Check for updates

SOFTWARE TOOL ARTICLE

REVISED Social Network: a Cytoscape app for visualizing co-authorship networks

[version 2; peer review: 2 approved with reservations]

Victor Kofia[1], Ruth Isserlin[1], Alison M.J. Buchan[2], Gary D. Bader[1]

[1]The Donnelly Centre, University of Toronto, Toronto, ON, M5S 1A8, Canada
[2]Faculty of Medicine, University of Toronto, Toronto, ON, M5S 1A8, Canada

v2 **First published:** 05 Aug 2015, **4**:481
https://doi.org/10.12688/f1000research.6804.1
Second version: 08 Oct 2015, **4**:481
https://doi.org/10.12688/f1000research.6804.2
Latest published: 23 Dec 2015, **4**:481
https://doi.org/10.12688/f1000research.6804.3

Abstract

Networks that represent connections between individuals can be valuable analytic tools. The Social Network Cytoscape app is capable of creating a visual summary of connected individuals automatically. It does this by representing relationships as networks where each node denotes an individual and an edge linking two individuals represents a connection. The app focuses on creating visual summaries of individuals connected by co-authorship links in academia, created from bibliographic databases like PubMed, Scopus and InCites. The resulting co-authorship networks can be visualized and analyzed to better understand collaborative research networks or to communicate the extent of collaboration and publication productivity among a group of researchers, like in a grant application or departmental review report. It can also be useful as a research tool to identify important research topics, researchers and papers in a subject area.

Keywords

social network analysis , data visualization , collaborative networks , co-publication networks

This article is included in the Cytoscape gateway.

This article is included in the Bioinformatics gateway.

Open Peer Review

Approval Status ✔ ? ?

	1	2	3
version 3 (revision) 23 Dec 2015	✔ view		? view
version 2 (revision) 08 Oct 2015	? view		
version 1 05 Aug 2015	? view	? view	

1. **Michael Bales**, Weill Cornell Medical College, New York, USA
   Terrie Wheeler, Weill Cornell Medical College, New York, USA
2. **Jiang Bian**, University of Florida, Gainesville, USA
3. **Shahadat Uddin**,

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Gary D. Bader (gary.bader@utoronto.ca)

Competing interests: No competing interests were disclosed.

Grant information: Financial support was provided by the Faculty of Medicine at the University of Toronto.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2015 Kofia V *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kofia V, Isserlin R, Buchan AMJ and Bader GD. **Social Network: a Cytoscape app for visualizing co-authorship networks [version 2; peer review: 2 approved with reservations]** F1000Research 2015, **4**:481
https://doi.org/10.12688/f1000research.6804.2

First published: 05 Aug 2015, **4**:481 https://doi.org/10.12688/f1000research.6804.1

REVISED **Amendments from Version 1**

IMPLEMENTATION

- An incorrect citation was fixed.
- Revised Figure 1 to contain the full user panel (network summary panel is located in bottom half). Mentioned how users can specify time intervals and visualize co-authorships as bar charts.
- Mentioned use of POST calls to query PubMed.

OTHER

- Switched to using the term "co-authorship" instead of "co-publication" because it is more common in the literature.

See referee reports

Introduction

A scientist's research output and collaborative tendencies - at least those that can be measured based on publications - can be visually summarized as a network where each node denotes an author and edges link authors who have co-published. Such a network facilitates determining who publishes with whom and in what topics, and identifying key individuals and organizations within collaborative research networks. It is useful to create and visualize a network showing a broad overview of collaborative research publications to communicate the extent of collaboration and impact of publications. As another example, creating a collaboration network from a set of publications for a specific topic, for example "Alzheimer's", can help highlight experts in the field and could be useful as a research tool to help identify important topics, researchers and papers.

Previously, creating co-authorship networks required users to manually retrieve the relevant data and transform it into either a formatted text file or an excel workbook that defined all the individual nodes and connections. Users would then have to import the text file or workbook into Cytoscape or another network visualization tool. To streamline this workflow, we developed the Social Network app, a Cytoscape 3 app that is capable of automatically generating visual summaries of individuals connected in academia. In the simplest mode of interaction, the user supplies the first initial and last name of the individual whose network they would like to visualize, and a co-authorship network is generated automatically from one of three currently supported bibliographic databases: PubMed, Scopus and Web of Science (via InCites). Users can also provide more complex and larger sets of publications, for example using the PubMed query system.

Methods and implementation

User interface

The Social Network App supports both text and file-based inputs. As Figure 1(1–4) shows, co-authorship networks can be created in four ways. With the search box, users can run queries against PubMed. A co-authorship network is automatically generated from any results that are retrieved. Alternatively, users can go directly

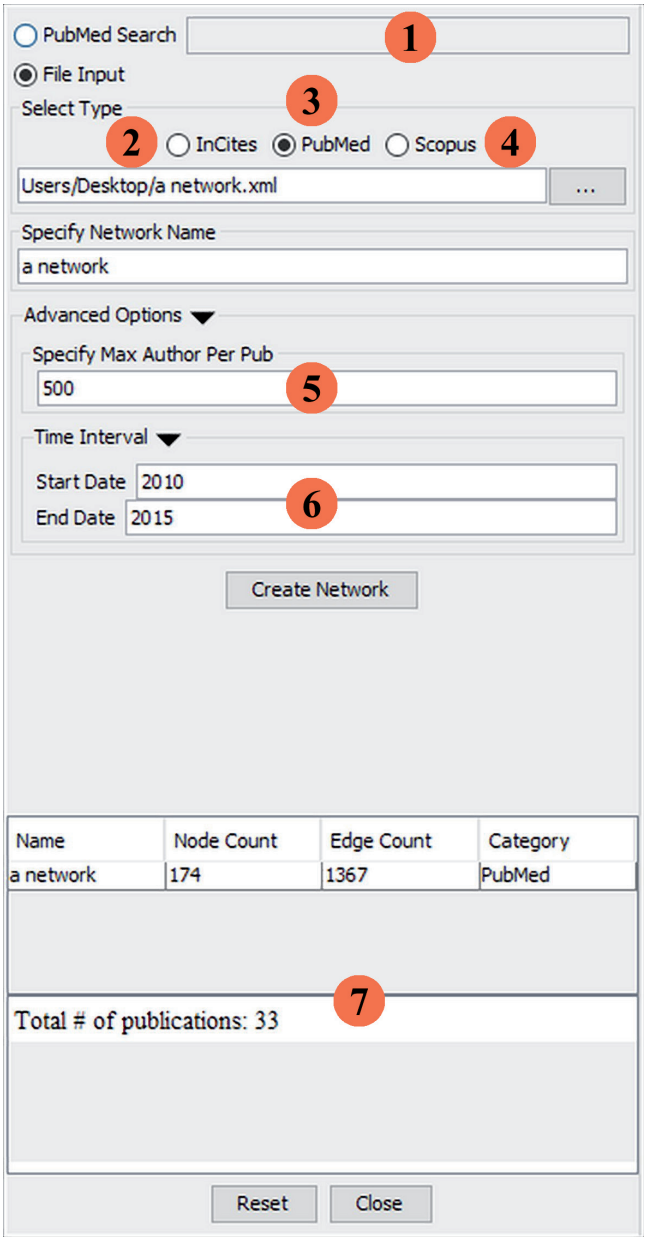


Figure 1. Snapshot of the Social Network user interface. Co-authorship networks can be generated in four ways: (1) By entering a query into the PubMed search box. (2) Loading an InCites report (XLSX format). (3) Loading a PubMed XML file containing query results retrieved from the PubMed web interface. (4) Loading a Scopus CSV file. (5) Users can also set the maximum # of authors allowed for a publication to filter out very large author lists that may clutter the network. (6) Users can specify a time interval for the co-authorships, which can optionally be visualized as bar charts on each author node showing number of publications plotted against publication year. (7) Extra information associated with each network is displayed in the network summary panel.

to PubMed, Scopus or InCites web sites, search for publications, export them to a specified file format and visualize them using the app. See the user guide for detailed instructions on how to do each of these tasks (http://baderlab.org/UserguideSocialNetworkApp);

Often when running queries with common names (e.g. "Smith J"), the query returns publications that contain more than 500 authors. Visualizing networks containing these types of publications is challenging because generating $n \cdot (n - 1)/2$ edges for each publication (n refers to the # of authors in the publication) is resource intensive and the large clusters created are difficult to visualize and interpret. To avoid this issue, the user panel includes a maximum author per publication field that allows users to specify the author number threshold at which publications are excluded. By default the threshold is set to 500.

A co-authorship network summary panel is also included in the user panel (Figure 1(7)). For PubMed and Scopus networks, the panel displays the total number of publications parsed and the number of excluded publications (publications are excluded if the number of authors they have exceeds the threshold). For direct PubMed queries the panel also includes the query translation automatically performed by PubMed. Scopus and InCites networks contain institutional affiliations for all the authors of a given publication. For InCites networks, charts that summarize the total number of publications and citations by location can be viewed by clicking on links in the panel that navigate to summary charts created with the Google Chart API (https://developers.google.com/chart/).

Implementation

Social Network is written in the Java programming language as an app for Cytoscape 3[1] and is based on the Cytoscape 3-supported OSGi (Open Services Gateway Initiative) software architecture. To facilitate development, we developed a set of coding guidelines and defined them as Eclipse templates. The Eclipse templates and instructions on how to import them into an existing workspace are available at (http://baderlab.org/Software/SocialNetworkApp/Development). We also used the Maven project management tool (https://maven.apache.org/) to retrieve and organize the dependencies required by the app. An outline of the required dependencies is provided in a pom file that is located in the project source code (https://github.com/BaderLab/SocialNetworkApp).

The design of the app followed Object Oriented Principles (OOP), reflected in the following class hierarchy; To facilitate future support for other social network types, we defined a set of flexible data structures, namely: AbstractNode, AbstractEdge and SocialNetwork. These data structures are used to represent networks and conveniently associated data created by the app in a general form. They are also specialized for each network type (e.g. PubMed, InCites). We also implemented the PubMed search feature generally to support future social network sources. Implementation details are provided in the source code (https://github.com/BaderLab/SocialNetworkApp).

Networks built from different bibliographic databases require their own unique visual styles. We implemented a standard visual style that all other visual styles extend. The standard visual style describes styles for attributes that all social networks share (e.g. *name, label*) and it is used for PubMed and Scopus networks but not InCites networks. InCites networks require their own specialized visual style because they contain additional attributes (*location*) that PubMed and Scopus networks typically do not possess. We implemented a new InCites visual style by extending the standard visual style and adding new style descriptions for the locations of authors. See the source code for implementation details.

Database evaluation and implementation

Multiple bibliographic databases were evaluated for support by the app: PubMed, Scopus, Web of Science/InCites, and Google Scholar. Evaluation was based on application program interface (API) availability, data export capabilities, coverage, citations and update frequency (see Table 1). PubMed is developed and maintained by the National Center for Biotechnology Information (NCBI) as part of the U.S. National Library of Medicine and is accessible through the Entrez query system (http://www.ncbi.nlm.nih.gov/pubmed/). Web of Science is a literature citation index created by Thomson Reuters containing over 90,000,000 records from all fields of science (http://wokinfo.com/citationconnection/realfacts/). Web of Science data is also accessible via the Thomson Reuters InCites web-based search engine, which facilitates access to additional information, such as author institution (http://researchanalytics.thomsonreuters.com/incites/). Scopus contains over 57,000,000 records, 27 million of which are patent records and 6.8 million of which are conference papers or proceedings (http://www.elsevier.com/solutions/scopus/content) and date back as far as 1823 (http://www.elsevier.com/__data/assets/pdf_file/0007/69451/sc_content-coverage-guide_july-2014.pdf).

Scopus contains author profiles that include, among other things, the institutional affiliations of an author. These profiles are helpful when disambiguating authors with very similar or identical names. Web of Science, InCites and Scopus access requires a paid subscription, which large academic institutions often provide.

Table 1. Coverage of various bibliographic databases. Scopus temporal coverage was retrieved here http://www.elsevier.com/__data/assets/pdf_file/0007/69451/sc_content-coverage-guide_july-2014.pdf.

Database	Indexed Citations	Temporal Coverage
PubMed	over 24 million	1946-present
Web of Science	over 90 million	1900-present
Scopus	over 57 million	1823-present
Google Scholar	100 million–160 million[2,3]	1700-present[3]

Google Scholar is a freely available and automatically updated database of citations with associated author pages (https://scholar.google.com/intl/en/scholar/about.html).

Database content was evaluated by selecting a specific publication and comparing its citation counts among the different databases. Update frequency was determined by checking whether a newly published paper (published on January 1st 2015 or later) had been indexed by the database and by examining citation counts and verifying that newer citations had been captured. Prior to app development, data from PubMed, Web of Science and Scopus was available for an internal project. Thus, development was oriented towards supporting content from these three databases.

Aside from Google Scholar, every database we examined had an API. Developers can access the APIs provided by both Scopus and Web of Science but a subscription is required. On the other hand, PubMed content and API access is free, easing implementation. Although both Scopus and Web of Science require paid subscriptions to view their data over the web, often large institutions have licenses to query this data which makes it accessible to many users. Scopus and Web of Science also both provide an intuitive web-based user interface that enables users to export the data to file formats that are recognizable by our app (CSV for Scopus and XLSX for Web of Science). Based on our evaluation, we chose to support PubMed (via file export and API), Scopus and Web of Science (via file export). We would have supported Google Scholar if a public API or file export was available.

PubMed is the default search engine used by the app because of the accessibility of its content and the straightforward nature of its associated retrieval mechanisms: its web-based interface and the Entrez Programming Utilities (eUtils) API. Eutils enables URL-based (non-RESTful) programmatic access to data contained in PubMed as well as any other databases linked to Entrez (http://www.ncbi.nlm.nih.gov/books/NBK1058/). Standard PubMed queries, for example "LastName First Initial"[Au], including recognized PubMed search tags (http://www.nlm.nih.gov/bsd/mms/medlineelements.html), can be entered into the PubMed search field in the app, which retrieves XML results using the eUtils web service. The results are parsed using the SAX (Simple API for XML) API included in the Java standard library and are transformed into a co-authorship network using the Cytoscape API. Nodes in the network represent authors, edges represent co-authorship and how frequently authors collaborate is indicated by the thickness of an edge.

Because data is retrieved from the NCBI servers through POST calls there is no restriction on the length of queries passed to PubMed through the app (http://www.ncbi.nlm.nih.gov/books/NBK25499/). Users can also construct networks from XML files exported directly from PubMed. Instructions for this workflow are provided in the app user guide: http://baderlab.org/UserguideSocialNetworkApp#PubMed. XML results obtained through eUtils differ slightly from XML results directly exported from PubMed. In particular, XML results exported from PubMed do not contain citations, whereas XML results retrieved by eUtils do. To correct this, the app retrieves this information using eUtils. Since the citation counts ultimately come from the same source regardless of how the initial data was obtained (PubMed or eUtils), networks generated via either method are equivalent. There is also a limit on the amount of data that can be retrieved at one time from eUtils. NCBI recommends that no more than 500 publications be retrieved from a single eUtils query (http://www.ncbi.nlm.nih.gov/books/NBK1058/). To circumvent this restriction, large data sets can be retrieved incrementally (i.e. 500 records at a time). There is also a limit set on the frequency of eUtils requests. A maximum of three requests is allowed per second (http://www.ncbi.nlm.nih.gov/books/NBK25497/). Violating these suggested limits may result in NCBI blocking the IP address of the offender. There is no limit when exporting data from the PubMed website directly. Users can safely export up to tens of thousands of records to XML.

Scopus and Web of Science (via InCites) are supported via file import. A user must manually export query results via the respective web interface. Scopus CSV exports are supported by the app. InCites reports must be saved in Excel 2007 (XLSX) format to be input into the app. The app can recognize InCites spreadsheets with exactly six columns in the following order (from left to right): *times cited, expected citations, publication year, subject area, all authors* and *document title*. Instructions on how to export results from InCites to this format are provided at http://baderlab.org/UserguideSocialNetworkApp#InCites.

Results and discussion

Use cases

We demonstrate the app using an example from the Hughes *et al.* study[6] in which social network analysis was used to determine whether Alzheimer Disease Centers (ADCs) based in the United States foster collaborative research. As part of the analysis, the study authors constructed multiple co-authorship networks using publication data collected from PubMed. In the original publication, the authors created Ruby scripts to query PubMed for co-authorships for a set of over 2000 researchers affiliated with ADCs.

The simplest way to interact with the app is to create either an individual researcher's publication network or a co-authorship network for an individual organization. Using an individual author from a single ADC, Rush University Medical Center, Figure 2 shows an individual's publication network. We created the co-authorship network by entering the researcher's name (last name <space> first initial, as expected by Pubmed) into the PubMed search bar (see Figure 1) and clicking on the 'Create Network' button. In order to generate the co-authorship network the app parses the results returned from the query. We assumed that all the authors with the specified last name and first initial in the results correspond to a single author. However, we made no attempt at name disambiguation, thus authors with common last names may be associated with inflated numbers of publications. For an individual author the same process can be performed on the Scopus or Incites websites to retrieve output files that can be loaded by the app. Conflicting author names may still be present although having institution affiliations available – as is the case for Scopus and InCites exported data – can help in disambiguating authors. Until such time that databases become cleaner or

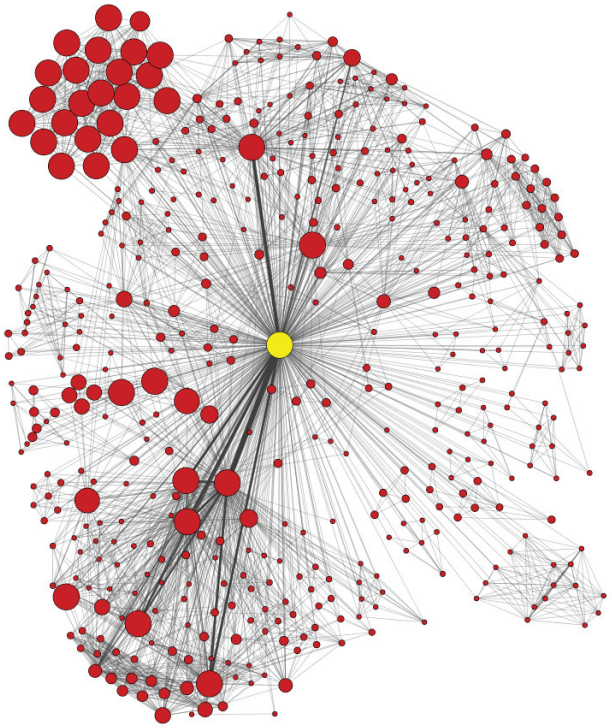


Figure 2. Publication network for an individual researcher at Rush University Medical Center. Each node represents a co-author of the original query author (the highlighted yellow node). The network was created by entering the author's last name and first initial into the PubMed query bar within the Social Network App. The network was then automatically created. The yFiles Organic layout was applied to better visualize the network. Node size represents the cumulative number of the author's publication citation counts as automatically retrieved from PubMed based on the set of publications associated with the node (the count only includes citations of publications that are in PubMed Central). Thickness of the edges connecting the nodes represents the number of publications the two authors have published together.

reliable automatic name disambiguation services become available, we recommend that users manually clean their data to resolve errors and name ambiguities before relying on co-authorship network results to support important decisions. The main difference between networks generated by PubMed, Scopus and Incites is the number of citations attributed to each author. PubMed counts paper citations only for articles found in the freely available PubMed Central literature archive whereas Scopus and Incites use a much larger set of publications stored in their databases. Thus, Scopus and Incites provide more accurate citation counts.

Extending the simple use case, using all authors from a single institution, such as Rush University Medical Center, as a query, Figure 3A shows the resulting co-authorship network (searching PubMed for publications that have at least two Rush University ADC researchers). Figure 3B shows the co-authorship network for only the Rush University ADC researchers. The length of the query depends on the number of researchers in the query set and would have the following format:

Example Department PubMed Query

(("LastName1 FirstInitial1"[Au] AND "LastName2 FirstInitial2"[Au])
OR
("LastName1 FirstInitial1"[Au] AND "LastName3 FirstInitial3"[Au])
OR
("LastName1 FirstInitial1"[Au] AND "LastName4 FirstInitial4"[Au])
OR
("LastName2 FirstInitial2"[Au] AND "LastName3 FirstInitial3"[Au])
OR
("LastName2 FirstInitial2"[Au] AND "LastName4 FirstInitial4"[Au])
OR ...)

AND "Rush University Medical Center"

A university department, faculty or a collaborative group typically desires to visualize and analyze all publications from the organizational unit over a period of time to help evaluate research

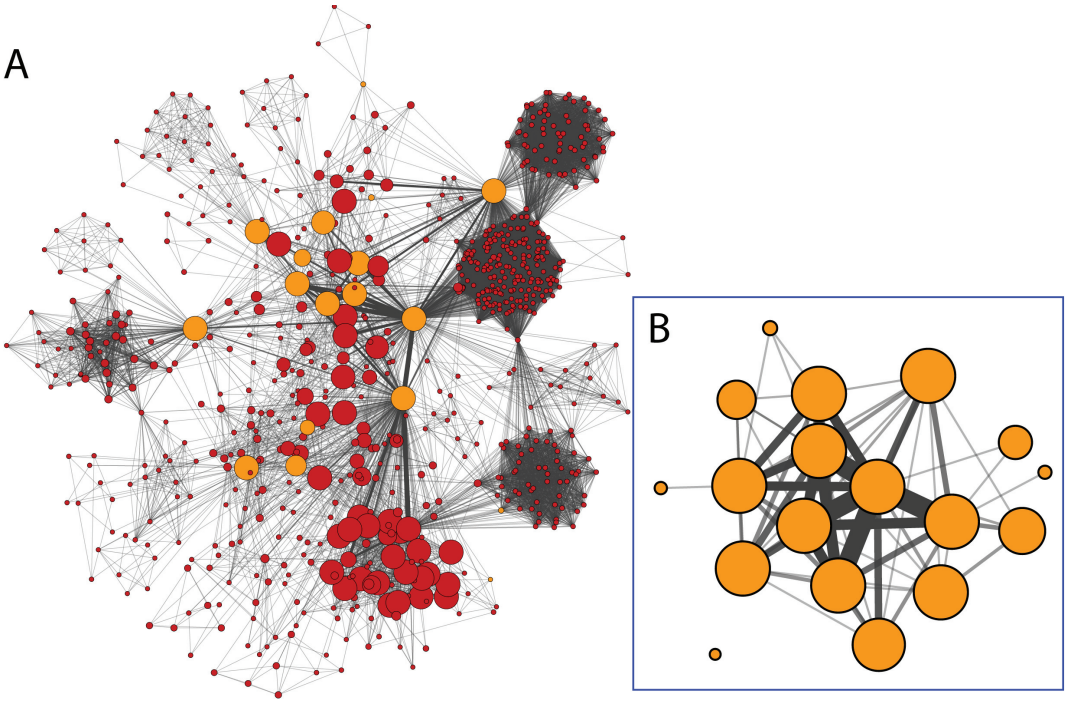


Figure 3. Co-authorship networks for ADC researchers at Rush University Medical Center. (**A**) Each node represents an author from the set of publications that have at least one Rush ADC researcher. Orange nodes are Rush ADC researchers and red nodes are non-Rush ADC researchers. Rush ADC researchers were selected manually and their node fill color was modified using the style bypass option and set to orange. This can also be achieved by importing a node attribute mapping only to the query authors and using the imported attribute to select the nodes. Node size represents the cumulative number of the author's publication citation counts as automatically retrieved from PubMed based on the set of publications associated with the node (the count only includes citations of publications that are in PubMed Central). Edge thickness represents the number of publications the two authors have published together. (**B**) Subset of the network in (**A**) containing only the ADC researchers. Author names are not shown to reduce visual clutter and to protect anonymity. Large cliques represent many-author publications.

productivity and effectiveness. Also, users may be interested in visualizing all of the publications and their topics in a particular research area. To demonstrate how the app can be used for a more sophisticated use case that also highlights how Cytoscape features can be used as part of a workflow, a simple comparison to the original broad analysis Hughes *et al.* was performed. We queried PubMed for the same set of researchers as used in the Hughes *et al.* study. Each author was queried along with their institution to reduce false positives and the entire query was limited to publications containing "alzheimer". The set of authors was large, leading to the creation of a large PubMed query, thus the PubMed web interface was used to execute the query. Both the Scopus and InCites web interfaces were unable to process the query and it was too long to pass to eUtils. Limiting the query to papers published in 2010 returned a set of 382 publications. Using the PubMed XML file downloaded from the PubMed website, we constructed a co-authorship network. By using Cytoscape's filtering capabilities, we reduced the network to just the authors used in the original query (see Figure 4). With Cytoscape's Styles, we colored nodes by institution as specified in the original dataset. To summarize this network we used a feature in the Enrichment Map App[7] that makes use of two other Cytoscape apps (clusterMaker[4] and WordCloud[5]) to automatically cluster and annotate the network based on the word summaries of a given attribute. Each cluster was annotated using frequent words found in the titles of publications within each cluster. This automatically highlights the collaborative research topics included in the network. The network can be further reduced by creating groups associated with each cluster. By collapsing the groups to an individual node the complexity in the network would be substantially reduced and the resulting network would highlight research themes found in this set of publications.

This workflow also illustrates the challenges of working with large co-authorship networks and large networks in general. There are many Cytoscape features and apps that can be used to reduce complexity of the network and help summarize the results. Given the limits of searching in PubMed, Scopus and Incites, a broad global analysis similar to the one conducted by Hughes *et al.* likely requires multiple queries, possibly automated by scripts to retrieve different data from the databases along with a process to collate

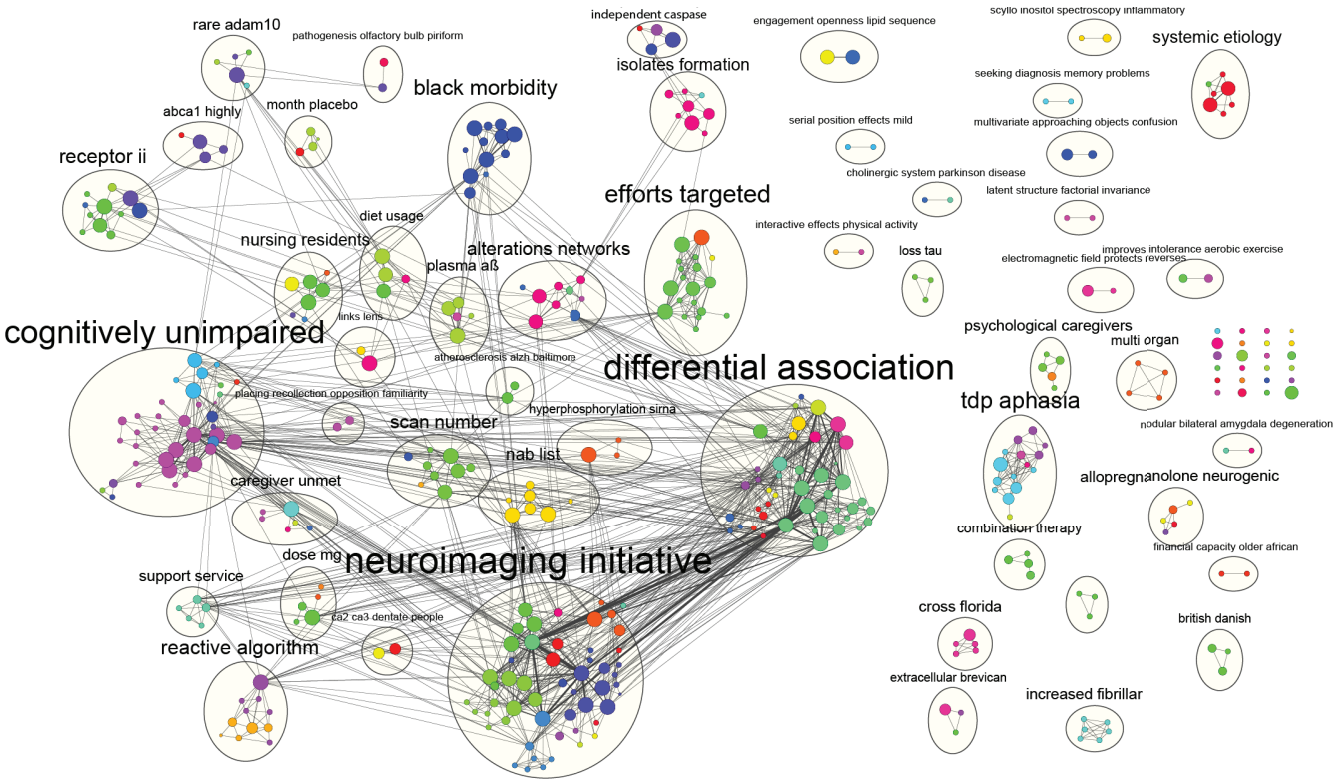


Figure 4. Co-authorship network for all ADC researchers from Alzheimer's publications in 2010. Each node represents an ADC researcher and its color represents their institution (as specified in the original dataset). Node colors were automatically generated in the visual style for the 'original institution' attribute that was available for this author set and loaded onto the co-authorship network after its creation. Node size represents the cumulative number of the author's publication citation counts as automatically retrieved from PubMed based on the set of publications associated with the node (the count only includes citations of publications that are in PubMed Central). Thickness of the edges connecting the nodes represents the number of publications the two authors have published together. Each cluster of authors, as calculated by the clusterMaker app[4] , is annotated with an additional circle around its members. Cluster labels aim to summarize the group with the two most frequent words in the author's set of publication titles as computed by the WordCloud App[5]).

and filter the results to generate the final set of publications to be analyzed. This is currently beyond the scope of this app, but more complex and automated query functionality could be added in the future.

Conclusions

Users interested in creating visual summaries of individuals connected via co-authorship links in academia will benefit from the Social Network App. The app aids users unfamiliar with Cytoscape by providing an intuitive and navigable user interface to query multiple bibliographic databases. Advanced users who wish to analyze large networks can take advantage of many powerful Cytoscape features. In the future we plan to expand the Social Network App to enable the creation of co-authorship networks made from multiple PubMed queries or files as well as support for Cytoscape commands which would enable scripted access to the app. We also plan to implement a system that will help disambiguate authors with common names and add support to visualize connections formed by Twitter, LinkedIn and Facebook users.

Software availability

Software available from

http://apps.cytoscape.org/apps/socialnetworkapp

Latest source code

https://github.com/BaderLab/SocialNetworkApp

Source code as at the time of publication

https://github.com/F1000Research/SocialNetworkApp

Archived source code as at time of publication

http://dx.doi.org/10.5281/zenodo.19825[8]

License: Lesser GNU Public License 2.1
https://www.gnu.org/licenses/old-licenses/lgpl-2.1.html

Tutorials
http://baderlab.org/UserguideSocialNetworkApp

Author contributions
VK drafted the manuscript with assistance from RI. Both VK and RI designed and developed the software. The project was initiated by GDB and AMJB. GDB supervised the project. All authors have read and approved the final manuscript.

Competing interests
No competing interests were disclosed.

Grant information
Financial support was provided by the Faculty of Medicine at the University of Toronto.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
We thank the authors of the Hughes *et al.* paper for sharing their social network data for analysis.

References

1. Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res.* 2003; **13**(11): 2498–2504.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Khabsa M, Giles CL: **The number of scholarly documents on the public web**. *PLoS One.* 2014; **9**(5): e93949.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Orduña-Malea E, Ayllón JM, Martín-Martín A, *et al.*: **About the size of Google Scholar: playing the numbers**. arXiv preprint arXiv: 1407.6239, 2014.
Reference Source
4. Morris JH, Apeltsin L, Newman AM, *et al.*: ***clusterMaker:* a multi-algorithm clustering plugin for Cytoscape**. *BMC Bioinformatics.* 2011; **12**(1): 436.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Oesper L, Merico D, Isserlin R, *et al.*: **WordCloud: a Cytoscape plugin to create a visual semantic summary of networks**. *Source Code Biol Med.* 2011; **6**(1): 7.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Hughes ME, Peeler J, Hogenesch JB, *et al.*: **The growth and impact of alzheimer disease centers as measured by social network analysis**. *JAMA Neurol.* 2014; **71**(4): 412–420.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Merico D, Isserlin R, Stueker O, *et al.*: **Enrichment map: a network-based method for gene-set enrichment visualization and interpretation**. *PLoS One.* 2010; **5**(11): e13984.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Kofia V, Isserlin R, Buchan AMJ, *et al.*: **F1000Research/SocialNetworkApp**. *Zenodo.* 2015.
Data Source

Open Peer Review

Current Peer Review Status: ? ?

Version 2

Reviewer Report 15 October 2015

https://doi.org/10.5256/f1000research.7716.r10730

© **2015 Bales M et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Michael Bales
Weill Cornell Medical College, New York, NY, USA
Terrie Wheeler
Weill Cornell Medical College, New York, NY, USA

Thank you for responding to our comments. We just have one concern remaining, in reference to the following discussion:

In our initial review we wrote, "On a related note, later in the sixth paragraph of the 'Database evaluation and implementation' section, the authors state that 'There is also a limit on the amount of data that can be retrieved at one time from eUtils. NCBI recommends that no more than 500 publications be retrieved from a single eUtils query'. The citation given is http://www.ncbi.nlm.nih.gov/books/NBK25498/. However, this citation does not directly support this assertion.

You responded: "Our apologies for the wrong citation. In “Building Customized Data Pipelines Using The Entrez Programming Utilities” (http://www.ncbi.nlm.nih.gov/books/NBK1058/) under the “Handling Large Datasets” subheading, it is stated that large lists should be split into smaller batches of around 500 records. We have updated the manuscript to reflect this change."

However, the new citation you provided refers to uploading UIDs, not downloading retrieved data. While it may be necessary when using Scopus, it is not necessary when using eUtils to modify queries so as to retrieve records in small batches. The actual maximum number of records that can be retrieved by a single query is 100,000, as indicated at http://www.ncbi.nlm.nih.gov/books/NBK25499/: "Increasing retmax allows more of the retrieved UIDs to be included in the XML output, up to a maximum of 100,000 records." One of us (M.B.) has used eUtils within the last year to download many thousands of records at a time, and can personally confirm that this is indeed possible.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 08 Dec 2015

Victor Kofia

Thanks for pointing out this documentation page. Our app uses esearch for UID retrieval and esummary for document summaries which we use to gather detailed information for the app. These have retmax limits of 100,000 and 10,000, respectively and we also confirmed these. We have switched the reference link to http://www.ncbi.nlm.nih.gov/books/NBK25499/ and updated our manuscript in this regard.

Competing Interests: No competing interests.

Version 1

Reviewer Report 21 August 2015

https://doi.org/10.5256/f1000research.7315.r9866

© **2015 Bian J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Jiang Bian
[1] Department of Health Outcomes & Policy, University of Florida, Gainesville, FL, USA
[2] Department of Health Outcomes & Policy, University of Florida, Gainesville, FL, USA

The article describes a software as an extension of the Cytoscape that can automatically query popular citation databases (PubMed only) and derive co-authorship networks based on the query results. It's a laudable goal, and will be a welcomed tool for researches on collaboration networks. Especially, the tool is disseminated as a open-source tool. However, there are a number of concerns.

1. The software does not have any process for disambiguation. This is problematic. In general, citation databases do not provide disambiguation services. Google Scholar attempts to "learn" which publications belong to a specific author when creating the author profile. However, it is not very accurate either at the beginning for common names. For studying social networks, getting accurate information is important, especially for studying Ego networks that focus on a specific person.

2. It is not very user friendly in terms of gathering data. Only for Pubmed, users can enter queries through the tool. With other citation databases (InCites and Scopus), users will need to query the databases directly and then export the results. It is understandable that this is

due to InCites and Scopus only provide paid API services. However, it is unclear whether the authors have implemented such integration for users who have paid those API services.

3. Further, even for Pubmed search, the software relies on the users to understand Pubmed query syntax. It would be useful if the authors could provide (in addition to text Pubmed query) guided search (see Pubmed Advanced search) for common use cases in studying collaboration networks. It would be very useful, if the authors could allow users to use the same guided search interface to three all citation databases.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Sep 2015

Victor Kofia

The software does not have any process for disambiguation. This is problematic. In general, citation databases do not provide disambiguation services. Google Scholar attempts to "learn" which publications belong to a specific author when creating the author profile. However, it is not very accurate either at the beginning for common names. For studying social networks, getting accurate information is important, especially for studying Ego networks that focus on a specific person.

Thank you for the comment. We agree that name disambiguation is a key concern with co-authorship networks. This is something we hope to address in the app in the future. In the meantime we have added to the paper some tips to minimize this issue. As noted in response to referee 1, until such time that databases become cleaner or reliable automatic name disambiguation services become available, we recommend that users manually clean their data to resolve errors and name ambiguities before relying on co-authorship network results to support important decisions.

It is not very user friendly in terms of gathering data. Only for Pubmed, users can enter queries through the tool. With other citation databases (InCites and Scopus), users will need to query the databases directly and then export the results. It is understandable that this is due to InCites and Scopus only provide paid API services. However, it is unclear whether the authors have implemented such integration for users who have paid those API services.

Our apologies for the lack of clarity. Right now the app does not support API querying for users who have paid services, but we have added this to a list of planned features in our github issue tracker at https://github.com/BaderLab/SocialNetworkApp/issues

Further, even for Pubmed search, the software relies on the users to understand Pubmed query syntax. It would be useful if the authors could provide (in addition to text Pubmed query) guided search (see Pubmed Advanced search) for common use

cases in studying collaboration networks. It would be very useful, if the authors could allow users to use the same guided search interface to three all citation databases.

Thanks for the comment. We have improved the user interface of the app and included a link to the PubMed query syntax. We have added a feature request to our issue tracker to develop a user friendly interface to assist users who may not be familiar with PubMed query syntax. For now we have included in the manuscript links to the relevant PubMed tutorials.

Competing Interests: No competing interests.

Reviewer Report 14 August 2015

https://doi.org/10.5256/f1000research.7315.r9859

© 2015 Bales M et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Michael Bales
[1] Weill Cornell Medical College, New York, NY, USA
[2] Weill Cornell Medical College, New York, NY, USA
Terrie Wheeler
[1] Weill Cornell Medical College, New York, NY, USA
[2] Weill Cornell Medical College, New York, NY, USA

The authors describe a tool that allows users to create co-authorship network diagrams within the Cytoscape application. This tool simplifies the network production process into as few as two steps: users can enter a PubMed query and then click the “search” button. A co-authorship network is generated automatically and displayed in Cytoscape; users can then use features available within Cytoscape to carry out additional tasks, such as adjusting visual properties and measuring topological measures of network structure. The system also supports searches to Scopus and Web of Science (via InCites).

The authors illustrate the system’s design and functions, describe how to use it, and present a use case in which they use the tool to replicate a search carried out in a study by Hughes *et al*. In so doing they highlight several challenges that may occur when working with large co-authorship networks.

The authors contend that co-authorship network visualization can be useful for understanding collaborative research networks or for “communicating the extent of collaboration and publication productivity among a group of researchers”. In this paper the authors do not conceive of a method to evaluate their app, for example to assess satisfaction with the app among system users, or to see whether system users are able to integrate the app into meaningful workflows. The authors report that they plan to continue developing the app, for example, by extending it so that it can display data from Facebook, Twitter, and LinkedIn.

The availability of the software is a welcome addition to existing tools for co-authorship network production. Its automated features and relatively seamless integration into Cytoscape will (we expect) make it an appealing option for analysts; to the extent possible it takes care of tedious steps that analysts may be accustomed to carrying out manually.

We have identified several minor points for the authors to take into consideration.

First, in the sixth paragraph of the “Database evaluation and implementation” section, the authors mention that E-utilities queries are limited to several hundred characters. While this may be the case for standard E-utilities queries, it is possible by using an HTTP Post call to make significantly longer queries. From Sayers E. The E-Utilities In-Depth: Parameters, Syntax, and More, http://www.ncbi.nlm.nih.gov/books/NBK25499/: “For very long queries (more than several hundred characters long), consider using an HTTP POST call.”

In the version of the Social Network that was current as of the time of this review, one of us (M.B.) attempted a query containing more than 1,000 characters and it was successful. In any case we would like to request clarification, as it appears that the current version of the system may be capable of longer queries, possibly by doing an HTTP Post request.

On a related note, later in the sixth paragraph of the “Database evaluation and implementation” section, the authors state that “There is also a limit on the amount of data that can be retrieved at one time from eUtils. NCBI recommends that no more than 500 publications be retrieved from a single eUtils query”. The citation given is http://www.ncbi.nlm.nih.gov/books/NBK25498/. However, this citation does not directly support this assertion. It is true, as the authors also point out, that requests should be limited to a maximum of three per second. Additionally, large jobs are to be limited to nights and weekends Eastern time (http://www.ncbi.nlm.nih.gov/books/NBK25497/). However, if there is a stated recommendation that no more than 500 publications be retrieved from a single eUtils query, we request that the authors identify a different source in which this is indicated in writing.

In the fourth paragraph of the “Database evaluation and implementation” section the authors mention that “Scopus and Web of Science APIs require paid subscriptions”. It is our understanding that Web of Science has an API that is free to subscribers, in addition to a paid API with more data fields. We would like clarification on whether the authors were referring to the API that is free to subscribers.

In the second paragraph of the “Results and discussion” section the authors state, “We created the co-publication network by entering the researcher’s name (last name <space> first initial, as expected by PubMed) … and clicking on the search button”. Due to the problem of ambiguous names in PubMed, it should be noted that this approach, without an attempt at name disambiguation, may result in many false drops, leading to invalid networks.

In the third paragraph of the “Methods and implementation” section the authors mention that the user panel includes a co-publication network summary panel. Later in this paragraph they mention “charts that summarize the total number of publications and citations by location can be viewed by clicking on links in the panel that navigate to summary charts created with the Google Chart API.” We have thus far been unable to locate the network summary panel or the links to the

charts, so further detail would be helpful here to describe the circumstances under which these features may be used, and/or how to activate and find them.

We also have some minor editorial suggestions.

First, we wanted to point out that “co-authorship” networks is far more common in the literature than “co-publication” networks, so the authors may wish to switch to this term if desired.

Second, in the third paragraph of the “Methods and Implementation” section the authors point out that “Because InCites networks contain institutional affiliations for all the authors of a given publication, they have richer summaries.” It may also be worth pointing out here that Scopus does this as well.

Third, the sixth paragraph of the “Database evaluation and implementation” section contains a broken link (http://baderlab.org/Software/SocialNetworkApp#PubMed).

We applaud the authors for making this app available for use within a freely available tool that has an active user base and community of users, and are hopeful that the authors will continue with active development of this tool, so that they may be responsive to user suggestions that may further improve the user experience and integration into workflows.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 27 Sep 2015

Victor Kofia

In the version of the Social Network that was current as of the time of this review, one of us (M.B.) attempted a query containing more than 1,000 characters and it was successful. In any case we would like to request clarification, as it appears that the current version of the system may be capable of longer queries, possibly by doing an HTTP Post request.

Thanks for pointing this out. We have now modified the app to use POST for all queries, which will be part of the next release.

On a related note, later in the sixth paragraph of the “Database evaluation and implementation” section, the authors state that “There is also a limit on the amount of data that can be retrieved at one time from eUtils. NCBI recommends that no more than 500 publications be retrieved from a single eUtils query”. The citation given is http://www.ncbi.nlm.nih.gov/books/NBK25498/. However, this citation does not directly support this assertion.

Our apologies for the wrong citation. In “Building Customized Data Pipelines Using The

Entrez Programming Utilities" (http://www.ncbi.nlm.nih.gov/books/NBK1058/) under the "Handling Large Datasets" subheading, it is stated that large lists should be split into smaller batches of around 500 records. We have updated the manuscript to reflect this change.

In the fourth paragraph of the "Database evaluation and implementation" section the authors mention that "Scopus and Web of Science APIs require paid subscriptions". It is our understanding that Web of Science has an API that is free to subscribers, in addition to a paid API with more data fields. We would like clarification on whether the authors were referring to the API that is free to subscribers.

We were referring to both APIs. Since users have to be subscribed to Web of Science to access the 'free' API we classified it as 'requiring a paid subscription'. We have clarified this point in the revision.

In the second paragraph of the "Results and discussion" section the authors state, "We created the co-publication network by entering the researcher's name (last name <space> first initial, as expected by PubMed) ... and clicking on the search button". Due to the problem of ambiguous names in PubMed, it should be noted that this approach, without an attempt at name disambiguation, may result in many false drops, leading to invalid networks.

We agree with the reviewer. Name disambiguation is definitely a problem when constructing co-authorship networks in this way. And it is a problem that affects all databases. We have updated the text to reflect this as it is something that all readers should be aware of. Until such time that databases become cleaner or reliable automatic name disambiguation services become available, we recommend that users manually clean their data to resolve errors and name ambiguities before relying on co-authorship network results to support important decisions.

In the third paragraph of the "Methods and implementation" section the authors mention that the user panel includes a co-publication network summary panel. Later in this paragraph they mention "charts that summarize the total number of publications and citations by location can be viewed by clicking on links in the panel that navigate to summary charts created with the Google Chart API." We have thus far been unable to locate the network summary panel or the links to the charts, so further detail would be helpful here to describe the circumstances under which these features may be used, and/or how to activate and find them.

The network summary panel is located at the bottom of the user panel. Instead of a figure that only shows the top half of the panel we have included a new figure that contains the entire panel as this will enable readers to view the sections we are referring to. If no network summary panel exists in your setup and the problem persists then it may be a bug, in which case we would appreciate it being filed as a bug report on GitHub (https://github.com/BaderLab/SocialNetworkApp/issues) or emailed to us so we can fix it.

Also note that even with the network summary panel being visible, links to the charts will

only appear after an InCites network has been created. No charts are available for Scopus and PubMed networks at the moment. So please verify that you are using an InCites document to build the network. An example InCites document is provided in the user guide: http://baderlab.org/Software/SocialNetworkApp. We have made this clearer in the text.

We also have some minor editorial suggestions. First, we wanted to point out that "co-authorship" networks is far more common in the literature than "co-publication" networks, so the authors may wish to switch to this term if desired.

Thank you for pointing this out. We have switched to using the more common "co-authorship" term throughout the manuscript.

Second, in the third paragraph of the "Methods and Implementation" section the authors point out that "Because InCites networks contain institutional affiliations for all the authors of a given publication, they have richer summaries." It may also be worth pointing out here that Scopus does this as well.

It is indeed true that Scopus provides institutional affiliations. The default setting for exporting files from Scopus is "Citation Information Only" which does not include the institutional affiliations of the co-authors. We have extended the functionality so that the app is capable of parsing Scopus reports that contain additional information (like institutional affiliations). We have made a note of this in the online manual.

Third, the sixth paragraph of the "Database evaluation and implementation" section contains a broken link (http://baderlab.org/Software/SocialNetworkApp#PubMed).

Thanks for noticing this. We have updated the manuscript to include the correct link: http://baderlab.org/UserguideSocialNetworkApp#PubMed

We applaud the authors for making this app available for use within a freely available tool that has an active user base and community of users, and are hopeful that the authors will continue with active development of this tool, so that they may be responsive to user suggestions that may further improve the user experience and integration into workflows.

Thank you for taking the time to review our app.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research