MGDB: a comprehensive database of genes involved in melanoma

Di Zhang1,†, Rongrong Zhu1,†, Hanqian Zhang1,†, Chun-Hou Zheng2,3, and Junfeng Xia1,3,*

1Institute of Health Sciences, School of Computer Science and Technology, 2College of Electrical Engineering and Automation and 3Center of Information Support and Assurance Technology, Anhui University, Hefei, Anhui 230601, China

*Corresponding author: Tel: +86 551 63861990, Fax: +86 551 63861965; Email: jfxia@ahu.edu.cn

†These authors contributed equally to this work.


Received 18 June 2015; Revised 24 August 2015; Accepted 7 September 2015

Abstract

The Melanoma Gene Database (MGDB) is a manually curated catalog of molecular genetic data relating to genes involved in melanoma. The main purpose of this database is to establish a network of melanoma related genes and to facilitate the mechanistic study of melanoma tumorigenesis. The entries describing the relationships between melanoma and genes in the current release were manually extracted from PubMed abstracts, which contains cumulative to date 527 human melanoma genes (422 protein-coding and 105 non-coding genes). Each melanoma gene was annotated in seven different aspects (General Information, Expression, Methylation, Mutation, Interaction, Pathway and Drug). In addition, manually curated literature references have also been provided to support the inclusion of the gene in MGDB and establish its association with melanoma. MGDB has a user-friendly web interface with multiple browse and search functions. We hoped MGDB will enrich our knowledge about melanoma genetics and serve as a useful complement to the existing public resources.

Database URL: http://bioinfo.ahu.edu.cn:8080/Melanoma/index.jsp

Introduction

Melanoma is a malignant tumor that develops from specialized pigment cells called melanocytes and considered as one of the most aggressive human cancers. The incidence of melanoma is increasing at a rate faster than any other solid malignancy, with an estimated 232 100 new cases and 55 000 deaths reported worldwide in 2012 (1). Due to its aggressive behavior and limited treatment options, the overall survival for patients with metastatic melanoma is poor. Early detection is critical to improve the overall prognosis of melanoma and an important way to reduce deaths from melanoma.

Similar to many other cancers, melanoma is initiated by activation of oncogenes or inactivation of tumor suppressor genes. Recent advances in molecular profiling and genome sequencing have revealed that the melanoma is a
gene expressions from hundreds of tumor samples and detailed annotations for each melanoma genes, such as references to support inclusion of a gene in the database, information from NCBI, manually curated literature data including basic gene information, sequencing omoma genes, we have created a user-friendly interface in the database as evidenced from the biomedical literature. To gather the genes known to be involved in melanoma carcinogenesis, we have created the Melanoma Gene Database (MGDB) that catalogs melanoma related genes to construct an integrated database yet available. Therefore, we have collected melanoma database focusing on genes involved in melanoma causation and would be helpful for the research community.

Data collection and content

In the last few years, a number of gene databases have emerged with a focus on a specialized cancer type as exemplified by Breast Cancer Gene Database (10), Oral Cancer Database (11), Prostate cancer (12), Lung Cancer Database (13), Cervical Cancer Gene Database (14) and Renal Cancer Gene Database (15), etc. However, there is no database focusing on genes involved in melanoma causation yet available. Therefore, we have collected melanoma related genes to construct an integrated database termed Melanoma Gene Database (MGDB) that catalogs the genes known to be involved in melanoma carcinogenesis as evidenced from the biomedical literature. To gather and uniformly present the available information on melanoma genes, we have created a user-friendly interface in the form of MGDB. The database integrates heterogeneous data including basic gene information, sequencing information from NCBI, manually curated literature references to support inclusion of a gene in the database, and detailed annotations for each melanoma genes, such as gene expressions from hundreds of tumor samples from the Cancer Genome Atlas (TCGA) project webpage (http://cancergenome.nih.gov), methylation data from GEO database (16), somatic mutations from COSMIC (17), interacting partners from PINA (18), pathway information from KEGG (19), and drug information from DGIdb (20). In addition, the database provides search facility for querying the database and an online BLAST interface for sequence search. It serves multiple purposes for the user, including: (i) obtaining literature-based gene lists for melanoma; (ii) browsing detailed information, including gene expression levels, methylation levels, gene mutation sites, protein–protein interactions, involved pathways and drug information; (iii) a resource for therapy melanoma; (iv) a resource for high-throughput genetic screening to find melanoma-related gene variants. Overall, MGDB serves as the most comprehensive data resource that will enable the exploration of relevant information for human melanoma related genes, and would be helpful for the research community.
To better understand the function of these melanoma genes in our database, we collected their extensive functional information. The representative annotations in the MGDB database are summarized in Table 1, which includes a basic description of the gene, gene expression, methylation, somatic mutation information, protein–protein interactions, pathway information and drug information (Figure 1). Details of these databases can be found through the cited references as well as from MGDB.

### Biological features of 422 protein-coding genes in MGDB

As the majority of melanoma genes are protein-coding genes (422 human protein-coding genes in Supplementary Table S1), we performed the pathway enrichment analysis to investigate their involved pathways using the KOBAS server (21). Statistically significant pathways were determined by a corrected $P < 0.05$ (Benjamini–Hochberg corrected hypergeometric test). Pathways playing important roles in the pathogenesis of melanoma are at the top enriched pathway list, such as ‘Pathways in cancer’, ‘Melanoma’, ‘Signaling by SCF-KIT’, ‘Oncogene Induced Senescence’, ‘p53 signaling pathway’, and ‘Downstream signal transduction’ (Supplementary Table S3).

To further investigate biological processes, we conducted enrichment tests on 422 human protein-coding genes using the online tool DAVID (Supplementary Table S4) (22). We selected those GO terms or InterPro domains with an adjusted $P < 0.05$ (Benjamini–Hochberg corrected hypergeometric test). Using the online DAVID server, 422 protein-coding genes were markedly overrepresented in regulation of cell proliferation, anatomical structure development and developmental process according to GO Biological Processes terms (Table 2). In addition, we observed that the pigmentation relevant GO Biological Processes are enriched in 422 protein-coding genes, including ‘pigmentation during development’, ‘pigmentation’ and ‘regulation of pigmentation during development’ (Supplementary Table S4). These results highlight that these biological processes may be involved in transition of pigment cell progression to cancerous state. Lastly, assignment of 422 genes to InterPro domains (Supplementary Table S4) revealed that the top four categories represented in the dataset are: ‘Protein kinase, core’, ‘Protein kinase, ATP binding site’, ‘Tyrosine protein kinase’ and ‘Tyrosine protein kinase, active site’, thereby suggesting importance of kinase activity in melanoma cells.

### Prioritization of protein-coding genes and its enriched dense network module

To further assess the importance of each protein-coding gene in MGDB, we performed gene prioritization analysis using the ToppGene web server (23). ToppGene requires two types of gene set, i.e. training gene set, and test gene set. The training gene set contains already well-known melanoma genes, while the test gene set are the remaining interesting genes in MGDB. Here, we compiled a training gene list that included 10 well-studied melanoma genes ($\text{BRAF}$, $\text{NRAS}$, $\text{CDK4}$, $\text{MITF}$, $\text{PTEN}$, $\text{CDKN2A}$, $\text{TP53}$, $\text{KIT}$, $\text{GNAQ}$ and $\text{GNA11}$). ToppGene ranks candidate genes (test genes) based on functional similarity to training gene list through integration of various types of data such as gene expression, regulatory information, and literature mining data. Based on the gene ranking results of ToppGene (Supplementary Table S5), the top ranked genes tend to have multiple evidences from different data sources. Besides 10 well-studied melanoma genes in the training set, $\text{CDKN1B}$ was top ranked melanoma gene in remaining 412 genes from the test set. Functional analyses on the top 100 melanoma genes show similar functional

---

**Table 1.** Annotation entry statistics for 527 melanoma genes

<table>
<thead>
<tr>
<th>Data category</th>
<th>Related entries</th>
<th>Annotated melanoma genes</th>
<th>Content/sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Melanoma genes</td>
<td>527</td>
<td>527</td>
<td>Gene names, full name, genomics position, synonym, definition from Entrez gene database</td>
</tr>
<tr>
<td>Literature</td>
<td>9788</td>
<td>997</td>
<td>Literature evidence for Human Melanoma genes</td>
</tr>
<tr>
<td>Expression of protein-coding genes</td>
<td>420</td>
<td>420</td>
<td>Expression in 470 SKCM$^a$ samples and 80 UVM$^a$ from TCGA database</td>
</tr>
<tr>
<td>Expression of Non-coding genes</td>
<td>89</td>
<td>89</td>
<td>Expression in 452 SKCM$^a$ samples and 80 UVM$^a$ from TCGA database</td>
</tr>
<tr>
<td>Methylation</td>
<td>1926</td>
<td>396</td>
<td>Methylation in melanoma-related samples from GEO (16)</td>
</tr>
<tr>
<td>Genomic variation</td>
<td>2418</td>
<td>300</td>
<td>COSMIC database</td>
</tr>
<tr>
<td>Protein-Protein Interaction</td>
<td>10 667</td>
<td>352</td>
<td>PINA database</td>
</tr>
<tr>
<td>Pathway</td>
<td>1833</td>
<td>185</td>
<td>KEGG database</td>
</tr>
<tr>
<td>Drug</td>
<td>1236</td>
<td>133</td>
<td>DGIdb database</td>
</tr>
</tbody>
</table>

$^a$SKCM is short for Skin Cutaneous Melanoma, and UVM is short for Uveal Melanoma.
distribution with the total 422 melanoma genes. These highly overlapping relationships of top ranked genes support the accuracy of our data.

Employing the software GenRev (24), we further searched for dense modules enriched with the top 100 melanoma associated genes through their protein–protein interactions (PPIs) and identified 1 module comprising of 94 genes (Figure 2). Of the 94 genes, 92 of them are from the top 100 melanoma associated genes. The remaining two are GTF3C1 and UBC. Previous study have found that GTF3C1 was differentially expressed between melanocytic nevi with and without the V600E BRAF Mutation (25), which may play an important role in melanoma cell. Network analysis highlights UBC, a polyubiquitin precursor, as a hub gene in the module, which suggests that the majority of melanoma protein-coding genes are involved in ubiquitination pathway via interacting with UBC (26). Frequent alteration of ubiquitination process has been reported in human cancer (27). In conclusion, the majority of the top ranked melanoma-related genes connects each...
Table 2. Top 20 enriched GO biological processes of the 422 protein-coding genes

<table>
<thead>
<tr>
<th>GO term</th>
<th>P value</th>
<th>Benjamini–Hochberg corrected P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulation of cell proliferation</td>
<td>3.5167E–50</td>
<td>1.1893E–46</td>
</tr>
<tr>
<td>Anatomical structure development</td>
<td>9.3799E–49</td>
<td>1.5861E–45</td>
</tr>
<tr>
<td>Developmental process</td>
<td>9.4470E–48</td>
<td>1.065E–44</td>
</tr>
<tr>
<td>System development</td>
<td>1.3688E–47</td>
<td>1.1573E–44</td>
</tr>
<tr>
<td>Organ development</td>
<td>5.8100E–46</td>
<td>3.9299E–43</td>
</tr>
<tr>
<td>Positive regulation of cellular process</td>
<td>7.1393E–46</td>
<td>4.0242E–43</td>
</tr>
<tr>
<td>Positive regulation of biological process</td>
<td>7.3029E–45</td>
<td>3.5283E–42</td>
</tr>
<tr>
<td>Multicellular organism development</td>
<td>6.2323E–42</td>
<td>2.6347E–39</td>
</tr>
<tr>
<td>Negative regulation of cellular process</td>
<td>1.3053E–41</td>
<td>4.9053E–39</td>
</tr>
<tr>
<td>Negative regulation of biological process</td>
<td>1.5792E–41</td>
<td>5.3408E–39</td>
</tr>
<tr>
<td>Cell differentiation</td>
<td>9.9281E–36</td>
<td>3.0524E–33</td>
</tr>
<tr>
<td>Cellular developmental process</td>
<td>5.6154E–35</td>
<td>1.5826E–32</td>
</tr>
<tr>
<td>Regulation of cell death</td>
<td>4.4543E–34</td>
<td>1.1588E–31</td>
</tr>
<tr>
<td>Regulation of programmed cell death</td>
<td>1.8354E–33</td>
<td>4.4338E–31</td>
</tr>
<tr>
<td>Response to chemical stimulus</td>
<td>2.9200E–33</td>
<td>6.5836E–31</td>
</tr>
<tr>
<td>Regulation of apoptosis</td>
<td>4.5486E–33</td>
<td>9.6147E–31</td>
</tr>
<tr>
<td>Multicellular organismal process</td>
<td>1.1631E–31</td>
<td>2.314E–29</td>
</tr>
<tr>
<td>Anatomical structure morphogenesis</td>
<td>1.9127E–31</td>
<td>3.5938E–29</td>
</tr>
<tr>
<td>Regulation of cell differentiation</td>
<td>1.7245E–30</td>
<td>3.0697E–28</td>
</tr>
<tr>
<td>Negative regulation of cell proliferation</td>
<td>4.7244E–30</td>
<td>7.9890E–28</td>
</tr>
</tbody>
</table>

Figure 2. The enriched dense network module for the top ranked melanoma-related genes based on protein-protein interaction data. The 92 genes in black are genes from the top 100 ranked melanoma-related genes. The remaining two genes in red are the linker genes that bridge the 92 genes.
other and forms a dense network, which may further confirm the importance of the gene ranking results.

**Biological features of microRNAs in MGDB**

The majority of non-coding genes in MGDB are belonging to microRNAs (miRNAs). We retrieved these miRNAs (Supplementary Table S2) predictive targets from miRTarBase (version 4.5, November 2013) (28). We required that a functional miRNA target interaction was limited in human species, which generated 748 target genes from 84 miRNAs (Supplementary Table S6). Among of the 748 genes, 117 target genes were overlapped with the 422 protein-coding genes in MGDB. Using the KOBAS server (21), we found that these genes are enriched in cancer related pathways, such as melanoma pathway (Supplementary Table S7).

To obtain more reliable miRNA-target genes, we also extracted 202 target genes for functional analysis that were regulated by at least 2 miRNA melanoma genes (Supplementary Table S7). We can observe that 2.97% of the target genes are involved in pigmentation.

**Database construction and web interface**

**Database organization**

We stored and managed all the data in MySQL (version 5.1.73), which is a popular and open source widely used in biomedical database development. A user-friendly web interface for browsing and searching implemented by JavaScript, the main function were shown in Figure 1, which is freely and systematically retrieve the information.

**Data search**

MGDB supports versatile searching, including quick search, advanced search and BLAST search. Firstly, a quick search in the Home page was provided for searching by gene ID or gene names. Secondly, advanced search was incorporated to query gene by filtering multiple options (i.e. gene ID, gene type, official symbol, position or full name) in the Search page. In addition, a search interface to access melanoma genes related literatures provided a window for users to find more comprehensive melanoma genes descriptions from original literature sources. Thirdly, user can utilize an online BLAST interface to input an interesting sequence in FASTA format and query against the melanoma gene nucleotide or protein sequences in our MGDB. Lastly, using the chromosome browser, users can obtain melanoma gene lists that include a summary of the genes and hyperlinks to detailed gene evidence and annotation pages.

**Representative entry in MGDB**

The output page displays comprehensive information about the gene entries searched by the user. Figure 1A shows a sample output page generated by querying MGDB using the keyword ‘PTEN’. Annotations of each gene can be obtained by clicking the label ‘General Information’, ‘Expression’, ‘Methylation’, ‘Mutation’, ‘Interaction’, ‘Pathway’ and ‘Drug’ on top of the output page.

In the ‘General information’ page, gene ID, gene official symbol, gene full name, synonym, gene location on chromosome and gene type are displayed. Besides, evidence section includes the literature evidence and the related PubMed ID. Users can click on the hyperlinks of literature evidence to access corresponding PubMed page. DNA, cDNA, and protein sequence can also be found in a tabular view in the ‘General Information’ page. In the ‘Expression’ page, MGDB collects RNA-seq-based level 3 expression data from TCGA. Output in MGDB is a box plot which shows the two melanoma subtypes [Skin Cutaneous Melanoma (SKCM) and Uveal Melanoma (UVM)] expression levels (Log 2 fold change of normalized counts). In the ‘methylation’ page, we use the textual interfaces to demonstrate the methylation patterns in melanoma tissues. In the ‘mutation’ page, full mutation details have been curated from COSMIC (17). Genomic variants in a specific gene were described for its genomic information and the corresponding histology of melanoma. In the ‘interaction’ page, the interaction table for storing gene corresponding UniProt accession number and its interacting partner, which extracted from PINA database (18). In the ‘pathway’ page, genes within its involved pathways were described. Through the entry, we can link to the corresponding pathway in KEGG database (19). In the ‘Drug’ page, we extracted the gene targeted drug information through integrating the DGIdb database (20).

**Conclusions and perspectives**

In this study, we applied a comprehensive literature mining method to extract 422 protein-coding genes, 95 microRNAs and 10 lncRNAs in melanoma. MGDB is the first literature-based melanoma database based on literature extraction by integrating multi-dimensional data of genomic, epigenetic, pathway and drug information. It provides a user-friendly web interface, where users can easily access to the database resources through text querying, sequence searching and so on. We believed that MGDB will greatly facilitate cancer researchers’ effort of surveying melanoma genes and pathways.
the literature on genes and their involvement in melanoma. MGDB will be routinely updated to follow research progresses of melanoma biology. We anticipated that MGDB could provide a valuable resource for further studies and understanding of melanoma genetics.

**Supplementary Data**

Supplementary data are available at Database Online.

**Acknowledgements**

We thank Min Zhao for giving us constructive suggestions.

**Funding**

National Natural Science Foundation of China (31301101, 61272339); the Anhui Provincial Natural Science Foundation (1408085QF106); the Specialized Research Fund for the Doctoral Program of Higher Education (20133401120011); Technology Foundation for Selected Overseas Chinese Scholars from Department of Human Resources and Social Security of Anhui Province (No. 61272339); the Anhui Provincial Natural Science Foundation of Human Resources and Social Security of Anhui Province (No. 31301101, 1508085QF106); the Specialized Research Fund of the Doctoral Program of Higher Education (20133401120011); Technology Foundation for Selected Overseas Chinese Scholars from Department of Human Resources and Social Security of Anhui Province (No. 31301101).

**Supplementary Data**

Supplementary data are available at Database Online.

**Conflicts of interest**

None declared.

**References**