Osteoarthritis Data Integration Portal (OsteoDIP): A web-based gene and non-coding RNA expression database

Chiara Pastrelloa,b, Mark Abovskya,b, Richard Lu1,a, Zuhaib Ahmeda,b, Max Kotlyara,b, Christian Veillettea, Igor Jurisica

A Osteoarthritis Research Program, Division of Orthopedic Surgery, Schroeder Arthritis Institute, University Health Network, Toronto, Canada
B Data Science Discovery Centre for Chronic Diseases, Krembil Research Institute, Toronto Western Hospital, University Health Network, Toronto, Canada
C Departments of Medical Biophysics and Computer Science, University of Toronto, Toronto, ON, Canada
D Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia

ABSTRACT

Objective: OsteoDIP aims to collect and provide, in a simple searchable format, curated high throughput RNA expression data related to osteoarthritis.

Design: Datasets are collected annually by searching “osteoarthritis gene expression profile” in PubMed. Only publications containing patient data and a list of differentially expressed genes are considered. From 2020, the search has expanded to include non-coding RNAs. Moreover, a search in GEO for “osteoarthritis” datasets has been performed using ‘Homo sapiens’ and ‘Expression profiling by array’ filters. Annotations for genes linked to osteoarthritis have been downloaded from external databases.

Results: Out of 1204 curated papers, 63 have been included in OsteoDIP, while GEO curation led to the collection of 28 datasets. Literature data provides a snapshot of osteoarthritis research derived from 1924 human samples, while GEO datasets provide expression for additional 1012 patients. Similar to osteoarthritis literature, OsteoDIP data has been created mostly from studies focused on knee, and the tissue most frequently investigated is cartilage. GEO data sets were fully integrated with associated clinical data. We showcase examples and use cases applicable for translational research in osteoarthritis.

Conclusions: OsteoDIP is publicly available at http://ophid.utoronto.ca/OsteoDIP. The website is easy to navigate and all the data is available for download. Data consolidation allows researchers to perform comparisons across studies and to combine data from different datasets. Our examples show how OsteoDIP can integrate with and improve osteoarthritis researchers’ pipelines.

ARTICLE INFO

Keywords: Data integration Gene expression Long non-coding RNA microRNA integrative computational biology

1. Introduction

High throughput data have become essential for unbiased investigations of important biomedical research questions in the last two decades. The type, amount and quality of data analyzed has progressively increased, providing biomedical researchers with the resources needed to create a molecular view of the system being studied gradually closer to a “whole picture”. Every set of data collected contributes to advance our understanding of complex diseases, including osteoarthritis (OA). Nonetheless, it is only one stroke in a much bigger painting, mostly because OA is not homogeneous but rather represents a spectrum of diseases: patients with the same disease may respond differently to the same treatment and the disease may progress differently due to patient heterogeneity. Individual biological assays can detect a high number of molecules (for example proteins, microRNAs, metabolites, gene expression quantification) or their status (for example mutation, methylation or post-transcriptional modification), but in a cell all these molecules act in concert, and individual patients are characterized by a combination of...
these perturbations. To fathom how they influence health and disease state, we need to combine them using corresponding networks – microRNA:gene, protein:protein, pathways, etc. Furthermore, even considering only one type of data, different datasets can include patients with the same disease, but might be trying to answer different questions, or include patients with different clinicopathological features. However, even when the clinical question and the clinicopathological features are the same, technical and biological heterogeneity can lead to different results – most of which are still valuable, but represent small (and sometimes redundant) pieces of the disease’s molecular puzzle [1].

Intuitively, data integration across datasets and molecules is key to gather more information and depict a more complete picture. Many disease-specific databases already harness this potential, especially for cancer (the most well-known being cBioPortal, that includes, among others, the huge collection of data from TCGA [2]).

In OA, many proteomics, gene and non-coding RNA expression, methylation, metabolomics and genome-wide association studies (GWAS) have been performed (reviewed in Ref. [3]). Resources collecting and annotating omics data in OA, though, are limited. SkeletonVis collects and re-analyses transcriptomics datasets linked to skeletal diseases, including, but not focusing on, OA [4]. A researcher can use the database to calculate fold change and related enrichment analyses in one dataset at a time, or compare different datasets one gene or one signature at a time. OA Targets collects data on model organisms and the effect of gene manipulation on their OA phenotypes [5]. In this paper we present OsteoDIP, a database collecting and annotating OA-specific omics data from human studies.

2. Method

2.1. Data collection

OsteoDIP primarily focuses on genes found to be linked to OA using high-throughput techniques (for example microarray or RNAseq). To collect such data, we performed this exact search in PubMed: “osteoarthritis gene expression profile”. Only papers that included patient diagnosis annotation for at least 4 patients and that provided the list of differential genes were collected. The search was first performed in October 2016, with annual updates. At each update cycle, we verify whether any paper has been retracted and, in such case, remove it from the database. In 2020, the search was extended to include also non-coding RNA molecules, using the same collection of articles as the ones used in the recent review [6]. Gene symbols are updated every release to the latest HGNC version [7], microRNAs to the latest miRBase [8] and long non-coding RNAs to the latest LNCipedia version [9].

With OsteoDIP, we provide a platform and initial curation with the aim to make available high-quality data for translational research in osteoarthritis. To ensure both high data quality and coverage, we opened the platform for contributed curation. For example, OsteoDIP includes 18 low-throughput protein expression and SNP data sets and radiographic biomarkers data from 5 studies curated and provided by Dr. Stuart Faulkner, Center for the Advancement of Sustainable Medical Innovation, Oxford University.

2.2. GEO data collection and normalization

Datasets were curated from GEO [10] using ‘osteoarthritis’ as a search term and ‘Homo sapiens’ and ‘Expression profiling by array’ as filters. For curation, the same criteria used for data collection were applied. Moreover, we checked that no overlap among samples (based on GSM identifier) was present. Series matrix files were then downloaded for the datasets of interest, and expression and clinical data were separated. For each expression table, if probeset to gene symbol mapping was not provided, it was created using the annotations present in GEO platform pages. For each clinical table, data was consolidated to be searchable in the database (for example, sex was transformed to F and M for all datasets, replacing “Female”, “female”, “Male” or “male”).

If raw data was available, normalization was performed using R 4.0.3 [11] with packages limma 3.46.0 [12], affy 1.68.0 [13] or oligo 1.54.1 [14]. Most of the datasets were RNA-normalized, but the full list of datasets and their normalization methods is available in Supplementary Table 1.

2.3. Annotation

OsteoDIP provides, for each gene of interest present in at least one of the curated papers, a set of annotation data from external databases. Disease annotation is collected from DisGeNET [15], protein secretion data from The Human Protein Atlas [16] and MetaZeeKB [17], SNPs from the GWAS Catalog [18], human protein-protein interaction (PPI) data from the Integrated Interactions Database (IID) ver. 2020–11 (with interaction annotations synovial macrophages, chondrocytes, growth plate cartilage, synovial membrane or articular cartilage) [19]. We also provide the number of conserved PPIs per species: conserved PPIs are determined by mapping experimentally detected human PPIs to orthologous protein pairs in 17 other species. Mappings are based on 1:1 orthologs from Ensembl [20] release 100.

2.4. Database

The web interface to the OsteoDIP database is implemented in the Java Server Faces framework running on IBM WebSphere Application Server (ver. 9.0). The backend storage deployed IBM DB2 database (ver. 11.1) engine. For performance improvement, the WebSphere and DB2 are placed on different virtual instances of IBM P770 and P750, running AIX (ver. 7.2). The OsteoDIP Data Integration Portal (DIP) is freely available at http://ophid.utoronto.ca/OsteoDIP, and online documentation provides more details for every page.

2.5. Use cases

Descriptive analyses. Descriptive analyses of OsteoDIP have been performed using search results and tables obtained from the website at the “Matrix” page. Top deregulated genes were extracted from the matrix of all deregulated molecules in OsteoDIP, and the microRNAs targeting them were obtained from mirDIP [21] (http://ophid.utoronto.ca/mirDIP) using the threshold “very high”. Human PPI data among top deregulated genes were obtained from the Integrated Interactions Database (IID) ver. 2020–11 (http://ophid.utoronto.ca/iid).

OA signature. An OA signature was collected from Ref. [22]. Gene names were used to query the GEO page in OsteoDIP, and the downloaded sets of tables (clinical, expression and normalized expression data) were used in R 4.0.3 to calculate fold changes and moderated t-test p-values using the limma package ver. 3.44.3. Datasets were used if they had patients and healthy controls, at least 3 independent samples per group, and patients were not receiving treatments. Genes that had a significant p-value were plotted using ggplot2 ver. 3.3.3.

MicroRNA review. A recent review listed the microRNAs that have a protective or destructive role in OA [24]. We collected the microRNAs, converted them to miRbase v.22 IDs using miRBaseConverter 1.12 in R 4.3.0, and searched for them in the microRNA search page of OsteoDIP. Furthermore, we investigated the targets of such microRNAs present in the network depicted in Fig. 2 of the review. All the genes were searched for in OsteoDIP. We next investigated the overlap between the two lists of microRNAs and the microRNAs targeting top deregulated genes from OsteoDIP. Hypergeometric distribution test was performed in R 4.0.3. A network depicting the microRNAs from the two lists targeting OA genes has been built using NAViGaTOR 3.0.14 [25]. Conservation of the network’s protein:protein interactions in different species was obtained from IID ver. 2020–11.

Hip and knee cartilage comparison. To investigate possible differences between hip and knee expression, we analyzed the matrices
including all the protein coding genes that were deregulated in cartilage hip OA and the ones deregulated in cartilage knee OA. Overlap and differences between the two sets of genes were calculated in R 4.0.3. Pathway enrichment analysis for each specific set (genes deregulated only in hip or genes deregulated only in knee) was performed in pathDIP 4.1 [26] selecting BioCarta, EHMN, HumanCyc, INOH, IPAVS, NetPath, Panter_Pathways, PID, REACTOME, Signalink2.0, SIGNOR 2.0, Spike, STKE, systems.biology.org, Uniprot_Pathways and WikiPathways as sources. Enrichment was conducted using the sources separately, and only pathways with adjusted p-value (False Discovery Rate, BH method) < 0.01 were retained.

MALAT1, MALAT1 targets were retrieved from LncTarD [27]. Of the retrieved targets, microRNAs were then used to query mirDIP (threshold “very high”) to obtain microRNA:gene interactions. Only gene targets of the microRNAs or of MALAT1 that were present in OsteoDIP were retained. A network was created using NAViGaTOR 3.0.14. Pathway enrichment analysis of gene targets was performed in pathDIP 4.1 using the same databases listed above. Enrichment was computed using the combined pathway source databases, and only pathways with adjusted p-value (False Discovery Rate, BH method) < 0.01 were retained. Each gene was then annotated with the pathway from the enriched list with the lowest p-value.

3. Results

3.1. Data

We have considered 1204 papers as of December 2020 (Fig. 1A), 63 of which have been collected in OsteoDIP after necessary exclusions. While many papers have been excluded due to our curation choices (for example non-original data, model organism data), it stands out that 135 papers were excluded because the data were not available – highlighting once more how the lack of data sharing affects curation efforts worldwide. As visible in Fig. 1B, OsteoDIP reflects the literature distribution, with most papers focusing on knee OA, and more frequently on cartilage tissues.

GEO curation led to the collection of 28 datasets. Consolidated clinical data from such datasets is available at the OsteoDIP GEO Clinical page, which shows that the datasets provide expression data for 1012 patients with heterogeneous types of data. Curation revealed that less than half of the datasets (13/28) include age annotation for the samples and only one (GSE15227) includes grade. The collection of papers, on the other hand, provides a view of OA obtained from 1924 samples (from different comparisons, but most frequently – in 39 papers – OA samples are compared to healthy controls). Four studies are present both in the curated and in the GEO dataset pages (PMID: 16508983, 24229462, 29258882, 29973527).

Molecules deregulated in at least one study include 8905 genes, 402 lncRNAs, 56 microRNAs and 58 circRNAs. The distribution of gene deregulation and its direction is shown in Supplementary Fig. 1.

3.2. Use cases

OsteoDIP can be used for different types of studies, for example:

- **Study specific genes of interest, where they have been published, to which conditions they have been linked, and which of their interactions are conserved across species.**
- **Study genes linked to specific tissues and/or joints and/or comparisons.** We show this in: “Hip and knee cartilage comparison” case.
- **Study specific noncoding RNAs of interest, where they have been published, to which conditions they have been linked and, in the case of microRNAs, which genes they target. Two cases are shown: “MALAT1” and “microRNA review”.
- **Perform analyses on OA related datasets, using consolidated annotation data that facilitates comparisons across datasets.** We provide an example in “OA signature”

### Descriptive analyses.

The most frequently downregulated molecule is the gene APOD (9 studies), while the most frequently upregulated is the gene COL5A1 (14 studies). Searching for APOD in OsteoDIP, we can see that it is secreted, and that it interacts with 108 other proteins. Of these interactions, 84 are conserved in mouse, 83 in cat and guinea pig, and 82 in cow and rat, suggesting they would be the best animal models to study APOD’s effect on OA. Of the 108 interactors, 52 are listed as deregulated in at least one study in OsteoDIP. Similarly, searching for COL5A1, we can see that it is secreted as well, and that it is annotated with a score of 0.73 with the disease Ehlers-Danlos syndrome type 1. There are 317 COL5A1 protein interactions reported in IID. Of the 317 interactors, 256 are annotated with at least one study where they are deregulated, but the highest number of conserved interactions is only 43, in mouse and pig.

We then explored top deregulated molecules. Focusing on those deregulated in at least 8 studies, we identified 138 genes that are connected by 624 PPIs, 607 of which are annotated with synovial or cartilage specific tissues. Table 1 shows that most PPIs are conserved across mammals, with cat being the species with the highest number of conserved PPIs.

**MicroRNA review.** Using mirDIP, we identified 701 microRNAs that target the top genes described above. OsteoDIP includes 41 of these microRNAs, providing further evidence of their importance to OA. To
further annotate the microRNAs, we looked at the overlap between them and the protective and destructive microRNAs [24]. Hypergeometric test provides evidence that the top genes are significantly targeted by the reviewed microRNAs (p-value 5.485815e-08 for destructive microRNAs and 3.682239e-10 for protective ones). A network built using such overlapping microRNAs and their gene targets present in OsteoDIP shows 18 genes targeted only by protective microRNAs, among which the most downregulated gene is CHI3L1, a gene that supports OA progression facilitating ECM degradation through MMP9 and that degrades key proteins such as proteoglycan, collagen and osteonectin [28]. Furthermore, 10 genes are targeted only by destructive microRNAs, among which the most downregulated gene is NQO1, an antioxidant enzyme regulated by Nrf2 and involved in preventing cartilage degradation [29] (Fig. 2).

We then looked for the microRNAs present in the review: 6 cartilage protective (hsa-miR-24–3p, hsa-miR-27a-3p, hsa-miR-27b-3p, hsa-miR-193b-3p, hsa-miR-210–3p and hsa-miR-30a-5p) and 10 cartilage destructive (hsa-miR-139–5p, hsa-miR-181a-5p, hsa-miR-23a-3p, hsa-miR-34a-5p, hsa-miR-4454, hsa-miR-203a-3p, hsa-miR-223–3p, hsa-miR-302b-3p, hsa-miR-381–3p and hsa-miR-483–5p) are found in OsteoDIP, all being annotated as deregulated in only one paper each. Among the 997 genes identified in the review targeted only by cartilage-destructive microRNAs, 390 were found to be deregulated in one or more studies. Interestingly, RNF34, highlighted in the text for its cartilage-destructive link, has been shown to be downregulated in two studies comparing OA knee cartilages to healthy controls. 833 of 1854 genes targeted only by cartilage-protective microRNAs were also found to be deregulated in at least one study. HAS3 has been highlighted for its link to protective microRNAs and it was found downregulated in one study where normal knee synovial tissue was compared to inflamed areas of OA knee. Almost 47% of all the targets of the reviewed microRNAs are described as deregulated in at least one OA study, suggesting a strong connection between the microRNAs, the genes and OA.

**OA signature.** A recent blood diagnostic signature has been shown to separate OA from healthy samples [22]. A researcher might be interested to know whether the identified genes have a role in OA pathogenesis; thus, we tested in which datasets such genes were differentially expressed in OA compared to healthy individuals. Using the 4 listed genes, we collected from OsteoDIP their expression from 28 GEO datasets. All 4 genes were present, and 6 datasets passed our filtering criteria. Table 2 shows the number of samples and distribution of expression in healthy and OA samples in each dataset. As expected, there was variability across the datasets, and IL18 and SRSF2 were the only genes with significant differential expression (Fig. 3). SRSF2 was differential in a dataset that studied meniscal tissue (GSE98918) and IL18 in a dataset that examined synovial tissues (GSE82107). Interestingly, the two datasets with the strongest differential expression for the genes of interest, GSE82107 and

### Table 1

| Species | # PPIs |
|---------|--------|
| Cat     | 599    |
| Horse   | 574    |
| Sheep   | 573    |
| Cow     | 571    |
| Pig     | 555    |
| Mouse   | 549    |
| Dog     | 544    |
| Guinea pig | 521    |
| Rabbit  | 472    |
| Rat     | 464    |
| Chicken | 404    |
| Turkey  | 357    |
| Duck    | 309    |
| Worm    | 7      |
| Fly     | 3      |
| Yeast   | 1      |
| Alpaca  | 0      |
GE143514, were both derived from synovial tissues, while the remaining datasets were derived from meniscus, synovial fibroblasts and cartilage tissues. This suggests a possible connection between synovial tissue OA molecular landmark and the OA blood signature genes. This also further highlights the benefit of diverse samples and their rich annotation.

**Hip and knee cartilage comparison.** Joint-specific OA pathogenesis has been hypothesized [31], but not many studies compare data from and mechanisms related to different joints [32–36]. In OsteoDIP, as in the literature, most studies focus on knee, but other joints are available for comparison. For example, we compared the genes deregulated in at least one study using knee cartilage (6793 genes) and hip cartilage (1248 genes) samples. 859 genes are in common between the two sets, and one study using knee cartilage (6793 genes) and hip cartilage (1248 genes) samples. 859 genes are in common between the two sets, and 5936 are deregulated only in studies that focus on knee cartilage. Metabolic differences in synovial fluid of knees and hips have been described, and in particular N-acetylated molecules, glycosaminoglycans, citrate and glutamine [35]. Pathway enrichment analysis results of non-overlapping pathways are available in Supplementary Table 3.

MALAT1. 786 out of 795 long non-coding RNAs present in OsteoDIP have been annotated as deregulated only in one study. Of the few long non-coding RNAs identified in multiple studies, MALAT1 is the top deregulated (three studies). We obtained experimental interactions of MALAT1 targets are also found in OsteoDIP, we created the network in Fig. 4: five MALAT1 targets are also targets of the microRNAs, creating regulation loops. One microRNA (hsa-miR-9-5p) is present in OsteoDIP and 4 microRNAs (hsa-miR-9-5p, hsa-miR-127–5p, hsa-miR-145–5p, hsa-miR-146–5p) have been linked in the literature to OA via MALAT1 regulation. In OA papers [40] [41] [45], MALAT1 is shown to affect proliferation as well as ECM degradation. Pathway enrichment analysis of OsteoDIP targets of the 4 microRNA finds 27 enriched pathways, mainly linked to ECM, collagen formation and degradation and integrin signaling (available in Supplementary Table 4).

4. Discussion

Data curation provides a scientific asset to any research topic, but curation remains challenging due to frequent data unavailability, missing and limited clinical and biological annotation, limited biological assays or improper informatics workflows [46]. In OA many high-throughput studies have been and are being conducted to identify molecular pathogenesis paths, characterize patient heterogeneity and predict new (and effective) OA treatments. One feature of high-throughput studies is the

---

Table 2: Features of datasets used in OA signature example.

| Feature | GSE117999 | GSE143514 | GSE19060 | GSE29746 | GSE82107 | GSE98918 |
|---------|-----------|-----------|-----------|-----------|-----------|-----------|
| OA samples | 12 | 5 | 5 | 11 | 10 | 12 |
| Healthy samples | 12 | 3 | 3 | 11 | 7 | 12 |
| Mean CCR6 OA | 13.0832 | 0.4914 | 3.6073 | 8.2111 | 3.4547 | 13.2728 |
| SD | 1.0046 | 0.089 | 0.1007 | 0.2236 | 0.1211 | 0.3484 |
| coefficient of variation, SD | 0.0712 | 0.774 | 0.0317 | 0.0068 | 0.14 | 0.0323 |
| mean CLEC7A OA | 8.0407 | 13.177 | 4.0879 | 7.9584 | 8.3056 | 7.9595 |
| SD | 0.5727 | 10.1985 | 0.1294 | 0.0543 | 1.163 | 0.2572 |
| coefficient of variation, SD | 0.0317 | 0.0117 | 0.0793 | 0.0307 | 0.0117 | 0.0223 |
| mean IL18 OA | 7.9868 | 22.6703 | 3.4045 | 8.5381 | 4.7223 | 8.1281 |
| SD | 0.3155 | 3.79 | 0.0946 | 0.348 | 0.4194 | 0.2498 |
| coefficient of variation, SD | 0.039 | 0.1672 | 0.0278 | 0.0408 | 0.0888 | 0.0307 |
| mean SRSF2 OA | 11.9468 | 101.7188 | 10.9003 | 14.7229 | 7.2814 | 11.8503 |
| SD | 0.9364 | 10.4502 | 0.3949 | 0.2297 | 0.5771 | 0.3822 |
| coefficient of variation, SD | 0.0313 | 0.578 | 0.0471 | 0.0196 | 0.0472 | 0.0448 |
| mean CCR6 healthy | 13.3711 | 0.4989 | 3.6359 | 8.0657 | 3.5601 | 13.3441 |
| SD | 0.455 | 0.3481 | 0.0246 | 0.1585 | 0.1682 | 0.5972 |
| coefficient of variation, SD | 0.034 | 0.6977 | 0.0068 | 0.0196 | 0.0472 | 0.0448 |
| mean CLEC7A healthy | 7.9932 | 22.4676 | 4.0984 | 7.9655 | 7.2286 | 7.6234 |
| SD | 0.4049 | 6.8424 | 0.1047 | 0.0443 | 1.7401 | 0.2268 |
| coefficient of variation, SD | 0.0507 | 0.3045 | 0.0255 | 0.0056 | 0.2407 | 0.0296 |
| mean IL18 healthy | 8.0039 | 15.2815 | 3.3938 | 8.6473 | 4.0956 | 7.9297 |
| SD | 0.2509 | 8.8324 | 0.16 | 0.5676 | 0.2324 | 0.3027 |
| coefficient of variation, SD | 0.0313 | 0.578 | 0.0471 | 0.0656 | 0.0567 | 0.0382 |
| mean SRSF2 healthy | 11.7984 | 121.4576 | 15.0048 | 7.1226 | 12.2954 |
| SD | 0.4921 | 19.476 | 0.1832 | 0.9023 | 0.5038 | 0.5752 |
| coefficient of variation, SD | 0.0417 | 0.1604 | 0.0162 | 0.0601 | 0.0706 | 0.0468 |

c.o.v = coefficient of variation, SD = standard deviation, OA = osteoarthritis.
amount of data collected – it being related to any set of molecules (i.e., the entire genome, the entire proteome, the entire transcriptome, metabolome). Obtaining high throughput data is expensive, though, leading to the application of these methods to only a reduced number of patient samples. While these data are still valuable, patient heterogeneity, lack of standard formats and annotation for data release (when available), different assays, and different questions being investigated can impede comparisons across studies. Still, each study provides a step to get closer to a more complete picture of OA molecular background, patient subtyping, and precision treatment.

One aim of data curation is to collect and integrate multiple datasets that are scattered across different locations (such as different databases or, as in our case, different publications), and annotate them with the same ontology and rigour. To satisfy this aim, we curated and collected all available literature on gene expression studies, and we curated the most recent studies on non-coding RNA expression. A second aim of data curation is to consolidate the data so that it is comparable across studies and, if possible, patients, and to annotate it with relevant information, such as tissue, disease, interactions and pathways. To this aim, all the data collected have been annotated with standard labels so that it is easily searchable but also, being structured data, amenable to computational analyses.

As highlighted with examples and use cases, OsteoDIP can support diverse translational research projects in OA. If a researcher has already identified molecules of interest (as in the microRNA example), OsteoDIP can provide the researcher with literature and multiple annotations for protein coding genes – all in one database. Such annotations can assist with simple tasks, e.g., providing context for genes of interest, or can provide the basis for further research steps. For example, knowing what part of a PPI network involving the proteins of interest is conserved across species (and which species) can suggest the best animal model where to test a specific mechanism, investigate OA pathogenesis or validate a hypothesized treatment. It is well known that there is not a single animal model that mimics all molecular and clinicopathological aspects of human OA (or indeed any complex disease), and that different models need to be used to answer different questions [47]. Thus, it would be useful to predict beforehand which model organism best recapitulates relevant biological context required for in vivo studies and pre-clinical validation.

Some genes of interest could be queried across different datasets and their gene expression compared (as in the signature genes examples). If needed, the expression could be linked to structured clinical data to provide support for external validations, reducing the time a researcher would need to spend to find the same type of data in more generic data repositories – and the time needed to consolidate the data that, due to lack of standard labels in many databases, is usually quite different from one dataset to another. In our example, we attempted to link genes present in an OA diagnostic blood signature to a possible OA mechanism in other tissues, but other researchers could be interested in using the same data to investigate and validate prognostic or predictive signature performance. Most gene expression signatures do not generalize to new data; over 150,000 studies have reported gene signatures, but fewer than 100 are in clinical use [48]. The best way to find an effective signature is to validate its performance across many independent datasets. Such testing and analysis can answer several key questions about a candidate signature: does the signature reflect biological mechanisms or technical

---

**Fig. 4. MALAT1 network.** MALAT1:target interactions are collected from LncTarD, while microRNA:gene ones are obtained from mirDIP. Blue nodes show miRNAs, while blue outline shows MALAT1 targets described in the literature as linked to OA. Blue edges illustrate the interactions between these microRNAs and their targets. Gene targets of the OA linked microRNAs are color coded according to the pathway with the lowest p-value they belong to.
artifacts, does it work across independent cohorts, and if not, does it work in specific subsets of samples. Testing a signature in more than one dataset greatly reduces the chances that it is based on technical artifacts, such as a protocol for gathering or processing data. Testing on many datasets with heterogeneous samples, including different age groups and both sexes, can indicate whether the signature can easily generalize to new cohorts. A signature may not work in all datasets, because many datasets include multiple conditions, such disease status, drug treatments and comorbidities, that all greatly affect gene expression. It is also important to pay attention to sample independence, that can strongly affect the results obtained. If a signature works in at least a few datasets, it may be possible to determine the context where the signature is effective; but it is equally valuable to know which patient cohorts cannot be reliably analyzed due to signature bias.

Finally, researchers that have OA molecular questions but no genes to start with, could query OsteoDIP and identify genes frequently associated with a feature of interest, or compare genes associated with different characteristics (as in the joint specific OA and the top deregulated genes examples).

We designed OsteoDIP to be flexible, so that many types of searches could be performed, open access, so that any kind of data collected and stored in OsteoDIP is immediately available to researchers, and modular, so that any kind of data of interest for the OA community can be included and the types of curation provided expanded.

Author contribution

CP participated in conception and design of the study, curated the data, performed analyses and data interpretation, drafted the article, and approved the final version of the manuscript. MA and RL created and update the database, collected annotation data, update nomenclature, drafted the article, and approved the final version of the manuscript. ZA collected and normalized GEO data, revised critically the article, and approved the final version of the manuscript. CV final version of the manuscript. RL and MA created and update nomenclature, update nomenclature, final version of the manuscript. ZA final version of the manuscript. RL and MA created and update nomenclature, update nomenclature, final version of the manuscript. CV final version of the manuscript. RL and MA created and update nomenclature, update nomenclature, final version of the manuscript.

Role of the funding source

This work was in part supported by the Krembil Research Foundation (grant to CV and LJ), Schroeder Arthritis Institute via the Toronto General and Western Hospital Foundation, University Health Network. LJ was supported in part by funding from Natural Sciences Research Council (NSERC #203475), Canada Foundation for Innovation (CFI #225404, #30865), Ontario Research Fund (RDI #34876), Buchan Foundation, Ian Lawson van Toch Fund and IBM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of competing interest

At the end of the text, under a subheading “Conflict of interest statement” all authors must disclose any financial and personal relationships with other people or organisations that could inappropriately influence (bias) their work. Examples of potential conflicts of interest include employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications/registrations, and research grants or other funding.

Acknowledgments

We thank Haroon Chaudhry, Zara Malik, Dr. Anne-Christine Hauschild, Andrea Rossos and Dr. Stuart Faulkner for their contribution to data curation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ocarto.2022.100237.

References

[1] L. Cantini, L. Calzone, L. Martignetti, M. Rydenfelt, N. Blüthgen, E. Barillot, A. Zinovyev, Classification of gene signatures for their information value and functional redundancy, Npj Syst. Biol. Appl. 4 (2018) 2. https://doi.org/10.1038/ s41545-017-0038-8.
[2] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobcen, C.J. Byrne, M.L. Heuer, E. Larsson, Y. Antipin, B. Reva, A.P. Goldberg, C. Sander, N. Schultz, The cibio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (2012) 401–404, https://doi.org/10.1158/2159-8290.CD-12-0095.
[3] C. Ruiz-Romero, I. Rego-Perez, F.J. Blanco, What did we learn from “omics” studies in osteoarthritis, Curr. Opin. Rheumatol. 30 (2018) 114–120. https://doi.org/10.1097/BOR.0000000000000460.
[4] J. Soul, T.E. Hardingham, R.P. Boot-Handford, J.-M. Schwartz, SkeletalVis: an exploration and meta-analysis data portal of cross-species skeletal transcriptomics data, Bioinformatics 35 (2019) 2283–2290, https://doi.org/10.1093/bioinformatics/bty947.
[5] J. Soul, M.J. Barker, C.B. Little, D.A. Young, OnTarge: a knowledge base of genes associated with osteoarthritis joint damage in animals, Ann. Rheum. Dis. 80 (2020) 376–383, https://doi.org/10.1136/annrhemdis-2020-218344.
[6] A. Ratemswaran, M. Kapoor, Osteoarthritis year in review: genetics, genomics, epigenetics, Osteoarthrits Cartilage 29 (2021) 151–160. https://doi.org/10.1016/j.joca.2020.11.003.
[7] S. Tweedie, B. Bracchi, K. Gray, T.E.M. Jones, L. Seal, B. Yates, E.A. Bradford, Genenames.org: the HGNC and VGNC resources in 2021, Nucleic Acids Res. 49 (2021) D939–D946. https://doi.org/10.1093/nar/gkaa980.
[8] A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miBase: from microRNA sequences to function, Nucleic Acids Res. 47 (2019), https://doi.org/10.1093/nar/gkz1141. DI155–DI162.
[9] P.J. Volders, J. Anckaert, K. Verheggen, J. Nyuteen, L. Martens, P. Mestdagh, J. Vandesompele, LNCipedia 5: towards a reference set of human long non-coding RNAs, Nucleic Acids Res. 47 (2019), https://doi.org/10.1093/nar/gkz1031. D135–D139.
[10] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (2002) 207–210. https://doi.org/10.1093/nar/30.1.207.
[11] R Core Team, R.A. Language, And Environment for Statistical Computing. 2014. http://www.r-project.org/.
[12] M.E. Ritchie, B. Hipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-seq and microarray studies, Nucleic Acids Res. 43 (2015) e79, https://doi.org/10.1093/nar/gkv107.
[13] L. Goutier, L. Cope, B.M. Boltad, R.A. Izzry, riffy-analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20 (2004) 307–315, https://doi.org/10.1093/bioinformatics/btg405.
[14] B.S. Carvalho, R.A. Izzry, A framework for oligonucleotide microarray preprocessing, Bioinformatics 26 (2010) 2363–2367, https://doi.org/10.1093/bioinformatics/btq431.
[15] J. Pintor, J.M. Ramírez-Anguita, J. Satish-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L.I. Furlong, The DistGenET knowledge platform for disease genomics: 2019 update, Nucleic Acids Res. 48 (2019) D845–D855. https://doi.org/10.1093/nar/gkz1021.
[16] M. Uhlén, M.J. Karlsson, A. Hober, A.-S. Svensson, J. Scheffel, D. Kotel, W. Zhong, A. Tehani, L. Strandberg, F. Edfors, E. Sjöstedt, J. Mulder, A. Mardingh, A. Berling, S. Eklad, M. Damnemeyer, S. Kanje, J. Rockberg, M. Lindsjö, M. Malm, A.-L. Volk, P. Nilsson, A. Isanberg, T. Dodig-Crnkovic, E. Pin, M. Zwahlen, P. Oksvold, K. von Felitzen, R.S. Häusler, M.-G. Hong, C. Lindskog, F. Ponten, B. Katona, J. Vuu, E. Lindström, J. Nielsen, J. Robinson, B. Ayoglu, D. Mahdessian, D. Sullivan, P. Thul, F. Danielsson, C. Stadler, E. Lundberg, G. Bergstrom, A. Gunnesson, B.G. Voldborg, H. Tegel, S. Hofer, B. Forrester, J.M. Schwenk, L. Fagerberg, Å. Sivertsson, The human secretome, Sci. Signal. 12 (2019),https://doi.org/10.1126/scisignal.aaz0274.
[17] J. Meikin, G. Walker, C.R. Cooper, X.J. Min, MetaSecrEk: the Human and Animal Secretome and Subcellular Proteome Knowledgebase, Database (2015), https://doi.org/10.1093/database/bav077, 2015.
[18] A. Bunisello, J.A.L. MacArthur, M. Cerezo, L.W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Soliss, D. Suvess, O. Vrousgou, P.L. Whetzel, R. Amode, J.A. Guillen, H.S. Riat, S.J. Trevanion, P. Hall, H. Junkin, P. Flice, T. Burdett, L.A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and
C. Pastrello et al. Osteoarthritis and Cartilage Open 4 (2022) 100237

summary statistics 2019, Nucleic Acids Res. 47 (2019) D1005-D1012, https://doi.org/10.1093/nar/gky1120.

[29] T. Zhao, Z. Su, Y. Li, X. Zhang, Q. You, Chitinase-3-like-protein-1 function and its role in diseases, Signal Transduct. Target. Ther. 5 (2020) 201, https://doi.org/10.1038/s41397-019-00305-7.

[30] X. Gao, S. Jiang, Z. Du, A. Ke, Q. Liang, X. Li, KL2 protects against osteoarthritis by repressing oxidative response through activation of Nrf2/ARE signaling in vitro and in vivo, Oxid. Med. Cell. Longev. (2019) 1–18, https://doi.org/10.1155/2019/8564861 (2019).

[31] R. Meliconi, L. Pulastrelli, Are mechanisms of inflammation joint-specific in osteoarthritis? Rheumatology 58 (2019) 743–745, https://doi.org/10.1093/rheumatology/key300.

[32] Y. Xu, M.J. Barter, D.C. Swan, K.S. Rankin, A.D. Rowan, M. Santibanez-Koref, J. Loughlin, D.A. Young, Identification of the pathogenic pathways in osteoarthritis hip cartilage: commonality and discord between hip and knee OA, Osteoarthritis Cartilage 20 (2012) 1029–1038, https://doi.org/10.1016/j.joca.2012.05.006.

[33] G. Barreto, J. Sandelin, A. Salem, D.C. Nordstrom, E. Waris, Toll-like receptors and their soluble forms differ in the knee and thumb basal osteoarthritic joints, Acta Orthop. 88 (2017) 326–323, https://doi.org/10.1080/17453674.2017.1281058.

[34] G. Barreto, R. Solymani, M. Baumann, E. Waris, K.K. Eklund, M. Zenobi-Wong, M. Lalowski, Functional analysis of synovial fluid from osteoarthritic knee and cartilage samples unravels different molecular profiles, Rheumatology 58 (2019) 897–907, https://doi.org/10.1093/rheumatology/key252.

[35] P. Akhbari, M.K. Jaggard, C.L. Boulange, U. Vaghela, G. Graa, R. Bhattacharya, J.C. Lindon, H.R.T. Williams, C.M. Gute, Differences in the composition of hip and knee synovial fluid in osteoarthritis: a nuclear magnetic resonance (NMR) spectroscopy study of metabolic profiles, Osteoarthritis Cartilage 27 (2019) 1768–1777, https://doi.org/10.1016/j.joca.2019.07.017.

[36] W. den Hollander, Y.F.M. Ramos, S.D. Bos, N. Bomer, R. van der Breggen, N. Lakensberg, W.J. de Dijker, B.L. Dujiníveld, P.E. Slagboom, R.G.H.L. Nelissen, I. Meulenbelt, Knee and hip articular cartilage have distinct epigenomic landscapes: implications for future cartilage regeneration approaches, Ann. Rheum. Dis. 73 (2014) 2208–2212, https://doi.org/10.1136/annrheumdis-2014-209580.

[37] S. Decka, S. Heymans, A. Papageorgiou, The diverse functions of osteoglycin: a deceitful dwarf, or a master regulator of disease? Faseb. J. 30 (2016) 2651–2661, https://doi.org/10.1002/ff.15000968.

[38] C.-H. Chou, D.E. Attarian, H.-G. Winiwitzer, P.A. Band, V.B. Kraus, TSG-6 - a double-edged sword for osteoarthritis (OA), Osteoarthritis Cartilage 26 (2018) 245–254, https://doi.org/10.1016/j.joca.2017.10.019.

[39] A. Tserou, Osteoarthritis year in review 2014: genetics and genomics, Osteoarthritis Cartilage 22 (2014) 2017–2024, https://doi.org/10.1016/j.joca.2014.07.024.

[40] C. Shen, K. Gan, P. Zhang, L. Huang, lncRNA MALAT1 promotes chondrocyte proliferation by inhibiting miR-127-5p, Int. J. Clin. Exp. Med. 13 (2020) 3978–3988, https://www.ijcmm.com/. (Accessed 25 February 2021). accessed.

[41] G. Zhang, H. Zhang, W. You, X. Tang, X. Li, Z. Gong, Therapeutic effect of Resveratrol in the treatment of osteoarthritis via the MALAT1/miR-9-5p/eIF2 signaling pathway, Exp. Ther. Med. 19 (2020) 2343–2352, https://doi.org/10.3892/etm.2020.8471.

[42] H. Li, S. Xie, H. Li, R. Zhang, H. Zhang, LncRNA MALAT1 mediates proliferation of LPS treated-articular chondrocytes by targeting the miR-146a-PI3K/akt/mTOR axis, Life Sci. 254 (2020) 116801, https://doi.org/10.1016/j.lfs.2019.116801.

[43] F. Jiang, X. Li, Z. Li, Y. Chen, J. Wang, H. Zhang, H. Wang, LncRNA MALAT1 promotes osteoarthritis by modulating miR-150-5p/akt3 axis, Cell Bioni. 9 (2019) 54, https://doi.org/10.1080/23259243.2019.1682639.

[44] J. Liang, L. Xu, F. Zhou, A.-M. Liu, H.-X. Ge, Y.-Y. Chen, M. Tu, MALAT1/miR-127-5p regulates osteopontin (OPN)-Mediated proliferation of human chondrocytes through PI3K/akt pathway, J. Cell. Biochem. 119 (2018) 431–439, https://doi.org/10.1002/jcb.26200.

[45] C. Liu, S. Ren, S. Zhao, Y. Wang, LncRNA MALAT1/MIR-145 induces IL-1-induced chondrocytes viability and cartilage matrix degradation by regulating ADAMTS5 in LPS treated-articular chondrocytes by targeting the miR-146a-PI3K/akt/mTOR pathway, J. Cell. Biochem. 119 (2018) 431–439, https://doi.org/10.1002/jcb.26200.

[46] A.H. Poole, How has your science data grown? Digital curation and the human factor: a critical literature review, Arch. Sci. 15 (2015) 101–139, https://doi.org/10.3349/j.ymjl.2019.11.1081.

[47] P.J. Cope, K. Ourradi, Y. Li, M. Sharif, Models of osteoarthritis: the good, the bad and the promising, Osteoarthritis Cartilage 27 (2019) 230–239, https://doi.org/10.1016/j.joca.2019.09.016.

[48] G. Poste, Bring on the biomarkers, Nature 469 (2011) 156–157, https://doi.org/10.1038/469156a.