Date submitted: 22/06/2010

Impact of Open Access on stem cell research: An author co-citation analysis

Andreas Strotmann and **Dangzhi Zhao**
University of Alberta
Edmonton, AB, Canada

Meeting: **155. Science and Technology Libraries**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

We explore the impact of Open Access (OA) on stem cell research through a comparison of research reported in OA and in non-OA publications. Using an author co-citation analysis method, we find that (a) OA and non-OA publications cover similar major research areas in the stem cell field, but (b) a more diverse range of basic and medical research is reported in OA publications, while (c) biomedical technology areas appear biased towards non-OA publications. It appears that OA helps maintain diversity of research in this highly interdisciplinary field, and hence contributes to a healthy balance of scientific advancement.

INTRODUCTION

Since S. Lawrence's 2001 Nature paper claiming that “Free online availability increases a paper's impact” (Lawrence, 2001), many studies have investigated whether Open Access (OA) publication of research results has a positive effect on the citation ranking of those publications, on the assumption that early and wide availability of research will result in a citation advantage. Swan (2010) surveys more than 30 studies testing, measuring, or otherwise analyzing this effect. Other studies ask how popular the OA publishing idea is among scientists as readers and as writers, respectively. (Mann, et al., 2009).

In this paper, we approach the comparison between OA and non-OA publishing of research results from a somewhat different perspective. We explore whether there are substantial differences between the intellectual structure of a research field when viewed from either the point of view of the OA publications in that field or from that of its non-OA publications. Results from this comparison may shed light on issues such as which research areas of a scientific field tend towards OA and how OA publishing has impacted the research in a field.

METHODOLOGY

We use author co-citation analysis (ACA) to map the stem cell research field based on two subsets of the literature of this field published 2004-2009: OA and non-OA.

ACA has long been known to produce robust maps of the intellectual structure of research fields (White & Griffith, 1981). By comparing ACA maps of a research field produced from OA and from non-OA publications in this field, respectively, we can readily identify significant commonalities as well as significant differences between them.

We use ACA to map the stem cell research field, a sizable sub-field of recent biomedical research with in the order of 10,000 journal publications annually. The stem cell field is chosen for this study because OA movement has been quite strong in the biomedical fields. This field is highly multidisciplinary and highly collaborative. It is rapidly developing from seminal research conducted a couple of decades ago, as it promises important ramifications both with respect to basic biological research into the structure and development of individuals and with respect to understanding and potentially healing a wide range of important human diseases.

The stem cell research field is very well covered in the U.S. National Library of Medicine's PubMed bibliographic database. PubMed Central is also the major open access repository of OA publications in this field. We therefore use PubMed to retrieve bibliographic records of stem cell research papers, and to identify the OA and non-OA subsets of these records. We then use the Elsevier's Scopus to retrieve the references cited in these papers, which are required for conducting a citation analysis.

Data Collection

In order to study the intellectual structure of a research field using ACA, a set of publications in this field during a certain time period needs to be collected to represent this research field. The intellectual structure of the field can then be studied based on the perceptions of authors of these publications as expressed in their citation behaviors recorded in citation links they provide in these publications. Clearly, the more complete and clean this set of publications is (i.e., including as many papers as possible on this research field and as few as possible on research outside of this field), the better a research field is represented and therefore the more confidently its intellectual structure can be studied. The citation links in these publications are an essential part of the dataset, and a complete list of authors of each cited reference should be included in order to take into account all contributions of the authors regardless of their positions in the by lines.

In order to collect a complete and clean set of publications in the stem cell research field, we developed and employed a multi-step process to overcome the limitations of existing citation databases (i.e., Scopus and the citation indexes by the Institute for Scientific Information (ISI)) for studying this large and highly collaborative, highly multidisciplinary field. Details of this process and the underlying reasoning can be found in Strotmann, Zhao, & Bubela (2010). A summary of the key points is provided here.

Limitations of Existing Citation Databases for Studying Highly Collaborative, Multi-disciplinary Research Fields

(a) The highly collaborative nature of the stem cell research requires all-author citation

counting, which requires a complete list of authors of each cited reference. ISI citation databases only index the first author of a cited reference and Scopus provides up to eight authors. Scopus may be good enough for low- to-medium-level collaborative research fields, but does not suffice for the highly collaborative stem cell research field in which there are many papers with more than eight authors.

(b) The stem cell field is highly multidisciplinary, with research ranging from biology to therapy, across all organs to a variety of diseases, and from biomedical sciences to social sciences and law. Journals that publish stem cell research are highly diverse on the one hand, and cover non-stem cell research extensively as well on the other. The traditional way of defining a research field by a few core journals therefore does not work for the stem cell field.

(c) The stem cell field is large and extremely fast-growing. The number of publications within a year in this field is already beyond the limit that Scopus puts on search results for downloads (i.e., 2000). Refining the search by journal does not work for reasons in (b).

Creation of a Complete and Clean Dataset of Stem Cell Research

Step 1: We used a MeSH term search on “stem cell” in PubMed to retrieve records on stem cell research. We selected a citation window of six years from 2004 and 2009. The actual searches were carried out from December 2008 to May 2010 to allow sufficient time for PubMed to index the papers. A total of 31 040 papers was retrieved.

Step 2. We created a set of search strings from these PubMed records, and issued these search strings in Scopus manually in order to retrieve these papers along with their cited references. About 98% of the papers were found in Scopus, and were subsequently kept in the dataset for our study.

Step 3. 2,281,584 (or 95%) of the 2,405,522 cited references were found in PubMed and their full PubMed records were retrieved and used in the analysis. Those that were not found there were added by parsing the Scopus cited reference information, which includes the names of up to eight authors.

OA vs. Non-OA Publications

For the purpose of this study, we define both OA and non-OA publications in a relatively strict sense. OA publications are those full text journal articles freely available in PubMed Central. Non-OA publications are those journal articles that are not linked from PubMed to a freely available full text copy. The OA publications in a looser sense, i.e., those journal articles that were linked from PubMed to a freely available full text copy outside of PubMed Central, typically not included there for reasons of restrictive copyrights, are not counted in either OA or non-OA publications in this study.

Specifically, we downloaded from PubMed the list of PubMed identifiers for the strict OA subset of the full literature we retrieved earlier. We identified the corresponding records through these identifiers in our full dataset. These records form the strict OA sub-dataset for this study. We also downloaded from PubMed the list of PubMed identifiers for the loosely OA subset of the literature we retrieved earlier, and formed the strict non-OA sub-dataset by removing records identified in that list from the full dataset.

Data Analysis

Author Name Disambiguation

In a highly diverse and multidisciplinary field like stem cell research, the problems with author names (e.g., spelling variations of the same names, same author with different names and same names for multiple authors) are extremely pronounced. Author name disambiguation therefore became a necessary component of author-based citation and co-citation counting. Details of the method we used for author name disambiguation can be found in Strotmann, Zhao, & Bubela (2009). An updated version of that method was applied to the full stem cell dataset, and then applied to the OA and non-OA subsets by simple extraction from the full dataset.

Citation and Co-citation Counting

We ranked cited authors by the number of times they are cited by papers in the two datasets based on fractional all-author counting, the most preferred counting method for allocating credit in the case of multi-authored works (Lindsey, 1980; van Hooydonk, 1997; Zhao, 2005; 2006a). The 100 most highly cited authors of each dataset were selected for an ACA, and their co-citation counts were calculated and put into two matrices, one based on OA and one on non-OA citing publications. We use exclusive all-author co-citation counting to calculate co-citation matrices whose diagonal values are the authors' exclusive co-citation counts with themselves, a method that has been shown to be the most preferred both theoretically and in practice (White, 2003; Zhao, 2006b; Zhao & Strotmann, 2008a).

To clarify, when an article by N authors is cited, each of these N authors' fractional citation counts increase by 1/N. The exclusive all-author co-citation count of authors A and B increases by 1 whenever a paper cites at least one paper from A's oeuvre and at least one *different* paper from B's oeuvre. An author's oeuvre is defined as the collection of all papers written by this author as an author listed in any position in the byline.

Factor Analysis and Visualization

Each of the two author co-citation matrices was factor analyzed (Hair et al., 1998) and the results visualized as described in Zhao & Strotmann (2008b). To compare these two sets of results, several relevant features of the ACA maps are examined, including which specialties are identified, which specialties are most active, how these specialties are related to each other, and how clear the specialty structure is (White & McCain, 1998; White, 1990).

Ranking

In addition, the 20 most highly cited authors and journals in each set are reported below. In the case of journals, the journal identifier field in the PubMed XML record for each cited reference was used for unique and disambiguated journal identification and improved citation count accuracy. The author citation counts are based on the author name disambiguation briefly described above, and utilize fractional counting as described earlier (i.e., each author accrues 1/N citation counts per citation to a publication with N authors).

RESULTS

Based on Kaiser's rule of eigenvalue greater than 1, the factor analysis routine in SPSS 7.0

produced a 14- and 12-factor model from the author co-citation matrices generated from the OA and non-OA publications, respectively. Both factor models have a very good model fit: 89.6% and 92.4% of the total variance are explained by the OA and non-OA based models, respectively, and the differences between observed and implied correlations are smaller than 0.05 for the most part (99%) in both cases. The difference in model fit indicates that the picture from OA is more diverse while that from non-OA is more concentrated.

Major and minor specialties of the Stem Cell (SC) research field

Table 1 provides the label and size of each of the factors in the two sets of results. The label of a factor is determined based on an examination of the highly cited papers written by authors who load highly on this factor. The highest loading in each factor is provided in the table as an indicator of clarity and distinctiveness of a factor. The size of a factor is the number of authors who primarily load on this factor, but in two cases (marked with ** in the table) where the highest loading is very low, we report the total number of primary and secondary loadings that are 0.3 or higher.

As factors are interpreted as research specialties in ACA, Table 1 shows the major and minor specialties in the stem cell field and how these differ in the views represented in OA and non-OA publications.

Table 1. Specialty sizes and clarities of the stem cell research field represented in OA and non-OA publications, 2004-2009

Factor	OA		Non-OA	
	Size	Max. Loading	Size	Max. Loading
Pluripotency and differentiation	17	0.94	24	1.05
Adult neurogenesis	17	1.01	23	1.00
Mesenchymal SCs / Cytotherapy	2	1.03	11	1.00
Cancer SCs	17	0.86	12	0.98
Angiogenesis	4	0.94	9	1.00
Fibroblasts / Wound healing	2	0.96	3	0.93
Muscle / Heart regeneration	3	0.98	5	0.87
Skin SCs / Cancer	5	0.95	5	0.99
Neural SC Development	6	0.67	6**	0.48
Embryonic SC biotechnology	--	--	6**	0.49
Tissue Engineering	--	--	4	0.8
*Cell senescence	*	*	3	0.8
Cell senescence / Tumour suppression	9	0.9	*	*
Telomeres	3	0.95	*	*
Signalling / receptors	3	0.83	--	--
Embryonic SC Development	5	1.01	--	--
Cell Matrix	7	0.92	--	--

Notes:
-- Does not exist.
* This factor (Cell senescence) of the non-OA analysis corresponds to the two below (Telomeres; Cell senescence/ Tumour suppression) in the open-access analysis.
**Number of all loadings rather than only primary loadings. The highest loading is also a secondary loading.

Pluripotency and differentiation, Adult neurogenesis and Cancer SCs are the most active specialties in both views. In the non-OA view, Mesenchymal SCs / Cytotherapy and Angiogenesis specialties are quite active as well whereas the OA view sees the Cell senescence / Tumour suppression specialty quite active.

There are 3 small specialties that are clear in both views: Fibroblasts / Wound healing, Muscle / Heart regeneration, and Skin SCs / Cancer, and four specialties that only appear clearly in the OA view: Telomeres, Signalling / receptors, Embryonic SC Development, and Cell Matrix.

These specialties, and their relationships as well as the membership of researchers in these specialties are shown in Figures 1 and 2 which are visual representations of factor analysis results from non-OA and OA citing publications, respectively.

In these maps, the larger nodes represent specialties and the smaller nodes authors. The size of a specialty node is accumulated from loadings and serves as an approximate indicator of its overall significance in the map. The width and the greyscale value of lines connecting nodes are proportional to the value of the author's loading on this factor and represent the degree of relatedness, with darker and thicker lines representing closer ties. Only loadings that are 0.3 or higher are counted as in White & McCain (1998). For example, the left most author in Figure 1, Judith Campisi, belongs to the Cell senescence specialty quite strongly, but is only slightly related to the Cancer stem cell specialty.

Non-OA view of the intellectual structure of the stem cell research field

The non-OA map shows horizontally across its center an arc of specialties in the stem cell research field that, broadly, appear to focus on medical implications and applications of stem cell research. One side of that arc can be categorized as regenerative medicine, which aims to utilize the potential of stem cells to grow new tissue for repair or other treatment. The other side of the arc focuses on cancer research, where the goal is to understand and control the proliferative potential of cancer stem cells. The two sides are generally bridged by research on the (re-)growth of blood vessels, which has an obvious connection to regenerative medicine, where vascularization of new tissue (angiogenesis) is a universal requirement, and a less obvious one to cancer medicine where inhibition of neovascularization of cancerous tissue is a possible target for treatment.

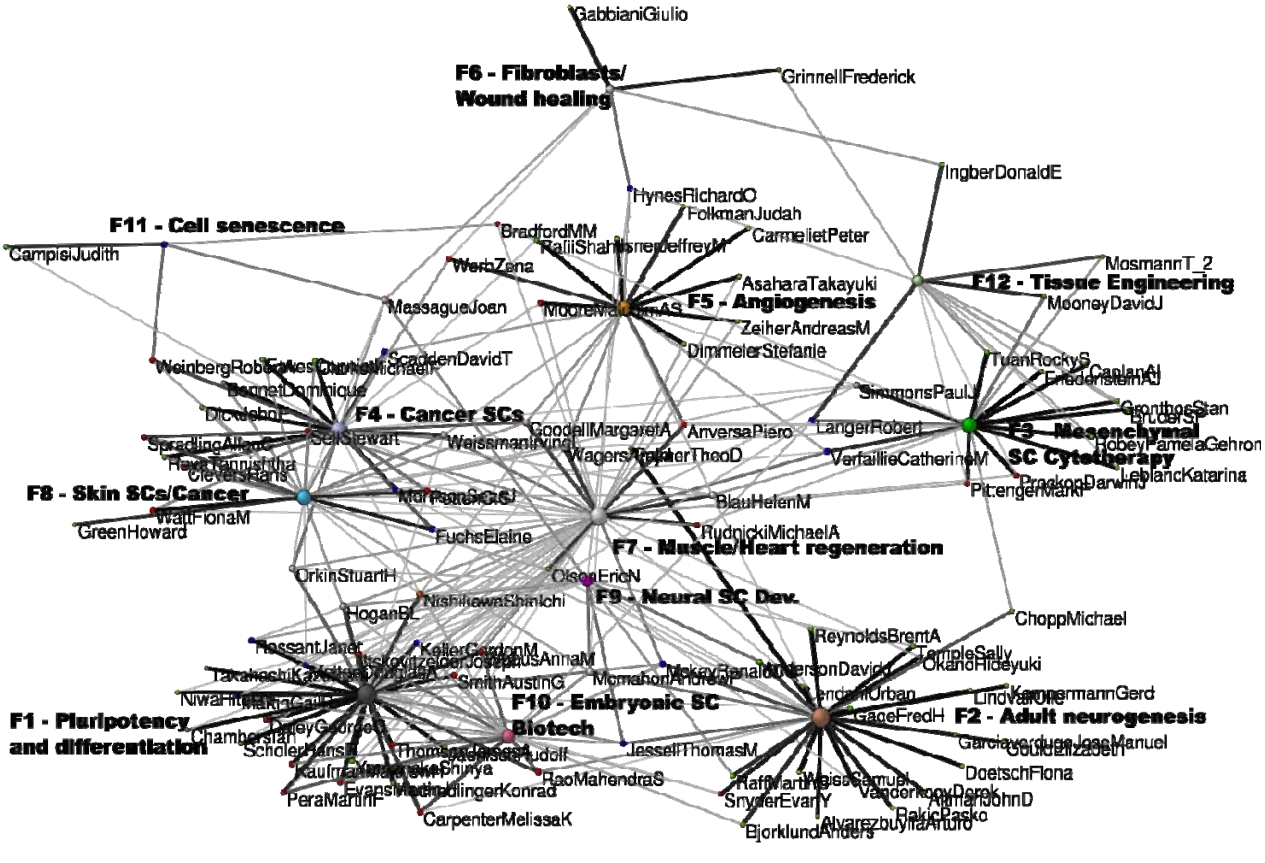

Figure 1: Authors co-cited in non-OA stem cell literature

Below this arc, we see two clusters of research specialties, one on neurogenesis (adult and embryonic, resp.) and one on embryonic and pluripotent stem cells.

Across the top of the figure, and more or less loosely connected to the different parts of the central arc, we see a very loosely connected outer arc of small specialties. On the cancer end of the map, we find research on cell senescence and cell death (which stem cells, like cancer cells, are able to avoid). On the regenerative end, we see wound healing and tissue engineering as peripheral specialties.

OA view of the intellectual structure of the stem cell research field

While they are arranged in a slightly different pattern, the major factors identified in the non-OA literature appear in the OA literature as well. This is true for the main medical applications research areas (both cancer and regenerative medicine ends as well as the bridging Angiogenesis specialty) and for the two main clusters found below them (Embryonic / pluripotent stem cell research and Neuronal stem cell research, resp.).

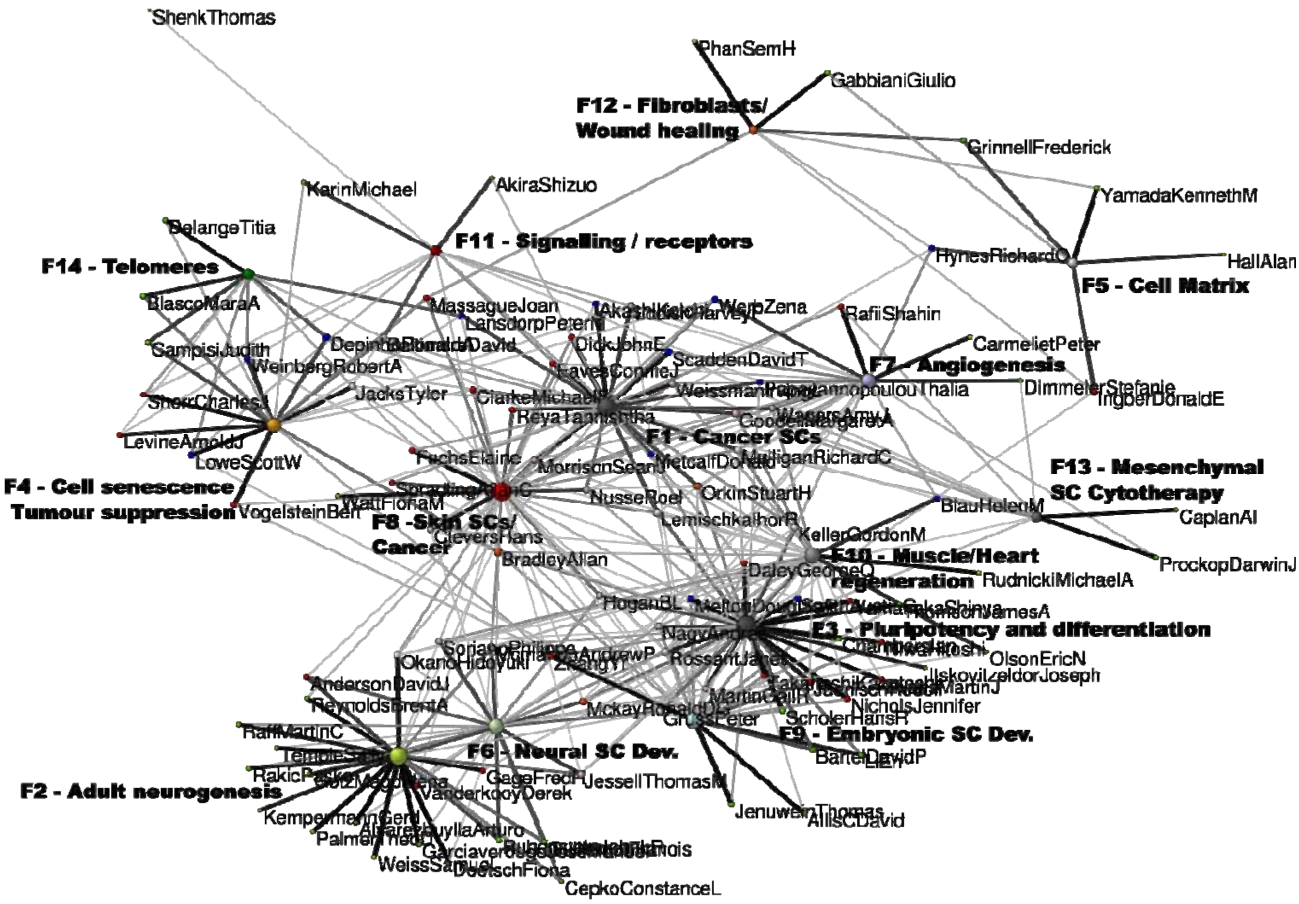

Figure 2: Authors co-cited in the OA stem cell literature

The arc of minor specialties in the non-OA publications becomes a major aspect of the map in the OA picture, however. Both in number and in size, these specialties become significantly more pronounced here. The tiny non-OA specialty on cell senescence (and its avoidance in stem and cancer cells through telomere maintenance) blossoms into three fields – a large area on tumour suppression through reactivation of cell senescence and cell death mechanisms in tumour stem cells, a medium size one on the role of telomeres in cell senescence, and a third small one on the signals and cell receptors involved in these mechanisms.

In addition, the arc of small specialties shows a major shift in focus – in addition to the shift to a focus on tumour suppression from cell senescence just described, the non-OA tissue engineering area is replaced by the related but significantly less applied research area on the interaction between stem cells and the surrounding cell matrix instead.

Similarly, the embryonic and pluripotent stem cell research cluster does not add an embryonic stem cell biotechnology research area, but instead adds a less applied specialty on the biology of embryonic stem cell development.

Rankings

Top 20 highly cited authors					
Table 2. in the Non-OA literature			**Table 3. in the OA literature**		
Name	**Non-OA Rank**	**OA Rank**	**Name**	**OA Rank**	**Non-OA Rank**
Gage, Fred H.	1	2	Weissman, Irving L	1	2
Weissman, Irving L	2	1	Gage, Fred H.	2	1
Prockop, Darwin J	3	9	Smith, Austin G	3	6
Alvarez-buylla, Arturo	4	7	Jaenisch, Rudolf	4	9
Caplan, A I	5	68	Morrison, Sean J	5	7
Smith, Austin G	6	3	Orkin, Stuart H	6	33
Morrison, Sean J	7	5	Alvarez-buylla, Arturo	7	4
Mckay, Ronald D G	8	29	Fuchs, Elaine	8	23
Jaenisch, Rudolf	9	4	Prockop, Darwin J	9	3
Thomson, James A	10	16	Keller, Gordon M	10	21
Dick, John E	11	17	Olson, Eric N	11	78
Yamanaka, Shinya	12	12	Yamanaka, Shinya	12	12
Weiss, Samuel	13	48	Weinberg, Robert A	13	41
Reynolds, Brent A	14	51	Soriano, Philippe	14	142
Clarke, Michael F	15	18	Sherr, Charles J	15	109
Itskovitz-eldor, Joseph	16	62	Thomson, James A	16	10
Verfaillie, Catherine M	17	105	Dick, John E	17	11
Rakic, Pasko	18	25	Clarke, Michael F	18	15
Garcia-verdugo, Jose Manuel	19	43	Gabbiani, Giulio	19	25
Martin, Gail R	20	31	Massague, Joan	20	30

Tables 2 and 3 present the 20 authors cited most highly by the OA and non-OA publications, and their corresponding ranks by fractional citation counts. Simply put, when a paper by N authors is cited, the fractional citation count of each of the N authors increases by 1/N.

Only 12 researchers are among the 20 most highly cited both in the non-OA and in the OA publications, or barely half. This confirms the findings from the two ACA maps above that there are significant overlaps, but also significant differences, between the two parts of the stem cell literature that we examine here. It should be interesting to examine the overlapping and non-overlapping authors in more detail to see how these two representations of the stem cell field differ.

Table 4 presents the top 20 journals ranked by how many times they have been cited in OA and non-OA publications.

Table 4. Most highly cited journals and their ranks

Journal	Non-OA Rank	OA Rank
Proc Natl Acad Sci U S A	1	2
Nature	2	3
Blood	3	6
J Biol Chem	4	1
Science	5	5
Cell	6	4
Development	7	7
Cancer Res	8	12
J Neurosci	9	16
Stem Cells	10	21
J Cell Biol	11	10
J Clin Invest	12	19
Genes Dev	13	9
Nat Med	14	22
Circulation	15	31
Biomaterials	16	56
Dev Biol	17	17
Mol Cell Biol	18	8
Biochem Biophys Res Commun	19	26
J Immunol	20	15
J Virol	41	11
EMBO J	24	13
Oncogene	21	14
Nat Genet	30	18

Sixteen of the 20 most highly cited journals are common to both the OA and the non-OA publications, and none of the non-overlapping highly cited journals are in the top ten. We therefore see little or no significant difference with respect to the highly cited journals, most of which are among the generalist science or medicine research journals. However, it is interesting to note that the only journal that is specific for stem cell research and the only technology-oriented journal Biomaterials are both ranked high in the non-OA ranking (10th and 16th respectively), but neither made to the top 20 in the OA publications. In fact, none of the most highly cited journals in the OA publications is highly specialized. These observations appear to provide additional evidence that biotechnology is more highly represented in the non-OA domain, as well as for a less specialized view of the stem cell research field from the OA than from the non-OA perspective.

DISCUSSION

The bird's eye view of stem cell research has many commonalities whether it is viewed exclusively from a strict open-access view or from a strictly non-OA view. That means that OA publishing has become a major part of the scholarly communication system in the stem cell research field, covering all the major areas of studies.

There are two major differences, however, between the OA and non-OA views.

(a) Major stem cell research areas labelled as “engineering” or “technology” (e.g., embryonic stem cell biotechnology) are picked up only by the map that is based on non-OA publications. This difference seen from the maps is supported by the rankings of highly cited journals. Technology oriented journals are only highly cited in the non-OA publications. At this point, we can only speculate why this is the case. It appears that the level of commercialization potential of stem cell research has significant influence on the degree to which it is published in OA journals.

(b) Several of the small specialties only play a significant role in the OA literature with one exception being an area closely related to tissue engineering and thus, again, to technology. Considering the fact that the strictly OA literature in this field is smaller than the strictly non-OA one by a factor of about three, it is actually quite remarkable that the smaller specialties are picked up by analysing the smaller rather than the larger literature. We may speculate that OA publications in the stem cell field covers a somewhat more diverse research community than does the non-OA literature.

CONCLUSIONS

While the strictly OA and non-OA journal publications in the stem cell research field account for less than 20% and roughly two thirds of the journal literature in the field, respectively, the intellectual structure of the field viewed from the non-OA publications appears to be significantly impoverished on the basic science and medical applications front of the field, while the OA perspective on the field tends to de-emphasize the biomedical technology aspect of research in favour of its underlying biomedical science.

It thus appears that the OA literature is instrumental in maintaining research diversity on the non-technological end of this field's research spectrum, at least in the stem cell research field studied here. OA publishing hence contributes to a healthy balance of scientific advancement. Further research would be required to see if this is true in other fields as well.

REFERENCES

Hair, J.F. Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate Data Analysis* (5th edition). Upper Saddle River, NJ: Prentice Hall.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature, 411*(6837), 521.

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science, 10*, 145-162.

Mann, F., Von Walter, B., Hess, T., & Wigand, R.T. (2009). Open Access publishing in science. *Communications of the ACM, 52* (3), 135-139.

Strotmann, A., Zhao, D., & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of The American Society for Information Science and Technology 2009 Annual Meeting*, November 6-11, 2009, Vancouver, British Columbia, Canada.

Strotmann, A., Zhao, D., & Bubela, T. (2010). Combining commercial and Open Access citation databases to delimit highly interdisciplinary research fields for citation analysis studies. *Journal of Informetrics, 4*(2), 194-200.

Swan, A. (2010). The Open Access citation advantage: studies and results to date. Technical Report, School of Electronics & Computer Science, University of Southampton. http://eprints.ecs.soton.ac.uk/18516/

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science, 48*, 944-945.

White, H. D. (1990). Author co-citation analysis: Overview and defense. In C. L. Borgman (ed.), *Scholarly communication and bibliometrics* (pp. 84-106). Newbury Park, CA: Sage.

White, H.D. (2003). Author cocitation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 54*, 1250-1259.

White, H.D., & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*, 163-171.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*, 327-355.

Zhao, D. (2005). Challenges of scholarly publications on the Web to the evaluation of science - A comparison of author visibility on the Web and in print journals. *Information Processing and Management, 41*(6): 1403-1418

Zhao, D. (2006a). Dispelling the myths behind straight citation counts. Information Realities: Shaping the Digital Future for All - *Proceedings of the American Society for Information Science and Technology 2006 Annual Meeting*, November 3 - 8, 2006, Austin, Texas, USA

Zhao, D. (2006b). Towards all-author co-citation analysis. *Information Processing and Management, 42*: 1578-1591.

Zhao, D., & Strotmann, A. (2008a). Comparing all-author and first-author co-citation analyses of Information Science. *Journal of Informetrics, 2*(3), 229-239

Zhao, D., & Strotmann, A. (2008b). Information Science during the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology, 59*(6), 916-937.