes' (http://   ) displays information about homologs of a gene for four species namely *P. troglodytes*, *C. lupus familiaris*, *M. musculus* and *R. norvegicus* from Homologene (17) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene).

Information stored in the database was integrated with the information obtained from literature research. The literature search was sourced from the National Center for Biotechnology Information (NCBI) PubMed database (http://www.ncbi.nlm.nih.gov) via a query data tool that our group develped based on the NCBI "Entrez Programming Utilities". The NCBI database was queried for each gene using the following keywords:

("Gene Symbol" OR "Gene Alias" OR "Gene Alias", ...) AND mammal AND cancer.

These queries produced list of 588,727 abstracts that were loaded into the "Dragon Biomedical Text Miner" of OrionCell (http://www.orioncell.org). This text-mining tool represents a complete redeveloped and significantly enhanced system based on the concepts published initially in Dragon Plant Biology Explorer (29) and Dragon TF Association Miner (30). By this tool we indexed the text document by vocabularies for Nuclear Proteins, Pathways, Enzymes and Mammalian Genes, and produced a database of associations that is integrated into DDOC.



Number of documents submitted: 1439
Number of documents which contain names: 1437
Selected vocabularies: Nuclear Proteins Pathways Enzymes Mammalian Genes

Fig 1 Screenshot of the interactive network generated using text-mining. The color coded vocabulary entries are interconnected with weighted links representing frequency of appearance of a term and its neighbors in our abstract list. By clicking on the node the user will retrieve the abstracts containing the selected term, and the associated terms.

3) Database structure

The DDOC is based on the three-tier (layer) (data, logic and presentation) architecture (Fig 2). The presentation layer is web-based and implemented in DHTML and Javascript. The logic layer was implemented as a number of server side PHP and Perl modules interfaced with the data layer. Data layer is MySQL, and for the text-mining purposes, file system based. The relational database design strictly distinguishes between tables that contain data entities and tables that establish logical connections between these data entities. The central data entity is the gene, to which most other data entities are linked. Other important data entities are transcription related such as transcription start sites (TSSs) and transcription factors (TFs). This is reflected in the two entry points that a user can chose between on the top level of the web-interface (Genes http:// …. vs. Transcription http://...).

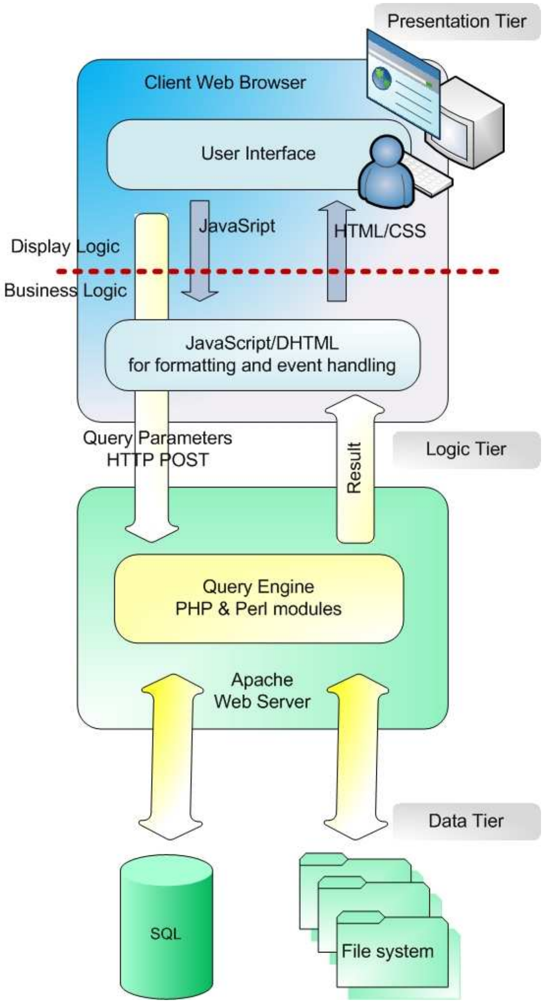
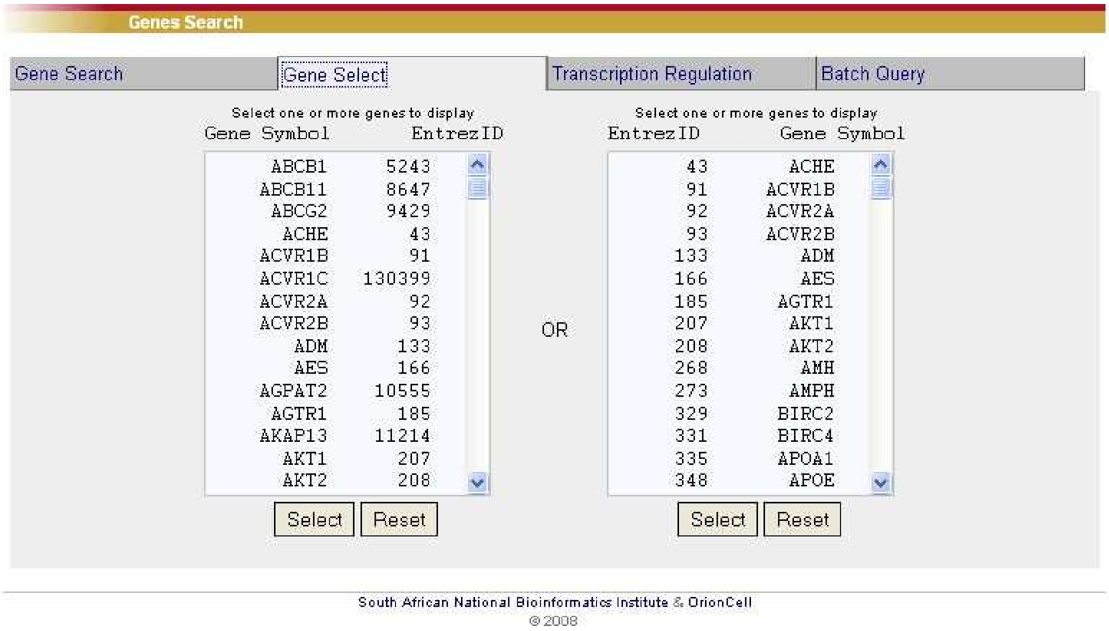


Fig 2 The schematic representation of the DDOC structure.

4) Browsing and searching DDOC

The front page of the database has three different menu types: Gene search, Gene select and Transcription regulation..

a/ Gene Select: In gene select option, the genes available in DDOC are displayed in an alphabetical order. Information related to a particular gene or genes can be searched by highlighting the genes of interest and clicking select button (Fig 3). The window will display the general information about the gene (Fig 4) and the gene can be further explored by clicking on the details icon. The window displayed next is the 'Gene Details' window (Fig 5) and it summarizes the relevant incorporated details of the genes which has been classified into several self explanatory sub-menus like General Information, Gene in other resources, Experimental evidence, Related Proteins, eVOC Ontologies, Gene Ontologies, Associated Pathways, Associated Diseases, Ortholog Genes, Regulations and Text mining reports. Required information can be displayed in the same window by selecting any of these sub-menus. To find out more about DDOC and trouble shooting users are referred to FAQ page (http://apps.sanbi.ac.za/ddoc/faq.php).



Select one or more genes to display

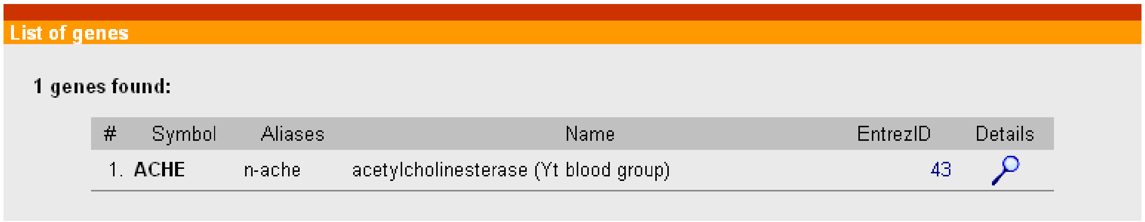
Gene Symbol	EntrezID
ABCB1	5243
ABCB11	8647
ABCG2	9429
ACHE	43
ACVR1B	91
ACVR1C	130399
ACVR2A	92
ACVR2B	93
ADM	133
AES	166
AGPAT2	10555
AGTR1	185
AKAP13	11214
AKT1	207
AKT2	208

Select one or more genes to display

EntrezID	Gene Symbol
43	ACHE
91	ACVR1B
92	ACVR2A
93	ACVR2B
133	ADM
166	AES
185	AGTR1
207	AKT1
208	AKT2
268	AMH
273	AMPH
329	BIRC2
331	BIRC4
335	APOA1
348	APOE

South African National Bioinformatics Institute & OrionCell
© 2008

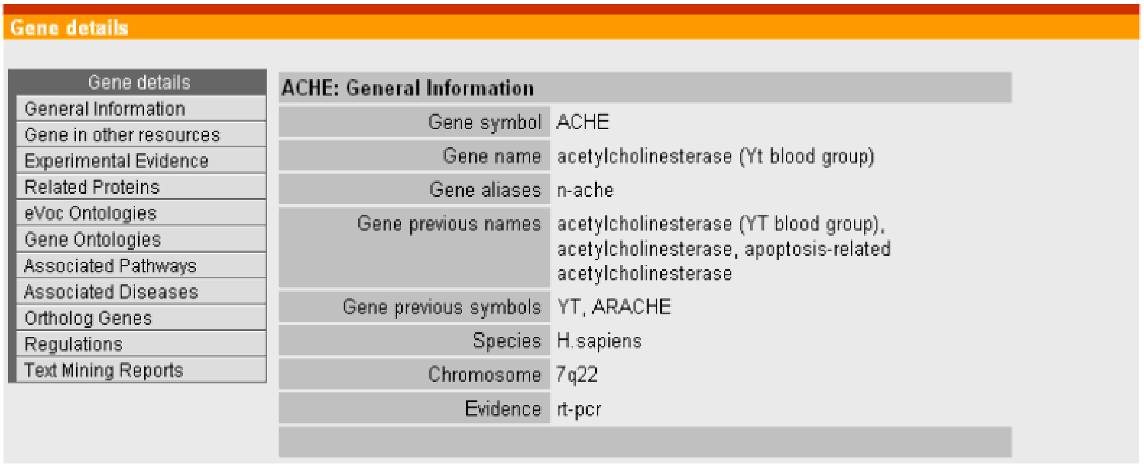
Fig 3 Screenshot of the 'Gene select' option in DDOC



1 genes found:

#	Symbol	Aliases	Name	EntrezID	Details
1.	ACHE	n-ache	acetylcholinesterase (Yt blood group)	43	🔍

Fig 4 Screenshot of the 'list of genes' page in DDOC

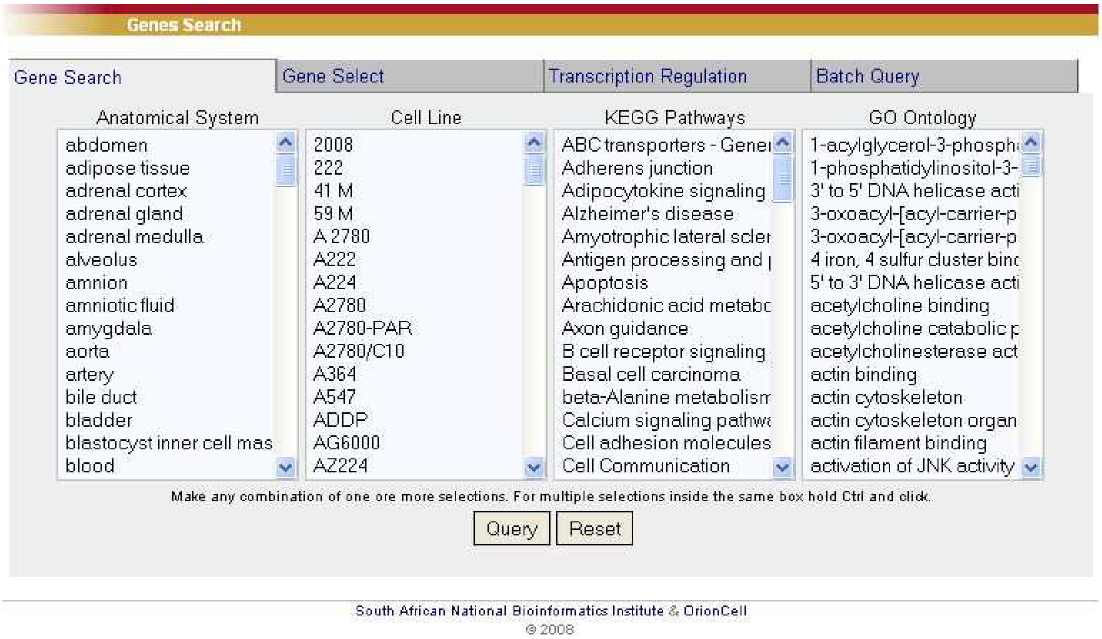


ACHE: General Information

Gene symbol	ACHE
Gene name	acetylcholinesterase (Yt blood group)
Gene aliases	n-ache
Gene previous names	acetylcholinesterase (YT blood group), acetylcholinesterase, apoptosis-related acetylcholinesterase
Gene previous symbols	YT, ARACHE
Species	H.sapiens
Chromosome	7q22
Evidence	rt-pcr

Fig 5 Screenshot of the 'Gene details' page in DDOC

b/ **Gene Search:** This option adds maximum flexibility and functionality for the biologists. Four categories are included in this option (Fig 6). The anatomical system (where the gene is expressed), cell line in which the genes were studied with respect to OC, KEGG pathways and GO categories related to the genes. User can select one or multiple categories to display the list of genes involved in the category/categories respectively. Multiple selection will retrieve the list of genes involved in all the processes and categories selected by user. For example- if one wants to see the genes involved in cell cycle as well as transcription, one should select the 'cell cycle' form pathways and 'transcription' from GO ontologies and press query. System will retrieve and display the list of genes involved in both types of processes and each of these genes can be further queries for detailed analysis.

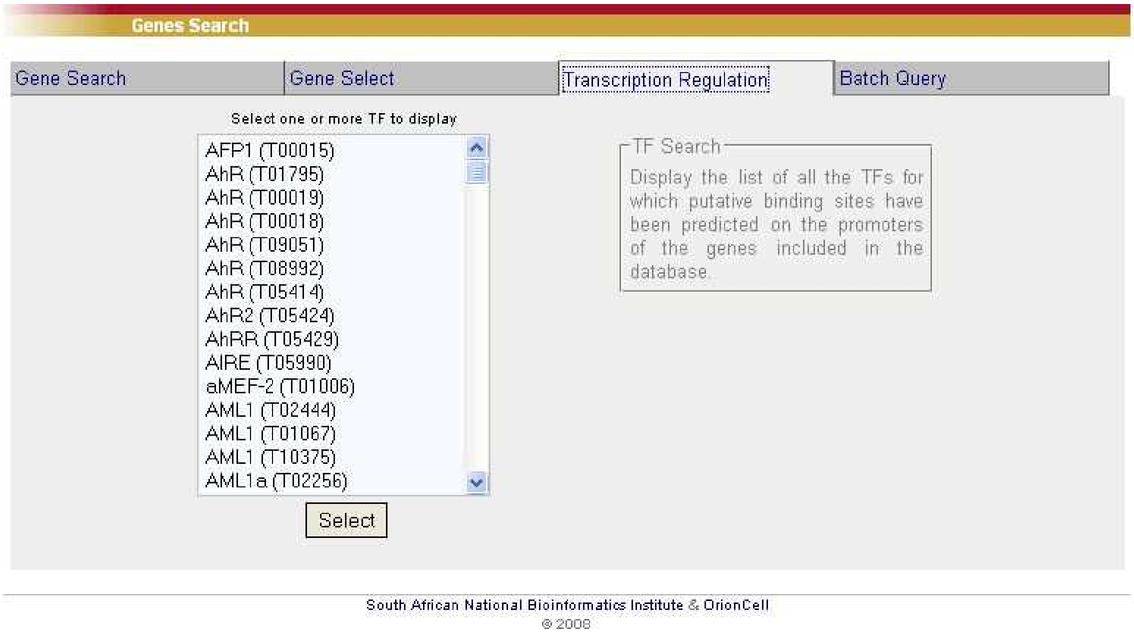


Make any combination of one ore more selections. For multiple selections inside the same box hold Ctrl and click.

South African National Bioinformatics Institute & OrionCell
© 2008

Fig 6 Screenshot of the 'Gene search' option of DDOC

c/ **Transcription regulation:** This option displays the list of all the TFs for which putative binding sites have been predicted on the promoters of the genes included in the database (Fig 7). User can highlight any TF of interest, for example AhRR and click the select button. The system will lead the user to the page where all the genes having binding sites for that particular TF are displayed (Fig 8). Need to explain method to predict TFBS.

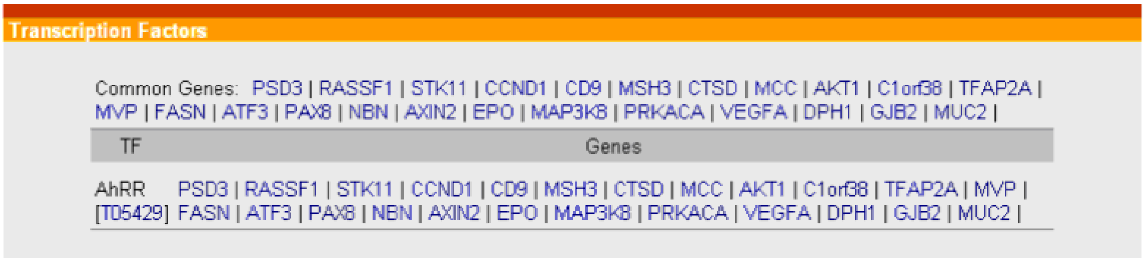


AFP1 (T00015)
AhR (T01795)
AhR (T00019)
AhR (T00018)
AhR (T09051)
AhR (T08992)
AhR (T05414)
AhR2 (T05424)
AhRR (T05429)
AIRE (T05990)
aMEF-2 (T01006)
AML1 (T02444)
AML1 (T01067)
AML1 (T10375)
AML1a (T02256)

Display the list of all the TFs for which putative binding sites have been predicted on the promoters of the genes included in the database.

South African National Bioinformatics Institute & OrionCell
© 2008

Fig 7 Screenshot of the 'Transcription regulation' option of DDOC



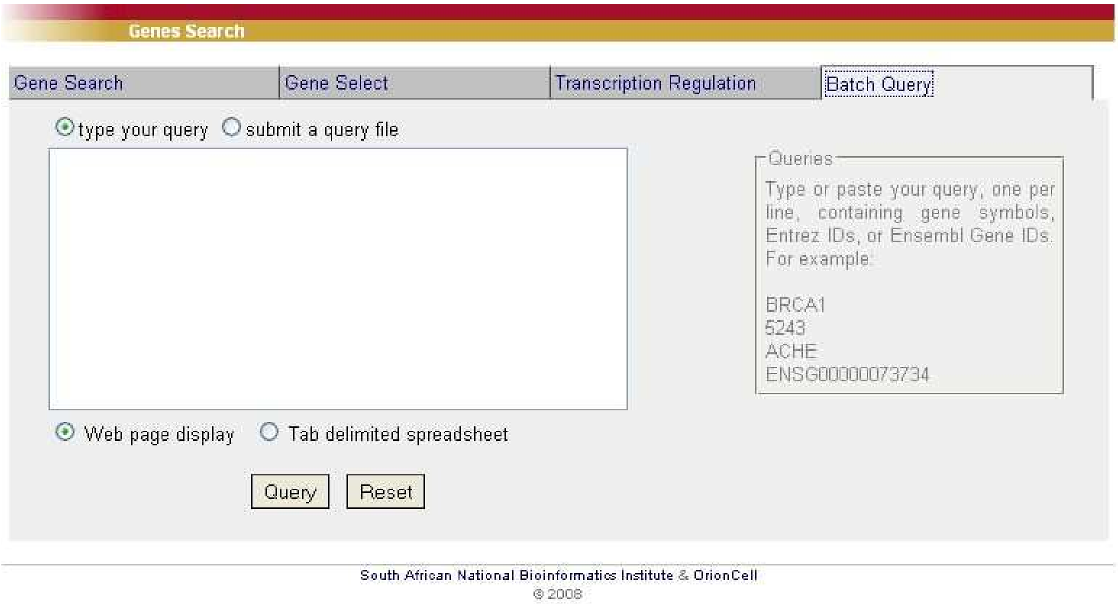
Common Genes: PSD3 | RASSF1 | STK11 | CCND1 | CD9 | MSH3 | CTSD | MCC | AKT1 | C1orf38 | TFAP2A | MVP | FASN | ATF3 | PAX8 | NBN | AXIN2 | EPO | MAP3K8 | PRKACA | VEGFA | DPH1 | GJB2 | MUC2 |

TF	Genes
AhRR [T05429]	PSD3 | RASSF1 | STK11 | CCND1 | CD9 | MSH3 | CTSD | MCC | AKT1 | C1orf38 | TFAP2A | MVP | FASN | ATF3 | PAX8 | NBN | AXIN2 | EPO | MAP3K8 | PRKACA | VEGFA | DPH1 | GJB2 | MUC2 |

Fig 8 Screenshot of the page showing list of genes having binding sites for selected TF

d) Batch Query

This option allows the user to extract the required information for selected set of genes from the database (Fig 9). The user can submit a list of mixed identifiers_(Entrez Gene IDs, Gene Symbols and Ensembl IDs) as shown in Fig 9 (a). Moreover, the user can select any type of information he/she may want to see in the reports (Fig 9 (b)) and the system will extract that information in the 'tsv' format , which is a portable format in excel in windows, and csv in linux systems.



● type your query ○ submit a query file

Queries

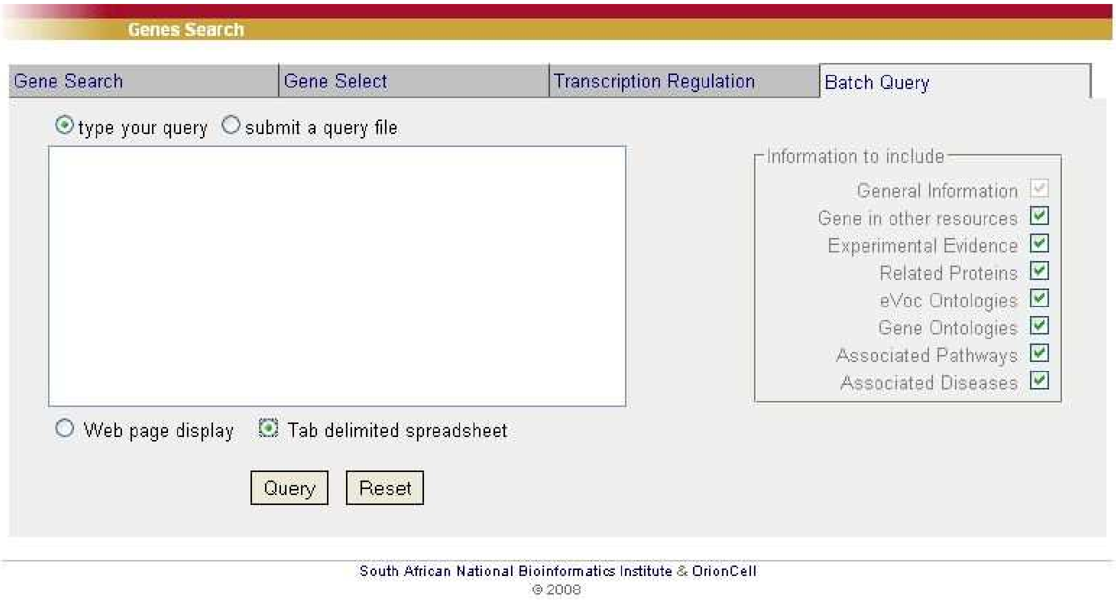
Type or paste your query, one per line, containing gene symbols, Entrez IDs, or Ensembl Gene IDs. For example:

BRCA1
5243
ACHE
ENSG00000073734

● Web page display ○ Tab delimited spreadsheet

South African National Bioinformatics Institute & OrionCell
© 2008

Fig 9 (a) Screenshot of the 'Batch Query' option



● type your query ○ submit a query file

Information to include

General Information ☑
Gene in other resources ☑
Experimental Evidence ☑
Related Proteins ☑
eVoc Ontologies ☑
Gene Ontologies ☑
Associated Pathways ☑
Associated Diseases ☑

○ Web page display ● Tab delimited spreadsheet

South African National Bioinformatics Institute & OrionCell
© 2008

Fig 9 (b) Screenshot of the 'Batch Query' option

5) Data visualization

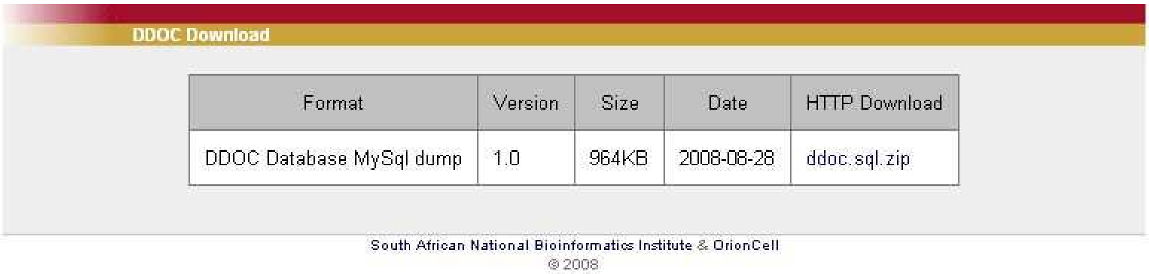
The text mining results have been presented in the form of networks which helps the user to view the interactions of the gene of interest with other biological entities, genes and pathways described in literature. The networks can be generated by selecting the number of abstracts in which the terms appeared together. This is a convenient and efficient method to explore the huge amounts of literature in few seconds time and visualize the important interactions among different terms in an easy to follow graphical representations (for example - Fig 1).

6) Retrieval of data from DDOC

User can download as "*csv*" file the information contained in "regulations and gene ontology" menus. The lists of genes displayed by the system can directly be copied and pasted in excel sheets. All other types of information provided is linked with resources which can be directly accessed to download relevant information.

As described earlier, we have provided 'Batch query' option in the search menu and user can select different types of information to be extracted from DDOC. This facility is available for downloading of all the available categories within DDOC. We will provide the search option using 'chromosome' and 'ovarian cancer type' during updates of the database.

We have also provided the MySql dump of DDOC (Fig 10). The data is available at 'Download' page (http://apps.sanbi.ac.za/ddoc/download.php) and is accessible to all the users, and can be used to integrate data into other resources. Moreover, users can make any usual programmatic access to DDOC using wrappers.



DDOC Download

Format	Version	Size	Date	HTTP Download
DDOC Database MySql dump	1.0	964KB	2008-08-28	ddoc.sql.zip

South African National Bioinformatics Institute & OrionCell
© 2008

Fig 10 Screenshot of the 'DDOC Download' page

Utility of DDOC

What type of queries can be performed using DDOC?

Scientists and clinicians can use DDOC to query questions such as:

i) Is my gene of interest expressed in OC and under which experimental conditions has it been verified?

ii) Which pathway has my gene of interest been implicated in and have the other genes in these pathways been verified as being expressed in OC under the same conditions?

iii) Which transcription factors influences the expression of my gene of interest and which other genes' expression are likely transcribed by the same transcription factors?

iv) Are there any orthologs for my gene of interest and what is the chromosomal location and species specific IDs of orthologous genes?

v) What are the different sites of expression for my gene of interest and what functional annotations have been assigned to it?

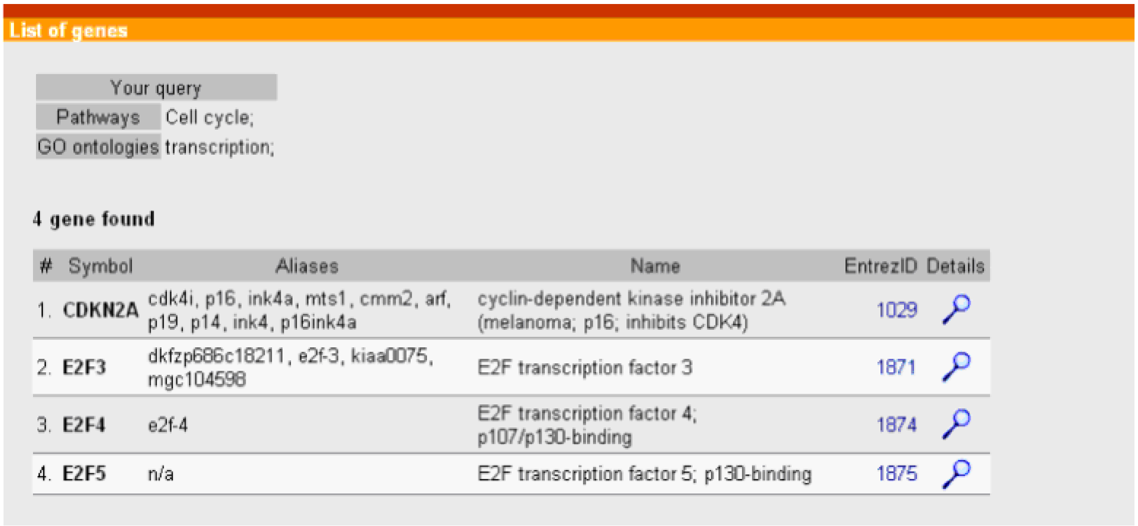
vi) Is there any way I can look out for interactions my gene of interest can have with other genes and proteins?

Examples of use

Question: I am interested in studying genes involved in cell cycle pathway, particularly with a role in transcription and also like to study their regulation and biological interactions.

Solution: The step-wise analysis is detailed below:

1. Go to the 'Gene search' page (http://apps.sanbi.ac.za/ddoc/search.php).
2. In *Gene search* tab, select the '*cell cycle*' pathway and query.
3. The resulting window will display the list of 25 genes involved in cell cycle pathway.
4. Since I am interested in genes with role in transcription and cell cycle, using the functions provided in the database, the genes having both these roles can be extracted without much effort. For this, simply select the '*cell cycle*' pathway and using the cntl key select '*transcription*' from GO ontology tab and query.
5. This exercise will result in the list of four genes, which are involved in cell cycle as well as in transcription (Fig 11).
6. The details of each resulting gene can be explored individually.
7. Let us take E2F3 as an example from the above list. The 'Gene details' page has all the information related to E2F3. The 'regulations' tab will display the list of transcription factors for which putative binding sites have been predicted on the promoter region of E2F3. This gives us an idea about the regulation of E2F3 gene at molecular level and provides a narrow list of TFs for further experimental evaluation, if interested. The binding sites are putative and should not be treated as a proof of regulation by respective TFs.



List of genes

Your query	
Pathways	Cell cycle;
GO ontologies	transcription;

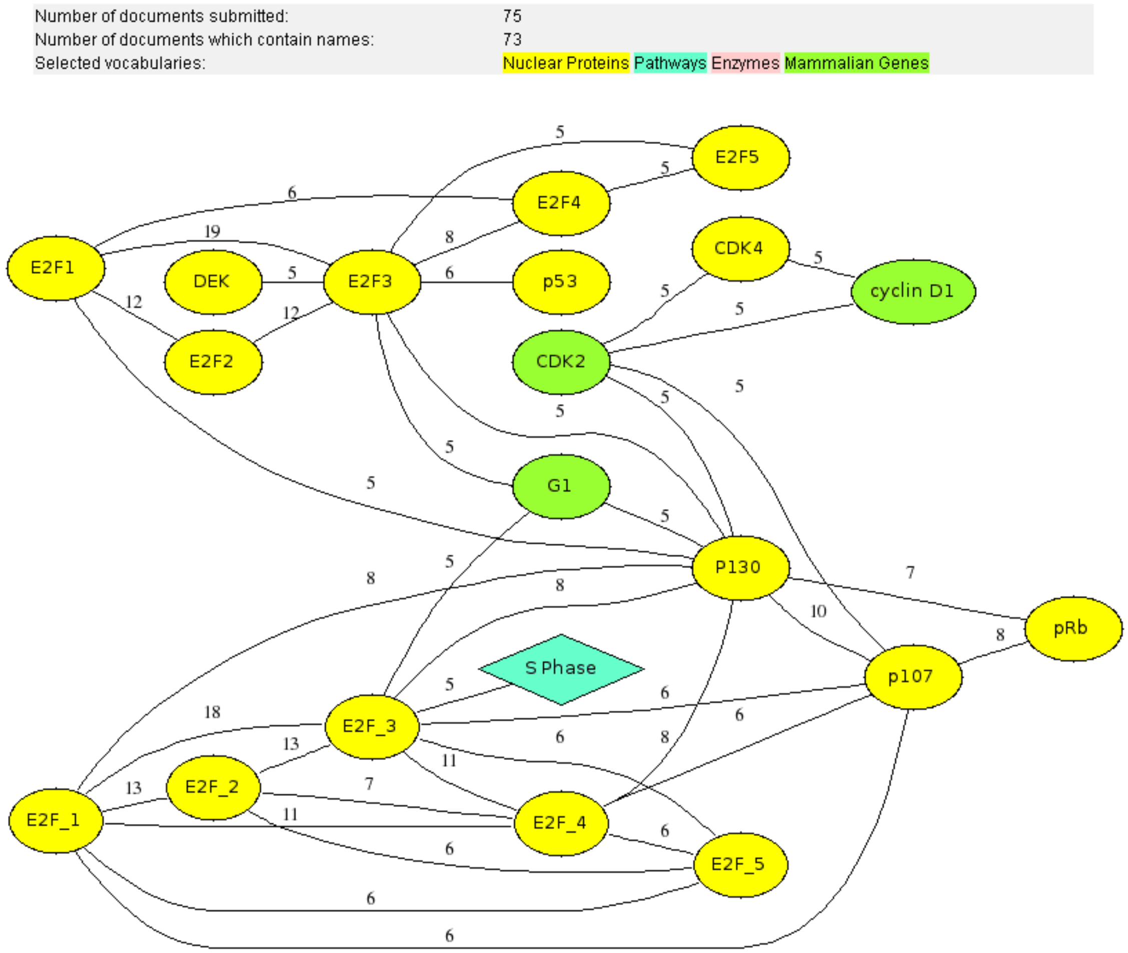
4 gene found

#	Symbol	Aliases	Name	EntrezID	Details
1.	CDKN2A	cdk4i, p16, ink4a, mts1, cmm2, arf, p19, p14, ink4, p16ink4a	cyclin-dependent kinase inhibitor 2A (melanoma; p16; inhibits CDK4)	1029	🔍
2.	E2F3	dkfzp686c18211, e2f-3, kiaa0075, mgc104598	E2F transcription factor 3	1871	🔍
3.	E2F4	e2f-4	E2F transcription factor 4; p107/p130-binding	1874	🔍
4.	E2F5	n/a	E2F transcription factor 5; p130-binding	1875	🔍

Fig 11 Screenshot of 'list of genes' involved in cell cycle pathways as well as in transcription.

8. The interactions predicted in the above step can be further analyzed using text-mining tool. Text-mining helps to deduce the types of interactions each gene might have at

biological level. We have predicted binding sites for E2F4 on the regulatory region of E2F3 gene. The text mining results show that E2F4 has appeared eight and 11 times with E2F3 (Fig 12) in the literature (depending on the symbol used).



Number of documents submitted:	75
Number of documents which contain names:	73
Selected vocabularies:	Nuclear Proteins Pathways Enzymes Mammalian Genes

Fig 12 Screenshot of the network generated using text-mining tool

The appearance of E2F3 and E2F4 in same abstract does not mean that E2F4 is a regulator of E2F3, rather it shows that E2F3 and E2F4 along with E2F5 are involved in similar tumor related processes (cell cycle- S phase) and are regulated by similar kind of factors pRb, p53 and p130 etc. By clicking on each entity in the network, a list of relevant PubMed abstracts is opened, which demonstrate the relationships as presented in the network.

Conclusions: Several conclusions can be drawn from above exercise:

a/ The list of genes involved in cell cycle as well as transcription (i.e. the genes responsible for control of cell cycle at transcription level).

b/ The list of putative regulators of the genes in question. This will allow us to study the molecular processes involved in regulation of cell cycle pathway and a network of all the genes and factors involved in transcriptional control of cell cycle can be generated.

c/ The system provide support (literature and experimental) for the genes under study and various observed biological interactions can be confirmed using literature provided, that certainly speeds up the process and enhances deeper understanding of the biological mechanisms by the use of simple but effective tools.

d/ The thorough analysis could help to identify new targets for future evaluations. The associations for three (E2F3, E2F4 and E2F5) out of four genes (Fig 11) are demonstrated using text-mining (Fig 12). The interactions with other mammalian genes cyclinD1 and CDK2 have also been observed which can be studied in more details, thus providing the clues for future investigations and also helps to draw known interations in an easy to interpret visual representations of the data.

Reference List

1. Jemal,A., Tiwari,R.C., Murray,T., Ghafoor,A., Samuels,A., Ward,E., Feuer,E.J. and Thun,M.J. (2004) Cancer statistics, 2004. *CA Cancer J Clin.*, **54**, 8-29.

2. Agarwal,R. and Kaye,S.B. (2003) Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nat.Rev.Cancer*, **3**, 502-516.

3. Pritchard,C.C., Hsu,L., Delrow,J. and Nelson,P.S. (2001) Project normal: defining normal variance in mouse gene expression. *Proc.Natl.Acad.Sci U.S.A*, **98**, 13266-13271.

4. Smyth,G.K., Yang,Y.H. and Speed,T. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol.Biol*, **224**, 111-136.

5. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat.Rev.Cancer*, **4**, 177-183.

6. Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben Dor,U., Esterman,N., Rosen,N., Peter,I. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542-1543.

7. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc.Natl.Acad.Sci U.S.A*, **101**, 6062-6067.

8. Baxevanis,A.D. (2003) Searching Online Mendelian Inheritance in Man (OMIM) for information for genetic loci involved in human disease. *Curr.Protoc.Hum.Genet.*, **Chapter 9**, Unit9.

9. Leo,C.P., Vitt,U.A. and Hsueh,A.J. (2000) The Ovarian Kaleidoscope database: an online resource for the ovarian research community. *Endocrinology*, **141**, 3052-3054.

10. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, **35**, D26-D31.

11. Frezal,J. (1998) Genatlas database, genes and development defects. *C R.Acad.Sci III*, **321**, 805-817.

12. Eyre,T.A., Ducluzeau,F., Sneddon,T.P., Povey,S., Bruford,E.A. and Lush,M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res*, **34**, D319-D321.

13. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L., Browne,P. *et al.* (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*, **35**, D16-D20.

14. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res*, **36**, D707-D714.

15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61-D65.

16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res*, **36**, D25-D30.

17. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **36**, D13-D21.

18. (2008) The universal protein resource (UniProt). *Nucleic Acids Res*, **36**, D190-D195.

19. Gasteiger,E., Jung,E. and Bairoch,A. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr.Issues Mol.Biol*, **3**, 47-55.

20. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, **35**, D301-D303.

21. Alibes,A., Yankilevich,P., Canada,A. and Diaz-Uriarte,R. (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics*, **8**, 9.

22. Khatri,P., Desai,V., Tarca,A.L., Sellamuthu,S., Wildman,D.E., Romero,R. and Draghici,S. (2006) New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate. *Nucleic Acids Res*, **34**, W626-W631.

23. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat.Genet.*, **25**, 25-29.

24. 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.

25. Kruger,A., Hofmann,O., Carninci,P., Hayashizaki,Y. and Hide,W. (2007) Simplified ontologies allowing comparison of developmental mammalian gene expression. *Genome Biol*, **8**, R229.

26. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res*, **30**, 42-46.

27. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*, **8**, R39.

28. Krull,M., Pistor,S., Voss,N., Kel,A., Reuter,I., Kronenberg,D., Michael,H., Schwarzer,K., Potapov,A., Choi,C. *et al.* (2006) TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res*, **34**, D546-D551.

29. Bajic,V.B., Veronika,M., Veladandi,P.S., Meka,A., Heng,M.W., Rajaraman,K., Pan,H. and Swarup,S. (2005) Dragon Plant Biology Explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists. *Plant Physiol*, **138**, 1914-1925.

30. 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.