Comparing citation numbers between articles at two stages of a Model Organism Database curation workflow

Michael Lauruhn
Elsevier Labs
3214 NE 25th Ave
Portland, OR 97212 U.S.A.
m.lauruhn@elsevier.com

Gillian Millburn
FlyBase
Department of Physiology, Development and Neuroscience, University of Cambridge
Cambridge CB2 3DY, United Kingdom
gm119@cam.ac.uk

ABSTRACT

Model organism databases (MODs) facilitate the connections between published research papers with genes and other biological information. MODs aim to make research data easier to access to the research community, especially for researchers relying on genetic data and other information about a specific species. This paper follows previous research (Beradini, et al. 2016) that attempted to use quantitative data to determine if and how literature curated by a MOD makes a difference to the access and reuse of the curated data. The research addresses whether articles that have been through the detailed curation process of a MOD are more likely to be cited when compared to 'similar' articles that are not curated. For this research, citations for articles curated by FlyBase, a MOD for genetic and molecular data for the Drosophilidae insect family, were compared with articles identified as having similar genetic and molecular data, but not yet given a detailed curation by FlyBase. In addition, citation counts from a larger set of articles retrieved through a title and keyword search for Drosophilidae are also compared.

Introduction

There is increasing awareness about issues related to reproducible research. Researchers are encouraged to be intentional about being better stewards of their research data. In scientific literature, there are multiple initiatives to annotate papers and other research outputs with references to resources such as antibodies and model organisms. Model organism databases (MODs) contain genetic and molecular data about a species with the intent of learning about other organisms. They are widely used for curating academic journal articles across the biological domain. This study is a second to use citation counts in an attempt to assess the accessibility and discoverability of articles that have been curated by a MOD when compared to "similar" articles that have not been curated. As the curation processes and workflows vary at each MOD, it is not feasible to create exact comparisons between them. Instead, data collection was conducted in a manner that reflected the resources available from the MOD.

FlyBase

FlyBase began in 1992, and provides an online repository of biological information and data about *Drosophila* (fly) species with a focus on *Drosophila melanogaster,* the common fruit fly and model organism species. Much of the data in FlyBase is from research literature and is curated manually (Marygold, et al 2016). Some of the data-types that are in FlyBase are: “sequence-level gene models, molecular classification of gene product functions, mutant phenotypes, mutant lesions and chromosome aberrations, gene expression patterns, transgene insertions” (Drysdale, et al 2008).

The FlyBase curation process begins with a weekly search of the PubMed database to identify any new publications that are related to *Drosophila*. Citation information is collected for the FlyBase bibliography and authors are contacted with an invitation for them to "fast track" their article for inclusion in FlyBase. Next, authors or FlyBase curators will make links between publication and genes that will appear on the FlyBase website (Marygold, et al 2013, Bunt et al., 2012). The data-types mentioned above are flagged and the articles are placed in a queue for detailed curation. During detailed curation, FlyBase staff will "extract detailed genetic and molecular information" (McQuilton, et al 2012).

Methodology

The first step in collecting data was to identify sets of similar articles, one set that had been curated in FlyBase and another that has not. The first subset of FlyBase articles are papers that received detailed curation for one or more of the following 'genetic' data-types: descriptions of new alleles, descriptions of new transgenes, phenotypic information, or GO (Gene Ontology) curation. (These articles did not have detailed curation of 'molecular' data-types such as expression or physical interaction data.) The dataset was limited to a publication date between 2008 and 2016 inclusive, since flagging of data-types for this date range by authors or FlyBase curators is complete. There were 4,882 articles in this set on the date collected. The second subset of FlyBase articles are papers that were identified by data-type for one or more of the genetic data-types from the above subset, however they had not received a detailed curation for any of the data-types. There were 3,392 articles in this set. Both of these subsets were collected in April 2017. In addition, the top ten most prevalent Journal titles that appear in FlyBase-curated articles were identified by their ISSN numbers and both of the subsets were split into additional sets as to whether the article appeared in one of the prevalent titles or not.

As an additional comparison, a search the Scopus bibliographic database was conducted. The search parameters were for Articles, with the search string "Drosophil*" in the title, abstract, or keyword fields, with the publication dates between 2008 and 2016. This search was intended to bring in a set of articles that mentioned the Drosophilidae insect family, though most would not contain gene or molecular information that would make them candidates to be curated in detail by FlyBase. These articles would serve as a baseline for comparison with the two FlyBase subsets. There were 31,066 articles in this result set when it was collected in April 2017.

For the FlyBase subsets, a list of PubMed Identifiers (PMIDs) for the articles was used to query Scopus to retrieve Scopus identification numbers that were then used to query Scopus for citation counts. Once all the citation data had been collected, duplicate records from the FlyBase subsets and the Scopus search set were put together as one large set. This complete article set had 36,827 unique articles.

Findings

Within each set, the number of times articles were cited was counted and calculated for both the average and median numbers (Figure 1; Table 1).

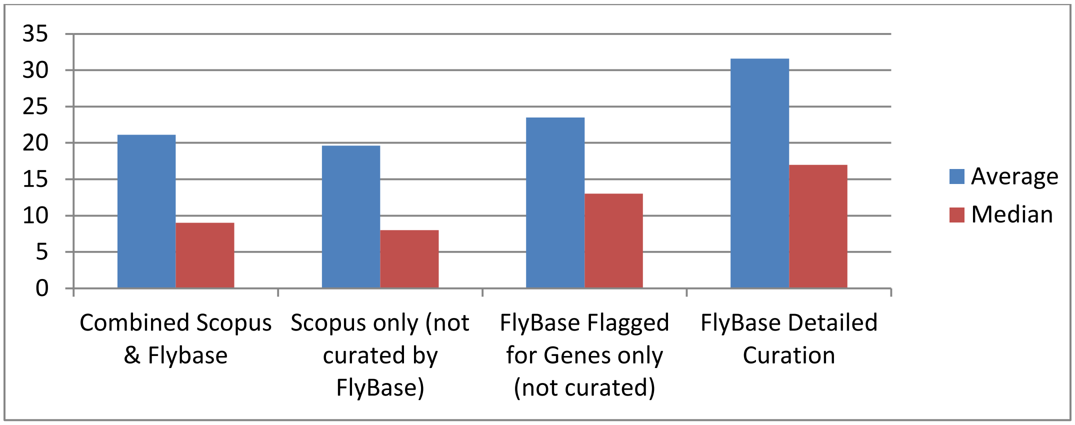

Figure 1. Average and Median Citation Counts for Scopus and FlyBase-flagged articles

Citation Counts	Average	Median
Combined Scopus & FlyBase	21.1	9
Scopus only (not curated by FlyBase)	19.6	8
FlyBase Flagged for Genes only (not curated)	23.5	13
FlyBase Detailed Curation	31.6	17

Table 2. Average and Median Citation Counts for Scopus and FlyBase-flagged articles

The combined Scopus Search and FlyBase collection of articles were cited an average of 21.1 times with a median of 9. When the articles that FlyBase identified as having gene data in them are removed from the Scopus search collection, the citation count is slightly lower, with the Scopus-only collection being cited 19.6 times with a median at 8 citations.

For the FlyBase-flagged sets, the collection of articles that were Flagged for Genes, but not given a detailed curation, the average citation count was 23.5 and the median was 13. Finally, the articles that were flagged for Gene data *and* have detailed curation from FlyBase have the most citations from within the subsets, with an average of 31.6 citations and a median of 17. The increase in average citations from flagged-only articles to curated articles was 34%. For each set, it is worth noting many of the articles had zero citations as part of a significant long tail of articles with two or fewer citations. This is consistent with findings and discussions about the volume of published articles that are yet to be cited (phys.org; Eveleth, 2014).

When comparing the additional subsets of FlyBase-flagged articles, similar citation count behavior to the sets above can be observed when looking at whether or not articles are only flagged for Genes or have had detailed curation.

Articles from the top ten most occurring journal titles that were only flagged for Gene data had an average of 23 citations, while similar articles that had undergone the Detailed Curation had an average of 28.1 Citations – an increase of 22%. However, Articles from the from *outside* the top ten most occurring journal titles that were only flagged for Gene data had an average of 23.9 citations, but articles that had undergone the Detailed Curation had an average of 34.1 Citations – an increase of almost 43% (Figure 2; Table 2)

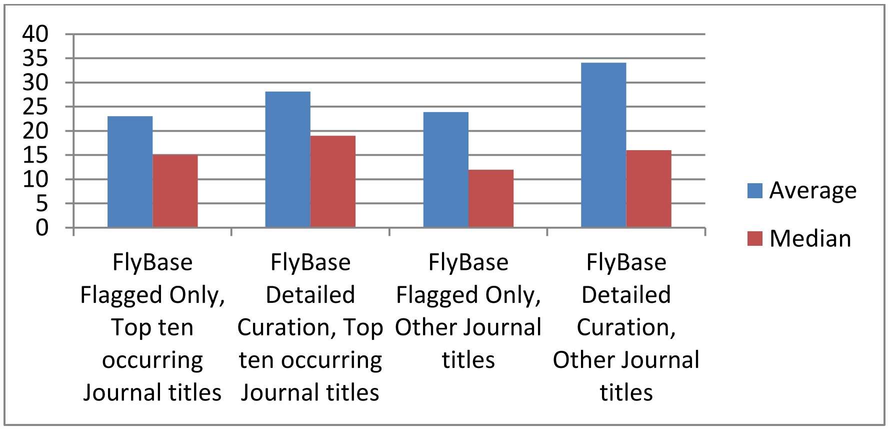

Figure 3 Average and Median citation counts comparing articles that appear in the most frequently occurring journal titles and those outside the frequently occurring titles.

Citation Counts	Average	Median
FlyBase Flagged Only, Top ten occurring Journal titles	23	15
FlyBase Detailed Curation, Top ten occurring Journal titles	28.1	19
FlyBase Flagged Only, Other Journal titles	23.9	12
FlyBase Detailed Curation, Other occurring Journal titles	34.1	16

Table 4 Average and Median citation counts comparing articles that appear in the most frequently occurring journal titles and those outside the frequently occurring titles.

Discussion

The results of this research show that articles that underwent a detailed curation from FlyBase have been cited more frequently than articles that have not had a detailed curation from FlyBase. These findings echo the findings from the TAIR research on citation analysis of Arabidopsis literature. The results of that work also showed an increase in citation counts going upward from a general population of Scopus articles mentioning Arabidopsis, articles identified by TAIR as being valid but not annotated and articles annotated in TAIR. In regards to those results, it was noted that even though “data show that articles used for annotations have more citations than those that are not annotated, it is not possible to state conclusively that there is a direct correlation between annotation and citation rates. This is due to the many challenges in collecting data for this experiment, identifying control data sets, and selection bias" (Bernardini, et al 2016).

Similarly, some amount selection bias will likely occur within any workflow process that involves human curation. Articles can be or selected for a number reasons based on a curator's expertise or time constraints. Likewise, it is not practical to set up a true control group of scientific literature presumably by removing a selection of articles from the workflow for a specified period of time.

Thus said, the consistent increase in citations in articles that are curated by FlyBase as well as TAIR is worth noting. FlyBase provides an access point to a specific community of researchers. They have a very specific information need with a significant likelihood of citing the resource they are using. FlyBase also provides tools like the ability to download RIS records that make citing the resources easier. It is conceivable that an article being curated by FlyBase would increase its likelihood to being cited.

As stated in previous research, future research in this area should account for other bias factors with an emphasis as to whether articles appearing in open access publications have different citation behavior than those which require purchase or a subscription. Another aspect worth investigating might be web traffic and user activity on the MOD portals.

Acknowledgments

We would also like to thank members of Elsevier Labs and the Elsevier Scopus team for technical assistance.

Funding

GM is supported by the FlyBase NIH/NHGRI grant U41HG000739 (N. Perrimon, Harvard University, PI; N.H. Brown, University of Cambridge, coPI).

References

Berardini, T., Daniel, R., Lauruhn, M., Reiser, L. (2016). Preliminary Study on the Impact of Literature Curation in a Model Organism Database on Article Citation Rates. D-Lib Magazine vol. 22 doi:10.1045/september2016-berardini

Bunt, S.M., Grumbling, G.B., Field, H.I., Marygold S.J., Brown, N.H., Millburn, G.H. and the FlyBase Consortium (2012) Directly e-mailing authors of newly published papers encourages community curation. Database (Oxford) 2012; bas024. doi:10.1093/database/bas024

Drysdale, R. and the FlyBase Consortium (2008). FlyBase: a database for the Drosophila research community. *Methods Mol. Biol*., 420 (2008), pp. 45–59 https://link.springer.com/protocol/10.1007%2F978-1-59745-583-1_3

Eveleth, Rose (2014), Academics Write Papers Arguing Over How Many People Read (And Cite) Their Papers. Smithsonian.com http://www.smithsonianmag.com/smart-news/half-academic-studies-are-never-read-more-three-people-180950222/ [Accessed April 2017]

Marygold, S.J., Crosby, M.A., Goodman, J.L. and the FlyBase Consortium. (2016). Using FlyBase, a Database of *Drosophila* Genes & Genomes. *Methods in Molecular Biology (Clifton, N.J.)*, *1478*, 1–31. http://doi.org/10.1007/978-1-4939-6371-3_1

Marygold, S.J., Leyland, P.C., Seal, R.L., Goodman, J.L., Thurmond, J., Strelets, V.B., Wilson, R.J.; FlyBase consortium (2013) FlyBase: improvements to the bibliography. Nucleic Acids Res. 41(Database issue):D751-D757 http://dx.doi.org/10.1093/nar/gks1024

McQuilton, P. and the FlyBase Consortium; Opportunities for text mining in the FlyBase genetic literature curation workflow. Database (Oxford) 2012; 2012 bas039. doi: 10.1093/database/bas039

Phys.org (2017), "Increasing visibility and enhancing impact of research". Phys.org https://phys.org/news/2017-04-visibility-impact.html [Accessed April 2017]