academicJournals

Vol. 12(51), pp. 7043-7053, 18 December, 2013
DOI: 10.5897/AJB2013.12928
ISSN 1684-5315 ©2013 Academic Journals
http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

A bibliometric analysis of global research on genome sequencing from 1991 to 2010

Yidan Sun[1], Hui-Zhen Fu[2,3] and Yuh-Shan Ho[2,3]*

[1]Longping Branch, Graduate School of Central South University, Changsha, Hunan 410125, People's Republic of China.
[2]Trend Research Centre, Asia University, No. 500, Lioufeng Road, Wufeng, Taichung County 41354, Taiwan.
[3]Department of Environmental Sciences, Peking University, Beijing 100871, People's Republic of China.

Accepted 12 September, 2013

This study was carried out to evaluate the global scientific production of genome sequencing research to assess the characteristics of the research performances and the research tendencies. Data were obtained from Science Citation Index Expanded database during 1991-2010. Conventional methods including document types, journals, categories, countries and institutions were used to analyze publication output to reveal the global performance. The development of genome sequencing research during last 20 years was described by synthetically analyzing the distribution of words in article title, author keywords, and *KeyWords Plus* in different periods. The results show that disease and protein related researches were the leading research focuses, and comparative genomics and evolution related research had strong potential in the near future.

Key words: Genome sequencing, research trend, scientometrics, science citation index expanded (*SCI-Expanded*), word cluster analysis, keywords.

INTRODUCTION

Genome sequencing is a laboratory process to determine the order of chemical base pairs which make up DNA or RNA at a single time. Earlier attempts for genome sequencing research mainly concentrated on small genomes such as *Bacteriophage MS2* (Fiers et al., 1976) and *Phage Φ-X174* (Sanger et al., 1977). As the sequencing methods developed, researchers considered to take on longer and more complicated genomes (Edwards and Caskey, 1991; Roach et al., 1995). The first complete genome sequences for representatives from all three domains of life were released in mid-1990s including *Haemophilus influenzae* (Fleischmann et al., 1995), budding yeast *Saccharomyces cerevisiae* (Goffeau et al., 1996), and *Methanococcus jannaschii* (Bult et al., 1996). Lately in 2001, *Nature* and *Science* published a rough draft of the human genome, marked a milestone in genome sequencing history (Lander et al., 2001; Venter et al., 2001). With success in human genome, the sequencing of model organisms, including fruit flies (Adams et al., 2000), *Arabidopsis* (Kaul et al., 2000), Algae (Douglas et al., 2001), rice (Goff et al., 2002), microbial organisms (Stover et al., 2000; Ivanova et al., 2003), and parasites (Gardner et al., 2003) have been studied in the first five years following the human genome project (HGP). Furthermore, genomes from more organisms were sequenced in a faster pace after 2005 with dramatic leaps in sequencing technology and a preci-

*Correponding author. E-mail: ysho@asia.edu.tw. Tel: 886 4 2332 3456 ext. 1797. Fax: 886 4 2330 5834.

Abbreviations: HGP, Human genome project; **AIDS,** acquired immunodeficiency syndrome; ***SCI-Expanded*,** Science Citation Index Expanded; **JCR,** Journal Citation Reports; **IF,** impact factor; **SPA,** single country article; **SNPs,** single nucleotide polymorphisms.

pitous drop in costs (Mardis, 2011). As of October 2011, the complete sequences were available for: 2,719 viruses, 1,115 archaea and bacteria, and 36 eukaryotes (available in www.NCBI.com).

Despite the massive success of genome sequencing achieved in 21st century, there have been few attempts at gathering systematic data on genome sequencing research. A common research instrument for this analysis is the bibliometric method which has been widely used to measure scientific progress in many disciplines of science and engineering, such as acquired immunodeficiency syndrome (AIDS) (Macias-Chapula, 2000), and cancer molecular epidemiology (Ugolini et al., 2007). Moreover, the Science Citation Index Expanded (*SCI-Expanded*) database is used to analyze research performance from a more comprehensive perspective (Li et al., 2009). Conventional methods concerning bibliometrics mainly investigated the publication characteristics, including countries (Braun et al., 1995), institutions (Rodríguez and Moreiro, 1996), journals (Colman et al.,1995), and categories (Ugolini et al., 1997) may not be adequate to indicate the future orientation of research field (Chiu and Ho, 2007). More information, closer to the research itself, such as paper titles, author keywords, *KeyWords Plus*, and abstracts have been introduced (Xie et al., 2008; Li et al., 2009; Zhang et al., 2010) for the indepth information. Furthermore, an innovative method named "word cluster analysis" was successfully applied for finding the hotspots to evaluate research emphasis and trend (Mao et al., 2010).

In this study, bibliometric methods involving both the conventional and innovative ones were used to quantitatively and qualitatively assess the global performance and trend of genome sequencing research between 1991 and 2010. The results could give insights into the characteristics of the genome sequencing literature. More importantly, it could provide not only a potential guide for novice researchers, but also a basis for better understanding the global development tendency of genome sequencing research.

MATERIALS AND METHODS

The data were based on the online version of the *SCI-Expanded* database. According to Journal Citation Reports (JCR), it indexes 8,073 journals with citation references across 174 scientific disciplines in 2010. The database was searched using the keywords including "genome sequencing", "genome sequence", "genome sequences", "genome-sequenced", "genome sequency", and "genome sequencings" in terms of topic (title, abstract, author keywords, and *KeyWords Plus*) within the publication year limitation from 1991 to 2010. Document information including names of authors, title, abstract, author keywords, *KeyWords Plus*, address, year of publication, categories, and journals were downloaded into spreadsheet software.

Additional coding was performed manually for the number of origin country and institution of the collaborators, and impact factors of the publishing journals. Impact factors were taken from the JCR published in 2010. Articles originating from England, Scotland, Northern Ireland, and Wales were reclassified as being from the United Kingdom (UK) (Chiu and Ho, 2005). USSR and Russia were also reclassified as being from Russia. Articles from Hong Kong published before 1997 were included in the China category (Chuang et al., 2011). Collaboration type was categorized and determined by the addresses of the authors as: Single country articles with addresses from the same country; internationally collaborative articles with author addresses from more than one country or territory (Li et al., 2009); single institution articles with addresses from the same institution; and inter-institutionally collaborative articles with author addresses from more than one institution (Malarvizhi et al., 2010). All keywords, both those reported by authors and those attributed by the Web of Science, as well as words in title and abstract were identified and separated into 5-year span (1991-1995, 1996-2000, 2001-2005, and 2006-2010). Then their ranks and frequencies were calculated. A word cluster analysis was performed in the combination of the words in titles, author keywords, *KeyWords Plus*, and words in abstracts, in which different words with identical meaning and misspelled keywords were grouped and considered as a group for one research focus (Li et al., 2009; Mao et al., 2010).

RESULTS AND DISCUSSION

Altogether 20,462 publications consist of 16 document types. Articles (15,722) dominate with the highest percentage of 77%, followed by reviews (2,984; 15%), proceedings paper articles (859, 4.2%), and editorial materials (396; 1.9%). The other 12 documents types with the percentages less than one percent were meeting abstracts, news items, letters, corrections, notes, software reviews, book chapter articles, reprints, database reviews, addition corrections, and biographical-item. Only 15,722 journal articles were extracted for subsequent analyses for its dominant position and including whole research ideas and results (Ho et al., 2010).

Publication outputs

The annual number of articles is shown in Figure 1. The annual number of articles increased nearly 100 times from only 22 articles in 1991 to nearly 2,000 articles in 2010. To be specific, the annual number of articles first exceeded 100 in 1996, and rocketed over 1,000 in 2003. The development in the past two decades could be primarily attributed to the strong support of HGP in 1990. HGP, regarded as the third massive science project after Manhattan Project and Apollo Project, was invested three billion dollars by the US Department of Energy and the National Institutes of Health (Barnhart, 1989). The massive government concern and economic investment strongly prompted to the development of genome sequencing area (Lander, 1996). Another possible reason for the fast increase is that alternative sequencing methods and instruments have been produced to reduce time and cost (Margulies et al., 2005). In the 21st century, fierce commercial competition will force manufacturers to create new faster and cheaper sequencing machines, which will benefit genome

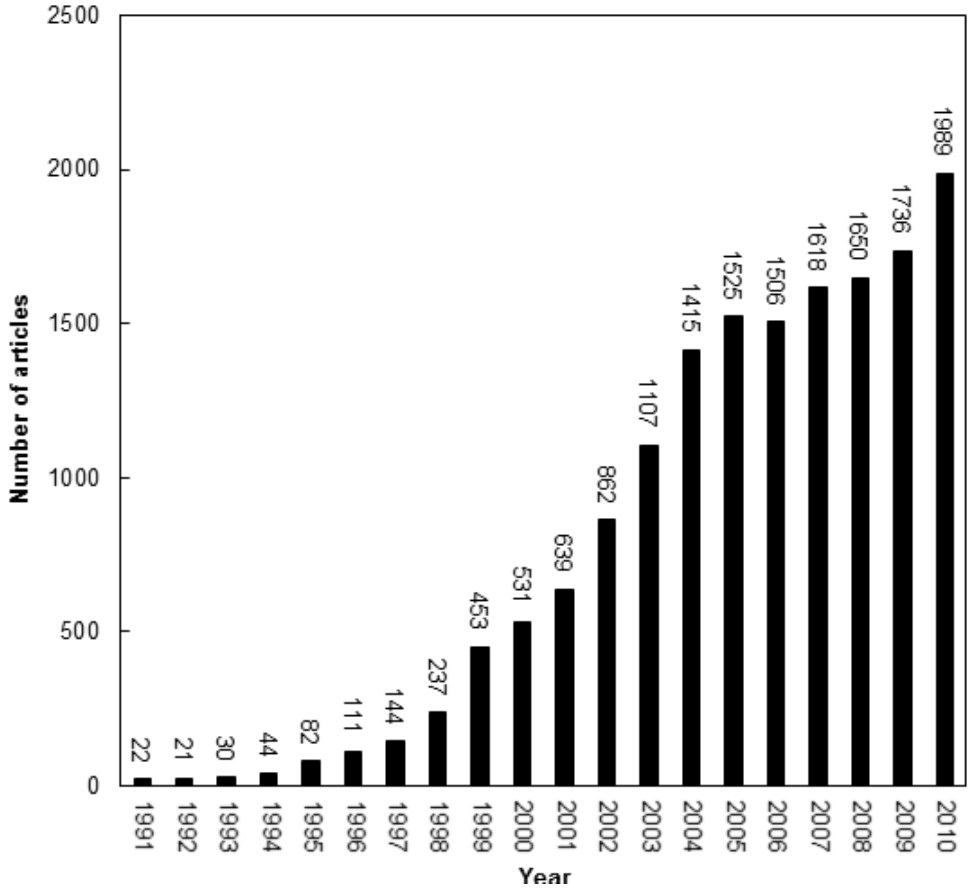

Figure 1. The growth trends of annual articles in genome sequencing research.

sequencing area deeper (Mardis, 2009).

Journals and Web of Science categories

The total articles (15,722) were published in 1,308 journals, among which, 555 (42%) journals contained only one article and 206 (16%) contained two. Table 1 shows 13 core journals that published more than 200 articles, accounting for one third of all articles. *Journal of Bacteriology* ranked first with 1,043 papers (6.6%), followed by *BMC Genomics* with 471 articles. However, the percentage of the top journal was not high, which indicated the breadth of genome sequencing research as well as the broad interest in genome sequencing from various research angles (Li et al., 2011; Wolfe and Li, 2003). In addition, as regards to the impact factor (IF) for all journals, *New England Journal of Medicine Proceedings of National Academy of Sciences of the United States of America, Nucleic Acid Research, and Genome Research* which ranked 3rd, 4th and 6th in the number of total articles, respectively, having the IFs greater than seven. Since the IF is used to evaluate a journal's relative importance of one field, these IF statistical results will help researchers select journals when publishing articles on genome sequencing related research (Ho, 2008).

Based on the classification of Web of Science categories in JCR in 2010, the publication output data was distributed in 141 Web of Science categories in science edition. Biochemistry and molecular biology, microbiology, genetics and heredity, and biotechnology and applied microbiology were the four most popular categories, which exceeded the other categories in both the cumulative number and the annual number. It is noticable that the category of biochemistry and molecular biology held primacy from 1991 to 2005, but started to decrease rapidly from 2006; while microbiology grew fast and became the first in 2010. Growth trends also appeared in the two categories of biotechnology and applied microbiology, and genetics and heredity. The annual number of articles of genetics and heredity increased continually in the study period and exceeded that of biochemistry and molecular biology after 2006. The shifting position of categories indicates that the mainstream of research is no longer restricted to the original one (Wolfe and Li, 2003), and more attention of mechanism has been transferred to the application in genome sequencing related research.

Table 1. The 13 core journals on genome sequencing, including the ranking, percentages, impact factors.

Journal	IF2010	TP (%)	Web of Science categories	Rank
Journal of Bacteriology	3.726	1,043 (6.6)	Microbiology	25/107
BMC Genomics	4.206	471 (3.0)	Biotechnology and applied microbiology	24/160
			Genetics and heredity	34/156
Proceedings of the National Academy of Sciences of the United States of America	9.771	458 (2.9)	Multidisciplinary sciences	3/59
Nucleic Acids Research	7.836	448 (2.8)	Biochemistry and molecular biology	30/286
Applied and Environmental Microbiology	3.778	431 (2.7)	Biotechnology and applied microbiology	32/160
			Microbiology	24/107
Genome Research	13.588	352 (2.2)	Biochemistry and molecular biology	8/286
			Biotechnology and applied microbiology	3/160
			Genetics and heredity	6/156
Molecular Microbiology	4.819	336 (2.1)	Biochemistry and molecular biology	60/286
			Microbiology	16/107
Infection and Immunity	4.098	325 (2.1)	Immunology	33/134
			Infectious diseases	11/58
Microbiology-SGM	2.957	310 (2.0)	Microbiology	39/107
Journal of Biological Chemistry	5.328	299 (1.9)	Biochemistry and molecular biology	50/286
PLoS One	4.411	238 (1.5)	Biology	12/86
FEMS Microbiology Letters	2.040	222 (1.4)	Microbiology	62/107
Journal of Virology	5.189	211 (1.3)	Virology	5/33

IF2010: impact factor in 2010; TP: number of total articles.

National and institutional contributors

Each author of an article has made an independent contribution to the manuscript (Coats, 2009), and therefore the institution and country the author affiliated could be consider the important contributors for the evaluation of research. Publication counts of countries is a reference for evaluating countries/territories research performance in a field, and has been used in many aspects of research such as tsunami (Chiu and Ho, 2007) and risk assessment (Mao et al., 2010). Excluding 35 articles with no author address information on the Web of Science, the 15,687 articles originated from 139 countries/territories. The distribution of the genome sequencing articles all around the world is displayed in Figure 2. America, West Europe, Japan, and China were the main production areas. Of all the 15,687 articles with author information, 4,629 (30%) were international collaborative publications and 11,058 (70%) were independent publications. The international collaborative rate of genome sequencing research is higher than that in certain studies, such as 14% biosorption technology for water treatment (Ho, 2008) and acupuncture research (Han and Ho, 2011), 16% of desalination research (Tanaka and Ho, 2011) and solid waste (Fu et al., 2010). Table 2 reveals the characteristics of the top 20 productive countries. Five indicators including the number of total articles, single country articles, internationally collaborative articles, first author articles, and corresponding author articles were displayed. The table also presented the percentage of independent articles in total articles, total articles per number of authors, and single country articles per number of authors. The USA ranked top one by all indicators. Single country articles were authored by 67 different countries, and 27 countries contributed less than ten single country articles. Furthermore, the developing countries such as India (36%), Brazil (30%), Russia (36%) had relatively low percentages of single country articles (%S); while developed countries USA and UK (%S = 61%) were more inclined or able to conduct research independently. It also appeared that the lowest average number of authors per total article (TPA) and single country article (SPA) was found to be 7.1 and 2.5 authors per article for India, while Italy, Brazil and Belgium had higher value over 14 for TPA and 4.5 for SPA.

The growth trends of the top eight productive countries are displayed in Figure 3. The USA was also dominant in the annual production, ranked first every year except 2005.

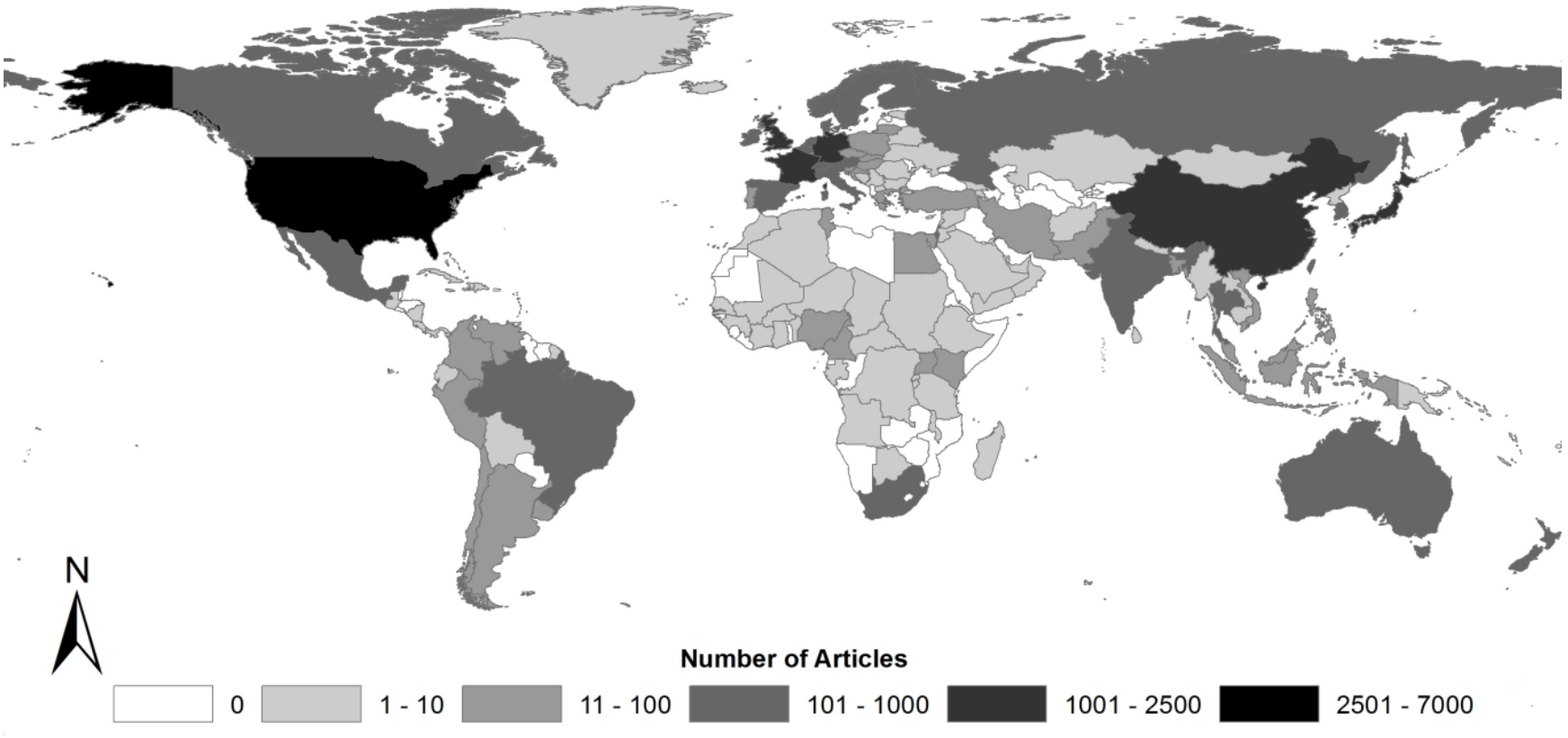

Figure 2. Distribution of genome sequencing articles in the world.

The elite performance of the USA may be due to its greater economic investment in biotech industry than other countries. For example, there were more than 1,500 biotech companies in Europe and approximately 1,300 in the United States in 2001, but revenues for European biotechs were less than one-third that of US revenues (Nagle et al., 2003). Notably, China published less than ten articles before 2001, but the annual number of articles grew sharply and ranked 4th in 2010. China is also the only one developing country which was involved in the top eight most productive countries/ territories (Figure 3). The outstanding energy of China was not surprising, because as reported, it not only experienced a sustained and remarkable increase in scientific production, as the world's second largest producer of scientific publications since 2006 (Zhou and Leydesdorff, 2008), but has also been taking a world-leading position in various fields, such as chemistry (Zhou and Leydesdorff, 2009), and nanotube technology (Kostoff, 2012). Of 15,687 articles with author addresses in Web of Science, 9,796 (62%) were inter-institutionally collaborative articles, while 5,891 (38%) were institutionally independent articles. The percentage of collaboration among institutions (62%) was twice more than that among countries (30%). The inter-institutionally collaborative rate was equal to that of global climate change with 62% (Li et al., 2011), but was larger than 53% of acupuncture research (Han and Ho, 2011), 44% of solid waste research (Fu et al., 2010), and 37% of desalination research (Tanaka and Ho, 2011). As for the top 10

Table 2. Characteristics of the top 19 productive countries/territories (TP □ 200).

Country	TP	TPR (%)	SPR (%)	CPR (%)	FPR (%)	RPR (%)	S%	TPA	SPA
USA	6,607	1 (42)	1 (37)	1 (55)	1 (34)	1 (34)	61	8.2	5.3
UK	2,016	2 (13)	3 (7.4)	2 (26)	3 (8.5)	3 (8.5)	61	11	5.1
Germany	1,728	3 (11)	4 (6.7)	3 (21)	4 (7.3)	4 (7.2)	47	12	4.8
Japan	1,649	4 (11)	2 (11)	7 (8.9)	2 (9.1)	2 (9.0)	45	8.3	5.7
France	1,412	5 (9.0)	6 (5.5)	4 (17)	5 (6.0)	5 (5.9)	46	11	5.8
China	1,079	6 (6.9)	5 (5.8)	6 (9.4)	6 (5.4)	6 (5.4)	56	10	7.0
Canada	888	7 (5.7)	7 (3.6)	5 (11)	7 (3.6)	7 (3.6)	44	10	4.3
Australia	600	8 (3.8)	10 (2.1)	8 (8.0)	8 (2.4)	8 (2.4)	41	11	4.5
Netherlands	509	9 (3.2)	12 (1.6)	9 (7.2)	10 (1.9)	10 (1.9)	46	12	5.1
Spain	457	10 (2.9)	11 (1.6)	10 (6.0)	11 (1.7)	11 (1.7)	50	12	4.4
Italy	411	11 (2.6)	13 (1.3)	11 (5.7)	13 (1.6)	13 (1.6)	43	14	5.9
South Korea	402	12 (2.6)	8 (2.2)	17 (3.3)	9 (2.0)	9 (2.0)	44	11	6.1
Sweden	356	13 (2.3)	15 (1.0)	13 (5.2)	15 (1.3)	15 (1.3)	41	12	3.6
Switzerland	335	14 (2.1)	18 (0.78)	12 (5.4)	16 (1.1)	16 (1.1)	40	16	4.2
India	317	15 (2.0)	9 (2.1)	24 (1.8)	12 (1.7)	12 (1.7)	36	7.1	2.5
Brazil	302	16 (1.9)	14 (1.2)	15 (3.6)	14 (1.3)	14 (1.3)	30	15	7.0
Belgium	293	17 (1.9)	17 (0.80)	14 (4.4)	17 (1.0)	17 (1.0)	30	15	4.9
Denmark	240	18 (1.5)	19 (0.65)	15 (3.6)	18 (0.76)	18 (0.74)	36	19	3.8
Russia	202	19 (1.3)	22 (0.55)	18 (3.0)	20 (0.62)	21 (0.63)	36	8.8	4.2

TP, Number of articles; TPR, the rank of total articles; SPR, the rank of single institution articles; CPR, the rank of inter-institutionally collaborative articles; FPR, the rank of first author articles; RPR, the rank of corresponding author articles; S%, the percentage of single institution articles in each institution; TPA, total articles per number of authors; SPA, single country articles per number of authors.

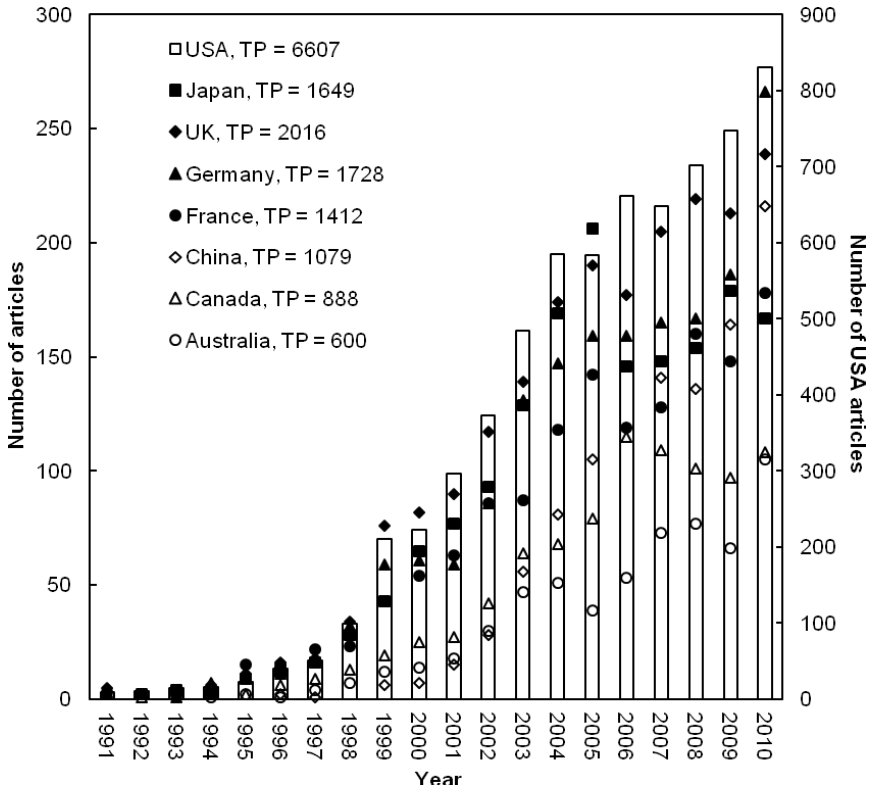

Figure 3. Comparison the growth trends of the top eight productive countries during 1991-2010.

Table 3. Characteristics of the top ten productive institutions.

Institution	TP	TPR (%)	SPR (%)	CPR (%)	FPR (%)	RPR (%)	S%
Chinese Academy of Sciences, China	289	1 (1.8)	7 (0.78)	2 (2.5)	2 (1.0)	1 (1.0)	22
University of Tokyo, Japan	288	2 (1.8)	1 (1.1)	3 (2.3)	1 (1.0)	1 (1.0)	22
University of California, Berkeley, USA	273	3 (1.7)	34 (0.42)	1 (2.5)	23 (0.49)	23 (0.47)	21
Institut Pasteur, France	264	4 (1.7)	12 (0.70)	3 (2.3)	3 (0.85)	3 (0.84)	19
INRA, France	263	5 (1.7)	6 (0.81)	6 (2.2)	4 (0.83)	4 (0.82)	19
Harvard University, USA	258	6 (1.6)	13 (0.68)	5 (2.2)	6 (0.68)	6 (0.67)	19
CNRS, France	247	7 (1.6)	19 (0.58)	7 (2.2)	16 (0.52)	16 (0.52)	19
University of Maryland, USA	243	8 (1.5)	15 (0.65)	8 (2.1)	5 (0.75)	5 (0.77)	18
University of California, Davis, USA	218	9 (1.4)	31 (0.44)	9 (2.0)	32 (0.4)	31 (0.41)	20
Washington University, USA	213	10 (1.4)	8 (0.75)	12 (1.7)	7 (0.65)	9 (0.63)	21

TP, Number of articles; TPR, the rank of total articles; SPR, the rank of single institution articles; CPR, the rank of internationally collaborative articles; FPR, the rank of first author articles; RPR, the rank of corresponding author articles; S%, the percentage of single institution articles in each institution.

institutions, a half of them were located in the USA and three were in France (Table 3). The USA, the UK, Germany, Japan, and France were the top five most productive countries. However, from Table 3, no institutions in Germany and UK could be found. Chinese Academy of Science ranked first in the total number of publications, but there is a bias because it is made up of many relatively independent institutions distributed throughout China. At present, the articles of these branches were pooled under one heading, and rankings would be different if these branches are counted as independent ones (Li et al., 2009). Thus, except Chinese Academy of Science, the leading was University of Tokyo in Japan, which also ranked first in the single insitution articles, first author articles and corresponding author articles (Table 3). University of California, Berkeley ranked first in internationally collaborative articles, but stood relatively low positions in the single institutions articles, first author articles and corresponding author articles.

Distribution of author keyword analysis

Author keywords are the words that expose the internal structure of an author's reasoning, and are used in a specific period as a bibliometric method (Chiu and Ho, 2007). Using the author keywords to analyze the trend of research is much more frequent in recent years, and proved to be important for monitoring development of science and programs (Xie et al., 2008; Li et al., 2009). Analysis of author keywords revealed that 18,030 author keywords were used from 1991 to 2010, of which 14,173 (79%) appeared only once and 1,828 (13%) appeared only twice. These once or twice only author keywords might not be standard or widely accepted by researchers (Ugolini et al., 2001). Author keywords appeared in the articles referring to genome sequencing were calculated and ranked by total 20-year period and four 5-year sub-periods (Table 4). The most frequently used keywords were identified, such as "evolution", "phylogeny", and "comparative genomics". The analysis of author keywords revealed a notable growth trend in "phylogeny", "comparative genomics", "bioinformatics", "phylogenetic analysis", "proteomics", and "functional genomics". Phylogeny is a discipline describing evolutionary relationships. It is a remarkable fact that the rank and percentage of "phylogeny" and "phylogenetic analysis" rose from 10th (1.9%), 17th (1.3%) during 1996-2000 to 1st (3.5%) and 6th (1.9%) during 2006-2010, indicating phylogeny research has been greatly prompted by the abundance of genome sequencing data (Wolfe and Li, 2003). The amount of phylogenetic information will be immense as the degree of similarities and differences between gene sequences is used as one of the most common and reliable methods to perform phylogenetic analysis, and more organisms' gene sequences information will be available from genome sequencing (Brooker, 1999; Wolfe and Li, 2003). Comparative genomics has become one of the most powerful strategies for analyzing genome sequencing data (Nelson and Cox, 2008), and its rank in author keywords increased to 1st in 2001-2005. Meanwhile, the rank and percentage of articles with "proteomics" and "functional genomics" went up respectively, which did not show up during 1991-1995, rose to 8th (1.9%) and 13th (1.4%) during 2001-2005. Proteomics is functional genomics at the protein level (Anderson and Anderson, 1998). Better understanding of proteomics would greatly aid the biological interpretation of the genome sequencing data and accelerate its medical exploitation (King et al., 2000; Weinberg, 2010).

Distribution of article titles, *KeyWords Plus*, and abstracts analysis

Article title, which always contained the information of the whole paper, is a useful tool to evaluate trend recently (Xie et al., 2008; Zhang et al., 2010). All the single words

Table 4. Top 20 most frequently used author keywords during 1991-2010 and four five-year sub-periods.

Author keyword	TP	91-10 Rank (%)	91-95 Rank (%)	96-00 Rank (%)	01-05 Rank (%)	06-10 Rank (%)
Genome sequencing	217	1 (3.2)	1 (41)	1 (14)	11 (1.5)	10 (1.4)
Evolution	211	2 (3.1)	9 (3.0)	7 (2.7)	3 (3.4)	2 (3.0)
Phylogeny	211	2 (3.1)	N/A	10 (1.9)	5 (2.9)	1 (3.5)
Comparative genomics	203	4 (3.0)	N/A	29 (0.94)	1 (4.1)	3 (2.7)
Genome	190	5 (2.8)	32 (1.0)	8 (2.4)	4 (3.1)	3 (2.7)
Bioinformatics	157	6 (2.3)	N/A	37 (0.78)	2 (4.0)	8 (1.6)
Genome sequence	152	7 (2.2)	5 (4.0)	8 (2.4)	6 (2.6)	5 (1.9)
Genomics	120	8 (1.8)	32 (1.0)	6 (2.8)	10 (1.6)	7 (1.7)
Archaea	105	9 (1.5)	N/A	4 (4.7)	9 (1.9)	19 (0.86)
Phylogenetic analysis	103	10 (1.5)	N/A	17 (1.3)	18 (1.0)	6 (1.9)
Gene expression	101	11 (1.5)	15 (2.0)	29 (0.94)	7 (2.0)	14 (1.3)
Proteomics	100	12 (1.5)	N/A	29 (0.94)	8 (1.9)	11 (1.3)
Bacillus subtilis	92	13 (1.4)	3 (17)	2 (7.5)	27 (0.91)	281 (0.16)
Microarray	78	14 (1.1)	N/A	N/A	15 (1.3)	11 (1.3)
Virulence	72	15 (1.1)	N/A	324 (0.16)	23 (1.0)	11 (1.3)
Rice	70	16 (1.0)	N/A	120 (0.31)	19 (1.0)	15 (1.2)
Saccharomyces cerevisiae	70	16 (1.0)	2 (19)	3 (6.0)	123 (0.30)	281 (0.16)
Mass spectrometry	66	18 (1.0)	32 (1.0)	29 (0.94)	17 (1.1)	18 (0.91)
Escherichia coli	62	19 (0.91)	N/A	14 (1.4)	23 (1.0)	21 (0.83)
Functional genomics	62	19 (0.91)	N/A	17 (1.3)	13 (1.4)	40 (0.59)

TP, Number of articles; N/A, not available.

in the title of genome sequencing related articles were statistically analyzed in this study. Some words which have no usefulness for the analysis of research trend were discarded such as prepositions, conjunctions. “Protein”, “virus”, “evolution”, “comparative”, and “proteins” presented in the 20 most frequently used keywords in title also appeared in the top 20 of author keywords. Mean while, the rank and percentage of “evolution” increased steeply from 106th (1%) during 1991-1995 to 14th (4.2%) during 2006-2010, similar to the results of analysis of author keywords, from 9th (3%) during 1991-1995 to 2nd (3%) during 2006-2010.

However, there is a disparity that authors might choose their title words to attract a more general or particular audience (Peters and van Raan, 1994). As a supplement, an abstract appeared as it is a brief summary of a research paper of any in-depth analysis of a particular subject or discipline, and is often used to help the reader quickly ascertain the subjective focus and emphasis specified by authors (Zhang et al., 2010). Through key words analysis in abstracts, it can be concluded that continual attention was given to “protein” and “proteins”, whose rank is 6th and 8th, respectively. Proteins have been paid much attention as they perform most life functions and even make up the majority of cellular structures (Nelson and Cox, 2008). Again, “phylogenetic” showed a notable increasing trend in genome sequencing research, rising from 156th (6.4%) during 1991-1995 to 23rd (16%) during 2006-2010.

In recent years, *KeyWords Plus* were separated into different year periods to analyze the variations of trends on research topics (Xie et al., 2008). *KeyWords Plus* can provide additional search terms extracted from the titles of articles cited by authors in their bibliographies and footnotes in the ISI database, thus to augment title words and author keywords indexing (Garfield, 1990). As with the distribution of article titles, abstracts, and author keywords, “evolution”, “protein”, and “proteins” were emphasized in *KeyWords Plus* analysis. Moreover, “SNP” and “SNPs” exhibited growth trends in *KeyWords Plus* as well as other three kinds of keywords analysis. So far, over 1.4 million locations where single nucleotide polymorphisms (SNPs) occur in humans have been identified (Sachidanandam et al., 2001). This will allow genome-wide, high-resolution analysis of amplifications and deletions, and means significantly due to the fact that genetic variants can be examined for association with phenotypes and interpreted in clinical settings (Lander, 2011).

Research emphases and trends

The distribution of words in the article title and abstract, author keywords, and *KeyWords Plus* in different periods could provide important information for research emphases. Each research emphase related synonymic single words and congeneric phrases were summed and grouped

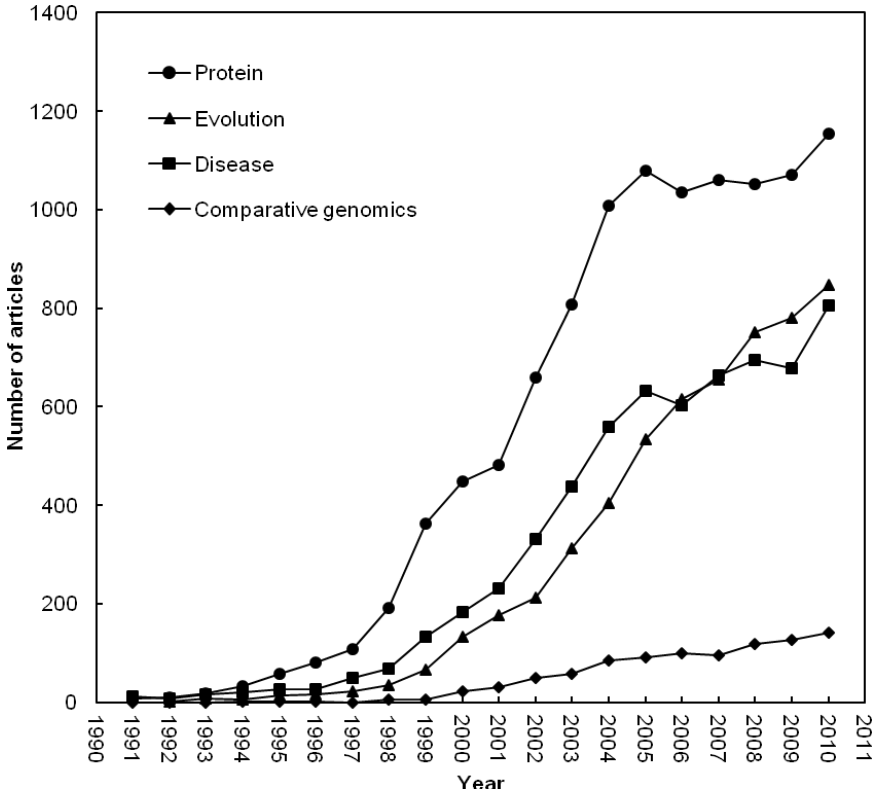

Figure 4. Growth trends of hotspot-related articles during 1991-2010.

into word clusters, so as to analyze the historical development of the science more completely and precisely, and more importantly, to discover the directions the science is taking. It was an innovative method that has been successfully used to analyze the research hotspot in several fields of science (Mao et al., 2010; Tanaka and Ho, 2011). Each hot issue in Figure 4 was supported by a word cluster, which was composed of several supporting words, including their plural forms, abbreviation, other transformations, as well as near synonyms. For example, the topic "comparative genomics" included "comparative (-) genomic (s)", "array comparative", "genome-wide association (s)", "single (-) nucleotide polymorphism (s)", "SNP (s)", and "array CGH", "blast". The growth trends of the articles concerning the supporting word clusters are displayed in Figure 4. Research emphases in genome sequencing were extracted and separated into four topics: protein, disease, evolution, and comparative genomics, among which, the topic "protein" was the most attractive. The predominant position is mainly attributed to that the success of genome sequencing speeded up protein primary structure study (Nelson and Cox, 2008). In addition, the large amount of genomic data were available for a variety of organisms facilitates proteome development, and has brought an urgent need for systematic proteomics to decipher the encoded protein networks that dictate cellular function (Ho et al., 2002). The rise in the study of disease in the field of genome sequencing started later than those of protein and is relatively slow after 2005. It indicates that although genome sequencing provides new avenues for disease genes discovery, the application is limited as gene findings from genome sequencing studies failed to explain much of the heritability of the diseases being studied (Maher et al., 2008; Ioannidis et al., 2008). Thus, it is expected that the following research in this field will turn to search evidences for the findings and elucidate the underlying mechanisms of disease in the next decade.

More attention was paid to the research on "evolution", especially after 1999. The number of articles related to "evolution" already exceeded that of "disease" in 2007. Along with gene sequences provided by whole genome sequencing, gene duplication and horizontal gene transfer (also called lateral gene transfer) are predicted to be the most familiar of the gene formation mechanisms and probably accounts for most new genes (Yang, 1998; Suzuki and Gojobori, 1999; Suzuki and Nei, 2001; McLysaght et al., 2002). Thus genome sequencing data have had an impact on the explanation for gene evolution at the scale of molecular. As the cost of genome sequencing falls and the capacity of sequencing centers grows, genome sequencing data will also allow the evolution of regulatory

elements studied in unprecedented detail (Lander, 2011; Mardis, 2011). Studies of regulatory elements will lead to answer more mysterious questions about of evolution since King and Wilson (1975) suggested that evolution of species depends more on innovation in regulatory sequences than changes in gene sequences.

Like the topic “evolution”, the research on comparative genomics increased remarkably in 21st century. Interspecific and intraspecific comparative genomics have widely been applied in many aspects, such as annotated the genome (Birney et al., 2007), and identified DNA variation including SNP (Kingsmore et al., 2008). The application of comparative genomics has increased the introduction of different ideas into genomics area, including concepts from systems and control, information theory, made the genome sequencing data better utilized (Via et al., 2011). Therefore, it is expected that research on comparative genomics will grow constantly since the genome sequencing data will be larger and the utilization will be more concerned in the next decades.

Conclusion

To gain a clearer insight into research focus and forecast on genome sequencing field, bibliometric analyses of annual publication outputs, journals, categories, countries, institutions, author keywords, title words, abstract words, and *KeyWords Plus* provide a synthetical overview. A total of 15,722 genome sequencing *SCI-Expanded* articles were analyzed over a period from 1991 to 2010. A fast increase was observed in the study period. ‘Journal of Bacteriology’ led the total 1,308 journals in the 141 Web of Science categories. Genome sequencing research tends to be utilized in a wide extent of areas. Shiftings among top categories indicated that the attention on application has been getting more popular. The national collaboration occurred more in genome sequencing in comprison with other fields. The USA held primacy using relatively less people per article, while China with a great growth rate was the most productive one among developing countries. University of Tokyo in Japan was actually the lead among the institutions. Moreover, the comprehensive analysis of author keywords, title words, abstract words, and *KeyWords Plus* provide important clues to word cluster for research emphases. The newly developed bibliometric method, “word cluster analysis”, can help researchers realize the panorama of global genome sequencing research and establish future research directions. Disease and protein related research obtained stable focus on a high degree in this field. The issues “comparative genomics” and “evolution” were active during the study period and will deserve increasing concern in the future.

REFERENCES

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et al. (2000). The genome sequence of Drosophila melanogaster. Aaohn J. 287 (5461):2185-2195.

Anderson NL, Anderson NG (1998). Proteome and proteomics: New technologies, new concepts, and new words. Electrophoresis 19(11): 1853-1861.

Barnhart BJ (1989) DOE human genome program, Human Genome Quarterly.

Birney E, ENCODE Project Consortium,Stamatoyannopoulos JA, Dutta A, Guigó R et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146): 799-816.

Braun T, Glänzel W, Grupp H (1995). The scientometric weight of 50 nations in 27 science areas, 1989-1993. Part I. All fields combined, mathematics, engineering, chemistry and physics. Scientometrics 33(3): 263-293.

Brooker RJ (1999), Genetics: analysis and principles, Addison-Wesley.

Bult CJ, et al. (1996). Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273(5278): 1058-1073.

Chiu WT, Ho YS (2005). Bibliometric analysis of homeopathy research during the period of 1991 to 2003. Scientometrics 63(1): 3-23.

Chiu WT, Ho YS (2007). Bibliometric analysis of tsunami research. Scientometrics 73(1): 3-17.

Chuang KY, Wang MH, Ho YS (2013), High-impact papers published in journals listed in the field of chemical engineering. Malays. J. Libr. Sci. 18(2): 47-63.

Coats AJS (2009). Ethical authorship and publishing. Int. J. Cardiol. 131(2): 149-150.

Colman AM, Dhillon D, Coulthard B (1995). A bibliometric evaluation of the research performance of British university politics departments: Publications in leading journals. Scientometrics 32(1): 49-66.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu XN, Reith M, Cavalier-Smith T, Maier UG (2001). The highly reduced genome of an enslaved algal nucleus. Nature 410(6832): 1091-1096.

Edwards A, Caskey CT (1991). Closure strategies for random DNA sequencing. Methods 3(1): 41-47.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Minjou W, Molemans F, Raeymaekers A, Vandenberghe A, Volckaert G, Ysebaert M (1976). Complete nucleotide sequence of bacteriophage MS2-RNA: Primary and secondary structure of replicase gene. Nature 260(5551): 500-507.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269(5223): 496-512.

Fu HZ, Ho YS, Sui YM, Li ZS (2010). A bibliometric analysis of solid waste research during the period 1993-2008. Waste Manage. 30(12): 2410-2417.

Gardner MJ, Hall N, Fung E, White O, Berriman M et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419(6906): 498-511.

Garfield E (1990). KeyWords Plus: ISI’s breakthrough retrieval method. Part 1. Expanding your searching power on Current Contents on Diskette. Curr. Contents 32: 325-9.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp japonica). Science 296(5565): 92-100.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996). Life with 6000 genes. Science 274(5287): 546-567.

Han JS, Ho YS (2011). Global trends and performances of acupuncture research. Neurosci. Biobehav. Rev. 35(3): 680-687.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868): 180-183.

Ho YS (2008). Bibliometric analysis of biosorption technology in water treatment research from 1991 to 2004. Int. J. Environ. Pollut. 34(1-4): 1-13.

Ho YS, Satoh H, Lin SY (2010). Japanese lung cancer research trends and performance in Science Citation Index. Intern. Med. 49(20): 2219-2228.

Ioannidis JP, Boffetta P, Little J, O'Brien TR, Uitterlinden AG et al. (2008). Assessment of cumulative evidence on genetic associations: Interim guidelines. Int. J. Epidemiol. 37(1): 120-132.
Ivanova N, Sorokin A, Anderson I, Galleron N et al. (2003). Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature 423(6935): 87-91.
Kaul S, et al. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408(6814): 796-815.
King MC, Wilson AC (1975). Evolution at two levels in humans and chimpanzees. Science 188(4184): 107-116.
King RD, Karwath A, Clare A, Dephaspe L (2000). Genome scale prediction of protein functional class from sequence using data mining. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edited by Ramakrishnan R, Stolfo S), Proceedings. KDD-2000. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 384-389.
Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD (2008). Genome-wide association studies: Progress and potential for drug discovery and development. Nat. Rev. Drug Discov. 7(3): 221-230.
Kostoff RN (2012). China/USA nanotechnology research output comparison-2011 update. Technol. Forecast. Soc. Chang. 79(5): 986-990.
Lander ES (1996). The new genomics: Global views of biology. Science 274(5287): 536-539.
Lander ES (2011). Initial impact of the sequencing of the human genome. Nature 470(7332): 187-197.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al. (2001). Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.
Li JF, Wang MH, Ho YS (2011). Trends in research on global climate change: A Science Citation Index Expanded-based analysis. Glob. Planet. Change 77(1-2): 13-20.
Li LL, Ding GH, Feng N, Wang MH, Ho YS (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. Scientometrics 80(1): 39-58.
Macias-Chapula CA (2000). AIDS in Haiti: A bibliometric analysis. Bull. Med. Libr. Assoc. 88(1): 56-61.
Maher B (2008). Personal genomes: The case of the missing heritability. Nature 456(7218): 18-21.
Malarvizhi R, Wang MH, Ho YS (2010). Research trends in adsorption technologies for dye containing wastewaters. World Appl. Sci. J. 8(8): 930-942.
Mao N, Wang MH, Ho YS (2010). A bibliometric study of the trend in articles related to risk assessment published in Science Citation Index. Hum. Ecol. Risk Assess. 16(4): 801-824.
Mardis ER (2009). New strategies and emerging technologies for massively parallel sequencing: Applications in medical research. Genome Med. 1(4): Article Number:40.
Mardis ER (2011). A decade's perspective on DNA sequencing technology. Nature 470(7333): 198-203.
Margulies M, Egholm M, Altman WE, Attiya S et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376-380.
McLysaght A, Hokamp K, Wolfe KH (2002). Extensive genomic duplication during early chordate evolution. Nat. Genet. 31(2): 200-204.
Nagle T, Berg C, Nassr R, Pang K (2003). The further evolution of biotech. Nat. Rev. Drug Discov. 2(1): 75-79.
Nelson DL, Cox MM (2008). Lehninger principles of biochemistry WH Freeman.
Peters HPF, van Raan AFJ (1994). A bibliometric profile of top-scientists: A case study in chemical engineering. Scientometrics 29(1): 115-136.
Roach JC, Boysen C, Wang K, Hood L (1995). Pairwise end sequencing: A unified approach to genomic mapping and sequencing. Genomics 26(2): 345-353.
Rodríguez K, Moreiro JA (1996). The growth and development of research in the field of ecology as measured by dissertation title analysis. Scientometrics 35(1): 59-70.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822): 928-933.
Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74(12): 5463-5467.
Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature 406(6799): 959-964.
Suzuki Y, Gojobori T (1999). A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16(10): 1315-1328.
Suzuki Y, Nei M (2001). Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 18(12): 2179-2185.
Tanaka H, Ho YS (2011). Global trends and performances of desalination research. Desalin. Water Treat. 25(1-3): 1-12.
Ugolini D, Cimmino MA, Casilli C, Mela GS (2001). How the european union writes about ophthalmology. Scientometrics 52(1): 45-58.
Ugolini D, Parodi S, Santi L (1997). Analysis of publication quality in a cancer research institute. Scientometrics 38(2): 265-274.
Ugolini D, Puntoni R, Perera FP, Schulte PA, Bonassi S (2007). A bibliometric analysis of scientific production in cancer molecular epidemiology. Carcinogenesis 28(8): 1774-1779.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. (2001). The sequence of the human genome. Science 291(5507) 1304-1351.
Via A, De Las Rivas J, Attwood TK, Landsman D, Brazas MD, Leunissen JAM, Tramontano A, Schneider MV (2011). Ten simple rules for developing a short bioinformatics training course. Plos Comput. Biol. 7(10): Article Number:e1002245.
Weinberg RA (2010). Point: Hypotheses first. Nature 464(7289): 678
Wolfe KH, Li WH (2003). Molecular evolution meets the genomics revolution. Nat. Genet. 33(S): 255-265.
Xie SD, Zhang J, Ho YS (2008). Assessment of world aerosol research trends by bibliometric analysis. Scientometrics 77(1): 113-130.
Yang ZH (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15(5): 568-573.
Zhang GF, Xie SD, Ho YS (2010). A bibliometric analysis of world volatile organic compounds research trends. Scientometrics 83(2): 477-492.
Zhou P, Leydesdorff L (2008). China ranks second in scientific publications since 2006. ISSI Newletter 13: 7-9.
Zhou P, Leydesdorff L (2009). Chemistry in China - a bibliometric view. Chim. Oggi-Chem. Today 27: 19-22.