Journal of Visualized Experiments

Lung microRNA profiling across the estrous cycle in ozone exposed mice

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE58664R2
Full Title:	Lung microRNA profiling across the estrous cycle in ozone exposed mice
Keywords:	Air pollution; Lung miRNome; bronchoalveolar lavage; lung inflammation, sex hormones, PCR arrays, estrous cycle, vaginal smear
Corresponding Author:	Patricia Silveyra Penn State College of Medicine Hershey, PA UNITED STATES
Corresponding Author's Institution:	Penn State College of Medicine
Corresponding Author E-Mail:	ptrcslvr@gmail.com
Order of Authors:	Nathalie Fuentes
	Patricia Silveyra
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	500 University Drive, room C7744. Hershey, PA 17033

Cover Letter

August 7, 2018

Jaydev Upponi, Ph.D.
Science Editor | Immunology and Infection
Editorial Department
JoVE

Dear Dr. Upponi,

Please find enclosed our revised manuscript titled “Lung microRNA profiling across the estrous cycle in ozone exposed mice” by Nathalie Fuentes, and Patricia Silveyra (corresponding author) to be evaluated for publication in JoVE Immunology and Infection.

We appreciate the comments received by the editor and have addressed every point.

Thank you for your consideration.

Yours sincerely,

Patricia Silveyra, M.Sc., Ph.D.
Associate Professor of Pediatrics, Biochemistry and Molecular Biology, and Humanities
Pulmonary Immunology and Physiology Laboratory (PIP)
Department of Pediatrics. The Pennsylvania State University, College of Medicine
500 University Dr. MC H085
Hershey, PA 17033-0850
Tel: 717-531-5605

TITLE:

Lung microRNA Profiling Across the Estrous Cycle in Ozone-Exposed Mice

AUTHORS AND AFFILIATIONS:

Nathalie Fuentes[1], Patricia Silveyra[1,2]

[1]Pulmonary, Immunology and Physiology Laboratory, Department of Pediatrics, The Pennsylvania State University College of Medicine, Hershey, PA, USA
[2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA, USA

Corresponding Author:
Patricia Silveyra (pzs13@psu.edu)
Tel: (717) 531-5605

Email Address of Co-author:
Nathalie Fuentes (nfuentes1@pennstatehealth.psu.edu)

KEYWORDS:

Air pollution, lung miRNome, bronchoalveolar lavage, lung inflammation, sex hormones, PCR arrays, estrous cycle, vaginal smear

SUMMARY:

Here we describe a method to assess expression of miRNAs that are predicted to regulate inflammatory genes in the lungs of mice exposed to ozone at different estrous cycle stages.

ABSTRACT:

microRNA (miRNA) profiling has become of interest to researchers working in various research areas of biology and medicine. Current studies show a promising future of using miRNAs in the diagnosis and care of lung diseases. Here, we define a protocol for miRNA profiling to measure the relative abundance of a group of miRNAs predicted to regulate inflammatory genes in the lung tissue from of an ozone-induced airway inflammation mouse model. Because it has been shown that circulating sex hormone levels can affect the regulation of lung innate immunity in females, the purpose of this method is to describe an inflammatory miRNA profiling protocol in female mice, taking into consideration the estrous cycle stage of each animal at the time of ozone exposure. We also address applicable bioinformatics approaches to miRNA discovery and target identification methods using *limma*, an R/Bioconductor software, and functional analysis software to understand the biological context and pathways associated with differential miRNA expression.

INTRODUCTION:

microRNAs (miRNAs) are short (19 to 25 nucleotides), naturally occurring, non-coding RNA molecules. Sequences of miRNAs are evolutionary conserved across species, suggesting the importance of miRNAs in regulating physiological functions[1]. microRNA expression profiling has been proven to be helpful for identifying miRNAs that are important in the regulation of a variety of processes, including the immune response, cell differentiation, developmental processes, and apoptosis[2]. More recently, miRNAs have been recognized for their potential use in disease diagnostics and therapeutics. For researchers studying mechanisms of gene regulation, measuring miRNA expression can enlighten systems-level models of regulatory processes, especially when miRNA information is merged with mRNA profiling and other genome-scale data[3]. On the other hand, miRNAs have also been shown to be more stable than mRNAs in a range of specimen types and are also measurable with greater sensitivity than proteins[4]. This has led to considerable interest in the development of miRNAs as biomarkers for diverse molecular diagnostic applications, including lung diseases.

In the lung, miRNAs play important roles in developmental processes and the maintenance of homeostasis. Moreover, their abnormal expression has been associated with the development and progression of various pulmonary diseases[5]. Inflammatory lung disease induced by air pollution has demonstrated greater severity and poorer prognosis in females, indicating that hormones and the estrous cycle can regulate lung innate immunity and miRNA expression in response to environmental challenges[6]. In this protocol, we use ozone exposure, which is a major component of air pollution, to induce a form of lung inflammation in female mice that occurs in the absence of adaptive immunity. By using ozone, we are inducing the development of airway hyperresponsiveness that is associated with airway epithelial cell damage and an increase in neutrophils and inflammatory mediators in proximal airways[7]. Currently, there are not well-described protocols to characterize and analyze miRNAs across the estrous cycle in ozone-exposed mice.

Below, we describe a simple method to identify estrous cycle stages and miRNA expression in lung tissue of female mice exposed to ozone. We also address effective bioinformatics approaches to miRNA discovery and target identification, with an emphasis on computational biology. We analyze the microarray data using *limma*, an R/Bioconductor software that provides an integrated solution for analyzing data from gene expression experiments[8]. Analysis of PCR array data from *limma* has an advantage in terms of power over t-test based procedures when using small number of arrays/samples to compare expression. To comprehend the biological context of miRNA expression results, we then used the functional analysis software. In order to understand the mechanisms regulating transcriptional changes and to predict likely outcomes, the software combines miRNA-expression datasets and knowledge from the literature[9]. This is an advantage when compared with software that just look for statistical enrichment in overlapping to sets of miRNAs.

PROTOCOL:

All methods described here have been approved by the Institutional Animal Care and Use Committee (IACUC) of Penn State University.

1. Assessment of the Estrous Cycle Stage

1.1. Properly restrain a female C57BL/6 mouse (8-9 weeks old) using the one-handed mouse restraint technique described in Machholz *et al.*[10].

1.2. Fill the sterile plastic pipette with 10 µL of ultra-pure water.

1.3. Introduce the tip of plastic pipette into the vagina.

1.4. Gently flush the liquid 4-5 times to collect the sample.

1.5. Place the final flush containing vaginal fluid on a glass slide.

1.6. Observe the unstained vaginal flush under a light microscope with a 20X objective.

Note: Animals that do not show regular cycles due to pseudopregnancy or other causes need to be excluded from the experiment. It is recommended to perform daily vaginal secretions for at least three consecutive cycles to confirm cyclicity.

2. Exposure to Ozone

2.1. Place a maximum of 4 mice in each 2 L glass container (ozone, filtered air) with wire mesh lids containing bedding, food, and water *ad libitum*.

2.2. Put one glass container in the ozone chamber and the other in the filtered air exposure chamber.

2.3. Adjust the ozone concentration to 2 ppm and monitor ozone levels regularly.

Note: Ozone is generated by an electrical discharge ozonizer to ensure stability. To monitor the ozone levels, use an ultraviolet ozone analyzer and mass flow controllers.

2.4. Remove glass containers after 3 h of ozone/filtered air exposure.

Note: The ozone apparatus delivers a regulated airflow (> 30 air changes/h) with controlled temperature (25 °C) and relative humidity (50%). The system generates ozone by an electrical discharge ozonizer, which is monitored and controlled by an ultraviolet ozone analyzer and mass flow controllers, respectively, as described previously[11].

3. Lung Collection

3.1. 4 h after exposure, anesthetize animals with an intraperitoneal injection of a ketamine/xylazine cocktail (90 mg/kg ketamine, 10 mg/kg xylazine).

Note: To confirm a proper level of anesthesia, check the mouse for a pedal reflex (firm toe pinch) and adjust the anesthetic as needed.

3.2. Dampen the mouse skin with 70% ethanol.

3.3. Make a 2 cm midline incision using an operating scissor and surgical tweezers to expose the vena cava.

3.4. Sacrifice mice by transection of the vena cava and aorta. If needed, insert a 21G gauge needle into the vena cava above the renal veins to collect blood prior to exsanguination. Alternatively, collect blood via heart puncture following standard protocols.

3.5. Use a surgical scissor to cut open the abdominal cavity and remove skin/upper muscle, moving upwards toward the ribs.

3.6. Use a surgical scissor to puncture the diaphragm.

Note: Lungs will collapse away from the diaphragm.

3.7. Cut away the ribcage using a surgical scissor to expose the heart and lungs.

3.8. Using forceps, take a 1.5 mL RNase-free microcentrifuge tube and submerge it in liquid nitrogen to fill the tube.

Note: Use goggles and protective gloves to handle liquid nitrogen.

3.9. Remove the lungs, place them into the RNase-free microcentrifuge 1.5 mL tubes filled with liquid nitrogen to snap-freeze the tissue, and wait a few seconds until the liquid evaporates.

3.10. Close the tube lid and store the tissue at -80 ˚C until use.

4. RNA Preparation

4.1. Pulverize whole lungs using a stainless-steel tissue pulverizer.

Note: The tissue pulverizer needs to be place in liquid nitrogen prior to use. Clean the pulverizer after each use with RNAse solution.

4.2. Split the pulverized lungs and place into two 1.5 mL tubes (half lung each).

Note: Samples can be spiked with 5.6 x 10^8 copies of a small RNA spike-in control (from a different species) before proceeding with the extraction.

4.3. Use 500 µL of guanidinium thiocyanate per sample and homogenize each sample using 18G, 21G, and 23G needles, respectively.

4.4. Add 500 µL of ethanol to each sample and vortex for 15 s.

4.5. Load the mixture in a spin column in a collection tube and centrifuge at 12,000 x g for 1 min. Discard the flow-through.

4.6. For DNase I treatment (in-column);

4.6.1. Add 400 µL of RNA Wash Buffer and centrifuge at 12,000 x g for 1 min.

4.6.2. In an RNase-free tube, add 5 µL of DNase I and 75 µL of 1x DNA digestion buffer and mix. Add the mix directly to the column matrix.

4.6.3. Incubate at room temperature for 15 min.

4.7. Add 400 µL of RNA prewash solution to the column and centrifuge at 12,000 x g for 1 min. Discard the flow-through and repeat this step.

4.8. Add 700 µL of RNA wash buffer to the column and centrifuge at 12,000 x g for 1 min. Discard the flow-through.

4.9. Centrifuge at 12,000 x g for 2 min to remove remaining buffer. Transfer the column into an RNase-free tube.

4.10. To elute RNA, add 35 µL of DNase/RNase-free water directly to the column matrix and centrifuge at 12,000 x g for 1.5 min.

4.11. Measure total RNA concentration (260 nm) and purity using a spectrophotometer. Follow instructions to perform RNA quantification in a 1.5 µL sample aliquot. Blank the instrument with DNase/RNase-free water used for elution.

Note: A 260/280 ratio of ~2.0 is generally accepted as “pure” for RNA. Typical RNA concentration usually ranges between 750 and 2500 ng/µL.

4.12. Store at -80 °C.

5. miRNA Profiling

5.1. To retro-transcribe small RNAs, use 200 ng of total RNA.

5.1.1. Prepare the reverse-transcription reaction mix on ice (total volume per reaction is 20 µL). For each reaction, add 4 µL of 5x buffer, 2 µL of 10x nucleotides mix, 2 µL of reverse transcriptase, and 2 µL of RNase-free water. Mix all the components and aliquot in 600 µL RNAse-free plastic tubes (10 µL of mix per reaction).

Note: The reverse-transcription master mix contains all components required for first-strand cDNA synthesis except template RNA.

Note: Calculate excess volume (1 extra tube every 10 reactions) when preparing the master mix.

5.1.2. Add the template RNA (200 ng in 10 µL) to each tube containing reverse-transcription master mix. Mix, centrifuge for 15 s at 1000 x g, and store them on ice until placing in thermocycler or dry block.

5.1.3. Incubate for 60 min at 37 °C.

5.1.4. Incubate for 5 min at 95 °C and place the tubes on ice.

5.1.5. Dilute the cDNA by adding 200 µL of RNase-free water to each 20 µL reverse-transcription reaction.

5.2. Perform real-time PCR using the Mouse Inflammatory Response and Autoimmunity miRNA PCR Array.

5.2.1. Prepare a reaction mix (total volume 1100 µL): For each reaction, add 550 µL of 2x PCR master mix, 110 µL of 10x universal primer mix, 340 µL of RNase-free water, and 340 µL of template cDNA (diluted reaction from step 5.1.5).

5.2.2. Add 10 µL of reaction mix to each well of the pre-loaded miRNA PCR Array using a multichannel pipettor.

5.2.3. Seal the miRNA PCR Array plate with optical adhesive film.

5.2.4. Centrifuge the plate for 1 min at 1000 x g at room temperature to remove bubbles.

5.2.5. Program the real-time cycler, PCR initial activation step for 15 min at 95 °C, 3-step cycling containing denaturation for 15 s at 94 °C, annealing for 30 s at 55 °C, and extension for 30 s at 70 °C for 40 cycles number.

Note: Follow manufacturer's cycling conditions instructions to set up the real-time cycler. Perform the dissociation curve step built into the real-time cycler software.

5.2.6. Perform data analysis.

6. Data Analysis

6.1. Extract Ct values from the real time PCR software for each sample into an analysis software.

Note: A Ct value of 34 is considered as cutoff. If samples contain spike-in control (*e.g.,* cel-mir-39), normalize Ct values to the spike in control for each sample. Threshold values may need to be manually set. Baseline values are automatically set.

6.2. Normalize Ct values to the average Ct of six miRNA housekeeping controls: SNORD61, SNORD68, SNORD72, SNORD95, SNORD96A, RNU6-2, using the following equation:

$$\Delta\text{Ct} = (\text{Ct\_Target} - \text{Ct\_housekeeping})$$

6.3. For fold change calculations, calculate ΔΔCt values using a specific sample as the control, using the relative expression equation[12]:

miRNA relative expression:

$$2^{-\Delta\Delta}Ct, where - \Delta\Delta Ct = -[\Delta Ct\ test - \Delta Ct\ control]$$

Note: A fold change of 200 is considered as cutoff.

6.4. Export fold change expression values to perform statistical analysis on R using the *limma* package in Bioconductor[8].

6.5. Correct for multiple comparisons using the Benjamini-Hochberg method[13].

Note: A copy of the R script is available at: http://psilveyra.github.io/silveyralab/

7. Data Analysis: Functional Analysis Software

7.1. Arrange the dataset. Include miRNAs with their respective expression log ratios and p-values. See **Table 1** for specific dataset formatting.

7.2. Open functional analysis software (version 01-10).

7.3. Upload the dataset using the following format: File format: “Flexible Format”, Contains Column Header: “Yes”, Select Identifier type: “miRBase (mature)”, Array platform used for experiments: Choose the array platform.

Note: Accepted files formats are .txt (tab delimited text files), .xls (excel files), and .diff (cuffdiff files).

7.4. Select “Infer Observations” and verify that the experimental group labeling is correct.

7.5. Go to “Dataset Summary” to revise the total amount of mapped and unmapped miRNAs.

7.6. Click on the “new” button at the top left of program. Select “New MicroRNA Target Filter” and upload a microRNA dataset.

7.6.1. Set Source to: TarBase, Ingenuity Experts Findings, miRecords.

7.6.2. Set Confidence to: Experimentally observed or high (predicted).

7.6.3. Select “Add Columns” to include a variety of biological information about the targets such as species, diseases, tissues, pathways and more.

7.6.4. To focus on targets that have changed expression in the experiment, select “Add/ replace mRNA dataset”. Use “Expression Pairing” to locate microRNAs with the same or different expression levels.

Note: The filter analysis will provide microRNA names and symbols, mRNA targets, the source that describes the target relationship, and the confidence level of the predicted relationship (**Figure 2**).

7.7. Click “Add to My Pathway” to send the filtered dataset to a pathway canvas and explore more biological relationships.

7.8. Use Path Designer to create a publication-quality model of microRNA effects.

7.9. An alternate option is to create a core analysis.

7.9.1. For core analysis type, select “Expression Analysis”.

7.9.2. For measurement type, select “Expr Log Ratio”.

7.9.3. Analysis Filter Summary: consider only molecules and/or relationships where: (species = mouse) AND (confidence = experimentally observed) AND (data sources = “Ingenuity Expert Findings”, “Ingenuity ExpertAssist Findings”, “miRecords”, “TarBase”, OR “TargetScan Human”).

7.9.4. Select a cutoff of p-value = 0.05.

7.9.5. Run the analysis.

Note: The report will include: canonical pathways, upstream regulators analysis, diseases and functions, regulator effects, networks, molecules and more.

REPRESENTATIVE RESULTS:

The different cell types observed in smears are used to identify the mouse estrous cycle stage (**Figure 1**). These are identified by cell morphology. During proestrus, cells are almost exclusively clusters of round-shaped, well-formed nucleated epithelial cells (**Figure 1A**). When the mouse is in the estrus stage, cells are cornified squamous epithelial cells, present in densely packed clusters (**Figure 1B**). During metestrus, cornified epithelial cells and polymorphonuclear leukocytes are seen (**Figure 1C**). In diestrus, leukocytes (small cells) are generally more prevalent (**Figure 1D**).

We extracted RNA from four mouse lungs following the protocol previously described. The nucleic acid concentrations (ng/ µL) ranged between 1197.9 and 2178.1 with an average of 1583.1 ± 215 (**Table 1**). The average A260/A280 ratio fluctuated from 2.010 to 2.020 with an average of 2.016 ± 0.002. On the other hand, the observed A260/A230 ratios oscillated between 2.139 and 2.223 with an average of 2.179 ± 0.018.

Table 2 shows differential expression results obtained with *limma* on R. We calculated top differentially expressed miRNAs between mice exposed to ozone or filtered air in proestrus (using the command toptable)[14]. The first column gives the value of the log2-fold fold change in miRNA expression between ozone and filtered air exposed animals. The column t represents the moderate t-statistic calculated for each miRNA in the comparison. The columns p.value and adj.p.value represent associated p-values for each comparison before and after multiple testing adjustment, respectively. Adjustment for multiple comparisons was done with the Benjamini and Hochberg's method to control the false discovery rate[15]. Column B represents the log-odds that the miRNA is differentially expressed[8].

We performed the miRNA target filter and core analysis that includes the enrichment pathway analysis. After uploading a list of 14 miRNAs with the significant expression log ratio and p-value, all of them were mapped by the miRNA target filter (**Table 3**). The results were filtered and sorted to get to certain pathways, in this case the "Cellular Immune Response". The core analysis provided information about canonical pathways, diseases and function, regulators, and networks (**Table 4**). The functional analysis software produced a network analysis that shows the relationship between the miRNAs of interest and other molecules (**Figure 3**).

FIGURE AND TABLE LEGENDS:

Figure 1: **Identification of estrous cycle stages.** (A) Proestrus (predominantly nucleated epithelial cells); (B) estrus (predominantly anucleated cornified cells); (C) metestrus (all three types of cells); and D) diestrus 2 (majority of leukocytes). Scale bar = 100 µm. Magnification = 20X.

Figure 2: Functional analysis software representative results: miRNA target filter. Comprehensive profile of miRNAs at different stages of the estrous cycle. After performing miRNA filter, the software delivers detailed listings of genes and compounds implicated in diseases and other phenotypes, which can be filter and sort to get to certain pathways, in this case "Cellular Immune Response".

Figure 3: Functional analysis software representative results: networks. Comparison of networks affected by filtered air or ozone exposure in females at different stages of the estrous cycle. Diagram of biological networks associated with miRNAs in the lungs of female mice exposed to filtered air vs. ozone in proestrus (A) or non-proestrus stages (B). This figure has been modified from Fuentes *et al.*[6].

Table 1: Example of RNA concentrations and absorbance ratios at 260, 230, and 280 nm from purified lung tissue samples from four mice. Concentrations were measured with a spectrophotometer.

Table 2: *Limma* analysis results for differentially expressed miRNAs in females exposed to ozone *vs*. filtered air in the proestrus stage.

Table 3: Example format for multi-observation upload of datasets to functional analysis software. Multiple experimental differential expressions can be grouped into a single spreadsheet and uploaded, and as many observations as needed can be added. Columns: 1) miRNAs ID; 2) Observation 1: Expr Log Ratio; 3) Observation 1: P Value; 4) Observation 2: Expr Log Ratio; 5) Observation 2: P Value.

Table 4: Functional analysis software summary of females exposed to ozone in the non-proestrus *vs*. proestrus stages. The functional analysis software allows the analysis of top canonical pathways, upstream regulators, diseases & functions, top functions, regulator effect networks and more. This table has been modified from Fuentes *et al.*[6].

DISCUSSION:

MicroRNA profiling is an advantageous technique for both disease diagnosis and mechanistic research. In this manuscript, we defined a protocol to evaluate the expression of miRNAs that are predicted to regulate inflammatory genes in the lungs of female mice exposed to ozone in different estrous cycle stages. Methods for the determination of the estrous cycle, such as the visual detection method, have been described[16]. However, these rely on one-time measurements, and therefore are unreliable. To accurately identify all estrous cycle stages in females that cycle regularly, the method described here is recommended. In addition, this simple protocol can also be used to indirectly estimate daily hormonal fluctuations in mice. To avoid activation of unwanted inflammatory responses due to vaginal irritation, sampling needs to be performed just once daily. Because of variability in the cycle length and housing influences, it is important to perform the protocol for two to three complete cycles before using animals in an experiment considering the cycle stage.

For the successful extraction of RNA from lung tissue, an accurate procedure is critical. This protocol describes a one-day method to isolate RNA from lung tissue that yields high-quality RNA. Modifications to the manufacturer's protocol were required to efficiently extract RNA from lungs. We added an additional centrifugation step after addition of the wash buffer to remove as much buffer as possible. We also eluted RNA with 35 µL of DNase/RNase-free water, centrifuging the column for 1.5 min to ensure high concentration levels. Spectrofluorometer results confirmed the effectiveness of our RNA extraction protocol. The ratio of absorbance at 260 and 280 nm (A260/280 ratio) is frequently used to assess the purity of RNA preparations. The maximum absorbance for nucleic acids is 260 and 280 nm, respectively. It is accepted as "pure" for RNA if the ratio is about 2.0[17]. Likewise, for the A260/A230 contamination absorbance ratio, the values for purity are in the range of 2.0-2.2[18]. In this study, the average A260/A280 and A260/A230 ratios observed were 2.016 ± 0.002 and 2.179 ± 0.018, respectively (**Table 1**). Therefore, our RNA extraction protocol was successful. Another advantage of the protocol used is the addition of the DNAse treatment. This is important to avoid genomic-DNA contamination[19]. A limitation of this RNA isolation protocol is the use of purification columns to discard waste through precipitation using alcohol because some lung debris may obstruct the membrane either partially or completely, resulting in low yields. Also, if the homogenization step is not carefully performed, large quantities of lung RNA can be easily lost or degraded. If low RNA yields are obtained, RNA can be re-purified and eluted in a smaller volume. Alternatively, RNA can be precipitated overnight following published protocols[20].

Microarray technologies applied to miRNA profiling are promising tools in many research fields. In our study, we used PCR arrays, which provide the advantage of higher detection threshold, and normalization strategies for detection of differentially expressed miRNAs *vs.* other technologies such as probe-based miRNA arrays[21]. A limitation of this protocol is that it requires a minimum amount of starting RNA material, and availability of specific sets of primers for the miRNAs of interest, as opposed to other available techniques such as RNAseq. Another advantage of PCR-based arrays is the option of using non-miRNA reference genes for qPCR normalization (such as small nucleolar RNAs or SNORDs) to calculate differential expression of miRNAs. Finally, using PCR arrays provides several options for data analysis, ranging from online tools provided by the manufacturers, to conventional methods to detect differential expression though Real-Time PCR. Statistical analysis with *limma* is convenient for both microarrays and PCR-based arrays and uses the empirical Bayes moderated *f*-statisitcs[22]. Here, we show that both p-values and q-values (adjusted for multiple testing) can be obtained with the command toptable adjusting the false discovery rate threshold and identifying differentially expressed miRNAs.

Functional analysis software is a web-based application for data analysis in pathway context. The software gives researchers powerful search abilities that can help to frame data sets or specific targets in context, within a bigger picture of biological significance. Although the software environment is flexible to different types of analysis (*i.e.,* metabolomics, SNPs, proteomics, microRNA, toxicology, *etc*.), our goal here is to highlight aspects of miRNA analysis. After uploading a list of 14 miRNAs with significant expression log-ratios and p-values, all were mapped by the software. We performed the miRNA target filter and core analysis, which includes the enrichment pathway analysis. However, such analyses consider genes for which the 14 miRNAs are predicted to target and not the miRNAs themselves. The results section lists outputs such as: canonical pathways, diseases and function, genes targeted by differentially expressed miRNAs, physiological system development, regulators, and networks (**Table 4**). The pathway visualization is shown under the network tab, where miRNAs and molecules are shown as clickable nodes that are linked with information associated to the gene of interest (**Figure 3**). An advantage of the functional analysis software is the high-quality miRNA-related findings, including both experimentally validated and predicted interactions. The functional analysis software databases include: experimentally validated microRNA-mRNA interactions databases[23,24], predicted microRNA-mRNA interaction database with low-confidence interactions excluded (*e.g.*, Target Scan)[25], experimentally validated human, rat, and mouse microRNA-mRNA interactions databases (*e.g.*, miRecords)[26], and literature findings (*e.g.,* microRNA-related findings manually curated from published literature by scientific experts). Other studies comparing the effectiveness and usability of bioinformatics tools to analyze pathways associated with miRNA expression confirm the effectiveness of this software[27]. Overall, computational methods are cost-effective, less time-consuming, and can be easily validated by molecular methods. With the constant growth and accumulation of biomedical data, bioinformatics methods will become increasingly powerful in the discovery of miRNA-mediated mechanisms of biological and disease processes.

ACKNOWLEDGMENTS:

This research was supported by grants from NIH K01HL133520 (PS) and K12HD055882 (PS). The authors thank Dr. Joanna Floros for the assistance with ozone exposure experiments.

DISCLOSURES:

The authors declare that they have no competing interests.

REFERENCES:

1 Rebane, A., Akdis, C. A. MicroRNAs: Essential players in the regulation of inflammation. *Journal of Allergy and Clinical Immunology.* **132** (1), 15-26 (2013).

2 Cannell, I. G., Kong, Y. W., Bushell, M. How do microRNAs regulate gene expression? *Biochemical Society Transactions.* **36** (Pt 6), 1224-1231 (2008).

3 Pritchard, C. C., Cheng, H. H., Tewari, M. MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics.* **13** (5), 358-369 (2012).

4 Mi, S., Zhang, J., Zhang, W., Huang, R. S. Circulating microRNAs as biomarkers for inflammatory diseases. *Microrna.* **2** (1), 63-71 (2013).

5 Sessa, R., Hata, A. Role of microRNAs in lung development and pulmonary diseases. *Pulmonary Circulation.* **3** (2), 315-328 (2013).

6 Fuentes, N., Roy, A., Mishra, V., Cabello, N., Silveyra, P. Sex-specific microRNA expression networks in an acute mouse model of ozone-induced lung inflammation. *Biology of Sex Differences.* **9** (1), 18 (2018).

7 Aris, R. M., *et al.* Ozone-induced airway inflammation in human subjects as determined by airway lavage and biopsy. *American Review of Respiratory Disease.* **148** (5), 1363-1372 (1993).

8 Ritchie, M. E., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research.* **43** (7), e47 (2015).

9 Krämer, A., Green, J., Pollard, J., Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics.* **30** (4), 523-530 (2014).

10 Machholz, E., Mulder, G., Ruiz, C., Corning, B. F., Pritchett-Corning, K. R. Manual restraint and common compound administration routes in mice and rats. *Journal of Visualized Experiments.* (67), (2012).

11 Umstead, T. M., Phelps, D. S., Wang, G., Floros, J., Tarkington, B. K. In vitro exposure of proteins to ozone. *Toxicology Mechanisms and Methods.* **12** (1), 1-16 (2002).

12 Livak, K. J., Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods.* **25** (4), 402-408 (2001).

13 Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics.* **10** (2), 946-963 (2016).

14 Smyth, G. K., *et al.* Linear Models for Microarray and RNA-Seq Data User's Guide. *Bioconductor*. (2002).

15 Benjamini, Y., Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological).* **57** (1), 289-300 (1995).

16 Byers, S. L., Wiles, M. V., Dunn, S. L., Taft, R. A. Mouse estrous cycle identification tool and images. *Public Library of Science ONE.* **7** (4), e35538 (2012).

17 Alves, M. G., *et al.* Comparison of RNA Extraction Methods for Molecular Analysis of Oral Cytology. *Acta Stomatologica Croatica.* **50** (2), 108-115 (2016).

18 Wilfinger, W. W., Mackey, K., Chomczynski, P. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques.* **22** (3), 474-476, 478-481 (1997).

19 Bustin, S. A., *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry.* **55** (4), 611-622 (2009).

20 Walker, S. E., Lorsch, J. RNA purification- precipitation methods. *Methods in Enzymology.* **530,** 337-343 (2013).

21 Git, A., *et al.* Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA.* **16** (5), 991-1006 (2010).

22 Smyth, G.K. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. 397-420 (2005).

23 Griffiths-Jones, S. miRBase: the microRNA sequence database. Methods Mol Biol. 342, 129-38 (2006).

24 Sethupathy, P., Corda, B., Hatzigeorgiou, A. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*. **12** (2), 192–197 (2006).

25 Agarwal, V., Bell, G.W., Nam, J., Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**, e05005 (2015).

26 Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*. **37**, D105-D110 (2009).

27 Mullany, L. E., Wolff, R. K., Slattery, M. L. Effectiveness and Usability of Bioinformatics Tools to Analyze Pathways Associated with miRNA Expression. *Cancer Informatics.* **14,** 121-130 (2015).

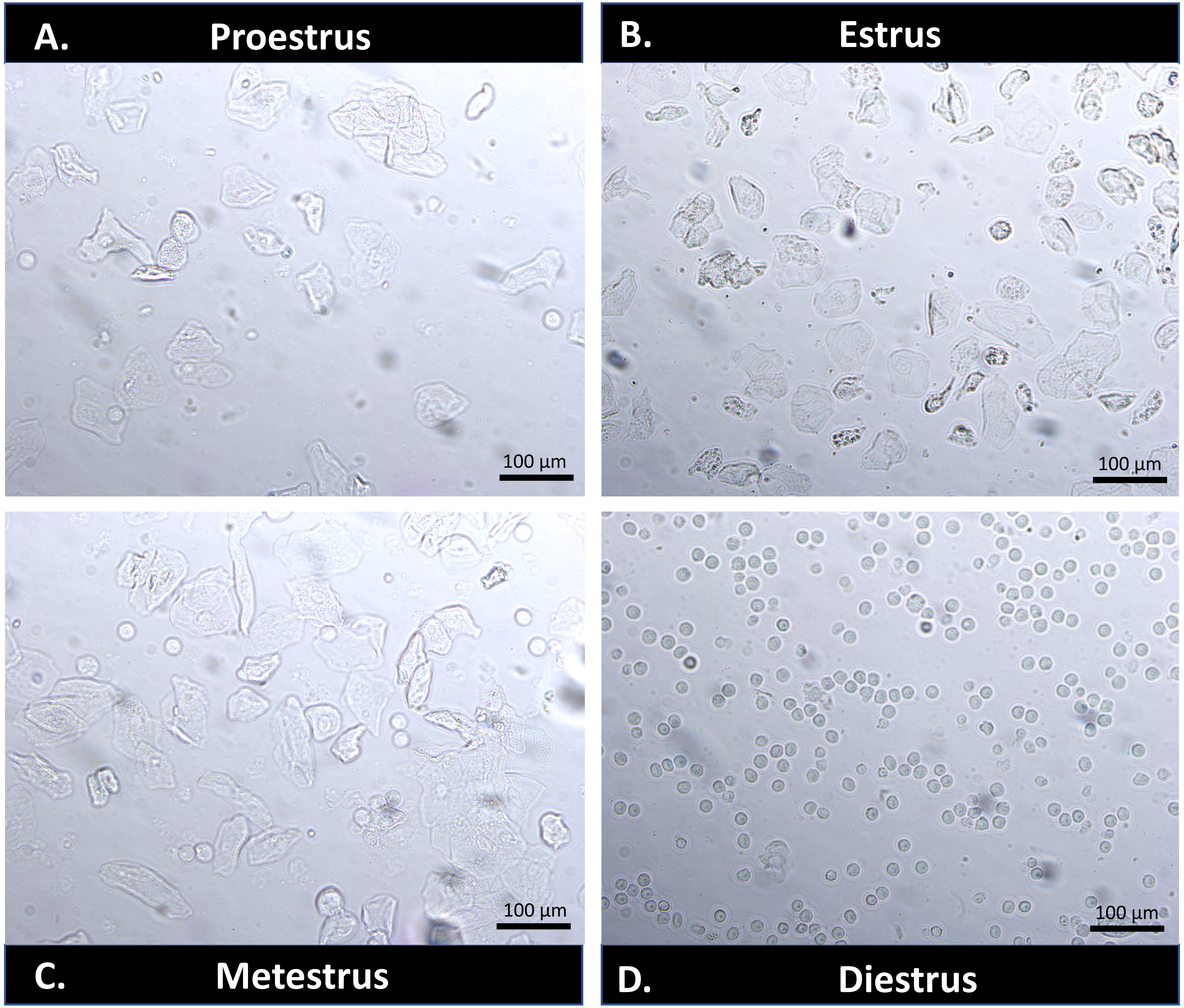
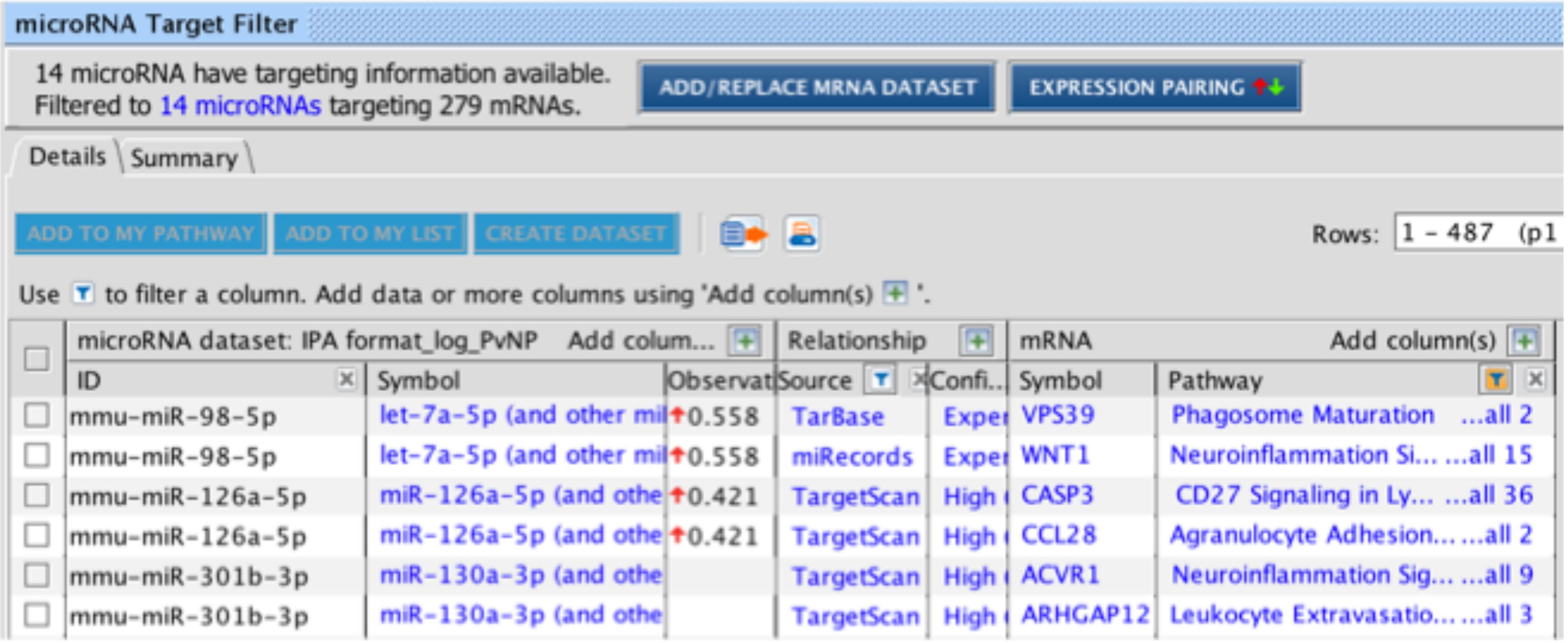

A.
Proestrus
100 µm
B.
Estrus
100 µm
100 µm
C.
Metestrus
100 µm
D.
Diestrus

Figure 2



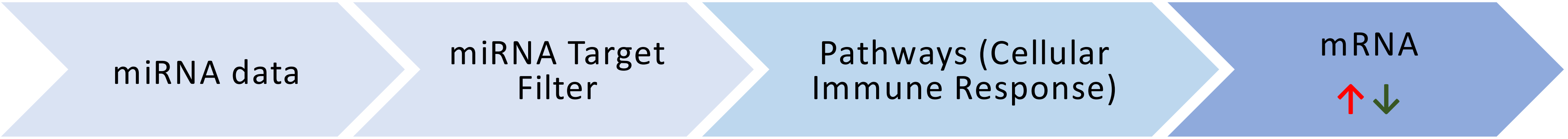
microRNA Target Filter

14 microRNA have targeting information available.
Filtered to 14 microRNAs targeting 279 mRNAs.

Rows: 1 - 487 (p1

Use to filter a column. Add data or more columns using 'Add column(s) '.

□	microRNA dataset: IPA format_log_PvNP Add colum...			Relationship		mRNA Add column(s)	
	ID	Symbol	Observat	Source	Confi..	Symbol	Pathway
□	mmu-miR-98-5p	let-7a-5p (and other mi	↑0.558	TarBase	Exper	VPS39	Phagosome Maturation ...all 2
□	mmu-miR-98-5p	let-7a-5p (and other mi	↑0.558	miRecords	Exper	WNT1	Neuroinflammation Si... ...all 15
□	mmu-miR-126a-5p	miR-126a-5p (and othe	↑0.421	TargetScan	High (	CASP3	CD27 Signaling in Ly... ...all 36
□	mmu-miR-126a-5p	miR-126a-5p (and othe	↑0.421	TargetScan	High (	CCL28	Agranulocyte Adhesion... ...all 2
□	mmu-miR-301b-3p	miR-130a-3p (and othe		TargetScan	High (	ACVR1	Neuroinflammation Sig... ...all 9
□	mmu-miR-301b-3p	miR-130a-3p (and othe		TargetScan	High (	ARHGAP12	Leukocyte Extravasatio... ...all 3



14 data input

14 targets

279 targets

Figure 3

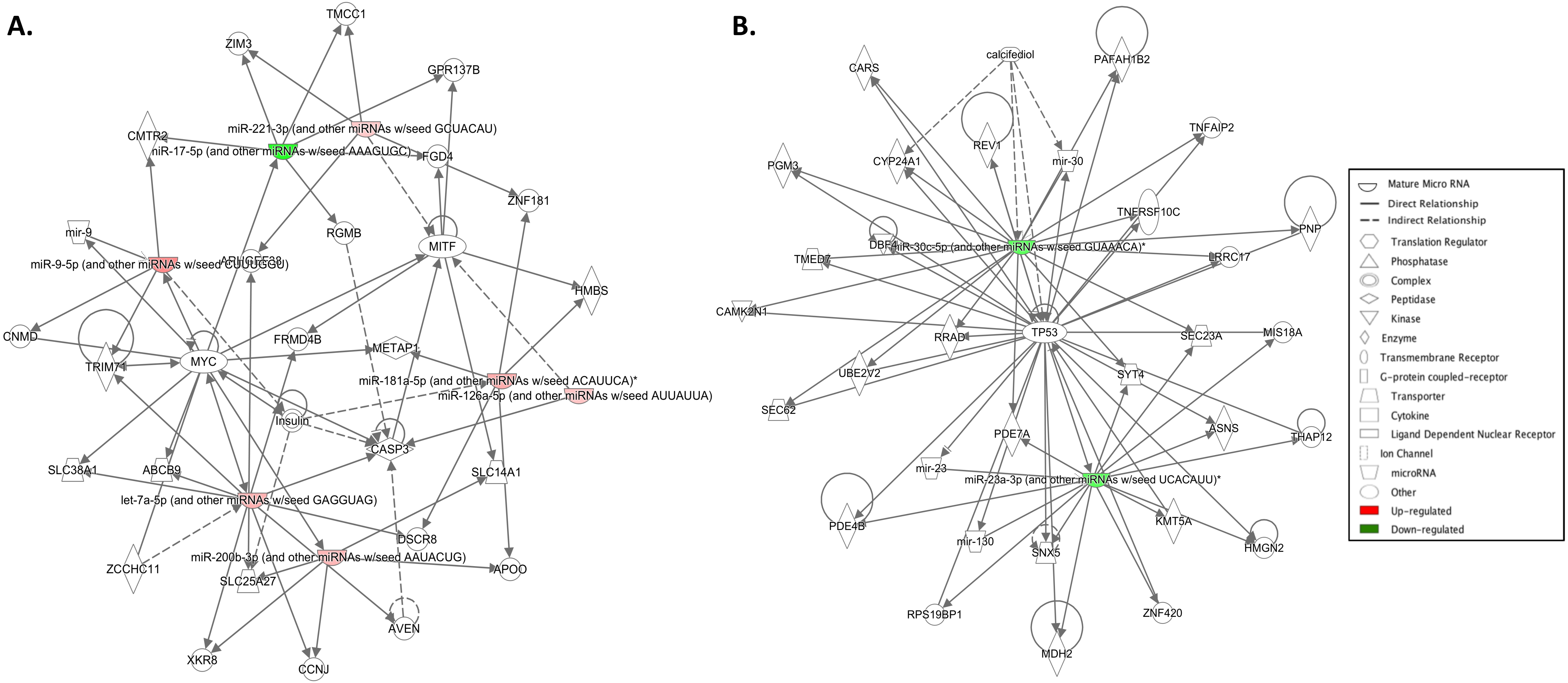

A.
TMCC1
ZIM3
GPR137B
miR-221-3p (and other miRNAs w/seed GCUACAU)
CMTR2
miR-17-5p (and other miRNAs w/seed AAAGUGC)
FGD4
ZNF181
mir-9
RGMB
MITF
miR-9-5p (and other miRNAs w/seed CUUUGGU)
ARHGEF38
HMBS
CNMD
FRMD4B
TRIM71
MYC
METAP1
miR-181a-5p (and other miRNAs w/seed ACAUUCA)*
miR-126a-5p (and other miRNAs w/seed AUUAUUA)
Insulin
CASP3
SLC38A1
ABCB9
SLC14A1
let-7a-5p (and other miRNAs w/seed GAGGUAG)
DSCR8
miR-200b-3p (and other miRNAs w/seed AAUACUG)
ZCCHC11
SLC25A27
APOO
AVEN
XKR8
CCNJ
B.
CARS
calcifediol
PAFAH1B2
REV1
TNFAIP2
PGM3
CYP24A1
mir-30
TNFRSF10C
PNP
DBF4
miR-30c-5p (and other miRNAs w/seed GUAAACA)*
TMED7
LRRC17
CAMK2N1
RRAD
TP53
SEC23A
MIS18A
UBE2V2
SYT4
SEC62
PDE7A
ASNS
THAP12
mir-23
miR-23a-3p (and other miRNAs w/seed UCACAUU)*
PDE4B
KMT5A
mir-130
SNX5
HMGN2
RPS19BP1
ZNF420
MDH2
Mature Micro RNA
Direct Relationship
Indirect Relationship
Translation Regulator
Phosphatase
Complex
Peptidase
Kinase
Enzyme
Transmembrane Receptor
G-protein coupled-receptor
Transporter
Cytokine
Ligand Dependent Nuclear Receptor
Ion Channel
microRNA
Other
Up-regulated
Down-regulated

Table 1

ID	Nucleic Acid (ng/µL)	A260/A280
Sample 1	1197.930	2.015
Sample 2	1355.703	2.018
Sample 3	2178.104	2.020
Sample 4	1600.837	2.010
Average	1583.144 ± 215	2.016 ± 0.002

A260/A230
2.192
2.223
2.163
2.139
2.179 ± 0.018

Table 2

	logFC	t	P.Value
mmu-miR-694	1.492	4.071	0.000153
mmu-miR-9-5p	0.836	3.916	0.000254
mmu-miR-221-3p	0.385	3.106	0.003014
mmu-miR-181d-5p	0.597	2.891	0.005516
mmu-miR-98-5p	0.558	2.699	0.009243
mmu-miR-712-5p	0.667	2.563	0.013169
mmu-miR-106a-5p	-0.528	-2.412	0.019278

adj.P.Val	B
0.009514	0.759
0.009514	0.289
0.075361	-1.982
0.103424	-2.526
0.138649	-2.987
0.164609	-3.299
0.206547	-3.632

Table 3

miRNAs ID	Observation 1 **Expr Log Ratio**	Observation 1 **P Value**
mmu-miR-694	1.492	0.000153
mmu-miR-9-5p	0.836	0.000254
mmu-miR-221-3p	0.385	0.003014
mmu-miR-181d-5p	0.597	0.005516
mmu-miR-98-5p	0.558	0.009243
mmu-miR-712-5p	0.667	0.013169
mmu-miR-106a-5p	-0.528	0.019278

Observation 2 **Expr Log Ratio**	Observation 2 **P Value**
0.543319208	0.0021385
0.677595421	0.004997439
0.342276659	0.106467657
0.455392799	0.034724699

	Non-proestrus
	A. (
CAMK2N1	PAFAH1B2
CARS	PDE4B
CYP24A1	PDE7A
DBF4	PGM3
HMGN2	PNP
KMT5A	REV1
LRRC17	RPS19BP1
MDH2	RRAD
MIS18A	SEC62
	B.
Diseases and Disorders	
Inflammatory disease	
Inflammatory response	
Organismal injury and abnormalities	
Molecular and Cellular Functions	
Cellular development	
Cellular compromise	
Cell cycle	
	D. *T*
Development and Function	
Organismal development	
Embryonic development	

Connective tissue development and function
Associated Network Functions
Cellular development, inflammatory disease, inflammatory response

Genes targeted by differentially expressed mi	
SEC23A	ABCB9
SNX5	APOO
SYT4	ARHGEF38
THAP12	AVEN
TMED7	CASP3
TNFAIP2	CCNJ
TNFRSF10C	CMTR2
TP53	CNMD
UBE2V2	DSCR8
ZNF420	

Differences in top diseases and biofunction

P Value	Diseases and Disorder
3.84E-02 - 3.84E-05	Organismal injury and
3.84E-02 - 3.84E-05	Reproductive system
4.17E-02 - 4.17E-05	Cancer

C. Top molecular and cellular functions

P Value	Molecular and Cellular
2.05E-02 - 5.26E-07	Cellular movement
3.75E-04 - 3.75E-04	Cellular death and surv
2.62E-03 - 2.62E-03	Cellular development

op physiological system development and fun

P Value	Development and Fun
4.17E-02 - 1.31E-03	Embryonic developme
1.29E-02 - 1.29E-02	Connective tissue deve

1.93E-02 - 1.93E-02	Tissue morphology

E. Top associated network functions

Score	Associated Network Fu
6	Organismal injury and reproductive system d

Proestrus	
RNAs	
FGD4	SLC25A27
FRMD4B	SLC38A1
GPR137B	TMCC1
HMBS	TRIM71
METAP1	XKR8
MITF	ZCCHC11
MYC	ZIM3
RGMB	ZNF181
SLC14A1	
ns	
s	*P* Value
abnormalities	4.96E-02 - 2.77E-14
disease	2.15E-02 - 2.77E-14
	4.96E-02 - 1.27E-10
Functions	*P* Value
	3.77E-02 - 4.47E-07
vival	4.91E-02 - 5.61E-06
	4.97E-02 - 1.38E-06
ction	
ction	*P* Value
ent	3.30E-02 - 2.12E-05
elopment and	1.79E-02 - 6.10E-05

	7.88E-05 - 7.88E-05
unctions	Score
abnormalities, lisease, cancer	19

Table of Materials

Click here to access/download;Table of Materials;JoVE_Materials_Reviewed.xls

Name of Material/ Equipment	Company	Catalog Number
C57BL/6J mice	The Jackson Laboratory	000664
UltraPure Water	Thermo Fisher Scientific	10813012
Sterile plastic pipette	Fisher Scientific	13-711-25
Frosted Microscope Slides	Thermo Fisher Scientific	2951TS
Light microscope	Microscope World	MW3-H5
Ketathesia- Ketamine HCl Injection USP	Henry Schein Animal Health	55853
Xylazine Sterile Solution	Lloyd Laboratories	139-236
Ethanol	Fisher Scientific	BP2818100
21G gauge needle	BD Biosciences	305165
Syringe	Fisher Scientific	329654
Operating Scissors	World Precision Instruments	501221, 504613
Tweezer Kit	World Precision Instruments	504616
-80 ˚C freezer	Forma	7240
Spectrum Bessman Tissue Pulverizers	Fisher Scientific	08-418-1
RNase-free Microfuge Tubes	Thermo Fisher Scientific	AM12400
TRIzol Reagent	Thermo Fisher Scientific	15596026
Direct-zol RNA MiniPrep Plus	Zymo Research	R2071
NanoDrop	Thermo Fisher Scientific	ND-ONE-W
miScript II RT kit	Qiagen	218161
Mouse Inflammatory Response & Autoimmunit	Qiagen	MIMM-105Z
Thin-walled, DNase-free, RNase-free PCR tubes	Thermo Fisher Scientific	AM12225
miRNeasy Serum/Plasma Spike-in Control	Qiagen	219610
Microsoft Excel	Microsoft Corporation	
Ingenuity Pathway Analysis	Qiagen	
R Software	The R Foundation	
Thermal cycler or chilling/heating block	General Lab Supplier	
Microcentrifuge	General Lab Supplier	
Real-time PCR cycler	General Lab Supplier	
Multichannel pipettor	General Lab Supplier	

RNA wash buffer	Zymo Research	R1003-3-48
DNA digestion buffer	Zymo Research	E1010-1-4
RNA pre-wash buffer	Zymo Research	R1020-2-25
Ultraviolet ozone analyzer	Teledyne API	Model T400
Mass flow controllers	Sierra Instruments Inc	Flobox 951/954

Comments/Description
8 weeks old
Capacity: 1.7mL
10X and 20X objective
90 mg/kg. Controlled drug.
10mg/kg. Controlled Drug.
Dilute to 70% ethanol with water.
1mL
14cm, Sharp/Blunt, Curved and 9 cm, Straight, Fine Sharp Tip
Capacity: 10 to 50mg
1.5 mL
for 20 µl reactions
https://office.microsoft.com/excel/
https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/
https://www.r-project.org/

48 mL
4 mL
25 mL
http://www.teledyne-api.com/products/oxygen-compound-instruments/t400
http://www.sierrainstruments.com/products/954p.html

jove JOURNAL OF VISUALIZED EXPERIMENTS

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:	Lung miRNA profiling across the estrous cycle in ozone exposed mice
Author(s):	Fuentes, Nathalie ; Silveyra, Patricia

Item 1: The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via:

☒ Standard Access □ Open Access

Item 2: Please select one of the following items:

☒ The Author is **NOT** a United States government employee.

□ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

□ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of The Journal of Visualized Experiments; "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and(c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this **Section 6** is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Protection of the Work.** The Author(s) authorize JoVE to take steps in the Author(s) name and on their behalf if JoVE believes some third party could be infringing or might infringe the copyright of either the Author's Article and/or Video.

9. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

10. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

11. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole

discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

12. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

13. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

14. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement is required per submission.

CORRESPONDING AUTHOR

Name:	Patricia Silveyra		
Department:	Pediatrics		
Institution:	Penn State College of Medicine		
Title:	Associate Professor		
Signature:		Date:	06/23/18

Please submit a **signed** and **dated** copy of this license by one of the following three methods:

1. Upload an electronic version on the JoVE submission site
2. Fax the document to +1.866.381.2236
3. Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02140

Editorial comments:

1. The editor has formatted the manuscript to match the journal's style. Please retain the same.
We thank the editor for the formatting and comments.

2. Please address all specific comments marked in the manuscript.
We have addressed every point in the revised version.

3. Please change the scale bar units for the Figure 1.
We appreciate the suggestion. The scale has been corrected.

4. Please remove all commercial language from your manuscript and use generic terms instead. All commercial products should be sufficiently referenced in the Table of Materials and Reagents.
We have revised the manuscript and removed all commercial language.

5. Please obtain explicit copyright permission to reuse any figures from a previous publication. Explicit permission can be expressed in the form of a letter from the editor or a link to the editorial policy that allows re-prints. Please upload this information as a .doc or .docx file to your Editorial Manager account.
BOSD editorial policy allows figure and table reproduction with proper citation under the Creative Commons CC0 waiver. Figure legends include appropriate citations. The editorial policy documents have been now uploaded (note these are .pdf files)

BMC Research Notes

EDITORIAL Open Access

Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals

Iain Hrynaszkiewicz* and Matthew J Cockerill

Abstract

Copyright and licensing of scientific data, internationally, are complex and present legal barriers to data sharing, integration and reuse, and therefore restrict the most efficient transfer and discovery of scientific knowledge. Much data are included within scientific journal articles, their published tables, additional files (supplementary material) and reference lists. However, these data are usually published under licenses which are not appropriate for data. Creative Commons CC0 is an appropriate and increasingly accepted method for dedicating data to the public domain, to enable data reuse with the minimum of restrictions. BioMed Central is committed to working towards implementation of open data-compliant licensing in its publications. Here we detail a protocol for implementing a combined Creative Commons Attribution license (for copyrightable material) and Creative Commons CC0 waiver (for data) agreement for content published in peer-reviewed open access journals. We explain the differences between legal requirements for attribution in copyright, and cultural requirements in scholarship for giving individuals credit for their work through citation. We argue that publishing data in scientific journals under CC0 will have numerous benefits for individuals and society, and yet will have minimal implications for authors and minimal impact on current publishing and research workflows. We provide practical examples and definitions of data types, such as XML and tabular data, and specific secondary use cases for published data, including text mining, reproducible research, and open bibliography. We believe this proposed change to the current copyright and licensing structure in science publishing will help clarify what users – people and machines – of the published literature can do, legally, with journal articles and make research using the published literature more efficient. We further believe this model could be adopted across multiple publishers, and invite comment on this article from all stakeholders in scientific research.

Introduction

Much has been written about, and support stated for, sharing and publishing scientific data, in recognition of the benefits for the economy [1], scientific discovery [2] and public health [3]. Maximizing the potential of scientific data sharing for the discovery of new knowledge involves reducing barriers to data dissemination, reuse, reproducibility and integration. Licensing, ownership, copyright and intellectual property present legal obstacles to data integration and reuse, which has led to the development of, and calls for, licensing standards for open data; where data are explicitly placed in the public domain with legal rights of the owners waived [4].

BioMed Central has previously stated that the concept of open data, analogous to its policy on open access to journals, goes beyond making data freely accessible. Data should also be free to distribute, copy, re-format, and integrate into new research, without legal impediments [5]. This position is consistent with the Panton Principles, which hold that for society to reap the full benefits of scientific research the published body of knowledge must be open – readily available such that it can be evaluated, reused, criticized and integrated with other knowledge without restrictions [6]. For the remainder of this article the term ‘open data’ is reserved exclusively for data available according to these principles.

Unfortunately much data – and other content – freely available on the web are available under restrictive or

* Correspondence: iain.hrynaszkiewicz@biomedcentral.com
BioMed Central Ltd, 236 Gray’s Inn Road, London WC1X 8HB, UK

ambiguous terms, which risks impeding or potentially criminalizing secondary users of scientific data. According to evidence submitted to the UK Government's intellectual property review by the Wellcome Trust, 87 per cent of the material housed in the full-text scholarly archive UK PubMed Central is unavailable for legal text and data mining [7]. A key finding of a more recent report, commissioned by JISC, was a need to overcome legal restrictions and uncertainties surrounding text mining of scientific literature [8].

Indeed, as recognition of the value of shared life science data has increased, so has recognition of intellectual property and copyright as barriers to progress. Writing in *Nature* in 2009, Schofield *et al*., urged that "any restrictions on use should be strongly resisted and we endorse explicit encouragement of open sharing" [9]; and Conway and VanLare in JAMA, in 2010, called for US health care data to be available without intellectual property constraints [10]. Waiver of all intellectual property rights in research data is central to the achievement of an "information commons", advocated by organisations such as Sage Bionetworks, to enhance the (slowing) pace of drug discovery.

The genomics community has shown leadership in establishing a framework for an "information commons", engrained in the Bermuda Principles, and have established built-in temporal latencies to data for knowledge (when data are released), and rights (when rights restricting use are removed) [11]. Researchers in this community typically must release their genetic sequence data immediately, and within 6–12 months release their exclusive rights in that data. During this relatively short embargo researchers have their opportunity to exploit the data for their discoveries, after which the community at large can benefit, if they wish, from the new data. A similar model for data release has since been proposed for clinical trials, although is probably far from implementation [12]. A number of factors seem to have led to a successful culture of sharing in the genomics community: a need to collaborate and share to achieve a major goal (the sequencing of the human genome); effective mechanisms and infrastructure for sharing large amounts of data (well-funded genetic sequence databases); scientific community and funding agency mandates to share data; and importantly, in the context of this article, successful collaborations with the publishing community. Journals, their editors and publishers, supported implementation of the Bermuda Principles by, for example, requiring accession number for data deposits as a condition of manuscript submission or publication.

BioMed Central in its August 2010 open data statement [5] and subsequent cross-publisher Publishing Open Data Working Group meeting identified that open data in journal publications could be implemented by specifying that, from a specific date, any author submitting to a journal or publisher agrees to dedicate the data elements of their article and supplementary material (in particular, additional data files; also known as "supplementary" data files) to the public domain [13]. Much of the contents of academic journals could be considered as data but licensing terms cannot be applied retroactively by publishers without authors' consent, and any changes to authors' agreements should ideally be made in consultation between authors and publishers.

This article aims to describe practically what is needed from publishers to explicitly dedicate data within open access journals to the public domain, and discusses the implications of this development for authors, editors, publishers and funders of research. Illustrative examples and use cases are provided throughout the article. In this article "open access" is defined according to the Budapest Open Access Initiative definition [14].

Applying the right license to published research and data

The internet has revolutionized the way we access and distribute information, enabling virtually anyone to post content online. There is much potential in rapidly sharing content on the web, but releasing content without information, or with ambiguous information, about if and how it can be shared and reused can also cause problems – especially for data.

Open access publishing of peer-reviewed journal articles commonly utilizes the legal tools – licenses – prepared by Creative Commons. BioMed Central, Public Library of Science, Nature Publishing Group, BMJ and many others publish open access articles where the authors retain the copyright to their work. Authors typically apply a Creative Commons attribution license (CC-BY), or variation of it, which means anyone is free to copy, reuse, distribute and make derivatives from their article provided that there is attribution of the original author(s). However, many "open access" publishers place restrictions on commercial reuse of published articles (papers) and on creation of derivative works, which can include text mining in some jurisdictions. Additionally, some commercial publishers' terms and conditions, by contract, can prevent text mining in any jurisdiction. Commercial use restrictions have been strongly discouraged – their use described as amounting to "pseudo open access" – as authors will not reap the full benefits of paying for open access publication (for example figures could not be uploaded to Wikipedia with commercial use restrictions) [15,16]. BioMed Central supports unrestricted use of open access content including commercial use and as such requires authors to apply a CC-BY license by default. BioMed Central's full text corpus of open access research articles published under CC-BY is available for free distribution, reuse and creation of derivatives with no commercial use restrictions – with data mining research strongly encouraged [17]. For data published by scholarly publishers, the Association of Learned and Professional Society Publishers and International Association of Scientific, Technical, & Medical Publishers (STM) issued a joint statement in 2006 supporting sharing of raw datasets among scholars and recommending that publishers do not require transfer of copyright in data submitted for publication [18].

Copyright and data

The policies and guidelines of many academic institutions advise researchers to establish intellectual property and copyrights at the start of any project (although whether the issue of data ownership is consistently addressed by researchers is unclear [19]). Copyright cannot generally be asserted in facts, only the ways in which they are presented. At a basic level raw data are merely simple, mathematical, descriptions of facts and to claim copyright a scientist would need to exert individual judgment, expression or skill in their representation. For example, Einstein could not claim copyright in the formula $E = mc^2$, but could in text explaining the theory behind it [20]. You could conclude from this that copyright and associated licenses and attribution requirements cannot legally be applied to data. However, there are many levels at which data – particularly digital data derived and integrated from different sources – and collections of data and metadata can operate and be represented, and many ways in which copyright law is applied in different jurisdictions.

In the US the law focuses on creativity ("Copyright does not protect facts, ideas, systems, or methods of operation, although it may protect the way these things are expressed") but in Australia originality is more important – and copyright may well apply to research data "in the same way that it applies to written works like books, journal articles and reports" [21]. In the European Union "*sui generis*" rights exist to protect data within digital databases – effectively, copyright – which can, furthermore, be implemented differently by member states. Because of these substantial international legal differences regarding how copyright can be applied to data, there are inherent difficulties in ascertaining the extent of copyright in a dataset. A more comprehensive summary of the different approaches to copyright in data and databases can be found in [22]. All of these issues compound the uncertainty about what an individual or machine (such as a computer crawling the web) can do, legally, with information they download from the internet, including from journals.

Licenses and waivers for data

A license is a legal instrument for a copyright holder or content producer to enable a second party to use their content, and apply certain conditions and restrictions to those uses. A waiver is also a legal instrument but is designed for a rights holder to *give up* their rights, rather than assert them. For a comprehensive guide to the different approaches to the licensing of research data see [23].

Placing restrictions on the reuse of scientific information, particularly data, slows down the pace of research. Furthermore, legal requirements for attribution ingrained in licenses such as CC-BY can prohibit future research across large collections of content – as commonly happens in data mining research. Consider the Human Genome Project: a watershed moment for scientific data sharing and collaboration. Without the collective effort of many different research institutions, commercial organizations and individual scientists the sequencing of the human genome would not have been possible. But if a researcher wishing to query the human genome database as part of a new research project was legally required to attribute all the – probably thousands – of data contributors, by providing a link back to or citation, this would be unmanageable, and probably un-publishable in the context of a traditional research paper's reference list.

International legal differences, described earlier, are another important reason to apply specific, appropriate legal tools to data. Also, it can be unclear what license to attach to copyright in a dataset or structure (for example a textual description of building the dataset could fall under CC-BY, but if source code were used rather than text it might not). This is an area of confusion where no licensing standard exists. Therefore, to eliminate legal impediments to integration and re-use of data, such as this stacking of attribution requirements in large collections of data, and to help enable long-term interoperability an appropriate license or waiver specific to data should be applied. There are a number of conformant licenses and waivers for open data [24], of which Creative Commons CC0 (http://creativecommons.org/publicdomain/zero/1.0/) is widely recognized. Under CC0, authors waive all of their rights to the work worldwide under copyright law and all related or neighboring legal rights they have in the work, to the extent allowable by law. Legal experts have recommended the use of standard, globally accepted licenses for data instead of developing *ad hoc* models [25].

The case for CC0 for scientific data

The Creative Commons' website catalogues a number of different organizations – publicly and privately funded – which use CC0 for data [26]. These include:

- Genomes Unzipped, which “aims to inform the public about genetics via the independent analysis of open genetic data, volunteered by a core group of genetics researchers and specialists”
- GlaxoSmithKline (GSK), a leading pharmaceutical company, has dedicated data on more than 13,500 compounds known to be active against malaria to the public domain [27].
- The British Library and Cologne-based Libraries, which have released large amounts of bibliographic data under CC0 [28]
- FigShare (http://figshare.com/), a freely-accessible repository for scientific content including images, video and data, uses CC0 for datasets

Data repositories are particularly relevant users of waivers and licenses for research data. Although there are many data repositories in life sciences (for a list see http://www.datacite.org/repolist), which are growing in size and number, not all scientific domains have a common repository and journals often function as repositories when data are included as additional files (supplementary material). Dryad (http://datadryad.org/) is an international repository for the datasets supporting published, peer-reviewed journal articles across the biosciences which requires authors to explicitly place deposited data in the public domain using the CC0 waiver. An entry on the Dryad weblog sets out cogently why CC0 is the most effective solution for achieving its goals:

“*By removing unenforceable legal barriers, CC0 facilitates the discovery, re-use, and citation of [that] data…*
“*Furthermore, Dryad’s use of CC0 to make the terms of reuse explicit has some important advantages:*

- *interoperability: Since CC0 is both human and machine-readable, other people and indexing services will automatically be able to determine the terms of use.*
- *universality: CC0 is a single mechanism that is both global and universal, covering all data and all countries. It is also widely recognized.*
- *simplicity: there is no need for humans to make, and respond to, individual data requests, and no need for click-through agreements. This allows more scientists to spend their time doing science.*” [29]

Dryad’s policy ultimately follows the Science Commons’ recommendations, set out in their Protocol for Implementing Open Access Data [30].

The online laboratory notebook software LabArchives (http://www.labarchives.com/), which includes the ability to share data privately and to publish datasets publicly and permanently online, also uses CC0 for public datasets [31].

Concerns about public domain dedication of data

Credit where credit’s due – attribution and citation

A common concern about moving from an attribution license, such as CC-BY, to CC0 for data and waiving attribution rights is that academic credit (citations) will be lost if there is no longer a legal requirement to attribute the original rights holder (author). While attribution can sometimes be achieved in the same way as citation the two practices serve different purposes. Attribution is a legal tool designed to permit copying, distribution, and creation of derivative works such as translations. As copyright does not protect ideas (in the US), to give scientists credit for their ideas the established norm in scholarly communication is citation [32].

Consider a scientist paraphrasing a concept put forward in a peer’s research article. He or she does not legally have to cite their peer’s published paper, but it is beneficial or possibly essential for the validity and reliability of the subsequent work to specify the source(s) of assertions made. Community norms are enforced by the community, and in science unacceptable citation practices are typically identified and resolved through peer review, and the publication ethics and editorial policies of peer-reviewed journals. See Table 1 for common citation and attribution events in scholarly communication. These examples are for illustrative purposes and do not constitute legal advice.

The examples in Table 1 demonstrate that, although they can sometimes be achieved in the same way, attribution and citation are not the same. Citations are much more important and relevant than attribution when tracking scholarly outputs and giving appropriate credit for individuals’ contributions.

Compared to legal requirements, cultural norms benefit from flexibility, and can evolve with the community which established them. In other words, using norms retains control and decision making within the research community, instead of taking it out of our control and handing it to lawyers and judges. Many scientific ideas, after a number of years, become undisputed and the community may deem it unnecessary for credit to be rigorously applied. For example, it would today be very unusual for an article describing a DNA sequencing experiment to cite the original work by Watson and Crick that elicited DNA’s structural properties. This is a cultural norm at work, where an idea is now so widely accepted, and the initial authors clearly recognized for their discoveries (in citations and prizes) a citation is not needed.

Jonathan Rees, formerly of the Creative Commons, said of community norms for influencing behavior over legal requirements: “*For widest latitude of use and best scalability, and therefore greatest return to the research community, the entirety of the data set, including any incidental*

Table 1 Attribution vs. citation in (re)uses of open access scientific content published under a Creative Commons Attribution license (CC-BY)

Activity	Attribution and/or citation?	Explanation
Printing an article for display at a conference	Attribution	Printing an article is redistribution so covered by copyright (and attribution is achieved inherently by the authors' names and copyright ownership being stated on the article)
Translating article for publication in another journal	Attribution + citation	Attribution is required as a translation is a derivative work, and most journal duplicate publication policies (an ethical requirement) require citation of the original paper for republications
Paraphrasing a concept or finding within a paper	Citation	If you rely on another scientists idea for your work credit is due to the previous author through citation
Reusing a figure, table or graph	Attribution + citation	Reusing a figure, table or graph is copying and redistribution, so requires attribution; by presenting another scientist's representation of their data you need to give credit to their original work
Publication of a reanalysis of data published as an additional file in a journal	Citation	The source of the data being reanalyzed may not legally need to be attributed if copyright does not apply (e.g. in the US), even if the data are included with the secondary publication, but for the reanalysis to stand up to scrutiny – and pass peer review – the source of the data must be cited

copyrightable elements, should be dedicated to the public domain. Note that public domain is not incompatible with a request for attribution or other terms of use following community norms. Such a request may be as effective – or more effective – at getting users to follow desired practices as any attempted legal restrictions [33]."

To our knowledge there have been no empirical studies of the citation of scholarly datasets assigned public domain dedication licenses compared to a comparable group available under attribution licenses. However, given public domain dedication – and specifically CC0 – is intended to maximize the potential for data discovery, reuse and therefore citation, it would be reasonable to hypothesize that citation potential of public domain data would be unaffected or might even increase. With CC0 there is no need for transfer agreements and preconditions, which inherently impede further (re)uses of data. Sharing of microarray experimental research data underlying journal articles has been associated with increased citation share [34], and increased reproducibility and repeatability of results [35]. In social science data collections funded by the National Institutes of Health and National Science Foundation in the US, data sharing has been associated with "many more times the publications" than collections where data were not shared [36]. Linking of publications to supporting datasets has also been associated with more citations to the linked paper in the marine science journal *Paleoceanography*, according to a conference abstract, and in the field of astronomy according to a preprint paper [37].

Competition

Researchers who apply CC0 to their data, or any other product of their scholarship, waive all rights in that data allowable by law. Such a waiver has been described as an "unattractive option for data whose creators have yet to fully exploit them, academically or commercially" [23]. This is true, but a waiver or license required by a journal or publisher generally applies only in the context of data submitted for publication. If a portion of a large database was analyzed and an additional data file included for publication the larger, unpublished, body of work would retain whichever license the researchers, their employers or institutions require. In other words, researchers remain in control of what they chose to publish – what they submit to a journal – and a change in the publishers' license does not affect this. Moreover, waivers and licenses for journal articles do not replace existing, established community norms for sharing of some data types (e.g. depositing microarray and genetic sequence data in appropriate databases) – nor do they affect requirements of many journals for sharing readily reproducible materials including raw data on request [38].

There is a trade-off between the additional opportunities which may result from transparency (such as new collaborations, secondary use) and the threat, improbable or otherwise, that opening up data may be valuable to competitors. Certain types of research, such as genetic sequencing to elucidate susceptibility to disease, generates far more data than one research team could conceivably analyze – which logically lead to sharing and collaboration. A number of companies have opened up some of their data and seen benefits [39]. A lot of data may have commercial value but much raw data, such as protein sequences of potential drug targets, are just the beginning of a knowledge-discovery process. More can be gained by "pre-competitive" sharing with the waiver of intellectual property. Such an approach is being championed by Sage Bionetworks in the US [40].

Plagiarism

Plagiarism is research misconduct and an unfortunate but ineliminable occurrence in scholarship. Plagiarism and the potential for plagiarism have increased with the proliferation of digital access to information [41]. Plagiarism is often not illegal, but it is certainly unethical, and undoubtedly damaging for the career of someone guilty of perpetrating plagiarism. Effective online tools for detecting plagiarism exist, such as CrossCheck (http://www.crossref.org/crosscheck/index.html), as does human detection via the peer-review process. Removal of a legal requirement for attribution for data elements of articles would be unlikely to impact on the potential for plagiarism. In addition, CC0 would not apply to the main, copyrightable text of articles.

Other safeguards

Public domain dedication of data does not mean that those who generated the data cannot express certain wishes about how the data are used. The Panton Principles frequently asked questions (FAQs) state: "*You should always aim to follow any reasonable requests made by the data owners/publishers. These may be explicit or may be implicitly understood by the community. You should make an effort to understand any relevant 'community norms' for the data you are using* [42]*.*"

A code of conduct has been proposed for those wishing to reuse clinical trial data obtained from other researchers [43] and a clinical trial dataset published in the journal *Trials*, by Sandercock *et al*., has requested that "any publications arising from the use of this dataset acknowledges the source of the dataset, its funding and the collaborative group that collected the data." [44].

Electronic publishing platforms provide further safeguards to ownership and authorship of published content, in the form of date stamping manuscript submissions and version control in some repositories, such as Edinburgh DataShare (http://datashare.is.ed.ac.uk/).

Citation of articles and datasets is facilitated through standard citation formats – such as those advocated by DataCite where persistent dataset identifiers, such as digital object identifier (DOI) names, are displayed as linkable, permanent URLs – and are increasingly supported by some publishers [37].

What do we mean by data?

There are numerous definitions of data. According to Wikipedia data "are qualitative or quantitative attributes of a variable or set of variables. Data are typically the results of measurements and can be the basis of graphs, images, or observations of a set of variables," [45]. Data can exist electronically or non-electronically, so a definition that includes electronic access is important, in the context of integration, reuse and data mining of online scholarly content. The Copyright, Designs and Patents Act, part of UK statute, uses a broad definition of databases incorporating electronic access:

> "*Databases(1)In this Part "database" means a collection of independent works, data or other materials which—*
> *(a)are arranged in a systematic or methodical way, and*
> *(b)are individually accessible by electronic or other means.*
> *(2)For the purposes of this Part a literary work consisting of a database is original if, and only if, by reason of the selection or arrangement of the contents of the database the database constitutes the author's own intellectual creation.*"

Other definitions, such as at the United States National Science Foundation 'DataNet' program, have been broader and implied anything capable of existing digitally, including publications and software, could be considered data [46]. The former definition is more broadly applicable to data which can be harvested, mined or downloaded from open access journals – and to which CC0 rather than CC-BY should apply. This inevitably means information which can be processed by machines as well as being transferred electronically by them (e.g. papers attached to emails) [47]. It is not possible to comprehensively define and account for all data and data file types, particularly given the rapidly evolving nature of data and text mining applications, but a number of general examples and definitions follow below. We strongly encourage readers to comment on these data definitions and provide additions and amendments. These examples intentionally do not include domain-specific data standards (agreed upon formats for disseminating and presenting particular types of scientific experiments), which are comprehensively catalogued by BioSharing (http://biosharing.org/?q=standards).

Tabular data

Data elements organized in columns and rows – a table – are extremely common in scientific publications. While attribution would be required for reproduction in whole or in part of a table as presented in a journal publication, the individual values and collection of values should be considered as data and therefore open. Data, in the course of a scientific experiment, are usually collected at a greater level of detail than are reported in a paper, with tables reporting summary or mean values. Although these data are aggregated from the raw data they remain numerical representations of a fact, and therefore data. Tables are furthermore often included as additional files in a variety of formats including PDF, HTML/XML, DOC and Excel/CSV. Ideally all tables included in the main body of a journal article should also be included as additional CSV files – an open, machine-readable format – but when they are not present as tables in journal articles CSV would represent good practice for tabular data. Proprietary file types, such as Microsoft Office, and formats which are not readily editable, such as PDF, are not recommended for tabular data provided as additional data files.

Graphs and graphical points

Graphs, graphical representations of relationships between variables, are ostensibly images and therefore not, when considered as a collective entity, data. However, the individual data points underlying a graph, similar to tables, certainly are. An example of best practice when submitting a manuscript with a graph to a peer-reviewed journal would be for authors to also submit accompanying CSV tables with the corresponding data points, so that graphs could be re-plotted. Although this practice is required by some specific journals it is not widespread. However, software tools exist that are capable of "scraping" underlying data points from graphs and images (for example http://www.chardta.com/) and can be useful, for example, for enhancing the discoverability of scientific information by exposing underlying data points to internet search engines.

XML

According to the World Wide Web Consortium (W3C), "Extensible Markup Language, abbreviated XML, describes a class of data objects called XML documents and partially describes the behavior of computer programs which process them" [48]. XML is widely used as a standard for data transfer and for creating versions of works intended for machine reading, and therefore to be used as data. Therefore for our purposes we can assume XML files are data. XML has many applications in science and is frequently published with journal articles as additional files in BioMed Central journals as well as underlying the online articles themselves. XML forms the basis of many domain-specific data standards such as Gating-ML in flow cytometry, FuGE-ML in functional genomics, GelML in gel electrophoresis and so on (see: http://biosharing.org/standards_view).

Bibliographic data

Bibliographic data have been historically described as information not included in the full text and images included with an article, which includes reference lists. "Core bibliographic data" have been further described as "data which is necessary to identify and / or discover a publication" and defined under the Open Bibliography Principles:

- names and identifiers of author(s) and editor(s)
- titles
- publisher information
- publication date and place
- identification of parent work (*e.g.* a journal)
- page information
- Uniform resource identifiers (URIs) [49]

Therefore, these core bibliographic data should be considered open data.

RDF

Resource Description Framework (RDF) is a standard language for encoding data and metadata on the web. It is designed to indicate the relationship between online objects in a human and machine-readable way, and facilitate merging of data between different sources even if the underlying schemas of the sources are different. RDF forms the basis of the semantic web, and is a core component of achieving Tim Berners-Lee's vision of Linked Data on the web [50]. RDF provides new opportunities for data and knowledge management in life sciences, chemistry, and publishing. All BioMed Central journal articles, for example, contain embedded RDF, which conveys harvestable information about content, such as authors, licensing information and the unique identifier for the article [51].

What aren't data?

Although source code may be represented as data and is certainly machine readable there are a wide range of existing licensing systems and community norms that exist around software. Therefore we choose to regard software, compiled code, and source code as a separate category and not as data. Specific licenses and repositories have been developed for source code for software and Open Source Initiative compliant licenses [52] are recommended. Files pertaining to programming languages can be included as additional files with journal publications, either directly in formats such as SQL or indirectly in compressed or packaged file formats such as ZIP.

There are myriad file types which can be published as additional files but amongst the most common, in BioMed Central journals, are those usually pertaining to text and written works – PDF and DOC/DOCX and HTML (a full list of published additional data files is available on request from BioMed Central [53]). Caution is recommended in the interpretation of these objects as open data.

Implementing a variable license for open access research and data

Setting date (CC) Zero

Creative Commons licenses (CC-BY, specifically) have provided an effective and penetrative solution for digital copyright in open access scientific works (papers). But as the nature of the published scientific paper (article) has evolved then so too should the copyright and licensing structure which authors apply to their content. Published articles are increasingly collections of different digital objects, perhaps including a few thousand words of text, half a dozen images and a similar number of CSV or XML (data) files.

The fact that CC-BY is a suboptimal license for data does not mean that the many thousands of published authors have done something wrong, as CC-BY was (and often still is) the best instrument available when copyright license agreements for open access journals were prepared (although, CC-BY version 4.0, currently in draft form, aims to tackle the issue of *sui generis* database rights [54], described earlier). A number of data repositories, including Dryad and FigShare, initially asked authors to make their deposited data available under a Creative Commons attribution license but have since changed their policy (https://twitter.com/#!/figshare/statuses/50241486796754944). However, licenses and waivers cannot be applied retroactively by a publisher without explicit consent of the copyright holder (s) – in the vast majority of cases at BioMed Central, the authors. A small number of datasets remain in Dryad which are not available under CC0, as explicit agreement from data depositors (rights holders) could not be obtained to change the terms of data release [29].

A change to BioMed Central's standard license agreement to include a CC0 waiver for published data would remove ambiguities about the copyright and attribution requirement status of parts of published articles and associated data files, and enable instead the application of scientific, cultural norms that meet the needs of scientists better than an inflexible legal instrument [33,55]. To implement open data in journal publications the new license agreement would need apply to all authors from a specific date, such that any author submitting to a journal/publisher agrees to dedicate the data elements of their article and additional files to the public domain. A proposal for how this could be reflected in published articles' copyright license statement follows:

> "© 2012 <Author> et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data included in this article, its reference list(s) and its additional files, are distributed under the terms of the Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/; http://www.biomedcentral.com/about/access)."

This would be further indicated in article metadata, RDF and on the journal and publisher's policy pages and author submission pages online. This addition to the statement aims to succinctly summarize that data in published articles generally originates from three sources – tabular data or text-minable factual data (e.g. numerical instances of a particular word or phrase, such as gene/protein names) in the main body of an article; additional (supplementary) files such as XML and CSV file extensions that include data; and the reference list (bibliographic data). From a legal perspective, we might need to be more comprehensive about our definitions of data (see 'What do we mean by data?', above) in the full legal code of a new license and in guidelines accompanying this proposed change. And practically, authors would need to agree to apply two different legal tools to content they submit for publication as part of the submission process – with links to both CC-BY and CC0, for example, in journal submission checklists.

The need for change in author behavior, as a result of this proposed modification, is minimal. CC0 would not apply to nor affect the availability of data not submitted for publication; we would instead be asking authors to apply different terms of use to a proportion of the content they already publish. By submitting a manuscript and all associated files to a BioMed Central journal authors already confirm that they agree to all terms of the BioMed Central Copyright and License Agreement [56], including that they are able to apply a CC-BY license. This proposed change to the license agreement should provide clarity to the licensing of specific components of published articles and does not represent a substantial change to the overall license agreement for authors' published work.

Open data by default – opt out

Public domain dedication of data, while universally desirable, is not always universally possible. Authors, for the most part, prepare research articles in the context of employment by a third party and can be subject to licensing terms which supersede the standard terms of a publisher. This already happens for a small proportion of the content published by BioMed Central and undoubtedly by other publishers, such as some articles funded by the World Health Organization and US government. Therefore, any submitting author who is not able to agree to all terms of the BioMed Central Copyright and License Agreement should contact the journal editorial office at the earliest opportunity – ideally before, during or immediately after manuscript submission. The onus is on the authors and, if applicable, their employers to decide on the applicability of the publisher's standard license agreement to their work and whether an exception is needed. Alternatives to the standard license agreement can be discussed. This process of checking suitability of the standard license agreement, and requesting alternatives where necessary, therefore already happens for all manuscripts submissions. The owners of the rights (if any) in the data must decide if their data can be made open. According to the Panton Principles FAQs "[i]n most cases, the people who make a decision to publish, and were intimately involved in the generation of the data, should be making this decision." [42]

Publishing platform developments

Implementing a new license agreement will have technical as well as policy and procedural implications. While it is not possible to specify these changes in detail here, and is beyond the scope of this document, the following would be essential for implementation:

- Tagging of articles and data files published under a non-standard license agreement (where authors have opted out of the new default open access-open data license)
- Editing standard embedded license information in article XML metadata and RDF and a tool to automate insertion of non-standard licensing terms
- Insertion of license information to additional files and associated metadata

Furthermore, the following would be desirable to enhance the discoverability and usefulness of open data in journal articles:

- Tagging and classification of published data files, for example by file type
- A tool to automatically discover and aggregate additional files
- A tool to (retrospectively) associate data objects with papers on the web
- Approaches to associating published datasets with journal articles which go beyond hyper-linking, such as through linked data methods
- Searching within and filtering of additional files

Open data in science use cases in published and unpublished contexts

There are many uses for open data but probably many more as yet unknown. As stated by Tim Berners-Lee and Nigel Shadbolt in *The Times* on New Year's Eve 2011, "*One reason that the worldwide web worked was because people reused each other's content in ways never imagined or achieved by those who created it. The same will be true of open data.*"

The examples that follow focus on licensing and reuse of data included with and/or harvestable from journal publications, in the context of the proposed change to BioMed Central's standard license agreement, above, for open access and open data journal articles.

Example #1 – analysis of a large clinical trial dataset

In April 2011 Sandercock *et al.*, published in *Trials* 'The International Stroke Trial Database' [44], which "aimed to make individual patient data from the International Stroke Trial (IST), one of the largest randomised trials ever conducted in acute stroke, available for public use, to facilitate the planning of future trials and to permit additional secondary analyses."

The "database", including 19,000 anonymized individual patient data, is available with the journal article as a 4.3 Mb CSV file (http://www.trialsjournal.com/content/supplementary/1745-6215-12-101-s1.csv) under a CC-BY license. With the new license agreement as proposed in this article the CSV file would be available for reuse without a legal requirement for attribution engrained in CC-BY. Secondary uses for this dataset might include a novel secondary analysis by a different group of researchers, and analysis and integration of the dataset in the context of a systematic review and meta-analysis of randomized trials of heparin and/or aspirin in acute ischemic stroke. Both of these activities might conceivably result in further publications. Although there would be no legal requirement for attribution, for any secondary article about this dataset – or indeed any systematic review – to be scientifically valid it would need to cite its source(s) of data.

Example #2 – Application of magnetic resonance techniques to cross-species comparative studies

Magnetic Resonance Imaging (MRI) techniques are used to better understand the evolution of specific traits in animals and cross-species comparisons (for example in primates) are particularly important. But due to ethical, practical and funding limitations single studies typically are only able to consider one or two species. There is only one publicly available dataset that has brains from multiple primate species scanned according to a common protocol and these scans (of 11 species) were recorded (in vivo) well over a decade ago, and so do not meet the quality criteria that underpin more recent brain morphometric algorithms of the kind required for cross-species studies of brain structure. However, a review of this area of research found that "the major barrier to cross-species MR-based brain morphometry is not the lack of data nor analytical tools but barriers preventing to combine them" [57]. Open data in this field would undoubtedly drive new discoveries.

Example #3 – Research utilizing text and data from journal publications

The copyright status of data obtained though text-mining is debatable. The numerical instances of a particular gene or protein name in a full-text corpus of articles could be valuable for secondary research and, in the US at least, are likely to be considered (non-copyrightable) facts. Some scholars take the position that mining does not violate copyright law because it does not meet the statutory definition of copying which requires “fixing” the work in a permanent form [15]. Yet text mining is often restricted by commercial publishers. In the study of small angle scattering (a “technique based on the deflection of a beam of particles, or an electromagnetic or acoustic wave, away from the straight trajectory after it interacts with structures that are much larger than the wavelength of the radiation” according to Wikipedia [58]), a researcher might be interested to harvest the data used in other publications to test their analysis tools and provide better teaching aids. Generally, in this area of research, the data are only presented as a graph, the data analysis is not spelt out and there is no specific license attached if the data are available (Cameron Neylon, personal communication).

Example #4 – Open bibliography

Online scientific publishing has driven a diversification of measures of research and researcher impact, extending the focus from journal impact factors to article and individually-led metrics. Bibliographic information (rather than copyright attribution), which enables identification of scholarly work and tracks citations to scientists’ work, is central to earning of academic credit for concepts and ideas. Many services are now available which enable individual authors to calculate their citation index, known as the Hirsch or h-index. Examples include Scopus, Thomson ISI, Google Scholar and Microsoft Academic Search. However, much of the data underlying these metrics is not available openly, leading to multiple scores for the same individual or paper – depending on the tool or service used, which have different corpuses and different algorithms for calculating impact scores. A common, open bibliography, as has been established by some leading libraries would enable anyone to assess, utilize and build applications based on the data [59]. And furthermore, from a researcher’s perspective this approach is far more efficient, negating the need to maintain and report multiple sources of data from multiple impact-measuring tools. As outlined by Jones *et al.* [49] the motivations for and opportunities for open bibliographies are many. The negative implications of open bibliography for an author of a paper are negligible. Under the license agreement proposed in this article CC0 would apply to the article title, author names and information, unique identifying and publishers’ information, and reference list. Given a primary use of bibliographic information is to track scholarly citation activity, authors could reasonably expect these open data to increase the visibility and impact of their work.

Example #5 – Reproduction/validation of results for teaching and further research

In September 2010 Tommi Nyman and colleagues published an article in *BMC Evolutionary Biology*, “How common is ecological speciation in plant-feeding insects? A 'Higher' Nematinae perspective” [60]. The article included, in addition to the sequence data used to reconstruct the phylogenetic trees, the background data used in the phylogeny-based ecological analyses as additional file 1 – an Excel file. The data are well labeled and readily understandable by other scientists and fully document how they sampled their insects. This informative approach means, for example, readers would not need work through the references to discover the sampling used. These data have potential usage for reproduction and validation of the article’s findings, for teaching purposes, and conceivably uses involving the processing and integration of the data using computer software. Explicit dedication of these data to the public domain minimizes barriers to these scientifically important activities and maximizes the reuse potential of the data, as we could be more confident that all future uses of the data will not be impeded by licensing restrictions.

Concluding remarks – and what next?

Legal issues present substantial barriers, in theory and reality, to the reuse and integration of research data which are free to access online, and data published in peer-reviewed journals. The implementation of a new license and waiver agreement, as per the protocol described in this article, in BioMed Central journals and in the future by other open access publishers should help further realize the benefits of open data for the scientific community – and beyond. We invite all our readers and authors to consider and comment on the implications of the proposed change to BioMed Central’s license agreement set out in this article.

Competing interests

MC is Managing Director of BioMed Central, part of Springer Science + Business Media. IH is an employee of BioMed Central.

Authors' contributions

The idea for this manuscript was conceived at a meeting (reported on the BioMed Central Blog, [13]) which was attended by MC and a number of those listed in the acknowledgements section of this article. The meeting was chaired by IH. IH wrote the first draft of the manuscript. MC was involved in critical review and editing of the original manuscript, and revised versions thereof, in response to comments received from the acknowledged contributors. Both authors read and approved the final manuscript.

Acknowledgements

This article builds on some of the concepts described in BioMed Central’s draft open data statement, and as such a small number of passages from it have been included verbatim in this article.
Thanks to Daniel Mietchen and Cameron Neylon for contributing practical examples of open data in life sciences for this paper.
For their comments on an earlier draft of this paper, thanks to: Victoria Stodden (Columbia University), Diane Cabell (Oxford Internet Institute), Ruth Wilson (Nature Publishing Group), Daniel Mietchen (EvoMRI Communications), Cameron Neylon (Public Library of Science, PLoS), Heather Piwowar (University of British Columbia), Susanna Sansone (University of Oxford), John Wilbanks (Ewing Marion Kauffman Foundation).

Received: 11 June 2012 Accepted: 4 September 2012
Published: 7 September 2012

References

1. Fry J, Lockyer S, Oppenheim C, Houghton J, Rasmussen B: **Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes.** 2009, 1–89. [http://hdl.handle.net/2134/4600].
2. Wood J: **Riding the wave: How Europe can gain from the rising tide of scientific data.** [http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf]. Accessed 11th June 2012.
3. Walport M, Brest P: **Sharing research data to improve public health.** *Lancet* 2011, **377:**537–539.
4. **Open Knowledge Definition - Defining the Open in Open Data, Open Content and Open Information** [http://opendefinition.org/okd/]. Accessed 11th June 2012.
5. Hrynaszkiewicz I: **BioMed Central's position statement on open data.** [http://blogs.openaccesscentral.com/blogs/bmcblog/resource/opendatastatementdraft.pdf]. Accessed 11th June 2012.
6. **Panton Principles** [http://pantonprinciples.org/]. Accessed 11th June 2012.
7. Hargreaves I: **Digital Opportunity: A Review of Intellectual Property and Growth.** [http://www.ipo.gov.uk/ipreview-finalreport.pdf]. Accessed 11th June 2012.
8. McDonald D, Kelly U: **The Value and Benefits of Text Mining.** [http://www.jisc.ac.uk/media/documents/publications/reports/2012/value-text-mining.pdf]. Accessed 11th June 2012.
9. Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, Einhorn D, Tocchini-Valentini G, Hrabe AM, Rosenthal N: **Post-publication sharing of data and tools.** *Nature* 2009, **461:**171.
10. Conway PH, VanLare JM: **Improving access to health care data: the open government strategy.** *JAMA* 2010, **304:**1007–1008.
11. Contreras J: **Prepublication data release, latency, and genome commons.** *Science* 2010, **329:**393–394.
12. Gøtzsche PC: **Why we need easy access to all data from all clinical trials and how to accomplish it.** *Trials* 2011, **12:**249.
13. BioMed Central Blog: **Report from the Publishing Open Data Working Group meeting, 17th June 2011.** [http://blogs.openaccesscentral.com/blogs/bmcblog/entry/report_from_the_publishing_open]. Accessed 11th June 2012.
14. **Budapest Open Access Initiative** [http://www.soros.org/openaccess/read]. Accessed 11th June 2012.
15. Carroll MW: **Why full open access matters.** *PLoS Biol* 2011, **9:**e1001210.
16. Lyubomir P, Daniel M, Vishwas C, Gregor H, David R, Vincent S, David S: **Data Publishing Policies and Guidelines for Biodiversity Data.** [http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf]. Accessed 11th June 2012.
17. **Using BioMed Central's open access full-text corpus for text mining research** [http://www.biomedcentral.com/about/datamining]. Accessed 11th June 2012.
18. STM, ALPSP: **Databases, data sets, and data accessibility – views and practices of scholarly publishers.** [http://www.stm-assoc.org/2006_06_01_STM_ ALPSP_Data_Statement.pdf]. Accessed 11th June 2012.
19. Swan A, Brown S: **To Share or not to Share: Publication and Quality Assurance of Research Data Outputs.** [http://eprints.soton.ac.uk/266742/]. Accessed 11th June 2012.
20. Jones RH: **Is there a property interest in scientific research data?** *High Technology Law Journal* 1987. [http://www.law.berkeley.edu/journals/btlj/articles/vol1/jones.html]. Accessed 11th June 2012.
21. **Guidelines - Ownership, copyright and intellectual property, Monash University Research Data Management** [http://www.researchdata.monash.edu/guidelines/ownership.html]. Accessed 11th June 2012.
22. **Guide to Open Data Licensing** [http://opendefinition.org/guide/data/]. Accessed 11th June 2012.
23. Ball C: **How to License Research Data | Digital Curation Centre.** [http://www.dcc.ac.uk/resources/how-guides/license-research-data]. Accessed 11th June 2012.
24. **Conformant Licenses.** [http://opendefinition.org/licenses/]. Accessed 11th June 2012.
25. Dalgleish R, Molero E, Kidd R, Jansen M, Past D, Robl A, Mons B, Diaz C, Mons A, Brookes AJ: **Solving bottlenecks in data sharing in the life sciences.** *Hum Mutat* 2012, doi:10.1002/humu.22123.
26. **CC0 use for data - CC Wiki** [http://wiki.creativecommons.org/CC0_use_for_data]. Accessed 11th June 2012.
27. Butler D: **GlaxoSmithKline goes public with malaria data.** *Nature* 2010, doi:10.1038/news.2010.20.
28. **Cologne-based libraries release 5.4 million bibliographic records via CC0 - Creative Commons** [http://creativecommons.org/weblog/entry/21344]. Accessed 11th June 2012.
29. Schaeffer P: **Why does Dryad use CC0?** [http://blog.datadryad.org/2011/10/05/why-does-dryad-use-cc0/]. Accessed 11th June 2012.
30. Science Commons: **Protocol for Implementing Open Access Data.** [http://sciencecommons.org/projects/publishing/open-access-data-protocol/]. Accessed 11th June 2012.
31. Hrynaszkiewicz I: **LabArchives and BioMed Central: a new platform for publishing scientific data.** [http://blogs.openaccesscentral.com/blogs/bmcblog/entry/labarchives_and_biomed_central_a]. Accessed 11th June 2012.
32. Wilbanks J: **Attribution v. Citation.** [http://scienceblogs.com/commonknowledge/2009/06/attribution_v_citation.php]. Accessed 11th June 2012.
33. Rees J: **Recommendations for independent scholarly publication of data sets.** [http://neurocommons.org/report/data-publication.pdf]. Accessed 11th June 2012.
34. Piwowar HA, Day RS, Fridsma DB: **Sharing detailed research data is associated with increased citation rate.** *PLoS ONE* 2007, **2**(3):e308.
35. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, Noort VV: **Repeatability of published microarray gene expression analyses.** *Nat Genet* 2009, **41:**149–155.
36. Pienta AM, Alter GC, Lyle JA: **The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data.** [http://deepblue.lib.umich.edu/bitstream/2027.42/78307/1/pienta_alter_lyle_100331.pdf]. Accessed 11th June 2012.
37. Hrynaszkiewicz I: **Citing and linking data to publications: more journals, more examples. . .more impact?** [http://blogs.openaccesscentral.com/blogs/bmcblog/entry/citing_and_linking_data_to]. Accessed 11th June 2012.
38. Hrynaszkiewicz I: **The need and drive for open data in biomedical publishing.** *Serials* 2011, **24:**31–37.
39. Tapscott D, Williams AD: **Wikinomics: How mass collaboration changes everything.** *Portfolio*; 2008.
40. Derry JM, Mangravite LM, Suver C, Furia M, Henderson D, Schildwachter X, Izant J, Sieberts SK, Kellen MR, Friend SH: **Developing predictive molecular maps of human disease through community-based modeling.** *Nature Precedings* 2011, doi:10.1038/npre.2011.5883.1.
41. Errami M, Garner H: **A tale of two citations.** *Nature* 2008, **451:**397–399.
42. Panton Principles: **FAQ.** [http://pantonprinciples.org/faq/]. Accessed 11th June 2012.
43. Vickers A: **Whose data set is it anyway? Sharing raw data from randomized trials.** *Trials* 2006, **7:**15.
44. Sandercock PAG, Niewada M, Członkowska A: **The International stroke trial database.** *Trials* 2011, **12:**101.
45. **Wikipedia definition of "Data"** [http://en.wikipedia.org/w/index.php?title=Data&oldid=506274317]. Accessed 11th June 2012.
46. Roberts D, Moritz T: **A framework for publishing primary biodiversity data.** *BMC Bioinformatics* 2011, **12:**11.
47. Wallis R: **Linked Open Data and Pavlova.** [http://blogs.talis.com/nodalities/2010/08/the-linked-open-data-and-pavlova.php]. Accessed 11th June 2012.
48. **Extensible Markup Language (XML) 1.0 (Fifth Edition)** [http://www.w3.org/TR/REC-xml/]. Accessed 11th June 2012.
49. Jones R, Macgillivray M, Murray-Rust P, Pitman J, Sefton P, O'Steen B, Waites W: **Open bibliography for science, technology, and medicine.** *Journal of Cheminformatics* 2011, **3:**47.
50. Berners-Lee T: **Linked Data.** [http://www.w3.org/DesignIssues/LinkedData.html]. Accessed 11th June 2012.
51. Willighagen EL, Brändle MP: **Resource description framework technologies in chemistry.** *Journal of Cheminformatics* 2011, **3:**15.
52. **Open Source Licenses | Open Source Initiative** [http://www.opensource.org/licenses/index.html]. Accessed 11th June 2012.
53. Hrynaszkiewicz I, Cockerill M: **In defence of supplemental data files: don't throw the baby out with the bathwater.** [http://blogs.openaccesscentral.com/blogs/bmcblog/entry/in_defence_of_supplemental_data]. Accessed 11th June 2012.

54. **4.0/Draft 1 - CC Wiki** [http://wiki.creativecommons.org/4.0/Draft_1]. Accessed 11th June 2012.
55. Thaney K: **Sharing Data on the Web**. [http://blogs.talis.com/nodalities/2010/02/sharing-data-on-the-web.php]. Accessed 11th June 2012.
56. **BioMed Central copyright and license agreement:** [http://www.biomedcentral.com/about/license]. Accessed 11th June 2012.
57. Mietchen D, Gaser C: **Computational morphometry for detecting changes in brain structure due to development, aging, learning, disease and evolution.** *Frontiers in Neuroinformatics* 2009, **3**:25.
58. **Wikipedia definition of "Small angle scattering"** [http://en.wikipedia.org/w/index.php?title=Small-angle_scattering&oldid=506086467]. Accessed 11th June 2012.
59. Fenner M: **Google Scholar Citations, Researcher Profiles, and why we need an Open Bibliography**. [http://blogs.plos.org/mfenner/2011/07/27/google-scholar-citations-researcher-profiles-and-why-we-need-an-open-bibliography/]. Accessed 11th June 2012.
60. Nyman T, Vikberg V, Smith DR, Boevé J-L: **How common is ecological speciation in plant-feeding insects? a "higher" nematinae perspective.** *BMC Evol Biol* 2010, **10**:266.

doi:10.1186/1756-0500-5-494
Cite this article as: Hrynaszkiewicz and Cockerill: **Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals.** *BMC Research Notes* 2012 **5**:494.

- ⌵ News and events
- Careers
- ⌵ Contact us

License agreement

In submitting an article to any of the journals published by BMC I certify that;

1. I am authorized by my co-authors to enter into these arrangements.
2. I warrant, on behalf of myself and my co-authors, that:
    - the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights;
    - I am/we are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BMC are not in breach of any other obligation;
    - the article contains nothing that is unlawful, libellous, or which would, if published, constitute a breach of contract or of confidence or of commitment given to secrecy;
    - I/we have taken due care to ensure the integrity of the article. To my/our - and currentlv accepted scientific - knowledge all statements contained in it purporting to be facts are true and any formula or instruction contained in the article will not, if followed accurately, cause any injury, illness or damage to the user.
3. I, and all co-authors, agree that the article, if editorially accepted for publication, shall be licensed under the Creative Commons Attribution License 4.0. In line with BMC's Open Data Policy, data included in the article shall be made available under the Creative Commons 1.0 Public Domain Dedication waiver, unless otherwise stated. If the law requires that the article be published in the public domain, I/we will notify BMC at the time of submission, and in such cases not only the data but also the article shall be released under the Creative Commons 1.0 Public Domain Dedication waiver. For the avoidance of doubt it is stated that sections 1 and 2 of this license agreement shall apply and prevail regardless of whether the article is published under Creative Commons Attribution License 4.0 or the Creative Commons 1.0 Public Domain Dedication waiver.

[End of BMC’s license agreement]

Explanatory notes regarding BMC’s license agreement

As an aid to our authors, the following paragraphs provide some brief explanations concerning the Creative Commons licenses that apply to the articles published in BMC-published journals and the rationale for why we have chosen these licenses.

The Creative Commons Attribution License (CC BY), of which CC BY 4.0 is the most recent version, was developed to facilitate open access as defined in the founding documents of the movement, such as the 2003 Berlin Declaration. Open access content has to be freely available online, and through licensing their work under CC BY authors grant users the right to unrestricted dissemination and re-use of the work, with only the one proviso that proper attribution is given to authors. This liberal licensing is best suited to facilitate the transfer and growth of scientific knowledge. The Open Access Scholarly Publishers Association (OASPA) therefore strongly recommends the use of CC BY for the open access publication of research literature, and many research funders worldwide either recommend or mandate that research they have supported be published under CC BY. Examples for such policies include funders as diverse as the Wellcome Trust, the Australian Governments, the European Commission's Horizon 2020 framework programme, or the Bill & Melinda Gates Foundation.

The default use of the Creative Commons 1.0 Public Domain Dedication waiver (CC0 or CC zero) for data published within articles follows the same logic, facilitating maximum benefit and the widest possible re-use of knowledge. It is also the case that in some jurisdictions copyright does not apply to data. CC0 waives all potential copyrights, to the extent legally possible, as well as the attribution requirement. The waiver applies to data, not to the presentation of data. If, for instance, a table or figure displaying research data is reproduced, CC BY and the requirement to attribute applies. Increasingly, however, new insights are possible through the use of big data techniques, such as data mining, that harness the entire corpus of digital data. In such cases attribution is often technically infeasible due to the sheer mass of the data mined, making CC0 the most suitable licensing tool for research outputs generated from such innovative techniques.

It is important to differentiate between legal requirements and community norms. It is first and foremost a community norm, not a law, that within the scientific community attribution mostly takes the form of citation. It is also a community norm that researchers are expected to refer to their sources, which usually takes the form of citation. Across all cases of research reuse (including data, code, etc), community norms will apply as is appropriate for the situation: researchers will cite their sources where it is feasible, regardless of the applicable license. CC0 therefore covers those instances that lie beyond long-established community norms. The overall effect, then, of CC0 for data is to enable further use, without any loss of citations. For further explanation, we recommend you refer to our Open Data page.

In the following, we provide the licenses' summaries as they can be found on the Creative Commons website.