User guide of MGEx-Tdb

Contents:

➢ *Introduction:*

Understanding almost every aspect of cellular and molecular biology requires establishing gene expression patterns. We have initiated a novel approach to create tissue-specific databases for the prediction of gene expression patterns and make maximum use of the available information.

Testis, where the cell differentiation is unique among all adult tissues, was taken as the tissue of interest for database creation. <u>MGEx-Tdb</u> (<u>M</u>ammalian <u>G</u>ene <u>Ex</u>pression <u>T</u>estis data<u>b</u>ase) is first among the series of databases to be created.

➢ *How is it created?*

Lists of genes reported to be transcribed or dormant, in several genome-wide expression studies on testis or specific testicular cells under different conditions were gathered from the literature (PubMed) and repositories (GEO, ArrayExpress, Oncomine, SMD, CIBEX).

The gene list was collected along with the necessary cognate information, such as,
1. the expression status (expressed or dormant),
2. species,
3. tissue-area or cell type, and
4. specific physiological or experimental condition.

This set of basic parameters is referred to as 'Expression Status under specific Location and Condition (ESLC)'. They were then systematically formatted and entered into the database.

No attempts were made to perform fresh statistical analysis. Instead, the 'expressed' or 'not expressed' call by the author(s) was relied upon. In any study with multiple hybridizations, the expression status indicated in the majority of cases was used as the final expression status. For example, any gene reported to be 'expressed' in majority of the hybridizations/samples, was identified as 'expressed'. In several cases, the gene-sets reported by the authors had to be split into several other gene-lists. Thus, several gene-sets of variable sizes were collected, and were characterized by the transcription, or lack of it, in specific location and condition.

Database creation: MySQL Relational Data-Base Management System (RDBMS) was employed for storing the data. A table was dedicated to store the basic gene-related information including the gene name, locus and transcript details. Another table was used to store gene identifiers such as gene name, gene description, official gene symbol, NCBI gene identifier, etc.

Perl based CGI script has been used to create an interface for the entry of gene-lists. Specially designed Perl programs such as LWP modules were used to connect to NCBI and, with the aid of NCBI E-utilities (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html), the required information was downloaded. The information included the NCBI gene id, official gene symbol, aliases, gene sequence, gene summary, chromosomal location, potential promoter

sequence [-1000 to +200 bp] and all transcript sequences (along with exon-intron details) corresponding to each gene. Using gene symbol and Swiss-Prot IDs downloaded from NCBI gene database, protein-related information encoded by that gene was downloaded from Swiss-Prot (http://ca.expasy.org/sprot). Similarly, transcription start sites were downloaded from dbTSS (ftp://ftp.hgc.jp/pub/hgc/db/dbtss/). When the information was not available in dbTSS, the 5' most position (a 5' end closest to that of the chromosome) of NCBI RefSeq gene models with was used to represent the TSS position. Perl codes were written to ensure automatic incorporation of the downloaded data into the database.

➤ *How it works?*　　　　　　　　　　　　　　　*Top*

The database predicts the ESLC of a gene using a novel scoring system. The score allotted for every ESLC of a gene is based on the consensus drawn from the expression status across multiple data-sets/gene-lists.

Note: A consensus for expression status has been drawn from only those studies that were comparable (i.e. those referring to the same species, physiological condition and/or location). However, there will be several differences within the experimental details and the data analysis process, across microarray studies. Such details include probe sets, samples and statistics used. Nevertheless, the simplicity of the current procedure allows us to bypass such details. In any case, the user can access the original data sets to refer to the details, using the PMID or other links provided for every data-set.

Each gene was assigned a score of two for a specific and definite expression status (transcribed or dormant) with respect to a single gene-list. With increasing evidence for an expression status of a gene, across the data-sets/gene-lists, the score increased, thereby providing a final 'reliability score'. For example, a gene reported to be 'not detected' in normal rat testis by two studies, would be shown as 'dormant' in normal rat testis with a reliability-score of four. Thus, a lower score would indicate either lack of supporting evidences or presence of contradicting reports for the specific expression status under consideration.

➤ *Logging into MGEx-Tdb :*　　　　　　　　　*Top*

1. Launch your web browser.
2. Enter the following in your browser's address bar:
   http://mgexdb.ibab.ac.in/cgi-bin/login/login.pl?pid=P00A08
3. Log in if you have already registered, if not, sign up as a new user.
4. Register, by providing the required information.
5. A confirmation mail is sent to you with the approval.
6. Log in with the user name and password.
7. Home page opens up.

The homepage

➢ *Start the search*

Searching MGEx-Tdb is as simple as can be.
The homepage in its entirety consists of two options of search.

Two search options provided. Click here
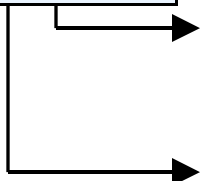
OPTION 1: Begin query with cell type/condition
1. Query:

To retrieve the genes and their details for a particular testicular cell type or for a physiological condition of a species, one can query the database using the **"**<u>Begin query with cell type, condition</u>**"** option**.**

The user can select the species of interest by using the drop down menu & can then select the cell type with the specific physiological condition. In case of no selection made for the cell type, the database uses the default option of testis. However, note that there is no default option for the physiological condition, and **must be** defined by the user before proceeding to the next step.

The **species** included in the database are **Mouse** (default), **Rat** and **Human.**

The testis cell types are further divided into two categories; Germ cells and Somatic cells. The user can either query with specific cell-type [eg: spermatogonia, sertoli cells] or collectively as germ cell, somatic cells or seminiferous tubules. While the former retrieves only those genes associated with the specified cell type, the latter retrieves all the genes that fall under the broader category.

The physiological conditions provided, include
Normal or disease state
Experimental conditions such as, 'hormone' or chemical/ drug treatment study
Developmental stages
Specific gene/protein studies- knock-out/knock-in/transgenic/mutant other miscellaneous conditions
The above physiological conditions further have sub-conditions that must be specified by the user.

- An option, ‘Click here for multiple conditions’ is provided, to facilitate the search for those microarray studies that deal with multiple conditions. For eg: FSH treated hypogonadal mice study is got by first choosing ‘hormone treatment’

- and then specifically, the study on ‘hypogonadism’.

Click submit

2. ***Results:***

- The result page displays a comprehensive list of both transcribed (expressed) and dormant (not-expressed) genes for the specified query. The genes are arranged based on their reliability scores (highest to lowest).
- They are further listed based on the order in which they were uploaded into the database (i.e., if 2 genes have same scores, the gene that is uploaded first will appear ahead of the other).
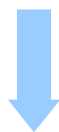
Genes in Mus musculus in Testis
Hormone treatment:FSH condition
### Others: Genetics – Hypogonadism condition

Top 20 Hits are shown below

Transcribed			Dormant		
Serial no.	Gene name	Reliability score	Serial no.	Gene name	Reliability score
1	Ren1	8.00	1	Tmem37	6.00
2	Nudcd2	8.00	2	Ank3	6.00
3	Vnn1	8.00	3	Dpysl4	4.00
4	Sbno1	8.00	4	Pck1	4.00
5	Cyp17a1	8.00	5	Ccl8	4.00
6	Actb	8.00	6	Plxnb2	4.00
7	Col4a1	8.00	7	Fbp1	4.00
8	Sp1	8.00	8	Rgs11	4.00
9	Rin2	8.00	9	Stra6	4.00
10	Fdx1	8.00	10	Myl1	2.00
11	Akr1b7	8.00	11	Hoxb9	2.00
12	Alas1	8.00	12	Ramp1	2.00
13	Erbb3	8.00	13	Rcan3	2.00
14	Nkx3-1	8.00	14	Itih2	2.00
15	Cyp11a1	8.00	15	Gata1	2.00
16	Rest	8.00	16	Myh6	2.00
17	Star	8.00	17	Scin	2.00
18	Lmna	8.00	18	Notch1	2.00
19	Aqp2	6.00	19	B4galnt1	2.00
20	Cidea	6.00	20	Nt5e	2.00

- A click on any gene in the list takes the user to the **'Gene Description Table'** for that particular species considered, as illustrated below.

Transcribed			Dormant		
Serial no.	Gene name	Reliability score	Serial no.	Gene name	Reliability score
1	Ren1	8.00	1	Tmem37	6.00
2	Nudcd2	8.00	2	Ank3	6.00
3	Vnn1	8.00	3	Dpysl4	4.00
4	Sbno1	8.00	4	Pck1	4.00
5	Cyp17a1	8.00	5	Ccl8	4.00

Click here
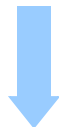
'Gene Description Table'

ASK QUESTIONS

renin 1 structural *(Mus musculus)* -Ren1	Gene Information	Transcript Information	Protein Information	Promoter Information
	Expression Information: tissue and/or cell, condition		Relevant PubMed Papers	Other Information

OPTION 2: Begin query with gene name/s   *Top*

1. Query:

- The "Begin query with gene name/s" option facilitates the user to query the database by individual gene name (gene symbol or full name) in the text box provided.
- Multiple genes can be queried by separating the entities with a comma.

Begin query with gene name/s

acrv1,prm1

Separate words by "," (comma); do not use space

Submit   Reset

Note: No other delimiter like semicolon/space/colon should be used.

The output page for such a query is generally divided into 2 panels, both of which are accompanied by the '**Gene Description Table**'. One panel consists of 'Identical matches', corresponding to exact/complete match (w.r.t the gene name/symbol) for the query. The other consists of 'Partial matches' that correspond to partial/incomplete match (w.r.t the gene name/symbol) for the query in the database.

- For eg: when queried with Prm1, the database displays the result as
  Prm1 under 'Identical match'
  Sprm1 & Oprm1 under 'Partial matches'
  For those where the 'partial hits' is not displayed, it only indicates that the database has only a perfect match for the given query and vice versa.

➢ *The 'Gene Description Table':*

The table provides links for information on following entities.

1. Gene Information
2. Transcript Information
3. Protein Information
4. Promoter Information
5. Expression Information
6. Relevant PubMed papers
7. Other Information

- Click on these links to get the complete details of each component.

acrosomal vesicle protein 1 *(Homo sapiens)* - ACRV1	Gene Information	Transcript Information	Protein Information	Promoter Information
	Expression Information: tissue and/or cell, condition		Relevant PubMed Papers	Other Information

1. Gene information:

- This provides complete information of a gene - official gene name, abbreviation, summary, organism in which the gene is reported, gene loci, the sequence, transcripts coded, the 'Transcription Start Site' (TSS), the potential promoter region (-1000 to +200 from TSS), etc. are provided by the database.
- The specific URLs in the map viewer and the genomic context section, link to the corresponding pages of NCBI.

Gene information

Gene name	acrosomal vesicle protein 1

Gene abbreviation	ACRV1	Gene aliases	D11S4365, SP-10, SPACA2,sperm protein 10.

Gene summary (copied from NCBI Gene)	This gene encodes a testis-specific, differentiation antigen, acrosomal vesicle protein 1, that arises within the acrosomal vesicle during spermatogenesis, and is associated with the acrosomal membranes and matrix of mature sperm. This gene consists of 4 exons and its alternative splicing generates multiple distinct transcripts, which encode protein isoforms ranging from 81 to 265 amino acids. The longest transcript is the most abundant, comprising 53-72% of the total acrosomal vesicle protein 1 messages; the second largest transcript comprises 15-32%; the third and the

NCBI gene ID	56

Organisms where the gene is reported	Macaca mulatta, Homo sapiens, Bos taurus, Rattus norvegicus, Mus musculus

Size of the gene	8713 nt

Gene loci and the position (start and end nucleotide numbers) on the chromosome.	125047438-1250561!

Gene sequence	CCCCAACACACTACCCTGCACACAGAAAATAATCATTGGTCTTCACTAGTGAAATAAGCAGTGGGTTGCT AAGGGCCAACTTGCCTGAACAGGCTACACAAGAACCTCAGAGCCCAACCCATTGTGAAGAAACATGGGTT ACTTCTGAGGTTCTAGAATTCCCAGAAGCTCTGCTTCAGCACTGGAAGCTTTTGCTCGCAGTTTGCTTCA TAGCTCTGTGAAGAAGCTGTGGCCCACACTGGGGTCCCCTCTTTTCCTAAATCCAGATGAACAGGTTTCT CTTGCTAATGAGTCTTTATCTGCTTGGATCTGCCAGAGGTGAGTGTTTGATCTAGGCTTCAGGTAAATAC AAAAGGGGTCTCCCTTAAGCACATTTCAGGCCTTGAAGAAGGAAGACCAGTCTCTCAGTCACCCAAACCAA TCTGAGTCAGAATCAAGAGCAGCTAAACAAGTCCAACTCGGAGAGCTCCTGAACCAGTTGATTTCTCTAG

Organization of the gene	minus/reverse strand	Chromosomal location	11p12-q13

NCBI transcript ID	NM_001612, NM_020115, NM_020113, NM_020111, NM_020110, NM_020109, NM_020108, NM_020107, NM_020069

TSS (Transcription Start Site)	TSS position	125056151
	TSS sequence	ACACTTTTAAAAGACAGTAGATATCTTTTAAACTTTATTACAAAAAATATGTGCTTTTTGGTA

OMIM ID (for disease related information)	102525

Potential promoter sequence position	125057151-125055951

Potential promotor sequence (-1000 to +200 nucleotides)	GGTCTTATGCTAGACAGAAAGCCTGGATTTGATGTAGAGCCGCCACTCTCTCTTTCTTTAACCCTCACCT TCCTTTCAAAATGTAACAATAAGTACATCTATTCTAGATTTAGCATTCAGTGGTATATAAAACAAAGTTG TTCAAGAGAGCTGAGCAGAGAAGCAGCGAGAAATGGAAGCTACCAGATCTTTGCCACTCAATCCTCCATA AGTTCTAGTTTTATATCAGGTGGATGCCTCACTGCATCCATCTTCTCAGCTTGGCCACCCCCTTGCTACC AAGTTGCAGCTCAAAACGCAGCTTTCTACTGTCTCTTGATCTCTTTACTCTGCCCATATCTGTGTAAACT CTCCTCAGTTACCCATTTGAGTTTGCCACCTGTTTCCTGCTAAAACCATGTTGATAAAGGTCTGTATGAA AAGACTATTTACAGAAGTGGGCAAGGTTAAGGAAACAAACAAACAAAAAAAGAGATAAGTTTGGGCATCC

Links	Physical map/genomic context Map viewer

From NCBI

From DBTSS

2. Transcript information:

- This provides the information for individual transcripts- the NCBI ID, the exon/intron specificities etc. for the queried gene.

- One can specify the exon/intron number in the '<u>view info about</u>' section to get the complete information about the length, position and the occurrence in other transcripts

Specify the number

Note: The details of all the transcripts (reported or hypothetical) for a particular gene, provided by the database are obtained from Entrez Gene.

3. Protein information:

- This includes the Swiss-Prot ID of the protein, along with its functional details, molecular weight and the amino acid sequence, are also displayed. All information provided is as per the ExPASy Swiss-Prot Database.

Protein information

Swiss ID	P26436		
Protein name	Acrosomal protein SP-10 [Precursor]		
Molecular weight	28156 Da	Length of the protein	265 AA
Function *(quoted from Swiss-Prot)*	NONE		
Protein sequence	MNRFLLLMSLYLLGSARGTSSQPNELSGSIDHQTSVQQLPGEFFSLENPSDAEALYETSSGLNTLSEHGSSEHGSSKHTVAEHTSGEHAESEHASGEPAATE		

Isoform ID	Details
P26436-1	This is the isoform sequence displayed in this entry.
P26436-2	Features which should be applied to build the isoform sequence: VSP_004127.
P26436-3	Features which should be applied to build the isoform sequence: VSP_004120.
P26436-4	Features which should be applied to build the isoform sequence: VSP_004121.
P26436-5	Features which should be applied to build the isoform sequence: VSP_004121, VSP_004127.
P26436-6	Features which should be applied to build the isoform sequence: VSP_004122.
P26436-7	Features which should be applied to build the isoform sequence: VSP_004120, VSP_004125.
P26436-8	Features which should be applied to build the isoform sequence: VSP_004121, VSP_004128.
P26436-9	Features which should be applied to build the isoform sequence: VSP_004121, VSP_004126.

4. Promoter information:

- This comprise of the details regarding the potential promoter region of the gene and its TSS position as per DBTSS. The specific URLs in the map viewer and the genomic context section, link to the corresponding pages of NCBI.

Promoter information

Gene name	acrosomal vesicle protein 1
Gene loci and the position (start and end nucleotide numbers) on the chromosome.	125047438-125056151
TSS (Transcription Start Site)	**TSS position** 125056151 **TSS sequence** ACACTTTTAAAAGACAGTAGATATCTTTTAAACTTTATTACAAAAAATATGTGCTTTTTGGTA
Potential promoter sequence position	125057151-125055951
Potential promotor sequence (-1000 to +200 nucleotides)	GGTCTTATGCTAGACAGAAAGCCTGGATTTGATGTAGAGCCGCCACTCTCTCTTTCTTTAACCCTCACCT TCCTTTCAAAATGTAACAATAAGTACATCTATTCTAGATTTAGCATTCAGTGGTATATAAAACAAAGTTG TTCAAGAGAGCTGAGCAGAGAAGCAGCGAGAAATGGAAGCTACCAGATCTTTGCCACTCAATCCTCCATA AGTTCTAGTTTTATATCAGGTGGATGCCTCACTGCATCCATCTTCTCAGCTTGGCCACCCCCTTGCTACC AAGTTGCAGCTCAAAACGCAGCTTTCTACTGTCTCTTGATCTCTTTACTCTGCCCATATCTGTGTAAACT CTCCTCAGTTACCCATTTGAGTTTGCCACCTGTTTCCTGCTAAAACCATGTTGATAAAGGTCTGTATGAA AAGACTATTTACAGAAGTGGGCAAGGTTAAGGAAACAAACAAACAAAAAAAGAGATAAGTTTGGGCATCC
Physigal map/ genomic context	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=full_report&list_uids=56#context
Map viewer	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?direct=on&idtype=gene&id=56

5. Expression information:

- This section provides the expression pattern of the queried gene, in testis tissue, with respect to its cell type, conditions and sub-conditions. The expression status is clearly indicated as 'transcribed' or 'dormant' or 'marginal' & is associated with a reliability score.
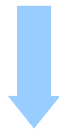
Expression information

Manually curated and automatically processed data from genome-wide studies for ACRV1 *(Homo sapiens)*

Species	Cell type	Sub Celltype	Condition	SubCondition	Expression Status	Reliability Score
Homo sapiens			Normal		Transcribed	16.00
Homo sapiens			Developmental Stages Disease	Embryonic stage Cancer	Dormant	2
Homo sapiens			Disease	Infertility - Azoospermia	Marginal	0.00
Homo sapiens			Disease	Infertility - Obstructive Azoospermia	Transcribed	2.00
Homo sapiens			Disease	Infertility - Non Obstructive Azoospermia	Marginal	0.00
Homo sapiens			Disease	Cancer	Dormant	8
Homo sapiens			Disease	Infertility	Marginal	0.00
Homo sapiens			Developmental Stages	Embryonic stage	Marginal	0.00
Homo sapiens	Somatic Cell	Leydig	Normal		Transcribed	4.00

- The user can click on specific '<u>Expression Status</u>' to get relevant literature, from PubMed or other repositories, from which the final reliability score was derived.

Expression information

Manually curated and automatically processed data from genome-wide studies for ACRV1 *(Homo sapiens)*

Species	Cell type	Sub Celltype	Condition	SubCondition	Expression Status	Reliability Score
Homo sapiens			Normal		Transcribed	16.00
Homo sapiens			Developmental Stages Disease	Embryonic stage Cancer	Dormant	2
Homo sapiens			Disease	Infertility - Azoospermia	Marginal	0.00
Homo sapiens			Disease	Infertility - Obstructive Azoospermia	Transcribed	2.00
Homo sapiens			Disease	Infertility - Non Obstructive Azoospermia	Marginal	0.00

Click here

Serial no	Identifier	Source
1	17312182	Pubmed
2	15774023	Pubmed
3	17074343	Pubmed
4	17473427	Pubmed
5	15075390	Pubmed
6	12029067	Pubmed
7	E-TABM-145	Array express
8	17177185	Pubmed
9	16158187	Pubmed
10	17363524	Pubmed
11	15950434	Pubmed
12	11015593	Pubmed
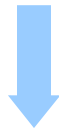
Click here

Cancer/testis genes in multiple myeloma: expression patterns and prognosis value determined by microarray analysis.

Condomines M, Hose D, Raynaud P, Hundemer M, De Vos J, Baudard M, Moehler T, Pantesco V, Moos M, Schved JF, Rossi JF, Rème T, Goldschmidt H, Klein B.

Institute of Research in Biotherapy, Centre Hospitalier Universitaire Montpellier, Montpellier, France.

Cancer-testis (CT) Ags are expressed in testis and malignant tumors but rarely in nongametogenic tissues. Due to this pattern, they represent attractive targets for cancer vaccination approaches. The aims of the present study are: 1) to assess the expression of CT genes on a pangenomic base in multiple myeloma (MM); 2) to assess the prognosis value of CT gene expression; and 3) to provide selection strategies for CT Ags in clinical vaccination trials. We report the expression pattern of CT genes in purified MM cells (MMC) of 64 patients with newly diagnosed MM and 12 patients with monoclonal gammopathy of unknown significance, in normal plasma cell and B cell samples, and in 20 MMC lines. Of the 46 CT genes interrogated by the Affymetrix HG-U133 set arrays, 35 are expressed in the MMC of at least one patient. Of these, 25 are located on chromosome X. The expression of six CT genes is associated with a shorter event-free survival. The MMC of 98% of the patients express at least one CT gene, 86% at least two, and 70% at least three CT genes. By using a set of 10 CT genes including KM-HN-1, MAGE-C1, MAGE-A3/6/12, MAGE-A5, MORC, DDX43, SPACA3, SSX-4, GAGE-1-8, and MAGE-C2, a combination of at least three CT genes-desirable for circumventing tumor escape mechanisms-is obtained in the MMC of 67% of the patients. Provided that the immunogenicity of the products of these 10 CT genes is confirmed, gene expression profiling could be useful in identifying which CT Ags could be used to vaccinate a given patient.

- The section also provides a link at the bottom of the page, to the expression status across other normal tissues for the user specified gene.

Homo sapiens	Germ Cell	Sperm	Disease	Infertility – Teratozoospermia	Transcribed
Homo sapiens	Seminiferous tubules		Normal		Transcribed
Homo sapiens	Germ Cell		Normal		Marginal

LINK SHOWING EXPRESSION OF ACRV1 ACROSS DIFFERENT TISSUES

Click here

Expression information for ACRV1

Tissue	Score	Expression type	Condition
Testis	16.00	Transcribed	Normal
Appendix	2	Dormant	Normal
Adrenal gland	0.00	Marginal	Normal
Brain Amygdala	4	Dormant	Normal
Brain Caudate nucleus	0.00	Marginal	Normal
Brain	4	Dormant	Normal
Brain Corpus callosum	2	Dormant	Normal
Brain Cerebellum	4	Dormant	Normal
Brain Cerebellum pedunculus	2	Dormant	Normal
Brain Globus pallidus	2	Dormant	Normal
Brain Cingulate cortex	2	Dormant	Normal

- At the end of the page one can find links to other gene expression databases

Links to gene expression databases

UniGene

BioGPS

Swiss-Prot

6. Relevant PubMed papers:

Links to PubMed.

7. Other information:

- This section provides links to literature information regarding the function, sub-cellular location or other general information, for the gene considered. The 'retrieve abstracts' option can be used to read the abstracts of the citations.

Other Details		
Features	**PubMed ID**	**Relevant statement**
Specific function	☐ 1.10775167	1. Human SP-10, expressed on the equatorial region of acrosome-reacted sperm, inde mediates sperm-oolemma binding in a beta(1) integrin-independent manner, but not sperm-zona binding.

Retrieve abstract

Relevant expression information papers

Click here to retrieve the abstracts