SOCIAL SCIENCES

Cross-disciplinary evolution of the genomics revolution

Alexander M. Petersen[1]*†, Dinesh Majeti[2]*, Kyeongan Kwon[2], Mohammed E. Ahmed[2], Ioannis Pavlidis[2]†

Born out of the Human Genome Project (HGP), the field of genomics evolved with phenomenal speed into a dominant scientific and business force. While other efforts were intent on estimating the economic impact of the genomics revolution, we shift focus to the social and cultural capital generated by bridging together biology and computing—two of the constitutive disciplines of "genomics". We quantify this capital by measuring the pervasiveness of bio-computing cross-disciplinarity (*XD*) in genomics research during and after the HGP. To provide interlocking perspectives at the career and epistemic levels, we assembled three data sets to measure *XD* via (i) the collaboration network between 4190 biology and computing faculty from 155 departments in the United States, (ii) cross-departmental affiliations within a comprehensive set of human genomics publications, and (iii) the application of computational concepts and methods in research published in a preeminent genomics journal. Our results show the following: First, research featuring *XD* collaborations has higher citation impact than other disciplinary research—an effect observed at both the career and individual article levels. Second, genomics articles featuring *XD* methods tend to have higher citation impact than epistemically pure articles. Third, *XD* researchers of computing pedigree are drawn to the biology culture. This statistical evidence acquires deeper meaning when viewed against the organizational and knowledge transfer mechanisms revealed by the data models. With cross-disciplinary initiatives set to dominate the agenda of funding agencies, our case study provides a framework for appreciating the long-term effects of these initiatives on science and its standard-bearers.

INTRODUCTION

With the coming of the 21st century, biology has emerged as the vanguard of the scientific enterprise and computing as the epicenter of engineering and technology (*1*). These two fields were primed as complements in the Human Genome Project (HGP). The successful completion of the HGP in the early 2000s ushered the genomics revolution that continues to transform our capacity to understand, predict, and edit life (*2*–*5*).

A significant amount of work has focused on estimating the impact of the HGP on human health and the return on investment in the U.S. economy (*6*–*8*). These efforts proved surprisingly difficult, highlighting the broader challenge of evaluating the socioeconomic impact of science policy (*9*, *10*).

Instead of focusing on economic and health outcomes we focus here on the evolution of social and cultural capital within the genomics revolution. To this end, we build on scholarship of epistemic (*11*) and network analysis (*12*–*16*) to quantify the factors and career incentives that contribute to the formation of new fields (*17*–*19*) in a team science context (*20*–*25*).

The HGP (1990–2003) was a singular opportunity for scientists from several disciplines, with biology and computing being prominent among them. For this reason, this project serves as a rich case study for science of science (*26*) to investigate the social and behavioral elements underlying cross-disciplinary research. Consequently, we adopt a mixed methods analytic approach that focuses not only at the epistemic level but also at the scholar level.

Specifically, we begin with a network analysis, focused on U.S. academia, of the administratively invisible or informal cross-disciplinary biology-computing collaborations that we dub a "college," to illustrate the explosion of the crossdisciplinary population parallel to, and also in the wake of, the HGP. By cross-disciplinary population, we refer to faculty from biology and computing who achieve their research objectives via collaboration across this disciplinary boundary. Upon further inspection, we find that the overwhelming majority (90%) of the faculty forming this biology-computing bridge have been active in genomics research. Building on this insight from the descriptive analysis, we then apply cross-sectional regression to the 4190 faculty in our data set, showing a positive correlation between one's inclination toward cross-disciplinary collaboration and total career citation impact. To pinpoint the source of this lifetime advantage, we implement a longitudinal panel regression, demonstrating that, within each scholarly career, the cross-disciplinary publications have significantly higher citation impact than the disciplinary ones.

As this scholar-centered result is based on a U.S. academic data set, to test for its broader relevance, we analyze a comprehensive set of human genomics publications from the international literature. We find that, in this set, publications with joint authorship from biology and computing scholars also have significantly higher citation impact than publications with biology authorship only.

Finally, we analyze cross-disciplinarity from an epistemological perspective, operationalizing the mixing of domain knowledge rather than the mixing of scholars. To be specific, we use publication-level keywords to identify the methods applied within research articles from a leading genomics journal. Our results demonstrate that articles with an explicit computational component have significantly higher citation impact than articles without an explicit computational component.

Together, these outcomes show that cross-disciplinarity in genomics is pervasive and impactful, creating upward mobility in morphed careers and generating dominant hybrid intellectual and social capital that have persisted long after the end of the HGP. The legacy of the HGP also survives in organizational capital that is fundamental to "consortium science," whereby teams of teams organize around central challenges, with a common goal to share benefits equitably within and beyond institutional boundaries.

[1]Ernest and Julio Gallo Management Program, Department of Management of Complex Systems, School of Engineering, University of California Merced, Merced, CA 95343, USA. [2]Computational Physiology Laboratory, University of Houston, Houston, TX 77204, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: ipavlidis@uh.edu (I.P.); apetersen3@ucmerced.edu (A.M.P.)

RESULTS

Cross-disciplinarity in genomics careers

Career data set

We anchor our analysis on individual scholars so that we can control for scholar-specific attributes and account for career-level decisions regarding one's orientation toward cross-disciplinary activity. A principal challenge in this endeavor was that there were no complete, validated, and publicly available data sets for the biology-computing college. Hence, we synthesized a large longitudinal data set using criteria that could, in principle, be generalized to other case studies.

We focused on the biology-computing college in the United States, because this is the source of the HGP, and it was a task that we could practically complete. Our entry point for classifying scholars was through their primary academic affiliation. Specifically, we accessed the websites of 155 biology and computing departments in the United States (table S1), matching the faculty listings to individual Google Scholar (GS) profiles. We consider the primary departmental affiliation of each faculty scholar to be a lifelong disciplinary trait, because while faculty may change institutions, the likelihood that they change from a computing department to a biology department, or vice versa, is very low. We verified this premise by examining publicly available curricula vitae (CVs) for many of the faculty.

As such, from this point forward, we will refer to these biology (n = 2077) and computing (n = 2113) faculty as $\mathcal{F}$, indexed by i (n = 4190 in total). We carefully examined the publication profile of each $\mathcal{F}_i$, removing spurious content (for example, articles that did not include the respective scholar's name). This process yielded a total of 413,565 publications that were incorporated into the faculty career data set.

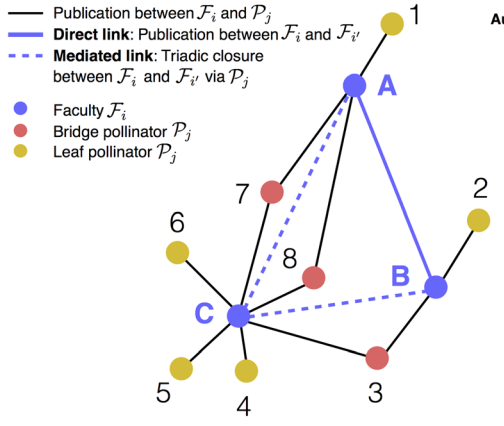
Collaboration network and analytical framework for the career data set

To build the collaboration network of the biology-computing college, we inspected each disambiguated $\mathcal{F}_i$ profile for instances of direct collaboration with another $\mathcal{F}_{i'}$; as a result, we identified 3900 $\mathcal{F}_i$ who collaborated with at least one other $\mathcal{F}_{i'}$, forming 16,799 links. Within the subset of connected $\mathcal{F}_i$, the size of the largest ("giant") connected component was 3869. Therefore, just 7.6% of the $\mathcal{F}_i$ were not part of the largest connected component, and only 6.9% of the $\mathcal{F}_i$ were completely disconnected. Description of the name disambiguation method and a thorough investigation of the network's structural properties can be found in sections S1 and S2, respectively.

We operationalized cross-disciplinarity by investigating both the direct collaborations and indirect associations within the $\mathcal{F}$ network, both of which are important to knowledge transfer and science development. In more detail, the network's nodes can be connected via two types of links, as illustrated in Fig. 1: "Direct collaboration" refers to a link between two faculty $\mathcal{F}_i$ and $\mathcal{F}_{i'}$ who appear together in at least one publication, and "mediated association" refers to a link between two faculty $\mathcal{F}_i$ and $\mathcal{F}_{i'}$ who are indirectly associated via a common non-$\mathcal{F}$ coauthor. This non-$\mathcal{F}$ coauthor creates the link via "triadic closure" between the two $\mathcal{F}$. Because many published researchers are not faculty in one of the 155 listed departments, the typical $\mathcal{F}_i$ has many more mediated associations than direct collaborations with other faculty in our data set (fig. S2).

We use the primary departmental affiliations, which we treat as time-invariant traits, to define three disciplinary orientations $\mathcal{O}$ for $\mathcal{F}$. If $\mathcal{F}_i$ collaborated with at least one $\mathcal{F}_{i'}$ from the opposite department, then we classify him/her as cross-disciplinary $\mathcal{O}(\mathcal{F}_i) \equiv XD_{\mathcal{F}}$; otherwise, $\mathcal{F}_i$ is classified as $\mathcal{O}(\mathcal{F}_i) \equiv BIO_{\mathcal{F}}$, or $CS_{\mathcal{F}}$, depending on her/his primary departmental affiliation. The group sizes are nearly equal: $BIO_{\mathcal{F}}$, n = 1353; $CS_{\mathcal{F}}$, n = 1590; and $XD_{\mathcal{F}}$, n = 1247. We further examined each member of the $XD_{\mathcal{F}}$ group by finding their corresponding Scopus author profile, which contains career-level keywords derived from their publications. We found that 90% of the $XD_{\mathcal{F}}$ faculty feature the Scopus keyword "genetics" in their curated profiles, indicating that the overwhelming majority of the $XD_{\mathcal{F}}$ group have been involved in genomics research. This consistency check confirms the soundness of our $XD_{\mathcal{F}}$ classification scheme.

As mentioned earlier, there are many collaborators of $\mathcal{F}$ who are not explicitly included in our starting sample, possibly because they are not faculty in one of the listed biology or computing departments (for example, PhD students, postdocs, and other international



Author index i / j	Total degree K_i	Coauthors	Department $\mathcal{F}_i$	Orientation $\mathcal{O}(\mathcal{F}_i)$, $\mathcal{O}(\mathcal{P}_i)$
A	4	B	*BIO*	$XD_{\mathcal{F}}$
B	3	A	*CS*	$XD_{\mathcal{F}}$
C	6	NULL	*CS*	$CS_{\mathcal{F}}$
1	1	A	N/A	$BIO_{\mathcal{P}}$
2	1	B	N/A	$CS_{\mathcal{P}}$
3	2	B,C	N/A	$CS_{\mathcal{P}}$
4	1	C	N/A	$CS_{\mathcal{P}}$
5	1	C	N/A	$CS_{\mathcal{P}}$
6	1	C	N/A	$CS_{\mathcal{P}}$
7	2	A,C	N/A	$XD_{\mathcal{P}}$
8	2	A,C	N/A	$XD_{\mathcal{P}}$

Fig. 1. Construction of the $\mathcal{F}$ network. Schematic network, serving as an instructive example of our method for classifying the faculty $\mathcal{F}_i$, their pollinator coauthors $\mathcal{P}_j$, and the links between them. The network corresponds to the table on the right. Two types of links connect the faculty nodes: a direct link ($\mathcal{F}_i - \mathcal{F}_{i'}$) if $\mathcal{F}_i$ and $\mathcal{F}_{i'}$ are coauthors of at least one publication together, and a mediated link ($\mathcal{F}_i - \mathcal{P}_j - \mathcal{F}_{i'}$) if there is at least one $\mathcal{P}_j$ that has coauthored separately with $\mathcal{F}_i$ and $\mathcal{F}_{i'}$, thereby mediating a triadic closure between the two $\mathcal{F}$. We classified each $\mathcal{F}_i$ according to her/his main discipline: $BIO_{\mathcal{F}}$ = biology and $CD_{\mathcal{F}}$ = computing unless they have collaborated with at least one $\mathcal{F}_{i'}$ from the other discipline, in which case the classification $XD_{\mathcal{F}}$ supersedes their original disciplinary classification. We classified the non-$\mathcal{F}$ coauthors $\mathcal{P}_j$ as bridge pollinators if they coauthored with two or more faculty; otherwise, these $\mathcal{P}_j$ are classified as leaf pollinators. Among the bridge pollinators, we classified those $\mathcal{P}_j$ who coauthor with faculty from both biology and computing as cross-pollinators. Thus, the solid link connecting A-B represents a direct cross-disciplinary link, the dashed link connecting C-A represents a mediated cross-disciplinary link, and pollinators 7 and 8 are cross-pollinators because they have collaborated with faculty from each discipline. N/A, not applicable.

researchers). These collaborators are still crucial for understanding the role of cross-disciplinarity in the genomics revolution, as they constitute the academic ecosystem or "invisible college" surrounding tenure-track faculty (*27*). We identify these non-$\mathcal{F}$ collaborators as pollinators $\mathcal{P}$, indexed by *j*.

In contradistinction to faculty $\mathcal{F}$, we do not have knowledge of the departmental affiliations of pollinators $\mathcal{P}$. Hence, we infer their disciplinary orientation by observing their coauthorship patterns with faculty $\mathcal{F}$. Specifically, (i) biology pollinators $\mathcal{O}(\mathcal{P}j) \equiv BIO_{\mathcal{P}}$, if they collaborated with $\mathcal{F}$ from biology departments only; (ii) computing pollinators $\mathcal{O}(\mathcal{P}j) \equiv CS_{\mathcal{P}}$, if they collaborated with $\mathcal{F}$ from computing departments only; and (iii) cross-pollinators $\mathcal{O}(\mathcal{P}j) \equiv XD_{\mathcal{P}}$, if they collaborated with $\mathcal{F}$ from both biology and computing departments. Those pollinators who appeared in just a single scholar profile are named "leaf pollinators;" they are not central to our analytic framework.

HGP and evolution of cross-disciplinarity in the biology-computing college

Figure 2A shows the evolution of the largest connected component of the biology-computing collaboration network from the pre-HGP era (around 1990) to the post-HGP era (beyond 2003), where nodes correspond to $\mathcal{F}_i$ and links represent only the direct collaborations. We sized the nodes according to their relative importance within the network, given by the centrality $\mathscr{C}_i(t)$ calculated up to time *t*. We calculated three different centrality measures: degree, PageRank, and betweenness. The degree centrality $\mathscr{C}_i^D(t)$ counts the number of faculty $\mathcal{F}$ connected to a given faculty $\mathcal{F}_i$. The PageRank centrality $\mathscr{C}_i^{PR}(t)$ self-consistently incorporates the centrality of the neighboring $\mathcal{F}$ into the centrality of $\mathcal{F}_i$ (*28*, *29*). The betweenness centrality $\mathscr{C}_i^B(t)$ counts the number of shortest paths between other nodes that intersect $\mathcal{F}_i$ and is an indicator of between-group brokerage (*30*). Although these three centrality variables quantify different properties of the nodes within the network, we found them to be correlated with each other: $r(\mathscr{C}^D, \mathscr{C}^{PR}) = 0.97$, $r(\mathscr{C}^D, \mathscr{C}^B) = 0.76$, and $r(\mathscr{C}^B, \mathscr{C}^{PR}) = 0.80$. For visual comparison, we illustrate these three measures simultaneously in fig. S3, which identifies Eric Lander, one of the leaders of the HGP, as the most prominent faculty according to all three measures.

In Fig. 2A, we chose to size the nodes according to the degree measure $\mathscr{C}_i^D(t)$, which is an intuitive count variable that facilitates comparisons across the different networks. We colored the nodes green if $\mathcal{F}_i$ belonged to the $BIO_{\mathcal{F}}$ group, magenta if they belonged to the $CS_{\mathcal{F}}$ group, and black if they belonged to the cross-disciplinary $XD_{\mathcal{F}}$ group. To illustrate the evolution of cross-disciplinarity within the $\mathcal{F}$ network, we initially classify (color) each faculty node according to her/his primary departmental affiliation and only change this classification (color) to $XD_{\mathcal{F}}$ once the year of her/his first direct cross-disciplinary collaboration is reached. As time passes by, the giant component of the $\mathcal{F}$ network experiences impressive growth in size and complexity; within it, the cross-disciplinary nodes grow in numbers and prominence. Part of the giant component's growth appears to be fueled by the increasing assimilation of formerly less connected scholars, as the diminishing set of nongiant components shown in fig. S4 suggests.

While Fig. 2A depicts the emergence and centrality of cross-disciplinary scholars in the network during and after the HGP, Fig. 2B quantifies this evolution. We determined the overall fraction of collaborations that are within- or cross-disciplinary, from the perspective of both the direct ($\mathcal{F}-\mathcal{F}$) and the mediated ($\mathcal{F}-\mathcal{P}-\mathcal{F}$) links. More specifically, we first disaggregated the publication data by nonoverlapping 2-year periods. Then, for each period, we tallied the number of direct $\mathcal{F}_i - \mathcal{F}_{i'}$ links in a given period that were within-discipline $L_{\mathcal{F},W}(t)$ or cross-discipline $L_{\mathcal{F},XD}(t)$, with $L_{\mathcal{F}}(t) = L_{\mathcal{F},W}(t) + L_{\mathcal{F},XD}(t)$ denoting the total number of direct links. Similarly, we constructed the total number of mediated links realized via pollinator connections: $L_{\mathcal{P}}(t) = L_{\mathcal{P},W}(t) + L_{\mathcal{P},XD}(t)$.

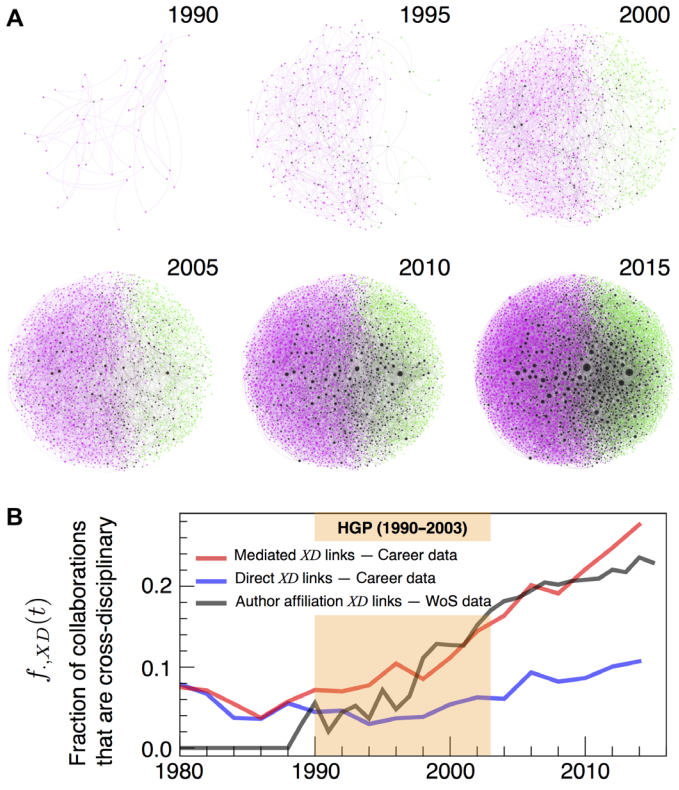


Fig. 2. Growth of cross-disciplinary social capital. (**A**) Evolution of the giant component in the U.S. biology-computing network. Green and magenta nodes represent faculty $\mathcal{F}_i$ with $BIO_{\mathcal{F}}$ and $CS_{\mathcal{F}}$ affiliation, respectively; black nodes represent faculty $\mathcal{F}_i$ that, by time *t*, published at least one cross-disciplinary publication and joined the $XD_{\mathcal{F}}$ group; node size is proportional to the logarithm of the degree centrality, $\ln \mathscr{C}_i^D$, of $\mathcal{F}_i$ at time *t*. (**B**) Evolution of the fraction of collaboration links in the $\mathcal{F}$ network that are cross-disciplinary. We calculated $f_{\cdot,XD}(t)$ using direct links $\mathcal{F}-\mathcal{F}$ between faculty (blue line) [that is, $f_{\mathcal{F},XD}(t)$] or association links $\mathcal{F}-\mathcal{P}-\mathcal{F}$ mediated by pollinators (red line) [that is, $f_{\mathcal{P},XD}(t)$]. For comparison, the black line shows the evolution of cross-disciplinary links in the human genomics literature per Web of Science (WoS); these values are divided by two to facilitate trend comparison. The orange area marks the HGP project period.

Next, we estimated the fraction of collaborations that are cross-disciplinary $f_{\cdot,XD}(t)$, with respect to two perspectives: $f_{\mathcal{F},XD}(t) = L_{\mathcal{F},XD}(t)/L_{\mathcal{F}}(t)$ using the direct collaboration links and $f_{\mathcal{P},XD}(t) = L_{\mathcal{P},XD}(t)/L_{\mathcal{P}}(t)$ using the mediated links (Fig. 2B). We complement these two estimates with a third estimation using a separate international data set of "Human Genome" publications, reporting the fraction of publications that include both *CS* and *BIO* author affiliations (see the "Assembly of the WoS data set" section).

The relative frequency of mediated cross-disciplinary associations shows marked growth during and in the wake of HGP, reaching ~30% of the total mediated associations by 2015. The relative frequency of direct cross-disciplinary collaboration shows slower growth. This feature may arise from the different competitive and leadership perspectives between the faculty $\mathcal{F}$ and the pollinators $\mathcal{P}$, leading to different capacities to explore cross-disciplinary projects. The difference between the mediated and direct $f_{\cdot,XD}(t)$ supports the importance of

mobility in the academic ecosystem as an underlying conduit for knowledge transfer, in addition to direct collaboration.

The impetus for this increasing rate of cross-disciplinarity is intriguing. Recent work identifies groundbreaking discoveries as one type of impetus leading to the densification and emergence of scholarly communities (*15*). However, in the case of the HGP (1990–2003), the evolution of the collaboration network was likely, to some degree, pulled ahead by the specification of a grand challenge that led to the organization of agents around a common agenda, not unlike the case of sustainability science (*18*). This alternative type of impetus is evident in the early growth of cross-disciplinarity *XD* (from the mid 1990s to early 2000s), when the HGP was in full swing, but the breakthrough of sequencing the human genome was not yet fully realized. As such, the co-occurrence of the start of the HGP and the increasing rate of cross-disciplinary activity provide preliminary evidence that incentivizing this activity around a unifying grand challenge was effective in bridging university disciplines. However, additional data and specifically tailored research design would be necessary to more conclusively estimate the magnitude of the HGP's impact on cross-disciplinary orientation in genomics research, which we leave for future work.

Career benefits of cross-disciplinarity

Figure 3 presents the descriptive statistics of the career data set. Figure 3A shows that the typical $\mathcal{F}_i$ career in all three faculty groups began in the early 1990s. This is ideal for studying the evolution of genomics, as HGP—arguably the field's constitutional project—was formally started in 1990. Figure 3B shows the significantly higher degree of collaboration in the $XD_{\mathcal{F}}$ group (370 ± 440) with respect to the $CS_{\mathcal{F}}$(122 ± 98) and $BIO_{\mathcal{F}}$(165 ± 175) groups.

Figure 3C shows that the $XD_{\mathcal{F}}$ group exhibits a significantly higher degree of cross-disciplinarity (0.3 ± 0.19) than the other two groups (0.1 ± 0.07 in both cases). The degree of cross-disciplinarity χ_i of $\mathcal{F}_i$ is defined as the fraction of her/his collaborators who are cross-disciplinary. Specifically, $\chi_i = k_{i,XD}/K_i \in [0, 1]$, where K_i is the total number of collaborators of $\mathcal{F}_i$, while $k_{i,XD}$ is the number of her/his cross-disciplinary collaborators; the collaborators include both other faculty $\mathcal{F}$ and pollinators $\mathcal{P}$ alike. We focus on one additional network characteristic, the scholar's PageRank centrality, which is measured relative to other members of the network. We use rescaled units, $N_{\mathcal{F}}\,\mathscr{C}_i^{PR}(t)$, so that the mean centrality value across all $\mathcal{F}_i$ is 1, which facilitates comparison. Figure 3D shows that the mean centrality of the

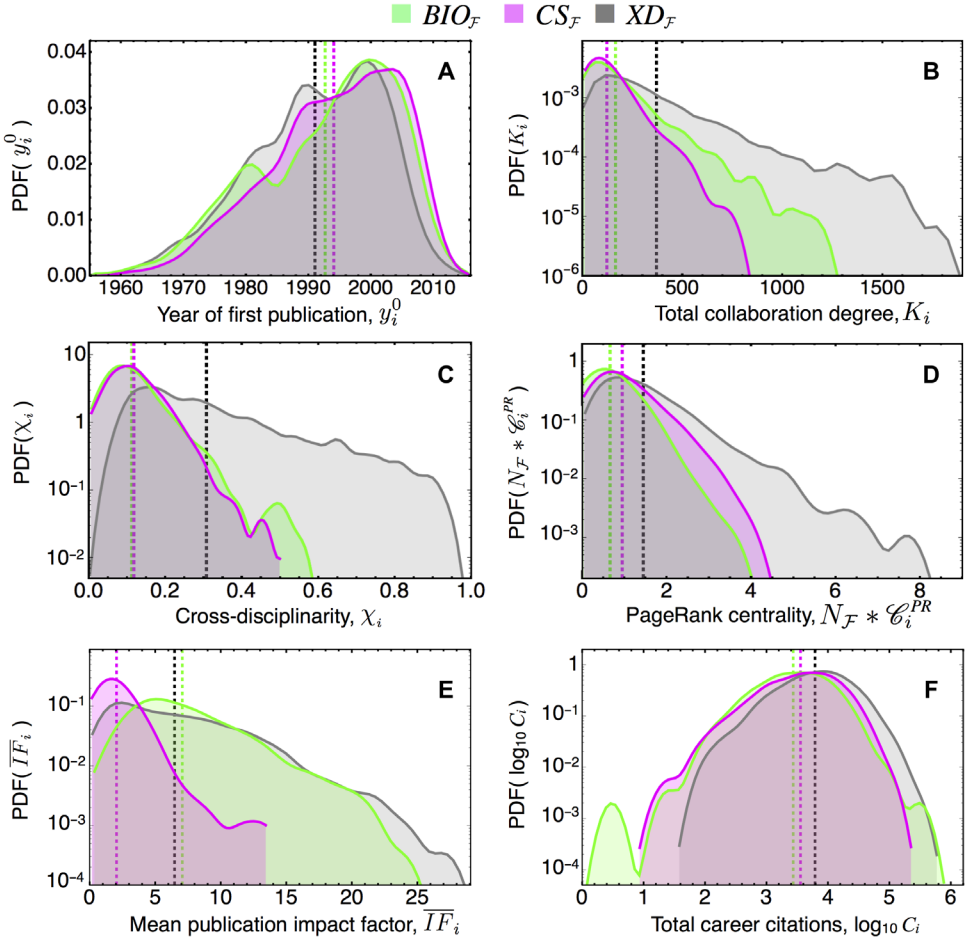


Fig. 3. Descriptive statistics for the career data set. Vertical lines indicate distribution means for the corresponding subsets. (**A**) Probability distribution of the year of first publication y_i^0 by $\mathcal{F}_i$. (**B**) Probability distribution of K_i, the total number of collaborators for a given $\mathcal{F}_i$. (**C**) Probability distribution of χ_i, the fraction of the collaborators of $\mathcal{F}_i$ who are cross-disciplinary. (**D**) Probability distribution of $\mathscr{C}_i^{PR}$, the PageRank centrality of $\mathcal{F}_i$; it is scaled by $N_{\mathcal{F}}$, the number of $\mathcal{F}_i$, so that the mean value of this scaled quantity across all $\mathcal{F}_i$, independent of the discipline subset, is 1. (**E**) Probability distribution of the mean impact factor ($\overline{IF}_i$) of the publication record of $\mathcal{F}_i$. (**F**) Probability distribution of the total citations $\log_{10} C_i$ of $\mathcal{F}_i$.

$XD_{\mathcal{F}}$ group (1.4 ± 1.1) is significantly higher than the mean centrality of the $BIO_{\mathcal{F}}$ (0.7 ± 0.5) and $CS_{\mathcal{F}}$(0.9 ± 0.6) groups.

Figure 3E indicates that $XD_{\mathcal{F}}$ faculty have similar publishing patterns with $BIO_{\mathcal{F}}$ faculty, that is, they tend to publish in high–impact factor (IF) journals. To be specific, we calculated the mean Journal Citations Report (JCR) impact factor among the publication set of each $\mathcal{F}_i$, denoted as $\overline{IF}_i$: The distribution of the $\overline{IF}_i$ among the $BIO_{\mathcal{F}}$ faculty is 7.1 ± 3.7; for the $CS_{\mathcal{F}}$ faculty, it is 2.0 ± 1.4; and for the $XD_{\mathcal{F}}$ faculty, it is 6.5 ± 4.5.

Given the relatively balanced composition of the cross-disciplinary group (n = 724 with biology pedigree versus n = 523 with computing pedigree), one would expect the $XD_{\mathcal{F}}$ mean to be more balanced in its distance from the $BIO_{\mathcal{F}}$ and $CS_{\mathcal{F}}$ mean values. Looking inside the $XD_{\mathcal{F}}$ group, we find that the cross-disciplinary subgroup with biology pedigree has $\overline{IF} = 8.58$, manifesting a small mean shift with respect to the core $BIO_{\mathcal{F}}$ faculty (+ 20.8%). The cross-disciplinary subgroup with computing pedigree has $\overline{IF} = 3.55$, manifesting a massive mean shift with respect to the core $CS_{\mathcal{F}}$ faculty (+ 77.5%). Hence, on the one hand, biology cross-disciplinary faculty maintain a publication culture that is on par with their disciplinary norms. On the other hand, computing cross-disciplinary faculty feature a publication culture that breaks away from their disciplinary norms and trends in the direction of biology. As a result, the overall mean of the $XD_{\mathcal{F}}$ group remains very close to the mean of the $BIO_{\mathcal{F}}$ group and far away from the mean of the $CS_{\mathcal{F}}$ group, revealing a degree of cultural assimilation.

Figure 3F shows the higher mean citation impact (in $\log_{10}$) in the $XD_{\mathcal{F}}$ group (3.8 ± 0.5) with respect to the $BIO_{\mathcal{F}}$ (3.4 ± 0.5) and $CS_{\mathcal{F}}$ (3.6 ± 0.6) groups. Because of the importance of total citation impact as a quantitative measure of career achievement (*31*), we begin by modeling C_i using cross-sectional analysis. Recent studies have demonstrated how collaboration factors can explain long-term success at the publication and career level (*25*, *32*, *33*). Consequently, here, we also account for the role of network attributes, reflecting the position of $\mathcal{F}_i$ in the collaboration network, in addition to controlling for standard *CV* attributes, such as her/his *h*-index, funding, and institutional prestige.

Our principal interest is to test whether $\mathcal{F}_i$ with stronger cross-disciplinary orientation (that is, higher χ_i) correlate with higher C_i. To this end, we used time-aggregated measures calculated through 2017 to estimate the parameters of the following cross-sectional ordinary least squares (OLS) regression model

$$\ln C_i = \beta_r \ln r_i + \beta_h \ln h_i + \beta_{\$1} \ln \$_i^{NSF} + \beta_{N1} \ln N_i^{NSF} + \beta_{\$2} \ln \$_i^{NIH} + \beta_{N2} \ln N_i^{NIH} + \beta_C \ln \mathscr{C}_i^{PR} + \beta_\chi \chi_i + D(O(\mathcal{F}_i)) + D(y_{i,5}^0) + \beta_o + \epsilon \quad (1)$$

where C_i is the total number of citations for $\mathcal{F}_i$, r_i is the ranking of her/his department, h_i is her/his *h*-index serving here as a productivity measure, and $N_i^{NSF}$and $N_i^{NIH}$are the total counts of her/his National Science Foundation (NSF) and National Institutes of Health (NIH) grants, while $\$_i^{NSF}$and $\$_i^{NIH}$are the total monies from the NSF and NIH grants deflated to constant 2010 USD, $\mathscr{C}_i^{PR}$ is her/his PageRank centrality within the $\mathcal{F}$ network, and χ_i is the fraction of her/his total K_i co authors who are cross-disciplinary. We include two dummy variables, the first capturing the three possible disciplinary orientations $\mathcal{O}(\mathcal{F}_i) = BIO_{\mathcal{F}}$ or $CS_{\mathcal{F}}$ or $XD_{\mathcal{F}}$, and the second capturing age cohort variation, where $y_{i,5}^0$ is the year of the faculty's first publication grouped into nonoverlapping 5-year intervals. Last, ϵ is white noise.

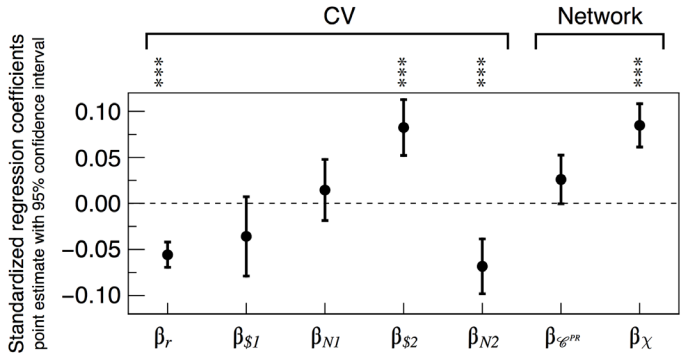


Fig. 4. Career cross-sectional regression model. OLS parameter estimates for the linear regression model in Eq. 1. The coefficients for the relevant covariates split into two categories are shown, depending on whether you might find the information in the researcher's CV or by analyzing her/his collaboration network. To facilitate comparison of the relative strength of the parameter estimates, the standardized beta coefficients are shown, representing the change in the dependent variable ln C_i that corresponds to a 1-SD shift in a given covariate. See table S2 for the complete list of parameter estimates. The levels of statistical significance are as follows: ***$P \leq 0.001$.

Table S2 shows the full-parameter estimates for the model expressed in Eq. 1, while Fig. 4 summarizes the relevant coefficient estimates for the funding and collaboration variables. The main result of this model shows that higher degrees of cross-disciplinary activity ($\beta_\chi > 0$, $P < 0.001$) correlate with higher career citations. To be specific, our estimates indicate that an increase in χ by 0.1 correlates to a $10 \times \beta_\chi$ = 5.7% increase in C_i.

We tested the robustness of this cross-sectional model by exploring several variations (table S3). In the first two variants, we replaced the PageRank $\mathscr{C}_i^{PR}$ centrality measure with one of two alternative centrality measures, that is, the betweenness centrality $\mathscr{C}_i^B$ and the degree centrality $\mathscr{C}_i^D$. In the third variant, we removed the variables N_i^{NSF} and N_i^{NIH} related to the number of grants, leaving only the variables $\$_i^{NSF}$ and $\$_i^{NIH}$ related to total funding, suspecting correlation effects. In the fourth variant, we removed the department rank variable r_i, because it is based only on the most recent university affiliation of a given $\mathcal{F}_i$ and thus could inaccurately represent her/his career. In all cases, the results of the modified regression estimates are not significantly different, indicating the robustness of our specification with respect to these adjustments.

The results of the cross-sectional model in Eq. 1, featuring the full set of funding variables (Fig. 4), point to a key career dilemma with respect to the pursuance of extramural grants. While our estimates confirm the benefit of total NIH funding ($\beta_{\$2} > 0, P < 0.001$), the correlation with the number of NIH grants is negative ($\beta_{N_2} < 0, P < 0.001$), pointing to the sunk costs associated with the management of several smaller grants (for example, R21) versus fewer bigger grants (for example, R01). Neither of the estimates for the NSF variables is significant, suggesting different levels of reliance on NIH/NSF between the biology and computing faculty.

Cross-disciplinary versus disciplinary production within careers

Motivated by the results of our pooled cross-sectional analysis, we implemented a panel regression model that leverages the longitudinal dimension of the career data disaggregated at the publication level. This enabled us to test whether the cross-disciplinary citation premium, indicated by $\beta_\chi > 0$ in the cross-sectional career model,

stems from the scholar's cross-disciplinary publications rather than other factors. In particular, we use a specification with individual $\mathcal{F}_i$ fixed-effects so that parameter estimates leverage the within-career comparison of publications that are cross-disciplinary with respect to those that are not. Hence, by identifying a clear counterfactual, this first panel model provides an estimate of the causal link between cross-disciplinary orientation and scientific impact.

To reduce false-positive (type I) classification errors, we do not use the disciplinary orientation of pollinators to classify individual publications. This is because the discipline of pollinators is not directly known and is based on inferences that may lead to overestimation. Consequently, for the classification of publications (hereafter denoted by p), we exclusively use the departmental affiliation of faculty, which are the only authors for whom we have ground-truth information. Within the profile of a faculty $\mathcal{F}_i$, for each p published in year t_p, we assign an indicator value $I_{i,p}^{XD} = 0$ if all of the faculty authors are from the same discipline or, conversely, $I_{i,p}^{XD} = 1$ if there is at least one faculty author from $CS_{\mathcal{F}}$ and at least one faculty author from $BIO_{\mathcal{F}}$. For example, a publication with three faculty authors classified as $\{CS_{\mathcal{F}}, CS_{\mathcal{F}}, \mathcal{F}_i\}$ with $\mathcal{F}_i = BIO_{\mathcal{F}}$ will have $I_{i,p}^{XD} = 1$; however, if instead $\mathcal{F}_i = CS_{\mathcal{F}}$, then $I_{i,p}^{XD} = 0$. Using this strict rule, out of the 413,565 publications (observations) in our faculty network sample, we classify with high confidence 4207 publications, or 1% of the entire sample, as cross-disciplinary.

Critical to the panel framework is the definition of a dependent variable measuring an article's long-term citation impact, one that is comparable across both different years t and disciplines s. This is a common difficulty in citation analysis and arises from a combination of three statistical biases: (i) varying citation rates across disciplines of different size, (ii) right censoring bias in the tallying of raw citation counts from a single census year (that is, the year in which citation data are downloaded from GS or another repository), and (iii) "citation inflation." The first bias reflects the fact that larger, more prolific disciplines produce more citations than smaller disciplines. The second bias refers to the fact that older publications have had more time to accrue citations than newer ones. The final bias arises from the change in the relative significance of a single citation over time, due to increasing publication rates and reference list lengths (*34*). By way of example, consider two publications, each cited 10 times in their first 10 years: If the first was published in 1980 and the second in 2007, in relative terms, then the first article has higher citation impact than the second.

The citation tallies reported by GS suffer from each of these problems. To neutralize these statistical biases, we applied a normalization formula that maps the GS citation count $c_{i,p,s,t}$—for an article p that was published in year t by a faculty $\mathcal{F}_i$ from discipline s—to a citation score $z_{i,p}$ that is comparable across s and t. A detailed description of the citation normalization formula is given in Materials and Methods.

Consequently, we formulate the following hierarchical panel regression model

$$z_{i,p} = \beta_i + \beta_\alpha \ln a_{i,p} + \beta_\tau \tau_{i,p} + \beta_I I_{i,p}^{XD} + D(t) + \epsilon_{i,p} \qquad (2)$$

The panel data encompass publications in the period 1970–2017 for the 3900 $\mathcal{F}_i$ that are connected within the $\mathcal{F}$ network, among which 1247 $\mathcal{F}_i$ are classified as $XD_{\mathcal{F}}$. This subset of cross-disciplinary publications is represented by the coefficient β_I, which provides an estimate for the impact of cross-disciplinarity at the publication level. By using author-specific fixed effects (β_i), which capture unobserved time-invariant researcher-specific characteristics, our model effectively compares the publications from the same $\mathcal{F}_i$ with $I_{i,p}^{XD} = 1$ using the counterfactual scenario $I_{i,p}^{XD} = 0$ as a baseline, after all other factors are held approximately constant. The other control variables in Eq. 2 include $a_{i,p}$ measuring the total number of coauthors listed on each publication p; the career age variable $\tau_{i,p} \equiv t_p - y_i^0 + 1$ referring to the number of years since the researcher's first publication, which controls for the career life cycle; the dummy year variable $D(t)$ controlling for year-specific shocks; and the residual white noise $\epsilon_{i,p}$.

The parameters in Eq. 2 are estimated using Huber-White robust SEs, which account for heteroscedasticity and serial correlation within the publication set of each $\mathcal{F}_i$. Table S4 shows the OLS estimates for models with and without $\mathcal{F}_i$ fixed effects. The sign and significance of the model variables are robust to the hierarchical specification, that is, with and without $\mathcal{F}_i$ fixed effects.

Figure 5A shows the model estimates for the three variables of principal interest. First, and most importantly, we estimate a statistically significant positive relationship between cross-disciplinarity and citation impact ($\beta_I = 0.145$, $P < 0.001$), meaning the average cross-disciplinary publication is more highly cited than the average disciplinary publication authored by the same $\mathcal{F}_i$. To translate the impact expressed by $z_{i,p}$ to the citation premium $c_{i,p}$, we calculate the percent change $100\,\Delta c_p/c_p$ when $I_{i,p}^{XD}$ goes from 0 to 1, which, due to the property of logarithms, is given by $100\,\Delta c_p/c_p = 100 \times \sigma_t \times (\partial z/\partial I^{XD}) \approx 100 \times 1.4 \times \beta_I = 20\%$ increase, which follows because the SD σ_t is approximately constant over time.

While our model specification does not focus on the effect of team size or author age on citation impact, the associated explanatory variables are also significant and worth discussing. Consistent with previous research, we observe a positive relationship between team size and citation impact ($\beta_a = 0.31$, $P < 0.001$) (*21*, *25*), which translates to a $\sigma_t \times \beta_a \approx 0.43\%$ increase in citations associated with a 1% increase in team size (as $a_{i,p}$ enters in ln in our specification), and finally, we observe a negative relationship with increasing career age ($\beta_\tau = -0.01$, $P < 0.001$), consistent with previous studies using different career data (*25*, *35*), which here translates to a $100 \times \sigma_t \times \beta_\tau \approx -1.3\%$ decrease in $c_{i,p}$ associated with every additional career year.

We tested the robustness of these results by introducing progressively stricter data selection criteria in two steps. First, we refined the faculty data set to include only the $\mathcal{F}_i$ with $\mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$, that is, we excluded from consideration the core $BIO_{\mathcal{F}}$ and $CS_{\mathcal{F}}$ faculty of the college (table S5). Second, within this $XD_{\mathcal{F}}$ subset, we became stricter as to what we considered fair comparison between their cross-disciplinary versus their other publications. Specifically, we used a matching scheme to pair each cross-disciplinary publication p ($I_{i,p}^{XD} = 1$) with a single disciplinary publication p' ($I_{i,p'}^{XD} = 0$) from the same faculty profile; the matched pair of publications (p, p') must also be within 2 years of one another and feature nearly identical number of coauthors (table S6). This matching procedure allows us to more accurately identify a counterfactual for each cross-disciplinary publication and thus to test the causal link between cross-disciplinarity and scientific impact (see Materials and Methods for further details on our matching procedure).

According to the Rubin causal model and potential outcomes framework (*36*) (see Materials and Methods), this matching procedure facilitates computing the average treatment effect in terms of cross-disciplinarity. Using the entire set of matched pairs (p, p'), we calculate the mean difference in normalized impact z corresponding to the mean treatment effect $\bar{T}_{XD}(z) = E[z_p - z_{p'} | \mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}] = 0.13$. An additional sign of robustness is that this estimate is consistent with the β_I estimates for the three panel scenarios we developed: (i) all faculty

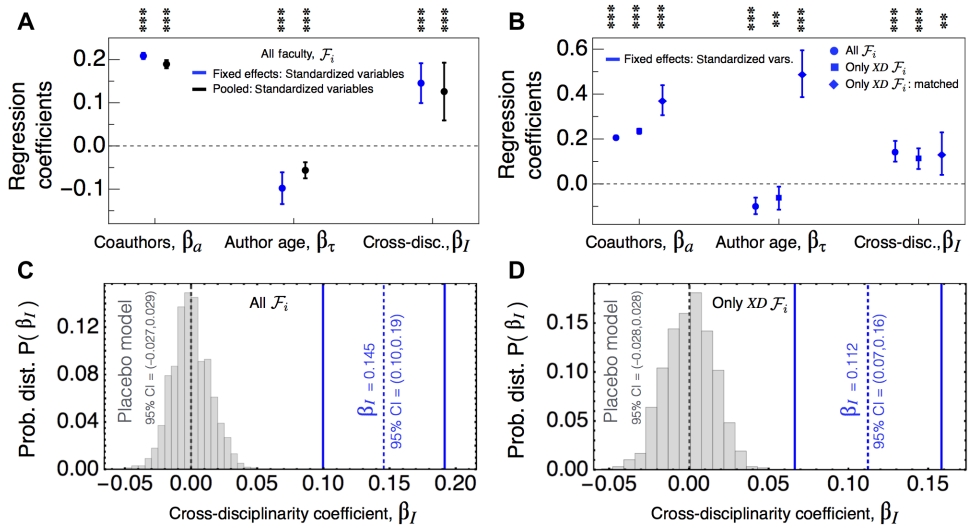


Fig. 5. Career panel regression model. (**A** and **B**) Parameter estimates for the three principal explanatory variables included in the fixed effects $\mathcal{F}$ career model defined in Eq. 2; see table S4 for the complete list of parameter estimates. (**C** and **D**) Robustness check of panel regression model. To test the possibility of spurious correlations leading to the significant estimates for the cross-disciplinary variables in the panel model (table S4), we ran this model using a randomized cross-disciplinary indicator variable $I^{XD}_{i,p}$, implemented by shuffling just that variable across the observations without replacement. (C) For n = 1000 shuffled data sets, we do not observe any (0%) coefficient estimates as large as the empirical value β_I = 0.145 corresponding to the dashed vertical blue line [solid vertical blue lines indicate 95% confidence interval (CI); see table S4, third column cluster]. (D) We repeated the same shuffling method for the panel model applied to only the 1247 $\mathcal{F}_i$ classified with orientation $\mathcal{F}(\mathcal{F}_i) \equiv XD_{\mathcal{F}}$, and again, we do not observe any (0%) coefficient estimates as large as the empirical value β_I reported in table S5 (third column cluster). The levels of statistical significance are as follows: $^{**}P \leq 0.01$, $^{***}P \leq 0.001$.

$\mathcal{F}$ (β_I = 0.145), (ii) cross-disciplinary faculty $\mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$ (β_I = 0.112), and (iii) cross-disciplinary faculty $\mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$ considering only their matched pairs of publications (β_I = 0.135). We can also use the matched pairs to estimate the average treatment effect in terms of percent change in citations, which we calculate to be, on average, a 10.6% increase in $c_p$over the counterfactual $c_{p'}$. Furthermore, by tallying the citation difference across all matched cross-disciplinary publications for each $\mathcal{F}_i$, we calculate the average treatment effect in terms of total net citations to be 630 citations over her/his career.

Figure 5B shows the fixed-effects model estimates for all three approaches: (i) using all faculty $\mathcal{F}_i$, (ii) using only cross-disciplinary faculty $\mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$, and (iii) using matched publication subsets for the cross-disciplinary faculty $\mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$. All parameter estimates are consistent across the three panel variants, with the exception of β_τ, which is positive for the specification (iii) and negative for specifications (i) and (ii); this inconsistency is due to the fact that the matched data in (iii) are a subset of the faculty's longitudinal profile, thus introducing bias in the subset selection with respect to career age.

We also explored the possibility that spurious correlations could give rise to the significance of $I^{XD}_{i,p}$ by using a placebo randomization scheme in which we shuffled the $I^{XD}_{i,p}$ values across the data set, without replacement, that is, conserving the total number of observations with $I^{XD}_{i,p}$= 1. We ran this placebo regression 1000 times, each time recording the value of β_I. Figure 5C shows the distribution of the placebo estimates, $P(\beta_I)$, when that data for all the faculty $\mathcal{F}_i$ are used; none (0%) of the placebo estimates were larger than the real estimate β_I = 0.145, thereby showing that it is unlikely that we obtained the magnitude and significance of β_I by chance alone. Figure 5D shows similar results for the distribution of the placebo estimates, $P(\beta_I)$, when the data for just the $XD_{\mathcal{F}}$ faculty are used.

By analyzing publications clustered within careers using fixed effects, we approach the problem differently than the bulk of recent relevant work on quantifying the correlation between interdisciplinarity and impact (*37*–*40*). There are relatively few studies we are aware of that use fixed-effects to net out unit-level variation (*41*, *42*), and none that uses the Rubin potential outcomes framework. Moreover, most relevant studies use as a proxy for interdisciplinarity the diversity of distinct journals or the diversity of distinct research areas cited within an article's reference list. While this is a reasonable approach, all by itself, it would not serve us well in the case of a team science field such as genomics, where we are preoccupied not only with mixed knowledge but also with behaviors in mixed teams.

The genomics story behind the numbers in the biology-computing college

Our study of the biology-computing college in the United States revealed significant cross-disciplinary activity in genomics during and after the HGP, with net career benefits for those involved, and a cultural shift for the cross-disciplinary faculty with computing pedigree. Motivated by these quantitative results, we further explored the collaboration-mediated pathways that trace knowledge transfer from the HGP to the present day. In addition to explicit knowledge pertaining to computing algorithms and biotechnology methods, this knowledge transfer also includes tacit organizational know-how that is fundamental to the management of consortium science—a paradigm in which teams of teams coordinate a common agenda around a single "grand challenge."

This emergent pattern began with the 2001 publication of the two seminal human genome papers in *Nature* (*43*) and *Science* (*44*). These parallel efforts, one public and one private, offer valuable insights into the economics of science (*8*). Yet, more germane to our focus on social and organizational capital formation in science is the intriguing

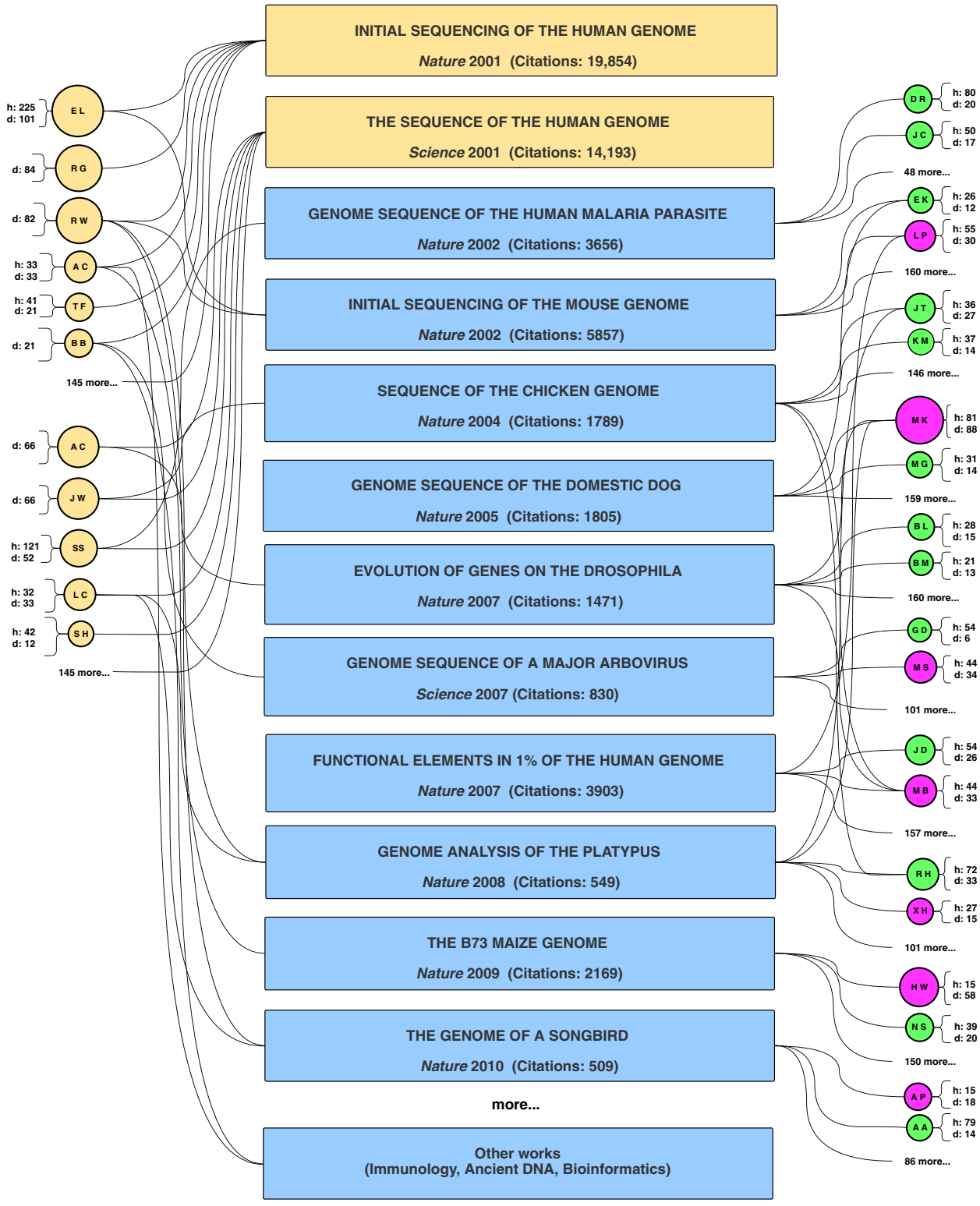


Fig. 6. The knowledge transfer story behind the numbers. Interactions of the HGP scholars with other faculty in the $\mathcal{F}$ network during the 2000s, and some of the landmark publications they produced, powering the genomics revolution. The scholar nodes bear the name initials. On the left panels, one can recognize some well-known HGP scholars, such as Eric Lander (EL) and Bruce Birren (BB). "d" stands for the network degree of a scholar and controls with the size of her/his node. "*h*" stands for the *h*-index of a scholar. Magenta nodes denote faculty affiliated with computing departments, while green nodes denote faculty affiliated with biology departments.

hereditary pattern of the consortium science model, as illustrated in Fig. 6. Namely, several of the authors in these two papers played a quintessential role in knowledge transfer and the evolution of genomics in the 2000s. Every year, starting from the culmination of the HGP circa 2000 and all the way to 2010 and beyond, subsets of the original HGP authorship seeded efforts to decode the genome of important animals, plants, and microorganisms. The faculty network $\mathcal{F}$ and its pollinating ecosystem $\mathcal{P}$ capture key aspects of this blooming period in genomics. In the left panel of Fig. 6, the faculty nodes and their "hidden" pollinator nodes that were members of the original HGP teams appear. In the middle panel of Fig. 6, some landmark genomics publications that were authored by these scholars appear, including the mouse genome (*45*), the chicken genome (*46*), and the dog genome (*47*). In the right panel of Fig. 6, the faculty nodes that were not in the original HGP teams but contributed in these subsequent genomics efforts appear, thus establishing coauthorship links with the original "HGP cohort" present in the network. Given that the authorship in all these papers was mixed, including authors from both biology and computing, all the coauthorship links presented in the figure are cross-disciplinary. The citations of the original HGP papers as well as the genomics papers that followed in their steps are impressive and testify to their impact. For HGP cohort members of the $\mathcal{F}$ network, the centrality and *h*-index attest to their apostolic role and status, respectively. The centrality and *h*-index of the non-HGP faculty who interacted with the HGP cohort suggest that these "HGP offspring" followed on the steps of their scholarly fathers/mothers, developing their own notable standing.

The sequencing of the human genome is an exemplary case of science for the public good, wherein the culminating achievement extended far beyond the organizational boundaries of the individuals and institutions centrally involved. Yet, in addition to far-reaching public health impacts, the development of the consortium approach cannot be understated, as it has subsequently served as the organizational model for the sequencing of other important genomes. Numerous prominent genomics papers in the 2000s are the products of consortia (for example, the "Mouse Genome Consortium" and the "Chicken Genome Consortium"), that is, analogs of the "Human Genome Consortium" first established in the HGP.

Finally, the evolutionary pathway depicted in Fig. 6 is not the only way genomics knowledge evolved through scholarly interactions in the $\mathcal{F}+\mathcal{P}$ network. For example, HGP scholars interacted with other faculty on genomic applications to immunology, cancer, and the decoding of ancient DNA, transforming medicine and evolutionary biology. These interactions, indicative of the broad impact of the HGP and the ever-expanding reach of genomics, are also captured in our data set and are lumped in the "Other works" box in Fig. 6.

Cross-disciplinarity as mixed authorship in the genomics literature

Analysis of the career data set above indicates that researchers in the U.S. biology-computing college achieve higher citation impact—both across and within faculty profiles—when they adopt a cross-disciplinary approach. As a consistency and robustness check, we further tested whether cross-disciplinarity, defined in this section as mixed bio-computing authorship in genomic publications, has value that transcends the U.S. biology-computing college.

We proceeded by collecting a comprehensive international data set consisting of 25,466 articles from the WoS using the topic query "Human Genome." We classified each article as cross-disciplinary (XD_g) if its affiliation list included both biology and computing departments, or biology (BIO_g) if its affiliation list included only biological sciences departments. The subscript *g* indicates that the *XD* and *BIO* attributes are linked to departmental affiliations of authors globally (not just U.S.-based), who have published in human genomics. As in our panel model, this operationalization establishes a clear counterfactual, that is, an article is either XD_g or BIO_g, reflecting researcher-level orientations. Consequently, the citation difference between the two publication subsets is associated with cross-disciplinary factors, net of other likely factors, such as funding levels or field size.

We calculated the mean citation impact $\bar{c}_{XD_g}(t)$ and $\bar{c}_{BIO_g}(t)$ for the nonoverlapping subsets of cross-disciplinary and biology publications, respectively (see Materials and Methods). The ratio

$$r_c(t) \equiv \bar{c}_{XD_g}(t)/\bar{c}_{BIO_g}(t) \quad (3)$$

measures the cross-disciplinary citation premium relative to the baseline established by the intradisciplinary biology publications. The value $\bar{r}_c = 1$ corresponds to the case in which there is no difference in citation impact between the two publication subsets.

Figure 7A shows the evolution of the citation premium $r_c(t)$ associated with cross-departmental collaboration in the international human genomics literature. We estimated the degree to which $r_c(t)$ could arise by chance using a random bootstrap sampling method to calculate the distribution of the randomized (null model) test statistic $r_{c,RND}(t)$ and thus to assess the likelihood of type I (false-positive) misestimation. To be specific, for a given year *t*, we randomly selected $N_{XD_g}(t)$ publications, independent of their departmental affiliations, and then calculated $\bar{c}_{XD_g,RND}(t)$ and $\bar{c}_{BIO_g,RND}(t)$ for this subset. We combined these two values to obtain a null model estimate $r_{c,RND}(t) \equiv \bar{c}_{XD_g,RND}(t)/\bar{c}_{BIO_gRND}(t)$. We repeated this randomization 10^6 times for each year and calculated the two-tailed 90, 95, and 99% thresholds for each distribution of $r_{c,RND}(t)$. It is worth noting that this null model, which is based on random sampling without replacement from the underlying citation distribution, conserves the overall proportion of publications with $N_{XD_g}(t)$ and also the total citations received by these publications. Thus, by sampling the empirical citation distribution, this randomization scheme demonstrates the range of r_c values one could obtain by chance.

Our results indicate significant citation premium $r_c(t)$ stemming from mixed authorship. Specifically, since 1999, the annual $r_c(t)$ values are significantly in excess of unity at the $P = 0.01$ level (false-positive rate; see Fig. 7A), with the mean $r_c(t)$ value standing at $\bar{r}_c = 1.1$. Because $r_c(t)$ is calculated using the logarithm of citation values, to temper the impact of outliers, we must convert this ratio to fully appreciate the magnitude of the effect. We can estimate the percent difference in raw citations drawing on the properties of the log-normal citation distribution. By assuming that cross-disciplinarity only affects the logarithmic mean of the citation distribution, one can estimate the percent difference in *c* for the XD_g group as compared to the BIO_g group as $\Delta c(\%) = 100(\exp[(r_c - 1)\overline{\ln(1+c)}] - 1)$. In these terms, the XD_g publications gained, on average, 37% more citations than those in the BIO_g group.

Cross-disciplinarity as mixed methods in a genomics journal

To test the value of cross-disciplinarity at the epistemic level of explicit and tacit knowledge, we constructed a third data set by collecting the 3516 research articles published over the period 1996–2014 in *Nature Biotechnology* (*NB*), a prestigious genomics-oriented journal. As in the

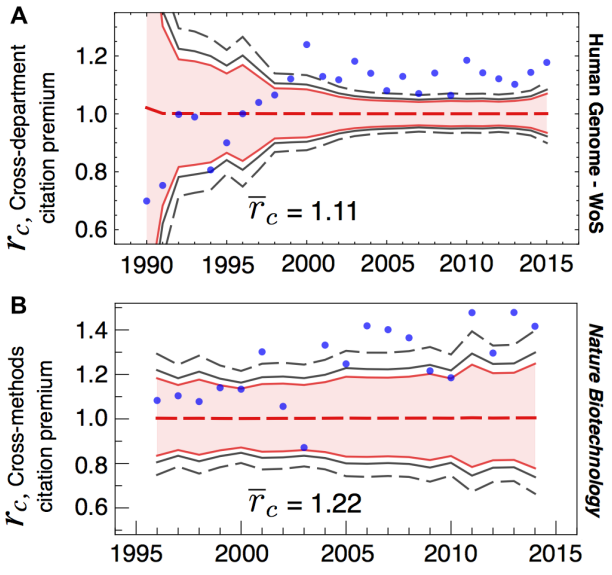


Fig. 7. Cross-disciplinarity beyond the faculty network $\mathcal{F}$. (**A**) Cross-disciplinarity XD_g as mixed authorship in the human genomics literature**:** Cross-disciplinarity is measured using the combinations of departmental affiliations on the set of Human Genome publications reported in the WoS. The mean value, weighted according to the publication volume each year, is $\bar{r}_c = 1.11$. (**B**) Cross-disciplinarity XD_e as mixed methods in *NB*: Cross-disciplinarity is measured by analyzing the combinations of computational and biological methods used within articles from the journal *NB*. The mean value, weighted according to the publication volume each year, is $\bar{r}_c = 1.22$. In both panels, blue dots represent the respective $r_c(t)$, calculated using real data to measure the additional citation impact of *XD* publications. The curves correspond to the respective null model test statistic distribution $P(r_{c,RND}(t))$, estimated from 1 million bootstrap randomizations, in which the expected value $r_{c,RND}(t) \equiv 1$ (that is, no difference between the mean citation impact of the subsets). The red curve and shaded region correspond to the 90% confidence interval for the respective randomized $r_{c,RND}(t) \equiv 1$, and the outer black curves correspond to the 95% (solid) and 99% (dashed) confidence intervals. Thus, empirical data above (or below) the null model confidence intervals are significantly different than the expected value $r_c = 1$ at the given significance level and demonstrate that it is highly unlikely to obtain these large values by chance alone.

previous analysis, it is important to verify that the cross-disciplinary citation premium persists among research articles of similar perceived novelty, that is, colocated in the same high–impact factor journal but differing to the extent to which they incorporate computational methods.

In this case, we assigned articles that featured computational methods to the cross-disciplinary group XD_e, as specified by the paper's Medical Subject Headings (MeSH), which are a controlled thesaurus of keywords implemented by PubMed (*48*). The remaining articles were assigned to the BIO_e group. Thus, in this *NB* analysis, the classification of articles as cross-disciplinarity is based on only epistemic and not authorship criteria (thus, the subscript *e* for *XD* and *BIO*). Nevertheless, statistical comparison of the two sets of research articles, corresponding to cross-disciplinary (XD_e) and biology (BIO_e), followed exactly the same method as in the case of the WoS human genomics data set.

Our results indicate significant citation premium $r_c(t)$ stemming from mixed research methods. Specifically, the annual $r_c(t)$ values are significant at the $P = 0.05$ level since 2004 (see Fig. 7B), with the mean $r_c(t)$ value standing at $\bar{r}_c = 1.22$. Translating this ratio, we find that XD_e publications gained, on average, roughly 126% more citations than those in the BIO_e group. Because we only compare publications within *NB*, the difference in the citation impact is net of journal-specific factors and represents the added value of computational knowledge and methods in research with genomic applications.

DISCUSSION

The merging trend among techno-scientific disciplines is bound to continue because of the nature of grand challenges faced by society. We know, however, little about what works and why in a cross-disciplinary fusion process. To start unlocking this problem in the context of team science (*21*), it is imperative to analyze not only scholarly knowledge production but also scholarly interactions that support scientific progress. To this end, the science of team science has contributed greatly to understanding how to accelerate scientific advancement via multiscale collaboration (*49*).

Less is known, however, about the factors that promote cross-disciplinary collaboration around a central challenge. Even in the case where the goals are well posed, agreeing on the best path forward can become contentious, especially when groups have different social and epistemological backgrounds. Consequently, harnessing the benefits of team science is often not just a matter of bringing primed stakeholders together. Recent work highlights a case of high-risk "gain-of-function" pathogen research, where the differing expertise of biomedical researchers affected their position around this politically charged dilemma (*50*). The issue of consensus formation acquires new urgency with the proliferation of social media, which sometimes undermine constructive dialogue between groups.

In general, whether between experts or nonexperts, there is a need to understand how to foster cultural bridging around scientific topics and narratives (*51*). It is within this overarching framework that we pursued mixed analysis for the field of genomics and two of its key constituent disciplines—biology and computing. First, we investigated cross-disciplinary versus disciplinary careers within the biology-computing college in U.S. universities. Strikingly, nearly all cross-disciplinary faculty in this college (~90%) have published research on genomics, and we show that the precipitation of this activity correlates with the onset of the HGP in 1990. Furthermore, we find that scholars with greater orientation toward cross-disciplinary collaboration tend to have higher career citation impact. We use several identification strategies, using publication-level data, to attribute this citation premium to the scholars' cross-disciplinary activity—net of other factors.

Germane to this discussion is the fact that cross-disciplinary computing scholars exhibit publication patterns that trend in the direction of biologists, with profiles that include papers in high-impact science journals. This is a sign of cultural assimilation, which gives cross-disciplinarity at the career level a fuller meaning.

Cross-disciplinarity, defined as joint authorship by biology and computing scholars, enjoys a premium that transcends the U.S. biology-computing college, being a feature of the international intellectual production in human genomics, as tracked by the WoS. Looking also at cross-disciplinarity as an epistemic fusion in the articles of a well-known genomics-oriented journal, we found that papers with explicit computational content enjoy a significant impact premium over papers without such content. These results are in agreement with recent work documenting the citation advantage that occurs when researchers innovate to form new within-discipline knowledge bridges (*52*) via measured combinations of novel and traditional concepts (*38*).

The latter represents a strategy that is generalizable to bridge-building in other domains, such as stimulating proactive public discourse (*51*).

One wonders about the reason behind the higher impact of cross-disciplinary publications in genomics. After all, this is what feeds the career advantage of the $XD_{\mathcal{F}}$ cohort and likely acts as a talent attractor, although other factors reportedly play a role in the decision to pursue cross-disciplinary collaboration (*53*).

Fast-paced and application-oriented techno-scientific disciplines, such as genomics, tend to be highly utilitarian. On the basis of this assumption, we can speculate for the moment that cross-disciplinary genomics publications are popular primarily because they are useful. One, however, should not underestimate the significant coordination cost associated with bridging disciplinary gaps within mixed teams (*54*). The fact that this coordination is done successfully in the biology-computing college suggests some compatibility between the two disciplines, with the assimilation of $XD_{\mathcal{F}}$ scholars of computing pedigree into biology's culture being a sign of it.

It is also generally true that mixed teams overcoming disciplinary communication barriers produce publications that are well posed, well framed, and well written, accessible to the union of the corresponding communities; all these likely contribute to higher citation rates (*55*).

The statistical results reported here can serve as an excellent springboard for science studies into the particular processes, artifacts, and personalities that powered genomics as well as the consortium science organizational model. In this direction, looking behind the numbers, we traced the collaborative pathways captured in our data model, bringing to the fore a key mechanism of the genomics revolution. Following the HGP, the data model points to several research efforts staggered over a decade, which led to the sequencing of important animals, plants, and organisms. The outcomes of these efforts were impactful publications in iconic journals, such as *Nature* and *Science*. The investigative teams included new coming faculty from both biology and computing, mixed with key members of the original HGP team in various configurations. Therefore, these projects shaped a new generation of cross-disciplinary researchers and helped them build their networks, their careers, and, along the way, genomics as we know it today. The work and authorship in the relevant genomic papers were structured around consortia, in the image of the HGP—a practice that ushered team science into the teams-of-teams science era.

In conclusion, as funding agencies are increasingly supporting cross-disciplinary investigations [for example, BRAIN initiative (*56*)] and associated scientific activity is on the rise (*42*, *57*), there is a growing need for insightful quantitative evidence from past cases to aid policy making (*8*, *10*). To this end, our findings show how a timely research initiative helped create cross-disciplinary human capital between two culturally complementary disciplines, and how inherent career incentives perpetuated this capital and contributed to its epistemic dominance. In modeling terms, science policy makers could view this as no different from the elements needed for a flame: a spark, a combustible medium, and a feeding system.

MATERIALS AND METHODS

The assembly of the career data set

We selected 155 biology and computing departments in the United States following the 2014 U.S. News & World Report (table S1). We confirmed that all the departments in the set had active PhD programs since the conception of HGP in the 1980s. Moreover, the ranking of academic departments is relatively rank-stable, as supported by theoretical and empirical evidence drawn from various other social systems characterized by positive feedback reinforcement mechanisms that temper large rank fluctuations (*58*). With respect to the latter, we found no significant differences in the ranks of these 155 departments between the 2014 and 2018 U.S. News & World Report ranking ($P > 0.05$, Wilcoxon test).

We accessed the home pages of these departments and recorded the listed faculty as of spring 2017. In this master list, we identified the faculty $\mathcal{F}_i$ that had GS pages, forming a database with their GS IDs, *h*-indices, departments, department rankings, and bibliometric data. We also indexed their NSF and NIH grant data from the corresponding repositories (*59*). We then applied a name disambiguation algorithm to $\mathcal{F}_i$ and their coauthors to reconcile their identities within and across $\mathcal{F}_i$ profiles (appendix S1). Figure 1 provides a visual example of how we constructed the biology-computing college network from the disambiguated $\mathcal{F}_i$ data.

The key motivator behind our data collection methodology for the career data set is the tendency of typical computing researchers to publish the bulk of their work in refereed conferences from where they receive most of their citations. Traditional bibliometric databases, such as Scopus and WoS, do not cover citations from many refereed conference publications, but GS does, thus emerging as the only viable alternative for fair career assessment.

Although the career data set covers a substantial portion of the biology-computing college in United States, it does not cover it all, and it does not explicitly cover the international biology-computing college. This limitation is tempered by two factors. First, it is important to clarify that, in our analysis, we are not seeking to measure the impact of the HGP on research outcomes, but rather the impact of cross-disciplinarity on research outcomes. Because the HGP had explicit cross-disciplinary alignment, we expect it to have had its strongest and most direct impact on the adoption of cross-disciplinary research orientation in the United States. Second, the construction of the mediated association network considerably expands the reach of the career data set, as it includes not only the faculty members in these 155 departments but all their collaborators, forming an impressive ecosystem. The representational power and validity of this ecosystem finds confirmatory evidence in two cases during the course of our analysis: (i) The evolution of the rate of cross-disciplinary collaborations in the U.S. biology-computing college mirrors the rate of cross-disciplinary collaborations gleaned via author affiliations in the human genomics literature at large. (ii) Entrance of faculty in the U.S. biology college crests in early 2000s, which is consistent with the doubling of NIH research funding in the period 1998–2003 (*8*, *60*).

Citation normalization

The citation normalization of publication p from faculty $\mathcal{F}_i$ leverages the universal log-normal properties of citation statistics (*61*), yielding a stationary, normally distributed citation measure $z_{i,p} \in N(0, 1)$ (fig. S5) that is well suited for identifying longitudinal patterns of citation impact in research careers (*25*, *35*).

To be specific, we disaggregated the articles by publication year and removed the time-dependent trend in the location and scale of the underlying log-normal citation distribution by defining

$$z_{i,p} \equiv [\ln(1 + c_{i,p,s,t}) - \mu_t]/\sigma_t, \quad (4)$$

where $\mu_t \equiv \overline{ln(1 + c_{s,t})}$ is the mean and $\sigma_t \equiv \sigma[\ln(1 + c_{s,t})]$ is the SD of the citation distribution, after adding 1 to each citation tally (to avoid the

divergence of ln 0 associated with uncited publications) and applying the natural logarithm. We calculated μ_t and σ_t within the subset of publications for a given year t and discipline s (*BIO* or *CS*). The SD $\sigma_t \approx 1.4$ is approximately constant across time and the two disciplines we analyzed.

Publication matching

We used the Rubin causal model framework (*36*) to provide additional evidence for a causal link between cross-disciplinary collaboration and increasing citation impact. According to the potential outcome notation, let $Y_{XD=1} = z_{i,p,1}$ represent the outcome, that is, scientific impact proxied by citations, of a publication drawing on cross-disciplinary collaboration, denoted in our data set by the indicator $I_{i,p}^{XD} = 1$; conversely, the counterfactual $\hat{Y}_{XD=0} = z_{i,p,0}$ represents the potential outcome of the same publication but without cross-disciplinary collaboration ($I_{i,p}^{XD} = 0$). To obtain counterfactual pairs from our data set $(Y_{XD=1}, \hat{Y}_{XD=0})$, for each *XD* publication p of each faculty $\mathcal{F}_i$ with $\mathcal{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$, we searched through just their profile for the most similar p' to pair with p. More specifically, for each p with $I_{i,p}^{XD} = 1$, we collected all publications from the same profile within ± 2 years ($|t_p - t_{p'}| \leq 2$). From this potential match set, we then selected the p' with the closest number of coauthors to a_p, and if $a_{p'}$ was larger or smaller than a_p by more than 20% ($|a_p - a_{p'}| \geq 0.2$), then we rejected this match and did not include p in the set of matched pairs. We produced matches without replacement so that each p' was included only once.

We then combined these matched pairs (p, p') into an observation subset and ran the same regression model as in Eq. 2 on this set of faculty with $N_{i,XD} \geq 10$ matched data pairs. Table S6 shows the model estimates for the resulting 53 $\mathcal{F}_i$. Using these matched publication pairs, we also estimated the mean cross-disciplinary "treatment effect," $T_{XD,i} = E[Y_1 - Y_0 | \mathcal{O}(\mathcal{F}_i) = XD] \approx N_{i,XD}^{-1} \Sigma_p (Y_{p,XD=1} - \hat{Y}_{p,XD=0})$. As such, the average value, $\bar{T}_{XD}$, is an estimate of the average treatment effect on the treated (ATET). In addition to comparing the outcome according to normalized citation impact, $Y_{XD=1} = z_{i,p,1}$, we also report the ATET calculated using the total citation difference, $Y_1 - Y_o \equiv \Sigma_p (c_{i,p} - c_{i,p'})$, and the percent citation difference, $Y_1 - Y_o \equiv 100(c_{i,p} - c_{i,p'})/c_{i,p}$.

Assembly of the WoS data set

We used the topic keyword Human Genome to query the WoS database. After excluding books and editorials, we arrived at a set of 25,466 publications, recording the total number of citations $c_{p,t}$ each publication p received through November 2016. We then defined cross-disciplinarity according to the diversity of departmental affiliations associated with each publication. Publications featuring researchers from both computing and biology departments were classified as XD_g, whereas publications featuring researchers from biology departments only were tagged as BIO_g.

Assembly of the NB data set

We downloaded from the WoS all publication records for articles published in the journal *NB* as of December 2015, resulting in a data set of 3516 items. We then used the MeSH of MEDLINE/PubMed, a unified and controlled vocabulary system of article keywords, to separate publications into the complementary XD_e and BIO_e subsets. The typical biomedical publication has roughly 10 to 20 MeSH descriptors assigned by professional MEDLINE experts and algorithms, which can then be used to position publications in a complex conceptual space composed of 16 top-level categories and more than 27,800 MeSH descriptors (*48*).

We leveraged this detailed ontology by tagging publications with at least one MeSH keyword from the "Information Science" category—the *L* branch—as XD_e articles. Three examples of MeSH keywords from the *L* branch are "Human Genome Project" (tree number L01.453.450), "Molecular Sequence Data" (tree number L01.453.245.667), and "Algorithms" (tree number L01.224.050). Seventy-one percent of the *NB* publications do not contain a single MeSH descriptor keyword belonging to the *L* branch; we tagged these as BIO_e articles. Thus, there are a significant number of publications with and without explicit computational methods, and we used the latter set as our baseline for comparison.

Calculating r_c

In the analysis of both the human genomics and *NB* articles, we calculated the mean citations per year for the $XD_{\cdot}$ ($\cdot \equiv g$ or e) subset, $\bar{c}_{XD_{\cdot}}(t) = [N_{XD_{\cdot}}(t)]^{-1} \Sigma_p \ln(1 + c_{p,t})$, where $N_{XD_{\cdot}}(t)$ is the total number of articles published in year t within the *XD* group. Similarly, we calculated the mean citations per year for the *BIO* ($\cdot \equiv g$ or e) subset, $\bar{c}_{BIO_{\cdot}}(t) = [N_{BIO_{\cdot}}(t)]^{-1} \Sigma_p \ln(1 + c_{p,t})$. We applied the logarithmic transformation to normalize the citation distribution within each year t. Assuming that citations follow a log-normal distribution and that the only difference between the *XD* and *BIO* groups is a multiplicative factor r_c affecting their logarithmic mean μ_{LN}, then the mean citations for the *XD* group is $\bar{c}_{XD} = \exp[r_c \mu_{LN} + \sigma_{LN}^2/2]$, and for the *BIO* group is $\bar{c}_{BIO} = \exp[\mu_{LN} + \sigma_{LN}^2/2]$, where μ_{LN} and σ_{LN} are the location and scale parameters of the underlying log-normal distribution. Thus, the percent difference between the mean citations is $\Delta c(\%) = 100(\bar{c}_{XD_{\cdot}}/\bar{c}_{BIO_{\cdot}} - 1) = 100(\exp[(r_c - 1)\mu_{LN}] - 1)$.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/8/eaat4211/DC1

Appendix S1. Author name disambiguation.
Appendix S2. Connectivity of the $\mathcal{F}$ network.
Fig. S1. Robustness of the $\mathcal{F}$ network with respect to link removal.
Fig. S2. $\mathcal{F}$ network distributions for direct and mediated associations.
Fig. S3. Three perspectives on the centrality of $\mathcal{F}_i$ in the direct collaboration network.
Fig. S4. Evolution of the nongiant components in the $\mathcal{F}$ network.
Fig. S5. Distribution of normalized citation impact by departmental affiliation and time period.
Table S1. Set of 155 biology and computing departments in the United States.
Table S2. Career data set: Pooled cross-sectional model.
Table S3. Career data set: Pooled cross-sectional model—robustness check.
Table S4. Career data set: Panel model on all faculty $\mathcal{F}$.
Table S5. Career data set: Panel model on the $XD_{\mathcal{F}}$ faculty.
Table S6. Career data set: Panel model on the $XD_{\mathcal{F}}$ faculty with matched pairs.
References (*62*–*64*)

REFERENCES AND NOTES

1. H. Stevens, *Life Out of Sequence: A Data-Driven History of Bioinformatics* (University of Chicago Press, 2013).
2. J. B. Hagen, The origins of bioinformatics. *Nat. Rev. Genet.* **1**, 231–236 (2000).
3. N. M. Luscombe, D. Greenbaum, M. Gerstein, What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **40**, 346–358 (2001).
4. F. Martin-Sanchez, I. Iakovidis, S. Nørager, V. Maojo, P. de Groen, J. Van der Lei, T. Jones, K. Abraham-Fuchs, R. Apweiler, A. Babic, R. Baud, V. Breton, P. Cinquin, P. Doupi, M. Dugas, R. Eils, R. Engelbrecht, P. Ghazal, P. Jehenson, C. Kulikowski, K. Lampe, G. De Moor, S. Orphanoudakis, N. Rossing, B. Sarachan, A. Sousa, G. Spekowius, G. Thireos, G. Zahlmann, J. Zvárová, I. Hermosilla, F. J. Vicente, Synergy between medical informatics and bioinformatics: Facilitating genomic medicine for future health care. *J. Biomed. Inform.* **37**, 30–42 (2004).

5. J. A. Doudna, E. Charpentier, The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
6. S. Tripp, M. Grueber, *Economic Impact of the Human Genome Project* (Battelle Memorial Institute, 2011).
7. J. Gitlin, *Calculating the Economic Impact of the Human Genome Project* (National Human Genome Research Institute, 2012); www.genome.gov/27544383/calculating-the-economic-impact-of-the-human-genome-project/.
8. P. E. Stephan, *How Economics Shapes Science* (Harvard Univ. Press, 2012).
9. B. H. Hall, J. Mairesse, P. Mohnen, *Handbook of the Economics of Innovation*, B. H. Hall, N. Rosenberg, Eds. (North-Holland, 2010), vol. 2, pp. 1033–1082.
10. K. H. Fealing, J. I. Lane, J. H. Marburger III, Ed., *The Science of Science Policy: A Handbook* (Stanford Business Books, 2011).
11. H. Torgersen, Fuzzy genes: Epistemic tensions in genomics. *Sci. Cult.* **18**, 65–87 (2009).
12. M. E. J. Newman, The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404–409 (2001).
13. K. Börner, J. T. Maru, R. L. Goldstone, The simultaneous evolution of author and paper networks. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5266–5273 (2004).
14. W. W. Powell, D. R. White, K. W. Koput, J. Owen-Smith, Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *Am. J. Sociol.* **110**, 1132–1205 (2005).
15. L. M. A. Bettencourt, D. I. Kaiser, J. Kaur, Scientific discovery and topological transitions in collaboration networks. *J. Informet.* **3**, 210–221 (2009).
16. C. T. Scott, J. B. McCormick, M. C. DeRouen, J. Owen-Smith, Democracy derived? New trajectories in pluripotent stem cell research. *Cell* **145**, 820–826 (2011).
17. L. M. A. Bettencourt, D. I. Kaiser, J. Kaur, C. Castillo-Chavez, D. E. Wojick, Population modeling of the emergence and development of scientific fields. *Scientometrics* **75**, 495–518 (2008).
18. L. M. A. Bettencourt, J. Kaur, Evolution and structure of sustainability science. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19540–19545 (2011).
19. E. Leahey, J. Moody, Sociological innovation through subfield integration. *Soc. Curr.* **1**, 228–256 (2014).
20. R. Guimerà, B. Uzzi, J. Spiro, L. A. N. Amaral, Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
21. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
22. A. M. Petersen, I. Pavlidis, I. Semendeferi, A quantitative perspective on ethics in large team science. *Sci. Eng. Ethics* **20**, 923–945 (2014).
23. I. Pavlidis, A. M. Petersen, I. Semendeferi, Together we stand. *Nat. Phys.* **10**, 700–702 (2014).
24. S. Milojević, Principles of scientific research team formation and evolution. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3984–3989 (2014).
25. A. M. Petersen, Quantifying the impact of weak, strong, and super ties in scientific careers. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4671–E4680 (2015).
26. S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A.-L. Barabási, Science of Science. *Science* **359**, eaao0185 (2018).
27. C. S. Wagner, *The New Invisible College: Science for Development* (Brookings Institution Press, 2009).
28. L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web* (Stanford InfoLab, 1998).
29. W. Xing, A. Ghorbani, *Proceedings of the Second Annual Conference on Communication Networks and Services Research* (IEEE, 2004), pp. 305–314.
30. R. S. Burt, Structural holes and good ideas. *Am. J. Sociol.* **110**, 349–399 (2004).
31. A. M. Petersen, S. Fortunato, R. K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. E. Stanley, F. Pammolli, Reputation and impact in academic careers. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15316–15321 (2014).
32. D. Rotolo, A. Messeni Petruzzelli, When does centrality matter? Scientific productivity and the moderating role of research specialization and cross-community ties. *J. Organ. Behav.* **34**, 648–670 (2013).
33. E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, F. Schweitzer, Predicting scientific success based on coauthorship networks. *EPJ Data Sci.* **3**, 1–16 (2014).
34. R. K. Pan, A. M. Petersen, F. Pammolli, S. Fortunato, The memory of science: Inflation, myopia, and the knowledge network. *J. Informetrics* **12**, 656–678 (2018).
35. A. M. Petersen, O. Penner, Inequality and cumulative advantage in science careers: A case study of high-impact journals. *EPJ Data Sci.* **3**, 1–25 (2014).
36. D. B. Rubin, Causal inference using potential outcomes. *J. Am. Stat. Assoc.* **100**, 322–331 (2005).
37. V. Larivière, Y. Gingras, On the relationship between interdisciplinarity and scientific impact. *J. Am. Soc. Inf. Sci. Technol.* **61**, 126–131 (2010).
38. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
39. A. Yegros-Yegros, I. Rafols, P. D'Este, Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PLOS ONE* **10**, e0135095 (2015).
40. V. Larivière, S. Haustein, K. Börner, Long-distance interdisciplinarity leads to higher scientific impact. *PLOS ONE* **10**, e0122565 (2015).
41. J. Wang, B. Thijs, W. Glänzel, Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLOS ONE* **10**, e0127298 (2015).
42. E. Leahey, C. M. Beckman, T. L. Stanko, Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Adm. Sci. Q.* **62**, 105–139 (2017).
43. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki; International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
44. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. Russo Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen,

K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
45. Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
46. International Chicken Genome Sequencing Consortium, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695 (2004).
47. K. Lindblad-Toh, C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas III, M. C. Zody, E. Mauceli, X. Xie, M. Breen, R. K. Wayne, E. A. Ostrander, C. P. Ponting, F. Galibert, D. R. Smith, P. J. deJong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C.-W. Chin, A. Cook, J. Cuff, M. J. Daly, D. DeCaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L. Goodstadt, A. Heger, C. Hitte, L. Kim, K.-P. Koepfli, H. G. Parker, J. P. Pollinger, S. Searle, N. B. Sutter, R. Thomas, C. Webber; Broad Sequencing Platform members, E. S. Lander, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
48. A. M. Petersen, D. Rotolo, L. Leydesdorff, A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of Medical Subject Headings. *Res. Policy* **45**, 666–681 (2016).
49. K. Börner, N. Contractor, H. J. Falk-Krzesinski, S. M. Fiore, K. L. Hall, J. Keyton, B. Spring, D. Stokols, W. Trochim, B. Uzzi, A multi-level systems perspective for the science of team science. *Sci. Transl. Med.* **2**, 49cm24 (2010).
50. A. Edelmann, J. Moody, R. Light, Disparate foundations of scientists' policy positions on contentious biomedical research. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6262–6267 (2017).
51. C. A. Bail, Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11823–11828 (2016).
52. J. G. Foster, A. Rzhetsky, J. A. Evans, Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
53. F. J. Van Rijnsoever, L. K. Hessels, Factors associated with disciplinary and interdisciplinary research collaboration. *Res. Policy* **40**, 463–472 (2011).
54. J. N. Cummings, S. Kiesler, J. N. Cummings, Collaborative research across disciplinary and organizational boundaries. *Soc. Stud. Sci.* **35**, 703–722 (2005).
55. J. F. Porac, J. B. Wade, H. M. Fischer, J. Brown, A. Kanfer, G. Bowker, Human capital heterogeneity, collaborative relationships, and publication patterns in a multidisciplinary scientific alliance: A comparative case study of two scientific teams. *Res. Policy* **33**, 661–678 (2004).
56. National Institutes of Health News Release, NIH embraces bold, scientific vision for BRAIN initiative (2014); www.nih.gov/news-events/news-releases/nih-embraces-bold-12-year-scientific-vision-brain-initiative.
57. A. L. Porter, I. Rafols, Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* **81**, 719–745 (2009).
58. N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.-P. Bouchaud, A.-L. Barabási, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.-P. Bouchaud, A.-L. Barabási, Dynamics of ranking processes in complex systems. *Phys. Rev. Lett.* **109**, 128701 (2012).
59. NSF and NIH Funding Data; http://www.nsf.gov/awardsearch/download.jsp; http://exporter.nih.gov/ExPORTER_Catalog.aspx (2015).
60. R. C. Larson, N. Ghaffarzadegan, M. G. Diaz, Magnified effects of changes in NIH research funding levels. *Serv. Sci.* **4**, 382–395 (2012).
61. F. Radicchi, S. Fortunato, C. Castellano, Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17268–17272 (2008).
62. D. Stauffer, A. Aharony, *Introduction to Percolation Theory* (CRC Press, ed. 2, 1994).
63. A. Bunde, S. Havlin, *Fractals and Disordered Systems* (Springer, ed. 2, 1996).
64. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332–7336 (2007).

Acknowledgments

Funding: The authors acknowledge funding from the Eckhard-Pfeiffer Distinguished Professorship Fund and from NSF grant 1738163 entitled "From Genomics to Brain Science." Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. **Author contributions:** A.M.P. developed methods and performed quantitative data analysis. D.M., K.K., and M.E.A. collected and curated data and also developed the software tools. I.P. designed research and linked epistemic analysis to quantitative results. A.M.P. and I.P. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. GS and NSF/NIH RePORTER data are openly available online; Impact Factor data were obtained from the Clarivate Analytics Journal Citations Report. WoS publication and citation data were also obtained from Clarivate Analytics. Scopus data were obtained via calls to the relevant Application Programming Interface. Supporting data is provided through the Open Science Framework repository (https://osf.io/7nb6d/). Additional data related to this paper may be requested from the authors.

Submitted 24 February 2018
Accepted 5 July 2018
Published 15 August 2018
10.1126/sciadv.aat4211

Citation: A. M. Petersen, D. Majeti, K. Kwon, M. E. Ahmed, I. Pavlidis, Cross-disciplinary evolution of the genomics revolution. *Sci. Adv.* **4**, eaat4211 (2018).

Cross-disciplinary evolution of the genomics revolution

Alexander M. Petersen, Dinesh Majeti, Kyeongan Kwon, Mohammed E. Ahmed and Ioannis Pavlidis

Sci Adv **4** (8), eaat4211.
DOI: 10.1126/sciadv.aat4211