Wikipedia as a tool for contemporary history of science: A case study on CRISPR

Omer Benjakob[#,1,2], Olha Guley[1,2], Jean-Marc Sevin[1,2], Leo Blondel[1,2], Ariane Augustoni[1,2], Matthieu Collet[1,2], Louise Jouveshomme[1,2], Roy Amit[3], Ariel Linder[1,2], Rona Aviram[#,1,2]

(1) Université Paris Cité, Inserm U1284, System Engineering and Evolution Dynamics, F-75004 Paris, France
(2) Learning Planet Institute, F-75004 Paris, France
(3) Bezalel academy of arts and design, Israel

[#] Corresponding Authors

Abstract

Rapid developments and methodological divides hinder the study of how scientific knowledge accumulates, consolidates and transfers to the public sphere. Our work proposes using Wikipedia, the online encyclopedia, as a historiographical source regarding contemporary science. We chose the high-profile field of gene editing as our test case, performing a historical analysis of the English-language Wikipedia articles on CRISPR. Using a mixed method approach, we qualitatively and quantitatively analyzed its text, sections and references, alongside 50 affiliated articles. These, we found, documented CRISPR's maturation from a fundamental scientific discovery to a biotechnological revolution with vast social and cultural implications. We developed automated tools to support such research generically and demonstrated its applicability on two other scientific fields we have previously studied - COVID-19 and Circadian clocks. This method makes use of Wikipedia as a digital and free archive, documenting the incremental growth of knowledge and the manner scientific research accumulates and translates into public discourse. Using Wikipedia in this manner compliments and overcomes some issues with contemporary histories and can also augment existing bibliometric research.

Keywords

Wikipedia, CRISPR, History of Science, Scientometrics, Digital Humanities, Science of Science.

Introduction

In recent years, the historically qualitative field of history of science has undergone a data revolution[1], with research increasingly making more use of big data and computational techniques for historical ends[2]. Despite the rise of digital humanities, a divide has persisted between quantitative historical research and textually rich qualitative work, resulting in a historiographic lacuna[3]. Meanwhile, a small but growing body of research based on Wikipedia has emerged at the intersection of bibliometrics[4], history[5], health[6], medical[7] and science[8]. We suggest the aforementioned lacuna can be partially addressed in the context of the history of contemporary science by systematizing research methods on an unlikely arena that is rich in both bibliometric data and historical text: Wikipedia.

Now over 20 years old, Wikipedia in English is, per its own definition, the largest and most popular reference work used by the general public[9]. Wikipedia's science articles top search engine results, making the open encyclopedia a key node in the transference of academic knowledge to the public sphere. Once ridiculed for being inherently unreliable, both academic research and the media have in recent years praised its coverage as being in lock step with science[10], especially in light of the COVID-19 pandemic[11].

Wikipedia requires “verifiable” sources to back all factual claims[9], and research has found that on medical, health[6] and science[8] topics it has an explicit bias towards academic sources. Wikipedia facilitates access to knowledge usually kept behind academic paywalls and jargon[12]. Unlike academic publications focused on the state-of-art of the field or review papers coverage of the aforementioned, Wikipedia does not aim to publish original research - it only reflects the scientific consensus based on already published sources. Here, we suggest Wikipedia can also play a bigger role, serving as a source of knowledge in its own right, regarding the history of contemporary science, which we demonstrate through a case study on the CRISPR field.

CRISPR-based gene-editing tools have been labeled the scientific “breakthrough” of the 21st century[13]. While CRISPRs were identified in the 1980's, and received their name in 2002[14], their function remained unclear for many years. In 2005, different labs deduced from *in silico* studies that CRISPR sequences were part of a bacterial adaptive immune system[15,16,17].

The academic studies that first performed CRISPR-based directed gene editing *in vitro* were famously published in 2012: First from the labs of Jennifer Doudna and Emmanuelle Charpentier[18] and shortly after in a paper of the Virginijus Šikšnys group[19]. These were rapidly followed by publications in February 2013 that performed genetic engineering in vivo in mammals, led by scientists Fang Zhang[20] and George Church[21]. Thus, the field matured from a basic science discovery into the ability to utilize CRISPR-associated proteins like Cas9 for genetic engineering, currently used by countless labs around the globe[22]. Doudna and Charpentier were awarded the 2020 Nobel Prize for Chemistry for their scientific contribution to genetic editing technologies, showcasing how the so-called CRISPR revolution has played out over the past 20 years.

In contrast to many other groundbreaking scientific discoveries which remain known only within scientific circles, human gene editing has also been in the spotlight of much public debate. For example, many news outlets have dedicated reports to developments in the field and debated the ethical implications of so-called designer babies[23]. Netflix has even broadcasted a documentary film dedicated to CRISPR (Human Nature, 2019), underscoring its iconic status in popular culture.

The CRISPR field's brief history has been riddled with controversies, and legal battles over credit and CRISPR patents were all covered extensively in the media[24]. Most famously, Eric Lander's perspective in Cell, the "Heroes of CRISPR"[25], was met with fierce criticism[26]. Critics claimed that the text offered a biased version of the field's history that minimized the roles of some scientists as part of the patent war raging between academic institutions[27] - going as far as to label Lander the "villain" of CRISPR[28]. This controversy underscores how scientific outlets, even those famous for publishing novel scientific research, may not necessarily serve as reliable historical sources on contemporary science itself.

CRISPR is a prime example of a scientific field that has undergone massive growth during Wikipedia’s lifespan. It is an ideal case study as its short history is multi-faceted: a highly scientific topic with wide-ranging technological and social ramifications. All of these, we found, were documented on Wikipedia and its different articles, supported by scientific, public and popular sources alike. Together, our findings - based on an analysis of the CRISPR article and 50 others with related content - suggest that Wikipedia can indeed serve as a tool in the history of contemporary science. To that end, we put forward a method for using Wikipedia, its articles, their edit histories and their references: we outline a methodology and provide some automated tools utilizing Wikipedia’s data. Our method relies on both quantitative and qualitative analyses that may help consolidate the aforementioned conflict between data and content dependent historical research.

Results

1. Delineating the research scope

The manner in which a scientific field is represented on Wikipedia requires clear delineation of scope and span - i.e., the articles that touch on it and the time frame being examined. While a single article can provide a rich source of textual and historical data, related articles may represent more nuanced facets of a field - like scientists' biographies or related events and technologies. Identifying these requires sieving through Wikipedia's massive body of articles - currently numbering well above 6 million in English alone.

For this aim, we propose a stepwise strategy for defining a research corpus about a topic. The first step utilizes Wikipedia's free-text search function to find all articles that contain the topic researched (Fig. 1A). In the present study, searching for "CRISPR" yielded 720 Wikipedia articles containing that term, as of June 2022 (Fig. 1B). Based on subjective reading of these articles, we found that many made only minor or incidental use of CRISPR. Thus, to permit qualitative analyses on a more focused pool, we designed the second stage of the research funnel, which calls for retaining only those articles with the term in either their title or one of their sections. With respect to CRISPR, this filtering yielded 51 articles (Table S1). Out of these, 10 had CRISPR in their title - and thus focused on it directly - and another 41 that only had it in the title of one of their sections, and thus touched on it indirectly through an intersection with another body of knowledge.

The main article/s, which we term the "anchor article/s", are those which in subject, text and focus are fully aligned with the topic being researched; while "auxiliary" articles, that make up the majority of the corpus, are those that represent secondary aspects of the topic or instances in which it is embedded within other fields. For this study, the anchor article was ”CRISPR", which was selected semantically based on its title and content. It ranked amongst the top 5 articles in terms of size, number of references, and number of edits (Fig. S1), while the other 50 served as auxiliary articles.

Within this CRISPR corpus, several auxiliary articles focused on scientific topics, for example the article for “CRISPR Activation”, “Cas9”, or “CRISPR gene editing”, while others had wider scientific topics, such as “Antibiotic”, “Gene knockdown”, and “Genome editing”. Also included were articles with broad topics, for example “Wheat” which had a section on CRISPR-edited strains of grain. Another group of articles were those dedicated for scientists, like the 2020 Nobel laureates Doudna and Charpentier, awarded the prize for their groundbreaking work in the field; or Šikšnys, who also played a pivotal role in CRISPR’s history. Other science-adjacent articles touched on CRISPR's social aspects e.g., “The CRISPR Journal” and “Designer baby”, showing how cultural aspects are also captured by this method.

We therefore concluded that these articles provide a good sample of CRISPR related knowledge.

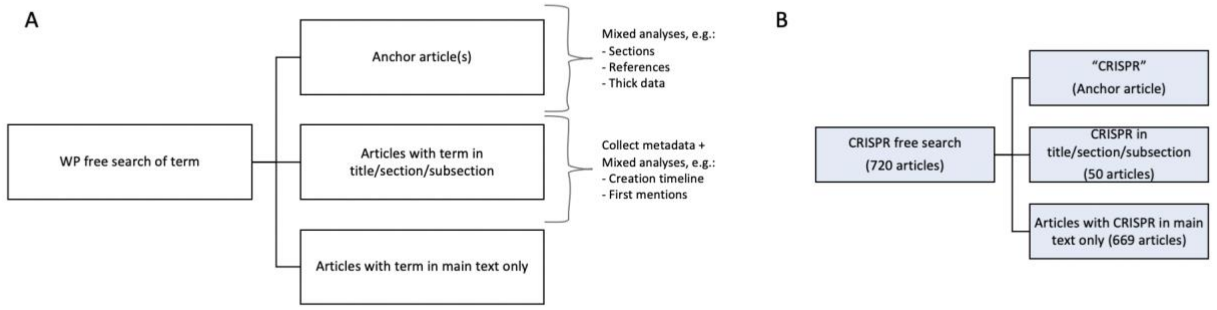

A
WP free search of term
Anchor article(s)
Articles with term in title/section/subsection
Articles with term in main text only
Mixed analyses, e.g.:
- Sections
- References
- Thick data
Collect metadata +
Mixed analyses, e.g.:
- Creation timeline
- First mentions
B
CRISPR free search (720 articles)
“CRISPR” (Anchor article)
CRISPR in title/section/subsection (50 articles)
Articles with CRISPR in main text only (669 articles)

Fig.1

Figure 1. Workflow for using Wikipedia to research the history of a specific field. A) Scheme of general flow. A free search of Wikipedia's English-language articles is conducted to identify all the relevant articles; these are then filtered to include only those that have the term in either their title or the title of a section. Next, different analyses can be performed on the anchor article and the corpus. B) Breakdown of flow scheme in the CRISPR case study, as of June 2022.

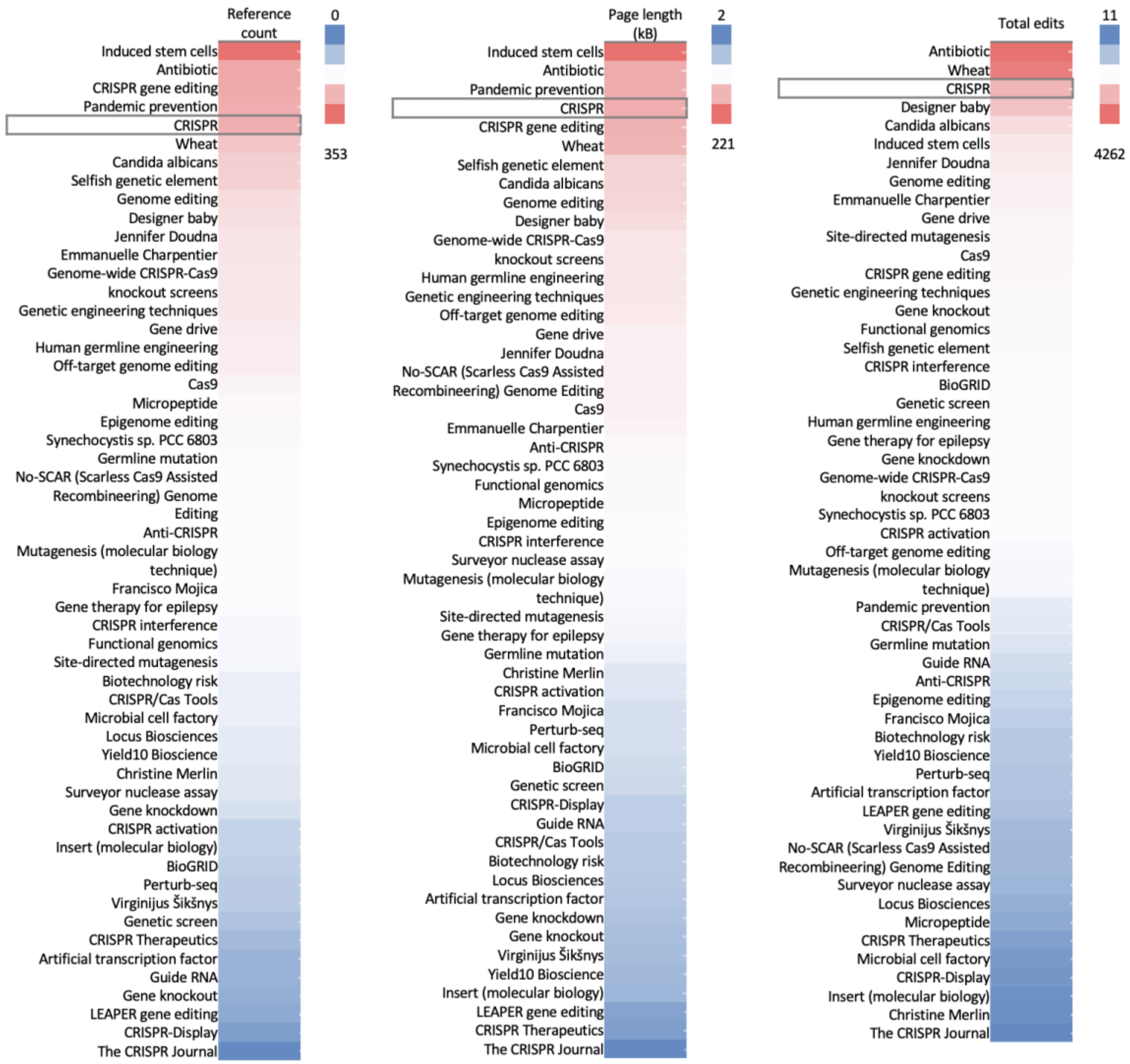


Fig. S1

Figure S1. The CRISPR corpus in numbers. The articles included in the corpus, sorted by number of references, size in kilobytes (kB) and number of edits. "CRISPR", highlighted, was among the top 5 articles of each category.

2. Mixed method analyses for understanding historical growth of knowledge

After having established our research scope, we first performed a comparative reading of the anchor article’s past versions, using annual intervals to sample textual and structural changes - at time narrowing the time frame to provide a more detailed account of the article’s historical textual growth. Thick description is a common methodology in the history and sociology of science. It is used for providing context and an interpretive framework for research based on multiple historical sources and diverse types of data. We suggest that unraveling scientific history through Wikipedia can be achieved by examining and then describing in rich detail the work of Wikipedia’s editors, the references they cited as well as the text these references supported. Here, this takes the form of reviewing the edit history and references of the CRISPR anchor article and understanding its interplay with auxiliary articles.

To augment the detailed thick description of the changes the article underwent throughout its development we used several mixed-method analyses. Mixed-methods research[29] combines quantitative and qualitative analyses and served as the basis for this research, with the data from Wikipedia and its subsequent analyses leading to textually rich examples interpreted to provide historical insight. This can be termed Wikipedia-focused “thick big data”[30], as opposed to content-agnostic big data approaches. This approach can be used both at the corpus level and that of specific anchor articles, and together provide a coherent system for researching other topics.

The article for CRISPR was created in June 2005, as what is termed a “stub” on Wikipedia - a short entry that calls for further elaboration (Fig 2A). This first version included but a single paragraph elucidating the CRISPR acronym and describing the genetic locus. At the time, there was no mention of its relation to bacterial immunity or gene editing, two points which would be integral to the field and as a result the article’s lead text in future versions (Fig. 2B).

We conducted the initial analysis on the CRISPR article's architecture, i.e., its table of contents, and mapped the shifts it underwent since the article's launch (2005). This "table of contents" or "section" analysis is a mixed-method: Quantitatively, we measured the overall number of sections and subsections (Fig. 3A); qualitatively, we reviewed their titles and documented the changes they underwent to provide insight into the content of the article, with the section titles serving as a proxy for new units of CRISPR-related knowledge (Table S2).

In addition, we examined the growth of the CRISPR corpus, by laying out the articles based on their Date Of Birth (DOB), (Fig. 3B). Opening new articles on Wikipedia requires the topic at hand to have a certain level of "notability"[31]. Here too, we combined a quantitative evaluation of the number of articles being created with a content-dependent reading of their titles. Finally, a side-by-side view of these two adds another layer of information, interpreted to provide a narrative to contextualize the findings, as described below.

Qualitative reading of the section titles showed that the structural changes were directly linked to shifts in the article's content, pertaining to either the accumulation of new knowledge or the restructuring of the growing field's representation on Wikipedia. For example, the first sections added in 2010 were "CRISPR Mechanism", "CRISPR Spacer and Repeats," "CAS Genes" and the reference section (Table S2). These sections pertain to CRISPR's genetic makeup, and can be collectively referred to as the basic science behind CRISPR.

In 2011, after a few months after a "Discovery of CRISPR" section was added to the article, a section termed "Evolutionary significance and possible applications" was created. For the next three years it included three proposed applications:

- *"Artificial immunization against phage by introduction of engineered CRISPR loci in industrially important bacteria, including those used in food production and large-scale fermentations.*
- *Knockdown of endogenous genes by transformation with a plasmid which contains a CRISPR area with a spacer, which inhibits a target gene.*
- *Discrimination of different bacterial strains by comparison of CRISPR spacer sequences (spoligotyping)"*

However, these would change in the following year. In a subsequent substantial edit to the article, in April 2013, a user called *Genomeengineering* made what would be their sole contribution to Wikipedia: Adding the 2012 paper by Doudna and Charpentier, and the two 2013 publications by Zhang and Church. They also amended the list of possible applications so it now included “genome engineering at cellular or organismic level by reprogramming of a CRISPR-Cas system to achieve RNA-guided genome engineering”. In November of that year the section’s title changed from “Possible applications” to “Applications”.

Alongside this section’s growth, which also saw the birth of the “further reading” section, and a section dedicated to “external links” was expanded, providing access to new utilities developed for CRISPR researchers. For example, a link to a “comprehensive software” for CRISPR guideRNA design was added as well as a link to a tool “for finding CRISPR targets.”

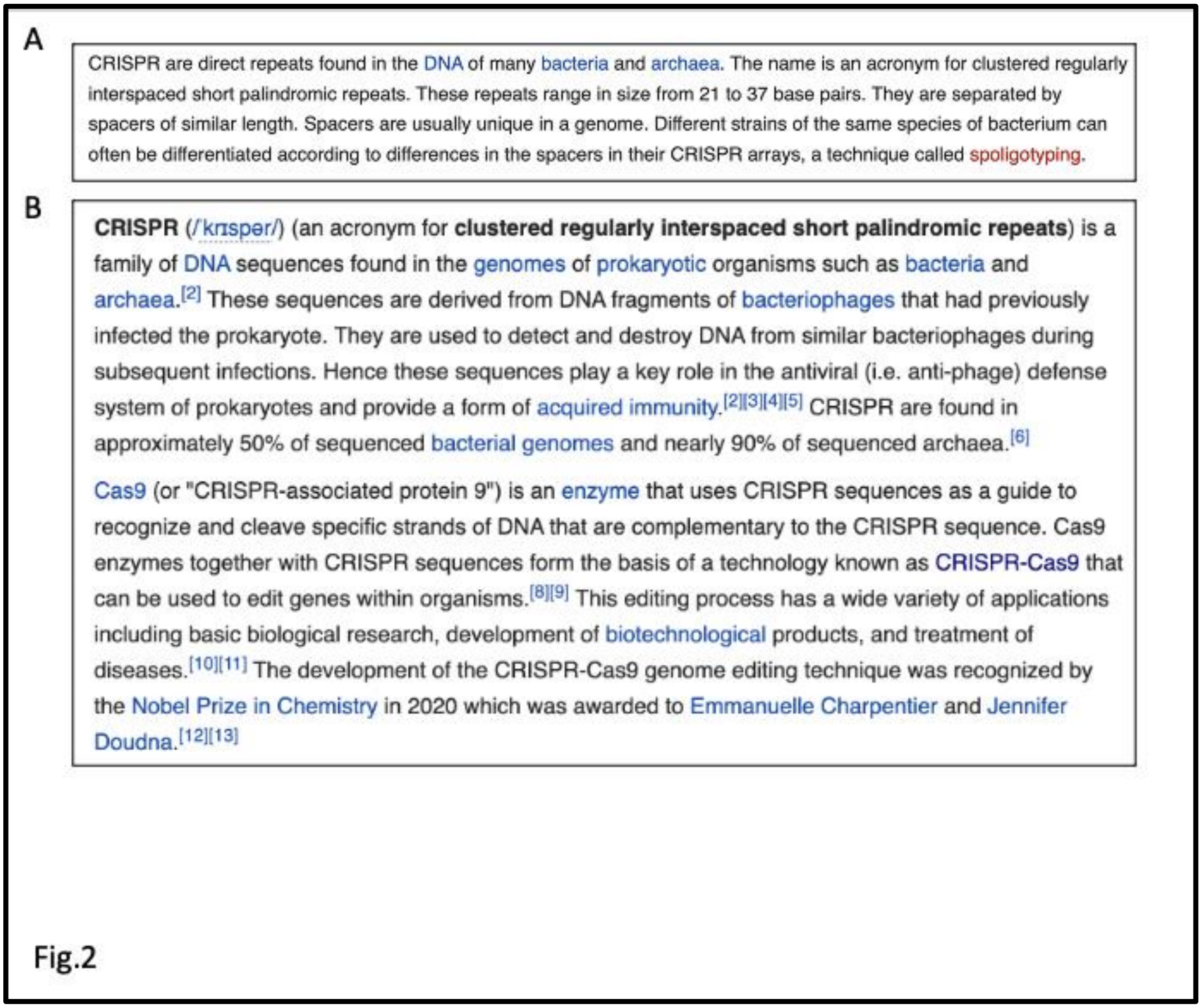
At the corpus level, this period also saw a spurt in article creation, with a number of CRISPR-related articles being created, like “CRISPR interference”. At this time, more articles directly based on or linked to CRISPR science and its applications were also created. For example, articles like “Genome editing” (2012) and “Cas9” (2013). It is also during this phase that the articles for scientists linked to its discovery were opened: an article about Doudna was created in 2012, coinciding with the publication of her landmark *Science* paper[18]. Soon thereafter, articles were created for “Epigenome editing” (2014) and “CRISPR/Cas tools” (2015). Thus, qualitatively, this period can be seen as covering the emergence and establishment of the applicative side of CRISPR.

On March 31, 2014, a few weeks after Doudna and Charpentier applied for a patent for their work, a “Patents” section was opened. In 2016, the section dealing with patents was expanded to include a “Patent and commercialization” subsection that included a detailed list of patent holders that at the time were fighting in the courts over legal ownership and in academic media over credit (Table S3). At the corpus level, we observed the creation of articles for Charpentier

(2015) and Šikšnys (in 2016), in tandem to the credit and patent wars raging over their respective discoveries.

In February 2019, with the patent wars reaching their resolution, the section (then four paragraphs long) was completely removed from the article. However, it was not deleted, but rather migrated to a new article called "CRISPR gene editing," opened that month in a big text-migration out of the anchor article. Also migrated was the section "Society and culture", which described the ability to conduct human gene editing in terms of the wider social debate about it and the policy changes it sparked.

Other migrations were seen throughout the article's history, also evident at the corpus level: In 2017, the "Knockdown/activation" subsection forked and expanded to an article of its own ("CRISPR interference"). A subsection about "Recognition" that attempted to attribute the CRISPR discovery to specific persons also moved to the new "CRISPR gene editing" article. The migration of key sections into "CRISPR gene editing" is evident in the drop in the number of sections in 2019 and is reflected in the uptick in the growth of the number of articles in the corpus, when, alongside the new fork article, "genome-wide CRISPR-cas9 knockout screens", "the CRISPR Journal" and "LEAPER gene editing" all got new articles that year or in 2020. This later phase also continued to document the growth of the biotech industry based on CRISPR, for example CRISPR Therapeutics, a company co-founded by Charpentier, received an article in 2021, further highlighting the field's maturation and growth in technology. Tellingly, 2020 also saw the creation of a "Pandemic prevention" article, which, in tandem with the COVID-19 pandemic, detailed all the medical and scientific attempts to preempt viral outbreaks - including those that could potentially make use of CRISPR. Articles like these raise an interesting question regarding the role of CRISPR in other bodies of knowledge and warrant an examination of the wider corpus.



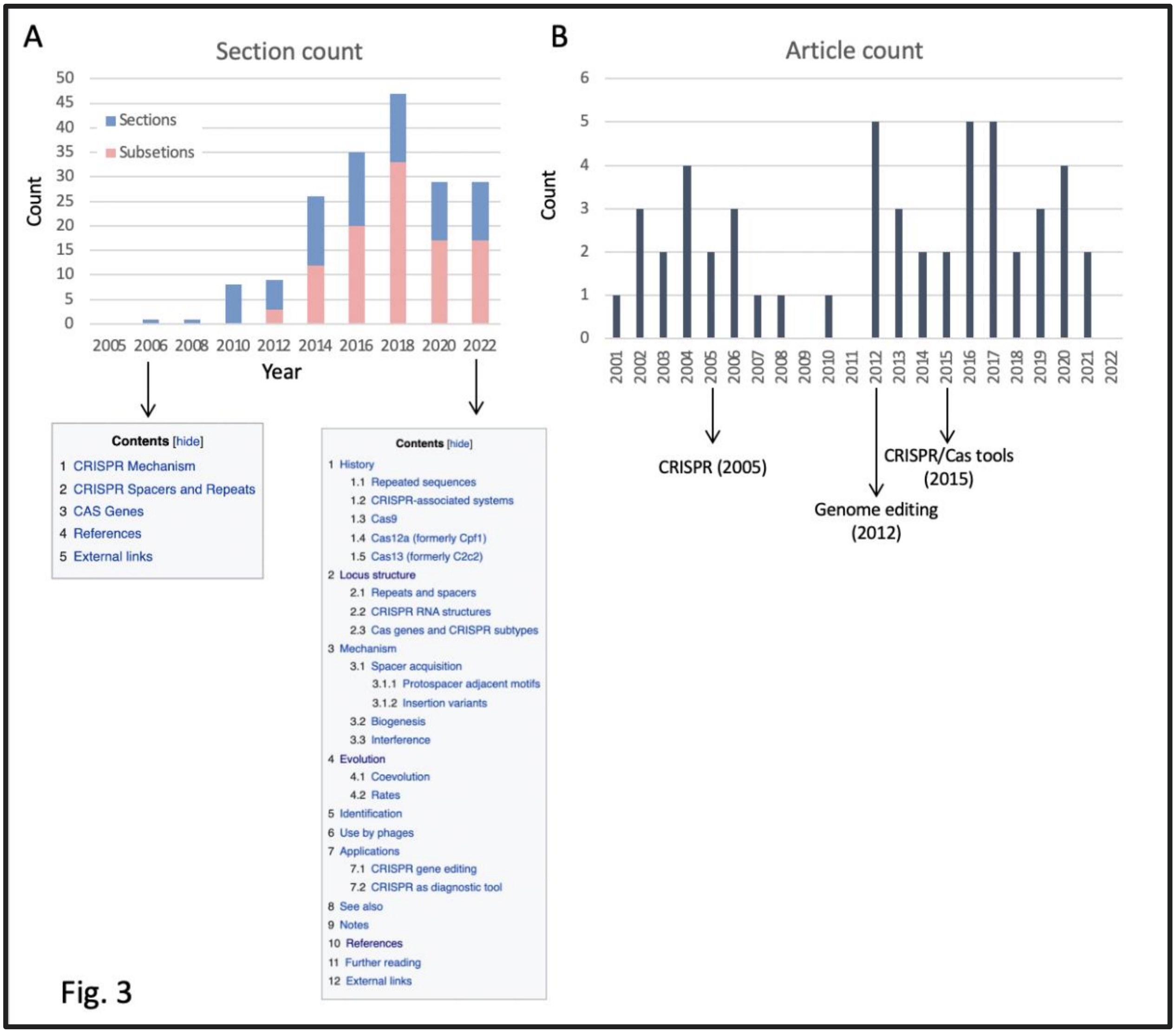
CRISPR are direct repeats found in the DNA of many bacteria and archaea. The name is an acronym for clustered regularly interspaced short palindromic repeats. These repeats range in size from 21 to 37 base pairs. They are separated by spacers of similar length. Spacers are usually unique in a genome. Different strains of the same species of bacterium can often be differentiated according to differences in the spacers in their CRISPR arrays, a technique called spoligotyping.

CRISPR (/ˈkrɪspər/) (an acronym for **clustered regularly interspaced short palindromic repeats**) is a family of DNA sequences found in the genomes of prokaryotic organisms such as bacteria and archaea.[2] These sequences are derived from DNA fragments of bacteriophages that had previously infected the prokaryote. They are used to detect and destroy DNA from similar bacteriophages during subsequent infections. Hence these sequences play a key role in the antiviral (i.e. anti-phage) defense system of prokaryotes and provide a form of acquired immunity.[2][3][4][5] CRISPR are found in approximately 50% of sequenced bacterial genomes and nearly 90% of sequenced archaea.[6]

Cas9 (or "CRISPR-associated protein 9") is an enzyme that uses CRISPR sequences as a guide to recognize and cleave specific strands of DNA that are complementary to the CRISPR sequence. Cas9 enzymes together with CRISPR sequences form the basis of a technology known as CRISPR-Cas9 that can be used to edit genes within organisms.[8][9] This editing process has a wide variety of applications including basic biological research, development of biotechnological products, and treatment of diseases.[10][11] The development of the CRISPR-Cas9 genome editing technique was recognized by the Nobel Prize in Chemistry in 2020 which was awarded to Emmanuelle Charpentier and Jennifer Doudna.[12][13]

Fig.2

Figure 2. Comparing versions of the CRISPR article. A snapshot from the Wikipedia archive of A) the full text of the CRISPR article when it first opened on June 30th 2005, and B) the lead section's opening paragraphs, as of July 6th, 2022.



Contents [hide]

1 CRISPR Mechanism
2 CRISPR Spacers and Repeats
3 CAS Genes
4 References
5 External links

Contents [hide]

1 History
1.1 Repeated sequences
1.2 CRISPR-associated systems
1.3 Cas9
1.4 Cas12a (formerly Cpf1)
1.5 Cas13 (formerly C2c2)
2 Locus structure
2.1 Repeats and spacers
2.2 CRISPR RNA structures
2.3 Cas genes and CRISPR subtypes
3 Mechanism
3.1 Spacer acquisition
3.1.1 Protospacer adjacent motifs
3.1.2 Insertion variants
3.2 Biogenesis
3.3 Interference
4 Evolution
4.1 Coevolution
4.2 Rates
5 Identification
6 Use by phages
7 Applications
7.1 CRISPR gene editing
7.2 CRISPR as diagnostic tool
8 See also
9 Notes
10 References
11 Further reading
12 External links

Fig. 3

Figure 3. Growth of CRISPR on Wikipedia - anchor article and corpus. A) The number of sections and subsections in the CRISPR article, since it was opened in 2005. B) The number of the corpus' articles opened since Wikipedia was launched (2001).

3. Cross-pollination: CRISPR as a body of knowledge

Shifts at the corpus level showed that knowledge on Wikipedia is rarely confined to a single article, but is rather stored in groups of articles that are constantly changing and cross-pollinate one another. On Wikipedia, this process can take on two distinct forms: new articles opening about the topic that directly address it, or existing articles changing to include new text, references or sections dedicated to the scientific topic's intersection with other bodies of knowledge. Tracking the migration between articles can illuminate how knowledge diffuses. To better understand the temporal aspect of CRISPR's representation across articles on Wikipedia we next compared the DOB of the different articles in our CRISPR corpus and the date the term CRISPR was first mentioned in them.

Of the 50 articles in the CRISPR corpus, 26 already had the term "CRISPR" in their first version (Fig. 4A). Among these were the articles for researchers like Charpentier, Šikšnys and Mojica. This group also included articles for scientific topics discovered in later stages of the CRISPR field's growth, like "Cas12", and articles reflecting CRISPR in culture, like the aforementioned academic journal. With few exceptions, like "CRISPR" and "CRISPR interference", opened in 2005 and 2010, respectively, articles that were created with CRISPR already mentioned in their first version were mostly opened post-2014 (Fig. 4B).

The 24 articles that lacked "CRISPR" in their inception provide insight into the growth of the field over time. Importantly, many concepts now associated with CRISPR did actually exist prior to its discovery or its application in gene editing was known. A prime example, "Gene knockout" and "Gene knockdown" existed as articles prior to CRISPR. However, as we saw, in a later stage their content was recast to take CRISPR into account and the articles were retroactively affiliated with the CRISPR field (in 2017 and 2013, respectively). Similarly, "Genome editing" was opened in 2012 but mentioned CRISPR only in 2014. The article "Designer baby", opened in 2005, initially only as a theoretical issue used in "popular scientific and bioethics literature." However, this changed with CRISPR's rise to prominence and since 2018 it directly referenced CRISPR, with a lengthy debate in wake of the "He Jiankui affair", in which the Chinese scientist created in 2018 the world's first so-called CRISPR babies in a widely reported incident.

We could also observe CRISPR's interface with other scientific fields through articles related to wider topics. For example, the two oldest articles in the corpus, "Wheat" and "Antibiotic", were opened in 2001, and were late to adopt "CRISPR" some twenty years later.

In sum, this analysis revealed a clear divide between articles that mentioned CRISPR from the onset and those that incorporated the term only in later stages: In general, this analysis underscores how CRISPR ramified across Wikipedia not just in the form of new articles, but also recasting older ones.

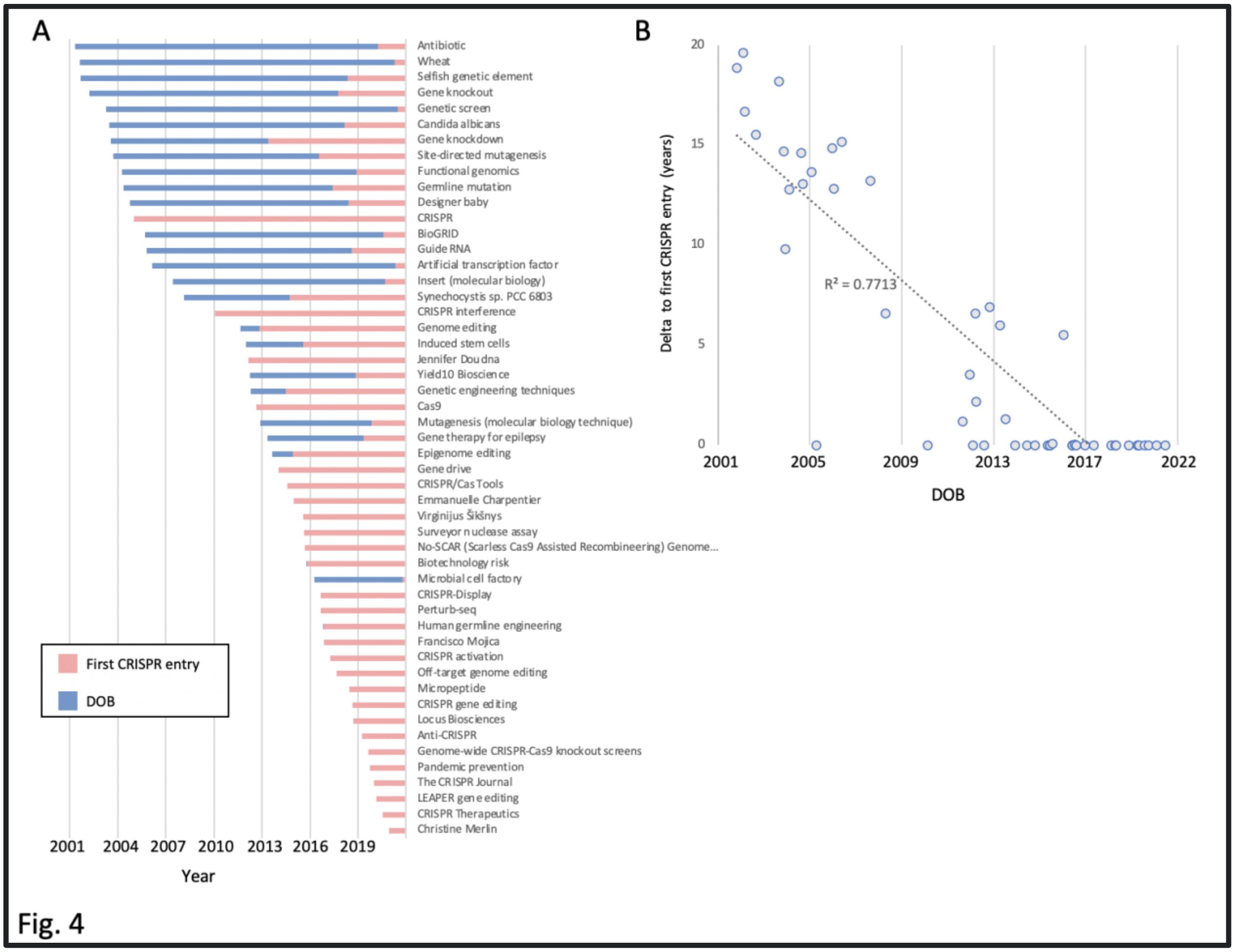


Fig. 4

Figure 4. Comparing an article's creation date and CRISPR's first mentions. A) An article's date of birth (DOB, blue) compared to the year of its first mention of the term CRISPR (red), sorted by the former. B) The relation between the DOB and the time it took for the first mention of CRISPR of each article. Displayed is a linear trendline and R^2.

4. From lab to public: Wikipedic bibliometrics map the diffusion of knowledge over time

All claims on Wikipedia need to be attributed to a verifiable source[32]. For our purposes, these references constitute substance for additional analyses: combining quantitative bibliometric analyses like citation count, with a content-dependent evaluation of the actual sources, to better understand the types of references supporting the “anchor” article. Quantitatively, we have previously developed two bibliometric analyses for Wikipedia articles - the “SciScore”, which gauges the ratio of academic to non-academic sources[11], and the “Latency”, which gauges the duration between an academic paper’s publication and when it was referenced in a Wikipedia article[33]. The reference list of each article in the corpus is parsed to break down the identity of its different sources: “.org”, “.com” and those containing DOIs/PMIDs/PMCs (i.e., scientific papers). Thus, we can assign a SciScore at both the corpus level and that of an individual article.

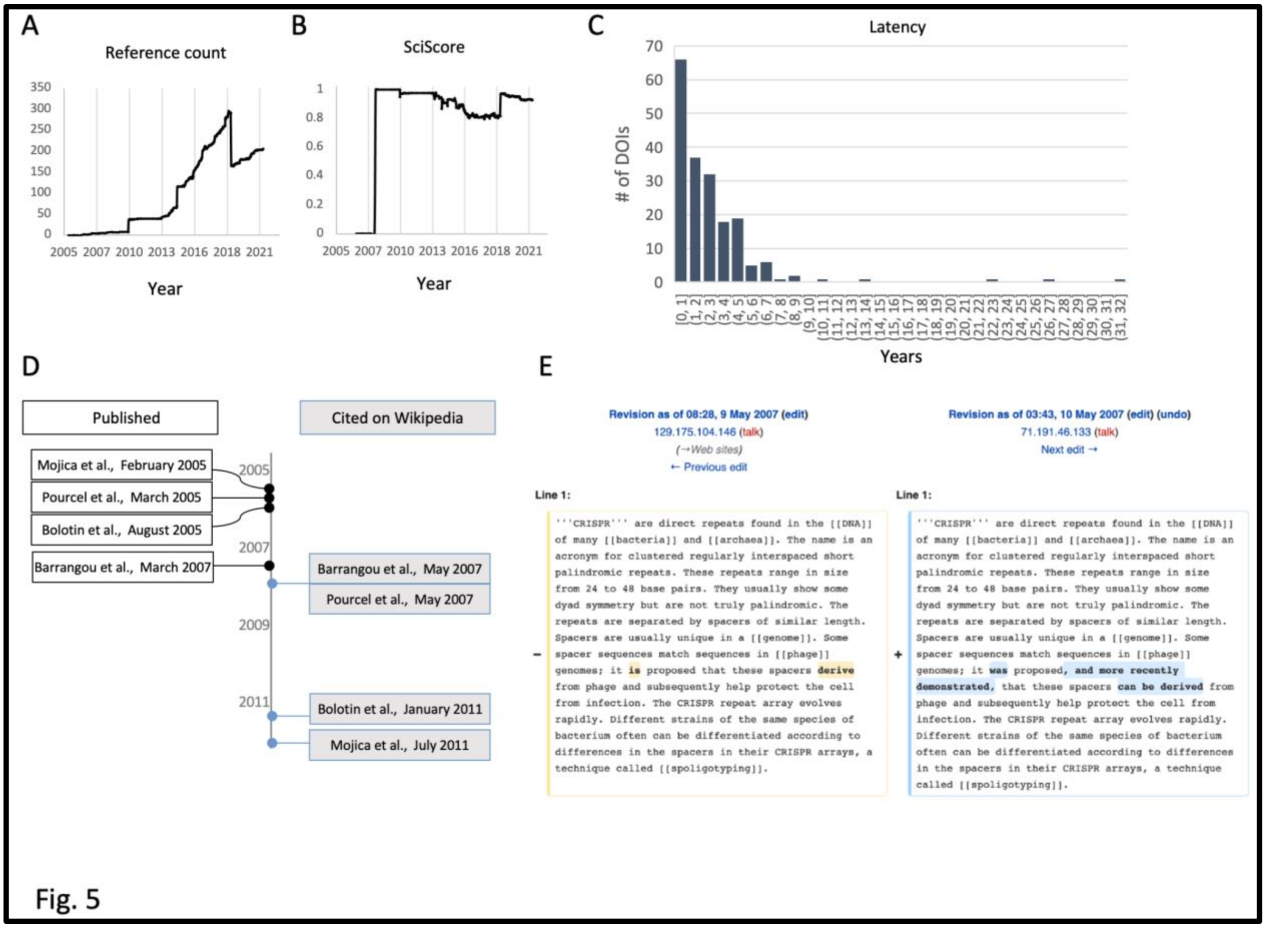
We found that the CRISPR anchor article was supported by 208 external sources in its “References” and “Further reading” sections (Fig. 5A). The article’s SciScore was 0.92 (out of 1), ranking 13/51 in the corpus (Figs. 5B and S2A). The top cited journal was *Science* (23 papers), followed by *Nature* and *Cell* (14 each), (Fig. S2B and S2C). These results are consistent with previous analyses of Wikipedia articles focused on scientific topics that show that these make use of peer reviewed, high-impact factor academic publications[4,8].

To attain a historical perspective, we next analyzed the temporal aspect of the above discussed bibliometric parameters, which were compared and contextualized to the changes in sections (Fig. 3A). We found that these metrics, and overlapping trends between them, served as markers for important events in the history of the field. A prime example of this can be seen in the aforementioned “Patents” section: on March 6, 2014 Doudna’s and Charpentier’s patent application was published online and a few weeks later the “Patents” section was opened in the

CRISPR article (Table S3). It cited the US Patent Office website. By 2015, after the Broad Institute was awarded its own patent and the appeal against it was filed by the universities representing Doudna and Charpentier, the article's text changed to indicate that, “As of December 2014, patent rights to CRISPR were still developing.” The text also noted that there was “a bitter fight over the patents for CRISPR”, a claim supported by this new type of citation which grew increasingly present in the CRISPR article: non-academic sources, in the form of both news articles about the legal cases and even the patents themselves. For example, the claim about the “bitter” legal battle was sourced to a story in MIT Technology Review, a popular science news site, while also referring directly to specific patents and or formal application documents made public online. Overall, the section included a laundry list of patent holders and claimants with a hodgepodge of popular and legal sources as citations. Throughout its entire existence, all the sources in this section were non-academic.

The fact that non-academic sources were deployed in the article to support non-academic aspects of the CRISPR history shows how these types of sources can document non-scientific ramifications of scientific developments. However, the entrance of non-academic sources was not limited to patent debates and also touched on CRISPR's growing social prominence. For example, the 2015 selection of CRISPR as “Breakthrough of the year”[34] was supported by links to popular media sources. Together with the patent links, these non-academic sources led to a decrease in the article’s SciScore during this phase (Fig. 5B).

Collectively, these highlight how bibliometric shifts are reflective of substantive changes in the article’s texts, which in turn are reflective of real-world developments in the field, both in terms of the science and of the social debates it inspires.



Revision as of 08:28, 9 May 2007 (edit)
129.175.104.146 (talk)
(→Web sites)
← Previous edit

Revision as of 03:43, 10 May 2007 (edit) (undo)
71.191.46.133 (talk)
Next edit →

Line 1:

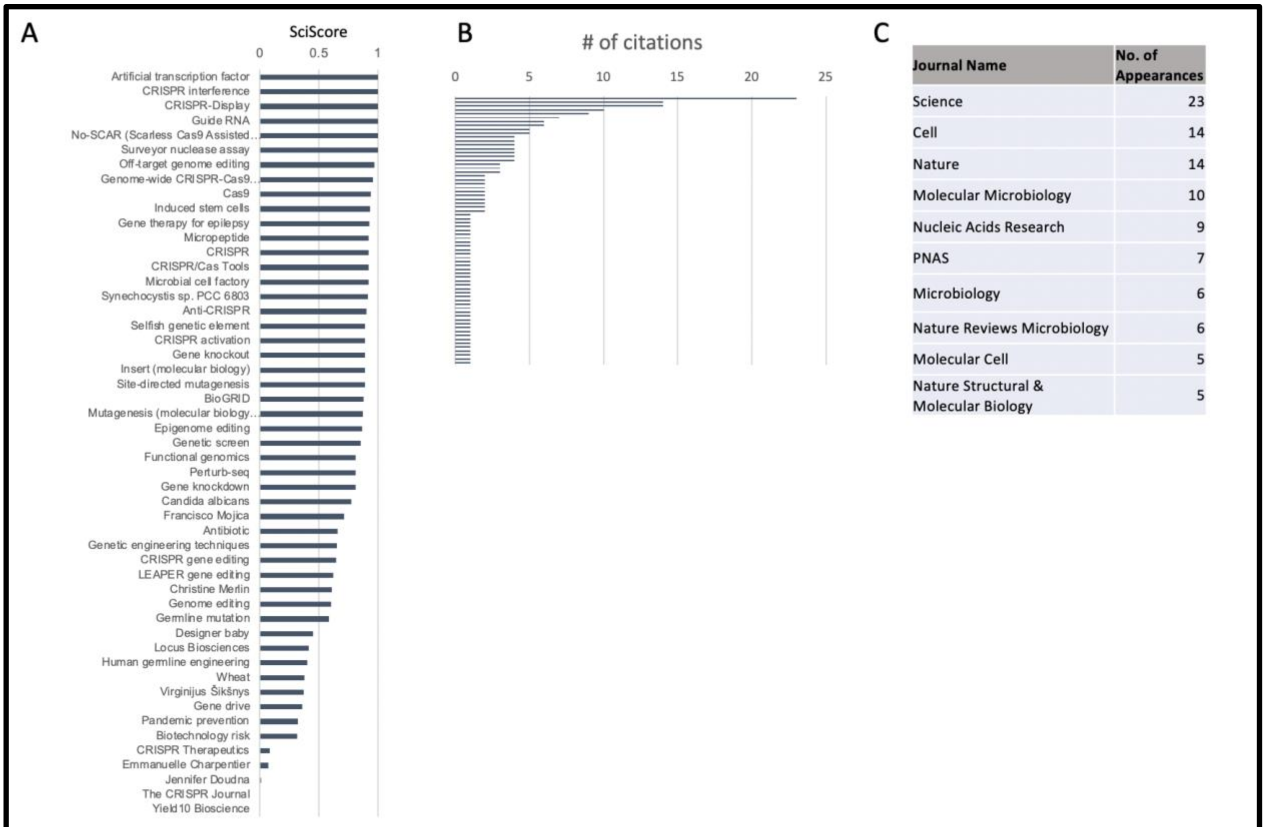
− '''CRISPR''' are direct repeats found in the [[DNA]] of many [[bacteria]] and [[archaea]]. The name is an acronym for clustered regularly interspaced short palindromic repeats. These repeats range in size from 24 to 48 base pairs. They usually show some dyad symmetry but are not truly palindromic. The repeats are separated by spacers of similar length. Spacers are usually unique in a [[genome]]. Some spacer sequences match sequences in [[phage]] genomes; it **is** proposed that these spacers **derive** from phage and subsequently help protect the cell from infection. The CRISPR repeat array evolves rapidly. Different strains of the same species of bacterium often can be differentiated according to differences in the spacers in their CRISPR arrays, a technique called [[spoligotyping]].

Line 1:

+ '''CRISPR''' are direct repeats found in the [[DNA]] of many [[bacteria]] and [[archaea]]. The name is an acronym for clustered regularly interspaced short palindromic repeats. These repeats range in size from 24 to 48 base pairs. They usually show some dyad symmetry but are not truly palindromic. The repeats are separated by spacers of similar length. Spacers are usually unique in a [[genome]]. Some spacer sequences match sequences in [[phage]] genomes; it **was** proposed**, and more recently demonstrated,** that these spacers **can be derived** from phage and subsequently help protect the cell from infection. The CRISPR repeat array evolves rapidly. Different strains of the same species of bacterium often can be differentiated according to differences in the spacers in their CRISPR arrays, a technique called [[spoligotyping]].

Fig. 5

Figure 5. CRISPR bibliometrics on Wikipedia. A) The number of references in the "CRISPR" article's reference section since it opened until December 2021. B) "CRISPR"s SciScore (until December 2021). C) The article's references latency distribution (i.e., delay between a scientific paper's publication and its integration into Wikipedia). D) A timeline comparing the date of selected publications (black frames, left) to their citation in the CRISPR article (blue frames, right). E) A side-by-side comparison of two versions of the CRISPR article from May 2007, showing how changes to the wording of the text were linked to the citation of Barrangou et al., 2007.



Journal Name	No. of Appearances
Science	23
Cell	14
Nature	14
Molecular Microbiology	10
Nucleic Acids Research	9
PNAS	7
Microbiology	6
Nature Reviews Microbiology	6
Molecular Cell	5
Nature Structural & Molecular Biology	5

Fig. S2

Figure S2. CRISPR article's references. A) The corpus SciScore. B) Peer-reviewed journals cited as references in the article as of June 2022, sorted by the number of references per publication. C) A list of the top cited journals (from B) with ≥5 appearances.

To better understand the relationship between Wikipedia and the sources supporting its articles we also conducted bibliometric analysis on the corpus, too. Thus, we found a number of articles with high SciScores (like "CRISPR interference" or "Cas9") alongside those with low percentage of academic sources, like that for Mojica or the concept of designer babies (Fig. S2A). This indicates a correlation between the scientificness of an article's topic and its SciScore, with biographical articles for scientists, for example, usually ranking lower than those for scientific concepts.

The "CRISPR" article ranked high in terms of SciScore. To gauge its current score with the state of the available research, we determined the latency of all the article's references. This analysis revealed a distribution varying between a single day to >30 years, with a median latency of 1.7 years (Fig. 5C). This bibliometric data can be contextualized through the example of the integration dynamics of publications relating CRISPR to bacterial immunity (Fig. 5D). Rodolphe Barrangou was the R&D director of genomics at DuPont chemicals manufacturer, who was first to have harnessed CRISPRs to provide immunity for their industrial bacterial strains. The resulting study was published in 2007, and was integrated into Wikipedia that year, a mere two months after going online. In this edit the text changed from "it **is proposed** that these spacers **…** protect the cell from infection" to "it **was** proposed, **and more recently demonstrated**, that these [can…] help protect the cell from infection" (bold added), (Fig. 5E). Only after this experimental demonstration were three landmark yet theoretical papers from 2005 that computationally supported the bacterial immune system hypotheses added to the article, and with a relatively large latency: Pourcel et al., 2005 was added two years after its publication, while Mojica et al., and Bolotin et al., were added only in 2011 - six years after publication. By this time, the text and the early references, as well as CRISPRs function in bacterial immunity and the experimental evidence - were all inserted into the article's lead section, too. These quantitative shifts in bibliometrics, we found, were the result of textual changes in the article, which reflected changes in the science itself.

5. Quantitative comparison between fields on Wikipedia

To examine whether the aforementioned methodology can provide insight into other scientific fields on Wikipedia, we developed an automated tool which generates corpuses along the aforementioned funnel (Fig. 1A) - and can be deployed on any search term of interest. The automated corpus creation is followed by a number of subsequent data collections that together form our suggested method and allow for cross field comparisons.

Alongside CRISPR, we deployed the tool on two additional terms- "Circadian" and "Coronavirus", which we have studied in different manners in earlier preliminary studies[33,11] and thus serve as control groups to some degree. We hence created three corpuses side by-side, at roughly the same time - June/July 2022, and demonstrated some of the aforementioned quantitative analyses.

As we observed for the CRISPR field, a substantial number of articles can be identified and selected to be part of the corpus - with 51, 138, and 306 articles for “CRISPR”, “Circadian”, and “Coronavirus”, respectively (Fig. 6, Tables S4 and S5). A subjective reading of the titles comprising these corpuses validated that they provide a diverse assortment of articles of different types that are relevant to each field - for example, articles for scientists alongside those for scientific terms or events. Thus, the corpus for “Circadian” yielded the articles “Circadian rhythms” and “Sleep”, and the corpus for “Coronavirus” yielded articles both about the pandemic like “COVID-19 pandemic in Japan” and more generally for “Virus”.

After an initial corpus creation, the first automated analysis generates a timeline based on each articles’ DOB. A side-by-side view of all three corpus timelines (Fig. 6A) illustrates how different fields display different modes of growth. For example, the “Coronavirus” timeline reveals a clear divide between scientific articles like “Pandemic” (2001) and “Spike protein” (2006), created early on in Wikipedia’s history, and post-pandemic articles like “Wuhan Institute of Virology” (2020). This timeline clearly shows how, with the outbreak of the pandemic, articles about the virus ballooned, but also how these were supported by a network of preexisting articles[9]. Meanwhile, the “Circadian” timeline exhibits a seemingly random distribution of article creation, with anchor articles (“Circadian Clock” and “Circadian Rhythms”), and auxiliary articles opening regularly over time. Some DOBs appear to tell a compelling scientific story - e.g., Paul Hardin, first author of the landmark paper highlighted in the 2017 Nobel declaration[35], received an article in 2017 - but these seem anecdotal. Interestingly, the biannual peaks are likely a product of the American chronobiologist Eric Herzog’s university course[36], selected according to the students’ personal inclination. This DOB pattern or lack thereof can be explained by the fact that unlike the timeliness of CRISPR or coronavirus, clocks are a more mature field whose growth, as our previous work has shown, is reflected in a more subtle manner on Wikipedia, with a paradigmatic shift in the field being documented in minute nuanced textual detail[33].

One similarity between all three timelines is an increase in article creation centered around 2005-7, a period which has been shown to have held a massive surge in article creation in Wikipedia in general[37].

Our tool also supports automated scraping of bibliometric data. This analysis showed that the top ten journal references in all three corpuses were dominated with high impact-factor academic peer-reviewed publications (Fig. 6B). Alongside prestigious scientific publications like Nature or PNAS, each corpus also included field-specific publications: For example, the Journal of Biological Rhythms in the Circadian list, Nature Biotechnology for CRISPR, or The Journal of Virology for coronavirus.

Non-academic references were also quite field-specific. As researchers from both the circadian clocks and CRISPR fields were awarded a Nobel Prize, the website for the prestigious award was among the most cited in the respective corpuses (Fig. 6C). In addition, the Sleep Foundation website was highly cited in the circadian corpus while three genome focused websites were highly cited in the CRISPR corpus. The International Committee on Taxonomy of Viruses (ICTV) was among the top 10 .org sites cited in the coronavirus corpus, which appears in the Wikipedia article for every variant.

In general, the CRISPR and Circadian corpuses relied more on scientific literature, while Coronavirus referenced mostly .com sources (Fig. 6D), which is also reflected in the different corpuses' SciScore (Fig. 6E). It appears the more prominent a scientific field is societally, the lower its scientific score: for example the non-scientifically focused CRISPR-corpus article about designer babies which had a relatively low score, as did the Circadian-corpus article of "Start school later movement." Meanwhile, the more clearly scientifically focused articles "Surveyor nuclease assay" and "CSNK1D" had high scores. The patterns of SciScore distribution show how different fields manifest differently and that comparing them can shed light, for example, on how much public, as opposed to purely scientific interest, a field has online. In summary, these analyses show how the same research tools and methods yield very different results for different research fields, all of which can facilitate the initial steps needed towards the creation of future case studies into how scientific knowledge is represented on Wikipedia over time.

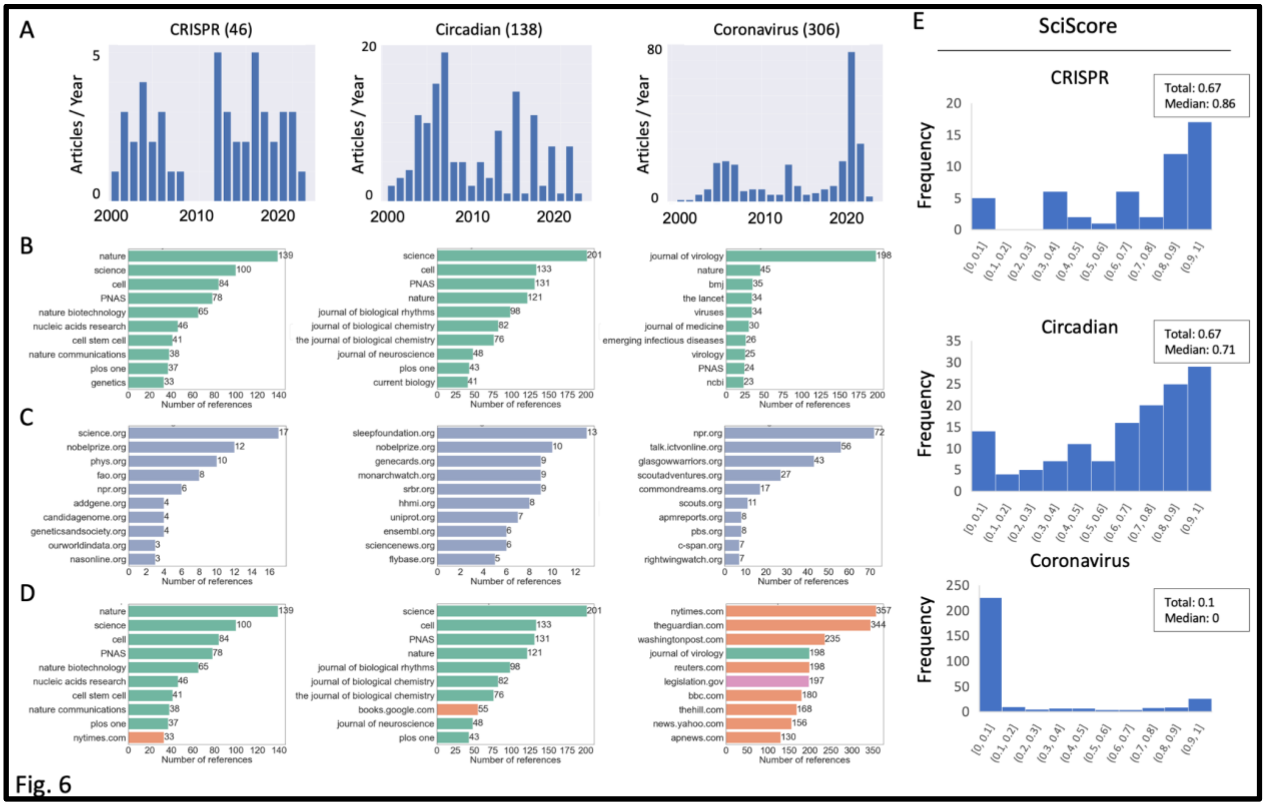


Fig. 6

Figure 6. Comparing Wikipedia corpuses: Different fields show different data. Corpuses were generated and quantitative metrics automatically collected in June-July 2022, for the terms "CRISPR", "Circadian" and "Coronavirus". The following data are presented: A) the number of articles opened each year, B) the top 10 most cited journals, C) the top 10 most cited .org websites, D) the top 10 most cited references altogether, E) SciScore distribution, along with the total (sum of all references) and median scores.

Discussion

Here, we examined the way CRISPR was represented on Wikipedia from the site's launch in January 2001 until 2022. By reviewing the CRISPR article's history, we saw that the article started off describing the "basic science" behind CRISPR, and was updated in the wake of the publication of canonical works in the field. Over time, the article grew, and with the emergence of gene editing technology it forked off into a number of affiliated articles with a more narrow focus, while the original CRISPR article offered a consolidated overview of the scientific narrative on CRISPR in bacterial systems. The article's text and its different citations served as a rich record of the growth of academic knowledge, the legal battles CRISPR sparked and the academic credit wars over what the journal *Science* called the "CRISPR Craze"[38], as well as the popular interest in the field.

We thus propose this method can be used to perform history of contemporary science on other topics using Wikipedia. This begins with corpus delineation, followed by a historical analysis of the sections of the anchor article and the timeline of all corpus articles. Both quantitative and qualitative methods are used to track these dynamics, augmented with bibliometric analyses - namely the SciScore and latency. Moreover, automated tools developed to support this research permit work on additional topics, though combining these with manual and semantic work are key to contextualizing findings and interpreting them to provide substantial historical insight.

Using Wikipedia for the history of science

Our findings join a small yet growing body of research dedicated to using Wikipedia for historical purposes. Previously, we analyzed the growth of two Wikipedia articles dedicated to the circadian clock field through their edit histories ("Circadian clocks" and "Circadian rhythms"), using them to ask whether the article's text reflected changes taking place in understanding how biological clocks work[33]. Within that more focused case-study we observed the importance of following the academic references, and developed the Latency metric. Meanwhile, our study on COVID-19 used large-scale quantitative bibliometrics to understand how the pandemic affected large swathes of articles during its “first wave”, putting forward metrics such as the SciScore to qualify hundreds of articles based on their reference list[11]. Collectively, these underscore the key role academic sources play on Wikipedia and serve as a wider proof-of-concept for the quantitative and qualitative underpinnings of this present study.

Wyatt suggested in a theoretical paper that Wikipedia could be used as a primary source in historical research[39]. From the edit history of articles, to metadata for traffic and even talk pages, he envisaged treating the open-source encyclopedia as an “endless palimpsest”. This is an idea that has also previously been expressed as an artwork: “The Iraq War: A Historiography of Wikipedia Changelogs” by artist James Bridle was 12-volume a book comprising all the versions of the article dedicated to the war in Iraq, with the online edit wars serving as a proxy for the real-world conflict. However, to our knowledge, no academic demonstration nor a clear method has yet been put forward as to how researchers can actually use Wikipedia to utilize Wikipedia’s historiographic potential to serve as this “endless palimpsest”, especially not in the interest of following shifts in science.

Different attempts to harness Wikipedia for historical ends were reported in recent years as computational methods permeate the non-exact and -natural sciences, including history and philosophy[40], through what is termed digital humanities[41]. For example, an algorithmic approach was deployed to mine the text of tens of thousands of Wikipedia articles to try to map the history of knowledge since the dawn of human history, using network science and semantic analysis to “put the ideas of Kuhn to the test”. The study, published as a preprint[5], makes interesting findings, but also shows the lack of a unification in methods in current Wikipedia-based historical research. There are numerous studies, for example, about Wikipedia and bibliometrics[4], even those that focus on science[8]; but none that clearly link scientometrics to historical methods[42]. Others from the more humanistic side of academia have worked to connect the digital arena to contemporary fields like discourse analysis, based on the works of Michele Foucault[43]. However, these too are all theoretical works and as of yet no programmatic paper has outlined how Wikipedia can be actually used for historical research. We hope our proposed method will encourage use of Wikipedia's ever-changing text as a rich historical source to augment existing work being done in the history of science and contribute to our understanding of the growth of scientific knowledge and its transference to the general public.

Why Wikipedia

Wikipedia easily lends itself to research of this type. A digital and open website that is easily searchable, it also provides a simple to use API for more complex queries and even a full dump of the entirety of Wikipedia in each language, including the full edit history of every article. Wikipedia's inherent structure allows comparable historical work across different fields, primarily since all articles are structured in a similar way: a lead text, table of contents, sections and then a reference list. Thus, cross-analyses of different subjects can yield results comparable through standardized metrics, like the DOB timelines, and the Latency or SciScore metrics for bibliometric comparisons. The structural similarity creates a sort of internal control that lays the groundwork for a rigid research system that can be utilized by others and applied to additional fields.

An initial method for selecting such future case studies could be to focus on the topics selected by Science and others as "Breakthrough of the Year" - these and their relevant Wikipedia articles are documented in a special list on Wikipedia[44] that could serve as the origin of many corpuses. Scientific developments that have garnered public interest over the past two decades, from the human genome project to Alpha Fold, could also serve as lucrative case-studies, each providing a unique dataset that could then be compared. Mapping out additional fields can eventually support theories/models of scientific growth in a resolution never before possible.

Moreover, unlike social media websites that collect user data, posing ethical dilemmas for researchers, Wikipedia collects no such information, making it and its data ideal for social research. Wikipedia's texts are not single-handedly written and are edited collectively in a form of what is termed peer-production. Though this system is not without its flaws, in the context of the contemporary history of science it proves a valuable resource: documenting the consensus regarding certain facts and fields' growth in real-time and in potentially minute details. Wikipedia provides a rich source of information as one can easily see past versions of these articles through what is termed the changelog. This continuum of text throughout time is a well-known historical practice using other sources, and compliments the classic analysis of historical scientific texts: reading changing versions of the same text as opposed to only comparing different scientific reviews and papers. This allows researchers to map the changes of specific parts of the article's text, structure and references and easily track new additions and deletions. Past versions that did not survive Wikipedia's mob review process or that included facts that were true at the time but have since been rendered obsolete prove especially interesting from the perspective of the history of science. For example, with CRISPR, a December 2005 version of the article described Cas1 as the "most important" of the cas systems, and one that is "present in almost every CRISPR/Cas system." This was more cautiously reworded in July 2010 so that, "The most important of the Cas proteins appears to be Cas1, which is ubiquitous" in CRISPR systems. In March 2011, Cas1's ubiquity was no longer said to be linked to its importance, and for the past decade the article has made due with noting in a subsection dedicated to CRISPR locus that "[m]ost CRISPR-Cas systems have a Cas1 protein." These changes were the result of new knowledge forcing a reevaluation of the preexisting scientific narrative regarding CRISPR: Cas1 was not falsified per se, rather its importance in CRISPR's story was reassessed. Another example from the CRISPR article can be seen in the shift in section title from "Potential Applications" to "Applications" regarding gene editing, which took place in November 2013. These are examples of what can be termed "negative" knowledge -

knowledge whose relevance was negated by new “positive” discoveries that outweighed it in significance. However, as such, its degradation of scientific status in CRISPR’s narrative, has much value from the historical perspective. Wikipedia, we suggest, is an inclusive media that documents both positive and negative knowledge, - the accumulation and the rejection of scientific facts through its edit history.

Wikipedic Bibliometrics

Bibliometrically, Wikipedia can be seen to be a much more inclusive than academic publications, making use of non-academic sources usually excluded from academic texts. As suggested above, we propose that the unique structure of Wikipedia facilitates comparison between different fields through the bibliometric analyses like SciScore and Latency. On CRISPR, for example, legal sources or popular media were added to support the “patent war”, which was also expressed in a drop in the article’s SciScore. The expansion and then contraction of the “Patents” section (Table S3), in tandem to the patent wars and their resolution, show how this historical inclusivity touches to both the text and to the sources. The SciScore reveals a different historical perspective when comparing the CRISPR and COVID-19 corpuses. We previously discovered a decrease in the SciScore as the pandemic grew to public prominence and more articles about it were opened[11]. This was because many of the new articles opened post-pandemic were about its social ramifications and outcomes, while the pre-pandemic articles focused on the science behind the virus. In the CRISPR anchor article, the SciScore revealed a completely different process: As CRISPR began as a purely scientific discovery, the decrease in SciScore (~2013-2018, Fig. 5A) was found to be the result of the appearance of the first non-academic sources about the looming “The CRISPR Craze”[38], followed by the much-publicized patent and credit wars, and finally the wider social, ethical and policy debates it sparked.

Latency analyses, which has yet to be successfully automated, revealed that CRISPR, a nascent field, was making use of extremely up-to-date papers, in some cases references were added within days of their publication. Meanwhile, the circadian clock article had a median latency of five years[33]. This coincides with the respective histories of the fields: CRISPR is a new emerging field, with advances in the field being mirror almost instantaneously on Wikipedia. On the other hand, clocks, which is a mature field that has been around for decades, was also found to be based on older research which predated Wikipedia. Meanwhile, COVID-19 had a major 17-year peak in latency, exactly in line with the SARS pandemic of 2003; hence, research from a preceding viral pandemic provided the backbone of the sourcing of the 2020 pandemic. Together these show how the character of each field is reflected in its bibliometrics.

One hypothesis regarding the potential of the SciScore and Latency is that this dynamic may also be taking place in other articles that began as purely scientific but are increasingly taking on social significance. Tracking articles that have short latencies and high SciScore which then begin to decrease could serve as a method for identifying new fields only now starting to make waves in terms of public interest.

Using Wikipedia bibliometrics also has value from the scientometric perspective. Measuring the impact of scientific research is a mature field that has in recent years expanded the metrics it works with - no longer just impact factor and citation counting, as new metrics like AltMetrics have emerged. In this sense, Wikipedia, too, can prove a valuable addition in the form of alternative metrics. Asking which papers are cited on Wikipedia and in which context, may provide insight into what parts of academic research are actually reaching the public. As such, our work can join and enrich existing studies on the history of contemporary science, augmenting their work in the field of bibliometrics or even Alt-Metrics, with Wikipedia.

The benefit of mixed methods

Our method can perhaps be best described as an example of “thick big data”[30], a data-driven sociological and semantically sensitive contextual reading. The data, in our case, is Wikipedia’s edit history and its sources, which are then analyzed through mixed methods and interpreted in a detailed manner.

The DOB timeline, for example, provides a qualitative dataset regarding the growth rate of the articles related to the topic, but a qualitative reading of their titles provides substantive context for this growth. The section analysis provides important quantitative insight regarding the article’s growth and structure while also permitting a semantic understanding of the architecture of knowledge and how it shifted over time as sections grew, contracted or migrated.

We suggest that employing these types of analyses is key to historical research into Wikipedia. The historical methods born with historian Derek J. de Solla Price that made use of publication data[42] joined the works of earlier thinkers like Robert K. Merton that laid the historiographic framework for historical research into the scientific revolution[45]. Later on, sociological works, written by historians like Robert Darnton on the history of books offered a qualitative detail-rich chronicle of the rise of scientific media during the Enlightenment, substantiating the scientometrics of history[46]. Along this line, we propose that analysis that is content-dependent and does not shy away from the semantic shifts is needed. Though tools, quantitative analyses and bibliometrics all help systematize research of Wikipedia, the historical work requires delving deep into the archive, so to speak. Hitherto, work of this type on Wikipedia was done either manually on a single article as aforementioned[33] - or others with a large-scale use of the entirety of Wikipedia as a dataset[47], analyzed for biometric trends[48], for example finding the most cited journals across English-language articles[4]. A mixed-method that meshes automatization and quantitative analyses with a textual reading to provide context and an “interpretive framework”[49] as suggested herein, has yet to be done with a focus on Wikipedia.

Limitations

For all its benefits, this method also has its shortcomings. To begin with, corpus lineation can exclude possibly valuable articles - for example, the article for George Church was absent from our corpus despite his seemingly important role in the history of CRISPR.

From a scientometric perspective, Wikipedia also poses some unique problems: Unlike bibliometric datasets created especially for such purposes, Wikipedia's footnotes are not all properly formatted and issues with their templates exist that make scrapping them consistently hard[50], especially with older articles. Initially, all footnotes on Wikipedia were added manually by editors working directly in wiki-code, the HTML markup language the website uses. Over time, bots and tools were put into place to help this menial task and unify footnotes formatting; in some cases, older articles with older footnotes that did not benefit from this unified new formatting will not be scrapped properly if one uses only Wikipedia's native bibliometric data. To overcome this issue in the present study, we scraped the references from the articles as simple text, regardless of how they were formatted by Wikipedia's volunteer editors. This list of references was then analyzed in search of DOIs/PMIDs/PMCs which were taken as a proxy for academic publications. Nonetheless, other issues exist, for example duplicate DOIs or DOIs included in article's texts and not just as footnotes. A manual validation of our method in random articles revealed this approach had a margin of error that was lower than 5 percent.

Moreover, our method also does not yet address all of Wikipedia's content: For example, the talk page, a key arena in Wikipedia and one that is rich in textual data, was not systematically included in this study, though debates about the patent war were found, and these included discussions of which type of sources (legal as opposed to scientific) should be cited on the article in this context. Another facet of Wikipedia we did not address touches on visual elements. Wikipedia's sister project, WikiCommons, supports multimedia, usually in the form of copyright-free images, and in this respect we also saw a growth: The first infographic explaining the CRISPR system was introduced to the article in 2009 and the file itself was updated in 2010 to show a more complex understanding of the "CRISPR prokaryotic antiviral defense mechanism", supported by a then-newly published review article[51]. Over time, additional more complex images were added to the article, for example those showing how CRISPR interference could be used for gene editing (Fig. 7). This multimedia aspect can serve in the future as a rich arena for like-minded research, for example by focusing on how infographics and scientific illustrations document growth of scientific knowledge overtime.

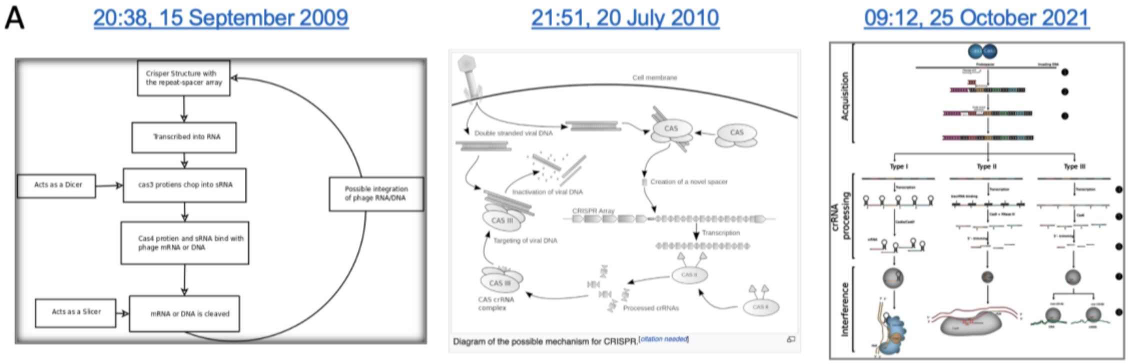


Diagram of the possible mechanism for CRISPR.[citation needed]

Fig. 7

Figure 7. Illustrations of the CRISPR model. Shown are a selection of screen grabs from the CRISPR article, reflecting the evolution of Wikicommons graphics of CRISPR's mechanism of action and key players. These are of different versions of the same illustration (A and B) and of a third illustration added later to the article.

Data accessibility

Our code for the corpus builder can be found at: https://github.com/RonaTheBrave/WikiCorpusBuilder

The article's data is accessible at https://zenodo.org/record/7206381#.Y1JoEezP23I DOI:10.5281/zenodo.7206381.

Competing interests

The authors declare no competing interests.

Acknowledgements

We want to thank Dusan Misevic, Bastian Greshake Tzovaras, Marc Santolini, Mad Price Ball and all those who provided feedback.

Funding

Thanks to the Bettencourt Schueller Foundation long term partnership, this work was partly supported by the LPI Research Fellowship, Université de Paris, INSERM U1284, to RAv and OB.

Footnotes

References

1. Sepkoski, D. Towards “A Natural History of Data”: Evolving Practices and Epistemologies of Data in Paleontology, 1800–2000. *J. Hist. Biol.* **46**, 401–444 (2013).

2. Rheinberger, H.-J. Infra-Experimentality: From Traces to Data, from Data to Patterning Facts. *Hist. Sci.* **49**, 337–348 (2011).

3. Maree, D. J. F. The Methodological Division: Quantitative and Qualitative Methods. in *Realism and Psychological Science* 13–42 (Springer International Publishing, 2020). doi:10.1007/978-3-030-45143-1_2.

4. Teplitskiy, M., Lu, G. & Duede, E. Amplifying the impact of open access: Wikipedia and the diffusion of science. *J. Assoc. Inf. Sci. Technol.* **68**, 2116–2127 (2017).

5. Ju, H. *et al.* The network structure of scientific revolutions. (2020).

6. Maggio, L. A., Steinberg, R. M., Piccardi, T. & Willinsky, J. M. Reader engagement with medical content on Wikipedia. *eLife* **9**, (2020).

7. Jemielniak, D., Masukume, G. & Wilamowski, M. The Most Influential Medical Journals According to Wikipedia: Quantitative Analysis. *J. Med. Internet Res.* **21**, e11429–e11429 (2019).

8. Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E. & Romero-Frías, E. Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE* **15**, e0228713–e0228713 (2020).

9. Wikipedia contributors. Wikipedia — Wikipedia, The Free Encyclopedia. (2022).

10. *Wikipedia @ 20: Stories of an Incomplete Revolution*. (The MIT Press, 2020). doi:10.7551/mitpress/12366.001.0001.

11. Benjakob, O., Aviram, R. & Sobel, J. A. Citation needed? Wikipedia bibliometrics during the first wave of the COVID-19 pandemic. *GigaScience* **11**, (2022).

12. Couto, L. & Teixeira Lopes, C. Equal opportunities in the access to quality online health information? A multi-lingual study on Wikipedia. in 1–13 (ACM, 2021). doi:10.1145/3479986.3480000.

13. Cohen, J. A cut above: pair that developed CRISPR earns historic award. *Science* **370**, 271–272 (2020).

14. Jansen, R., Embden, J. D. A. van, Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–75 (2002).

15. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–82 (2005).

16. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiol. Read. Engl.* **151**, 653–663 (2005).

17. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiol. Read. Engl.* **151**, 2551–2561 (2005).

18. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–21 (2012).

19. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2579-86 (2012).

20. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–23 (2013).

21. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–6 (2013).

22. Han, W. & She, Q. CRISPR History: Discovery, Characterization, and Prosperity. *Prog. Mol. Biol. Transl. Sci.* **152**, 1–21 (2017).

23. Marcon, A., Master, Z., Ravitsky, V. & Caulfield, T. CRISPR in the North American popular press. *Genet. Med.* **21**, 2184–2189 (2019).

24. Wolfe, J. *U.S. appeals court upholds MIT, Harvard patents on CRISPR gene editing*. https://www.reuters.com/article/us-ucberkeley-crispr-idUSKCN1LQ1XP (2018).

25. Lander, E. S. The Heroes of CRISPR. *Cell* **164**, 18–28 (2016).

26. Why Eric Lander's Controversial Paper "The Heroes of CRISPR" Is Not Solid Historical Research. *American Scientist* https://www.americanscientist.org/blog/macroscope/why-eric-lander%E2%80%99s-controversial-paper-%E2%80%9Cthe-heroes-of-crispr%E2%80%9D-is-not-solid-historical (2017).

27. "Heroes of CRISPR" Disputed. *The Scientist Magazine®* https://www.the-scientist.com/news-opinion/heroes-of-crispr-disputed-34188.

28. The Villain of CRISPR. https://www.michaeleisen.org/blog/?p=1825.

29. *Debates in the Digital Humanities 2019*. (University of Minnesota Press, 2019). doi:10.5749/j.ctvg251hk.

30. Jemielniak, D. *Thick big data: doing digital social sciences*. (Oxford University Press, 2020).

31. Wikipedia contributors. Notability in the English Wikipedia — Wikipedia, The Free Encyclopedia. (2022).

32. Wikipedia:Verifiability. *Wikipedia* (2022).

33. Benjakob, O. & Aviram, R. A Clockwork Wikipedia: From a Broad Perspective to a Case Study. *J. Biol. Rhythms* **33**, 233–244 (2018).

34. And Science's 2015 Breakthrough of the Year is... | Science | AAAS. https://www.science.org/content/article/and-science-s-2015-breakthrough-year.

35. The Nobel Prize in Physiology or Medicine 2017. *NobelPrize.org* https://www.nobelprize.org/prizes/medicine/2017/press-release/.

36. Chiang, C. D. *et al.* Learning chronobiology by improving Wikipedia. *J. Biol. Rhythms* **27**, 333–336 (2012).

37. Suh, B., Convertino, G., Chi, E. H. & Pirolli, P. The singularity is not near: slowing growth of Wikipedia. in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration - WikiSym '09* 1 (ACM Press, 2009). doi:10.1145/1641309.1641322.

38. Pennisi, E. The CRISPR Craze. *Science* **341**, 833–836 (2013).

39. Wyatt, L. Endless palimpsest: Wikipedia and the future's historian. *Stud. High. Educ.* **45**, 963–971 (2020).

40. Zurn, P. & Bassett, D. S. Seizing an opportunity. *eLife* **8**, e48336 (2019).

41. Su, F. & Zhang, Y. Research output, intellectual structures and contributors of digital humanities research: a longitudinal analysis 2005–2020. *J. Doc.* **78**, 673–695 (2022).

42. Price, D. J. D. S. *Little Science, Big Science*. (Columbia University Press, 1963). doi:10.7312/pric91844.

43. Gredel, E. Digital discourse analysis and Wikipedia: Bridging the gap between Foucauldian discourse analysis and digital conversation analysis. *J. Pragmat.* **115**, 99–114 (2017).

44. Breakthrough of the Year. *Wikipedia* (2022).

45. Merton, R. K. & Shapere, D. *The Sociology of Science: Theoretical and Empirical Investigation*. *Phys. Today* **27**, 52–53 (1974).

46. Porter, C. A. & Darnton, R. The Business of Enlightenment: A Publishing History of the Encyclopedie 1775-1800. *Eighteenth-Century Stud.* **13**, 335 (1980).

47. Yang, D., Halfaker, A., Kraut, R. & Hovy, E. Identifying Semantic Edit Intentions from Revisions in Wikipedia. in 2000–2010 (Association for Computational Linguistics, 2017). doi:10.18653/v1/D17-1213.

48. Mostafa, M. M. Two decades of Wikipedia research: a PubMed bibliometric network analysis. *Glob. Knowl. Mem. Commun.* (2021) doi:10.1108/GKMC-03-2021-0056.

49. Rowlands Bruce. Grounded in Practice: Using Interpretive Research to Build Theory. *Electron. J. Bus. Res. Methodol.* **3**, 81–92 (2005).

50. Pooladian, A. & Borrego, Á. Methodological issues in measuring citations in Wikipedia: a case study in Library and Information Science. *Scientometrics* **113**, 455–464 (2017).

51. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170 (2010).