Author disambiguation using multi-aspect similarity indicators

Gurney, T.; Horlings, E.; van den Besselaar, P.

published in
Scientometrics
2012

DOI (link to publisher)
10.1007/s11192-011-0589-1

document version
Publisher's PDF, also known as Version of record

Scientometrics (2011) 87:115–131
DOI 10.1007/s11192-010-0317-2

Intellectual structure of stem cell research: a comprehensive author co-citation analysis of a highly collaborative and multidisciplinary field

Dangzhi Zhao · Andreas Strotmann

Received: 21 July 2010 / Published online: 12 December 2010

Abstract This study is an attempt to approach the intellectual structure of the stem cell research field 2004–2009 through a comprehensive author co-citation analysis (ACA), and to contribute to a better understanding of a field that has been brought to the forefront of research, therapy and political and public debates, which, hopefully, will in turn better inform research and policy. Based on a nearly complete and clean dataset of stem cell literature compiled from PubMed and Scopus, and using automatic author disambiguation to further improve results, we perform an exclusive all-author ACA of the 200 top-ranked researchers of the field by fractional citation count. We find that, despite the theoretically highly interdisciplinary nature of the field, stem cell research has been dominated by a few central medical research areas—cancer and regenerative medicine of the brain, the blood, the skin, and the heart—and a core of cell biologists trying to understand the nature and the molecular biology of stem cells along with biotechnology researchers investigating the practical identification, isolation, creation, and culturing of stem cells. It is also remarkably self-contained, drawing only on a few related areas of cell biology. This study also serves as a baseline against which the effectiveness of a range of author-based bibliometric methods and indicators can be tested, especially when based on less comprehensive datasets using less optimal analysis methods.

Keywords Citation analysis · Author co-citation analysis · Scholarly communication · Intellectual structure · Research policy · Stem cell research · Biomedical research

Electronic supplementary material The online version of this article (doi:10.1007/s11192-010-0317-2) contains supplementary material, which is available to authorized users.

D. Zhao (✉) · A. Strotmann
School of Library and Information Studies, University of Alberta, Edmonton, AB T6G 2J4, Canada
e-mail: dzhao@ualberta.ca

A. Strotmann
e-mail: andreas.strotmann@ualberta.ca

Introduction

Recent years have seen stem cell research rising to the forefront of biomedical science, public health and research policy (Department of Health and Human Services 2001, 2006).

"A stem cell is a special kind of cell that has a unique capacity to renew itself and to give rise to specialized cell types. … Their proliferative capacity combined with the ability to become specialized makes stem cells unique" (Department of Health and Human Services 2001). Stem cell research investigates the biological and medical promises of stem cells. Its long-term major clinical goals include (a) improved understanding of cancers that develop from stem cells running amok, and (b) utilizing the ability of stem cells to differentiate into a large variety of tissue types to assist in healing a range of "diseases, conditions, and disabilities including Parkinson's disease, amyotrophic lateral sclerosis, spinal cord injury, burns, heart disease, diabetes, and arthritis" (Department of Health and Human Services 2001).

The Stem cell research field is highly multidisciplinary, given its huge implications for public health and the intense scrutiny it has received with respect to medical ethics. It is advancing at an incredible pace with new discoveries being reported in the scientific literature on a weekly basis. It should therefore be interesting and timely to take a look at how this field has been developing and what its social and intellectual structures are, to inform researchers, policy makers as well as laypersons.

The present study uses an author co-citation analysis (ACA) approach to identify major specialties, laboratories, researchers, and research groups in the stem cell research field, and to examine how they relate to each other. This is enabled by a citation dataset we compiled through a sophisticated multi-step process that collects much cleaner data on stem cell research publications and much more accurate and complete information about the cited references they contain, compared to data from citation indexes commonly used in citation analysis studies, i.e. Scopus and the databases by the Institute for Scientific Information (ISI).

Research questions and related studies

There is a large body of literature on citation analysis studies of research fields and specialties, including some biomedical research fields, which has been reviewed extensively from various perspectives (Borgman and Furner 2002; Morris and Van der Veer Martens 2009; White and McCain 1989, 1997). ACA is a central part of relational citation analysis studies.

These studies clearly show the power of citation analysis in the study of scholarly communication patterns in research fields, as well as the limitations brought by the citation data source that most of these studies relied on, i.e., the ISI citation indexes. For example, these ISI citation databases only index the first author of any cited reference although they index all authors of each source paper in the databases. As a result, only first-author citation counting is possible in a large-scale citation analysis using data downloaded from these databases, although when an individual author name is searched for manually in these databases, all references to papers written by this author are retrieved, including those that do not list this author name as the first author in the byline but are indexed in the ISI databases as source papers (i.e., references to works that are not indexed as source paper and do not list the author as first author will not be

retrieved). The lack of support for all-author citation counting is a serious limitation when it comes to identifying and assessing highly influential (i.e., highly cited) researchers, especially in research fields where coauthorship is commonplace. Scopus provides up to eight authors of a cited reference, supporting all-author-based citation analysis studies for many research fields, but is still quite insufficient for research fields where large-scale collaborations are common.

Biomedical fields are among these highly collaborative research fields and even exhibit occasional hyperauthorship (Cronin 2001). According to Newman (2001), the mean number of authors per paper in the biomedical fields documented by MEDLINE was 3.75 and the average total number of collaborators per author was 18, compared to 2.22 and 3.59 respectively in the computer science fields. The collaboration level in some biomedical subfields was even higher than seen in the entire MEDLINE (e.g., 6.18 authors/paper in the cardiovascular subfield) (Bordons et al. 1996). Our own data show that less than 10% of the stem cell publications in the past few years were single-authored, and about one in seven had more than eight authors.

It is generally understood that research and publication culture is different in different research fields. While in many fields such as Library and Information Science authors are ordered in the byline by their contribution to the paper, the meaning conveyed in authorship order is different in chemistry, biomedical fields and many other research fields. Researchers in these fields "work in individual laboratories in close association with their own group of students, postdoctoral fellows, and technicians" (Brown 2010, p. 307). The lab's head is often the principal investigator who develops initial ideas for research and procures the funding it requires. Its junior researchers often conduct the actual studies and perform the necessary experiments in the lab under the guidance of the lab head. Research results are published with that junior researcher as the publication's first author, the lab head as its last, and the other lab members or other collaborators involved inbetween (Sonnenwald 2008, p. 670). This wide-spread research and publication culture has not been taken into account sufficiently in citation analysis studies, due partly to the limited support provided by available citation indexes.

We have undertaken a full-scale study on the scholarly communication patterns of the stem cell research field that takes into account both this research and publication culture and the highly collaborative nature of this field, going to considerable lengths to put together as complete and clean an author citation dataset for the field as we can and to analyze that dataset with the best author co-citation methodology available. This paper reports the part of the results of this study that focuses on the scholarly communication patterns as seen from ACA.

Specifically, the present paper addresses the following research questions:

(1) What is the overall structure of the stem cell research field?
(2) Which are the central, peripheral or bridging specialties in this field?
(3) Who are the central, peripheral or bridging researchers in this field?

Results from this study should contribute to a better understanding of the stem cell research field that has recently been brought to the forefront of research, therapy, and political and public debates, which may in turn better inform research and policy. These results will also serve in related studies as a baseline for the evaluation and testing of bibliometric methodologies, especially as they are applied to highly collaborative research fields.

Methodology

Data collection

In order to study the scholarly communication patterns of a research field using a citation-based approach, a set of publications in this field during a certain time period needs to be collected to represent this research field. The scholarly communication patterns of the field can then be studied based on the perceptions of authors of these publications as expressed in their citation behaviours recorded in citation links they have made 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 better its scholarly communication patterns 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.

During the last few years, between five and ten thousand scholarly journal publications per year have been published in the stem cell research field. Less than 10% of these publications were single-authored, and one in seven had more than eight authors. Given the magnitude of the dataset, the pervasiveness of multi-authorship, the multidisciplinary nature of this field, and limitations of current citation databases (i.e., Scopus, ISI citation databases), traditional core journal- or keyword search based methods for collecting data using existing citation databases, especially the ISI data source, do not work well for this study for a number of reasons. We therefore developed and employed a multi-step process in order to build a dataset for this study that is close to complete, clean and accurate compared with a dataset gathered directly from existing citation databases. Details of the reasoning behind as well as the steps and algorithms of this process can be found in Strotmann et al. (2010). A summary of key points is provided below.

Limitations of existing citation databases for studying highly collaborative, highly multi-disciplinary research fields

(a) The highly collaborative nature of stem cell research requires all-author 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 research fields such as Library and Information Science, 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, covering all organs of the body and a wide variety of diseases, as well as research ranging from biomedical sciences to the social sciences and law. Journals that publish stem cell research are highly diverse on the one hand, and usually cover non-stem cell research extensively as well on the other. For example, stem cell research is reported extensively in the general science journals (e.g., Nature, Science) and general medical journals (e.g., Lancet). Examples of journals that most frequently appear in our dataset include, in addition, New England Journal of Medicine, Cell, Blood, Circulation Research, Neuron, Stroke, and Cancer. This means that traditional methods for retrieving a set of literature to represent a research field do not work for this field, whether basing it on a set of core journals or on a keyword search.

(c) The stem cell field is large and extremely fast-growing. The number of publications within a year in this field is already a multiple of the limit that Scopus puts on search results for download (i.e., 2,000). Refining a search by journal does not work for reasons in (b).

Creation of a close-to-complete and clean dataset of stem cell research

Retrieval of PubMed records of papers representing the stem cell research field We used a Medical Subject Heading (MeSH) term search on "stem cell" in PubMed. We selected a citation window of 6 years from 2004 to 2009 with the plan to study the development of this field during this period through a comparison of three 2-year time slices.

The actual searches for the years 2004–2007, 2008, and 2009 were carried out in December 2008, August 2009, and May 2010 respectively to allow sufficient time for PubMed to index the papers. A total of 62,081 papers were retrieved (8,215 papers for 2004, 9,304 for 2005, 10,475 for 2006, 10,915 for 2007, 11,561 for 2008, and 11,611 for 2009). The smaller increase in 2009 might be because some 2009 papers have not been indexed in PubMed by the time of retrieval.

Retrieval of Scopus records of stem cell papers and their cited references We created a set of search strings from these PubMed records, and issued these search strings in Scopus manually[1] in order to retrieve these papers along with their cited references. About 98% of these papers were found in Scopus, and were subsequently kept in the dataset for our study, including the cited references they contained.

Completion of information on cited references We mapped these Scopus records of cited references to their corresponding full PubMed records. 2,281,584 (or 95%) of the 2,405,522 cited references were found in PubMed. Those that were not found there were kept in the dataset by parsing the Scopus cited reference information, which includes the names of up to eight authors. In other words, only about 5% of cited references might not have all of their authors indexed in our dataset. Since Scopus does index up to eight authors for a cited reference, this means that we likely have the names of 98% or more of all cited authors in this dataset.

The fact that the vast majority of stem cell papers retrieved in PubMed was found in Scopus and the vast majority of cited references in these papers was found in PubMed indicates that stem cell research is covered very well by both PubMed and Scopus. The dataset built this way is thus almost certainly much more complete and clean than traditional core journal or keyword search based methods applied directly to a citation index, even considering the error rates reported above.

Data analysis

Author name disambiguation

In a highly diverse and multidisciplinary field like stem cell research, the well-known 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

[1] Scopus licensing forbids automated retrievals.

name disambiguation therefore became a necessary component of author-based citation and co-citation counting.

We summarize here the key points of the method we used for author name disambiguation. Details of this method can be found in Strotmann et al. (2009). We used an updated version of the algorithm described there.

In the first phase of the disambiguation process, our algorithm classifies a particular set of papers as belonging to an individual author's oeuvre if:

(1) The author names for this individual from all papers in this set are mutually compatible.
(2) There is positive evidence that indicates that all papers in the set have been co-authored by the same individual.
(3) There is more positive evidence for a paper to belong to this particular oeuvre than for potential alternative oeuvres.

Positive evidence for two papers to belong to the same individual's oeuvre includes full name equality; shared coauthors; and close similarity of topics (e.g., shared major MeSH term). Compatibility of author names requires equal normalized last names, and compatibility of full first names and/or initials.

In the second phase, the results of the first phase are converted to a collaboration graph, and two authors in this collaboration graph are identified as the same individual if

(1) their full names are mutually compatible,
(2) they share collaborators, and
(3) they do not collaborate with each other.

This step is repeated until the algorithm converges, usually five times.

The results of this disambiguation are not perfect, but quite comparable to other algorithms reported in the literature.

Citation and co-citation counting

We used fractional all-author counting to select top-cited researchers for the ACA, and exclusive all-author co-citation counting to calculate their co-citation matrix, with the diagonal values being the authors' exclusive co-citation counts with themselves. Studies have shown that these methods are the most preferred citation and co-citation counting methods both theoretically and in practice (Ahlgren et al. 2003; Lindsey 1980; van Hooydonk 1997; White 2003; Zhao 2005, 2006a, b; Zhao and Strotmann 2008a, c).

To clarify, when an article by N authors is cited, each of these N authors' *fractional citation count* increases by $1/N$. The *exclusive 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 paper from B's oeuvre that is different from the one in A's. An author's oeuvre is defined as the collection of all papers written by this author in any position in the byline.

We ranked cited authors by the number of times they are cited by papers in our dataset using fractional all-author citation counting. The top 300 authors were selected for an author co-citation analyses (ACA), and their co-citation counts were calculated using exclusive all-author co-citation counting and entered into a matrix. The diagonal values of the co-citation matrices are the authors' exclusive co-citation counts with themselves. For example, the diagonal value of author A is the number of papers that cited at least two *different* papers written by author A in any position in the byline.

Factor analysis and visualization

This author co-citation matrix was factor analyzed and results visualized using the method introduced in Zhao and Strotmann (2008b). We only included the top 200 authors in this 300 × 300 matrix in the factor analysis to meet the variable-case ratio requirement by the factor analysis method (Hair et al. 1998), resulting effectively in a matrix of 200 variables (cited authors) and 300 cases (their co-cited authors). Considering the size of the stem cell field, we chose a threshold here (i.e., 200) for author selection that is much larger than the largest seen in published ACA studies to date (Zhao and Strotmann 2008a).

Results and discussion

Using the factor analysis routine in SPSS 7.0, Kaiser's rule of eigenvalue greater than 1 resulted in a 19-factor model, which has an excellent model fit. The model explains 92.6% of the total variance, and the differences between observed and implied correlations were smaller than 0.05 for the most part (almost 100%). An oblique rotation (Direct Oblimin) of the factor analysis resulted in a pattern matrix available as Electronic Supplementary Material to this paper and in a structure matrix presented here as a two-dimensional map (Fig. 1).

Since large factors are interpreted as specialties in ACA (White and McCain 1998), the pattern matrix shows specialties (i.e., factors) and authors' unique contributions to each specialty (i.e., authors' loadings in each factor), and the structure matrix represents both the individual authors' contributions to specialties and the correlations between specialties (Hair et al. 1998). Each of these 19 factors is therefore labelled based on the pattern matrix from an examination of the highly cited papers written by authors who load highly on the factor (Zhao and Strotmann 2008b).

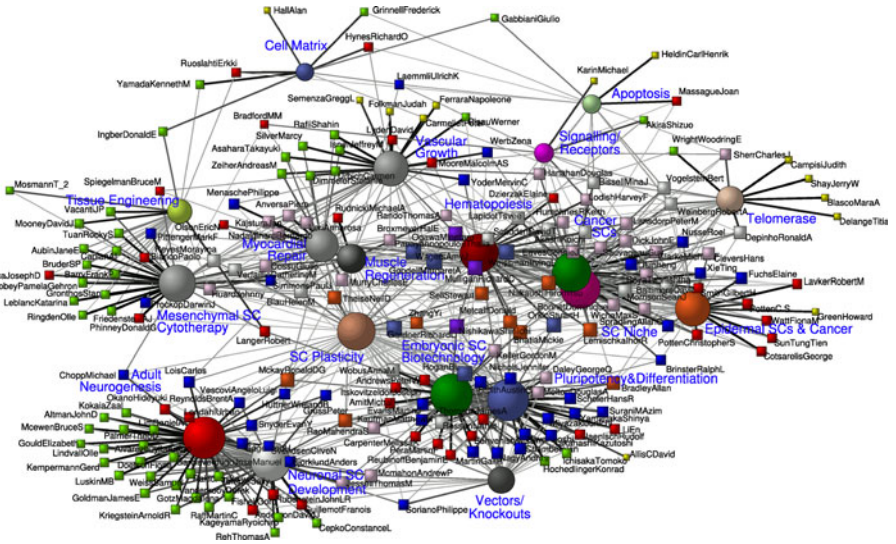

Fig. 1 Researchers, specialties and their interrelationships

The Electronic Supplementary Material is a large table in which factors are presented as columns, authors as rows, and author loadings on factors as cell values. As in White and McCain (1998), only loadings of 0.3 or greater are shown to focus on the authors' major memberships in specialties. The top 60 authors as ranked by fractional counts are highlighted in bold, and their ranks indicated. Factor names are listed at the top of the table. When factors are interpreted as specialties, these top ranked authors can be considered highly influential authors in each specialty whereas authors with highest loadings in each factor are considered focus researchers in the corresponding specialty.

Table 1 summarizes the pattern matrix presented as Electronic Supplementary Material, listing the label, the size, and the highest loading of each of the 19 factors. The size of a factor is the number of authors who primarily load on this factor in the pattern matrix. The highest loading is an indicator of clarity and distinctness of the factor. Table 1 also provides for each factor the names of its top three most focused researchers (i.e., its most highly loading authors) as well as the names and citation ranks of its three most highly influential researchers (i.e., the three most highly cited authors with fractional citation rank of 60 or higher who load primarily on the factor).

Table 1 shows that the Adult Neurogenesis specialty is the largest and most clear and distinct with 31 primary loadings, the highest of which is 1.06, followed by the Mesenchymal Stem Cell Cytotherapy, Embryonic Stem Cell Biotechnology, Pluripotency and Differentiation, and Vascular Growth specialties which have 17–19 primary loadings and 0.9 or larger highest loadings. In contrast, the Stem Cell Plasticity and the Knockouts factors have very few primary loadings (3 and 2) that are all quite low (the highest being 0.58 and 0.57 respectively). These factors are small, unclear, and indistinct, and therefore cannot be labelled or interpreted with much confidence (thus, undefined—UD).

The structure matrix of the factor analysis results is directly visualized in Fig. 1 in which the circular nodes represent specialties and the square nodes authors. The width and the grey-scale value of lines connecting author and specialty nodes are proportional to the value of the author's loading on this factor and represent the degree of relatedness of an author with the specialty, with darker and thicker lines representing closer ties. The layout uses the Kamada-Kawai implementation in Pajek with loadings as similarities. As in the pattern matrix, only loadings of 0.3 or higher are retained. The size of a specialty node is accumulated from loadings and serves as an approximate indicator of its overall significance in the map. Note that this size is a count of both primary and secondary loadings in each factor of the structure matrix, and is different from the factor size in Table 1 which is the number of primary loadings in each factor of the pattern matrix (Hair et al. 1998; Zhao and Strotmann 2008b).

In addition, the color of an author node (available only in the electronic version of the present paper) indicates the number of specialties in which this author has membership: yellow for authors who only have membership in a single specialty, green for two, red for three, blue for four, and other colors for more than four specialties.

Overall structure

Horizontally across the centre of the map in Fig. 1 we see a loosely connected arc of specialties in the stem cell research field that, broadly, appear to focus on medical implications and applications of stem cell research. The left half of the 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 right half of the arc focuses on cancer research, where the goal is to reduce the growth potential of cancer stem cells. The two

Table 1 Size, highest loading, focus researchers, and most highly influential authors of each factor

Factor	Size	Highest loading	Focus researchers[a]	Most highly influential authors (and their ranks)[b]
Adult Neurogenesis	31	1.06	Gould, E; McEwen, BS; Altman, JD	Gage, FH (2); Alvarez-Buylla, A (4); Weiss, S (14)
Mesenchymal SC Cytotherapy	19	0.99	Leblanc, K; Friedenstein, AJ; Ringden, O	Prockop, DJ (3); Caplan, AI (7); Verfaillie, CM (26)
Embryonic SC Biotechnology	19	0.95	Reubinoff, BE; Carpenter, MK; Pera, MF	Thomson, JA (9); Keller, GM (13); Itskovitz-Eldor, J (18)
Pluripotency and Differentiation	18	0.90	Allis, D; Li, E; Surani, A	Smith, A (5); Jaenisch, R (8); Yamanaka, S (12)
Vascular Growth	17	0.99	Carmeliet, P; Folkman, J; Ferrara, N	Dimmeler, S (25); Asahara, T (37); Rafii, S (39)
Hematopoiesis	15	0.81	Dzierzak, E; Humphries, K; Lodish, HF	Orkin, SH (17); Scadden, DT (46)
Cancer SCs	14	0.89	Wicha, MS; Clarke, MF; Dick, JE	Weissman, IL (1); Morrison, SJ (6); Dick, JE (11)
Neuronal SC Development	10	1.01	Cepko, CL; Reh, TA; Kageyama, R	Anderson, DJ (35); Jessell, TM (60)
Epidermal SCs and Cancer	8	1.03	Lavker, RM; Cotsarelis, G; Sun, TT	Fuchs, E (20); Watt, FM (28)
Myocardial Repair	8	0.84	Menasche, P; Murry, CE; Nadalginard, B	Goodell, MA (52); Wagers, AJ (54)
Telomerase	8	1.02	Delange, T; Shay, JW; Campisi, J	Weinberg, RA (29); Sherr, CJ (50); Campisi, J (59)
Muscle Regeneration	7	0.87	Rudnicki, MA; Rando, TA; Olson, EN	Olson, EN (33); Blau, HM (40); Rudnicki, MA (43)
Cell Matrix	6	0.90	Yamada, KM; Hall, A; Ruoslahti, E	
Signalling/Receptors	4	0.85	Karin, M; Akira, S; Baltimore, D	
SC Niche	4	0.81	Xie, T; Brinster, RL; Spradling, AC	
Tissue Engineering	4	0.80	Mooney, DJ; Vacanti, JP; Mosmann, T	Langer, R (53)
Apoptosis	3	0.93	Heldin, CH; Massague, J; Gabbiani, G	Gabbiani, G (19); Massague, J (23)
UD—Stem Cell Plasticity	3	0.58	Theise, N; Sell, S; Gardner, R	
UD—Vectors/ Knockouts	2	0.57	Soriano, P; Bradley, A	

[a] Top 3 authors ranked by their pattern matrix loadings (>0.3) in each factor

[b] Authors ranked by fractional citation count; only the 60 most highly cited authors are considered highly influential here, thus none are listed for some factors

halves are bridged by research on haematopoietic (i.e., bone marrow) stem cells, which has strong connections to (blood) cancer on the one hand but also provides a central ingredient for transplantation-style regenerative medicine. Research on the (re-)growth of blood vessels, too, bridges the two sides as it has an obvious connection to regenerative medicine, where vascularization of new tissue is a universal requirement, and a less obvious one to

cancer medicine where inhibition of neovascularization of cancerous tissue is a potential target for treatment.

Below, a central node connects this arc to two clusters of highly interconnected research specialties, one on neuronal stem cells and one on embryonic and pluripotent stem cells. The central node itself is hard to label, as it has many low loadings across nearly the entire field and no high loadings at all. It appears to represent the idea that stem cell research has a common theme, namely, the plasticity of stem cells, i.e., their ability to differentiate into any type of body tissue.

Across the top of the map, finally, we see a very loosely connected arc of small specialties that are also more or less loosely connected to one or more of the research fields in the central regenerative medicine/cancer research arc. On the cancer end of the spectrum, this includes research on cell senescence and cell death (which stem cells, like cancer cells, are able to avoid) and on extracellular regulation of stem cell differentiation and maturation (with implications for metastasis formation on the cancer end, and for targeting of damaged tissue on the regenerative medicine end of this arc). On the regenerative end, we see tissue engineering and cell matrix interactions of stem cells as two peripheral specialties.

Major clusters of specialties

Embryonic/induced pluripotent stem cell (SC) cluster

This cluster of research specialties deals primarily with embryonic stem cells (which are totipotent) and with pluripotent stem cells (which can grow into any kind of tissue). It consists of two large highly correlated research specialties and a small specialty that connects this cluster with the Neuro-cluster. This cluster of specialties has direct strong connections with the cancer research area and with the neuronal stem cells cluster. It indirectly connects to the regenerative medicine area through the central connecting factor.

The large specialty in this cluster labelled “Embryonic SC Biotechnology” is entirely focused on embryonic stem cells and the biotechnologies required to locate, produce, grow or otherwise handle or understand them in vitro. The journal *Nature Biotechnology* stands out among those that publish research relevant to this area, in addition to cell biology journals and the usual multidisciplinary ones. As seen in Table 1, focus researchers in this specialty include Benjamin Reubinoff, Melissa Carpenter, and Martin Pera. Highly influential researchers include James Thomson and Joseph Itskovitz-Eldor who are ranked top 9 and 18 respectively among all stem cell researchers by fractional counting. Together with their colleagues, they were the first to successfully isolate embryonic stem cells from human embryos and grow them in culture. This seminal paper was published in *Science* in 1998, and has been cited 1,879 times.

The second large factor in this cluster, labelled “Pluripotency and Differentiation”, deals primarily with pluripotent stem cells, in particular induced pluripotent stem cells (which are stem cells which, on the one hand, can grow into any type of tissue at all, but on the other hand are not derived from embryonic stem cells). A common theme of research in this specialty is the molecular biology machinery inside the cell that regulates pluripotency as well as the differentiation of pluripotent stem cells via less potent stem cells into tissue cells. Publications in this specialty tend to appear in the journal *Cell* in addition to *Nature* and *Science*.

As seen in Table 1, focus researchers in this specialty include David Allis, En Li, and Azim Surani. Most influential researchers in this area include Austin Smith, Rudolf

Jaenisch, and Shinya Yamanaka who are ranked 5, 8, and 12 respectively among all stem cell researchers. Yamanaka and his Japanese colleague Kazutoshi Takahashi co-authored several highly cited papers, the most famous of which is their 2006 paper in the journal *Cell*, reporting their discovery of how to induce pluripotent stem cells from mouse embryonic and adult fibroblast cultures. This discovery started a whole new fast-growing research area in the stem cell field, and has been cited 839 times.

The much smaller specialty labeled "Vectors/Knockouts" appears to deal with embryonic knockout mouse stem cells and the technologies (particularly viral vectors) required to create them. This factor has two primary loadings (from Philippe Soriano and Allan Bradley) and two secondary loadings (from Andras Nagy and BL Hogan) as shown in the Electronic Supplementary Material. Andras Nagy is a key member of the NorCOMM (North American Conditional Mouse Mutagenesis) knockout mouse consortium.

Neuronal stem cell research cluster

This cluster consists of two very strongly interconnected research specialties. Both are published in neuro-science journals in addition to the usual multidisciplinary and cell biology ones. This cluster of specialties is very self-contained as many researchers only connect with the other Neuronal Stem Cells specialty within this cluster, as seen from the many green author nodes in Fig. 1.

One specialty, labelled "Neuronal SC Development," traces the development and differentiation of neuronal stem cells from the embryo to adult neurons, along with the genetic and bio-chemical machinery that affects this process. Research in this specialty is published in genetics and/or biochemistry journals in addition to the ones mentioned earlier. As seen in Table 1, the top focus researcher in this area is Constance Cepko. The most influential researcher in this area appears to be David Anderson at Howard Hughes Medical Institute who is ranked 35 overall.

The other specialty, labelled "Adult Neurogenesis", focuses on identifying and tracing neuronal stem cells which actively differentiate into brain tissue in the adult brain, as well as on experimental therapeutic applications of these findings for stroke or for neurodegenerative disorders like Parkinson's or Alzheimer's. This specialty is the largest of all specialties. As shown in Table 1 and in the Electronic Supplementary Material, focus researchers in this area are Elizabeth Gould, Bruce McEwen, John Altman, and Gerd Kempermann. Most influential researchers in this area are Fred Gage, Arturo Alvarez-Buylla, Samuel Weiss, Brent Reynolds, Pasko Rakic, and Jose Manuel Garcia-Verdugo, who are ranked 2, 4, 14, 16, 22 and 24 overall respectively. Gage concentrates on the plasticity and adaptability of the adult brain, and Alvarez-Buylla is focused on neurogenesis of the adult brain. Weiss and colleagues discovered brain stem cells that differentiate in vitro and in the intact brain to produce the major neuronal cell types found in the brain.

Regenerative medicine

This cluster is divided largely by the class of stem cells involved—mesenchymal stem cells in one specialty, myocardial stem cells in another, muscle stem cells in a third, and endothelial progenitor cells in the fourth. The first of these targets general cytotherapy; the second focuses on repairing infarcts; the third, on repairing muscle tissue; and the fourth, on regrowing or repairing blood vessels. Specialties in this cluster are much more loosely connected compared to the embryonic SC cluster and the neuronal stem cells cluster.

The "Mesenchymal Stem Cell Cytotherapy" specialty is the largest in the cluster, and most researchers here (in green) only relate to the Tissue Engineering specialty, making it relatively separated from the rest of the cluster. Darwin Prockop, Al Caplan, Catherine Verfaillie, and Mark Pittenger are the most influential researchers here, ranked 3, 7, 26, and 30 respectively among all stem cell researchers, as shown in Table 1 and in the Electronic Supplementary Material. Pittenger and colleagues at Osiris Therapeutics published a seminal paper in *Science* in 1999 on "Multilineage Potential of Adult Human Mesenchymal Stem Cells" which signalled the start of this area of research, and has been extremely highly cited (2,511 times).

The second largest in this cluster is the "Vascular Growth" (angiogenesis and neovascularization) specialty. Like the Mesenchymal SC Cytotherapy specialty, many researchers in this specialty are either focused on this single specialty (yellow) or connect with only one other specialty (green), mostly Myocardial Repair. Stefanie Dimmeler, Takayuki Asahara, Shahin Rafii, and Jeffrey Isner are the most influential members. They are ranked 25, 37, 39 and 51 respectively. Isner, Asahara and their colleagues co-authored several highly cited papers published in journals such as Science and Circulation Research. Their 1997 Science paper on "Isolation of putative progenitor endothelial cells for angiogenesis" has been cited 1,110 times.

The other two major specialties in this cluster (especially the Muscle Regeneration specialty) are heavily integrated with the rest of the field as indicated in Fig. 1 by the lack of yellow or green or even red author nodes.

Cancer research

While it has long been understood that stem cells and cancer cells have much in common (in particular, their immortality and their ability to proliferate), the discovery that cancer stem cells (i.e., stem cells running amok) might be the primary initiators of tumours has greatly added to the interest in stem cell research in the cancer medicine community. This cluster of specialties is heavily interconnected as indicated in Fig. 1 by the lack of yellow or green author nodes, except the Telomerase specialty which is small but quite distinct.

Within this cluster, one specialty focuses on "Cancer stem cells", and on finding molecular markers that identify them (i.e., markers which distinguish (potentially) cancer-causing stem cells from well-behaving stem cells). The hope is that these markers could somehow be targeted in anti-cancer drugs and help avoid some of the terrible side effects of current chemotherapies, some of which are due to the fact that these therapies kill off all active stem cells rather than just the malignant ones. This subarea is particularly active with respect to leukemia (cancers of the blood), which connects it directly to the haematopoietic stem cell research area.

The second major type of cancers being studied in this area is skin cancer, and "Epidermal stem cells" are instrumental to this separate specialty.

A third specialty in this part of the stem cell research field studies the concept of a "Stem Cell Niche"—the idea that the cells surrounding a stem cell help determine the type of cells that it will differentiate into. Cancer stem cells evade this control mechanism, which keeps healthy stem cells in check. The ability to mimic the molecular signalling factors that effectuate this control would, on the one hand, offer a potential cancer drug target, and on the other hand, make it possible to "program" stem cells to produce specific desired tissue for regenerative medical purposes.

As seen in Table 1, the most influential cancer researchers in the stem cell field are Sean Morrison, John Dick, Michael Clarke, Elaine Fuchs, and Fiona Watt who are ranked 6, 11,

15, 20, and 28 overall respectively. The former three are in the Cancer Stem Cells specialty, which is the largest, while the latter two are in the Epidermal Stem Cells specialty, which is the most distinct in this cluster. Irving Weissman, who is most highly cited among all stem cell researchers, loads on the Cancer Stem Cells specialty slightly higher than on the Hematopoiesis specialty although both loadings are low.

Haematopoietic stem cells

The study of bone marrow stem cells forms a large specialty that connects regenerative medicine with cancer research (in particular, leukemia) as well as with research on embryonic stem cells. These haematopoietic (blood forming) stem cells have been used in medical research and therapies for a very long time, indeed long before the existence of stem cells as a separate and widespread class of cells in the body was discovered. Consequently, research both on and with these cells has been comparatively mature and widespread. In addition, these stem cells potentially have effects in any part of the body as they (or their immediate offspring) are mobile via blood vessels or lymphoid ducts.

Research in this area concentrates, on the one hand, on the genetics and the molecular biology that regulates the development of haematopoietic stem cells from embryonic origins, and on their differentiation into a host of different cell types on the other hand (in particular to help repair tissues and fight cancers).

This specialty is not distinct, with mostly medium to low loadings. As shown in the Electronic Supplementary Material, many researchers have secondary loadings on this factor, including Irving Weissman, the most highly cited author among all stem cell researchers. The top focus researcher of this area is Elaine Dzierzak, the only author who loads higher than 0.8 on this specialty. The most influential researcher who loads primarily in this specialty appears to be Stuart Orkin who is ranked 17 overall.

Integration of stem cell research

At the centre of Fig. 1 we see a factor with a large number of low-to-medium co-loadings with (almost) every other factor on the map, and without any high author loadings that would give it a defining core. The papers that constitute the closest thing to a core that this factor has can only be characterized as spanning the breadth of the entire field—all the organs (liver, muscle, brain, bone marrow) which characterize individual clusters of the regenerative medicine and cancer related research, and both the biology and the medical application of stem cells. “Stem cell plasticity” may be the common theme here—their ability to differentiate into a wide range of types of cells.

We suspect that this factor, at its core, represents studies that attempt to construct a cohesive landscape of stem cell research results from across its separate research areas, and thus to integrate findings from across the main dimensions of the field (embryonic and pluripotent vs. more specialized stem cells; medical applications that induce vs. inhibit stem cell proliferation and differentiation) into a coherent picture. In this regard, it is similar to the Haematopoietic SC specialty, which integrates the medical stem cell research.

Peripheral specialties

This bird’s-eye view of the intellectual landscape is rounded off by a number of smaller specialties, which form a wide arc across the entire upper part of the map in Fig. 1.

The three factors labelled "Signaling/Receptors", "Telomerase", and "Apoptosis" that connect to the cancer research side of the central arc appear to deal with different phases of the stem cell life cycle—cell maturation (especially its initiation via extracellular signals activating the TGF-beta receptors), cell senescence (particularly its absence in stem cells due to over-expression of telomerase), and cell death (especially apoptosis induced via signals activating mitochondrial cell death). These three specialties connect to each other in this order. As seen in Table 1, Giulio Gabbiani and Joan Massague, who are ranked 19 and 23 respectively among all stem cell researchers, are the most influential researchers in the Apoptosis specialty, although Gabbiani also loads on the Cell Matrix specialty, only slightly lower (0.44 vs. 0.53).

The peripheral specialties that are linked to the regenerative medicine aspect of stem cell research, labelled "Cell Matrix" and "Tissue Engineering", are respectively concerned with the biology of the attachment and adhesion of (stem) cells to the surrounding cell matrix (including wound healing) on the one hand, and with the biotechnology involved in mimicking the three-dimensional cell matrix and both its biomolecular and physical–mechanical properties for ex vivo tissue growth on the other. These specialties are connected to each other, but also (via a linkage between stem cell/cell matrix adhesion and stem cell maturation) to the peripheral cancer-related specialties.

Connecting and bridging researchers

We have identified the most influential researchers of major specialties of stem cell research as well as researchers who are focused on each specialty. Another interesting aspect of the scholarly communication pattern of the field is which researchers connect or bridge the various specialties, or more specifically, which researchers bridge between specialties in the same cluster, and which researchers connect specialties across clusters. We will focus our analysis on the major clusters of specialties.

As seen in the Electronic Supplementary Material, in the Neuronal SC cluster, seven researchers load on both the Adult Neurogenesis specialty (2) and the Neuronal SC Development specialty (18). That means that their work is recognized in both specialties, connecting these two major areas within this cluster. These researchers are Martin Raff, Arnold Kriegstein, Magdalena Gotz, and Sally Temple who are primarily located in the Adult Neurogenesis specialty, as well as John Rubenstein, Gord Fishell, and Wieland Huttner who are primarily in the other specialty.

Michael Chopp and Ronald McKay, who are primarily located in the Adult Neurogenesis specialty, also load on the Mesenchymal SC Cytotherapy specialty (4) and on the Embryonic SC Biotechnology specialty (11), respectively, and Peter Gruss, who is primarily in the Neuronal SC Development specialty, co-loads on the Pluripotency and Differentiation (3) and the Muscle Regeneration (9) specialties. This indicates that their work has impact in two or three specialties across clusters, serving as a bridge between these clusters of research areas, i.e., between the Neuronal SC cluster and the Regenerative medicine or the Embryonic/Induced Pluripotent SC cluster.

In the Embryonic/Induced Pluripotent SC cluster, some top researchers are recognized in both of the two major specialties in this cluster, Pluripotency and Differentiation (3) and Embryonic SC Biotechnology (11), some almost equally. These bridging researchers include Austin Smith, Janet Rossant, and George Daley in the Pluripotency and Differentiation specialty, who are ranked 5, 27, and 58 respectively among all stem cell researchers, and Gail Martin, Douglas Melton, and Martin Evans in the other specialty, who are ranked 21, 32 and 42 respectively overall.

Researchers who connect these two major specialties in this cluster with the small indistinct area of research Vectors/Knockouts (15) include Andras Nagy, BL Hogan, and Allan Bradley.

Researchers in this cluster who connect to specialties in other clusters include Mickie Bhatia, Gordon Keller, and Shinichi Nishikawa in the Embryonic SC Biotechnology specialty (11) who connect to the Haematopoietic Stem Cells specialty (12), Yi Zhang in the Pluripotency and Differentiation (3) specialty who connects to the Mesenchymal SC Cytotherapy specialty (4) in the Regenerative Medicine cluster, and Andrew McMahon who connects to the Neuronal SC Development specialty (18) in the Neuronal SC cluster.

More than half of the 14 researchers who primarily load on the Haematopoietic Stem Cells specialty (12) have secondary loadings, which provides further evidence for the connecting role discussed earlier that this large specialty plays between regenerative medicine, cancer research (in particular, leukemia) as well as research on embryonic stem cells.

Similarly, in the Regenerative Medicine cluster, more than half of the researchers who primarily load on the Muscle Regeneration specialty (9) have secondary loadings, some of whom have two secondary loadings each. In contrast, researchers who primarily load on the Mesenchymal SC Cytotherapy (4) or the Myocardial Repair (10) specialties are focused on their own single specialty except one researcher in each of these specialties who has a secondary loading.

In the Cancer cluster, the three smaller areas of research do not directly connect to each other, but all connect to the large Cancer SC specialty (19) through one or two researchers' secondary loadings, which in turn connects to the central Haematopoietic Stem Cells (12) specialty through five authors' secondary loadings, including the highly influential researcher Irving Weissman who is ranked first among all stem cell researchers. However, the smallest factor SC Niche (17) does connect directly to the Haematopoietic Stem Cells (12) and to the Epidermal Stem Cells (8) specialties through Linheng Li and Allan Spradling respectively, but not to the Telomerase (7) specialty.

Conclusions

Using a comprehensive ACA approach, this study examines the intellectual structure of the stem cell research field 2004–2009, a field that has been brought to the forefront of research and public policy in recent years. Based on a nearly complete and clean dataset of stem cell literature compiled from PubMed and Scopus for these years, and using automatic author disambiguation to further improve results, we perform an exclusive all-author co-citation analysis of the 200 top-ranked researchers of the field by fractional citation count.

We find that, despite the theoretically highly interdisciplinary nature of the field, this bird's-eye view of the research field is dominated by a few central medical research areas—cancer and regenerative medicine of the brain, the blood, the skin, and the heart—and a core of cell biologists trying to understand the nature and the molecular biology of stem cells along with biotechnology researchers investigating the practical identification, isolation, creation, and culturing of stem cells. The stem cell research field is also remarkably self-contained, drawing only on a few related areas of cell biology.

Specifically, stem cell research appears to have been focused on four broad areas of study: embryonic and pluripotent stem cells, neurogenesis, regenerative medicine, and cancer research. Efforts appear to have begun to integrate the experimental studies in these

different areas into a coherent whole. Some peripheral small specialties are quite interesting and promising, and may represent emerging specialties.

ACA provides a useful tool for helping non-experts of stem cell research to understand the overall intellectual structure of the stem cell research field. The ACA results appear consistent with published reviews of stem cell research (Department of Health and Human Services 2001, 2006), and provide a nice overview of the structure of major specialties and researchers and their interrelationships that those reviews do not provide.

Finally, the comprehensive ACA presented here serves as a baseline against which we can test the effectiveness of a range of author-based bibliometric methods and indicators, especially when based on less comprehensive datasets (e.g., keyword searching results downloaded from the ISI databases) and/or using less optimal analysis methods (e.g., first-author citation counting).

Acknowledgments This study was funded in part by the Social Sciences and Humanities Research Council (SSHRC) of Canada and by Genome Canada. The authors would like to thank Gencheng Guo for his assistance in the data collection process.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science, 54*, 550–560.

Bordons, M., Gomez, I., Fernandes, M. T., Zulueta, M. A., & Mendez, A. (1996). Local, domestic and international scientific collaboration in biomedical research. *Scientometrics, 37*(2), 279–295.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 3–72.

Brown, C. (2010). Communication in the sciences. *Annual Review of Information Science and Technology, 44*, 287–316.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*(7), 558–569.

Department of Health and Human Services. (2001). Stem cells: Scientific progress and future research directions. Retrieved May 21, 2010 from http://stemcells.nih.gov/info/2001report/2001report/.

Department of Health and Human Services. (2006). Regenerative medicine. Retrieved May 21, 2010 from http://stemcells.nih.gov/info/2006report/.

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

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

Morris, S. A., & Van der Veer Martens, B. (2009). Mapping research specialties. *Annual Review of Information Science and Technology, 42*, 213–295.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science of the USA, 98*(2), 404–409.

Sonnenwald, D. H. (2008). Scientific collaboration. *Annual Review of Information Science and Technology, 41*, 643–681.

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, BC, 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.

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

White, H. D., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology, 24*, 119–186.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 32*, 99–168.

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, TX, 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.

Zhao, D., & Strotmann, A. (2008c). Evolution of research activities and intellectual influences in Information Science 1996–2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086.