Insight into the workforce advancing fields of science and technology

Frank W. Pfrieger*

Centre National de la Recherche Scientifique, Université de Strasbourg, Institut des Neurosciences Cellulaires et Intégratives, F-67000 Strasbourg, France.

https://orcid.org/0000-0001-7085-1431

*Corresponding author

Frank W. Pfrieger, PhD

CNRS UPR 3212 University of Strasbourg

8 allée du Géneral Rouvillois

67000 Strasbourg, France

Email: fw-pfrieger@gmx.de or frank.pfrieger@unistra.fr

Phone: +33388456645

Abstract

Advances in biomedicine and other fields of science and technology depend on research teams and their peer-reviewed publications. The scientific literature represents an invaluable socio-economic resource guiding future research. Typically, this growing body of information is explored by queries in bibliographic databases concerning topics of interest and by subsequent scrutiny of matching publications. This approach informs readily about content, but leaves the workforce driving the field largely unexplored. The hurdle can be overcome by a transparent team-centered analysis that visualizes the teams working in a field of interest and that delineates their genealogic and collaborative relations. Context-specific, but citation-independent metrics gauge team impact and reveal key contributors valuing publication output, mentorship and collaboration. The new insight into the structure, dynamics and performance of the workforce driving research in distinct disciplines complements ongoing efforts to mine the scientific literature, foster collaboration, evaluate research and guide future policies and investments.

Keywords

Authorship, team science, science of science, citation, scientometrics, bibliometrics, meta-analysis, data science, genealogy, collaborations

Introduction

Progress in biomedicine and other fields of science and technology (S&T) depends on research teams working in specific fields and on the publication of their results by peer-reviewed articles. The rapidly growing body of scientific information reflects past and current states of the art and represents an invaluable socio-economic resource guiding future research activity, policies and

investments (Mukherjee et al., 2017; Fischhoff and Scheufele, 2019). Moreover, scientific publications are explored by the "Science of Science" aiming to understand the inner workings of science from global points of view (Clauset et al., 2017; Zeng et al., 2017; Fortunato et al., 2018; Fischhoff and Scheufele, 2019; Hardwicke et al., 2020). The utility of this information relies on bibliographic databases, on refined methods to search and analyse content (Lu, 2011; McLevey and McIlroy-Young, 2017) and on efficient science communication (Fischhoff and Scheufele, 2019). For global analyses complex algorithms process large data sets (Muller et al., 2004; Zeng et al., 2017; Miotto et al., 2018; Kastrin and Hristovski, 2019), whereas a typical user queries a bibliographic database on a specific topic using relevant keywords. From the resulting list of publications, scientific content can be extracted, but the workforce driving the field, its size, dynamics and key contributors remain largely inaccessible. This hurdle can be overcome by a team-centered approach, named TeamTree Analysis (TTA). Based on scientific articles related to a specific topic, this approach reveals instantly the teams working in a research field, visualizes workforce growth, delineates family and collaborative connections and gauges team performance in a citation-independent manner.

Results

To explore TTA, the workforce of an exemplary field in biomedical science was analyzed. A PubMed query using the term "circadian clock" (Clock) yielded a list of articles published between 1960 and 2020, from which TTA identified principal investigators (PIs)/teams working in the field based on last author names (Table 1; Supplementary Data 1).

Visualization and quantitative analysis of team publications, genealogy and collaborations

Plotting publication years of each PI/team against a chronologic team index with alternating sign (TI) creates a tree-like visual revealing each team's entry into the field and its publication count (PC) per year (Fig. 1*A*). The Clock field expanded steadily in terms of workforce and of publication output as indicated by annual counts of newly entering teams and of published articles, respectively (Fig. 1*A*). Individual PIs/teams published up to 113 articles (PC, publication count) as last authors (Last) with a maximum annual output of nearly 7 papers per year. The majority of teams (69%) contributed single articles (Fig. 1*B*). Ranking teams by numbers of publications revealed the top ten players in the Clock field with respect to publication record (Fig. 1*C*).

TTA exposes ancestor - offspring relations based on last - first authors of articles, respectively. A quarter of PIs/teams with last author articles published previously as first authors (PC First) thus qualifying as offspring in this field (Fig. 1*B*). About 11% of the teams qualified as ancestors that generated up to 24 offspring teams (OC, offspring count) and published up to 70 articles with their offspring (PCoff). Offspring teams and their articles represented a relatively constant fraction of the annual workforce and of the publication output (Fig. 1*D*). Overall, the Clock field comprised 543 families with up to 43 members (FS, family size) spanning 5 generations (TG, team generation; Fig. 1*E*). The genealogic analysis indicated the most prolific players in the field and their family relations (Fig. 1*D*, *F*).

Collaborations among teams in the Clock field were delineated from co-authorship of PIs/teams on PubMed articles. Figure 2*A* shows connections between teams (left; last authors) and their collaborators (right; coauthors) and the numbers of collaborators per team (CC, collaboration count). The analysis separates out- and in-degree connections, where teams are listed as last authors or co-authors, respectively. Collaborating teams represented a substantial

fraction of the workforce and contributed more than half of all publications (Fig. 2*A*). The increasing importance of collaborations was indicated by the steadily increasing mean number of authors per article published annually (Fig. 2*B*). About 60% of the PIs/teams working in the Clock field established up to 61 and 58 out- and in-degree collaborations, respectively. Collaborative articles represented 81% of the total publication output with individual PIs/teams publishing up to 68 and 95 collaborative papers as last and coauthor, respectively (Fig. 2*C*). Ranking teams based on CC values revealed the most strongly connected players in the field. They form a large network with a diameter of 10, a mean distance of 4.4 and an edge density of 0.0095 (Fig. 2*D*).

Team ranking based on publication output, mentorship and collaborations

A frequent goal in S&T is to identify key contributors to a field, who can serve as referees, experts, collaborators or awardees. TTA delivered three parameters (PC, OC and CC) allowing to estimate team performance (Figs. 1, 2). Intersection of the top 100 teams for each parameter revealed some overlap between PC-, OC- and CC-based rankings and a core of 35 teams that figured among the top in all three categories (Fig. 2*E*). Plotting individual teams in the three-dimensional parameter space showed that the top teams occupied distinct volumes (Fig. 2*F*). This suggested that the product of the three parameters, further referred to as POC, enables differentiated team ranking that values scientific production, offspring generation and cooperativity (Fig. 2*G*). Do teams with high POC values have a strong impact on the field? This was supported by the high ranks of distinguished scientists (Fig. 2*G*) including Nobel prize winners (M. Rosbash: 7th, M.W. Young: 17th, J.C. Hall: 30st out of 6,506; Table 1; Fig. 2*G*; Supplementary Data 1).

Analysis of distinct scientific disciplines and of highly productive teams

S&T comprise a wide range of disciplines raising the question, whether the field-specific workforce can be evaluated in disciplines other than biomedicine. To address this, exemplary fields in geoscience, computer science, chemistry, astronomy and physics were analyzed (Table 1; Fig. 3*A*; Supplementary Data 1). Synoptic graphs summarized the workforce expansion, genealogic relations and collaborative connections of each field and revealed field-specific differences. Notably, the teams with top ten POC values comprised winners of field-specific scientific awards (Table 1; Fig. 3*A*) corroborating that this parameter can identify high impact teams in different scientific disciplines.

Frequently, specific teams are of interest, for example those with exceptionally high publication output (Ioannidis et al., 2018). TTA allows to trace their publication activity over time while separating articles originating from offspring and collaborators. To illustrate this point, PubMed articles authored by selected PIs working in distinct areas of biomedicine were analyzed (Table 1; Fig. 3*B*-*D*). This revealed their distinct publication histories and the contributions from offspring and collaborators. For example, two PIs/teams, Lip and Raoult, showed strong increases of annual publication rates 6 and 8 years ago, respectively. Breakdown of contributions revealed that these changes were driven by increased numbers of in-degree collaborations, where the PIs are listed as co-authors (Fig. 3*B*,*C*).

Workforce impact on field development

An important question is how research fields develop over time. The dynamics can determine priorities for public funding, private investments and workforce allocation. To explore whether

and how the workforce impacts field development, exemplary fields in biomedicine showing distinct dynamics were analyzed (Table 1; Fig. 4*A*-*C*). Separation of "Newcomers" entering a field per year from "Established" teams working already in the field per year revealed their respective impact on the field's development (Fig. 4*A*). Many newcomers published only one article (SATs, single article teams; Fig. 4*A*) excluding a sustained contribution to the workforce. Their annual fraction was consistently lower in expanding compared to non-expanding fields (Fig. 4*B*). On the other hand, teams with collaborative and family connections showed consistently longer publication periods indicating that they influence the development of each field (Fig. 4*C*).

The visuals and quantitative measures introduced here provide a comprehensive view on the workforce behind a field of research. The choice of field subjected to TTA can be user-dependent (Figs. 1-3, Table 1) or it may be driven by public interest. An example for the latter is the Covid-19 pandemic, where TTA reveals the major contributors, their connections and the remarkable dynamics of the field during coronavirus-induced disease outbreaks (Table 1, Fig. 5).

Effects of subtopics, journals and affiliations

Teams focus on specific subtopics within a field, they publish in different journals and they are affiliated with distinct organizations. How do these factors impact research within a field? To address these points, TTA was applied to teams a) focusing on distinct subtopics related to Alzheimer's disease (Abeta, tau and ApoE; Fig. 6*A*-*C*), b) publishing in multidisciplinary journals in the climate field (J1: Science; J2: PNAS; J3: Nature; J4: PLoS One; J5: Nature Communications; J6: Scientific Reports; Fig. 6*D*-*F*) and c) working on artificial intelligence and machine learning at selected institutions (Affiliation 1: Harvard; A2: Stanford; A3: IBM; A4:

MIT; A5: Microsoft: A6: Google; Table 1; Fig. 6*G-I*). The graphics reveal the factor-dependent development of each field, notably the strong expansion of the Abeta-related workforce and publication output, the strong increase of climate-related teams publishing in recently founded journals and the successive and rather parallel growth of the AI/machine learning field at academic and private institutions (Fig. 6*A, D, G*). Most teams worked on single topics (Fig. 6*B*), published in single journals (Fig. 6*E*) and worked at single institutions (Fig. 6*H*). Increasing the numbers of these factors potentiated POC values (Fig. 6*B, E, H*) and thus probably the team impact. Teams with top ten POC values formed large networks connecting topics (Fig. 6*C*) and affiliations (Fig. 6*I*). Co-publishing and collaborating teams in the Climate field established selective connections between journals and affiliations, respectively. These connections differed in strength and depended in part on the total workforce and on the publication period (Fig. 6*F, I*).

Discussion

The approach introduced here provides new insight into the workforce that advances a field of interest in S&T based on peer-reviewed articles. The field- and team-related visuals and measures can be scrutinized ad-hoc following a database query. Thereby, users gain an accessible and transparent tool to mine the ever-growing scientific literature. Learning about the workforce of an important, but unfamiliar field facilitates the identification of experts and the establishment of collaborations crossing disciplinary boundaries (Trujillo and Long, 2018). The field-specific approach complements global analyses focusing on team impact (Sekara et al., 2018; Ahmadpoor and Jones, 2019), evolution (Milojevic, 2014) and affiliation (Jones et al., 2008; Way et al., 2019). Similar to other approaches, TTA faces the name ambiguity challenge, whose solution requires more refined approaches (Zeng et al., 2017). The measures introduced

here can help to evaluate context-specific team performance and to reveal the impact of the workforce on the development of a research area. Notably, the POC value takes into account three key activities in research, namely scientific production, mentorship and cooperativity, and reliably identifies key contributors. Thereby, this metric complements measures such as citations (Hirsch, 2005; Ioannidis et al., 2016; Zeng et al., 2017) without requiring proprietary databases. On the other hand, POC values are field-specific and depend on the period of activity therefore precluding absolute ranking of teams or evaluation of junior scientists. The delineation of family relations and of collaborative connections allows to categorize the publication output of individual PIs/teams and to identify the different types of contributors (e.g. offspring, collaborators). The automatic delineation of family connections based on first author-last author pairs provides an alternative to previous efforts requiring user input (David and Hayden, 2012; Hirshman et al., 2016; Lienard et al., 2018) (see also Mathematics Genealogy Project: https://www.genealogy.math.ndsu.nodak.edu/). However, this approach may underestimate offspring counts in the case of first or last co-authorship, when first authors change the field of interest and in the case of alphabetical author lists or of field-specific author ranking (Waltman, 2012). Genealogic and collaborative connections seem to enhance the impact of teams and to prolong their life-span within a field. This underlines the relevance of training and mentorship ensuring the continuity of research in S&T (Sauermann and Haeussler, 2017) and supports the importance of cooperations regardless of the field (Lu et al., 2011; Mukherjee et al., 2017; Wuchty et al., 2007; Stallings et al., 2013; Coccia and Wang, 2016; Parish et al., 2018). Considering these connections and the resulting publications within a research area adds an important component to evaluate team performance, inform about underlying networks and forecast the dynamics of the field.

Materials and Methods

TTA was implemented by custom-written software with additional data analysis and visualization accomplished using the open source software R (R Core Team, 2019) and selected R packages [data.tree: Hammil et al., 2018 https://cran.r-project.org/package=data.tree; data.table: Dowle, 2019 https://cran.r-project.org/package=data.table; dunn.test: Dinno, 2017 https://cran.r-project.org/package=dunn.test; eulerr: Larsson, 2019 https://cran.r-project.org/package=eulerr; ggplot2: Wickham, 2016; ggrepel: Slowikowski et al., 2018 https://cran.r-project.org/package=ggrepel; igraph: Csardi and Nepusz (2006); plot3D: Soetaert, 2017 https://cran.r-project.org/package=plot3D]. Code and additional data are available upon written request to the author. TTA was applied to lists of publications resulting from queries of bibliographic databases concerning topics of interest. PIs listed as last authors defined research teams. Last author names included initials to reduce author ambiguity (Milojevic, 2013). Articles with an arbitrary limit of 1000 authors were omitted from the analysis. TTA assigned a chronologic index (TI) according to the year of the first publication with alternating sign enabling a tree-like display of the workforce. Further, TTA attributed a color to each team and calculated parameters summarizing publication record, genealogy and collaborations. Genealogic relations were based on offspring - ancestor pairs, where offspring was defined as PIs that appeared initially as first authors on a publication of an ancestor team and that published subsequently as last (team) authors. Each offspring (first author; generation i+1) was assigned to the ancestor (last author; generation i) with the earliest common publication. Families were defined as progeny of first ancestors (i = 1) encompassing all subsequent generations. TTA derived collaborations based on co-authorship (Newman, 2004). Out- and in-degree connections

specified how often a team listed other teams as co-authors and how often the same team was listed as co-author, respectively. The research fields selected for analyses are summarized in Table 1. Statistical analyses were performed using the indicated tests.

Acknowledgments

The author thanks Drs. V. Demais, M. Muzet, V. Pallottini, H. Runz, and M. Slezak for helpful comments on previous versions of the manuscript.

References

Ahmadpoor, M. & Jones, B. F. (2019) Decoding team and individual impact in science and invention. *Proc Natl Acad Sci U S A* **116,** 13885-13890.

Clauset, A., Larremore, D. B. & Sinatra, R. (2017) Data-driven predictions in the science of science. *Science* **355,** 477-480.

Coccia, M. & Wang, L. (2016) Evolution and convergence of the patterns of international scientific collaboration. *Proc Natl Acad Sci U S A* **113,** 2057-2061.

Csardi, G. & Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal* ***Complex Systems,*** 1695.

David, S. V. & Hayden, B. Y. (2012) Neurotree: a collaborative, graphical database of the academic genealogy of neuroscience. *PLoS One* **7,** e46608.

Fischhoff, B. & Scheufele, D.A. (2019) The Science of Science Communication III. *Proc Natl Acad Sci U S A* **116**, 7632-7633.

Fortunato, S. et al. (2018) Science of science. *Science* **359,** eaao0185.

Hardwicke, T.E., et al. (2020) Calibrating the Scientific Ecosystem Through Meta-Research. *Annu Rev Stat Appl* **7,** 11-37.

Hirsch, J.E. (2005) An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A* **102,** 16569-16572.

Hirshman, B. R., et al. (2016) Impact of medical academic genealogy on publication patterns: An analysis of the literature for surgical resection in brain tumor patients. *Ann Neurol* **79,** 169-177.

Ioannidis, J. P. A., Klavans, R. & Boyack, K. W. (2018) Thousands of scientists publish a paper every five days. *Nature* **561,** 167-169.

Ioannidis, J. P., Klavans, R. & Boyack, K. W. (2016) Multiple Citation Indicators and Their Composite across Scientific Disciplines. *PLoS Biol* **14,** e1002501.

Jones, B. F., Wuchty, S. & Uzzi, B. (2008) Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science* **322,** 1259-1262.

Kastrin, A. & Hristovski, D. (2019) Disentangling the evolution of MEDLINE bibliographic database: A complex network perspective. *J Biomed Inform* **89,** 101-113.

Lienard, J. F., Achakulvisut, T., Acuna, D. E. & David, S. V. (2018) Intellectual synthesis in mentorship determines success in academic careers. *Nat Commun* **9,** 4840.

Lu, Z.Y. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database-Oxford*.

McLevey, J. & McIlroy-Young, R. (2017) Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *J Informetr* **11,** 176-197.

Milojevic, S. (2013) Accuracy of simple, initials-based methods for author name disambiguation. *J Informetr* **7,** 767-773.

Milojevic, S. (2014) Principles of scientific research team formation and evolution. *Proc Natl Acad Sci U S A* **111,** 3984-3989.

Miotto, R., Wang, F., Wang, S., Jiang, X. Q. & Dudley, J. T. (2018) Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* **19,** 1236-1246.

Mukherjee, S., Romero, D. M., Jones, B. & Uzzi, B. (2017) The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Sci Adv* **3,** e1601315.

Muller, H. M., Kenny, E. E. & Sternberg, P. W. (2004) Textpresso: An ontology-based information retrieval and extraction system for biological literature. *Plos Biology* **2,** 1984-1998.

Newman, M. E. J. (2004) Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci U S A* **101**, 5200-5205.

Parish, A. J., Boyack, K. W. & Ioannidis, J. P. A. (2018) Dynamics of co-authorship and productivity across different fields of scientific research. *PLoS One* **13,** e0189742.

R Core Team. (R Foundation for Statistical Computing, Vienna, Austria, 2019).

Sauermann, H. & Haeussler, C. (2017) Authorship and contribution disclosures. *Sci Adv* **3,** e1700404.

Sekara, V., et al. (2018) The chaperone effect in scientific publishing. *Proc Natl Acad Sci U S A* **115,** 12603-12607.

Stallings, J., Vance, E., Yang, J., Vannier, M. W., Liang, J., Pang, L., Dai, L., Ye, I. & Wang, G. (2013) Determining scientific impact using a collaboration index. *Proc Natl Acad Sci U S A* **110**, 9680-9685.

Trujillo, C. M. & Long, T. M. (2018) Document co-citation analysis to enhance transdisciplinary research. *Sci Adv* **4,** e1701130.

Waltman, L. (2012) An empirical analysis of the use of alphabetical authorship in scientific publishing. *J Informetr* **6,** 700-711.

Way, S. F., Morgan, A. C., Larremore, D. B. & Clauset, A. (2019) Productivity, prominence, and the effects of academic environment. *Proc Natl Acad Sci U S A* **116,** 10729-10733.

Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).

Wuchty, S., Jones, B. F. & Uzzi, B. (2007) The increasing dominance of teams in production of knowledge. *Science* **316,** 1036-1039.

Zeng, A. et al. (2017) The science of science: From the perspective of complex systems. *Phys. Rep. - Review Section of Physics Letters* **714,** 1-73.

Table 1. Summary of research areas analyzed.

Field / Discipline	Database	Pubs / Teams / Year
Alzheimer's Disease + Subtopics / Biomedicine	PubMed	37453 / 13837 / 1984
Aplysia (APL) / Biomedicine	PubMed	4657 / 1577 / 1898
Artificial Intelligence, Machine Learning + Affiliations / Computer Science	WOS	17410 / 9557 / 1962
Chirped Laser Pulses[1] / Physics	WOS	9437 / 4695 / 1969
Circadian Clock[2] / Biomedicine	PubMed	17,023 / 6506 / 1960
Climate + Journals / Multiple Disciplines	WOS	24125 / 16776 / 1883
Coronavirus OR corona-virus OR "corona virus" OR covid-19 OR sars-cov OR "severe acute respiratory syndrome" OR "Middle East respiratory syndrome"	PubMed	19414 / 9346 / 1949
Cosmic Inflation[3] / Astronomy	WOS	9018 / 3510 / 1981
Creutzfeldt Jacob Disease (CJD) / Biomedicine	PubMed	7950 / 4317 / 1946
CRISPR-CAS (CRI) / Biomedicine	PubMed	14255 / 7738 / 2002
Extracellular Vesicles (EVS) / Biomedicine	PubMed	13621 / 7034 / 1970
Freeze-Fracture (FRE) / Biomedicine	PubMed	8981 / 4665 / 1961
Ice core climate[4] / Geoscience	WOS	9727 / 5910 / 1956
Ind. Pluripotent Stem Cells (IPS) / Biomedicine	PubMed	16286 / 7504 / 1967
Quantum Computer[5] / Computer Science	WOS	29938 / 10575 / 1967
Supramolecular Chemistry[6] / Chemistry	WOS	19409 / 7820 / 1985
Gregory Y.H. Lip / Selected Team / Biomedicine	PubMed	1802 / 335 / 1994

Hagop Kantarjian / Selected Team / Biomedicine	PubMed	2029 / 324 / 1981
Didier Raoult / Selected Team / Biomedicine	PubMed	2220 / 197 / 1985

Selected awards: [1]Nobel Prize Physics (2018), [2]Nobel Prize Physiology/Medicine (2017); [3]Kavli Prize (2014); [4]Vetlesen Award (2013); [5]Micius Quantum Prize (2018, 2019); [6]Nobel Prize Chemistry (1987).

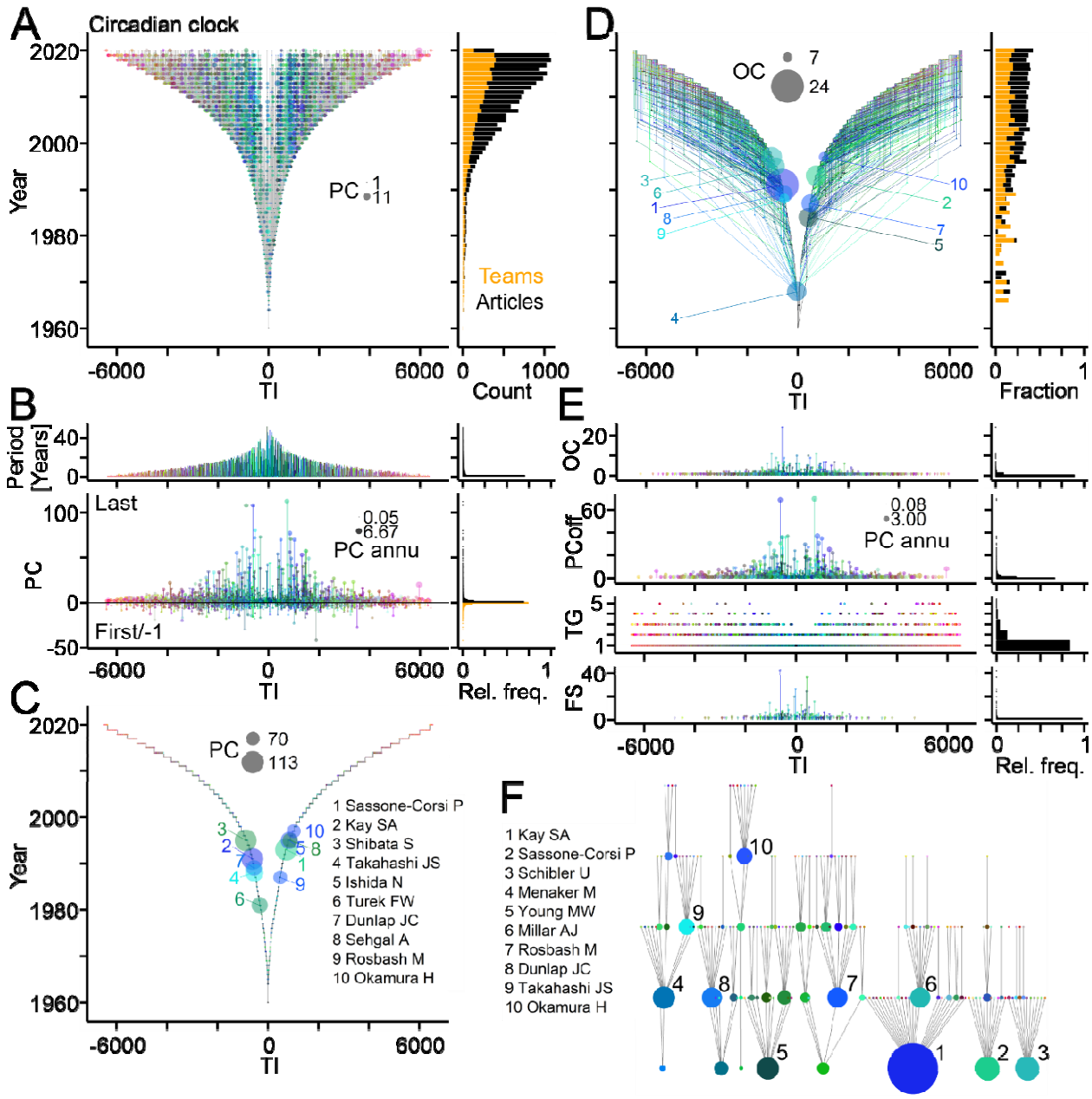

Figure 1. Publication record and genealogic relations of the workforce driving a field of biomedical research.

A, Publication records of individual teams working on "Circadian Clock". PIs/teams, represented by team index (TI), were identified based on last authors of PubMed articles. Symbol size, publication count (PC) per year (left). Number of teams entering the field per year (orange) and

of articles (black) published per year (right). *B*, Publication period (top) of individual teams and counts (bottom) of last (positive) and first (negative) author publications per PI/team. Symbol size, mean annual PC (PC annu) (left). Relative frequency distributions of parameters shown on the left (right). *C*, Teams with top ten PC Last values indicated by symbol size. *D*, Ancestor-to-offspring connections of teams with top ten offspring counts (OC) (left; names indicated in part F) and fraction of offspring teams (orange) and of offspring publications (black) compared to total number per year (right). *E*, on left side from top to bottom, counts of offspring (OC), of offspring publications (PC off, positive) per team, generation of each team (TP, middle) and family size (FS) of first-generation teams. Symbol size, mean annual counts of offspring publications (PC annu). On right side, relative frequency distributions of parameters shown on the left. *F*, Family trees and names of teams with top ten OC values.

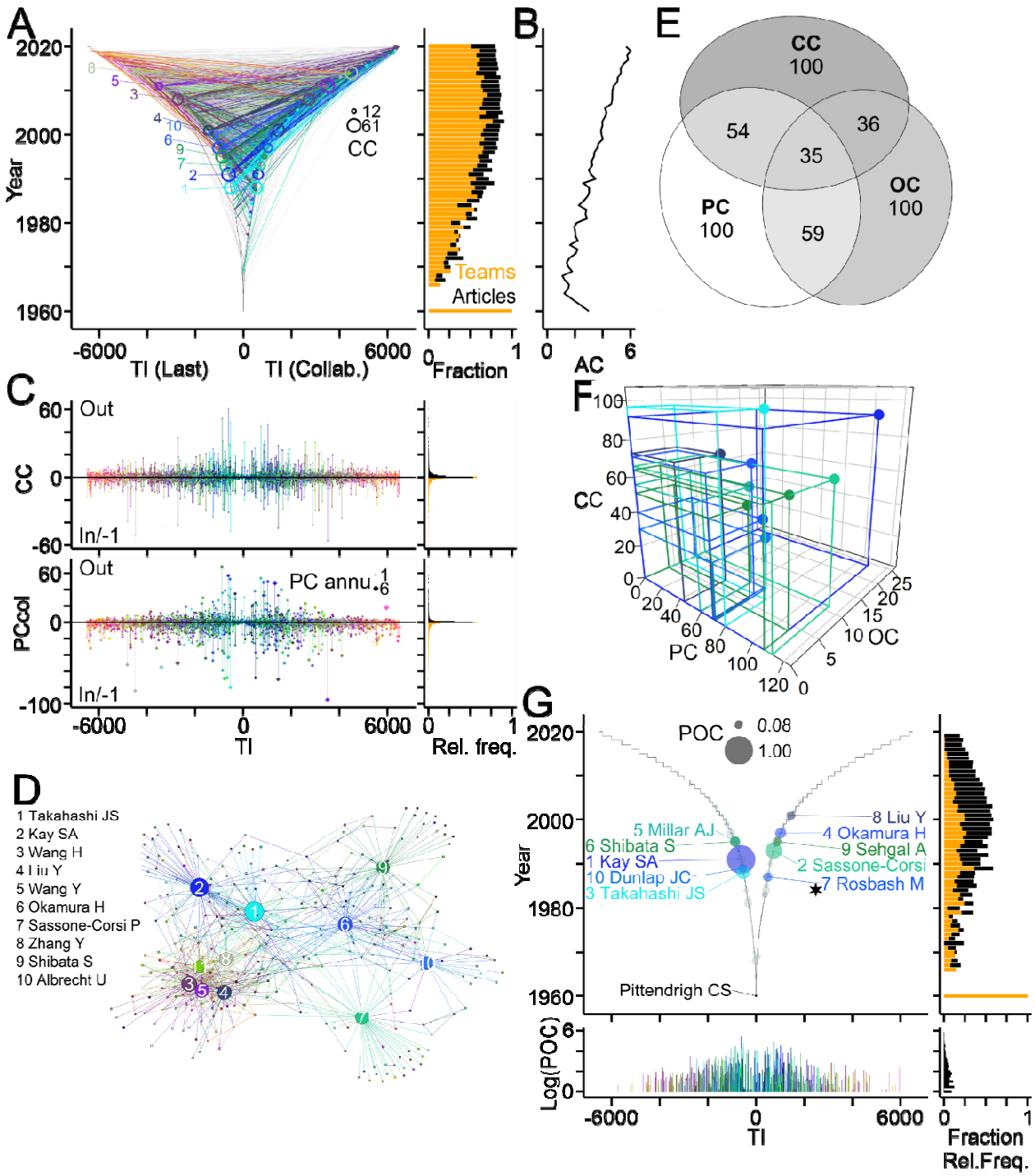

Figure 2. Collaborator network and impact of teams contributing to the Clock field.

A, Collaborative connections (left) between PIs/teams listed as last author (negative TI) and as collaborating co-authors (positive TI). Colored lines, connections of teams with top ten connection counts (CC). Symbol size, out (negative) and in (positive) degree values. Annual

fractions of collaborating teams (orange) and their publications (black) compared to total numbers (right). *B*, Mean number of authors (AC) per article published each year. *C*, Counts of out-degree (positive, last author) and of in-degree collaborations (negative, co-author; top) and of resulting articles (bottom) per PI/team (left) and corresponding relative frequency distributions (right). Symbol size (left), mean annual PC. *D*, Names of teams with top ten CC values and their networks. Circles and squares, offspring and non-offspring teams, respectively. Symbol size, CC values normalized to maximum. *E*, Numbers of intersecting teams ranking among top 100 for each parameter. *F*, 3D space occupied by teams with top ten POC values. *G*, POC values of teams (symbol size, POC normalized to max; symbol color: top ten teams, team color; all others, grey) with rank and names of top ten teams and name of team publishing first article in the field (top). Log10(POC) values per team (left) and their relative frequency distribution (right, bottom). Star, Nobel prize winner.

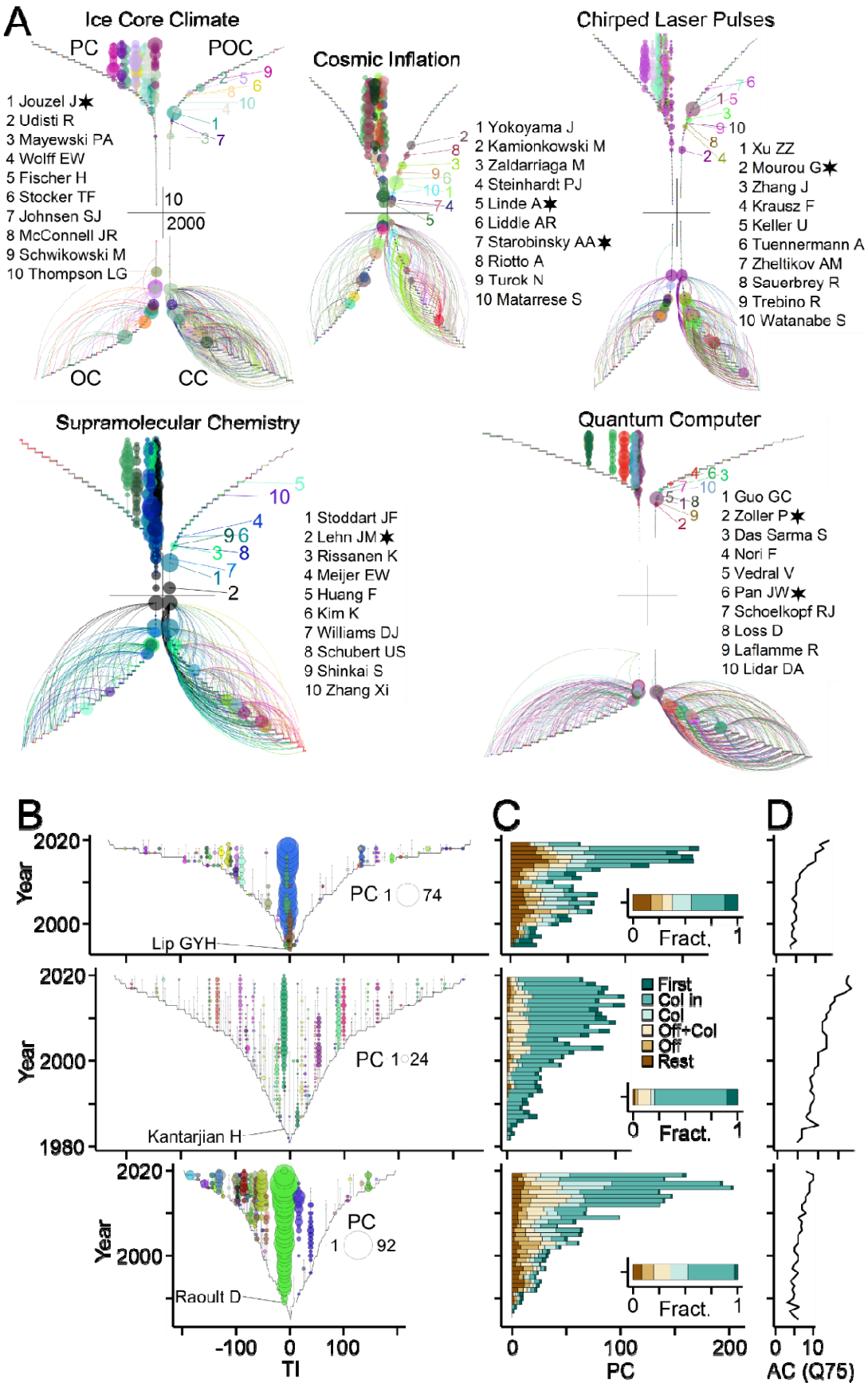
A
Ice Core Climate
PC
POC
OC
CC
10
2000
1 Jouzel J✶
2 Udisti R
3 Mayewski PA
4 Wolff EW
5 Fischer H
6 Stocker TF
7 Johnsen SJ
8 McConnell JR
9 Schwikowski M
10 Thompson LG
Cosmic Inflation
1 Yokoyama J
2 Kamionkowski M
3 Zaldarriaga M
4 Steinhardt PJ
5 Linde A✶
6 Liddle AR
7 Starobinsky AA✶
8 Riotto A
9 Turok N
10 Matarrese S
Chirped Laser Pulses
1 Xu ZZ
2 Mourou G✶
3 Zhang J
4 Krausz F
5 Keller U
6 Tuennermann A
7 Zheltikov AM
8 Sauerbrey R
9 Trebino R
10 Watanabe S
Supramolecular Chemistry
1 Stoddart JF
2 Lehn JM✶
3 Rissanen K
4 Meijer EW
5 Huang F
6 Kim K
7 Williams DJ
8 Schubert US
9 Shinkai S
10 Zhang Xi
Quantum Computer
1 Guo GC
2 Zoller P✶
3 Das Sarma S
4 Nori F
5 Vedral V
6 Pan JW✶
7 Schoelkopf RJ
8 Loss D
9 Laflamme R
10 Lidar DA
B
C
D
Year
2020
2000
1980
PC 1 74
Lip GYH
PC 1 24
Kantarjian H
PC
1 92
Raoult D
-100
0
100
TI
0
Fract.
1
First
Col in
Col
Off+Col
Off
Rest
0
100
200
PC
0
10
AC (Q75)

Figure 3. Workforce in distinct disciplines and teams associated with individual PIs.

A, Workforce of selected research fields from distinct disciplines showing teams with top ten values per indicated parameter. Symbol size, relative counts normalized to respective maximum. Names, teams with top ten POC values. Horizontal and vertical scale bars (half length) indicate number of teams and years, respectively. Stars, recipients of selected awards (Table 1). *B*, PC per year of indicated PIs and their associated teams detected by TTA of PubMed publications. *C*, Publication rates per year separating articles by contributions (First, first author; Col in, in-degree collaborations; Col, out-degree collaborations without offspring; Off+Col, offspring with collaborators; Off, offspring without collaborators; Rest, all other publications). Insets, contributions of indicated categories to total publication output. *D*, Mean number of authors (AC; Q75 values) per article published each year.

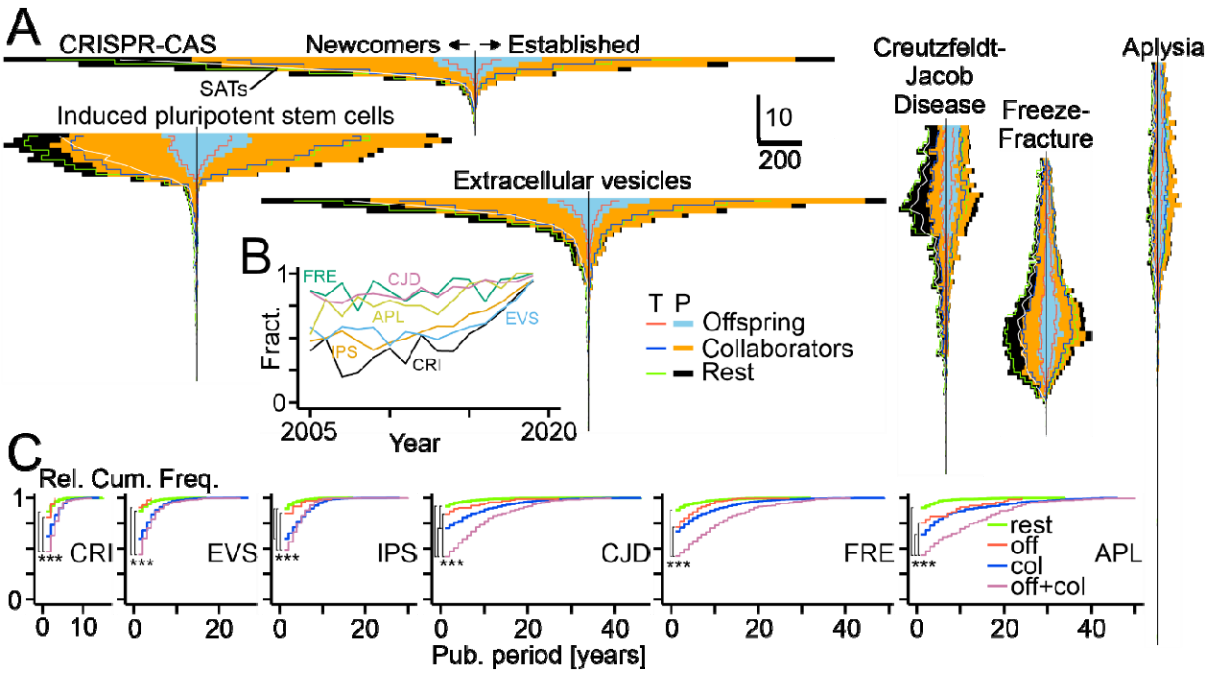

Figure 4. Workforce dynamics in selected fields of biomedicine

A, Annual publication (P, bars) and team (T, lines) counts distinguishing newcomers (left) and established teams (right) from the indicated fields and categories. Horizontal and vertical scale bars indicate number of teams and years, respectively. *B*, Annual fractions of SATs in indicated fields. *C*, Publication periods of teams from indicated categories and fields. Grey lines, differences among groups ($p < 0.001$, Kruskal-Wallis tests. Asterisks, $p < 0.001$, post-hoc Dunn test, Benjamini-Hochberg adjusted; group sizes adjusted to minimum by random selection; CRI: n =31; EVS: 42; IPS: 70; CJD: 91; FRE: 174; APL: 65).

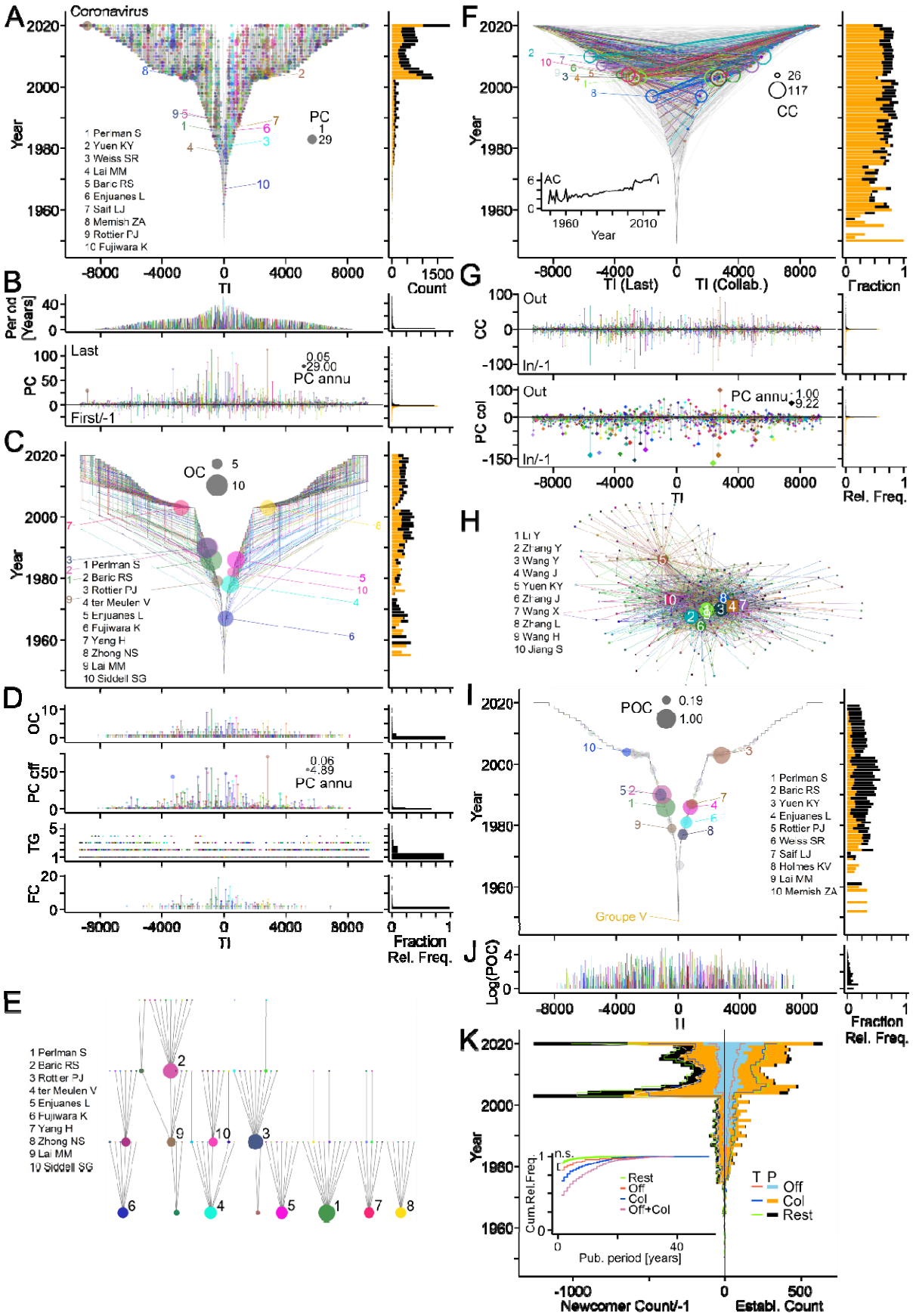
A
Coronavirus
Year
PC
1
29
1 Perlman S
2 Yuen KY
3 Weiss SR
4 Lai MM
5 Baric RS
6 Enjuanes L
7 Saif LJ
8 Memish ZA
9 Rottier PJ
10 Fujiwara K
TI
Count
B
Period [Years]
Last
First/-1
PC annu
C
OC
1 Perlman S
2 Baric RS
3 Rottier PJ
4 ter Meulen V
5 Enjuanes L
6 Fujiwara K
7 Yang H
8 Zhong NS
9 Lai MM
10 Siddell SG
D
OC
PC off
TG
FC
PC annu
Fraction
Rel. Freq.
E
1 Perlman S
2 Baric RS
3 Rotter PJ
4 ter Meulen V
5 Enjuanes L
6 Fujiwara K
7 Yang H
8 Zhong NS
9 Lai MM
10 Siddell SG
F
AC
CC
TI (Last)
TI (Collab.)
Fraction
G
Out
In/-1
CC
PC col
PC annu
Rel. Freq.
H
1 Li Y
2 Zhang Y
3 Wang Y
4 Wang J
5 Yuen KY
6 Zhang J
7 Wang X
8 Zhang L
9 Wang H
10 Jiang S
I
POC
1 Perlman S
2 Baric RS
3 Yuen KY
4 Enjuanes L
5 Rottier PJ
6 Weiss SR
7 Saif LJ
8 Holmes KV
9 Lai MM
10 Memish ZA
Groupe V
J
Log(POC)
Fraction
Rel. Freq.
K
Cum.Rel.Freq.
n.s.
Rest
Off
Col
Off+Col
Pub. period [years]
T P
Off
Col
Rest
Newcomer Count/-1
Establ. Count

Figure 5. TeamTree analysis of PubMed articles related to "coronavirus" and associated diseases.

A, Publication record of individual PIs/teams identified based on last authors of PubMed articles (Table 1) and represented by team index (TI). Symbol size, publication count (PC) per year (left). Number of teams entering the field per year (orange) and of articles (black) published per year (right). Names indicate teams with top ten total PC. *B,* Publication periods (top, left) of individual teams and their PC values as last author (positive) and as first author (negative; bottom, left) and relative frequency distribution of the indicated parameters (right). Symbol size, mean annual PC (PC annu). *C*, Ancestor-to-offspring connections of teams with top ten offspring counts (OC) (left) and fraction of offspring teams (orange) and of offspring publications (black) compared to total counts per year (right). *D*, left, from top to bottom, counts of offspring (OC), of offspring publications (PC off, positive) per team, generation of each team (TP, middle) and family size (FS) of first-generation teams. Symbol size, mean annual counts of offspring publications (PC annu). Right, relative frequency distributions of parameters shown on the right. *E*, Family trees and names of teams with top ten OC values. *F*, Collaborative connections (left) between PIs/teams listed as last author (negative TI) and as collaborating co-authors (positive TI). Colored lines, connections of teams with top ten connection counts (CC). Symbol size, out (negative) and in (positive) degree values. Annual fractions of collaborating teams (orange) and their publications (black) compared to total numbers (right). Inset, mean number of authors (AC) per article published per year. *G*, Counts of out-degree (positive, last author) and of in-degree collaborations (negative, co-author; top) and of resulting articles (bottom) per PI/team (left) and corresponding relative frequency distributions (right). Symbol size (left), mean annual PC. *H*, Names and collaborator network of teams with top ten CC values. Circles and squares, offspring

and non-offspring teams, respectively. Symbol size, CC values normalized to maximum. *I*, POC values of teams (symbol size, POC normalized to max; symbol color: top ten teams, team color; all others, grey). *J*, Log10(POC) values per team (left) and their relative frequency distribution (right). *K*, Annual publication (P, bars) and team (T, lines) counts distinguishing newcomers (left) and established teams (right) from the indicated categories. White line, count of single article teams. Inset, Publication periods of teams from indicated categories and fields. Groups showed statistically significant differences except for the indicated pair ($p < 0.001$, Kruskal-Wallis tests. Asterisks, $p < 0.001$, post-hoc Dunn test, Benjamini-Hochberg adjusted; group sizes adjusted to minimum by random selection).

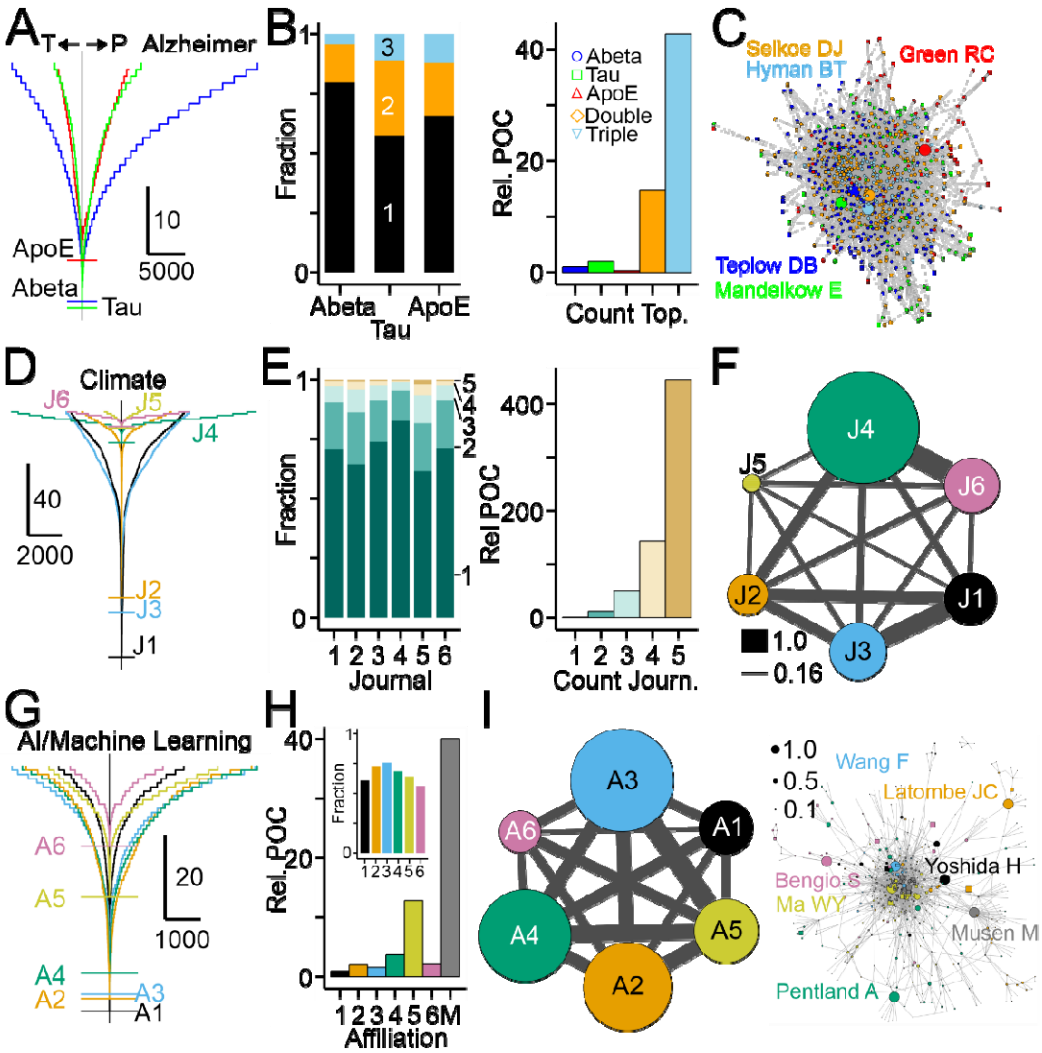

Figure 6. Segregation of the workforce by specific factors.

Exploration of subtopics (*A-C*), journals (J1: Science; J2: PNAS; J3: Nature; J4: PLoS One; J5: Nat. Comm.; J6: Sci. Rep.) (*D-F*) and affiliations (A1: Harvard; A2: Stanford; A3: IBM; A4: MIT; A5: Microsoft: A6: Google) (*G-I*) in indicated fields. Factor-specific workforce development (*A*, *D*, *G*), fractions of teams focusing on one or more subtopics (*B*, left: colors, number of topics), publishing in indicated number of journals (*E*, left: colors, number of journals) or working at a single affiliations (*H*, inset). Relative POC values compared to controls (*B*, right: Abeta; *E*, right: Journal 1; *H*, Affiliation 1; *M*, multiple). (*C* and *I* right) Networks of teams with top 10 POC values from each category (PI names). (*F* and *I* left) strength (thickness,

normalized to maximal numbers) of team- and collaboration-based connections between journals (*F*) and affiliations (*I*, left), respectively. Symbol sizes, number of teams per journal or affiliation (*F*, *I* left) and CC values normalized to maximum per affiliation (*I*, right; largest size, teams with top ten POC values per affiliation).

Supplementary Information

Insight into the workforce advancing fields of science and technology

Frank W. Pfrieger

Email: fw-pfrieger@gmx.de or frank.pfrieger@unistra.fr

Supplementary Data 1. Excel file with TeamTree data for selected fields of research.