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 for 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 (Sepkoski 2013), with research increasingly making more use of big data and computational techniques for historical ends (Rheinberger 2011). Alongside the rise of digital humanities, a divide has persisted between quantitative historical research and textually rich qualitative work. The result is a historiographic lacuna (Maree 2020) and a debate regarding the materials and methods which can be researched (Lean, Rivelli, and Pence 2021)(Pence and Ramsey 2018). Here, we suggest a new resource can be useful in the context of history of contemporary science by systematizing research methods on an unlikely arena that is rich in both bibliometric data and historical text: Wikipedia.

Wikipedia's volunteer-run-editorial process was long lambasted for being inherently unreliable. When in 2005 Wikipedia was found by a study by *Nature*'s newsite to be as reliable as Encyclopedia Britannica - the news was met with surprise (Giles 2005)(Heather Ford 2020). Issues of accuracy in the encyclopedia anyone could edit, along with "edit wars" between volatile editors, even became a topic of research (Wilson and Likens 2015)(Yasseri et al. 2012)(Heather Ford 2020). In recent years however, this narrative has reversed: both academic research and the media have praised Wikipedia's coverage and reliance on sources, which in some cases have been found to be in lock step with science (Jemielniak 2019)(Reagle and Koerner 2020)(Smith 2020), especially in regards to the COVID-19 pandemic (Benjakob, Aviram, and Sobel 2022)(Colavizza 2020).

If Wikipedia is meetings its self-defined goal of representing the scientific consensus on scientific knowledge while making use of scientific sources, then we wanted see if this was also true historically: Can Wikipedia articles document shifts in science *over time*? Using Wikipedia's edit history, we hypothesized, could allow to see changes in how a field was represented in the past and allow the tracking of citations, with new ones being added and old ones removed as changes took place and new knowledge accumulated, sparking reassessments of existing

paradigms. Our first study showed how two Wikipedia articles on the Nobel Prize-winning field of circadian clocks managed to accurately reflect the field for over 15 years, even as it underwent scientific revolutions (Benjakob and Aviram 2018). Others have used Wikipedia in a similar manner to focus on documentation of political events such as the 2011 Egyptian Revolution (H. Ford 2022).

Wikipedia provides a rich, open, and accessible 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 compliments the traditional historical practice of classic analysis of historical scientific texts, yet in this case the reading is of changing versions of the same text as opposed to comparing different scientific reviews and papers. For us, this feature raised the possibility to map the changes of specific parts of the article's text, structure and references and easily track new additions and deletions.

Academic sources and the study of academic publications are mainstays for the history of science. Both qualitative and quantitative researchers make use of them - be it for bibliometric analysis or thick description (Lean, Rivelli, and Pence 2021) (Darnton 1985).

The historical methods born with historian Derek J. de Solla Price that made use of publication data (Price 1963) joined the works of earlier thinkers like Robert K. Merton that laid the historiographic framework for historical research into the scientific revolution (Merton and Shapere 1974). 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 (Darnton 1982)(Porter and Darnton 1980).

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. In that, Wikipedia also provides a textually rich base for qualitative work regarding narrativization and the communication of

science outwards from academia to the public. Importantly, policies enforced by Wikipedia's editors require "verifiable" sources to back all factual claims ("Wikipedia:Verifiability" 2022), and every article has a reference list. A small but growing body of bibliometric research based on Wikipedia has also emerged (Teplitskiy, Lu, and Duede 2017)(Singh, West, and Colavizza 2021) and even found that on medical (Maggio et al. 2020) and science (Arroyo-Machado et al. 2020) topics English-language articles have an explicit bias towards using academic sources. Wikipedia thus easily lends itself to such research, providing both data and text that can be used for historical analysis. In this work, we demonstrate this through a case study on the CRISPR field.

In a relatively short period of time, CRISPR-based gene-editing tools have been labeled the scientific "breakthrough" of the 21st century (Cohen 2020). While CRISPRs were identified in the 1980's, and received their name in 2002 (Jansen et al. 2002), their function remained unclear until 2005, when different labs deduced from *in silico* studies that CRISPR sequences were part of a bacterial adaptive immune system (Bolotin et al. 2005)(Mojica et al. 2005)(Pourcel, Salvignol, and Vergnaud 2005). 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 (Jinek et al. 2012) and shortly after in a paper of the Virginijus Šikšnys group (Gasiunas et al. 2012). These were rapidly followed by publications in February 2013 that performed genetic engineering in vivo in mammals, led by scientists Fang Zhang (Cong et al. 2013) and George Church (Mali et al. 2013). 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 (Han and She 2017). 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. Told as such, CRISPR's history can seemingly be

deduced through academic publications. However, the science itself does not tell the full scientific story.

In contrast to many other groundbreaking scientific discoveries which remain known only within scientific circles, 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 (Marcon et al. 2019). Netflix has even broadcasted a documentary film dedicated to CRISPR, underscoring its iconic status in popular culture. By now, CRISPR is not a purely scientific phenomena. Wikipedia, a popular source read and compiled by the general public, also strives to document these facets as well. The encyclopedia's text and their sources, can thus be viewed as an inclusive media, one that can potentially help track the interaction with additional fields and allow a better understanding of how scientific knowledge ramifies well outside the realm of academic publications.

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 (Wolfe, J. 2018). Most famously, Eric Lander's perspective in Cell, the "Heroes of CRISPR" (Lander 2016), was met with fierce criticism ("Why Eric Lander's Controversial Paper 'The Heroes of CRISPR' Is Not Solid Historical Research" 2017). 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 ("'Heroes of CRISPR' Disputed" n.d.) - going as far as to label Lander the "villain" of CRISPR ("The Villain of CRISPR" n.d.). 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 history is short (i.e., parallel to Wikipedia's lifetime) and multi-faceted: a highly scientific topic with wide-ranging technological and social ramifications. These facets, 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 digital history of contemporary science. To that end, we put forward a methodology and provide some automated tools utilizing Wikipedia's data - its articles, their edit histories and their references. Our method relies on both quantitative and qualitative analyses that may help consolidate the aforementioned conflict between data and content dependent historical research.

Methods & Results

1. Delineating the research scope

Historical research requires a clear delineation of the field being studied, for instance gathering a collection of academic publications (Pence and Ramsey 2018). Similarly, 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 the majority 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, as these will likely contain a substantial amount of information directly related to the subject. To facilitate these steps of search and filtration we designed a tool to do this automatically for any given term of interest (WikiCorpusBuilder). Continuing with the term "CRISPR", this filtering yielded 51 articles (Table S1). Out of these, 10 had CRISPR in their title, and another 41 that only had it in the title of one of their sections.

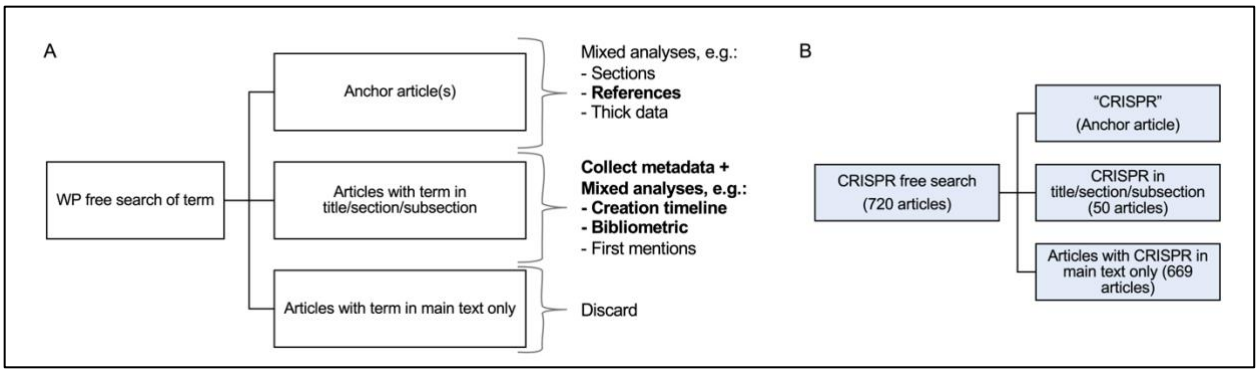

Fig 1. Workflow for using Wikipedia to research the history of a specific field. A) Scheme of proposed research flow as supported by our tool: A free search of Wikipedia's English-language articles is conducted to identify all the relevant articles; these are then filtered to include only those that 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. Of the listed examples, in bold are the data provided by our tool, the rest are performed manually. B) Breakdown of flow scheme in the CRISPR case study, as of June 2022.

Even among this list, a clear hierarchy arises between those which in subject, text and focus are fully aligned with the topic being researched (which we term the "anchor article/s"); and "auxiliary" articles, those that represent secondary aspects of the topic or instances in which it is embedded within other fields. The distinction is important as it allows us to focus the qualitative work described in the following sections. Here, the anchor article selected was "CRISPR", which was determined semantically based on reviewing its title and content.

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.

Another advantage Wikipedia provides is open access to these articles' multitude of data, which we harness using our tool to further characterize the articles of the corpus. For example, "CRISPR" ranked amongst the top 5 articles in terms of size, number of references, and number of edits (Fig S1).

Thus, we have composed a clearly defined research copus regarding our key term of interest. The corpus can be depicted through the titles (qualitative) or the data (quantitative), and further mixed analyses can take place as will be demonstrated below.

2. Mixed method analyses for understanding historical growth of knowledge

After having established the research scope, created our corpus and identified the anchor article(s) within it, we move to the analysis phase. We deployed three different complementary analyses: (1) A qualitative reading of anchor article(s) text and structure, both its current and past versions; a quantitative analysis of the (2) anchor article(s) and (3) of the corpus. All three are based on data and materials that are readily available on Wikipedia and many aspects of the quantitative analyses have been automated through our tool (Fig 1). The following describes how these were deployed in a mixed manner on CRISPR.

Mixed-methods research (Gold and Klein 2019) combines quantitative and qualitative analyses and served as the basis for this research. We employed this in what can be termed Wikipedia-focused "thick big data" (Jemielniak 2020), (as opposed to content-agnostic big data approaches), which meshes the world of thick description and big data. In our case, Wikipedia articles, their edit histories and sources are treated as an initial dataset, which are then analyzed semantically as well as through mixed methods and interpreted in a detailed manner.

First, using Wikipedia's "view history" tool, available at the top right of every article (Fig 2A), we can access the anchor article's past versions and perform a comparative reading. Here, we

used annual intervals to sample textual changes - at times narrowing the time frame to provide a more detailed account of the article's historical textual growth.

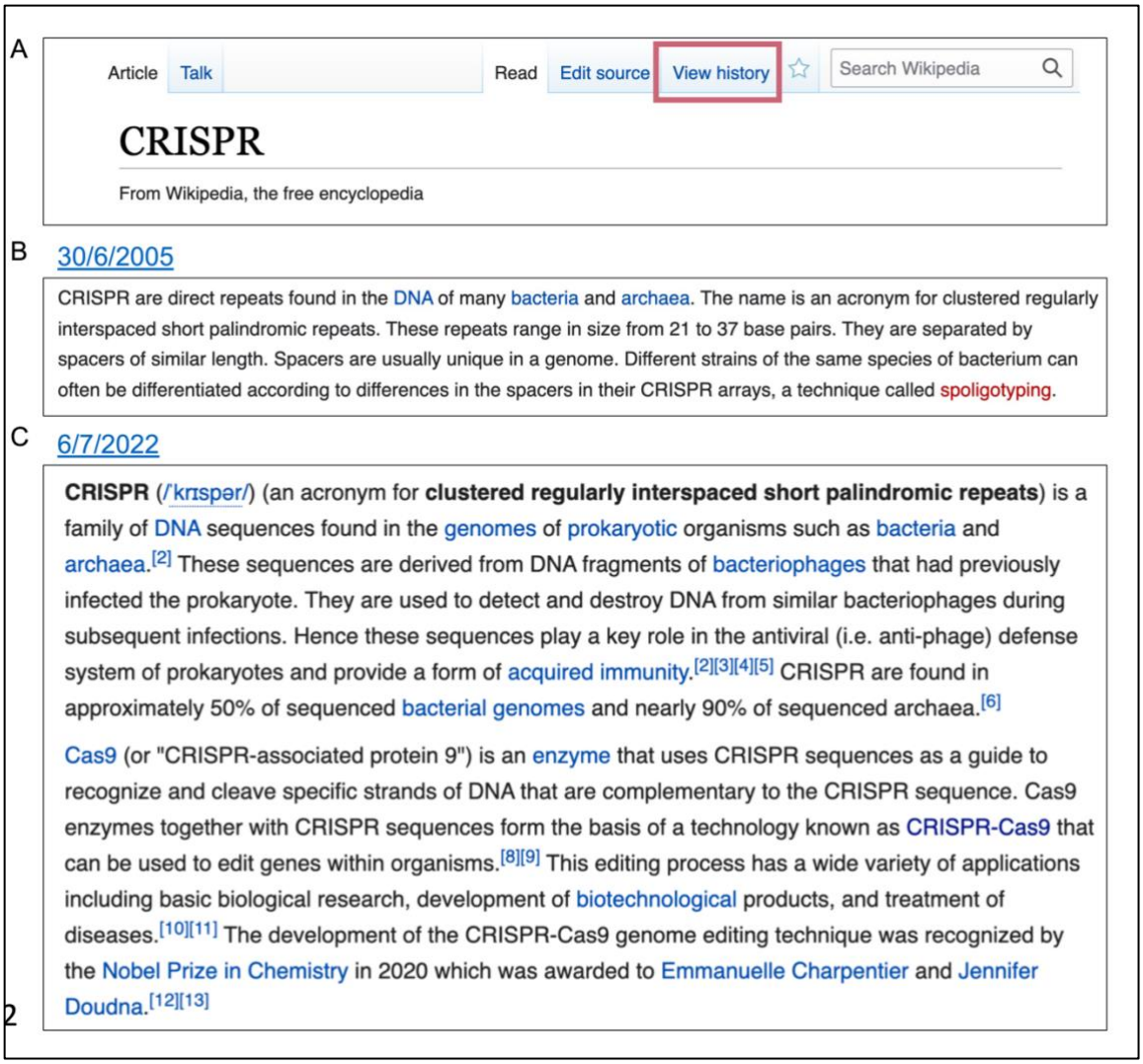

Fig 2. Comparing versions of the CRISPR article. A) An example of the top of a Wikipeida article, note the `View history` (frame added) link that enables accessing older versions of the text. Snapshots from the Wikipedia archive of the CRISPR article: B) the full text of the article when it first opened on June 30th 2005, and C) extract of the lead section's opening paragraphs, as of July 6th, 2022.

Comparing the article's past versions provided rich historical context: 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 2B). 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 be highlighted in the article's lead text in future versions (Fig 2C).

Next, we augmented this form of textual comparison with structural analyses of the CRISPR article's architecture, i.e., its Table of Contents (TOC), a basic feature of all Wikipedia's articles. This "table of contents" or "section" analysis is done as 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 (Fig 3B and Table S2). Due to the high semantic variation between sections, we prefer to gather this data and perform the comparison manually.

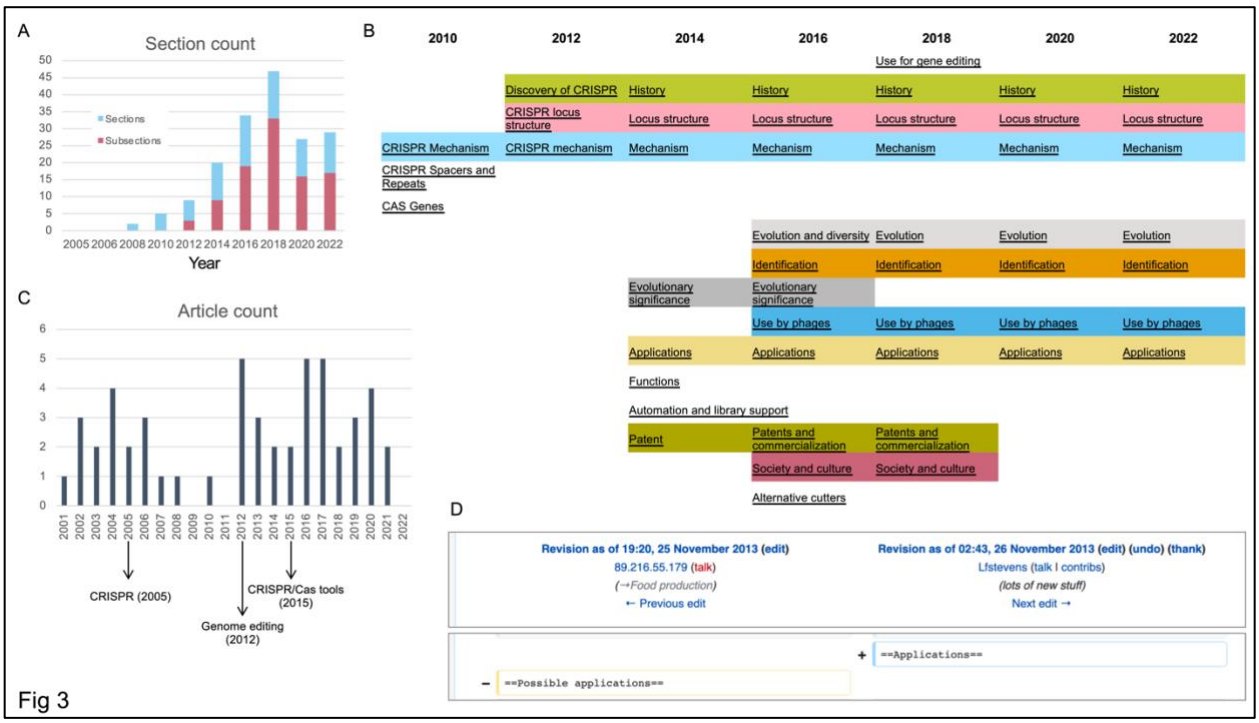

Fig 3

Fig 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) Titles of the article's sections throughout 2010-2022, sampled biannually. Subsections were removed to support clarity and can be found in Table S2. Row spacing of the content and coloring was

added manually to highlight sections repeating in consecutive revisions. C) Timeline of the number of the corpus' articles opened on each year since Wikipedia was launched (2001). D) Changelog as of November 25th, 2013, documenting the section title change from "Possible applications" to "Applications" (Other changes that occurred as part of that edit were removed for visibility, and can be found in the archive).

As we shall see the section analysis is intertwined with the corpus view of things. To understand the historical processes that took place in the CRISPR corpus, we can examine the articles based on their Date Of Birth (DOB), (Fig 3C and Table S1). Similar to the appearance of a new section - opening new articles on Wikipedia requires the topic at hand to have a certain level of "notability" (Wikipedia contributors 2022), and we therefore consider the creation of a new article as an indication that a critical abundance of knowledge and/or editor interest has been reached. Here too, we combined a quantitative evaluation of the number of articles being created with a content-dependent reading of their titles (Fig 3C and Table S1). Finally, a side-by-side view of these two timelines (sections and DOB) 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 (Fig 3B and Table S2). These sections pertain to CRISPR's genetic makeup, and can be collectively referred to as the basic science behind CRISPR.

In 2011, a "Discovery of CRISPR" section was added to the article, which was later renamed "History". The addition of an explicitly historical section in the article indicated a new phase in the scientific narrative it put forward, perhaps the result of a new self-consciousness or understanding that the emerging field was now old enough to have a history of its own. After a few months, 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". This serves as a prime example of how the article documented changes in the field as they took place, with Wikipedia's native "View history" tool's textual comparison function offering a snapshot of the "revolution" (Fig 3D).

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 detailed a 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, alongside a subsection, “Recognition”, that attempted to attribute the CRISPR discovery to specific persons. 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.

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

Our analyses thus far shows 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).

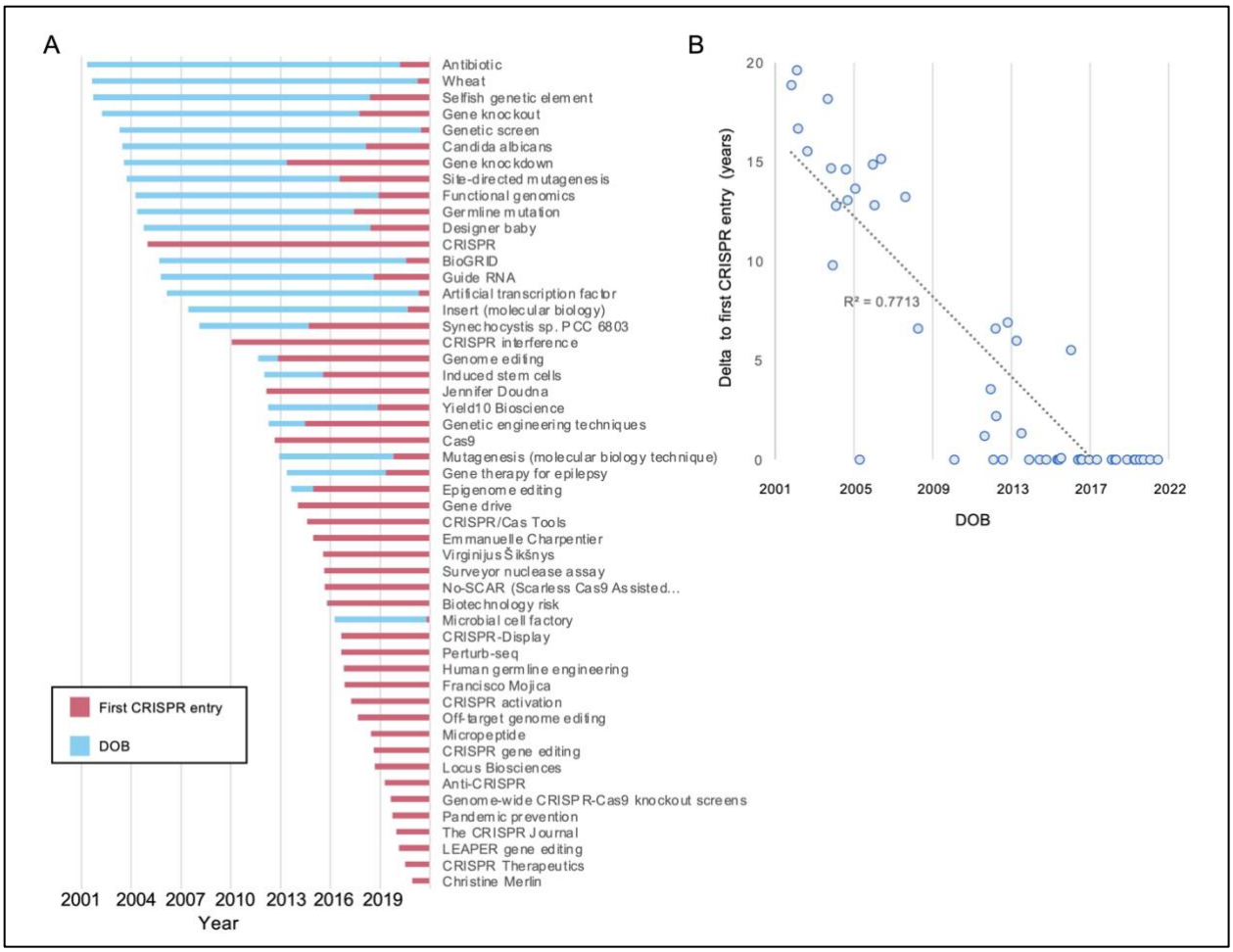

Fig 4. Comparing an article's creation date and CRISPR's first mentions. A) An article's date of birth (DOB, blue) compared to the year of its first mention of the term CRISPR (red), sorted by the former. B) The relation between the DOB and the time it took for the first mention of CRISPR of each article.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

The 24 articles that lacked "CRISPR" in their inception provide insight into the growth of the field over time. Importantly, this analysis shows how 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.

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 ("Wikipedia:Verifiability" 2022). 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 (ranges 0-1) (Benjakob, Aviram, and Sobel 2022), and the "Latency", which gauges the duration between an academic paper's publication and when it was referenced in a Wikipedia article (Benjakob and Aviram 2018).

Our automated tool detected the reference list of each article in the corpus, which is then further 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, ranking 13/51 in the corpus (Fig 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 (Arroyo-Machado et al. 2020)(Teplitskiy, Lu, and Duede 2017).

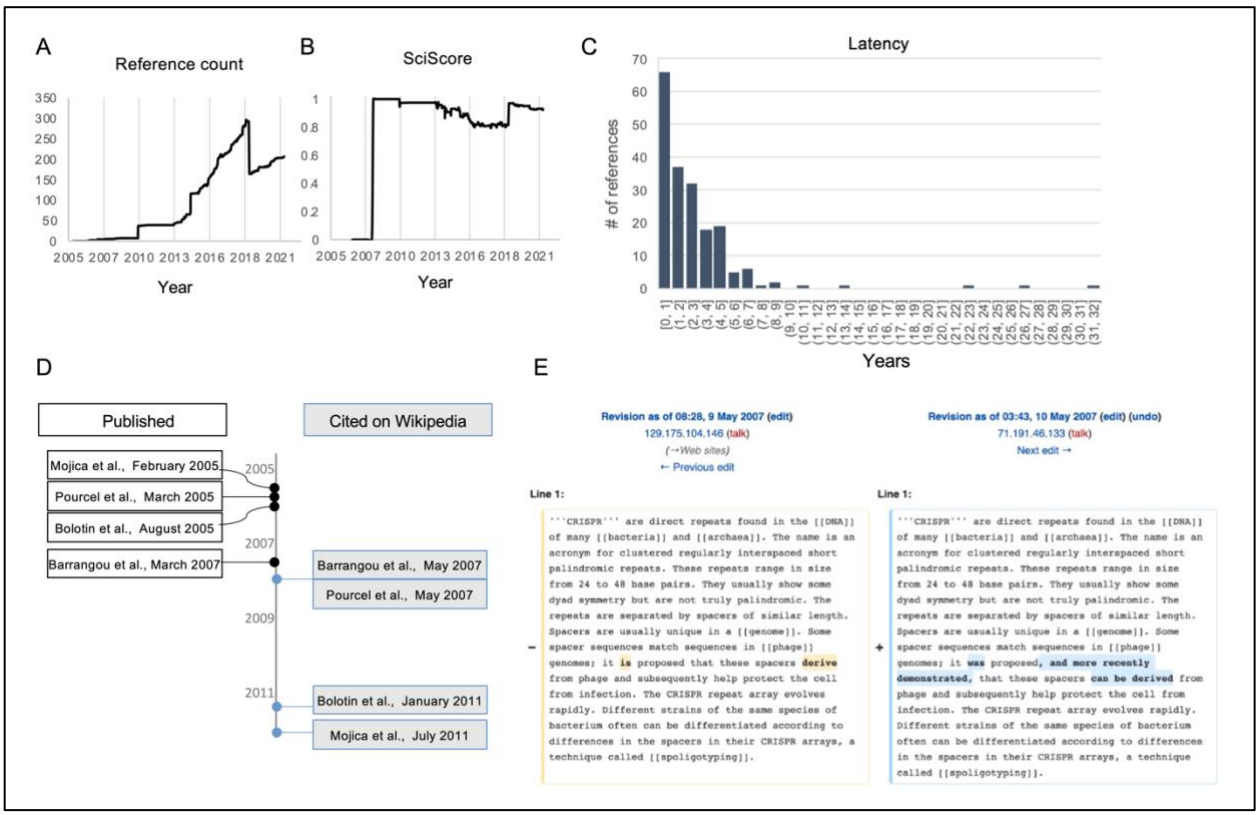

Fig 5. CRISPR-bibliometrics on Wikipedia. A) The number of references in the "CRISPR" article's reference section since it opened until December 2021. B) "CRISPR"s SciScore (until

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.

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.

We next conducted bibliometric analysis on the entire corpus. 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 among the top ten scores. To further 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 over 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.

In sum, these quantitative shifts in bibliometrics, we found, were the result of textual changes in the article. This links together our different forms of analyses: the bibliometrics are linked to the historical shifts in the text which together reflected changes in taking in the scientific field itself.

5. Quantitative comparison between fields on Wikipedia

We next aimed to examine whether the aforementioned methodology can provide insight into other scientific fields on Wikipedia. Therefore, we deployed our automated tool on two additional terms, "Circadian" and "Coronavirus", which we have studied in different manners in earlier preliminary studies (Benjakob and Aviram 2018)(Benjakob, Aviram, and Sobel 2022) 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 quantitative analyses described above.

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). While varying in size, all corpuses are within a range that allows for subjective reading and examination of the titles comprising them. This validated that they indeed 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. For example, 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".

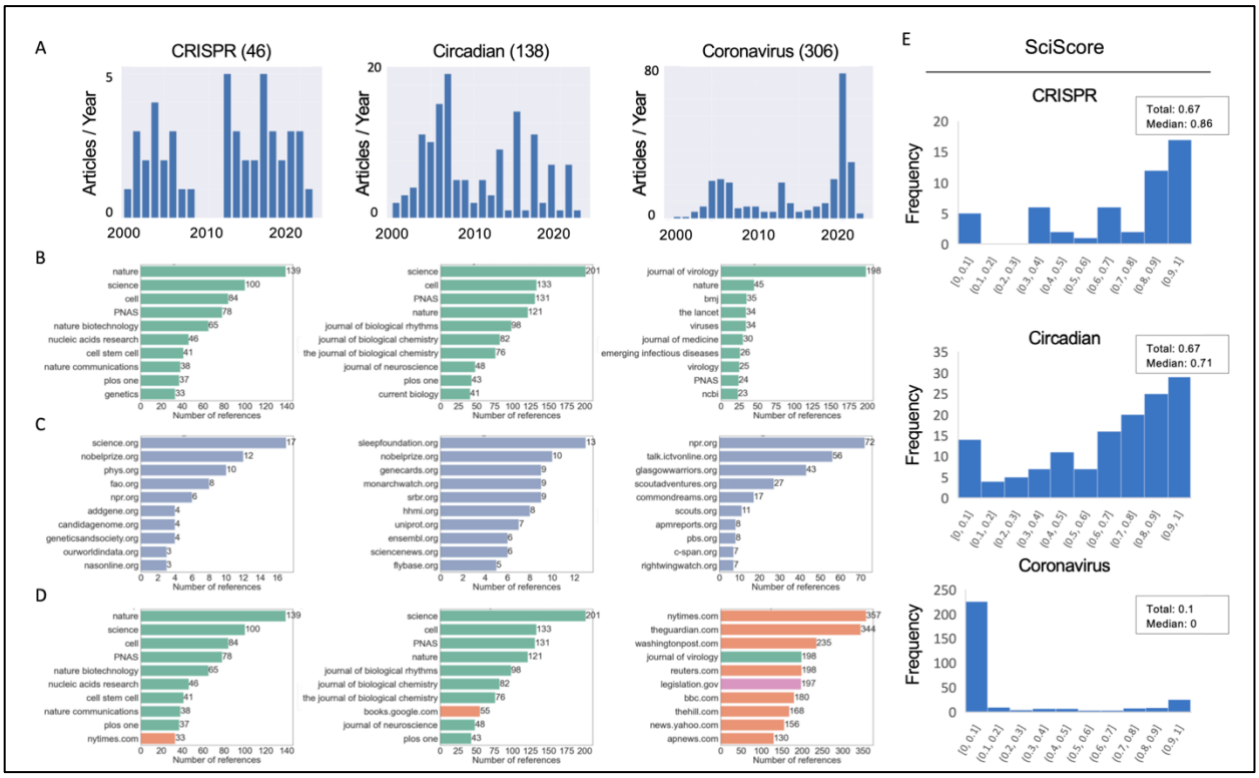

Fig 6. Comparing Wikipedia corpuses: Different fields show different data. Corpuses were generated and quantitative metrics automatically collected in June-July 2022, for the terms "CRISPR", "Circadian" and "Coronavirus". The following data are presented: A) the number of articles opened each year, B) 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 in all articles) and median scores of the articles' distribution.

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 (Benjakob, Aviram, and Sobel 2022). 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 (“The Nobel Prize in Physiology or Medicine 2017” n.d.), 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 (Chiang et al. 2012), 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 (Benjakob and Aviram 2018). Broadly, this suggests that article creation time is perhaps more applicable for contemporary and what can be termed “active” fields.

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 (Suh et al. 2009).

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, we can examine how each corpus the top 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 (i.e., websites) 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, we observed that 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 SciScore: 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.

Discussion

Here, we delved into Wikipedia's archives to examine the way a prominent scientific field, CRISPR, was represented 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" (Pennisi 2013), as well as the popular interest in the field.

This case study allowed us to flash out some essential metrics with which one can conduct such research, and we thus propose a method that can be used to perform history of contemporary science on other topics using Wikipedia. Automated tools were developed and are openly supplied 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 (Benjakob and Aviram 2018). 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 (Benjakob, Aviram, and Sobel 2022). Together, 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 (Wyatt 2020). 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 (2010) 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. The aforementioned study on the Egyptian protest movement attributed historical significance to the addition of the word “revolution” to Wikipedia articles’ titles, taken to be reflective of the real revolution playing out in the streets (H. Ford 2022). This is a captivating demonstration in support of attributing historical significance to semantic shifts, in line with our usage of sections and titles.

From the perspective of digital humanities, an algorithmic approach was previously 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 (Ju et al. 2020), makes interesting findings, while highlighting the lack of a unification in methods in current Wikipedia-based historical research. 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.

Numerous studies about Wikipedia and bibliometrics (Teplitskiy, Lu, and Duede 2017), even those that focus on science (Arroyo-Machado et al. 2020)(Maggio et al. 2017); but none that clearly link scientometrics to historical methods (Price 1963). 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 (Gredel 2017). 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.

Mapping out additional fields can eventually support theories/models of scientific growth in a resolution never before possible. 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 (“Breakthrough of the Year” 2022) 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.

The advantages of Wikipedia

Wikipedia easily lends itself to research of this type. A digital and open website that is easily searchable, it also succumbs to a simple to use API for more complex queries and even provides a full dump of the entirety of Wikipedia in each language, including the full edit history of every article.

Importantly, 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. This feature, in combination with the convenience of the

613 “View history” function, facilitates in-depth analysis of the same section over time (for example,
614 as was done here for “Patents”) in a manner difficult to imagine for comparing texts of academic
615 literature. Moreover, cross-analyses of different subjects can yield results comparable through
616 standardized metrics, like the DOB timelines, and the Latency or SciScore metrics for
617 bibliometric comparisons. The structural similarity creates a sort of internal control that lays the
618 groundwork for a rigid research system that can be utilized by others and applied to additional
619 fields.
620 Past versions that did not survive Wikipedia’s mob review process or that included facts that
621 were true at the time but have since been rendered obsolete prove especially interesting from
622 the perspective of the history of science. For example, with CRISPR, a December 2005 version
623 of the article described Cas1 as the “most important” of the cas systems, and one that is
624 “present in almost every CRISPR/Cas system.” This was more cautiously reworded in July 2010
625 so that, “The most important of the Cas proteins appears to be Cas1, which is ubiquitous” in
626 CRISPR systems. In March 2011, Cas1’s ubiquity was no longer said to be linked to its
627 importance, and for the past decade the article has made due with noting in a subsection
628 dedicated to CRISPR locus that “[m]ost CRISPR-Cas systems have a Cas1 protein.” These
629 changes were the result of new knowledge forcing a reevaluation of the preexisting scientific
630 narrative regarding CRISPR: Cas1 was not falsified per se, rather its importance in CRISPR’s
631 story was reassessed. Another example from the CRISPR article can be seen in the shift in
632 section title from “Potential Applications” to “Applications” regarding gene editing (Fig 3D).
633 These are examples of what can be termed “negative” knowledge - knowledge whose relevance
634 was negated by new “positive” discoveries that outweighed it in significance. However, as such,
635 its degradation of scientific status in CRISPR’s narrative, has much value from the historical
636 perspective. Wikipedia, we suggest, is an inclusive media that documents both positive and
637 negative knowledge - the accumulation and the rejection of scientific facts through its edit
638 history.

Moreover, unlike social media websites that collect user data for financial reasons, posing a privacy threat and creating ethical dilemmas for researchers, Wikipedia collects no such information as it has no such business model. This makes it not only attractive to volunteers willing to donate hours to writing and editing the site, but also makes Wikipedia and its data ideal material for social research. Wikipedia's texts are not single-handedly written and are edited collectively in a form of what is termed peer-production (Benkler 2006). 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.

Additional advantages that Wikipedia offers in respect to bibliometrics are numerous and deserve their own section.

Wikipedic Bibliometrics

Numerous studies analyzing Wikipedia's references, even those that focus on science (Arroyo-Machado et al. 2020) (Maggio et al. 2017) (Teplitskiy, Lu, and Duede 2017); exemplify the use of Wikipedia for bibliometric research, and to a degree support the view that Wikipedia is much more inclusive than academic publications, making use of non-academic sources usually excluded from academic texts. Here we implicitly study this using our SciScore, and contextualize its trends through the historical thick description. 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 Coronavirus 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” (Pennisi 2013), followed by the much-publicized patent and credit wars, and finally the wider social, ethical and policy debates it sparked.

Our latency analyses 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 (Benjakob and Aviram 2018). 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, Coronvirus 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 (Benjakob, Aviram, and Sobel 2022). 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. In light of emerging attempts to harness Wikipedia for trend detection (Kämpf et al. 2015) (Mittermeier et al. 2021), this idea remains to be examined as more case studies to be created in the future.

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 (Jemielniak 2019)(Mostafa 2021)(Mostafa 2023). 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 altmetrics, with 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. Being a prominent scientist with a broad scope, his Wikipedia article succinctly mentions his contribution to the CRISPR field under a section titled “Synthetic biology and genome engineering” that despite its topic does not use the key term itself.

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 (Pooladian and Borrego 2017), 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 yet untapped facet of Wikipedia 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 (Horvath and Barrangou 2010). 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 S3). 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.

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.

Code 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, Alex Webb 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.

References

Arroyo-Machado, Wenceslao, Daniel Torres-Salinas, Enrique Herrera-Viedma, and Esteban Romero-Frías. 2020. “Science through Wikipedia: A Novel Representation of Open Knowledge through Co-Citation Networks.” *PLOS ONE* 15 (2): e0228713–e0228713. https://doi.org/10.1371/journal.pone.0228713.

Benjakob, Omer, and Rona Aviram. 2018. “A Clockwork Wikipedia: From a Broad Perspective to a Case Study.” *Journal of Biological Rhythms* 33 (3): 233–44. https://doi.org/10.1177/0748730418768120.

Benjakob, Omer, Rona Aviram, and Jonathan Aryeh Sobel. 2022. “Citation Needed? Wikipedia Bibliometrics during the First Wave of the COVID-19 Pandemic.” *GigaScience* 11 (January). https://doi.org/10.1093/gigascience/giab095.

Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press. https://www.jstor.org/stable/j.ctt1njknw.

Bolotin, Alexander, Benoit Quinquis, Alexei Sorokin, and S Dusko Ehrlich. 2005. “Clustered Regularly Interspaced Short Palindrome Repeats (CRISPRs) Have Spacers of Extrachromosomal Origin.” *Microbiology (Reading, England)* 151 (Pt 8): 2551–61. https://doi.org/10.1099/mic.0.28048-0.

“Breakthrough of the Year.” 2022. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Breakthrough_of_the_Year&oldid=1101623902.

Chiang, C. D., C. L. Lewis, M. D. E. Wright, S. Agapova, B. Akers, T. D. Azad, K. Banerjee, et al. 2012. “Learning Chronobiology by Improving Wikipedia.” *Journal of Biological Rhythms* 27 (4): 333–36. https://doi.org/10.1177/0748730412449578.

Cohen, Jon. 2020. “A Cut above: Pair That Developed CRISPR Earns Historic Award.” *Science (New York, N.Y.)* 370 (6514): 271–72. https://doi.org/10.1126/science.370.6514.271.

Colavizza, Giovanni. 2020. “COVID-19 Research in Wikipedia.” *Quantitative Science Studies* 1 (4): 1349–80. https://doi.org/10.1162/qss_a_00080.

Cong, Le, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” *Science (New York, N.Y.)* 339 (6121): 819–23. https://doi.org/10.1126/science.1231143.

Darnton, Robert. 1982. “What Is the History of Books?” *Daedalus* 111 (3): 65–83.

———. 1985. *The Great Cat Massacre and Other Episodes in French Cultural History*. New York : Vintage Books. http://archive.org/details/greatcatmassacr000darn.

Ford, H. 2022. *Writing the Revolution Wikipedia and the Survival of Facts in the Digital Age*. MIT Press. https://opus.lib.uts.edu.au/handle/10453/163379.

Ford, Heather. 2020. “Rise of the Underdog,” October. https://doi.org/10.7551/mitpress/12366.003.0017.

Gasiunas, Giedrius, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. 2012. “Cas9-CrRNA Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (39): E2579-86. https://doi.org/10.1073/pnas.1208507109.

Giles, Jim. 2005. “Internet Encyclopaedias Go Head to Head.” *Nature* 438 (7070): 900–901. https://doi.org/10.1038/438900a.

Gold, Matthew K., and Lauren F. Klein, eds. 2019. *Debates in the Digital Humanities 2019*. University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk.

Gredel, Eva. 2017. “Digital Discourse Analysis and Wikipedia: Bridging the Gap between Foucauldian Discourse Analysis and Digital Conversation Analysis.” *Journal of Pragmatics* 115 (July): 99–114. https://doi.org/10.1016/j.pragma.2017.02.010.

Han, Wenyuan, and Qunxin She. 2017. “CRISPR History: Discovery, Characterization, and Prosperity.” *Progress in Molecular Biology and Translational Science* 152: 1–21. https://doi.org/10.1016/bs.pmbts.2017.10.001.

“‘Heroes of CRISPR’ Disputed.” n.d. The Scientist Magazine®. Accessed October 20, 2022. https://www.the-scientist.com/news-opinion/heroes-of-crispr-disputed-34188.

Horvath, Philippe, and Rodolphe Barrangou. 2010. “CRISPR/Cas, the Immune System of Bacteria and Archaea.” *Science (New York, N.Y.)* 327 (5962): 167–70. https://doi.org/10.1126/science.1179555.

Jansen, Ruud, Jan D A van Embden, Wim Gaastra, and Leo M Schouls. 2002. “Identification of Genes That Are Associated with DNA Repeats in Prokaryotes.” *Molecular Microbiology* 43 (6): 1565–75. https://doi.org/10.1046/j.1365-2958.2002.02839.x.

Jemielniak, Dariusz. 2019. “Wikipedia: Why Is the Common Knowledge Resource Still Neglected by Academics?” *GigaScience* 8 (12): giz139. https://doi.org/10.1093/gigascience/giz139.

———. 2020. *Thick Big Data: Doing Digital Social Sciences*. New product. New York: Oxford University Press.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science (New York, N.Y.)* 337 (6096): 816–21. https://doi.org/10.1126/science.1225829.

Ju, Harang, Dale Zhou, Ann S. Blevins, David M. Lydon-Staley, Judith Kaplan, Julio R. Tuma, and Danielle S. Bassett. 2020. “The Network Structure of Scientific Revolutions,” October.

Kämpf, Mirko, Eric Tessenow, Dror Y. Kenett, and Jan W. Kantelhardt. 2015. “The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks.” *PLOS ONE* 10 (12): e0141892. https://doi.org/10.1371/journal.pone.0141892.

Lander, Eric S. 2016. “The Heroes of CRISPR.” *Cell* 164 (1–2): 18–28. https://doi.org/10.1016/j.cell.2015.12.041.

Lean, Oliver M., Luca Rivelli, and Charles H. Pence. 2021. “Digital Literature Analysis for Empirical Philosophy of Science.” Preprint. March 21, 2021. https://www.journals.uchicago.edu/doi/10.1086/715049.

Maggio, Lauren A, Ryan M Steinberg, Tiziano Piccardi, and John M Willinsky. 2020. “Reader Engagement with Medical Content on Wikipedia.” *ELife* 9. https://doi.org/10.7554/eLife.52426.

Maggio, Lauren A., John M. Willinsky, Ryan M. Steinberg, Daniel Mietchen, Joseph L. Wass, and Ting Dong. 2017. “Wikipedia as a Gateway to Biomedical Research: The Relative Distribution and Use of Citations in the English Wikipedia.” *PLOS ONE* 12 (12): e0190046. https://doi.org/10.1371/journal.pone.0190046.

Mali, Prashant, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo, Julie E Norville, and George M Church. 2013. “RNA-Guided Human Genome Engineering via Cas9.” *Science (New York, N.Y.)* 339 (6121): 823–26. https://doi.org/10.1126/science.1232033.

Marcon, Alessandro, Zubin Master, Vardit Ravitsky, and Timothy Caulfield. 2019. “CRISPR in the North American Popular Press.” *Genetics in Medicine* 21 (10): 2184–89. https://doi.org/10.1038/s41436-019-0482-5.

Maree, David J. F. 2020. “The Methodological Division: Quantitative and Qualitative Methods.” In *Realism and Psychological Science*, by David J. F. Maree, 13–42. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-45143-1_2.

Merton, R. K., and Dudley Shapere. 1974. “*The Sociology of Science: Theoretical and Empirical Investigation*.” *Physics Today* 27 (8): 52–53. https://doi.org/10.1063/1.3128814.

Mittermeier, John C., Ricardo Correia, Rich Grenyer, Tuuli Toivonen, and Uri Roll. 2021. “Using Wikipedia to Measure Public Interest in Biodiversity and Conservation.” *Conservation Biology: The Journal of the Society for Conservation Biology* 35 (2): 412–23. https://doi.org/10.1111/cobi.13702.

Mojica, Francisco J M, César Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. 2005. “Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements.” *Journal of Molecular Evolution* 60 (2): 174–82. https://doi.org/10.1007/s00239-004-0046-3.

Mostafa, Mohamed M. 2021. “Two Decades of Wikipedia Research: A PubMed Bibliometric Network Analysis.” *Global Knowledge, Memory and Communication*, November. https://doi.org/10.1108/GKMC-03-2021-0056.

———. 2023. “Twenty Years of Wikipedia in Scholarly Publications: A Bibliometric Network Analysis of the Thematic and Citation Landscape.” *Quality & Quantity*, February. https://doi.org/10.1007/s11135-023-01626-7.

Pence, Charles H., and Grant Ramsey. 2018. “How to Do Digital Philosophy of Science.” *Philosophy of Science* 85 (5): 930–41. https://doi.org/10.1086/699697.

Pennisi, Elizabeth. 2013. “The CRISPR Craze.” *Science* 341 (6148): 833–36. https://doi.org/10.1126/science.341.6148.833.

Pooladian, Aida, and Ángel Borrego. 2017. “Methodological Issues in Measuring Citations in Wikipedia: A Case Study in Library and Information Science.” *Scientometrics* 113 (1): 455–64. https://doi.org/10.1007/s11192-017-2474-z.

Porter, Charles A., and Robert Darnton. 1980. “The Business of Enlightenment: A Publishing History of the Encyclopedie 1775-1800.” *Eighteenth-Century Studies* 13 (3): 335. https://doi.org/10.2307/2737989.

Pourcel, C, G Salvignol, and G Vergnaud. 2005. “CRISPR Elements in Yersinia Pestis Acquire New Repeats by Preferential Uptake of Bacteriophage DNA, and Provide Additional Tools for Evolutionary Studies.” *Microbiology (Reading, England)* 151 (Pt 3): 653–63. https://doi.org/10.1099/mic.0.27437-0.

Price, Derek J. De Solla. 1963. *Little Science, Big Science*. Columbia University Press. https://doi.org/10.7312/pric91844.

Reagle, Joseph, and Jackie Koerner, eds. 2020. *Wikipedia @ 20: Stories of an Incomplete Revolution*. The MIT Press. https://doi.org/10.7551/mitpress/12366.001.0001.

Rheinberger, Hans-Jörg. 2011. “Infra-Experimentality: From Traces to Data, from Data to Patterning Facts.” *History of Science* 49 (3): 337–48. https://doi.org/10.1177/007327531104900306.

Sepkoski, David. 2013. “Towards ‘A Natural History of Data’: Evolving Practices and Epistemologies of Data in Paleontology, 1800–2000.” *Journal of the History of Biology* 46 (3): 401–44. https://doi.org/10.1007/s10739-012-9336-6.

Singh, Harshdeep, Robert West, and Giovanni Colavizza. 2021. “Wikipedia Citations: A Comprehensive Data Set of Citations with Identifiers Extracted from English Wikipedia.” *Quantitative Science Studies* 2 (1): 1–19. https://doi.org/10.1162/qss_a_00105.

Smith, Denise A. 2020. “Situating Wikipedia as a Health Information Resource in Various Contexts: A Scoping Review.” *PLOS ONE* 15 (2): e0228786–e0228786. https://doi.org/10.1371/journal.pone.0228786.

Suh, Bongwon, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. “The Singularity Is Not near: Slowing Growth of Wikipedia.” In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration - WikiSym ’09*, 1. Orlando, Florida: ACM Press. https://doi.org/10.1145/1641309.1641322.

Teplitskiy, Misha, Grace Lu, and Eamon Duede. 2017. “Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science.” *Journal of the Association for Information Science and Technology* 68 (9): 2116–27. https://doi.org/10.1002/asi.23687.

"The Nobel Prize in Physiology or Medicine 2017." n.d. NobelPrize.Org. Accessed October 20, 2022. https://www.nobelprize.org/prizes/medicine/2017/press-release/.

"The Villain of CRISPR." n.d. Accessed October 20, 2022. https://www.michaeleisen.org/blog/?p=1825.

"Why Eric Lander's Controversial Paper 'The Heroes of CRISPR' Is Not Solid Historical Research." 2017. American Scientist. January 3, 2017. 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.

Wikipedia contributors. 2022. "Notability in the English Wikipedia — Wikipedia, The Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Notability_in_the_English_Wikipedia&oldid=1110309951.

"Wikipedia:Verifiability." 2022. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=1108181695.

Wilson, Adam M., and Gene E. Likens. 2015. "Content Volatility of Scientific Topics in Wikipedia: A Cautionary Tale." *PLOS ONE* 10 (8): e0134454. https://doi.org/10.1371/journal.pone.0134454.

Wolfe, J. 2018. "U.S. Appeals Court Upholds MIT, Harvard Patents on CRISPR Gene Editing." Reuters. https://www.reuters.com/article/us-ucberkeley-crispr-idUSKCN1LQ1XP.

Wyatt, Liam. 2020. "Endless Palimpsest: Wikipedia and the Future's Historian." *Studies in Higher Education* 45 (5): 963–71. https://doi.org/10.1080/03075079.2020.1749793.

Yasseri, Taha, Robert Sumi, András Rung, András Kornai, and János Kertész. 2012. "Dynamics of Conflicts in Wikipedia." *PLoS ONE* 7 (6): e38869. https://doi.org/10.1371/journal.pone.0038869.