Structuring osteosarcoma knowledge: an osteosarcoma-gene association database based on literature mining and manual annotation

Kathrin Poos1, Jan Smida2,3, Michaela Nathrath2,3, Doris Maugg2,3, Daniel Baumhoer2,4, Anna Neumann1 and Eberhard Korsching1,*

1Institute of Bioinformatics, University of Münster, Münster, Germany, 2Clinical Cooperation Group Osteosarcoma, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany, 3Children’s Cancer Research Center and Department of Pediatrics, Klinikum rechts der Isar, Technische Universität München, 81664 Munich, Germany and 4Bone Tumor Reference Center at the Institute of Pathology, University Hospital Basel, Basel, Switzerland

*To whom correspondence should be addressed. Tel: +49 251 8350651; Fax: +49 251 8353005; Email: korschi@uni-muenster.de.


Received 8 November 2013; Revised 4 March 2014; Accepted 23 April 2014

Abstract

Osteosarcoma (OS) is the most common primary bone cancer exhibiting high genomic instability. This genomic instability affects multiple genes and microRNAs to a varying extent depending on patient and tumor subtype. Massive research is ongoing to identify genes including their gene products and microRNAs that correlate with disease progression and might be used as biomarkers for OS. However, the genomic complexity hampers the identification of reliable biomarkers. Up to now, clinico-pathological factors are the key determinants to guide prognosis and therapeutic treatments. Each day, new studies about OS are published and complicate the acquisition of information to support biomarker discovery and therapeutic improvements. Thus, it is necessary to provide a structured and annotated view on the current OS knowledge that is quick and easily accessible to researchers of the field. Therefore, we developed a publicly available database and Web interface that serves as resource for OS-associated genes and microRNAs. Genes and microRNAs were collected using an automated dictionary-based gene recognition procedure followed by manual review and annotation by experts of the field. In total, 911 genes and 81 microRNAs related to 1331 PubMed abstracts were collected (last update: 29 October 2013). Users can evaluate genes and microRNAs according to their potential prognostic and therapeutic impact, the experimental procedures, the sample types, the biological contexts and microRNA target gene interactions. Additionally, a pathway enrichment analysis of the collected genes highlights different aspects of OS progression.
Introduction

Osteosarcoma (OS) is the most common primary malignant tumor of bone frequently affects children and young adolescents (1). It is a complex disease with manifold numerical and structural genomic alterations affecting multiple genes to a varying extent (2). Patients without clinical signs of systemic spread show 5-year survival rates of 60–80% (3), whereas patients with metastasis at diagnosis exhibit 5-year survival rates of 20–30%. Since 1980, the prognosis of patients has more or less stagnated and no significant therapy improvements have been achieved (4).

Massive research in the field of OS is ongoing to assess the prognostic and therapeutic impact of possible biomarkers and altered molecular pathways. For instance, several studies detected frequent genomic alterations of the tumor suppressor genes TP53 and RB1 in OS and correlated these findings with disease outcome (5–7). Other studies identified p-glycoprotein and ezrin that influence the response to chemotherapy and metastatic spread, respectively (8). Recently, attention has been paid to the value of small non-coding microRNAs in the pathogenesis of OS, e.g. the miR-17-92 cluster (9, 10) and miR-9-5p (11, 12). MicroRNAs represent interesting biomarkers for OS, as they are able to simultaneously regulate hundreds of target genes and several molecular pathways (13). However, the prognostic and therapeutic significance neither for distinct genes including their gene products nor for microRNAs has been determined in controlled clinical studies yet (3). The key prognostic determinants are still clinico-pathological factors and include tumor stage (14), patient age, tumor size and location and the response to neoadjuvant chemotherapy (15). Consequently, all patients are treated with multiagent chemotherapy irrespective of its individual efficacy (16). Moreover, new studies about OS are continuously published and complicate the acquisition of information for specific research purposes and questions.

To support the efforts in OS research and biomarker discovery, we constructed the Osteosarcoma Database. It provides a structured and review-like overview on current OS knowledge with the possibility to rank and sort the literature according to various parameters, including therapeutic and prognostic value of specific genes and microRNAs and the type of samples used. Information of genes and microRNAs in OS was collected by automated literature mining and manual review and annotation of PubMed abstracts. This information was further enriched by determining microRNA–target gene interactions (MTIs) of all collected candidates related to OS.

Database Construction

The Osteosarcoma Database aims to provide a high-quality collection of genes and microRNAs implicated in the pathogenesis of OS, reviewed by experts of the field. The data collection and processing steps are illustrated in Figure 1. The workflow comprised three major steps: automated dictionary-based gene and microRNA recognition, manual review and annotation and data storage. The pipeline was based on PubMed abstracts that contained the keywords ‘osteosarcoma*’ or ‘osteogenic sarcoma*’ in their titles and/or abstracts. They were downloaded with the R package XML (17) via NCBI’s E-utilities. Only abstracts written in English and involving human data or specimens were considered. The last download of abstracts was executed on 29 October 2013. In total, 9908 PubMed abstracts were obtained and served as initial corpus for further processing.

Dictionary-based gene and microRNA recognition

To reduce the time-consuming process of manual review and annotation, a dictionary-based gene and microRNA recognition was performed on the initial corpus of abstracts. The dictionary of human genes was compiled from the Human Genome Organisation (HUGO) gene nomenclature committee (18) and the National Center for Biotechnology Information (NCBI) Entrez gene database (19). Official symbols, aliases, synonyms, descriptions, names and database accessions of all genes were combined to generate the gene dictionary with the Entrez geneid as unique identifier. The gene dictionary was extended by...
Textual variants of genes (e.g. IL6, IL-6 or IL-6) to be as complete as possible. Ambiguous synonyms and frequent English words according to the stop words function of the R package tm (20) were excluded to avoid inaccurate gene recognitions. In case of microRNAs, regular expressions like ‘mir’, ‘miR’, ‘MIR’, ‘miRNA’ and ‘microRNA’ were used for entity recognition. The miRBase (21) accessions of mature microRNA sequences served as unique identifiers.

Genes included in the dictionary were identified in the initial corpus of abstracts by string matching and the microRNAs by regular expressions using the R package tm (20). Abstracts without any gene or microRNA occurrence were excluded from further processing, e.g. abstracts of epidemiologic studies. The remaining abstracts were manually reviewed and annotated according to their functional role in the OS.

Manual review and annotation

During the manual review and annotation step, the reviewers verified the specific genes and microRNAs recognized in the abstracts. Additionally, information about experimental settings, the biological context and therapeutic and prognostic impact was marked. The experimental settings comprised the experimental procedure, name of cell lines and kind of samples. Abstracts dealing with human OS cell lines but describing anything but OS biology were excluded.

To provide as much information as possible, we mapped OS-related genes and microRNAs to external databases like NCBI Entrez gene (19), Ensembl (22), Online Mendelian Inheritance in Man (OMIM) (23), Gene Ontology (24), Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway (25) and miRBase (21). Furthermore, the OS-related literature derived from PubMed (26) was linked to each gene and microRNA entry.

As microRNA regulation has become a major subject of OS research, we determined possible MTIs between OS-related genes and microRNAs. Predicted microRNA targets were computed by running the local perl scripts targetscan_60.pl and targetscan_61_context_scores.pl that were downloaded from the TargetScan Web site (http://www.targetscan.org/) (27). Mature microRNA sequences

Figure 1. Database construction pipeline. The database construction is performed in three major steps: automated dictionary-based literature mining, data review and annotation by reviewers and external data sources and data storage in a MySQL relational database with Web interface. The whole pipeline is based on PubMed-derived abstracts related to OS research.
were gained from miRBase release 20 (21). To obtain high-efficacy targets, we excluded target predictions with a context score $>-0.1$ (27).

**Data storage**

To store and access the collected information on OS-related genes including their gene products and microRNAs, we implemented a database and a user-friendly Web interface. The Osteosarcoma Database is a MySQL relational database. The database scheme is illustrated in Supplementary Figure S1. To easily access OS-related genes and microRNAs, users can search and browse via a Web interface at http://osteosarcoma-db.uni-muenster.de. It is built on PHP and JavaScript. For interactive visualization, we applied tagcanvas (http://www.goat1000.com/tagcanvas.php) and cytoscapeWeb (28). Alternatively, users can download the Osteosarcoma Database sql file to perform their own queries. The download link is provided at http://osteosarcoma-db.uni-muenster.de/download.php.

**Database Description**

The Osteosarcoma Database allows retrieving information of candidate genes including their gene products and microRNAs associated with the pathogenesis of OS to support their individual research purposes. Beside gene and microRNA information derived from external databases, manual annotations of OS-related abstracts are provided. Annotations include the number of abstracts focusing on the specific genes with their gene products and microRNAs, the experimental procedures conducted in distinct studies, the potential therapeutic and prognostic value of genes and microRNAs, the specific data types and the biological context investigated. Additionally, regulatory MTIs between collected microRNAs and genes were added. Currently, the database contains 911 genes including their gene products and 81 microRNAs associated with osteosarcoma biology according to 1331 abstracts. Between these microRNAs and genes, we determined 6305 regulatory MTIs due to TargetScan 6 (27).

The database can be searched using the Web interface (http://osteosarcoma-db.uni-muenster.de) with two possible input forms depending on the user’s research focus. For gene search, Entrez geneids and official gene symbols are accepted. MicroRNAs require miRBase accessions or names of mature microRNA sequences. A search for word components is also possible. After submitting the query, suggestions of genes or microRNAs are presented matching the search term. Users can select their requested entry and the results page is displayed.

The main results page lists general information of the requested gene or microRNA. Underscored entries provide links to respective external databases. Below the general gene or microRNA information, a table marks the abstracts describing the gene’s or microRNA’s involvement in the pathogenesis of OS. The abstracts can be filtered according to potential therapeutic and prognostic value and according to tumor samples. Further annotation of experimental settings and biological contexts is provided for download using the export button on top of the table. To note, even if the selection of abstracts was initially based on gene names, we also included experiments involving their gene products such as immunohistochemistry and western blots. However, gene symbols are used as unique identifiers for each gene and/or gene product. Moreover, regulatory MTIs of a specific query are accessible via the MTI button on top of the results page. This button directs the user to predicted microRNA target gene networks. For microRNAs, all target genes are visualized, and for genes, the microRNAs that regulate the respective genes are presented. The network can be explored by zooming in and out or drag and drop nodes. Below the network, details of TargetScan predictions are given. Figure 2 illustrates the main results page and the MTI network using the example of the gene CDKN1A.

Alternatively, the user can browse collected genes, microRNAs and abstracts stored in the database. The last column of all browse tables provides a link to the main results page of the respective gene or microRNA. To visually explore genes including gene products frequently mentioned in OS-related literature, a tagcloud of the top genes was implemented. Just genes mentioned in at least five PubMed abstracts are visualized as top genes. By clicking on gene names, the user is again directed to the main results page for the specific gene.

If we miss specific genes or publications about osteosarcoma, users are welcome to suggest them to us via a contact form, and we are pleased to add them to the database. A graphical guide through the Osteosarcoma Database is available for download on the database Web site at http://osteosarcoma-db.uni-muenster.de/php/tutorial.pdf.

**Discussion and Future Directions**

The ongoing research to detect genes or pathways frequently altered in OS and the search for new therapeutic and prognostic procedures is hampered by the genetic complexity of OS. It becomes even more complicated because of the ever increasing literature about studies of OS that make literature research highly time-consuming. Therefore, it is necessary to structure the existing knowledge of genes and microRNAs associated with OS.
On that account, we developed the Osteosarcoma Database to supply a review of the current state of OS research and made this information easily accessible to researchers.

Pathway enrichment analysis on osteosarcoma-related genes

To evaluate the content of the Osteosarcoma Database regarding its functional association to cancer, we performed a KEGG pathway enrichment analysis. All Entrez genes in the human genome were used as a background set. The hypergeometric test was computed to find significantly overrepresented categories (false discovery rate <0.05). The top 20 enriched pathways are listed in Table 1.

The enrichment results show that the collected OS genes are overrepresented in cancer-related pathways. This indicates that in OS, many well-known oncogenes (e.g. MYC) and tumor suppressor genes (e.g. TP53 and PTEN) are altered. Furthermore, the TGFB signaling pathway is discussed for its contribution to tumor suppression and progression, (29) and the terms apoptosis, cell cycle and focal adhesion represent key signaling pathways in cancer (hallmarks of cancer) (30). Interestingly, we also detected the osteoclast differentiation pathway. In a normal bone, there is a precisely regulated balance between osteoclastic and osteoblastic activity. In OS, this critical balance might be interrupted (31). Taken together, these results indicate OS to require pathways commonly deregulated in cancer as well as to feature OS-specific alterations comprising deregulated osteoclast differentiation.

All properties of OS mentioned earlier are included in the Osteosarcoma Database in terms of OS-related genes, supporting the quality of this collection.
The table shows the results of the hypergeometric test of KEGG pathways.

aFDR, false discovery rate.

Table 1. KEGG pathway enrichment analysis

<table>
<thead>
<tr>
<th>ID</th>
<th>KEGG pathway</th>
<th>Number of genes</th>
<th>Number of genes in pathway</th>
<th>P-value</th>
<th>FDR(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa05200</td>
<td>Pathways in cancer</td>
<td>158</td>
<td>327</td>
<td>5.74 × 10^{-48}</td>
<td>1.11 × 10^{-45}</td>
</tr>
<tr>
<td>hsa05215</td>
<td>Prostate cancer</td>
<td>59</td>
<td>89</td>
<td>8.83 × 10^{-28}</td>
<td>8.57 × 10^{-26}</td>
</tr>
<tr>
<td>hsa05219</td>
<td>Bladder cancer</td>
<td>33</td>
<td>42</td>
<td>9.62 × 10^{-20}</td>
<td>6.22 × 10^{-18}</td>
</tr>
<tr>
<td>hsa05212</td>
<td>Pancreatic cancer</td>
<td>44</td>
<td>70</td>
<td>1.30 × 10^{-19}</td>
<td>6.31 × 10^{-18}</td>
</tr>
<tr>
<td>hsa04510</td>
<td>Focal adhesion</td>
<td>82</td>
<td>200</td>
<td>4.58 × 10^{-19}</td>
<td>1.78 × 10^{-17}</td>
</tr>
<tr>
<td>hsa05222</td>
<td>Small-cell lung cancer</td>
<td>46</td>
<td>85</td>
<td>8.39 × 10^{-17}</td>
<td>2.62 × 10^{-15}</td>
</tr>
<tr>
<td>hsa05220</td>
<td>Chronic myeloid leukemia</td>
<td>42</td>
<td>73</td>
<td>9.45 × 10^{-17}</td>
<td>2.62 × 10^{-15}</td>
</tr>
<tr>
<td>hsa05210</td>
<td>Colorectal cancer</td>
<td>38</td>
<td>62</td>
<td>1.46 × 10^{-16}</td>
<td>3.55 × 10^{-15}</td>
</tr>
<tr>
<td>hsa04110</td>
<td>Cell cycle</td>
<td>58</td>
<td>128</td>
<td>3.74 × 10^{-16}</td>
<td>8.07 × 10^{-15}</td>
</tr>
<tr>
<td>hsa04330</td>
<td>TGF-beta signaling pathway</td>
<td>44</td>
<td>85</td>
<td>3.55 × 10^{-15}</td>
<td>6.89 × 10^{-14}</td>
</tr>
<tr>
<td>hsa05223</td>
<td>Non-small-cell lung cancer</td>
<td>33</td>
<td>54</td>
<td>1.72 × 10^{-14}</td>
<td>3.04 × 10^{-13}</td>
</tr>
<tr>
<td>hsa04115</td>
<td>p53 signaling pathway</td>
<td>38</td>
<td>69</td>
<td>2.01 × 10^{-14}</td>
<td>3.25 × 10^{-13}</td>
</tr>
<tr>
<td>hsa04210</td>
<td>Apoptosis</td>
<td>44</td>
<td>89</td>
<td>3.12 × 10^{-14}</td>
<td>4.66 × 10^{-13}</td>
</tr>
<tr>
<td>hsa05214</td>
<td>Glioma</td>
<td>36</td>
<td>65</td>
<td>7.85 × 10^{-14}</td>
<td>1.09 × 10^{-12}</td>
</tr>
<tr>
<td>hsa05213</td>
<td>Endometrial cancer</td>
<td>31</td>
<td>52</td>
<td>2.86 × 10^{-13}</td>
<td>3.70 × 10^{-12}</td>
</tr>
<tr>
<td>hsa05218</td>
<td>Melanoma</td>
<td>37</td>
<td>71</td>
<td>4.46 × 10^{-13}</td>
<td>5.41 × 10^{-12}</td>
</tr>
<tr>
<td>hsa05142</td>
<td>Chagas’ disease (American trypanosomiasis)</td>
<td>46</td>
<td>104</td>
<td>1.37 × 10^{-11}</td>
<td>1.57 × 10^{-11}</td>
</tr>
<tr>
<td>hsa05221</td>
<td>Acute myeloid leukemia</td>
<td>31</td>
<td>58</td>
<td>1.53 × 10^{-11}</td>
<td>1.65 × 10^{-10}</td>
</tr>
<tr>
<td>hsa04380</td>
<td>Osteoclast differentiation</td>
<td>50</td>
<td>128</td>
<td>4.04 × 10^{-11}</td>
<td>4.12 × 10^{-10}</td>
</tr>
<tr>
<td>hsa04012</td>
<td>ErbB signaling pathway</td>
<td>39</td>
<td>87</td>
<td>4.39 × 10^{-11}</td>
<td>4.26 × 10^{-10}</td>
</tr>
</tbody>
</table>

Many cancers and activate TP53 regulated pathways. This microRNA family was extensively tested for its therapeutic use in several tumors and might be the first microRNA family to reach the clinic (35).

Up to now, the prognostic prediction or therapeutic stratification of OS is not based on biomarkers. However, the table suggests many promising candidates that should be further investigated and sometime enter clinical studies.

**Osteosarcoma-related microRNA target gene regulation**

Much attention has been focused on microRNAs in the pathogenesis of OS as a new tool for assisting prognosis or therapy. They function through multiple pathways simultaneously, which is in accordance with the perspective on cancer as a disease affecting the whole cellular system. For the collected data, we determined potential MTIs by using TargetScan 6 (27). All microRNAs affecting the largest number of genes (≥100 targets) are shown in Table 3. Again, members of the microRNA family microRNA-34 are listed in the table. They regulate the highest number of target genes collected in the Osteosarcoma Database supporting a crucial role in OS as well as in other cancer types. Further, the remaining microRNAs are also known to function as tumor suppressors or oncomirs, e.g. the microRNA families microRNA-29 and -15. Both families have several members involved in various cancer subtypes (36, 37).
As already mentioned, microRNA research is a young field and not much is known about their function in OS. Thus, we provide detailed and up-to-date networks about possible MTIs to researchers for hypothesis generation and testing of individual models.

Future directions
Currently, the Osteosarcoma Database focuses on genes including their gene products and microRNAs associated with OS development and progression. However, the OS is a complex tumor with a huge amount of genomic instability that influences the expression and function of several genes and microRNAs. Hence, genomic alterations need to be added in future versions. We plan to include already known genomic positions marking regions of copy number variations, allelic imbalances and translocations, as it has been shown that structural chromosomal alterations could be used to predict prognosis at diagnosis (2). Moreover, observations of genome-wide changes from next-generation sequencing studies might further obtain new insights into OS biology and must be added as soon as they are available.

We plan to update the database biannually to provide state-of-the-art knowledge and keep track of improvements in the field. We hope that the Osteosarcoma Database will serve as a platform for information and hypothesis generation for the research community that helps to uncover the complexity of OS.

Supplementary data
Supplementary Data are available at Database Online.

Acknowledgements
K.P. and E.K. designed the study. K.P. implemented the Osteosarcoma Database and wrote the article. K.P., J.S., M.N., D.M., DB, A.N. and E.K. manually reviewed and annotated the PubMed abstracts and participated in discussions and preparation.
of the manuscript. The authors thank Christian Ehrhart from anderthalh.com for support in Web design.

FUNDING

This work was funded by the Translational Sarcoma Research Network [FKZ 01GM0870 to JS, MN, DM and FKZ 01GM0869 to KP, EK], the European TRANSCAN I consortium - PROSpective Validation of Biomarkers in Ewing Sarcoma for personalized translational medicine both supported by the BMBF [FKZ 01KT1310 to K.P.E.K.] and the Foundation for the Preservation of the Basel Bone Tumor Center [no DB]. The authors acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publication Fund of the University of Münster. Funding for open access charge: Deutsche Forschungsgemeinschaft (DFG) and Open Access Publication Fund of the University of Münster (WWU).

Conflict of interest. None declared.

References