Progress in genome research over fifteen years (1999-2013): longitudinal changes and organism distribution

Accepted 27th May, 2015

ABSTRACT

This study is aimed to describe the profile of genomic studies through a bibliographic analysis of articles published over the past 15 years, with a specific focus on i) outcomes of genomic articles, ii) journals involved, iii) authorship patterns (institutes and countries), and iv) research topics, as a means of tracking research on genome development. The primary literature data were obtained from the Science Citation Index-Expanded (SCI-E) database for the period 1999 to 2013. All articles with the terms “genome or genomic(s)” included in the title or topic field were analyzed. A total of 50,933 articles on genomes published in 3,113 journals by 19,844 institutes in 108 different countries were retrieved. The number of articles published each year increased over time, with 1,570 published in 1999 and 6,256 published in 2013. The following countries published the most articles: USA (n=18,142), CHN (n=4,052), JPN (n=3,304), GBR (n=3,156), GER (n=2,935), and FRA (n=2,458). Forty-seven institutes in 10 countries/regions published more than 150 papers from 1999 to 2013. Keyword analyses show that, in descending order, the most frequently appeared organism categories are microorganisms, plants, vertebrates and invertebrates (7,719, 5,183, 3,720, and 2,066 keywords, respectively). Articles on the topic of plant, which has the highest annual average growth rate (AAGR) of 13.85% among four organism categories, accounts for 10.18% of total publication. Escherichia coli (2,492), Arabidopsis thaliana (2,078) and mouse (1,818) were the most frequently investigated. In conclusion, more than 50,000 genomic studies focusing on various bio-organisms and published in a wide range of journals have been performed in major economies over the past 15 years, and an increasing number of articles are being published in developing countries/regions and in several new journals.

Key words: Genomics, organism, publication, country, institution, bibliometrics.

INTRODUCTION

It has been 15 years since a draft sequence of the human genome was first released in 2001 (Lander et al., 2001; Venter et al., 2001). In 2003, the full human genome sequence was determined through the Human Genome Project (Collins et al., 2003). Since then, genome researchers have made great progress in investigations of human development and disease and in the general realm of life science. The genome database currently contains over 11,300 items with sequence and map data on various organisms (http://www.ncbi.nlm.nih.gov/genome, accessed on Feb 1, 2015). Genome views of 338 organisms, including vertebrates (147), invertebrates (26), protozoa...
Sequencing technology advancements have allowed researchers to obtain sequences faster and at lower costs than before (Venter et al., 2001; Drmanac et al., 2010). The first report on the complete viral genome (φX174, 5,375 nucleotides) was published in 1977 (Anderson et al., 1981). Sequencing a bacterial genome using the same techniques employed in this seminal work would take more than 100 years. However, 18 years later, the first complete bacterial genome (1,830,137 bp) was revealed because innovative methods, e.g., polymerase chain reaction (PCR), were applied (Fleischmann et al. 1995; Mullis et al., 1986). Moreover, reductions in the cost of gene sequencing have allowed researchers worldwide to engage in genome research using individual or public platforms (Drmanac et al., 2010).

The field of genomics has allowed scientists to examine data by employing a novel yet fundamental and comprehensive perspective. Studies that apply genome sequencing methods to clinical and biological research have emerged as a new research focus (Ginsburg, 2014). Genomes that may play a crucial role in human health and disease genesis are being identified and examined. According to the Online Mendelian Inheritance in Man (OMIM) database, over 3,300 genes with phenotype-causing mutation traits have been identified (http://www.omim.org/, accessed on Feb 1. 2015). These phenotypes include single-gene Mendelian disorders and exhibit susceptibility to complex diseases and genetic somatic cell diseases. Furthermore, researchers worldwide have moved beyond just using sequencing to capture raw data to applying genomic information to clinical care. Some studies have presented new tests and drug targets for gaining insight into the basis of human disease and new biomedical products (Ginsburg, 2014; Simo-Riudalbas and Esteller 2014; Lander, 2011).

To offer additional background information on the current state of genomic research, we examined scientific literature focusing on the genome that has been published from 1999 to 2013. The goal of this bibliometric analysis was to assess the development and outcomes of genomic research over the past 15 years by addressing the following questions. How many studies on organism genomes have been conducted? Which journals have genomic studies been published in, and how important are these articles (based on the impact factor)? What are the authorship characteristics of these publications? What are the most popular research topics within the field of genomics? This article summarizes the results of this bibliographic analysis.

**METHODS**

Based on the questions listed above, a search of the SCI-Expanded database on the Web of Science (http://isiknowledge.com, accessed on May 1, 2014) was carried out to collect primary research data for January 1, 1999, to December 31, 2013. With * serving as a wildcard character, the search phrase, i.e. “genome” and “genomic”, was employed for various titles or topics for the study period. After restricting the search results to articles only, retrieved entries were saved in an Endnote library. The entries were then categorized for analysis.

The following productivity indicators were used: total papers published, journal distributions, institutes by country/region and research focus, and time-dependent (annual, two-year, or five-year intervals) evaluations, which were analyzed as previously reported (Zou et al., 2009; Zou et al., 2012). The institute and country/region of the first author’s primary institutional affiliation was identified as the geographic location (excluding co-author institutes) where a paper was produced.

A journal’s impact factor (IF) score is a measure of a journal’s average number of citations relative to the number of articles published in the journal over a given period (Borgman, 2002). To assess the perceived quality of journals in which articles were published, the IF of each journal was extracted from the Journal Citation Reports (http://www.isiknowledge.com/jcr, 2012) without accounting for annual fluctuations in IF scores. Genomic article keywords can signify interesting areas of scientific research focus. One SCI-E topic words (keywords) from each publication were also categorized (vertebrate, invertebrate, microorganism, and plant) with the organism terms (Zou et al. 2009). Specific diseases that each paper examined were denoted by the term “disease” to identify changes in research focus over the study period.

Average annual growth rates (AAGR) were calculated as follows: assuming that year A is x and year B is y, $AAGR = (y/x)(1/(B−A))−1$.

**RESULTS**

**Outcomes of genomic articles for 1999-2013**

Figure 1 reveals the fluctuations in genome-related articles published each year, journals that published these articles, institutes involved in producing these articles and countries/regions where studies were conducted over the 1999-2013 period. Over the 15-year period, 3,113 journals published 50,933 genomic papers. The number of articles focusing on genomes increased from 1570 in 1999 to 6256 in 2013 at a AAGR of 10.38%. A small but constant increase in the number of journals that published genomic papers was also observed for this period (i.e., 432 in 1999, 1,093 in 2013). A total of 788 institutes based in 49 countries/regions produced genomic papers in 1999, and this number increased to 2,005 institutes based in 76 countries/regions by 2013.
It is evident that a decrease in the number of published papers focusing on genomes has been associated with an increase in the number of journals that have published such articles (Figure 2 A). A significant number of journals (1,123) only published one genome-related paper. Forty-nine percent of the journals published between 2 and 15 papers, whereas 414 journals published between 16 and 150 papers. Only 53 journals published more than 150 articles over the 15-year period (Figure 2 B). A number of journals showed a sharp increase in the number of genomic papers published in recent years (Figure 2 C). Whereas the Journal of Virology published only 16 papers in the field of genomics in 1999, in 2013, it published 767. Similarly, PLOS ONE published 58 articles in 2006 and 1,288 in 2013. Thus, the increase in papers published per journal may be largely attributed to a small proportion of journals that have exhibited a strong interest in this field.

Impact factors of journals that have published genomic papers

Each journal’s 2012 impact factor was used to determine genomic article quality levels for the 15-year period, and this period was divided into three shorter periods: 1999-2003, 2004-2008 and 2009-2013 (Figure 3). A smaller collection of journals with an impact factor higher than 20, published 306, 505, and 841 papers in the first, second and third periods, respectively. Of the 2,631 journals that indexed an impact factor higher than zero, 81 journals with an impact factor of ≥10 published 5,119 articles. On the other hand, 89.5% (43,638 of 48,757) of the papers were published in 2,550 journals (96.9%) with impact factors of less than 10. Moreover, it is worth noting that papers published in 6 journals with IF values of <10 [PLOS ONE (2,192 papers, IF 3.73), BMC Genomics (1,671, 4.40), Journal of Bacteriology (1,661, 3.19), Journal of Virology (1,346, 5.08), Nucleic Acids Research (1,298, 8.28) and PNAS_USA (1,122, 9.74)] accounted for roughly one-fifth of all of the articles examined. When trends were studied over five-year periods, i.e., 1999-2003, 2004-2008, and 2009-2013, the volume of high-IF articles published remained largely constant over the three periods, whereas an increase in the publication of low-IF (≥1~<5) articles occurred from 2009 to 2013. However, the total number of published articles gradually increased among all journal of various IF levels (Figures 1 and3).

Geographic locations of genomic paper publication

All countries showed an increase in the number of articles published over time (Figure 4). The USA is the founder of this field of research and continues to lead by a large margin in terms of the number of articles published. The USA consistently contributed 36% of all articles over the 15-year period. Whereas the total number of articles published quadrupled, some countries exhibited an increase in articles published that deviated from the increase in the total number of articles published. China
Figure 2. Analyses of journals that published genomic papers from 1999 to 2013. A. Scatter plots indicating the journals and the number of papers each journal published. The x- and y-axes represent log scales. B. Pie chart ranking journals by genomic paper publication. Eighty-five percent of all journals published 15 papers or fewer. C. Increases in genomic papers published in the top 10 journals over 15 years. Four journals, i.e., PLOS ONE, BMC Genomics, Journal of Bacteriology, and Journal of Virology, exhibited a sharp increase in genomic paper publication over the last four to eight years.

showed a 37-fold increase in articles published (24 in 1999 to 902 in 2013), increasing its proportion of published articles from 1.5 to 14.4%. Over the 15-year period, six countries (USA, n=18,142; CHN, n=4,052; JPN, n=3,304; GBR, n=3,156; GER, n=2,935; FRA, n=2,458) each contributed more than 5% of all articles published (Figure 4 B). As pioneers of research in this field, Japan, Britain, Germany and France were ranked second to fifth, respectively, in terms of the number of papers published in 1999, with consistent increases occurring over the 15-year period, although all of these countries showed declines in shares of publications. China, Britain, Germany and Japan were ranked second to fifth, respectively, in terms of the number of papers published in 2013. Although these pioneering countries continue to play an important role in genomic research, other countries such as China and Korea that were initially less prominent in this field have shown an increase in the percentage of articles published over the 15-year period.

A total of 19,844 institutes based in 108 countries have released articles on genomics. There is an observable difference between percentages of articles and institutions by geographic region of publication (Figure 5 A, B). The USA, for example, accounted for 25% of all institutes that have published genomic papers over the time period, but institutes based in this country generated 36% of all papers examined. We calculated the number of papers produced per institute in each country and observed values that vary from less than 1.5 to greater than 3.5. The USA and China generated the highest ratios (roughly 3.5), whereas India and Italy generated the lowest ratios (roughly 1.5). Figure 5 C shows that 47 institutes based in 10 countries have published 150 articles over the 15-year period, accounting for 21.4% of all publications considered. Most of these institutes are based in the USA (30 of 47). The Chinese Academy of Sciences (555), Harvard University (519) and Stanford University (385) published the highest numbers of genomic research papers.
Figure 3. Impact factors (IF) of journals that published genomic papers over various time periods. The 15-year period was divided into three periods (early, 1999-2003; middle, 2004-2008; late, 2009-2013). A. Scatter plots present the number of papers and journal IF values. No major variations were observed among the distributions of the three periods for papers with IF values higher than 20. B. Percentage composition of genomic papers of various IF values for the three time periods. A clear increase in papers with an IF ≥1 and <5 were published from 2009 to 2013.

Genomic paper focuses

To identify major research interests that characterized the 15-year period, the number of keywords related to organisms was analyzed and diseases found in the articles examined. The keywords were sorted and, when possible, coded these keywords were into various categories, i.e., vertebrate, invertebrate, microorganism, plant, and disease (Figure 6). Organism keywords were identified the following number of times: microorganisms (7,719), plants (5,183), vertebrates (3,720) and invertebrates (2,066). During 1999-2013, articles with keywords of vertebrate, invertebrate, microorganism and plant grew at AAGR of 8.37, 8.73, 8.55 and 13.85%, respectively. Articles on the topic of plant account for 40.00% of plant papers and 10.18% of total publication. In addition, keywords classified under the disease category consistently increased in number from 141 in 1999 to 608 in 2013, accounting for a AAGR of 11.0 and 9.5% of all publications. In line with publication trends, all keywords were used more frequently over time. An examination of individual keywords shows that Escherichia coli (2,492), Arabidopsis thaliana (2,078) and mouse (1,818) were the most prominent (Figure 7).
DISCUSSION

The field of genomics has evolved and progressed considerably since its inception. The Human Genome Project initiated by the USA Department of Energy/NIH in the 1990s triggered considerable interest in genomic research worldwide (Burris et al., 1998; Wiechers et al., 2013). Our data indicate a significant increase in the number of genomic articles and journals that have focused on genomic research, denoting a significant increase in research efforts dedicated to this field. An overall increase in the number of genomic papers published per journal also indicates that editorial boards are becoming more interested in publishing genomics research papers. It is interesting to note that four journals, i.e., PLOS ONE, BMC Genomics, Journal of Bacteriology, and Journal of Virology, have published 6,822 genomics articles and have published an increasing number over the last four to eight years. Of these journals, the output of PLOS ONE, established as open access journal in 2005, exhibited explosive growth over a matter of years.

The IF score is used to rank, evaluate and categorize journals and published papers. It is interesting to note that with rampant increases in output, genomic articles published in high-IF (>20) journals increased in number. However, the degree of total production remained constant.
Figure 6. Prominent organism categories found in each genomic paper over the 15-year period. Meanwhile, articles focusing on diseases are denoted by the red line.

Figure 7. Major organism keywords found in genomic papers published from 1999-2013.

over the various periods (Figure 3). By contrast, a distinct increase in the number of papers with low IF (≥1~<5) scores was observed. This trend is attributed to the fact that with periods of innovation and technological development, the majority of genomic research has since involved gene data mining of various organisms and less innovative results may be presented in lower IF journals or uploaded to specific databases.

The majority of genomic research is conducted in developed countries/regions, raising issues surrounding research funds. Indeed, Moses reported that American research funding and life science patents have accounted
for roughly half of the global share over the past decade (Moses et al., 2015). Similarly, genomic research is largely funded and directed by institutions based in developed countries. Most western countries have shown a stable increase in the number of genome studies published (Figure 4). Moreover, researchers based in the USA founded this field and continue to publish the most research papers in genomics. The country’s genomic publication contributions account for more than four times that of the next most prominent country, China, which has shown rapidly increasing trends over the past five years. On the other hand, a rampant increase in publications produced in “other” countries is also evident, with economically developing countries such as India anticipated to generate numerous scientific papers in the near future. In addition, the present study only recorded the first author’s institution as a measure of authorship. It must be recognized that many genomic studies are conducted cooperatively between researchers based in numerous countries.

Our findings concerning keywords and keyword categories indicate that a broad range of genome analyses have been carried out on plants, animals, and microbes and on relationships between genomics and diseases, and new technologies are emerging for the examination of diverse life forms. A particular focus has been placed on identifying Escherichia coli, Arabidopsis thaliana and mouse sequences. Papers of studying on A. thaliana account for 40% of articles on the topic of plant that has the fastest growing one among four organism categories. This increased interest reflects the wide range of diversity in life sciences research and also suggests that numerous researchers of various scientific disciplines engage in genomic studies.

In addition, genomics and other "omics" have reinvented the entire biomedical field and, from the laboratory to the clinic, have become an important facet of modern medicine. For example, analyses of gene disturbances in cancer tissues have led to the development of several promising drugs; individual responses to various drugs can be predicted; and breast cancer mutation testing is now common practice (Ginsburg, 2014). In this study, 9.5% publications were found to focus on diseases. It must be recognized that genomic research is conducted for various purposes; including drug development, gene diagnosis and gene therapy, and genomics technologies also play a decisive role in other biological fields, such as agriculture and fishing.

Bibliographic indicators have been deemed accepted indices of scientific activities. Because our primary data are SCI-E-based only, the number of publications on genome studies may have been underestimated. Despite this limitation, publications analyzed in the present study were retrieved from the most prominent journals on genomics, and appropriate conclusions were made accordingly. Tamames et al. (2013) found that for most publications on microbiology species, there has been no significant change in the number of citations. We did not examine citations of retracted papers in the current bibliographic study. However, even without considering citations, genome information is still important for the purposes of charting complete protein families of the biosphere (Perez-Iratxeta et al., 2007). Integrated information regarding genome studies provides us with an appropriate means of fighting diseases and of obtaining superior clinical biomarkers for prevention, diagnosis and prognosis. Although we have yet to successfully describe the genomes of all organisms, this study illustrates that genomic research became more prominent from 1999 to 2013.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (grant numbers. 81400381) (http://www.nsfc.gov.cn) and the State Key Laboratory of Ophthalmology, China.

REFERENCES


