Warning signals for long-term innovative potential of genetic research under commercial pressure

Junghun Chae[1,17], WooJoong Kim[1,17], Woochul Jung[2,17], Dawoon Jeong[3–6,17], Roktaek Lim[1,7,17], Manoj Chamlagain[7], Giju Jung[2], Juneil Jang[2,8], Jae Won Lee[9,10], Nam Kyu Kang[10], Kwangryul Baek[10], Jonghyeok Shin[10], Ye-Gi Lee[9,10], Hyun Gi Koh[10], Chanwoo Kim[9,10], Sangdo Yook[9,10], Allen Ka Loon Cheung[7], Yong-Su Jin[9,10], Hyejin Youn[5,6,11–13,*], Pan-Jun Kim[7,14–16,*], and Cheol-Min Ghim[1,2,15,*]

[1]Department of Physics, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[2]Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[3]Technology Management, Economics and Policy Program, Seoul National University, Seoul 08826, Republic of Korea
[4]Institute of Engineering Research, Seoul National University, Seoul 08826, Republic of Korea
[5]Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA
[6]Northwestern Institute on Complex Systems, Evanston, IL 60208, USA
[7]Department of Biology, Hong Kong Baptist University, Kowloon, Hong Kong
[8]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[9]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[10]Department of Food Science and Human Nutrition, University of Illinois at Urbana−Champaign, Urbana, IL 61801, USA
[11]Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL 60208, USA
[12]London Mathematical Lab, London WC2N 6DF, UK
[13]Santa Fe Institute, Santa Fe, NM 87501, USA
[14]Center for Quantitative Systems Biology & Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Kowloon, Hong Kong
[15]Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk 37673, Republic of Korea
[16]Abdus Salam International Centre for Theoretical Physics, 34151 Trieste, Italy
[17]These authors contributed equally to this work.

[*]Correspondence and requests for materials should be addressed to H.Y. (e-mail: hyejin.youn@kellogg.northwestern.edu), P.-J.K. (e-mail: extutor@gmail.com), or C.-M.G. (e-mail: cmghim@unist.ac.kr).

Abstract

Current technological revolutions, involving artificial intelligence, mRNA vaccines, and quantum computing, are largely driven by industry. Despite the existing perception that commercial motives promote cutting-edge innovation, concerns may arise about their risk of limiting scientific exploration from diverse perspectives, which nurtures long-term innovative potential. Here, we investigate the interplay between scientific exploration and industrial influence by analyzing about 20 million papers and US, Chinese, and European patents in genetic research, a domain of far-reaching societal importance. We observe that research on new genes has declined since the early 2000s, but the exploration of novel gene combinations still underpins biotechnology innovation. Fields of highly practical or commercial focus are less likely to adopt the innovative approaches, exhibiting lower research vitality. Additionally, continuous scientific research creates exploratory opportunities for innovation, while industry's R&D efforts are typically short-lived. Alarmingly, up to 42.2–74.4% of these exploratory opportunities could be lost if scientific research is restrained by industry interests, highlighting the cost of over-reliance on commercially-driven research. Given the industry's dominance in recent technologies, our work calls for a balanced approach with long-term scientific exploration to preserve innovation vitality, unlock the full potential of genetic research and biotechnology, and address complex global challenges.

Introduction

Groundbreaking innovation emerges as the resonance of science and technology [1–3]. From the landmark developments at the Palo Alto Research Center (PARC) and Bell Labs to today's artificial intelligence (AI), robotics, mRNA vaccines, and quantum computing [4–8], industrial entrepreneurship has powered technology innovation. Given the particular dominance of industry in the ongoing technological shifts, it is imperative to discuss the positive and negative impacts of this dominance on the course of research and development (R&D) [6–8]. Ethical and regulatory aspects aside, different perspectives may exist for the lasting effects on the innovation itself [9–16], in relation to the following stories.

Recalling the history of biotechnology, the Human Genome Project decades ago was primarily led by government agencies with widespread support from academia, but Celera Genomics from the private sector played a pivotal role in accelerating the process [16]. On the other hand, commercially-supported research is vulnerable to cuts or compromise upon changes in management conditions. The storied Bell Labs, an industrial laboratory once

known as an innovation powerhouse across fundamental and applied sciences, could not maintain its standing amid company downturns [5].

More broadly, innovative attempts can face innovation barriers such as cost-effectiveness, particularly under commercial focus in favor of early profitability [6, 11–13]. This situation reflects the conflict between entrepreneurial risk-taking and economic rationality, and the latter can exert pressure on innovators to deliver quick and profitable marketable outcomes. This pressure, which we refer to as commercial pressure here, tends to discourage long-term exploration efforts and promote risk-averse approaches against uncertain returns and tight regulations [9–13]. It may in turn limit the spectrum of approaches from diverse perspectives and ultimately harm the long-term vitality of innovation itself, as our analysis will later indicate. In parallel to this issue, concerns have already been raised about the decline of innovative research output across academic disciplines and industrial sectors [17, 18]. Notably, it is the risk-averse nature of pharmaceutical firms that fosters underinvestment in groundbreaking innovation and thereby suppresses novelty in drug development [10, 19]. Besides, a recent COVID-19 R&D study suggests that non-monetary incentives can actually be more common drivers of innovation [9].

As a crucial step to facilitate further discussions, we here examine a gap between the activities of scientific exploration and industry-driven technology development from the aspect of commercial pressure and innovation vitality, unlike the previous data-driven studies that often focused on the flow of scientific or technological knowledge into future innovation [20–23]. We target the empirically observed patterns in the realm of biotechnology—specifically, genetic research. The recent year (2023) marked both the 70th anniversary of the discovery of the DNA double helix and the 20th anniversary of the completion of the Human Genome Project. Biotechnology has extraordinary impacts on modern society, with an undeniable role in overcoming the COVID-19 pandemic. Especially, research on genes and gene products forms the foundation of modern biotechnology and is recognized for its applicability in medicine, agriculture, food industry, energy supply, environmental remediation, and grand challenges such as climate change [16, 19, 24–26]. For the sake of simplicity in this study, both genes and gene products (proteins) will henceforth be referred to as genes, unless specified.

Results

Popular genes in scientific research and technology inventions

We traced research of 53,484 different genes for protein production through the systematic survey of 16 million papers and 2 million patent families, the latter from the U.S. Patent and

Trademark Office (USPTO), the China National Intellectual Property Administration (CNIPA), and the European Patent Office (EPO). The genes were curated from the UniProt Knowledgebase (UniProtKB)/Swiss-Prot (Methods) [27]. Our analysis focuses on the periods of 1990–2021 for paper publications (i.e., until the COVID-19 pandemic era) and 1990–2020 for patent applications. These periods are chosen because the Human Genome Project was launched in 1990 as a milestone in the history of genetic research, whereas the patent application records after about 2020 did not seem fully accessible at the time of this manuscript preparation (Methods).

Like previous studies, we view papers as carriers of scientific research and patents as carriers of technology inventions [20–22, 28–31]. Accordingly, we obtained the annual counts of papers and patents for each gene as the activity levels of scientific research and technology inventions with that gene, respectively (Methods). Because not all genes in the full text of a given document might directly contribute to its subject, we only considered genes in the title and abstract. According to our analysis, a patent is typically company-assigned with a 76.0% chance for each gene, whereas a paper rarely includes company-funded research, with only a 5.2% chance for each gene (Methods). In other words, the patent-filed inventions can be viewed primarily as industry-driven, but the paper-published research is not.

One example of our analysis is that the research on hemoglobin grew until 2021 with five times as many paper publications as in 1990. Similarly, hemoglobin-involving inventive activities grew with ten times as many patent applications as in 1990 (Fig. 1*A*). The examples of CRISPR associated protein 9 (Cas9), angiotensin-converting enzyme 2 (ACE2), and cyclooxygenase-2 (COX-2) are presented in Fig. 1*B*–*D*. Intriguingly, ACE2 is a binding target of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to enter host cells and cause COVID-19 [32, 33], and as expected, its research and inventive activities skyrocketed from 2020 (Fig. 1*C*). On the other hand, COX-2 showed the decline of the inventive activities in the 2000s (Fig. 1*D*), and this case will be elaborated on later. As an alternative way, among the multiple genes in a given paper or patent, we identified only the genes of the main focus in that paper or patent (Methods). In this study, when analyzing only those main genes, we will explicitly state it.

We found that interleukin-6 (IL-6) and tumor necrosis factor-alpha (TNF-α) are the top-ranked genes in the cumulative volumes of scientific research (Fig. 1*E*). The top ten genes for research included many immune genes, several cancer-related genes such as tumor suppressor p53 (TP53) and epidermal growth factor receptor (EGFR), and others (Fig. 1*E*). In the case of genes in inventive activities, serum albumin and EGFR were the top-ranked, but their positions were overtaken by Cas9 and programmed cell death protein 1 (PD-1)

during 2016–2020 (Fig. 1*F*). PD-1 and programmed cell death ligand 1 (PD-L1) together came to the front in this period and they are both immunotherapy-related genes. Similarly, the top genes among only the genes of the main focus in papers or patents are presented in Appendix 1–Fig. 1 (see also Methods).

Beyond those prominent genes, our analysis expands to all other genes for a comprehensive view of genetic research, as will soon be unfolded here.

Research on new genes slows down

One may suppose that the vitality in genetic research can be partially related to the pace of studying new genes. We observe that the total number of new genes in papers and patents each year, which were not homologous to any genes in prior studies, grew until 2002 and then turned to robust decline (Fig. 2*A* and Methods). A similar tendency has also been observed in a previous study with only human protein-coding genes [26]. Consequently, the exploration of new genes in research and inventive activities continued to slow down soon after the turn of the millennium, which saw the announcement of the first draft human genome and the near-end of the Human Genome Project. Compared to 2002, the new genes fell into one-third in 2021 (Fig. 2*A*). Despite this overall tendency, an exception is the continuous increase of new genes in Chinese patents each year, unlike US and European patents (Appendix 1–Fig. 2*A*–*D*). This exception resulted from the unusual growth of the total annual patents themselves from China. Still, whichever US, Chinese, or European patents, the *proportion* of patents with new genes (i.e., the probability of new gene occurrence per patent) began to decline just before 2000 and finally recorded 3.3–5.4% (Appendix 1–Fig. 3). Such a declining trend and the other key results throughout this study remain unchanged when we only consider the genes of the main focus in papers or patents (Methods). The overall decline of newly researched genes calls to mind the pre-existing concerns about conventionally-biased human gene studies and drug developments [10, 19].

For further analysis of this trend, we categorized genes according to their frequent research themes. Mainly based on the expert-examined classifications of journals and patents as the venues of genetic research, each gene was assigned to an appropriate thematic category (Methods, and Table S1–S3 in Supplementary File 1–3). As will be clear, commercially-oriented categorical genes exhibit some unique patterns amid the overall new gene decline. While newly researched genes in medical science mirror the aforementioned post-millennium decline trend, new genes in the field of metabolic engineering underwent an earlier and severer decline, even dating back to 1992 (Fig. 2*B*,*C*). Of note, metabolic engineering is based on commercial and economic motives to enhance the desirable performance of living organisms by altering their biochemical pathways [34, 35]. Likewise, its

184 parent field, applied biotechnology shows an early decline of newly researched genes (see
185 Appendix 1–Fig. 4 for all the gene categories). We will revisit these issues later. On the other
186 hand, those in plant science just started to decline after 2008 (Fig. 2*D*). Interestingly, 2008
187 was close to the end of the NSF-funded 2010 Project that aimed to determine the function of
188 as many genes in model plant *Arabidopsis thaliana* as possible [36]. From 2017, the further
189 decline of new genes in plant science stopped, mainly by the contributions of the genes
190 related to *Arabidopsis* (48.5%) and rice (16.6%) (Fig. 2*D* and Methods).
191 To account for the overall decline of new genes from 2002 in Fig. 2*A*, we developed a
192 mathematical model that describes the cumulative number of distinct genes in papers and
193 patents over time (Methods). In this model, an effectively finite pool of genes and their
194 copies across species becomes a target of current-day genetic research. For these gene
195 copies, we considered all the possible orders of their initial studies with approximately equal
196 probabilities, and then calculated the expected cumulative number of the distinct studied
197 genes each year in combination with annually-updated gene copy numbers from the
198 UniProtKB/Swiss-Prot data. This model accurately reproduces Fig. 2*A* and gives other
199 results compatible with empirical data over a broad range of conditions (Fig. 2*E*, Appendix
200 1–Fig. 5, and Methods; $P < 10^{-4}$). Our model suggests that the vast majority (71.3 ± 5.4%) of
201 taxonomically-dispersed genes studied as of 2021 had already been studied at least once
202 before 2002 (avg. ± s.d. and Methods). In the case of taxonomically-restricted genes [37, 38]
203 as of 2021, only their 8.0 ± 0.8% had been studied before 2002, though (avg. ± s.d.). Hence,
204 the time around the completion of the Human Genome Project was a turning point with the
205 substantial accumulation of the researched, taxonomically-dispersed genes, which then
206 rarified the exploration of the new ones.
207
208 **Novel gene combinations keep actively explored, but not in every field**
209
210 Despite the significance of new genes, a more common mode of novel genetic research
211 could be the use of genes in new combinations. In stark contrast to the slowdown of
212 exploring new genes from 2002, new combinations of genes tended to increase until recently
213 (Fig. 3*A* and Methods). There exist some variations between different patent sources: from
214 the 2000s, the numbers of novel gene combinations in Chinese and US patents were
215 accelerating and running steadily, respectively. Still, none of US, Chinese, and European
216 patents entered a decline phase in novel gene combinations and they all remained active for
217 those combinations (Appendix 1–Fig. 2*E*–*P*).
218 Regarding different thematic categories of genes, we observe an increase in novel gene
219 combinations in medical science and a similar but recently weaker trend in plant science.
220 About commercially-oriented categories, we observe a contrast in metabolic engineering and

applied biotechnology with drop-offs and stagnancy in the novel gene combinations from 2017, possibly warning for their low innovation vitality (Fig. 3*B*–*D* and Appendix 1–Fig. 6 for all the gene categories; see also Methods).

Back to the overall trend beyond regional and categorical patterns, this "innovation by combination" [22, 39–41] involves the succession of highly frequent genes in those combinations over time: the most frequent in the newly paired genes were β-galactosidase in 1990–1997, green fluorescent protein (GFP) in 1998–2015, and Cas9 in 2016–2021, along with less yet highly frequent genes such as mammalian target of rapamycin (mTOR) and PD-L1 (Fig. 3*E*–*H* and Appendix 1–Table 1). For example, Cas9 newly teamed with 4,141 different genes between 2016 and 2021. We view the genes of such frequent fresh partnership as have been the workhorses of biotechnology innovation, recapitulating the history of landmark ideas and technologies from DNA cloning (β-galactosidase) and live cell imaging (GFP) to genome editing (Cas9) (Fig. 3*E*–*G*) [42, 43].

Among the above prominent genes of biotechnology innovation, Cas9 and GFP are recognized for their utility across disparate domains such as medical and plant sciences. Therefore, we examined how many genes in each category teamed with Cas9 or GFP across papers and patents, especially during the early stages of Cas9 and GFP. We found that, in their infancy to peaks, Cas9 and GFP respectively teamed with 6.4–13.1% and 6.7–23.4% of medical science genes that debuted in papers or patents during similar periods (Fig. 3*I* and Methods). Interestingly, if we limit our focus to pharmaceutical genes, they fall short of this gross medical science result, with the particularly lower proportion for Cas9 (Fig. 3*I*; $P \leq 2 \times 10^{-4}$ and Methods). Of note, pharmaceutical research within medical science is known to prioritize marketable outcomes. Another observation from our data is that both metabolic engineering and (its parent) applied biotechnology rank among the lowest in the proportions of the genes teaming early with Cas9 or GFP (Fig. 3*I*, and Appendix 1–Fig. 7 for all the gene categories; $P \leq 10^{-4}$ and Methods). These patterns do not largely depend on specific periods of the early teaming (Appendix 1–Fig. 7). The patterns of pharmaceutical, metabolic engineering, and applied biotechnology fields indicate that the highly practical focus of these fields may be in fact limiting the test and adoption of newcomer technologies. There exist further field-specific issues, such as the emphasis on product price competitiveness in the cases of metabolic engineering and applied biotechnology, requiring the mindset of rational managers rather than risk-taking innovators. Such commercial pressure restraining the innovative potentials appears to be consistent with the previous, long-alarming trend in Fig. 2*C*, 3*C* and Appendix 1–Fig. 4*D*,*E*, 6*D*,*E*, which exhibited severe decline of both newly researched genes and novel gene combinations in metabolic

engineering and applied biotechnology. In a later part of this study, we will further elaborate on the downside of commercially-controlled research activities.

Short-lived industry interests and long-lived scientific research

Recall the very initial finding in this study that patent-filed technology inventions are typically industry-driven, while paper-published scientific research is not. Because industry R&D is closely aligned with market needs and demands, commercial pressure would influence patent trajectories more than paper trajectories. Our subsequent results do remain valid when considering only the company-assigned patents among all the patents.

Fig. 4*A* shows the paper and patent time-series of COX-2, which has been studied for anti-inflammatory treatment and pain relief [44–48]. Notably, its annual patent applications started to decline in 2004 while the paper publications still persisted (Fig. 4*A*). This patent falloff, in fact, coincided with the withdrawal of a COX-2 inhibitor drug for arthritis patients (Vioxx) from the market, due to the risk of heart attack and stroke [44]. Phosphodiesterase 5 (PDE5) from our data, well known for its inhibitor sildenafil to treat erectile dysfunction (Viagra), is another gene with a patent falloff but persistent paper publications. The use of sildenafil was approved for pulmonary arterial hypertension in 2005, following its previous approval for erectile dysfunction [49, 50]. Upon sildenafil's dominance in the market, patent applications with PDE5 started to fall in 2007. Therefore, the cases of COX-2 and PDE5 presumably exemplify the industry's R&D depression due to risk management and anticipated lower returns in a saturated market, respectively [10, 12, 13].

From now on, we define the paper or patent falloff as a decrease in its number by at least a quarter of the peak value, without a later full rebound (Methods). Alternations of this definition did not change our main results. Remarkably, the preceding patent falloff above is not limited to COX-2 or PDE5 but characterizes the general course of genetic research. For two-thirds of genes that debuted in papers or patents during similar periods, their patent falloffs preceded the paper falloffs (Fig. 4*B*; $P < 10^{-4}$ and Methods). For example, genes that debuted in 1990–2000 show their patent falloffs 6.5 ± 4.9 years (avg. ± s.d.) ahead of the paper falloffs, and a quarter of these genes even show ≥10-years earlier patent falloffs. Our further analysis excludes the possibility that the policy requirement of non-overlap of a patent content with prior art might be responsible for the rapider patent falloff (Appendix 1–Fig. 8 and Methods). We hence conclude that commercial pressure tends to disrupt the industry R&D on a given gene first, despite the ongoing scientific interest in that gene. One natural consequence of this relatively early patent falloff is that the total patent volume of each gene becomes capped by the total paper volume with some scaling (Appendix 1–Fig. 9 and Methods).

293 During the period of continuous scientific research without much of the industry's drive,
294 the pursuit of breakthroughs and impactful works did not stop. Fig. 4*C* shows that even
295 ~60% of genes with clinical trials underwent these trials during the times of their limited
296 patent levels (Methods). Besides, a trial with each gene was usually not sponsored by the
297 industry, with a ~70% probability whether it was the limited patent case or not (Fig. 4*C*; $P \leq$
298 0.05 and Methods). This fact reminds us of the previous report that 70% of COVID-19
299 clinical trials were driven by public research institutions [9]. To evaluate the technological
300 impact of scientific research, we measured it by the patent citations of each paper within a
301 defined period [20–22, 28, 30, 41]. We found that a substantial body of the genes (37.0–
302 50.7%) with the highly impactful papers did appear in those papers during the times of the
303 limited patent levels (Fig. 4*D* and Methods).
304 These general tendencies are repeated in the cases of COX-2 and PDE5. After the
305 patent falloff, COX-2 continued to be studied for its function in various research papers,
306 along with many clinical trials [45–48, 51]. They include the studies of COX-2, EGFR, and
307 mTOR relations in cancer, later followed by a clinical trial [45, 46], the relationship between
308 ω−3 fatty acids and COX-2, with many patent citations [47], and the relief of
309 hyperinflammatory responses in COVID-19 [48]. In the case of PDE5 after the patent falloff,
310 its inhibition continued to be tested in paper-published research and clinical trials from
311 various aspects, such as β-catenin-involving anti-cancer mechanism and muscle ischemia
312 alleviation, both followed by patent citations [52–54], cardiovascular risk reduction [55], and
313 ventilation improvement for single-ventricle patients [56]. In other words, the industry's
314 interest in these genes turned cold first, but the innovative scientific research persevered.
315
316 **Commercial pressure can even disrupt long-term innovative opportunities**
317
318 Our results indicate that genes with the fading industry's interest still retain the opportunities
319 for innovative approach and possible breakthrough explorations. To further examine this
320 issue, we preselected eight workhorse genes in biotechnology innovation, based on their
321 shares in novel gene combinations explained above: Cas9 for genome editing [43], PD-L1
322 for immunomodulation and immunotherapy [57], mTOR and TP53 mainly for cancer
323 research [58, 59], GFP and luciferase for live cell imaging [60, 61], β-catenin mainly for
324 regenerative medicine and cancer research [62, 63], and β-galactosidase for DNA cloning
325 [42] (Methods). We will henceforth refer to these genes as *select genes*. Interestingly, the
326 aforementioned COX-2 teamed with two select genes—PD-L1 and Cas9—for the first time
327 after its patent falloff, although the patent applications reduced to one-third and one-fourth of
328 the peak, respectively (Fig. 5*A*). In these cases, PD-L1 and Cas9 had not yet reached their
329 own peaks, but they respectively served for the investigations of the suppression of

transplant rejection and the mechanism of gastric cancer [64, 65]. Likewise, the aforementioned PDE5 newly teamed with three early-stage select genes—GFP, TP53, and mTOR—when its patent applications reduced to one-third to one-fourth of the peak. GFP, TP53, and mTOR here served for the monitoring of cyclic guanosine monophosphate (cGMP) levels inside living cells and the investigations of kidney injury protection and liver cancer treatment, respectively [66–68].

Indeed, we found that a substantial portion of genes teamed with new select genes after their patent falloffs, although these select genes were still in the early stages of work. Of genes that teamed with select genes in their infancy to peaks, the majority (54.5%) initially teamed with at least one of these select genes after the patent falloffs (Fig. 5*B* and Methods). About a half of these genes, including COX-2 and PDE5, initially teamed with two or even more of those select genes after the patent falloffs (Fig. 5*B* and Methods). In addition, for one-third to a half of the genes that teamed with Cas9, PD-L1, or mTOR in its infancy to peak, these teams were initially formed after the patent falloffs, and this fraction had already been appreciable during an earlier period of the teaming (Fig. 5*B*, Appendix 1–Fig. 10, and Methods).

These results suggest that the loss of the industry's active interest does not necessarily close the exploratory opportunities of genetic research. Further analysis reveals that sustained scientific research is a key to bring this possibility to life, despite the fading industry's interest. Specifically, the teaming of each gene with a new, early-stage select gene after the patent falloff predominantly occurred (≥72.9%) when the active paper publications still continued (Fig. 5*C*, Appendix 1–Fig. 11, and Methods; $P < 10^{-4}$).

These findings raise a fundamental question: if commercial pressure propagates from industry to scientific activities even outside the industry with financial or other impacts [12, 69], to what extent would this pressure disrupt the exploratory opportunities of genetic research? One instance of such propagation is the blow to biofuel research at leading universities in the mid-2010s due to the oil price drop and subsequent investment cuts [70–72]. To answer our question, we conducted a "thought experiment" that assumed the hindering of scientific research on a given gene due to a lack of the industry's active interest: specifically, we altered the temporal trend of the paper counts to match that of the patent counts, which had come to a low level (Methods). The lack of the industry's visible interest in a particular research area may not only affect its own investment but also possibly diminish the public and governmental support for that research [73]. Our thought experiment gives rise to a remarkable estimation that the vast majority (≥69.7%) of the genes, which initially teamed with a given early select gene amid the limited industry's interest and patent applications, would now lose this teaming (Fig. 5*D*, Appendix 1–Fig. 12, and Methods).

Among all the genes whether affected or not by our experimental scenario, still 42.2–74.4% of the genes that once teamed with a given early select gene would lose this teaming, with the exception of PD-L1 although it would also lose one-fourth of the teams (Fig. 5*D*, Appendix 1–Fig. 12, and Methods). In the example of COX-2 and PDE5 whose annual patents fell in 2005 and 2007, the hindering of their scientific research would drastically reduce the chances of the teaming with each early select gene by ≥83.5% at most (Methods). Hence, these estimations suggest the cost of over-reliance on industry-dependent research, which would miss the long-term exploratory and innovative opportunities once the industry's drive becomes limited.

Discussion

Through the macroscopic overview of decades-long genetic research courses, our work challenges the existing perception that commercial motives promote cutting-edge innovation. In fact, excessive commercial pressure can distort research trajectories and stifle the full exploration of the innovation landscape in genetic research. Consistently, the fields of highly practical or commercial focus, such as metabolic engineering, are less likely to embrace innovative approaches and exhibit lower research vitality (Fig. 3*C*,*I* and Appendix 1–Fig. 6, 7). Our analysis also shows that steady scientific activities create the long-term opportunities for exploration and innovation, while industry's R&D efforts are typically short-lived (Fig. 4*B*–*D*, 5*C* and Appendix 1–Fig. 11). Of note, prioritizing early returns and risk aversion has been recognized as common features of industry [6, 9–13]. We estimate that up to 42.2–74.4% of the exploratory opportunities with notable gene combinations would be lost if the scientific research becomes hindered due to the industry's limited interest (Fig. 5*D* and Appendix 1–Fig. 12). In other words, the propagation of the commercial pressure from the industry to the external scientific research can endanger long-term innovation efforts, suggesting the cost of over-reliance on industry-driven research. Regarding such propagation, we envisage that the lack of the industry's active interest may not only suppress its own investment but also diminish the societal support for that research [73]. Although our present data analysis takes various controls and null models to minimize the possibility of alternative data interpretations (Methods), a future work beyond the current scope should identify more specific causal relationships and mechanisms. This identification may benefit from the use of highly refined data [74–76] and is a prerequisite for setting relevant policies. Consideration of hidden biases in patent data due to commercial interests is also warranted [23].

Our work points to the downside of commercial pressure (or its analogous effects at large) on research activities, which can narrow the spectrum of approaches by restricting an intellectual scope and may ultimately harm the vitality of innovation in an entire field itself

(e.g., Fig. 2*C*, 3*C*). Regarding the issue of the industry that is supposed to spur cutting-edge innovation but can act contrary through the commercial pressure, a proper innovation policy might aim to preserve the entrepreneurship of innovators. One way could be governmental partnership to reduce innovation risks. For example, the government may offer the support for the commercialization of new technologies by providing firms with opportunities to test, apply, and publicize these technologies [77, 78]. Furthermore, the government may help establish the communication channels with consumers to identify the social needs for new technologies and thereby lower the uncertainty about market response [79, 80]. These policies would encourage the industry to remain risk-taking innovators rather than risk-averse rational managers, and thus contribute positively to the innovation ecosystem.

To summarize, despite the perceived positive role of commercial pressure in technology progress, it risks undermining long-term innovative potential. Our findings call for a balanced approach with stable support for exploratory research [1, 5, 73]. This balance will be essential to unlock the full potential of biotechnology and to address complex challenges such as climate change, pandemics, major technological shifts, and social inequality [6, 81–84], for which the need for innovative and creative solutions has never been greater.

Methods

Listing of gene names

Each entry of the UniProtKB/Swiss-Prot database presents a curated, non-redundant protein sequence mainly at the species level [27, 85]. For example, IL-6 protein has an entry for each single species such as human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), or pig (*Sus scrofa*). Across all entries, we collected gene and protein names and their synonyms from the attributes in Appendix 1–Table 2 (UniProtKB/Swiss-Prot XML file, Release 2022_04). These names and synonyms were later used for their search in paper and patent texts. For simplicity, we here refer to a set of homologous genes and their proteins as a gene, and refer to their names and synonyms as synonyms unless specified. The comprehensive synonym set of each gene was identified by the grouping of the entries with highly shared synonyms, as these entries often corresponded to homologous genes. Specifically, we excluded improper synonyms such as the names of cleaved chains or subdomains of main peptide bodies, and constructed the graph where each entry is a node and two nodes are linked if their entries share a substantial portion of the synonyms to satisfy $|A \cap B|/\min(|A|, |B|) \geq 0.75$ (A and B are the synonym sets of these entries). By applying the generalized Louvain method to this graph (v0.3; https://sourceforge.net/projects/louvain) [86], we detected densely-knitted subgraphs, i.e., network communities. For synonyms still shared by multiple network communities, each synonym was assigned only to the community with the largest share of the entries for that synonym. The whole synonyms in each community were then considered the gene's tentative synonym set. After the test run of synonym matches with paper and patent texts (described below), we identified some homonyms irrelevant to genes. To reduce these homonyms, we removed non-biological phrases in common use based on their frequent occurrences in published books in society, available from *Google Books Ngram Corpus* (version 3, February 2020) [87]: the phrases with higher counts than an empirically-chosen threshold (1.8×10^7 counts in the books published by 2020) were automatically removed from our gene synonyms. The remaining synonyms underwent extensive manual inspection and curation, and we then finalized our gene synonym sets.

Collection of paper and patent data

To survey research activities in genetic subjects, we targeted the papers in the journals of biology, applied biology, medicine, and other related disciplines. The categories of those disciplines and the associated journal lists were obtained from the Web of Science (WoS) Journal Citation Reports (JCR) (https://jcr.clarivate.com). We manually curated these journal

lists for better association with the journal categories and removed the journals irrelevant to contemporary genetic research. The resulting journal lists are provided by Table S1 in Supplementary File 1 with our expert-curated journal categories JG0, JG1, ⋯, and JG8: JG1 includes biomedical journals, along with its subsets JG2 and JG3 for veterinary medicine and pharmaceutics, respectively. JG4 includes journals in applied biotechnology, along with its subset JG5 for metabolic engineering. JG6 and JG7 include plant science and general microbiology journals, respectively (we here use the term “general microbiology” to distinguish it from applied microbiology covered by JG4). JG0 includes all the journals in JG1 to JG7, with additional journals for animal, algal, and aquatic research. JG8 includes broad-scope journals and other journals not fit to JG0.

At the same time, we downloaded the annual releases of the baseline data files from the PubMed FTP server (April 22, 2022; https://ftp.ncbi.nlm.nih.gov/pubmed/baseline) [88], which contained the titles and abstracts of the articles serviced by PubMed. Other than the titles and abstracts, we did not pursue the full text analysis because all the genes in the full texts were not necessarily relevant to the very subjects of the articles. We considered only the articles in the above JG0 and JG8 journals and published in 1990–2021 because the Human Genome Project was launched in 1990 and our gene synonyms derived from the UniProtKB/Swiss-Prot data dated May 2022 as noted above.

To survey inventive activities in genetic subjects, we collected the USPTO, CNIPA, and EPO patent application data (US, Chinese, and European, respectively) in the PATSTAT Global database (version 5.22, 2023 Autumn Edition) provided by the EPO. For each patent family or any set of patents with duplicated titles and abstracts, we selected only the earliest-dated patent to avoid data redundancy. Among the Cooperative Patent Classification (CPC) codes jointly managed by the EPO and the USPTO (dated May 1, 2022) and the International Patent Classification (IPC) codes managed by the World Intellectual Property Organization (WIPO) (dated October 1, 2023), we selected possibly gene-related codes based on expert opinions (Table S2 in Supplementary File 2) and considered the patents with these codes. There is a known delay between patent application and publication, and we did observe a roughly two-year delay with a spurious decrease in recent patent application records. Therefore, taking this delay and our data access in Fall 2023 into consideration, the patents applied by the end of 2020 were included in our further analysis. Throughout this study, we considered only the patent titles and abstracts because the genes there mostly overlapped with those in the claims and the rest included genes not necessarily relevant to the very subjects of the patents.

491 **Search of genes in papers and patents**
492
493 To find the synonyms of each gene in the paper or patent titles and abstracts, we used the
494 string-matching algorithm that directly compares the phrases in the target literature to the
495 strings from the gene synonyms, while incorporating the indexed tree and tokenizing scheme
496 for efficient computation [89]. Several rules were added to the search algorithms to consider
497 common variations in the synonyms, such as the use of Greek letters, plural forms, and
498 different letter cases. For each gene, papers and patents were counted when matching at
499 least one of its synonyms. However, papers and patents with too many genes (>10 genes
500 each) were excluded from our analysis, regarding their ambiguous research targets among
501 those genes. The precision and recall of the gene identification were 0.90 and 0.71,
502 respectively, according to the comparison with manually identified genes from 100 randomly
503 selected papers (Table S4 in Supplementary File 4). Although the existing resource,
504 PubTator3 also provides a list of genes in each article [90], the precision and recall (0.96 and
505 0.79) were only marginally better than our result above (Table S4 in Supplementary File 4).
506 Critically, PubTator3 serves only PubMed papers, and therefore we excluded its use for
507 consistency of our paper and patent analyses.
508 Throughout this study, we performed additional analyses that involve only the genes of
509 the main focus in papers or patents. We identified these genes with the following heuristic
510 rules: in the case of a paper, we considered a gene as of its main focus if this gene appears
511 in the title or appears ≥3 times in the abstract. In the case of a patent, we considered a gene
512 as of its main focus if this gene appears in the title or appears ≥2 times in the abstract, or if
513 this gene is the only gene in the title and abstract. We validated these rules by the manual
514 interpretation of 100 randomly chosen papers and patents. This manual inspection showed
515 the precision of 0.86 and the recall of 0.90 for papers and the precision of 0.88 and the recall
516 of 0.85 for patents (Table S5 in Supplementary File 5).
517
518 **Paper funding source and patent assignee analyses**
519
520 We manually identified the funding sources of 300 different papers in genetic research
521 (Table S6 in Supplementary File 6), selected evenly as will be explained below. If the
522 funding sources of a given paper included at least one company, the paper was considered
523 company-funded. In the case of other funding sources from this paper set, we confirmed
524 their public nature by reviewing the specific funding sources. For the selection of these 300
525 papers, we first randomly selected 50 different genes each (total 200), whether they debuted
526 in papers/patents in 1990–2000 or 2001–2011, and whether their patent time-series include
527 falloff years or are fully spanned by states II and III, which will be defined later. These four

cases were considered due to their high relevance to many analyses in this study. In the cases of the patent falloffs, one paper before the patent falloff and the other paper after (or at) this falloff were randomly selected for each gene; otherwise, any single paper was randomly selected for each gene. The probability of a paper being company-funded for each gene was obtained by the weighted average of the proportion of company-funded papers over all those six cases, reflecting two from each patent-falloff case (one before the falloff and the other afterward). In this calculation, we assigned a weight to each case using the actual proportions of genes and their papers from the entire dataset.

For patent assignees of unambiguous characters, we considered “COMPANY”, “GOV”, “NON-PROFIT”, “UNIVERSITY”, and “HOSPITAL” among the patent applicant sectors from the PATSTAT database. For the genes with patents assigned to these sectors, we calculated the average proportion of patents, each with at least one company assignee. This quantity represents the probability of a patent being company-assigned for each gene.

Thematic categorization of genes

The common categorization of genes based on molecular or biochemical contexts rarely allowed the best interpretation of our paper or patent counts with the genes. Therefore, we categorized the genes according to their frequent research themes. Mainly based on the journal categories and CPC/IPC codes of the papers and patents with each gene, the gene was assigned to one of the following thematic categories (Table S3 in Supplementary File 3). The methods described here were heuristically developed by the examination of the categorization results.

“Medical science”: for each gene, we calculated the ratio of its papers in the above JG1 journals to those in the JG0 journals (Table S1 in Supplementary File 1). Genes with these ratios greater than or equal to an empirically-chosen threshold (0.7) were assigned to category “medical science”. In addition, for each gene, we calculated the fraction of its patents with CPC or IPC codes “A61K38”, “A61K31”, “A61K39”, “A61K48”, and their child codes (Table S2 in Supplementary File 2). Genes with these fractions greater than or equal to an empirically-chosen threshold (0.5) were also assigned to the category “medical science”.

“Genetic tools”: the genes utilized as the tools for a range of molecular biology and genetic experiments were assigned to category “genetic tools” based on manual literature search and expert opinions, in the absence of their specific journals and patent classification codes. This category itself was not used for our analysis, but for excluding its genes from other categories below.

564 *"Applied biotechnology"*: this category was intended to include the genes mainly in non-
565 medical applied biotechnology. For each gene, we calculated the fraction of its patents with
566 CPC or IPC code "A61K" and its child codes (Table S2 in Supplementary File 2). For each of
567 the genes with these fractions lower than an empirically-chosen threshold (0.5), we counted
568 its papers in the JG4 journals (Table S1 in Supplementary File 1). Each gene with this count
569 ≥ 40% of the total number of its papers in the JG0 journals was assigned to category
570 "applied biotechnology". To avoid the overlap with the other categories, the genes in the
571 previous categories "medical science" and "genetic tools" were here excluded from the
572 outset.
573 *"Plant science"* and *"general microbiology"*: as noted above, we use the term "general
574 microbiology" to distinguish it from applied microbiology covered by "applied biotechnology".
575 For each gene, we calculated the ratio of its papers in the JG6 (JG7) journals to those in the
576 JG0 journals, and the genes with these ratios ≥ 0.5 were assigned to category "plant
577 science" ("general microbiology"). To avoid the overlap with the other categories, the genes
578 in "medical science", "genetic tools", and "applied biotechnology" were here excluded from
579 the outset. The overlap between "plant science" and "general microbiology" was avoided by
580 manual recategorization of their shared genes.
581 A part of genes in "medical science" were subcategorized as follows: for each gene in
582 "medical science", we calculated the ratio of its papers in the JG3 journals to those in the
583 JG1 journals. The genes with these ratios ≥ 0.25 were assigned to subcategory
584 *"pharmaceutical"*. Another subcategory *"immune"*, given its definite biological function and
585 existing database, was assigned the genes of Gene Ontology (GO) term "immune response"
586 (GO:0006955) and its child terms, but "hemocyte proliferation" (GO:0035172) or its child
587 terms [91]. Additionally, the genes associated with Anatomical Therapeutic Chemical (ATC)
588 codes "immunostimulants" (L03) and "immunosuppressants" (L04) in the DrugBank
589 database [92] were manually reviewed and thereby assigned to the subcategory "immune".
590 Note that the above subcategory "pharmaceutical" allows the overlap with this subcategory
591 "immune".
592 Each gene in "applied biotechnology" was assigned to subcategory *"metabolic*
593 *engineering"*, when the ratio of its papers in the JG5 journals to those in the JG4 journals
594 was ≥ 0.4.
595 The genes in the above categories and subcategories, particularly those near their
596 thresholds, underwent further manual reviews to adjust and finalize the categorization.
597 Naturally, there remain genes without any category assignment.

Annual count of newly researched genes

We counted new genes in papers and patents each year in the following ways: (*i*) based on the debut year of each gene in papers or patents, whichever earlier for that gene (Fig. 2*A*); (*ii*) based on the debut year of each gene in papers (Appendix 1–Fig. 2*A*); and (*iii*) based on the debut year of each gene in US, Chinese, or European patents, whichever the earliest for that gene (Appendix 1–Fig. 2*B*–*D*). Because of the low availability of the patent data after 2020 as noted above, we only considered the paper data for 2021 in the case of (*i*). Nevertheless, because 2021 is just one year and the paper volume far exceeds the patent volume, the result is unlikely to change much even when the 2021 patent data are considered.

To count the genes related to *Arabidopsis thaliana* or rice among new genes in the category of plant science, we considered the new genes whose paper or patent titles/abstracts matched one of words “Arabidopsis”, “A. thaliana”, and “Arabidopsis thaliana” for *A. thaliana*, and “rice”, “Oryza”, “O. sativa”, and “Oryza sativa” for rice, in a case-insensitive way.

Mathematical modeling of new gene research over time

In our mathematical model, we consider a total of *L* genes and their *M* non-redundant amino acid (AA) sequences across species (i.e., *M* gene copies across species). These sequences can correspond to the aforementioned UniProtKB/Swiss-Prot entries and are assumed to have the potential for prospective research. We label these *M* sequences by $1, 2, \cdots, M$ in the temporal order of their debuts in research in papers and patents, and consider all the possible configurations of this labeling with approximately equal *a priori* probabilities. If *X* and *Y* are the number of the member sequences of a given gene and the earliest label of these member sequences, respectively, then the probability of $Y \leq y$ for the gene with $X = x$ is given by

$$\Pr(Y \leq y | X = x) = \begin{cases} 1 - \binom{M-y}{x}\binom{M}{x}^{-1} & \text{if } 1 \leq y \leq M - x \\ 1 & \text{if } y > M - x \end{cases}, \quad (1)$$

where the binomial coefficient is given by $\binom{a}{b} = \frac{a!}{b!(a-b)!}$. If *f*(*x*) is the number of genes with *x* member sequences each, it satisfies the following relations:

$$\sum_{x=1}^{M} f(x) = L \ ; \quad \sum_{x=1}^{M} x f(x) = M \ . \tag{2}$$

Let $b(y)$ be the expected number of genes that debuted in research by the time of the yth-labeled sequence's debut in research. $b(y)$ is then expressed as

$$b(y) = \sum_{x=1}^{M} f(x) \Pr(Y \leq y | X = x) = \sum_{x=1}^{M} f(x) - \sum_{x=1}^{M-y} f(x) \binom{M-y}{x} \binom{M}{x}^{-1} . \tag{3}$$

$b(y)$ in Eq. (3) can be further expressed as the function of time t, i.e., $b(y(t))$ if $y(t)$ is the total number of the researched sequences by that time t. Note that the reporting of a given AA sequence to the scientific community does not necessarily coincide with the research of this sequence. After the reporting of the sequence, the event of its first research may be modeled as a Poisson process with the following $y(t)$:

$$y(t) = \sum_{t''=t_0}^{t} \sum_{t'=t_0}^{t''} \beta N(t') e^{-(t''-t')/\tau} \left(1 - e^{-1/\tau}\right) . \tag{4}$$

Here, $N(t')$ is the number of the newly reported, non-redundant sequences in year t' obtainable from the UniProtKB/Swiss-Prot entries, β is the fraction of the sequences for prospective research ($0 \leq \beta \leq 1$), t'' denotes the debut year of each sequence in research, τ is the characteristic latent period between the sequence's first report and research, t_0 is the earliest of year t' for the available $N(t')$ data (i.e., 1939 in the case of the UniProtKB/Swiss-Prot data), and $\left(1 - e^{-1/\tau}\right)$ serves as the normalization factor.

In practice, the reporting years of the sequences in the UniProtKB/Swiss-Prot entries often indicate their first sequencing years, which may not be the same as their actual earliest reporting years, particularly in the case of the early-reported genes before the wide popularization of sequencing technologies. In this regard, we examined the first reporting years in the UniProtKB/Swiss-Prot entries and found some cases before 1997 where each gene's reporting was spuriously preceded by its debut in research. Therefore, we considered only $t \geq 1997$ (year) in our modeling.

If there exists x' to satisfy $\sum_{x=x'}^{M} f(x) \ll \sum_{x=1}^{M} f(x)$ and $x' \ll M - y$, $b(y)$ in Eq. (3) is simplified as follows:

$$b(y) \approx \sum_{x=1}^{M} f(x) - \sum_{x=1}^{M} f(x)\left(1-\frac{y}{M}\right)^{x}. \tag{5}$$

To avoid the overfitting of our model to empirical data, we lump $f(x)$'s and reduce their number, as follows:

$$f(x)\begin{cases} \geq 0 & \text{if } x = 1 \text{ or } x = k\ (k > 1) \\ = 0 & \text{otherwise} \end{cases}. \tag{6}$$

As noted previously, the non-redundant sequence in each UniProtKB/Swiss-Prot entry is mainly valid at the species level, and therefore genes with $x = 1$ in Eq. (6) correspond to taxonomically-restricted genes (TRGs) or orphans that lack recognizable homologs in other species [37, 38]. Here, we call the other genes taxonomically-dispersed genes (TDGs), and they are lumped to $x = k$ in Eq. (6) where k denotes the effective number of the species for a TDG. Eq. (5) is now converted to

$$\begin{aligned} b(y) &\approx \sum_{x=1}^{M} f(x) - f(1)\left(1-\frac{y}{M}\right)^{1} - f(k)\left(1-\frac{y}{M}\right)^{k} \\ &\approx \sum_{x=1}^{M} f(x) - f(1)\left(1-\frac{y}{M}\right)^{1} - f(k)e^{-ky/M} \\ &= \frac{f(1)}{M}y + f(k)\left(1-e^{-ky/M}\right) = sy + i\left[1-e^{-(1-s)y/i}\right], \end{aligned} \tag{7}$$

where $s \equiv f(1)/M$, $i \equiv f(k)$, and we used this relation:

$$\frac{k}{M} = \frac{1}{f(k)}\frac{kf(k)}{M} = \frac{1-s}{i}. \tag{8}$$

Because the values of $b(y(t)) - b(y(t-1\text{ year}))$ are available from Fig. 2*A*, their summation over time leads to $b(y(t))$ with some constant shift. If we reinterpret $b(y)$ as this summation of the empirical values, then it is rewritten as

$$b(y) \approx sy + i\left[1-e^{-(1-s)y/i}\right] + c\ , \tag{9}$$

where c is a constant.

Next, we seek the parameters in Eq. (9). For each pair of τ and β in Eq. (4) with $(\tau, \beta) \in \{0, 1, 2, \cdots, 19\} \otimes \{0.1, 0.2, 0.3, \cdots, 1.0\}$ (τ takes the unit of year), we randomly selected 10^6 sets of the initial parameters of (s, i, c) in the ranges of $(0,1) \otimes [[, b(t=2021)] \otimes [[, b(t=1997)]$ ($b(t)$ denotes the above-mentioned empirical data of $b(y(t))$, t takes the unit of year, and we consider only $t \geq 1997$ as previously discussed). This random initial parameter

selection was conducted by the PCG algorithm in NumPy v1.19.1. Plugging the empirical values of $N(t')$ in Eq. (4) for $y(t)$, we minimized the following error functions by the sequential least squares programming (SLSQP in SciPy v1.6.2):

$$\chi_1 \equiv \frac{\sum_{t=1997}^{2021} |\Delta b(t) - \Delta b_{\mathrm{est}}(t)|}{\sum_{t=1997}^{2021} \Delta b(t)}; \tag{10}$$

$$\chi_2 \equiv \frac{\sum_{t=1997}^{2021} |b(t) - b_{\mathrm{est}}(t)|}{\sum_{t=1997}^{2021} b(t)}, \tag{11}$$

where $\Delta b(t) \equiv b(t) - b(t - 1\ \mathrm{year})$, $\Delta b_{\mathrm{est}}(t) \equiv b_{\mathrm{est}}(t) - b_{\mathrm{est}}(t - 1\ \mathrm{year})$, and $b_{\mathrm{est}}(t)$ is the right-hand side of Eq. (9) with $y(t)$ in Eq. (4). The summation in the error functions runs from the year 1997 for the aforementioned reason. Given τ and β, the minimization of χ_1 gives rise to the best-fit *s* and *i*, and the subsequent minimization of χ_2 with these *s* and *i* gives rise to the best-fit *c*. Fig. 2*E* presents the heatmap of the resulting χ_1 and the corresponding graphs of $b_{\mathrm{est}}(t)$ and $\Delta b_{\mathrm{est}}(t)$ as the functions of $y(t)$ and *t*, respectively. The comparison of $b_{\mathrm{est}}(t)$ and $b(t)$ shows their excellent agreement ($\chi_1, \chi_2 \leq 0.07$) over a broad range of τ and β: $P < 10^{-4}$ for either the resulting χ_1 or χ_2 (one-tailed test) from the null distribution of χ_1 or χ_2 with the direct use of 10^4 randomly selected sets of (*s*, *i*, *c*) for each (τ, β) in the same parameter ranges as above (PCG algorithm in NumPy v1.19.1). We stress that the observed match between $b_{\mathrm{est}}(t)$ and $b(t)$ is not a mere result of parameter overfitting but stands on the genuine agreement of the empirical data with our model output: $b(t)$ itself initially curves and approaches a straight line as a function of $y(t)$ (Fig. 2*E*) as exactly predicted by Eq. (9) in the model, and thereby always results in $\chi_2 < 0.01$ whenever $\chi_1 \leq 0.07$ is satisfied.

Our model output involves the quantities about TRGs and TDGs mentioned above. First, the fraction of TRG sequences among all the non-redundant sequences of research merit in the model is $f(1)/M = s$. Across all the τ and β pairs resulting in $\chi_1, \chi_2 \leq 0.07$, we obtained the probability distribution of *s* in Appendix 1–Fig. 5*A*. The expected fraction of the TRGs among all the researched genes as of year *t* is $f(1)\cdot[y(t)/M]/b_{\mathrm{est}}(t) = sy(t)/b_{\mathrm{est}}(t)$. In the same χ_1 and χ_2 conditions as above, we obtained the probability distribution of the latest available $sy(t)/b_{\mathrm{est}}(t)$ (i.e., *t* = 2021 (year)) as in Appendix 1– Fig. 5*B*.

Because the expected number of TDGs researched as of year *t* is $\sim f(k)\{1 - [1 - y(t)/M]^k\} \approx i\left[1 - e^{-(1-s)y(t)/i}\right]$, the expected fraction of the TDGs researched as of year t_1 among the TDGs as of year t_2 ($t_1 < t_2$) is

$F_{\mathrm{TDG}}(t_1, t_2) \equiv \left[1 - e^{-(1-s)y(t_1)/i}\right] / \left[1 - e^{-(1-s)y(t_2)/i}\right]$. Because the expected number of TRGs researched as of year t is $f(1)\cdot[y(t)/M] = sy(t)$, the expected fraction of the TRGs researched as of year t_1 among the TRGs as of year t_2 ($t_1 < t_2$) is $F_{\mathrm{TRG}}(t_1, t_2) \equiv y(t_1)/y(t_2)$. In the same χ_1 and χ_2 conditions as above, we computed $F_{\mathrm{TDG}}(t_1, t_2)$ and $F_{\mathrm{TRG}}(t_1, t_2)$ when $t_1 < 2002$ (year) and $t_2 = 2021$ (year), as presented in the main text for the discussion of the result in Fig. 2*A*.

Identification of gene combinations

To identify the combination of genes that are meaningfully linked in a given paper or patent, we considered the pair, triplet, or entire combination (n-tuple with $n \geq 2$) of the genes that include at least one gene of the main focus in that paper or patent (see above for the identification of the genes of the main focus in a paper or patent). Because of the low availability of the patent data after 2020 as noted above, we considered only the paper data in the case of gene combinations in 2021. Nevertheless, because 2021 is just one year and the paper volume far exceeds the patent volume, the result is unlikely to change much even when the 2021 patent data are considered.

If any gene in a given pair belongs to a certain thematic gene category, we treated this pair as affiliating with this thematic category (Fig. 3*B*–*D* and Appendix 1–Fig. 6).

For each year or period, we defined the share of a given gene in new pairs as the number of new pairs with this gene divided by the total number of new pairs in that year or period (Fig. 3*E*–*H* and Appendix 1–Table 1). Among the genes with the highest annual shares, Cas9 and GFP are recognized for their utility across disparate domains such as medical and plant sciences. Therefore, for each thematic gene category, we computed the proportion of genes paired with Cas9 or GFP. In this calculation, we focused on the pairings during the early stages of Cas9 and GFP, as these early stages indicate the novelty and innovativeness of the pairings. Specifically, for Cas9 or GFP, we considered the period up to the peak time of its annual paper and patent number. Additionally, we even considered an earlier stage, the period up to the midpoint of the cumulative paper and patent number at the peak time. As will be seen, the same definitions of the early pairing can apply to other genes of interest beyond Cas9 and GFP. If the annual paper and patent number of such a gene continued to grow until 2021, then 2021 is treated as the peak time in these definitions. Back to the proportion of genes paired with Cas9 or GFP, we considered genes that debuted in papers or patents during similar periods, in order to control for variations from different periods of debut years: one group of genes with debut years of 1990–2000 (Fig. 3*I*, left and Appendix 1–Fig. 7*A*,*C*) and the other with 2001–2011 (Fig. 3*I*, right and Appendix 1–Fig. 7*B*,*D*). More recent debut years (2012–) were not considered because the number of these

genes was relatively small and they may not yet have reached the levels of research stages for the full analysis.

For the statistical significance test of the lower proportion of pharmaceutical genes paired with Cas9 or GFP than that of the whole medical science genes (Fig. 3*I*), we generated 10^4 random subsets of the medical science genes and each subset included the same number of genes as pharmaceutical genes. Using these subsets, we obtained the *P* value of the observed proportion of pharmaceutical genes paired with Cas9 or GFP in its above-mentioned early stage (one-tailed test). In addition, we performed the statistical significance test of the relatively low proportion of metabolic engineering (applied biotechnology) genes paired with Cas9 or GFP, compared to those in other categories (Fig. 3*I* and Appendix 1–Fig. 7). Specifically, we randomly regrouped genes in major thematic categories—medical science, plant science, metabolic engineering (applied biotechnology), and general microbiology—while keeping the original sizes of these categories. In this way, we generated 10^4 null configurations of the category membership and obtained the *P* value of the observed proportion of metabolic engineering (applied biotechnology) genes paired with Cas9 or GFP in its above-mentioned early stage and the rank of this proportion among the categories (one-tailed test). All these statistical tests were performed for the genes that debuted in papers or patents during similar periods, as described above.

Paper and patent falloff patterns

If annual paper numbers of a given gene are decent enough (≥ 5 papers/year on average for certain three consecutive years), we smoothened the paper time-series with a moving window average (3-year window) to reduce its fluctuations. If the papers in the last window do not exceed a certain threshold (75% of the peak), we considered the latest year of meeting or crossing this threshold after the peak time as the paper falloff year. We defined the patent falloff year in a similar way. Alternations of our methods, however, did not much change our main results.

If annual papers of a given gene are scarce (without any three consecutive years of ≥ 5 papers/year on average), we averaged the paper numbers in year t–1, t, and t +1, and termed this average X_t here. To each year t, we assigned one of the following states based on the majority of X_{t-1}, X_t, and X_{t+1}: state I if the majority are ≥1 each; state II if the majority are 0 each; and otherwise, state III. In the case that t –1 is the debut year of the gene in papers or patents, or t +1 is the last time point, only two of X_{t-1}, X_t, and X_{t+1} are defined and we assigned one of the following states to year t: state I if any of two is ≥1 and both are >0; state II if any of two is 0 and both are <1; and otherwise, state III. In the case that t is the debut year of the gene in papers or patents, or t is the last time point, only either X_{t-1} or X_{t+1}

is defined and we assigned one of the following states to year *t*: state I if the value is ≥1; state II if the value is 0; and otherwise, state III. If the paper time-series includes state I sometime before the last point and state II or III at the last point, we considered the year next to the latest time of state I as the paper falloff year. We defined the patent falloff year in a similar way. Alternations of our methods, however, did not much change our main results.

To investigate paper and patent falloff patterns, we obtained the proportion of genes whose paper falloffs preceded the patent falloffs and vice versa (Fig. 4*B*). In this calculation, we used the same period (–2020) for both paper and patent time-series for their fair comparison and considered only the genes that satisfy the following conditions: (*i*) any of the paper and patent time-series includes the falloff year defined above; and (*ii*) neither the paper nor patent time-series is fully spanned by states II and III defined above. In addition, we considered the genes that debuted in papers or patents during similar periods (1990–2000 and 2001–2011, but not 2012–) for the same reasons explained earlier.

To examine whether the policy requirement of non-overlap of a patent content with prior art is responsible for the observed tendency of the paper and patent falloffs (Fig. 4*B*), we filtered out papers and patents with highly similar contents to their prior works and re-analyzed the data (Appendix 1–Fig. 8). This test ensures a fair paper and patent treatment, with the analyzed papers being as dissimilar to their prior works as the patents are. Specifically, we used PaECTER, a citation-informed sentence transformer [93] and obtained the cosine similarity of every pair of papers (patents) for each gene, based on the paper (patent) titles and abstracts. If the total number of papers (patents) exceeded 5,000 for a given gene, we only used 5,000 randomly selected papers (patents) for the similarity calculation, which would otherwise become computationally impractical. For the papers (patents) of each gene in a particular year, we filtered out the later papers (patents) with the similarity at or above a certain threshold, and repeated this process across all the considered years in chronological order. The similarity threshold here (0.97) was determined by the manual inspection of the contents of paper or patent pairs with varying threshold values. After the filtering, we obtained the proportion of genes whose paper falloffs preceded the patent falloffs and vice versa. Regarding the control group without the filtering, we used the aforementioned 5,000 papers (patents) for each gene if this gene occurred in more than 5,000 papers (patents), and otherwise, all its papers (patents).

For the statistical significance test of the dominance of genes whose patent falloffs preceded the paper falloffs (Fig. 4*B* and Appendix 1–Fig. 8), we randomly swapped the paper and patent time-series of each gene and generated 10^4 null configurations. Using these null configurations, we obtained the *P* value of the observed proportion of genes whose patent falloffs preceded the paper falloffs (one-tailed test). This statistical test was

performed for the genes that debuted in papers or patents during similar periods, as described above.

The total paper versus patent numbers of each gene

The practical upper limit of the patent number of a gene can be expressed as a function of its paper number, $y = \alpha x^{\gamma}$ where x and y are the paper and patent numbers, respectively (Appendix 1–Fig. 9). The parameters γ and α were initially found by the fitting of the function with the least square method (polyfit in NumPy v1.19.1) and then adjusted manually. To test the statistical significance of the observed fraction of the genes below or on the upper limit, we randomly permuted the paper numbers of the genes (or equivalently, their patent numbers) and obtained the probability of the fraction of the genes (below or on the same above-mentioned upper limit) being greater than or equal to the observed fraction. This probability serves as the *P* value of the observed fraction (one-tailed test) and was computed with 10^4 trials of the randomization by the PCG algorithm in NumPy v1.19.1.

Selection of genes that signify biotechnology innovation

We first calculated the share of each gene in new annual pairs of genes, as defined above. We then identified the genes of the highest shares for ≥5 consecutive years or among the top five share genes for ≥10 years in 1990–2021. They were a total of nine genes that signify biotechnology innovation: Cas9 for genome editing [43], PD-L1 for immunomodulation and immunotherapy [57], mTOR, phosphatidylinositol 3-kinase (PI3-K), and TP53 mainly for cancer research [58, 59, 94], GFP and luciferase for live cell imaging [60, 61], β-catenin mainly for regenerative medicine and cancer research [62, 63], and β-galactosidase for DNA cloning [42]. We excluded PI3-K from further analysis due to its strong contextual overlap with mTOR through the biochemical pathway sharing. The remaining eight genes are referred to as *select genes* in this study.

Clinical trial analysis

We sought the papers with a given gene as the main focus in the above-mentioned way, while each paper contains the titles/abstracts with “clinicaltrials.gov” (case-insensitive) and “NCT” followed by an eight-digit number. Using this National Clinical Trials (NCT) number, we identified the type of the sponsor for that clinical trial from ClinicalTrials.gov (June 6, 2024), managed by the U.S. National Library of Medicine as a registry with information on global clinical trial activities. Among the sponsor types, we only considered “INDUSTRY”, “NIH”, “OTHER_GOV”, “FED”, and “OTHER”. In contrast to “INDUSTRY”, we confirmed the

public nature of those other sponsor types by reviewing their specific sponsors.

Since September 2007, the Food and Drug Administration Amendments Act of 2007 has mandated the posting of trials to ClinicalTrials.gov. According to our observation, it typically took ~6 years between the debut of a gene in papers/patents and its first trial, suggesting 2002– as the appropriate debut years of genes for the analysis. In addition, the genes of relatively recent debut years (2012–) were not considered because the number of these genes was relatively small and they may not yet have reached the levels of research stages for the full analysis. As a result, we focused on the genes of the debut years 2002–2011.

Among these genes in clinical trial papers as their main focus, we obtained the proportion of the genes in those papers before the patent falloffs (Fig. 4*C*, left "Lively"), and the proportion after (or at) the patent falloffs or with the patent time-series fully spanned by states II and III defined above (Fig. 4*C*, left "Low"). To calculate the probability of industry (or other) sponsorship for a clinical trial with each gene before its patent falloff, we obtained the average proportion of clinical trials with industry (or other) sponsorship described above, over all the genes in the clinical trial papers before the patent falloffs (Fig. 4*C*, right "Lively"). Similarly, we calculated the probability of industry (or other) sponsorship for a clinical trial with each gene after (or at) its patent falloff or of the patent time-series fully spanned by states II and III defined above (Fig. 4*C*, right "Low").

For the statistical significance test of the high probability of non-industry sponsorship for a clinical trial with each gene compared to industry sponsorship, we randomly but equally reassigned industry or other sponsorship to each clinical trial and generated the 10^4 null configurations. From these null configurations, we obtained the *P* value of the observed probability of non-industry sponsorship in the case before the patent falloff, or in the case after (or at) the patent falloff or of the patent time-series fully spanned by states II and III (one-tailed test).

Patent citation analysis

We collected the data of patent citations of papers from SciSciNet [76]. For compatibility with our paper data, the PubMed reference numbers (PMIDs) of papers in SciSciNet were retrieved from OpenAlex (May 25, 2024) [75]. We focused on highly patent-cited works, the papers that received $\geq$3 patent citations each within the first three years of publications. However, alternative criteria of high patent citations did not qualitatively change our analysis result. Among the genes of the main focus in the highly patent-cited works, we obtained the proportion of genes in those works before the patent falloffs (Fig. 4*D*, "Lively"), and the proportion after (or at) the patent falloffs or with the patent time-series fully spanned by states II and III defined above (Fig. 4*D*, "Low"). To control for variations from different debut

880 years of genes, we considered the genes of similar debut periods, 1990–2000 and 2001–
881 2011 separately. More recent debut years (2012–) were not considered for the same reason
882 as the previous analysis.
883
884 **Analysis of the teaming with new select genes**
885
886 We considered genes that had the patent falloff years defined above and were paired with
887 any select genes until the aforementioned peak times of these select genes. Among them,
888 we obtained the proportion paired with ≥1, ≥2, or ≥3 new select genes after (or at) the patent
889 falloffs, until the peak times of the select genes (Fig. 5*B*). In addition, for a given select gene,
890 we considered genes paired with this select gene until its peak time. Among them, we
891 obtained the proportion of genes initially paired with that early select gene before/after the
892 patent falloffs (Fig. 5*B* and Appendix 1–Fig. 10*A*). Similarly, we obtained the proportion for
893 the above-defined midpoint of the peak time, instead of the peak time (Appendix 1–Fig.
894 10*B*). All these analyses were performed for the genes that debuted in papers or patents
895 during a similar period, 1990–2000. Genes with post-2000 debuts were not considered due
896 to the fundamentally low overlap of their active research periods with the early stages of the
897 select genes.
898 For a given select gene, we next focused on the genes that satisfy the following
899 conditions: (*i*) they were initially paired with this select gene until its peak time or the
900 midpoint of the peak time, after (or at) the patent falloffs; (*ii*) their paper time-series are not
901 fully spanned by states II and III defined above; (*iii*) the patent falloffs preceded or coincided
902 with the paper falloffs. Among these genes, we obtained the proportion of genes initially
903 paired with that early select gene before/after the paper falloffs (Fig. 5*C* and Appendix 1–Fig.
904 11). In addition, we performed the statistical significance test of the higher proportion of
905 those genes before the paper falloffs than after (or at) the paper falloffs, observed for most
906 select genes. Specifically, we randomly reassigned the initial pairing times of each select
907 gene with equal *a priori* probabilities over the range from the select gene's debut year or its
908 paired gene's patent falloff year (whichever later) to the latest point of at least one paper or
909 patent of this paired gene. Across all the relevant select genes and their paired genes, we
910 generated 10^4 null configurations of the initial pairing times and obtained the *P* value of the
911 observed, joint proportions of genes initially paired with the select genes before the paper
912 falloffs. All the analyses were performed for the genes that debuted in papers or patents
913 during a similar period (1990–2000), as discussed earlier.

Missing exploratory opportunities under commercial pressure

To examine the effect of commercial pressure on the exploratory opportunities of genetic research, we conducted the "thought experiment" for each select gene, as follows.

Define set *A* as the genes that teamed with this select gene until its peak time and satisfy the following condition (*i*) or (*ii*): (*i*) they were initially paired with that early select gene after (or at) the patent falloffs and the patent falloffs preceded or coincided with the paper falloffs; or (*ii*) the patent time-series are fully spanned by states II and III defined above but the paper time-series are not. In addition, define set *B* as a subset of *A*, where genes were initially paired with that early select gene before the paper falloffs. Also, define set *C* as a subset of *A*, where genes were initially paired with that early select gene after (or at) the paper falloffs.

For the genes initially paired with a given early select gene with their limited patent volumes (i.e., set *A*), we consider a hypothetical scenario that the papers undergo immediate falloffs due to the limited patent volumes. This scenario allows us to estimate the maximum extent to which the early teaming with this select gene would be lost if the research activities become hindered by the commercial pressure from the industry's limited interest. The estimated proportion of genes that would lose this early teaming is given by $(|A| - x_0)/|A|$ (Fig. 5*D* and Appendix 1–Fig. 12), where x_0 is the number of genes that would still maintain the teaming despite the paper falloffs. Here, $x_0 \equiv P|B| + |C|$ with $P \equiv |C|/|A|$. P denotes the conditional probability that each gene in B would still maintain the teaming. By definition, all genes in C would clearly maintain the teaming.

We further define set D as all genes that teamed with a given early select gene. We also define set E as $E \equiv D - B$. Among all the genes in D whether affected or not by the above hypothetical scenario, the estimated proportion of genes that would lose the early teaming with the select gene is given by $(|D| - x_1)/|D|$ (Fig. 5*D* and Appendix 1–Fig. 12), where $x_1 \equiv P|B| + |E|$ with P defined previously. This derivation is similar to the above formulation.

Additionally, we conducted the same thought experiment, but for the teaming with each select gene until the above-defined midpoint of its peak time. All these analyses were performed for the genes that debuted in papers or patents during a similar period (1990–2000), as discussed previously. In the example of COX-2 or PDE5, the estimated proportion of the chances to lose the early teaming with a particular select gene is given by $1 - P$, calculated for this select gene.

Code availability

We used custom codes for our data analysis and modeling (Python 3.10.4). For all random number generations in statistical significance tests, we used the Mersenne Twister in random.py of Python 3.10.4. Our source codes have been deposited to the public repository GitHub (https://github.com/WooJoongKim0107/GeneBib_demo).

Acknowledgments

We thank Taekyun Kim, Hong Lok Lung, Seolmin Yang, Daehyun Kim, Liming Xiong, Yiji Xia, Adam G. Craig, and Anna Oi Wah Leung for useful discussions. This work was supported by the National Research Foundation of Korea Grants RS-2023-00263411 and RS-2024-00353098 funded by the Ministry of Science and ICT (J.C., W.J.K., R.L., and C.-M.G.), Hong Kong Baptist University Start-up Grant for New Academics (R.L. and P.-J.K.), the National Science Foundation Grant Award EF-2133863 and the National Research Foundation of Korea, Global Humanities and Social Sciences Convergence Research Program 2024S1A5C3A02042671 (H.Y.), and the UNIST Supercomputing Center for computing resources.

References

1. Bush V. Science the Endless Frontier: A Report to the President. Washington, DC: United States Office of Scientific Research and Development; 1945.
2. Stokes DE. Pasteur's Quadrant: Basic Science and Technological Innovation. Washington, DC: Brookings Institution; 1997.
3. Mokyr J. The Gifts of Athena: The Origin of the Knowledge Economy. Princeton, NJ: Princeton University; 2002.
4. Hiltzik MA. Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age. New York, NY: HarperCollins; 1999.
5. Georgescu I. Bringing back the golden days of Bell Labs. Nat Rev Phys. 2022;4(2):76-8.
6. Ahmed N, Wahed M, Thompson NC. The growing influence of industry in AI research. Science. 2023;379(6635): 884-6.
7. Kumar A, Blum J, Le TT, Havelange N, Magini D, Yoon IK. The mRNA vaccine development landscape for infectious diseases. Nat Rev Drug Discov. 2022;21(5): 333-4.
8. Gibney E. Quantum gold rush: the private funding pouring into quantum start-ups. Nature. 2019;574: 22-4.
9. Agarwal R, Gaule P. What drives innovation? Lessons from COVID-19 R&D. J Health Econ. 2022;82: 102591.

10. Krieger J, Li D, Papanikolaou D. Missing novelty in drug development. Rev Financ Stud. 2022;35: 636-79.
11. Teece DJ. Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. Res Policy. 1986;15: 285-305.
12. Blumenthal D, Causino N, Campbell E, Louis KS. Relationships between academic institutions and industry in the life sciences—an industry survey. N Engl J Med. 1996;334(6): 368-73.
13. Pellegrino G. Barriers to innovation in young and mature firms. J Evol Econ. 2018;28: 181-206.
14. Azoulay P, Ding W, Stuart T. The impact of academic patenting on the rate, quality, and direction of (public) research output. J Ind Econ. 2009;57(4): 637-76.
15. Lee ED, Kempes CP, West GB. Idea engines: Unifying innovation & obsolescence from markets & genetic evolution to science. Proc Natl Acad Sci U S A. 2024;121(6): e2312468120.
16. McElheny VK. Drawing the Map of Life: Inside the Human Genome Project. New York, NY: Basic Books; 2010.
17. Park M, Leahey E, Funk RJ. Papers and patents are becoming less disruptive over time. Nature. 2023;613(7942): 138-44.
18. Bloom N, Jones CI, Van Reenen J, Webb M. Are ideas getting harder to find? Am Econ Rev. 2020;110: 1104-44.
19. Gates AJ, Gysi DM, Kellis M, Barabási A-L. A wealth of discovery built on the Human Genome Project - by the numbers. Nature. 2021;590(7845): 212-5.
20. Ahmadpoor M, Jones BF. The dual frontier: Patented inventions and prior scientific advance. Science. 2017;357(6351): 583-7.
21. Poege F, Harhoff D, Gaessler F, Baruffaldi S. Science quality and the value of inventions. Sci Adv. 2019;5(12): eaay7323.
22. Shi F, Evans J. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nat Commun. 2023;14(1): 1641.
23. Kwon S. Underappreciated government research support in patents. Science. 2024; 385(6712): 936-8.
24. McGuire AL, Gabriel S, Tishkoff SA, Wonkam A, Chakravarti A, Furlong EEM, et al. The road ahead in genetics and genomics. Nat Rev Genet. 2020;21(10): 581-96.
25. Wu W, Zhang HH, J. WM, B KP. Gene Biotechnology. 2nd ed. Boca Raton, FL: CRC Press; 2003.

1020 26. Stoeger T, Amaral LAN. The characteristics of early-stage research into human genes
1021 are substantially different from subsequent research. PLOS Biol. 2022;20(1): e3001520.
1022 27. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic
1023 Acids Res. 2021;49(D1): D480-9.
1024 28. Narin F, Noma E. Is technology becoming science? Scientometrics. 1985;7: 369-81.
1025 29. Narin F, Noma E, Perry R. Patents as indicators of corporate technological strength. Res
1026 Policy. 1987;16: 143-55.
1027 30. Meyer M. Does science push technology? Patents citing scientific literature. Res Policy.
1028 2000;29: 409-34.
1029 31. Sampat B, Williams HL. How do patents affect follow-on innovation? Evidence from the
1030 human genome. Am Econ Rev. 2019;109(1): 203-36.
1031 32. Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al.
1032 SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically
1033 proven protease inhibitor. Cell. 2020;181(2): 271-80.
1034 33. Stoeger T, Amaral LAN. Meta-Research: COVID-19 research risks ignoring important host
1035 genes due to pre-established research patterns. eLife. 2020;9: e61981.
1036 34. Stephanopoulos G, Aristidou AA, Nielsen JH. Metabolic Engineering : Principles and
1037 Methodologies. San Diego: Academic Press; 1998.
1038 35. Sun L, Lee JW, Yook S, Lane S, Sun ZQ, Kim SR, et al. Complete and efficient
1039 conversion of plant cell wall hemicellulose into high-value bioproducts by engineered
1040 yeast. Nat Commun. 2021;12(1): 4975.
1041 36. Provart NJ, Alonso J, Assmann SM, Bergmann D, Brady SM, Brkljacic J, et al. 50 years
1042 of *Arabidopsis* research: highlights and future directions. New Phytol. 2016;209(3): 921-
1043 44.
1044 37. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. Orphans as taxonomically
1045 restricted and ecologically important genes. Microbiol-Sgm. 2005;151: 2499-501.
1046 38. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans:
1047 are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):
1048 404-13.
1049 39. Youn H, Strumsky D, Bettencourt LMA, Lobo J. Invention as a combinatorial process:
1050 evidence from US patents. J R Soc Interface. 2015;12(106): 20150272.
1051 40. Fink TMA, Reeves M. How much can we influence the rate of innovation? Sci Adv.
1052 2019;5(1): eaat6107.
1053 41. Ke Q. Technological impact of biomedical research: The role of basicness and novelty.
1054 Res Policy. 2020;49: 104071.

1055 42. Ullmann A, Jacob F, Monod J. Characterization by *in vitro* complementation of a peptide
1056 corresponding to an operator-proximal segment of beta-galactosidase structural gene of
1057 *Escherichia coli*. J Mol Biol. 1967;24(2): 339-43.
1058 43. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable
1059 dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science.
1060 2012;337(6096): 816-21.
1061 44. Topol EJ. Failing the public health—rofecoxib, Merck, and the FDA. N Engl J Med.
1062 2004;351: 1707-9.
1063 45. Shin DM, et al. Chemoprevention of head and neck cancer by simultaneous blocking of
1064 epidermal growth factor receptor and cyclooxygenase-2 signaling pathways: preclinical
1065 and clinical studies. Clin Cancer Res. 2013;19(5): 1244-56.
1066 46. Saba NF, et al. Chemoprevention of head and neck cancer with celecoxib and erlotinib:
1067 results of a phase Ib and pharmacokinetic study. Cancer Prev Res. 2014;7(3): 283-91.
1068 47. Groeger AL, et al. Cyclooxygenase-2 generates anti-inflammatory mediators from
1069 omega-3 fatty acids. Nat Chem Biol. 2010;6(6): 433-41.
1070 48. Chen JS, et al. Nonsteroidal anti-inflammatory drugs dampen the cytokine and antibody
1071 response to SARS-CoV-2 infection. J virol. 2021;95(7): e00014-21.
1072 49. Jackson G, Gillies H, Osterloh I. Past, present, and future: a 7-year update of Viagra®
1073 (sildenafil citrate). Int J Clin Pract 2005;59(6): 680-91.
1074 50. Bhogal S, Khraisha O, Al Madani M, Treece J, Baumrucker SJ, Paul TK. Sildenafil for
1075 pulmonary arterial hypertension. Am J Ther. 2019;26(4): e520-6.
1076 51. U.S. National Institutes of Health. ClinicalTrials.gov. Available from:
1077 http://www.clinicaltrials.gov.
1078 52. Tinsley HN, et al. Colon tumor cell growth-inhibitory activity of sulindac sulfide and other
1079 nonsteroidal anti-inflammatory drugs is associated with phosphodiesterase 5 inhibition."
1080 Cancer Prev Res. 2010;3(10): 1303-13.
1081 53. Tinsley HN, et al. Inhibition of PDE5 by sulindac sulfide selectively induces apoptosis and
1082 attenuates oncogenic Wnt/β-catenin–mediated transcription in human breast tumor cells.
1083 Cancer Prev Res. 2011;4(8): 1275-84.
1084 54. Nelson SF, Miceli MC, Elashoff RM, Sweeney HL, Victor RG. PDE5 inhibition alleviates
1085 functional muscle ischemia in boys with Duchenne muscular dystrophy. Neurology.
1086 2014;82(23): 2085-91.
1087 55. Wang R, Lei X, Rakesh CK. PDE5 inhibitor tadalafil and hydroxychloroquine cotreatment
1088 provides synergistic protection against type 2 diabetes and myocardial infarction in mice.
1089 J Pharmacol Exp Ther. 2017;361(1): 29-38.

1090 56. Amedro P, et al. Efficacy of phosphodiesterase type 5 inhibitors in univentricular
1091 congenital heart disease: the SV-INHIBITION study design. ESC Heart Fail. 2020;7(2):
1092 747-56.
1093 57. Ohaegbulam KC, Assal A, Lazar-Molnar E, Yao Y, Zang XX. Human cancer
1094 immunotherapy with antibodies to the PD-1 and PD-L1 pathway. Trends Mol Med.
1095 2015;21(1): 24-33.
1096 58. Zou ZL, Tao T, Li HM, Zhu X. mTOR signaling pathway and mTOR inhibitors in cancer:
1097 progress and challenges. Cell Biosci. 2020;10(1): 31.
1098 59. Mantovani F, Collavin L, Del Sal G. Mutant p53 as a guardian of the cancer cell. Cell
1099 Death Differ. 2019;26(2): 199-212.
1100 60. Lippincott-Schwartz J, Patterson GH. Development and use of fluorescent protein
1101 markers in living cells. Science. 2003;300(5616): 87-91.
1102 61. Liu S, Su YC, Lin MZ, Ronald JA. Brightening up biology: advances in luciferase systems
1103 for *in vivo* Imaging. Acs Chem Biol. 2021;16(12): 2707-18.
1104 62. Liu JQ, Xiao Q, Xiao JN, Niu CX, Li YY, Zhang XJ, et al. Wnt/beta-catenin signalling:
1105 function, biological mechanisms, and therapeutic opportunities. Signal Transduct Tar.
1106 2022;7(1): 3.
1107 63. Zhang Y, Wang X. Targeting the Wnt/beta-catenin signaling pathway in cancer. J
1108 Hematol Oncol. 2020;13(1): 165.
1109 64. English K, Barry FP, Field-Corbett CP, Mahon BP. IFN-γ and TNF-α differentially
1110 regulate immunomodulation by murine mesenchymal stem cells. Immunol Lett.
1111 2007;110(2): 91-100.
1112 65. Liu X, Ji Q, Zhang C, Liu X, Liu Y, Liu N, et al. miR-30a acts as a tumor suppressor by
1113 double-targeting COX-2 and BCL9 in H. pylori gastric cancer models. Sci Rep. 2017;7(1):
1114 7113.
1115 66. Biswas KH, Sopory S, Visweswariah SS. The GAF domain of the cGMP-binding, cGMP-
1116 specific phosphodiesterase (PDE5) is a sensor and a sink for cGMP. Biochemistry.
1117 2008;47(11): 3534-43.
1118 67. Küçük A, Yucel M, Erkasap N, Tosun M, Koken T, Ozkurt M, et al. The effects of PDE5
1119 inhibitory drugs on renal ischemia/reperfusion injury in rats. Mol Biol Rep. 2012;(39):
1120 9775-82.
1121 68. Tavallai M, Hamed HA, Roberts JL, Cruickshanks N, Chuckalovcak J, Poklepovic A, et
1122 al. Nexavar/Stivarga and viagra interact to kill tumor cells. J Cell Physiol. 2015;230(9):
1123 2281-98.
1124 69. Krimsky S. Science in the Private Interest: Has the Lure of Profits Corrupted Biomedical
1125 Research? Lanham, MD: Rowman & Littlefield; 2003.

70. BP cuts funding for ‘most promising’ biofuel. Investigate Midwest. August 20, 2015; https://investigatemidwest.org/2015/08/20/bp-cuts-funding-for-most-promising-biofuel-2.
71. Stocker M, Baffes J, Vorisek D. What triggered the oil price plunge of 2014–2016 and why it failed to deliver an economic impetus in eight charts. World Bank Blogs. January 18, 2018; https://blogs.worldbank.org/en/developmenttalk/what-triggered-oil-price-plunge-2014-2016-and-why-it-failed-deliver-economic-impetus-eight-charts.
72. Asase RV, Okechukwu QN, Ivantsova MN. Biofuels: present and future. Environ Dev Sustain. 2024; https://doi.org/10.1007/s10668-024-04992-w.
73. Mandt R, Seetharam K, Cheng CHM. Federal R&D funding: the bedrock of national innovation. MIT Sci Policy Rev. 2020;1: 44-54.
74. Marx M, Fuegi A. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. J Econ Manage Strat. 2022;31(2): 369-92.
75. Priem J, Piwowar H, Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [Preprint]. 2022. Available from: https://arxiv.org/abs/2205.01833
76. Lin Z, Yin Y, Liu L, Wang D. SciSciNet: A large-scale open data lake for the science of science research. Sci Data. 2023;10(1): 315.
77. Haessler P, Giones F, Brem A. The who and how of commercializing emerging technologies: A technology-focused review. Technovation. 2023;121: 102637.
78. Levinthal DA. The slow pace of rapid technological change: Gradualism and punctuation in technological change. Ind Corp Change. 1998;7(2): 217-47.
79. Scaringella L, Miles RE, Truong Y. Customers involvement and firm absorptive capacity in radical innovation: The case of technological spin-offs. Technol Forecast Soc Change. 2017;120: 144-62.
80. Arthur WB. The structure of invention. Res Policy 2007;36: 274-87.
81. Arnold C. Death by climate change. Nat Clim Change. 2022;12(7):607-9.
82. Perry BL, Aronson B, Pescosolido BA. Pandemic precarity: COVID-19 is exposing and exacerbating inequalities in the American heartland. Proc Natl Acad Sci U S A. 2021;118(8): e2020685118.
83. Yun J, Kim P-J, Jeong H. Anatomy of scientific evolution. PLOS ONE. 2015;10(2): e0117388.
84. Chen CY, Kahanamoku SS, Tripati A, Alegado RA, Morris VR, Andrade K, et al. Systemic racial disparities in funding rates at the National Science Foundation. eLife. 2022;11: e83071.
85. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). Nucleic Acids Res. 2005;33: D154-D9.

1162 86. Browet A, Absil P-A, Dooren PV. Fast community detection using local neighbourhood
1163 search. arXiv:1308.6276 [Preprint]. 2013. Available from: https://arxiv.org/abs/1308.6276
1164 87. Lin Y, Michel J-B, Aiden EL, Orwant J, Brockman W, Petrov S. Syntactic annotations for
1165 the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the
1166 Association for Computational Linguistics. 2012: 169-74.
1167 88. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database
1168 resources of the National Center for Biotechnology Information in 2023. Nucleic Acids
1169 Res. 2022;50: D20-6.
1170 89. Fredkin E. Trie Memory. Commun Acm. 1960;3(9): 490-9.
1171 90. Wei CH, Allot A, Lai PT, Leaman R, Tian S, Luo L, et al. PubTator 3.0: an AI-powered
1172 literature resource for unlocking biomedical knowledge. Nucleic Acids Res, 2024;52(W1):
1173 W540-6.
1174 91. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine.
1175 Nucleic Acids Res. 2021;49(D1): D325-34.
1176 92. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a
1177 major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):
1178 D1074-82.
1179 93. Ghosh M, Erhardt S, Rose ME, Buunk E, Harhoff D. PaECTER: patent-level
1180 representation learning using citation-informed transformers. arXiv:2402.19411 [Preprint].
1181 2024. Available from: https://arxiv.org/abs/2402.19411
1182 94. Fruman DA, Christian R. PI3K and cancer: lessons, challenges and opportunities. Nat
1183 Rev Drug Discov. 2014;13(2): 140-56.

Figures

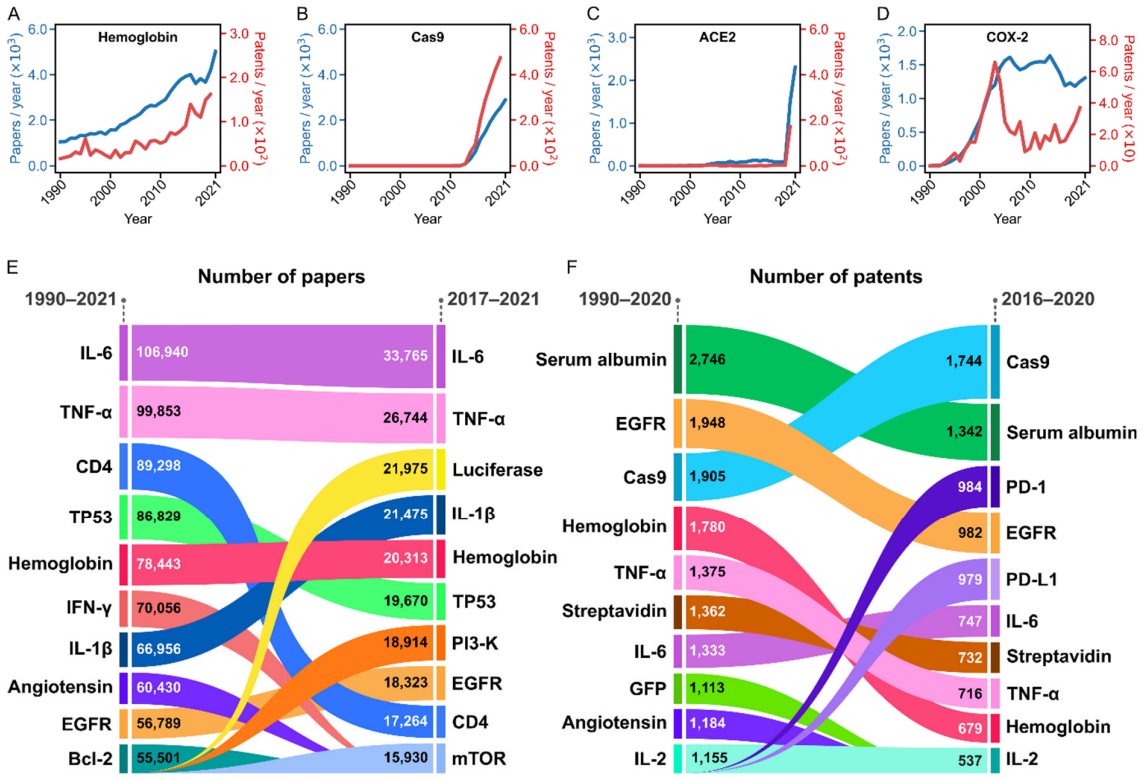


Fig. 1. Genetic research trend since the launch of the Human Genome Project (1990). ***A***–***D***, Annual counts of papers (blue) or patents (red) whose titles/abstracts include a given gene. ***E***, ***F***, Ten most frequent genes in papers or patents of all time (1990–2021 for papers and 1990–2020 for patents) and latest five years (2017–2021 for papers and 2016–2020 for patents). The reason for selecting different periods for paper and patent counting is explained in Methods. The vertical width of each ribbon is proportional to the paper or patent number of the respective gene in that period.

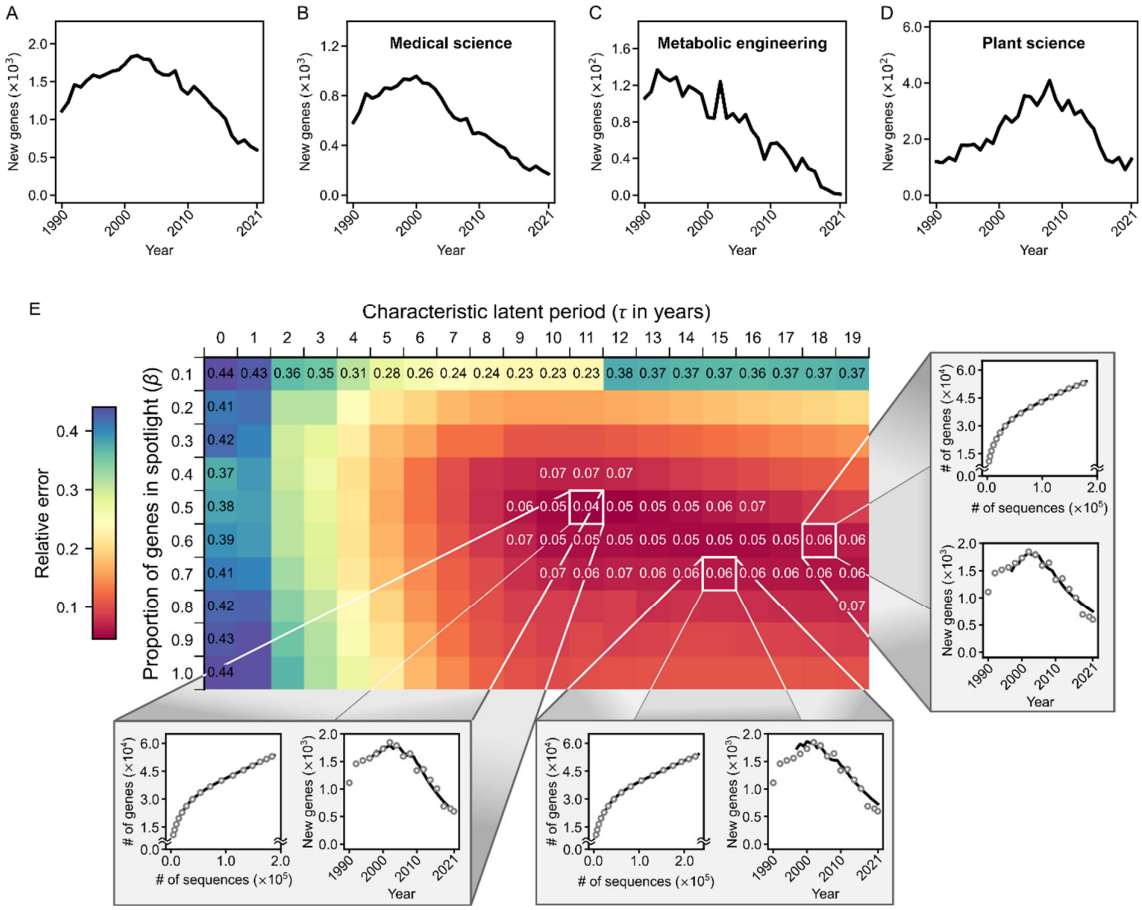


Fig. 2. Research on new genes. *A*, Annual counts of new genes in papers and patents. ***B***–***D***, Annual counts of new genes in each thematic gene category (Methods). ***E***, Output of our mathematical model for the annual counts of newly researched genes. A heatmap shows the relative errors of the annual counts of the new genes between the model output and actual data over parameters τ (horizontal) and β (vertical), while the model assumes that fraction β of newly reported amino acid (AA) sequences would be studied sometime later with characteristic latent period τ (Methods). For particular τ and β with the small relative error, the heatmap is projected to the graph of the cumulative number of the studied genes as a function of the number of their non-redundant AA sequences, i.e., individual gene copies (left) and the graph of the annual counts of newly studied genes (right), which is relevant to ***A***: the linear interpolation of the model output (solid line) and the real data points (circle) are shown together in these graphs. Over a broad range of τ and β, the model produces the quantitative pattern in ***A***.

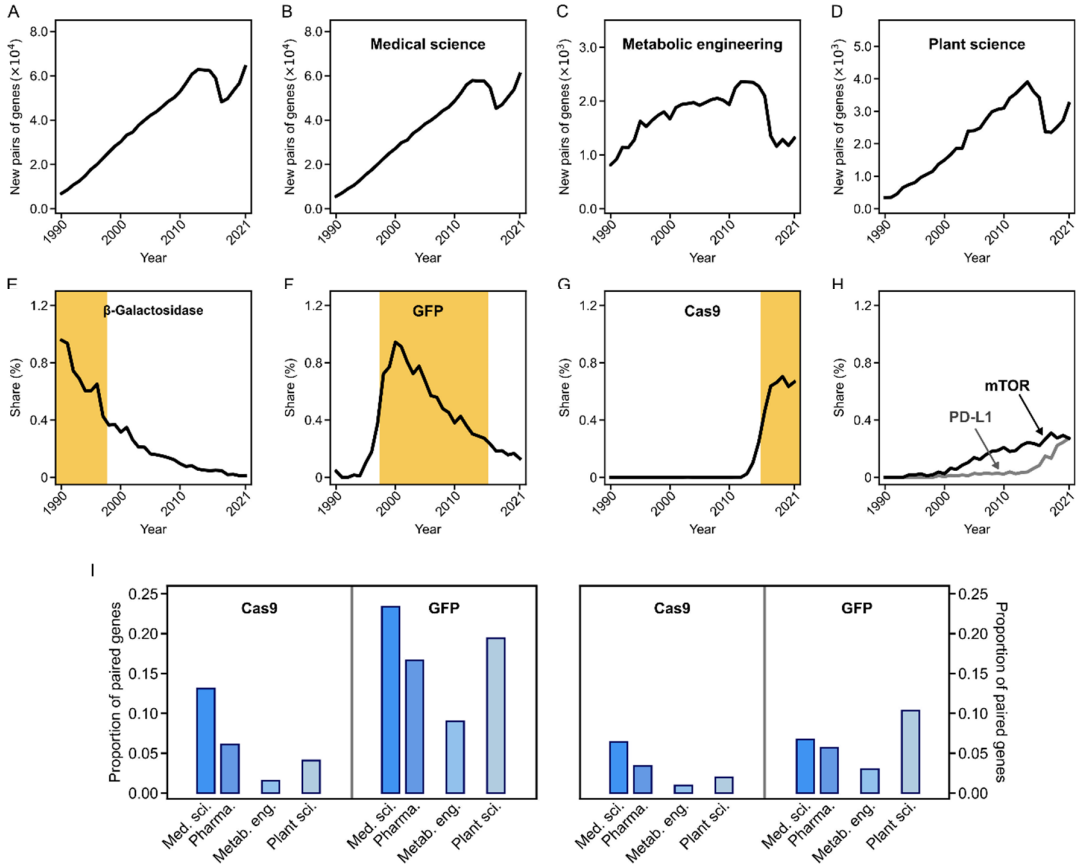


Fig. 3. Research on novel gene combinations. ***A***, Annual counts of new pairs of genes in papers and patents. For separate counting of paper and regional patent cases, and for larger gene combinations beyond a pair, see Appendix 1–Fig. 2*E–P*. ***B–D***, Annual counts of new pairs of genes in each thematic gene category (Methods). ***E–G***, For each gene that once had the highest share (relative frequency) in new pairs of genes, its share is presented over time. Shaded is the period when this gene had the highest shares. ***H***, The shares of PD-L1 and mTOR in new pairs of genes over time. They were not the highest but still ranked high. ***I***, The proportion of genes in each thematic gene category, paired with Cas9 or GFP until the peak time of its annual paper and patent counts. In this pairing, we considered genes that debuted in papers/patents in 1990–2000 (left) and 2001–2011 (right), separately (Methods). Abbreviations: med. sci., medical science; pharma., pharmaceutical; metab. eng., metabolic engineering; plant sci., plant science.

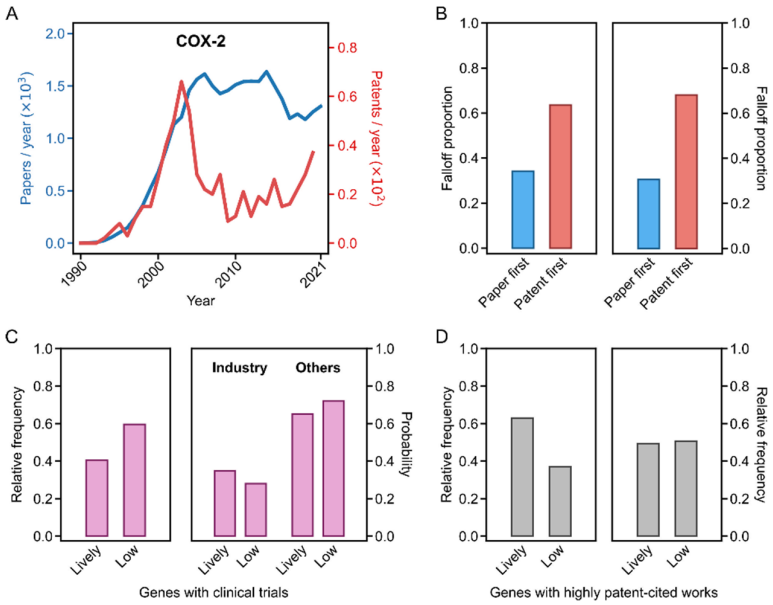


Fig. 4. Paper–patent falloff relationship and research continuation. ***A***, Annual paper (blue) or patent (red) counts of COX-2. ***B***, The proportion of genes whose paper falloffs precede the patent falloffs (blue) and the proportion of genes whose patent falloffs precede the paper falloffs (red). Here, we considered genes that debuted in papers/patents in 1990–2000 (left) and 2001–2011 (right), separately. For a fair comparison between paper and patent falloffs, we set the endpoint of the paper time-series to the same year (2020) as the patent time-series. ***C***, On the left side, the proportion of genes with clinical trial papers before ("Lively") or after/at ("Low") their patent falloffs, among the clinical-trial genes that debuted in papers/patents in 2002–2011. These debut years were chosen regarding the first year of the mandated trial posting to ClinicalTrials.gov (Methods). On the right side, the probability of industry or other sponsorship for a clinical trial with each gene in a given case ("Lively" or "Low") from the left side. ***D***, The proportion of genes with papers before ("Lively") or after/at ("Low") their patent falloffs, if these papers received ≥3 patent citations each within the first three years of publications. We considered the genes that debuted in papers/patents in 1990–2000 (left) and 2001–2011 (right), separately. Alternative criteria for high patent citations did not qualitatively change the observed result here.

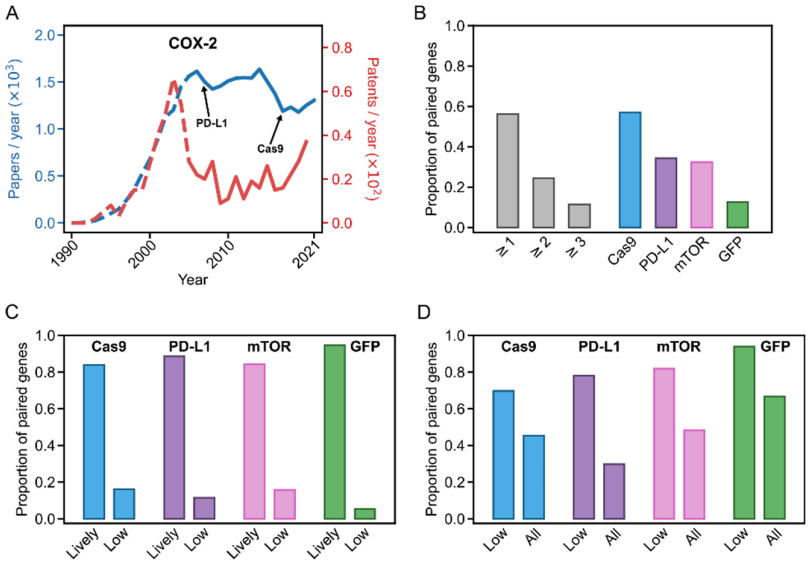
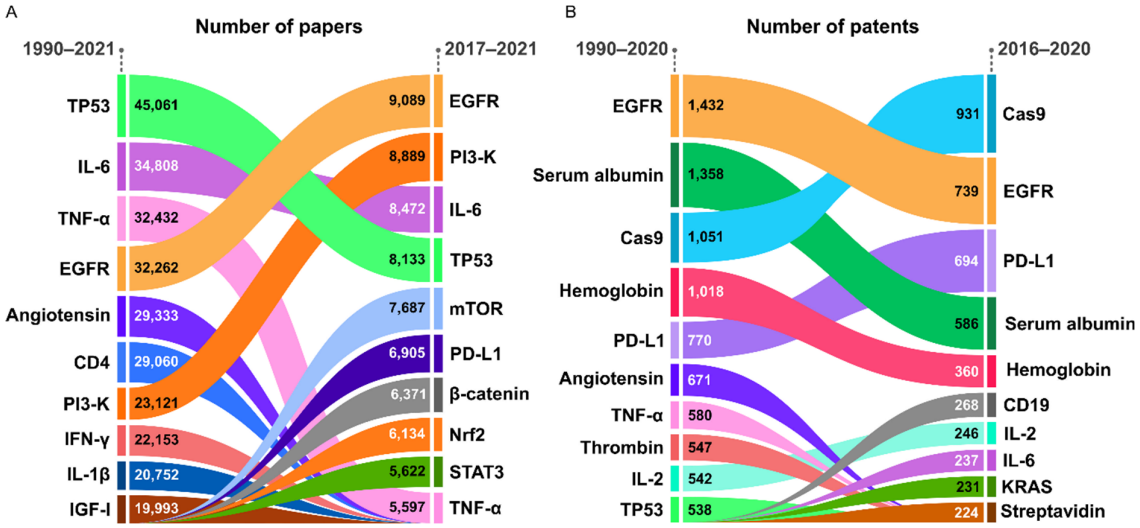
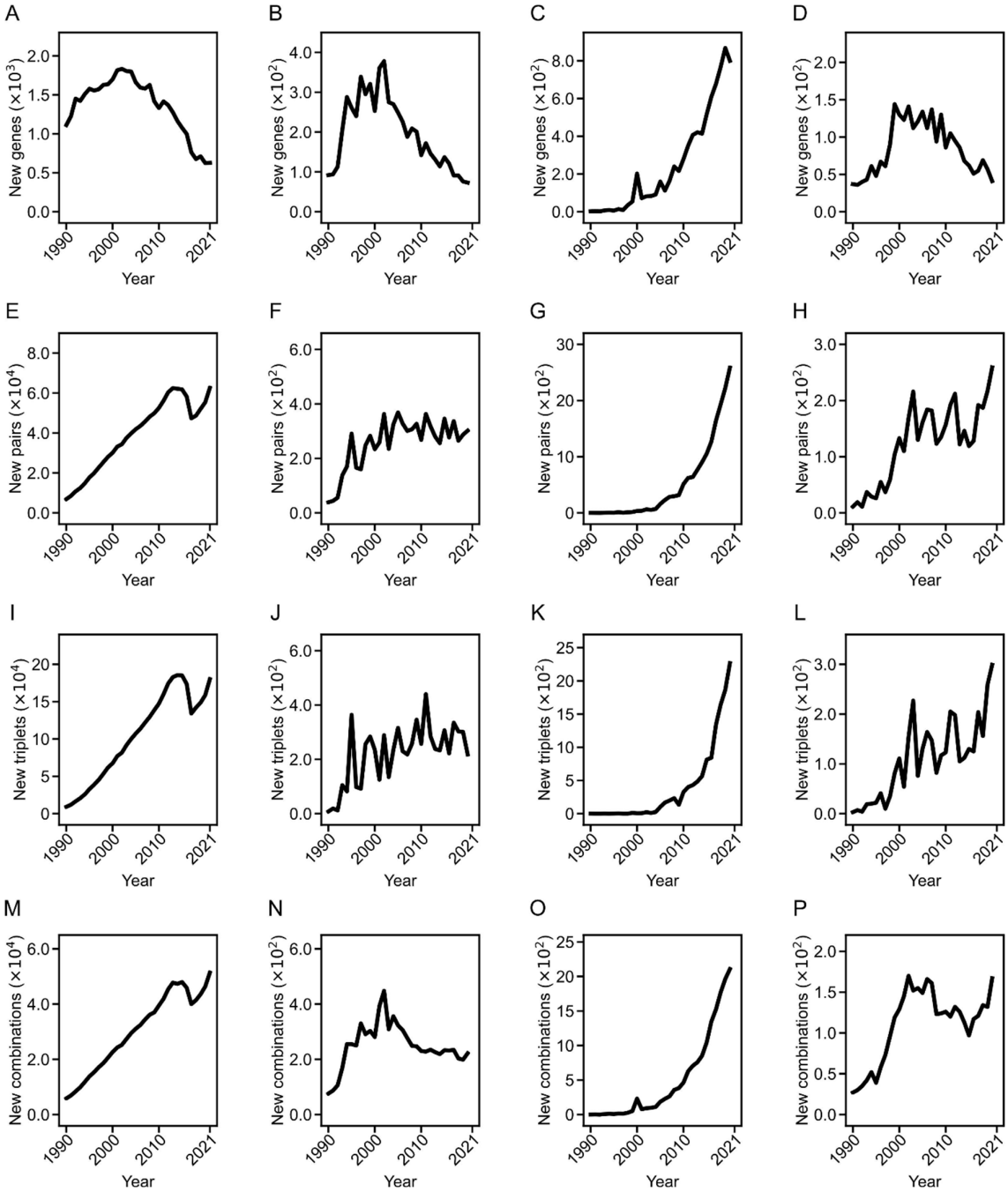


Fig. 5. Teaming with new select genes in their early stages. ***A***, Annual paper (blue) or patent (red) counts of COX-2 (dashed line for a period before its patent falloff). An arrow indicates the initial time of the teaming with each select gene if this teaming comes after or coincides with the patent falloff. ***B***, The proportion of genes teaming with ≥1, ≥2, or ≥3 new select genes after (or at) the patent falloffs, and the proportion of genes initially teaming with a given select gene—Cas9, PD-L1, mTOR, or GFP—after (or at) their patent falloffs (Methods). We only considered the teaming with the select genes until the peak times of their annual paper and patent counts. ***C***, The proportion of genes initially teaming with a given select gene before ("Lively") or after/at ("Low") their paper falloffs, among the genes with this teaming after/at the patent falloffs (Methods). We only considered the teaming with the select gene until the peak time of its annual paper and patent counts. ***D***, The estimated proportion of genes that would lose the teaming with a given select gene, among the genes that initially teamed with this select gene amid their limited patent applications ("Low") or among all the genes that teamed with this select gene ("All") (Methods). We only considered the teaming with the select gene until the peak time of its annual paper and patent counts. In ***B***–***D***, we considered genes that debuted in papers/patents in 1990–2000. Genes with post-2000 debuts were not considered due to the fundamentally low overlap of their research periods with the early stages of the select genes.

Appendix 1

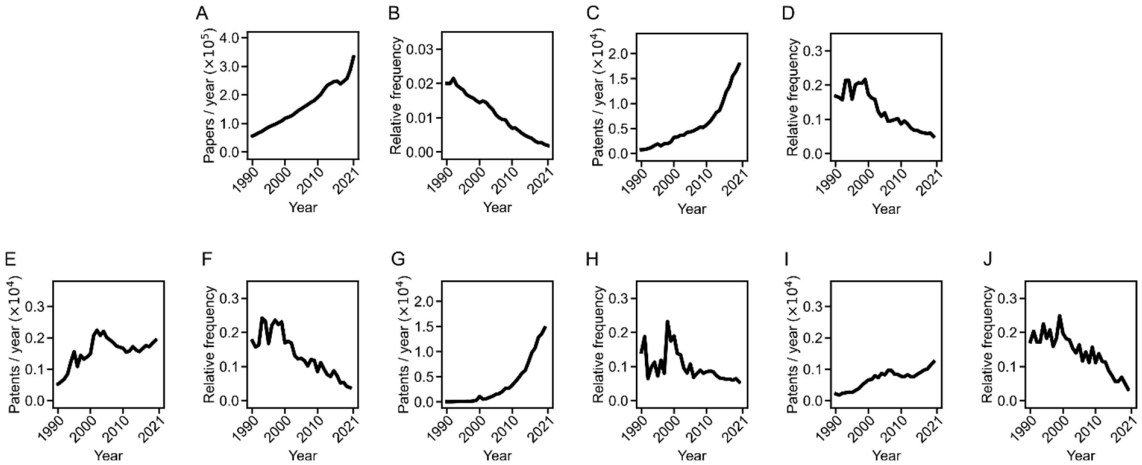


Appendix 1–Fig. 1. Genes of frequent research focus. Related to Fig. 1*E*,*F*. Ten most frequent genes of the main focus in papers or patents of all time (1990–2021 for papers and 1990–2020 for patents) and latest five years (2017–2021 for papers and 2016–2020 for patents). For the identification of the genes of the main focus in papers or patents, refer to Methods. The vertical width of each ribbon is proportional to the paper or patent number of the respective gene in that period.

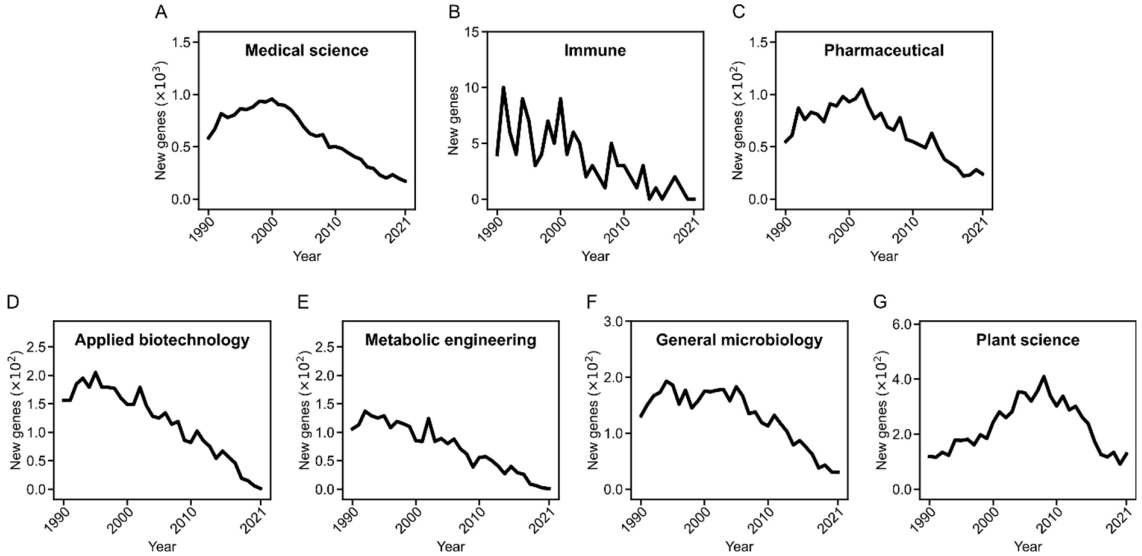


Appendix 1–Fig. 2. Annual counts of newly researched genes and novel gene combinations. Related to Fig. 2*A*, 3*A*. ***A–D***, Annual counts of new genes in only papers (***A***), and only patents from the US (***B***), China (***C***), or Europe (***D***) whichever the earliest. ***E–P***, Annual counts of new pairs (***E–H***), new triplets (***I–L***), or new entire combinations (***M–P***) of genes in only papers (***E***, ***I***, and ***M***), and only

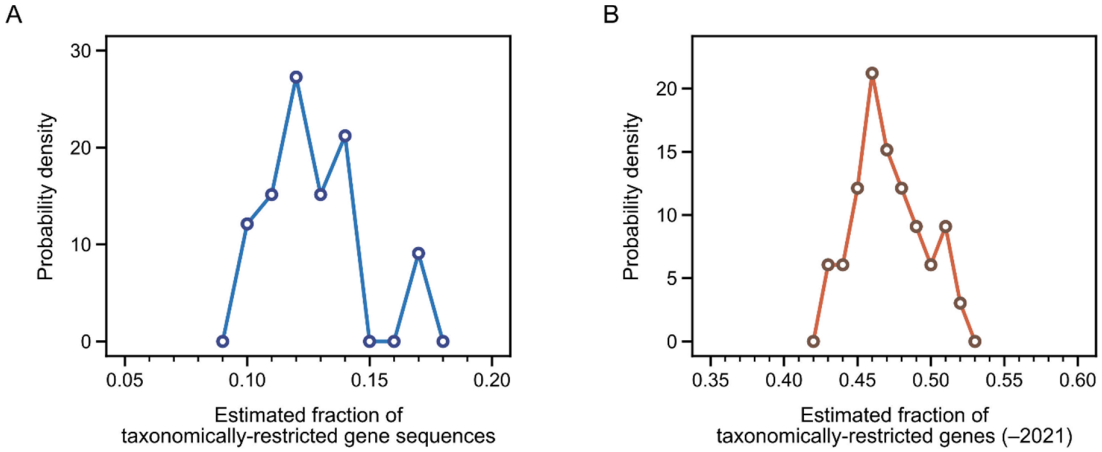
patents from the US (***F***, ***J***, and ***N***), China (***G***, ***K***, and ***O***), or Europe (***H***, ***L***, and ***P***) whichever the earliest.



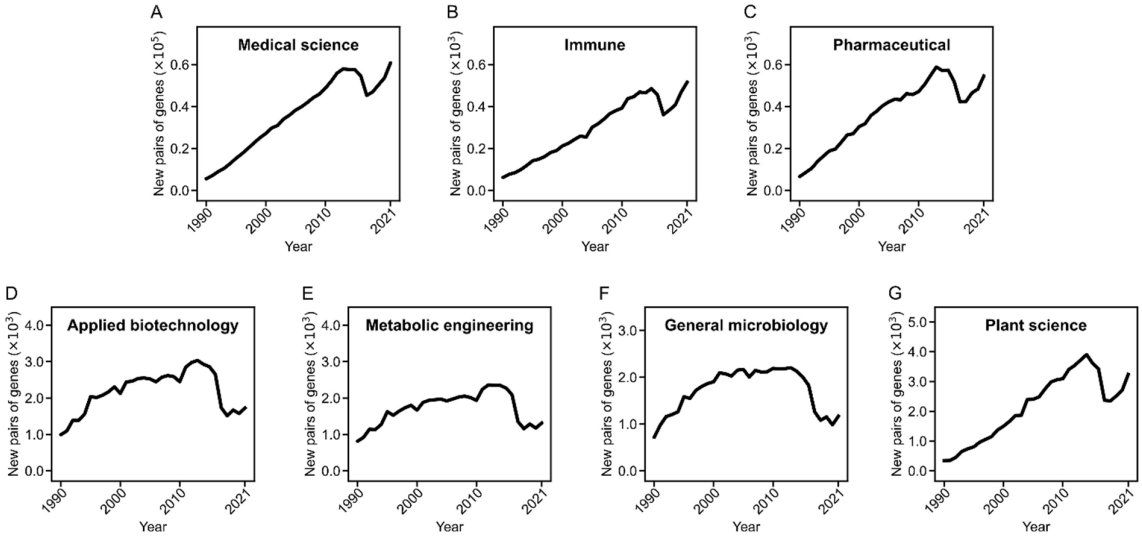
Appendix 1–Fig. 3. New gene occurrence per paper or patent. Related to Fig. 2*A*. ***A***, ***B***, The total annual counts of papers (***A***) and their proportion with new genes (***B***). ***C***, ***D***, The total annual counts of patents (***C***) and their proportion with new genes (***D***). ***E***, ***F***, The total annual counts of only US patents (***E***) and their proportion with new genes (***F***). ***G***, ***H***, The total annual counts of only Chinese patents (***G***) and their proportion with new genes (***H***). ***I***, ***J***, The total annual counts of only European patents (***I***) and their proportion with new genes (***J***).



Appendix 1–Fig. 4. Research on new genes in each thematic gene category. Related to Fig. 2*B*–*D*. ***A***–***G***, Annual counts of newly researched genes in each thematic gene category.

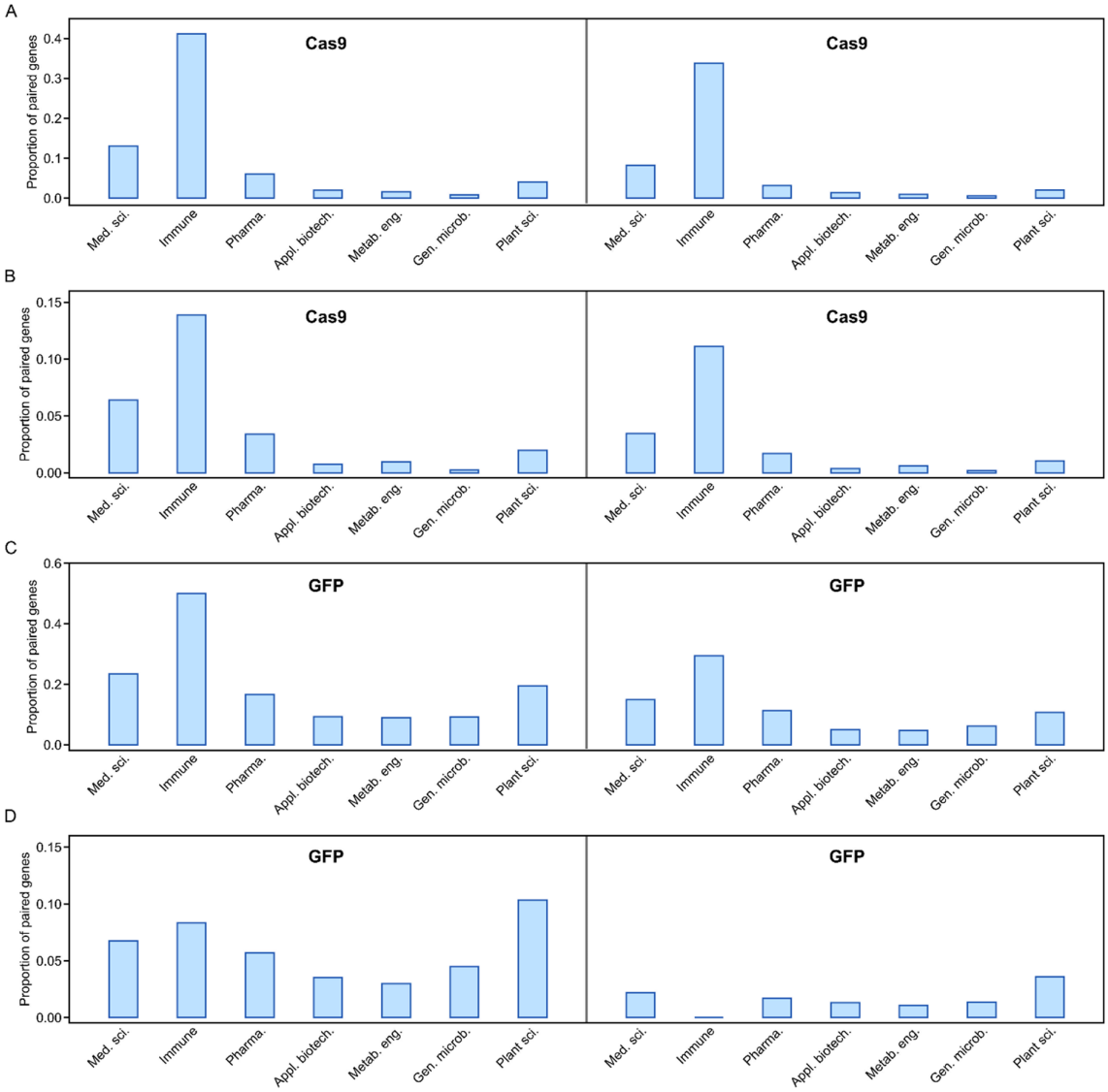


Appendix 1–Fig. 5. Some estimates from our mathematical model. Related to Fig. 2*E*. ***A***, The probability distribution of the fraction of taxonomically-restricted gene (TRG) sequences among all the possible non-redundant sequences of research merit, estimated from our mathematical model (Methods). It shows that TRGs makes up 9.9–17.3% of the sequences of research merit. This result is compatible with the existing knowledge that each sequenced genome contains 10–20% of genes as TRGs [38], as this known fraction is likely to set an upper bound of the TRGs of research merit given the TRGs' tendency of being less researched. ***B***, The probability distribution of the fraction of TRGs among all researched genes until 2021, estimated from our mathematical model (Methods). It shows that TRGs constitute 43.1–51.8% of the researched genes as of 2021. This result is compatible with the fraction (59.1%) of the genes present in single species (i.e., single entries) in UniProtKB/Swiss-Prot, because the fraction from UniProtKB/Swiss-Prot is possibly an overestimate with its missing species that may have the homologs of those single-entry genes.

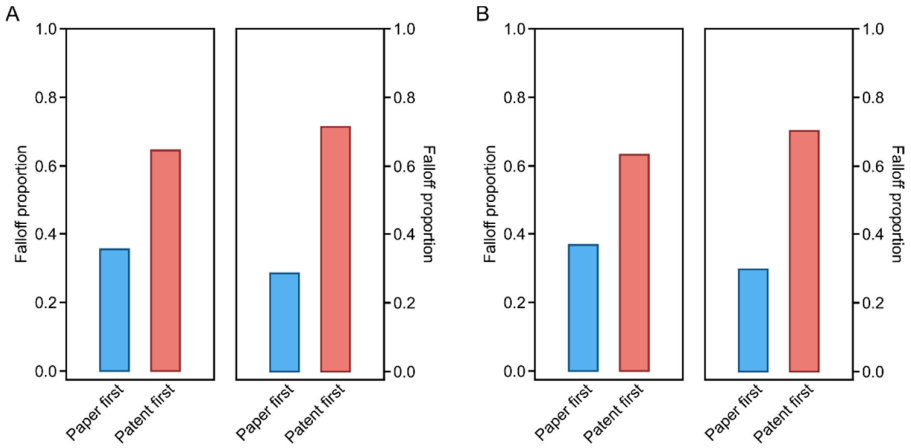


Appendix 1–Fig. 6. Novel gene combinations in each thematic gene category. Related to Fig.

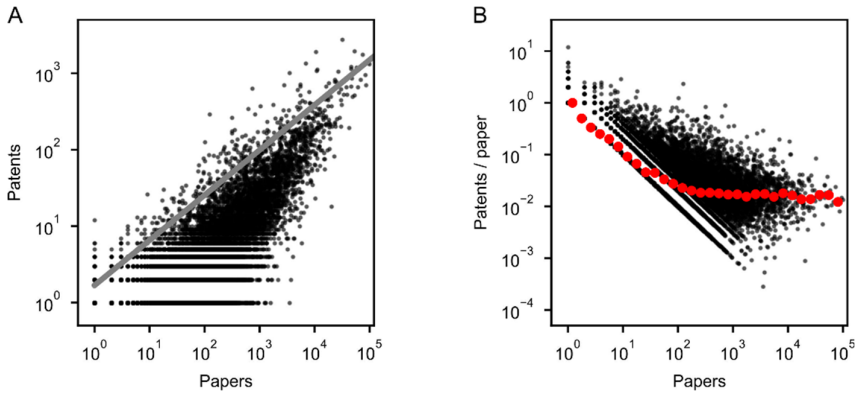
3*B*–*D*. ***A***–***G***, Annual counts of new pairs of genes in each thematic gene category.



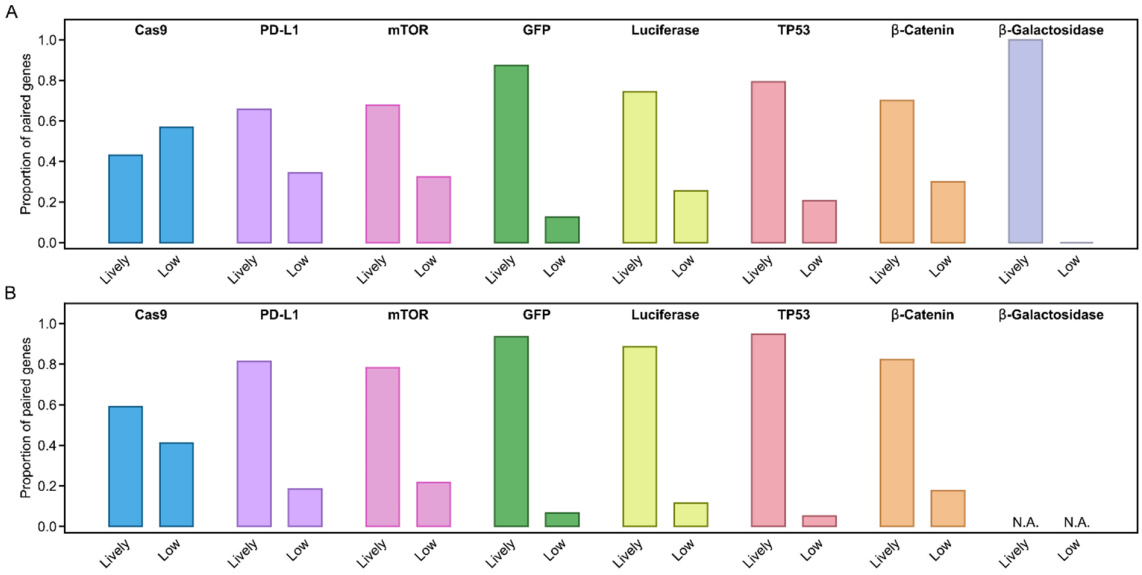
Appendix 1–Fig. 7. Teaming with Cas9 or GFP in each thematic gene category. Related to Fig. 3*I*. The proportion of genes in each thematic gene category, paired with Cas9 (***A*** and ***B***) or GFP (***C*** and ***D***) until its peak time (left) or earlier midpoint (right). For the definitions of the peak time and midpoint, refer to Methods. In this teaming, we considered genes that debuted in papers/patents in 1990–2000 (***A*** and ***C***) and 2001–2011 (***B*** and ***D***), separately. Abbreviations: med. sci., medical science; pharma., pharmaceutical; appl. biotech., applied biotechnology; metab. eng., metabolic engineering; gen. microb., general microbiology; plant sci., plant science.



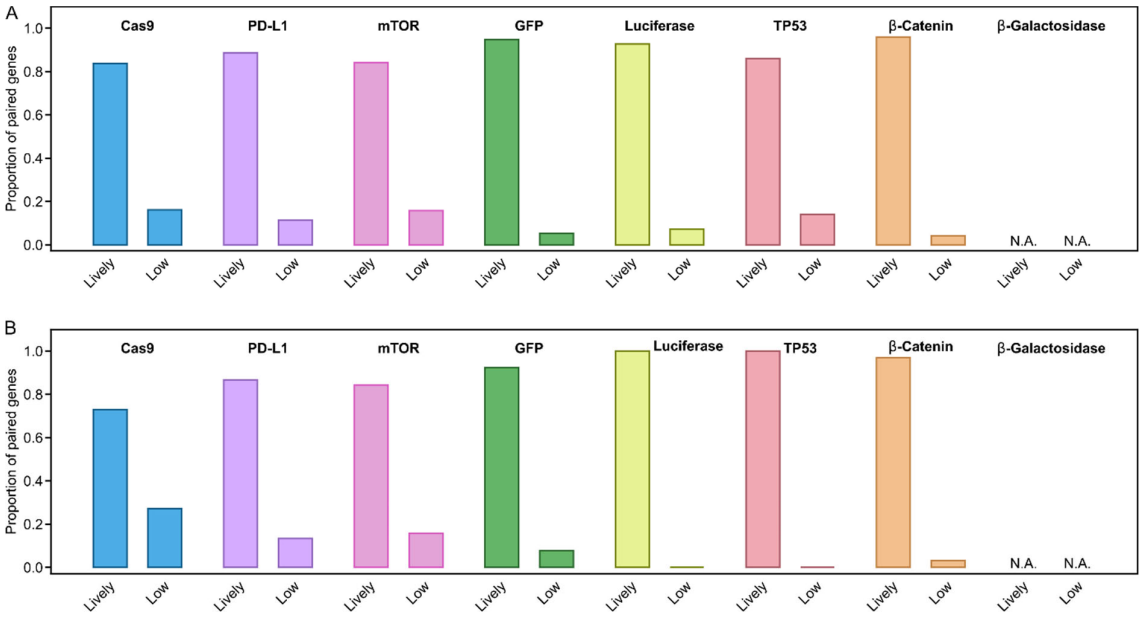
Appendix 1–Fig. 8. Paper–patent falloff relationship with or without highly similar prior works. Related to Fig. 4*B*. ***A***, ***B***, The proportion of genes whose paper falloffs precede the patent falloffs (blue) and the proportion of genes whose patent falloffs precede the paper falloffs (red). In this analysis, we first measured the content similarity of every pair of the papers (or patents) with each gene, using PaECTER [93]. If the total number of the papers (or patents) exceeds 5,000 for that gene, we only used 5,000 randomly selected from them to ensure computational feasibility. ***A*** shows the results with all these papers and patents, and ***B*** shows the results after filtering out the papers and patents highly overlapping with their prior works (Methods). In ***A*** and ***B***, we considered genes that debuted in papers/patents in 1990–2000 (left) and 2001–2011 (right), separately. For a fair comparison between the paper and patent falloffs, we set the endpoint of the paper time-series to the same year (2020) as the patent time-series.



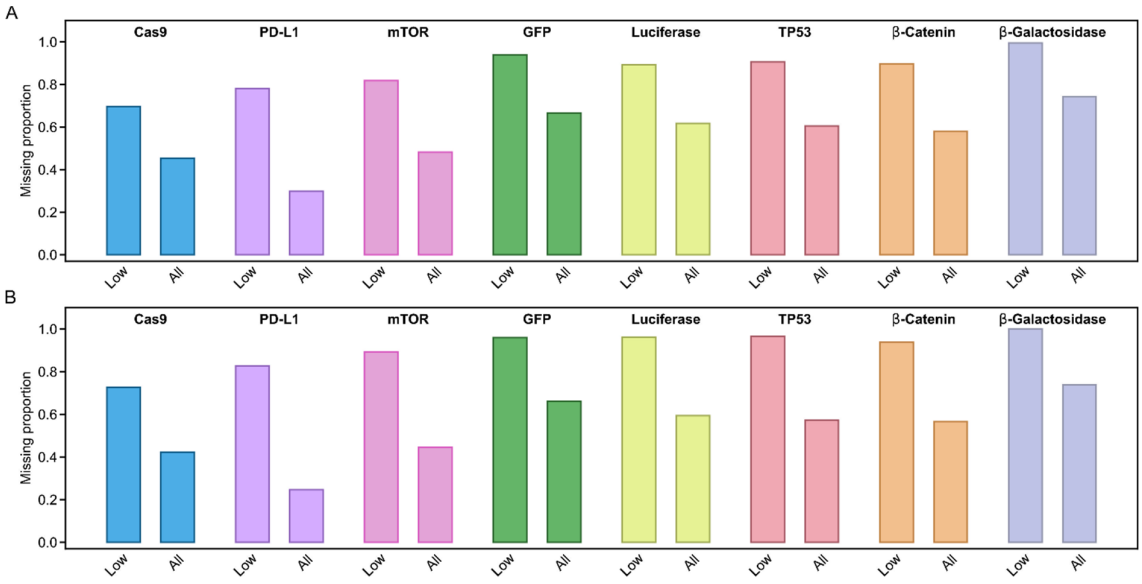
Appendix 1–Fig. 9. Paper versus patent volumes. Related to Fig. 4*B*. ***A***, A scatter plot between the total paper and patent numbers of each gene. A gray straight line indicates $y = \alpha x^{\gamma}$ where x and y are the paper and patent numbers of a gene, respectively, $\gamma = 0.59$, and $\alpha = 1.68$. Most genes (98.8%) are placed below or on this line ($P < 10^{-4}$ and Methods). In other words, the paper number of each gene tends to put an *upper limit* or a cap on its patent number. ***B***, A scatter plot between the total paper number and patent-to-paper ratio of each gene (the median in each interval of the paper number is plotted in red). The more papers published for a gene, the fewer patents tend to be applied per paper, eventually at a level of ~1.6 patents/100 papers. In ***A***, ***B***, papers and patents are counted over the same periods (1990–2020) for their fair comparison. Only the genes with non-zero hits in both papers and patents are considered here.



Appendix 1–Fig. 10. Patent falloff and teaming with new select genes. Related to Fig. 5*B*. ***A***, ***B***, The proportion of genes initially paired with a given select gene before ("Lively") or after/at ("Low") their patent falloffs. We here focus on the teaming until the peak time of this select gene (***A***) or the earlier midpoint (***B***) (Methods). In the case of too small a number of genes itself whether paired with the select gene before or after the patent falloffs, this case is marked with "N.A." All the results were obtained with genes that debuted in papers/patents in a similar period (1990–2000). Genes with post-2000 debuts were not considered due to the fundamentally low overlap of their research periods with the early stages of the select genes.



Appendix 1–Fig. 11. Paper publications and teaming with new select genes. Related to Fig. 5*C*. ***A***, ***B***, The proportion of genes initially paired with a given select gene before ("Lively") or after/at ("Low") their paper falloffs, among the genes with this teaming after/at the patent falloffs. We here focus on the teaming until the peak time of this select gene (***A***) or the earlier midpoint (***B***) (Methods). In the case of too small a number of genes itself whether paired with the select gene before or after the paper falloffs, this case is marked with "N.A." All the results were obtained with genes that debuted in papers/patents in a similar period (1990–2000). Genes with post-2000 debuts were not considered due to the fundamentally low overlap of their research periods with the early stages of the select genes.



Appendix 1–Fig. 12. Missing opportunities for the teaming with select genes. Related to Fig. 5*D*. ***A***, ***B***, The estimated proportion of genes that would lose the teaming with a given select gene, among the genes that initially teamed with this select gene with their limited patent volumes ("Low") or among all the genes that teamed with this select gene ("All"). We here focus on the initial teaming until the peak time of that select gene (***A***) or the earlier midpoint (***B***) (Methods). The results were obtained with genes that debuted in papers/patents in a similar period (1990–2000). Genes with post-2000 debuts were not considered due to the fundamentally low overlap of their research periods with the early stages of the select genes.

1455

1456

Year / Rank	1990–1994	1995–1999	2000–2004	2005–2009	2010–2014	2015–2019	2020	2021
1	β-Gal	GFP	GFP	GFP	GFP	Cas9	Cas9	Cas9
2	IL-2	β-Gal	Luc	TP53	Luc	Luc	Luc	Luc
3	IL-6	Luc	TP53	Luc	TP53	mTOR	mTOR	PD-L1
4	CAT	TP53	β-Gal	PI3-K	PI3-K	PI3-K	PI3-K	mTOR
5	EGF	TNF-α	PI3-K	β-Catenin	mTOR	β-Catenin	IL-6	β-Catenin

1457

1458 **Appendix 1–Table 1. Frequent genes in novel gene pairs in each period.** Five genes of the
1459 largest shares in new pairs of genes in papers or patents in each 5-year period up to 2019, and in
1460 2020 and 2021 separately. Abbreviations: β-Gal, β-galactosidase; IL-2, interleukin-2; IL-6, interleukin-
1461 6; CAT, chloramphenicol acetyltransferase; EGF, epidermal growth factor; GFP, green fluorescent
1462 protein; Luc, luciferase; TP53, tumor suppressor protein p53; TNF-α, tumor necrosis factor-α; PI3-K,
1463 phosphatidylinositol 3-kinase; mTOR, mammalian target of rapamycin; Cas9, CRISPR associated
1464 protein 9; PD-L1, programmed cell death ligand 1.

1465

1466

1467

1468

Section	Attribute	Remark
Protein description	Recommended name	
	Alternative name	
	Allergen	
	CD antigen	
	INN (International Nonproprietary Names)	
Gene names	Primary name	
	Synonyms	
	Ordered locus number	Excluded from our analysis
	ORF (open reading frame) number	Excluded from our analysis

1469

1470 **Appendix 1–Table 2. Available attributes of gene and protein names in UniProtKB/Swiss-Prot**
1471 **entries.** Both **"**Ordered locus numbers" and "ORF numbers" were excluded in the listing of our gene
1472 names, regarding their frequent use in the cases of not-fully-identified gene products. All the other
1473 attributes were considered for the listing of the gene names.