How to Cite:
Umang, U., Bharti, P. K., & Husain, A. (2022). Bibliometric analysis of research aspects of simple sequence repeats in cultivars. *International Journal of Health Sciences*, *6*(S4), 7922–7933. https://doi.org/10.53730/ijhs.v6nS4.11756

Bibliometric analysis of research aspects of simple sequence repeats in cultivars

Umang
School of Computer Science, Shri Venkateshwara University, Gajraula, 244236, Uttar Pradesh, India
Corresponding author email: anilumang@yahoo.co.in

Dr. P K Bharti
School of Computer Science, Shri Venkateshwara University, Gajraula, 244236, Uttar Pradesh, India
Email: vc@svu.edu.in

Dr. Akhtar Husain
Department of Computer Science & IT, MJP Rohilkhand University, Bareilly, 243006, Uttar Pradesh, India
Email: akhtarhusain@mjpru.ac.in

***Abstract*---**Simple sequence repeats are ubiquitous patterns found in prokaryotes and eukaryote genome sequences. Identification and analysis of simple sequence repeat in chloroplast and mitochondrial genome sequences of several crop cultivars and environmentally significant plants have been performed to study genetic diversity, genetic variations, molecular breeding, gene discovery, disease identification, and hypervariability. Considering the versatile nature of microsatellites, a bibliometric analysis of the published global literature related to the study of microsatellites in cultivars from the web of science core collection database was analyzed by VOSviewer to gain insight into the frontier research topics, key researchers, institutes, current trends, technology, full citations and, lead publisher etc. A total of 1700 publications were retrieved, with an average article citation of 22.91 and an H index of 77. In addition, the frontier topics of research and technologies involved were explored from the abstract and title field. This study gives insight into the current and future scientific trends in the identification and analysis of microsatellites in cultivar's genome sequences. Also, it may help professionals and other stakeholders to fill the gap in microsatellites-related research in crops.

***Keywords*---**bibliometric studies, citation analysis, cultivars, simple sequence repeats, VOSviewer.

Introduction

The term Simple sequence repeats (SSRs) was coined and discovered by Litt and Luty in 1989 and applied to humans by Edwards et al. 1991 and to plants by Akkaya et al. 1992. These are tracts of repetitive DNA motifs from 1 to 6 nucleotides in length, such as TAATAATAATAA. Simple sequence repeats are also known as microsatellites. They are present abundantly at different locations in all eukaryotes and prokaryotes genomes with elevated levels of polymorphism compared to other molecular markers. Microsatellite repeat numbers can range from two (CA) 2 or three (GT) 3 to a few dozen (GCTT) U, whereas the minisatellites are composed of a few dozens to a few hundreds of repeated motifs. Significantly, Simple sequence repeats mutate at notably higher rates than non-repetitive sequences: 10−2 to 10−3 per locus, per gamete, per generation, leading to higher polymorphism. In addition, replication slippage, sister chromatid exchange, unequal crossing-over, and gene conversion may result in microsatellite diversity.

Simple sequence repeats (SSRs) are widely used tools in plant breeding & genetics research, and evolutionary studies, because of their high ability to show diversity among cultivars (Adato et al., 1995; Mhameed et al., 1996; Levi and Rowland, 1997). However, in vitro SSR identification, analysis and construction of a high-resolution linkage map for SSR markers in crops are expensive and time-consuming. Due to the easy availability of next-generation sequencing tools, simple sequence repeats analysis and identification in crop cultivars and other genome sequences has boosted research in plant breeding, gene resistance, genetics and diversity analysis in almost all crops, vegetables and fruits. The advantages of SSR include co-dominant inheritance, analytical simplicity and transferability (Weber, 1990; He et al., 2003). Simple sequence repeats are used to study genetic variations, population structure, paternity, phylogenetic studies, gene mapping, identifying genes or mutations, genetic linkage analysis, QTL and fingerprinting applications. These are potent tools for discriminating species and assigning them to geographically defined populations.

By analyzing the published research, there is much more potential to find new research frontiers in crop cultivars using microsatellites. Also, scholars use bibliometric analysis for various reasons, such as to uncover emerging trends in research, article and journal performance, collaboration patterns among authors and institutes and research constituents, and to survey the intellectual structure of a particular domain in the extant literature (N. Donthu, 2021). For example, a keyword and abstract analysis identifies the most popular topics covered by the bibliometric study and interdisciplinary articles with the highest impact. This method has a lot of potential for discovering up-and-coming fields. Many Software tools for conducting science mapping bibliometric analysis are available such as Bibexcel, Biblioshiny, CiteSpace, Content, VOSviewer etc. Many support Windows, Linux, or Mac platforms as a standalone tool. These tools extract data from many databases in various formats using visualization tools and clustering algorithms and can load and export information from many sources of scientific literature. For example, a bibliometric approach to analyze the articles published in the journals Buildings and Tulsi prajna was used to visualize the scope of research, article volumes over time, citations, most active contributors, countries,

primary language, and robust keywords (Xiao et al. 2022 & Tyagi, S and Bharadwaj, S N., 2021). Also, the COVID19-related published articles were analyzed (Yuetian et al. 2020) using VOSviewer to study the critical aspects of research.

Visualization provides information by representing voluminous data; visualization tools involve statistical analysis, graphics interaction etc. The primary purpose is to represent data appropriately to understand and collect the parameters in an understandable format; also, visualization helps improve human cognition and analytical reasoning. We can immediately grasp the minimum and maximum observations in any data, the clusters, the scatters etc.

Clustering algorithm relies on unsupervised machine learning; these are used to group similar examples and similar data; for example, we can apply it in market segmentation, Social Network Analysis, Medical imaging, Image segmentation etc. A similar clustering feature has been used in VOSviewer (Van Eck, N.J., Waltman, L, 2010) software to find out the related terms used in detailed research, or we can say what kind of research revolves around a specific field or where researchers are focusing upon, say, for example, if we are talking of simple sequence repeats then, Researchers are connecting simple sequence repeats to which other study fields. Clustering tools used in visualization help organize data into hierarchical and non-hierarchical clusters, density-based clusters etc.

Published literature from the web of science database was explored using the bibliometric quantitative analytical methods to explore the research aspects of simple sequence repeats in crop cultivars. Earlier, such a study was not performed to identify the current status and possibly new research frontiers till now.

Therefore, this study uses a bibliometric approach to analyze the following points.

- To analyze the metadata of all the papers related to simple sequence repeats identification research aspects in cultivars indexed in the Web of Science Core Collection since it is a multidisciplinary platform and allows accessing almost 1.9 billion cited references from over 171 million records.
- To find prominent research fields or interest in the Simple Sequence Repeats identification in cultivars.
- To identify the most prolific authors, publishers and funding institutions.
- To identify significant contributors and the major language of published literature.
- To identify the most cited research article.
- To find out the year in which maximum literature was published.
- To Analyze co-authorship with the unit of analysis as authors.
- Co-occurrence and keywords analysis.
- To identify and visualize the keywords of interest where authors are working.
- To visualize and analyze title and abstract fields.

Materials and Methods

i. For Bibliometric analysis, the Web of Science (WOS) Core Collection database has been selected since it is the most often used database in management and Organizations. The web of science core collection was searched for the query-
   - Results for "simple sequence repeat identification in cultivars (All Fields)".
   - Period 2002 to 2022.

ii. Preferred reporting items for Systematic Reviews and Meta-Analysis were performed for valid data extraction.

iii. All valid results were exported in RIS format for analysis by VOSviewer version 1.6.18 for visualizing scientific landscapes and exploring network data related to Bibliometric analysis of simple sequence repeats research aspects.

iv. Construction and visualization of data related to Web of Science research categories, publications and citations over time, TreeMap Chart view of research areas was analyzed using the web of science inbuilt features.

v. Authors-co-authorship from various institutes working in specialized fields, co-occurrence- co-citation analysis, and analysis of the keywords from the title and abstract field was performed using VOSviewer to obtain clustering information, links, and link strength.

Results and Discussion

The web of science core collection database was searched for the query "Simple sequence repeats identification in cultivars (all fields)". A total of 1728 articles were retrieved after using the PRISMA method (figure1); 1700 articles were found suitable for the analysis using the web of science core collection inbuilt tools. Of these, 1680 were research articles followed by 18 review articles and only a few editorial and book chapters.

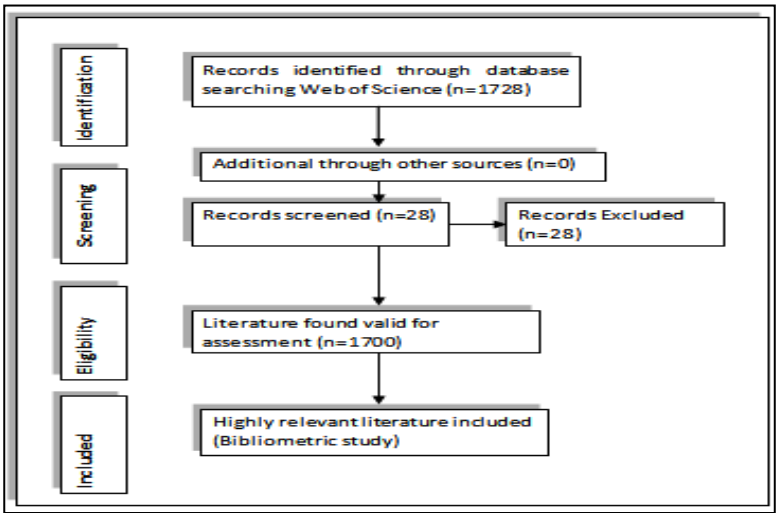

Figure1. PRISMA flow diagram for data extraction and refined process of the query

Analysis of Web of Science categories and Research areas

Among the web of science categories, 731 articles were published in Plant Sciences, 585 in agronomy, 574 in horticulture, 465 in Genetics heredity and 189 in biotechnology and applied microbiology. The TreeMap chart visualization is shown in figure 2. A maximum of 56.58% of research was carried out in agriculture, 43% in Plant Sciences and 27.35% in genetics and heredity and the least in multidisciplinary sciences. By visualizing this chart, it has been observed that maximum research on genetic studies on plants and horticulture-related to agriculture. A similar analysis was performed in mining literature from the Web of Science category related to applying big data tools in poll campaigns (Umang and Pandey D.K., 2022).

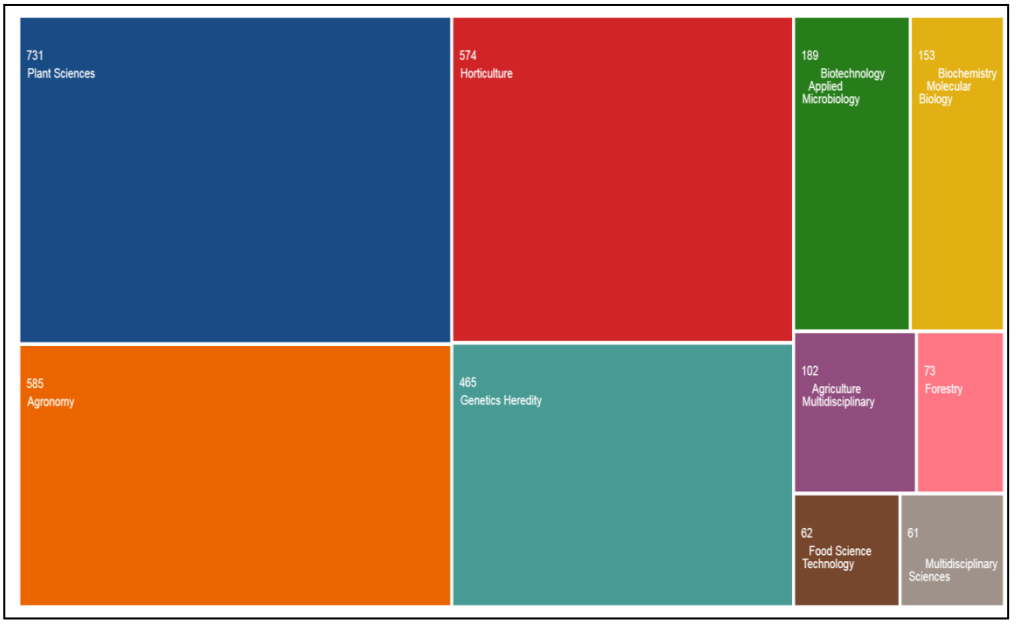

Figure 2 TreeMap chart visualization showing the web of science categories

Figure 3 reflects that the maximum literature about research on simple sequence repeat identification in cultivars was published by China, followed by USA and India. Topmost populated, developed, and developing countries contribute to studying diversity by simple sequence repeat identification research in crop cultivars.

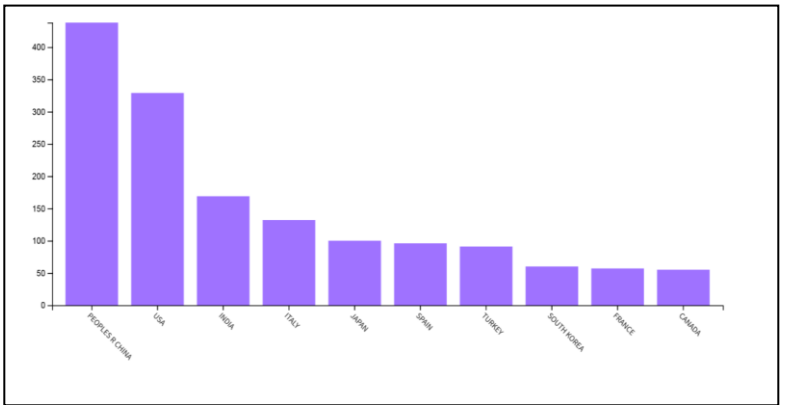

Figure 3 Bar chart visualization of countries publishing utmost research articles

A similar approach of bibliometric analysis in hypertension-related articles was performed from articles listed in PubMed and the Web of Science database (Devos, P and Menard, J in 2019), where article volumes and citations were observed.

Annual publishing trend and citations

In publications analysis from 2002 to 2022, it has been observed that 22972 articles were without self-citations reflecting the originality of the research. Overall the articles were cited 38924 times with an average of 22.91 per item with an H index of 77. This H index reflects the productivity and citations of the research articles are genuinely exceptional. Utmost research articles were published in 2012 & 2014, followed by 2015. This concluded that the research on simple sequence repeats in crop cultivars was at an all-time high from 2012 to 2015. The highest citation of the articles was in the year 2021, showing the maximum impact of the research work (figure 4). The article “Exploiting EST databases for the development and characterization of gene-derived SSR- markers in Barley (Hordeum vulgare L”, authored by Theil, T et al., had maximum citations, followed by “Development and characterization of 140 new Microsatellites in Apple (Malus x domestica Borkh)” written by Liebhard, R et al.; it was cited 433 times reflecting the research interest in simple sequence repeats and crop cultivars.

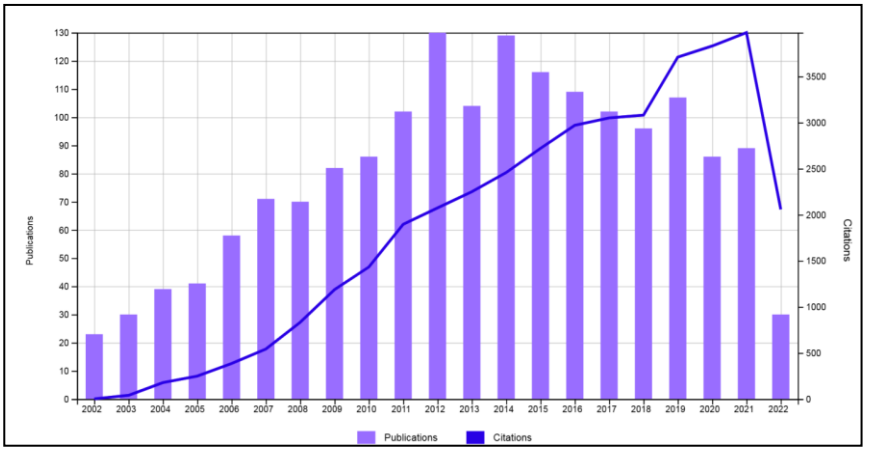

Figure 4 The Bar and Line chart visualization of the article's volume and citations over time

The citations of the articles ranged from 0 to 1663 and increased over the years but utmost after 2018. Published volume of literature per year, growth, and research themes were explored using the Scopus database on climate change and human health, emphasizing infectious diseases using VOsviewer (WM Sweileh, 2020).

Analysis of most prolific authors, publishers and funding institutions

The researchers who contributed most were Yamamoto T from National Agriculture & Food Research Organization - Japan, Ercisli S from Turkey, Chen XM from USA, and Wang L. This show the active authors belong to agricultural institutes in Japan and USA. Actively engaged researchers were affiliated with the United States Department of Agriculture, followed by the Chinese Academy of Agriculture Sciences and then the Indian Council of Agriculture Research.The prominent journals were Theoretical and Applied Genetics, Molecular Breeding, Euphytica, Scientia Horticulturae and Crop Science. These renowned journals mainly publish research on genetics breeding and biotechnology in crops and have impact factors from 2 to 5.

Major publishers are Springer Nature, Elsevier, and Wiley, followed by MDPI. These are lead research publishers in their respective fields. The funding agency which contributed maximum funds to research projects was the National Natural Science Foundation of China, followed by the United States Department of Agriculture, then USDA Agricultural Research Service. People's Republic China, USA, India, Italy followed by Japan are countries actively engaged in microsatellites studies in crop cultivars. The language in which maximum literature was published was English, and the Web of Science index was Science Citation Index Expanded. From the above analysis, the further research can be determined for fetching grants, collaboration and identifying prominent publishers.

Analysis of co-authorship with the unit of analysis as authors

On co-authorship and author's research, the maximum number of authors was set to 6 per document and minimum to 01; then, we retrieved 1000 authors, 1785 links and total link strength of 2286. This shows the prominent presence of single authors strongly connected to a similar type of research in 203 clusters, as shown in figure 5.

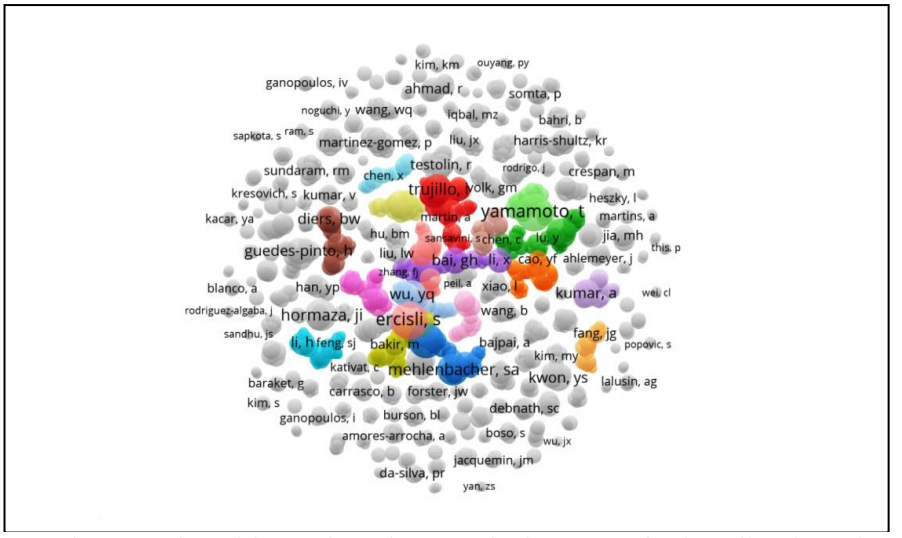

Figure 5 Co-authorship and author analysis network visualization chart.

Co-occurrence and keywords analysis

Nine clusters were obtained upon co-occurrence and keywords analysis, as shown in figure 6. The most significant cluster had 189 terms; it was related to identifying simple sequence repeats for candidate genes in Brassica species for studying association analysis, mapping, diseases like bacterial wilt, blast resistance, DNA markers, gene mapping, and genetic linkage, map, marker-assisted selection etc. Cluster 2 had 86 items related to DNA extraction, diversity, cultivar identification, biodiversity, identifying issrs, microsatellite, microsatellites DNA, varieties, variety tag etc. Cluster 3 had 84 terms related to database, DNA fingerprinting, ests SSR, and cross-species amplification. Cluster 4 had 71 items related to cluster analysis, cultivar diversity analysis, DNA polymorphism, Genetic diversity, resistance trait and several diseases. Cluster 5 had 56 things that reflected study related to Barley, wheat bread, common wheat, and disease resistance. Cluster 6 had 43 items in which chloroplast DNA, fingerprinting, and microsatellite markers study was conducted in apple cultivars, peach, apricot, cherry, almond, sweet cherry etc. Cluster 7 had 33 items identifying fingerprinting rapid analysis and sequence repeats. Cluster 8 had 22 things; majorly, the research was performed in Brassica, cabbage for genetic resistance, inter-specific hybrids etc. Finally, cluster 9 had one item fragment length polymorphism.

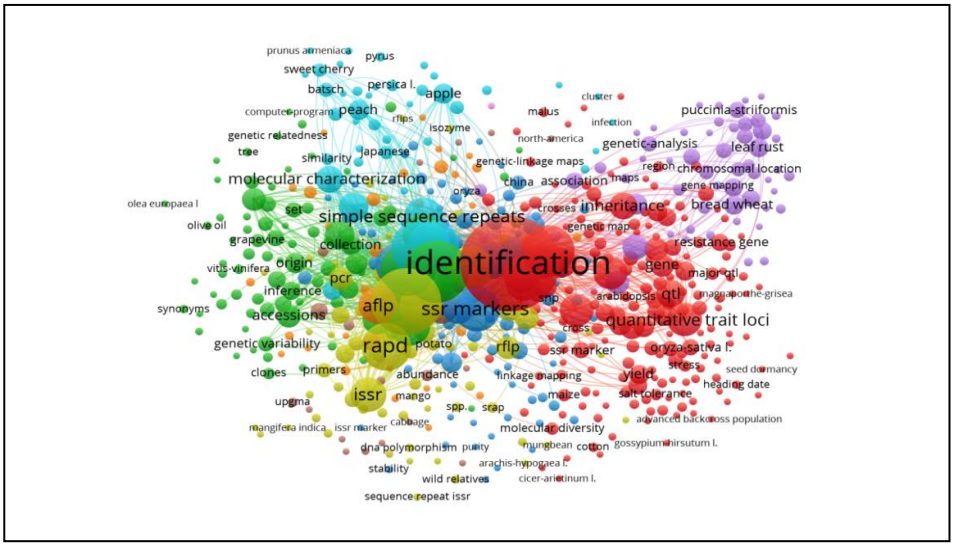

Figure 6 Co-Occurrence and Keyword network visualization chart showing various clusters with minimum terms set to 05

Identification of the key terms of interest where authors were working

The key terms from the abstract field retrieved revealed that the researchers were working on the related fields like they were using simple sequence repeats to study resistance, reaction, mapping, linkage group, issr, rapd, relatedness for the abundant type, amplified region, single dominant gene, synonym, QTL, migration, amplified band etc. Therefore, these terms were grouped into 4 clusters, and clusters are none overlap. Therefore, future research aspects can include words from non-overlapping clusters, thus filling the gap.

To visualize and analyze title and abstracts fields

When a map based on text data was created using terms from the title and abstract field analysis, five significant clusters were obtained with 1166 terms and total link strength of 150779. These were a genetic relationship, gene, resistance gene, and chloroplast genome. The keywords reflected vital research on rice, tritici, apricot, powdery mildew, and resistance. In addition, simple sequence repeats research related to water, salinity, fibre quality, weak growth, greenhouse condition, and adult plant stage was observed. These keywords show that the articles were mainly related to investigating cultivar's variation, genes, population structure, and environmental tolerance in various crops using microsatellites. A bibliometric approach utilizing a web of science database and VOSviewer program was applied to explore research investigating acupuncture for pain treatment and the current status of acupuncture analgesia (Lee I et al. 2020).

In the present analysis, it has been observed that simple sequence repeats played an essential role in studying genetic diversity, species variations, genetic linkage, mapping and mutations etc., in various grains, vegetables and fruit varieties distributed geographically. The title and keywords analysis retrieved the utmost research categories from the literature. The web of science analysis retrieved

names of active authors, institutions, countries, publishers, journals, and papers with minimum to total citations. The number of single and multiple authors working was observed. The crops in which the SSRs were used for studying various parameters were marked as essential words from the title and abstract fields of the papers.

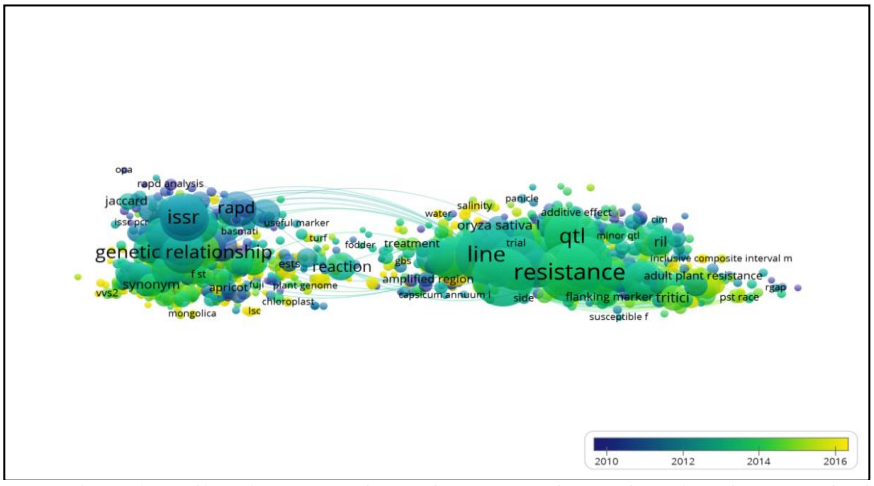

Figure 7 Overlay visualization map based on text data showing keywords from the title and abstract fields

Roadmap for further research in crop cultivars using simple sequence repeats

The next-generation Sequencing tools can be applied for research on crop cultivars for species identification, improving crop yields, and finding diseases resistant and climatic stress tolerant genes. Such analysis has been used on banana, citrus fruits, peach, wheat, rice, maize, Barley, carrot, spinach, coconut, Apple, Sweet Cherry and many other cultivars. The researchers can work on association, inheritance and genetic relatedness in other pulses and vegetable crops where less research was observed. Due to the changing climate pattern, the crop sowing patterns are also changing, so there is a need to study climatic stress resistance genes and mutations. New computing tools and technologies can be applied to exploit voluminous genomic data from NCBI. The researchers can use them to develop databases related to the analysis and identification of simple sequence repeats coding, non-coding regions, primer pairs, t-RNA, and rRNA locations. These can help naive researchers to use non-overlapping terms from various clusters to design new research. They can find collaborators, funding agencies and experienced researchers to proceed. Knowledge of the available tools and techniques can help to work out gaps.

Conclusion

The present study showed significant research categories: plant science, agriculture, genetics and breeding. Furthermore, diversity analysis, gene resistance, population structure, mapping, QTL, gene resistance, environmental tolerance, etc., were performed using simple sequence repeat identification in crop

cultivars. The leading journal was found to be Theoretical and Applied Genetics, with publishers such as Springer Nature, the utmost contributing author from Japan, and only contributing countries such as the USA, China, Japan and India. The language used mainly was English. However, the focused keywords and terms can identify gaps using non-overlapping clusters and correlating terms in future research.

Acknowledgments

Authors are grateful to two anonymous reviewers for their valuable comments on the earlier version of this paper.

Financial support
Nil

Conflict of Interest
The author declares no conflict of interest.

References

Adato, A., Sharon, D., Lavi, U., Hillel, J., & Gazit, S. (1995). Application of DNA fingerprints for identification and genetic analyses of mango (Mangifera indica) genotypes. *Journal of the American Society for Horticultural Science*, *120*(2), 259-264.

Akkaya, M. S., Bhagwat, A. A., & Cregan, P. B. (1992). Length polymorphisms of simple sequence repeat DNA in soybean. *Genetics*, *132*(4), 1131–1139. https://doi.org/10.1093/genetics/132.4.1131

Devos, Patrick & Menard, Joël. (2019). Bibliometric analysis of research relating to hypertension reported over the period 1997–2016. *Journal of Hypertension*: Volume 37 - Issue 11 - p 2116-2122. DOI: 10.1097/HJH.0000000000002143.

Edwards Al, et al. (1990). Automated DNA sequencing of the human HPRT locus, *Genomics*, Volume 6, Issue 4, Pages 593-608,https://doi.org/10.1016/0888-7543(90)90493

He, C., Poysa, V., & Yu, K. (2003). Development and characterization of simple sequence repeat (SSR) markers and their use in determining relationships among Lycopersicon esculentum cultivars. *Theoretical and Applied Genetics*, *106*(2), 363-373.

Lee, I. S., Lee, H., Chen, Y. H., & Chae, Y. (2020). Bibliometric Analysis of Research Assessing the Use of Acupuncture for Pain Treatment Over the Past 20 Years. *Journal of pain research*, *13*, 367–376. https://doi.org/10.2147/JPR.S235047

Levi, A., & Rowland, L. J. (1997). Identifying blueberry cultivars and evaluating their genetic relationships using randomly amplified polymorphic DNA (RAPD) and simple sequence repeat-(SSR-) anchored primers. *Journal of the American Society for Horticultural Science*, *122*(1), 74-78.

Litt, M. & Luty, J.A. (1989). A hypervariable microsatellite revealed by in-vitro amplification of dinucleotide repeat within the cardiac muscle actin gene. American of *Journal of Human Genetics*, 44(3), 397-401.

Mhameed, S., Sharon, D., Hillel, J., Lahav, E., Kaufman, D., & Lavi, U. (1996). Level of heterozygosity and mode of inheritance of variable number of tandem

repeat loci in avocado. *Journal of the American Society for Horticultural Science*, *121*(5), 768-772.

N. Donthu, S. Kumar, N. Pandey, & W.M. Lim. (2021). Research constituents, intellectual structure, and collaboration patterns in Journal of International Marketing: An analytical retrospective. *Journal of International Marketing*. Available at doi:10.1177/1069031X211004234 (in press).

Sweileh, W.M. (2020). Bibliometric analysis of peer-reviewed literature on climate change and human health with an emphasis on infectious diseases. *Global Health* 16, 44. https://doi.org/10.1186/s12992-020-00576-1

Tyagi, S and Bharadwaj, S N., (2021). Bibliometric Analysis of Papers Published During 2016-2020 in 'Tulsi Prajna' Research Journal". *Library Philosophy and Practice (e-journal).* 5165. https://digitalcommons.unl.edu/libphilprac/5165

Umang & Pandey, D. K. (2022). A bibliometric analysis of application of big data tools in electoral campaigns by using VOSviewer. *Third Concept*, Vol. 35(420), 46-49.

Van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84, 523–538.

Weber, J. L. (1990). Informativeness of human (dC-dA) n·(dG-dT) n polymorphisms. *Genomics*, *7*(4), 524-530.

Xiao, Z., Qin, Y., Xu, Z., Antucheviciene, J., & Zavadskas, E. K. (2022). The Journal Buildings: A Bibliometric Analysis (2011–2021). *Buildings*, 12(1), 37. https://doi.org/10.3390/buildings12010037

Yu, Yuetian, Li, Yujie, Zhang, Zhongheng Gu, Zhichun, Zhong, Han Zha, et al. (2020). A bibliometric analysis using VOSviewer of publications on COVID-19. *Annals of Translational Medicine*; Vol 8, No 13 July 2020: Annals of Translational Medicine 2020. https://atm.amegroups.com/article/view/46197