RHODRI LENG, GIL VIRY, MIGUEL GARCÍA-SANCHO, JAMES LOWE, MARK WONG AND NIKI VERMEULEN*

The Sequences and the Sequencers: What Can a Mixed-Methods Approach Reveal about the History of Genomics?

ABSTRACT

This special issue on sequences and sequencers uses new analytical approaches to re-assess the history of genomics. Historical attention has largely focused on a few central characters and institutions: those that participated in the Human Genome Project (HGP), especially its final stages. Our analysis—based on an assessment of almost 13.5 million DNA sequence submissions and 30,000 publications of human, yeast, and pig DNA sequences—followed overlapping chronologies starting before and finishing after the concerted efforts to sequence the genomes of each species: 1980 to 2000 in yeast, 1985 to 2005 for the human, and 1990 to 2015 for the pig. Our main conclusion is that when broader sequencing practices—especially those addressed to nonhuman species—are taken into account, the large-scale center model that characterized the organization of the HGP falls short in representing genomics as a whole. Instead of taking the HGP as a model, we describe an iterative process in which the practices of sequence submission and publication were entangled. Analysis of co-authorship networks between institutions derived from our data shows how linked sequence submission and publication were to medical, biochemical, and agricultural research. Our analysis thus reveals the utility of big data and mixed-methods approaches for addressing science as a multidimensional endeavor with a history shaped by co-constitutive, synchronic interactions among different elements—such as communities, species, and disciplines—as much as diachronic

*Miguel García-Sancho, Science, Technology and Innovation Studies, University of Edinburgh, Old Surgeons' Hall, High School Yards, Edinburgh, EH1 1LZ, United Kingdom. miguel.gsancho@ed.ac.uk

The following abbreviations are used: API, Application Programming Interface; DNA, deoxyribonucleic acid; ENA, European Nucleotide Archive; PMC, Europe; HGP, Human Genome Project; YGSP, Yeast Genome Sequencing Project.

Historical Studies in the Natural Sciences, Vol. 52, Number 3, pps. 277–319. ISSN 1939-1811, electronic ISSN 1939-182X. DOI: https://doi.org/10.1525/hsns.2022.52.3.277.

trajectories over time. This perspective enables us to better capture interdisciplinary and interspecies work, and offers a more fluid portrayal of the connections between scientific practices and agricultural, industrial, and medical goals. This essay is part of a special issue entitled The Sequences and the Sequencers: A New Approach to Investigating the Emergence of Yeast, Human, and Pig Genomics, edited by Miguel García-Sancho and James Lowe.

KEY WORDS: genomics, DNA sequencing, genome sequencing, *Saccharomyces cerevisiae*, *Homo sapiens*, *Sus scrofa*, mixed methods, co-authorship networks, social network analysis

1. BACKGROUND: THE HISTORIOGRAPHICAL TASK

In this special issue, we investigate the practice of DNA sequencing by combining historical research with the analysis of institutional co-authorship networks. We do this with the aim of opening up new vistas in the historiography of genomics and contributing to ongoing attempts at placing genomics within a broader history of contemporary life sciences research. Earlier historical scholarship has uncovered multiple genealogies of genomics, including the redefinition of the questions, theories, and experimental approaches of molecular biology; the practice of sequencing biological molecules, especially proteins, and their use in evolutionary phylogenetics research; the history of computing and its connections with biology, as in the creation of the discipline of bioinformatics; and the hybridization of the comparative and experimental ways of knowing throughout the twentieth century.[1] These narratives emphasize continuity over time: the continuity of genomics with molecular biology; with pre-genomic practices of sequencing and sequence comparison; and with existing ways of producing knowledge that may be manifested and combined in distinct ways in genomics research. Yet, as we will show here and in the other papers of this special issue, there are alternative ways of historicizing genomics.

The uniqueness and novelty of genomic endeavors has long fascinated scholars. Building on spatially and longitudinally extensive ethnographic work,

1. Michel Morange, *The Black Box of Biology: A History of the Molecular Revolution* (Cambridge, MA: Harvard University Press, 2020); Edna Suárez-Díaz, "The Long and Winding Road of Molecular Data in Phylogenetic Analysis," *Journal of the History of Biology* 47 (2014): 443–78; Hallam Stevens, *Life Out of Sequence: A Data-Driven History of Bioinformatics* (Chicago: University of Chicago Press, 2013); Miguel García-Sancho, *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA, 1945–2000* (Basingstoke: Palgrave-Macmillan, 2012); Bruno J. Strasser, *Collecting Experiments: Making Big Data Biology* (Chicago: University of Chicago Press, 2019).

Stephen Hilgartner has argued that the success story behind the Human Genome Project (HGP) was driven by the proposal of a distinctive "knowledge-control regime"—one that portrayed genomics as distinct from earlier molecular biological research while not threatening to displace or question the status of molecular biology. The knowledge-control regime of genomics relied on what we call the large-scale center model: an organizational regime characteristic of the HGP proposed by a self-conscious "genomics vanguard" in the late 1980s. This vanguard included both Nobel Prize–winning molecular biologists and champions of the emerging biotechnology industry. It envisaged the HGP as an effort that a small number of high-throughput and industrially organized facilities called genome centers would undertake over a set period of time. According to Hilgartner, the vanguard conceptualized these centers as sequence data producers, with an intended unidirectional flow of the data from the centers to global, open-access repositories, and from these databases to eventual downstream users in laboratories.[2]

In this large-scale center model, production and use of sequence data were spatially, temporally, and intellectually separated. The genome centers produced and packaged the data for journeys to data repositories that applied metadata to the sequences, thus enabling them to be found and downloaded by end-users.[3] The advent of this model reflected the perceived need to constantly upscale production of sequence data to meet the ambitious HGP targets. Within this regime and model, the creation of fewer and larger sequencing centers, and a division of labor between sequence production and use, seemed a logical progression for genomics research. This tendency was further advanced by the 1996 advent of the Bermuda Principles that mandated daily public release of DNA sequence production to global, open-access databases ahead of its publication in the scientific literature. In 2003, the Fort Lauderdale report explicitly distinguished the practices of submitting sequences to databases and those concerned with analyzing them and recognized the right of the original submitter to be the first in publishing the sequence.[4]

2. Stephen Hilgartner, *Reordering Life: Knowledge and Control in the Genomics Revolution* (Cambridge, MA: MIT Press, 2017).

3. To use the analytical framework of Sabina Leonelli, *Data-Centric Biology: A Philosophical Study* (Chicago: University of Chicago Press, 2016) and Sabina Leonelli and Niccolò Tempini, eds., *Data Journeys in the Sciences* (Cham, Switzerland: Springer International Publishing, 2020).

4. Kathryn Maxson Jones, Rachel A. Ankeny, and Robert Cook-Deegan, "The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project," *Journal of the History of Biology* 51 (2018): 693–805.

The large-scale center model was adopted gradually. Hilgartner observes that during the early years of the HGP, in the early to mid-1990s, the distinction between production and use of sequence data was fuzzier, to the extent that they made less sense as separable categories even within the genome centers. Yet the novelty of those centers and the knowledge-control regime they embodied has focused scholarly attention. Studies that combine historical research with anthropological observations have unpacked the trajectory and operation of the Sanger Institute and the Center for Genome Research at the Whitehead Institute, the two main HGP contributors in the UK and US.[5] Another body of literature has compared the human genome sequence these two centers produced, along with the other HGP participants, with that of corporate counterpart, Celera Genomics.[6] These rival, large sequencing efforts and their legacy in open-access databases also shape more recent investigations of the phenomenon of postgenomics.[7]

The genome center model and its successful mobilization in the HGP has thus informed the historical approach of scholars and prevailed as the standard definition of genomics in reconstructions of sequencing efforts, narratives of continuity with earlier life sciences endeavors, and explorations of subsequent postgenomic research. Without wanting to overlook the significance of this form of genomics, we argue that scholarly emphasis has produced a series of historiographical gaps that we address throughout the special issue.

First, a substantial number of institutions contributed to genomics research despite not adopting the genome center model. The first use of the term "genomics," in 1987, predated the launch of the HGP and designated the collective gathering of map and sequence data. Rather than constituting a selective club of large-scale producers, this early incarnation of genomics involved a wider range of institutions with varied sequencing technologies and approaches, among them human genetics, medical genetics, and cytogenetics laboratories. The article specifically dealing with human genomics in this special issue details the collective efforts of these institutions, which operated

5. On the Sanger Institute, see Andrew Bartlett, "Accomplishing Sequencing the Human Genome" (PhD dissertation, Cardiff University, Cardiff, Wales, 2008). On the Center for Genome Research at the Whitehead Institute, see Stevens, *Life Out of Sequence* (n.1), esp. chap. 3.

6. Adam Bostanci, "Sequencing Human Genomes," in *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics*, eds. Jean-Paul Gaudillière and Hans-Jörg Rheinberger (Abingdon, UK: Routledge, 2004), 158–79.

7. Sarah S. Richardson and Hallam Stevens, eds., *Postgenomics: Perspectives on Biology after the Genome* (Durham, NC: Duke University Press, 2015).

in parallel to the HGP and forged connections with Celera's sequencing strategy.[8] For now, we will note that unlike the genome centers, they did not focus primarily on submitting sequence data to databases but placed far more emphasis on exploring the biological potential of their results by publishing them in the scientific literature.

Second, the large-scale center model should not metonymically stand in for genomics as a whole, given that beyond the HGP a substantial number of concerted nonhuman genome efforts adopted other organizational regimes. This special issue devotes specific articles to the sequencing of the yeast (*Saccharomyces cerevisiae)* and pig (*Sus scrofa)* genomes, thus following Rachel Ankeny's call for multispecies, interdisciplinary studies to counter the dominance of a small number of model organisms in the historiography of genetics and genomics.[9] As well as genome centers, consortia of laboratories played substantial roles in the yeast and pig whole-genome sequencing projects. These laboratories used the sequences they produced or co-produced for the genome efforts in evolutionary, biochemical, cell biological, and agriculturally oriented research. Another difference between the HGP and the European yeast genome effort—which largely predated the formulation of the Bermuda Principles—was that the latter allowed sequence data to be held and deployed by the laboratories that had determined them for a period of time before public dissemination.[10] In this introductory essay, we will show that both yeast and pig sequencers followed the pattern of human and medical geneticists by combining sequence submission and publication. However, the institutions prominent in yeast and pig whole-genome sequencing often displayed a more even balance between the two practices than those involved in human whole-genome sequencing.

Third, in sharply demarcating sequence production, the genome center model black-boxes the notion of the sequence *user.* This figure tends to be projected into a never-accomplished future and is seldom directly addressed by

8. Miguel García-Sancho, Rhodri Leng, Gil Viry, Mark Wong, Niki Vermeulen, and James Lowe, "The Human Genome Project as a Singular Episode in the History of Genomics," this issue.

9. Rachel Ankeny, "Historiographic Reflections on Model Organisms: Or How the Mureaucracy May Be Limiting our Understanding of Contemporary Genetics and Genomics," *History and Philosophy of the Life Sciences* 32 (2010): 91–104.

10. Miguel García-Sancho, James Lowe, Gil Viry, Rhodri Leng, Mark Wong, and Niki Vermeulen, "Yeast Sequencing: 'Network' Genomics and Institutional Bridges," this issue; James Lowe, Rhodri Leng, Gil Viry, Mark Wong, Niki Vermeulen, and Miguel García-Sancho, "The Bricolage of Pig Genomics," this issue.

historical, scientific, and policy literature. Early genomic practitioners were both sequence producers and users: they contributed genomic data—and other genomic tools and resources—as well as availing themselves of that data to further their research programs in biochemistry, cell biology, and medical and agricultural genetics. Recognizing this enables us to unpack the notion of a sequence user and, at the same time, problematize its separation from sequence producers.

The history of genomics has largely been a winner's history shaped by representatives of the winners, including administrators and scientists leading the sequencing efforts, and the ethnographers studying them.[11] Throughout this special issue, we offer a more inclusive and historically sensitive conception in which genomics was not only a new field of research epitomized by the genome center and connected to—albeit different from—earlier *pregenomic* and later *postgenomic* research. Genomics brought together different communities operating outside the HGP framework—working on both human and nonhuman DNA—toward the twin objectives of characterizing genomes and solving more immediate research issues. A key way in which we have captured these other forms of practicing genomics has been by systematically reviewing all sequence submissions and looking at the entanglement between this practice and that of publishing the resulting sequence data.

2. A MIXED-METHODS APPROACH TO THE HISTORY OF GENOMICS

In this special issue, we analyze the practice of DNA sequencing over a 35-year period (1980–2015) and across three species: *S. cerevisiae* (brewer's yeast), *Homo sapiens* (human), and *S. scrofa* (pig). These three species were subject to concerted genome projects but also bespoke sequencing work to identify genes and variants linked to, among others, particular biochemical pathways, hereditary diseases, or commercially relevant traits in brewing and agriculture. Examining all three allows us to explore different incarnations of DNA sequencing beyond the HGP and the genome center model. As we will show, this comprehensive and multispecies survey constitutes the foundation of our new strategy to

11. On challenging winner's narratives of genomics, see Edna Suárez-Díaz, "Making Room for New Faces: Evolution, Genomics and the Growth of Bioinformatics," *History and Philosophy of the Life Sciences* 32 (2010): 65–89.

characterize the connections between the practice of DNA sequencing and different historical configurations of genomics research.[12]

Our central argument is that the forms of sequencing conducted by the genome centers and the regimes of data production, curation, and re-use embodied in the HGP and other genome projects did not represent the full complexity of genomics research as a whole, whether oriented to humans or nonhuman species. In our research, we sought to obtain an inclusive systematic survey of the practice of DNA sequencing across yeast, human, and pig using a mixed-methods approach. We gathered large amounts of data to comprehensively trace sequencing activity in each species. We then analyzed the resulting datasets in the light of qualitative research on the history of the institutions and scientific endeavors represented within them. Moving between the quantitative and qualitative dimensions of our work has allowed us to broaden the lens beyond that with which historians and social scientists have traditionally addressed genomics research.

We began by collecting DNA sequence submissions from the European Nucleotide Archive (ENA), a major global repository for sequence data, and linking these submissions to the first peer-reviewed journal publications describing that DNA sequence in the scientific literature. We assembled a dataset of approximately 13.5 million yeast, human, and pig DNA sequence submissions and approximately 30,000 publications in which these sequences made their first appearance in scientific journals. This approach enabled us to avoid the limitations of starting our research with the HGP, the modes of operation of large-scale genome centers, or any other pre-selected dimension of genomics that we could have spelled out in a search term or any other form of keyword-based database query.

The resulting datasets formed the evidentiary basis of three institutional co-authorship networks—one for each species—and subsequent qualitative historical work. We sought to visualize the strength of inter-institutional relationships, as measured by the frequency with which two or more institutions published a DNA sequence together in the scientific literature, and analyze the

12. On DNA sequencing as a constitutive practice of genomics research and the earlier field of molecular evolution, see García-Sancho, *Biology* (n.1); Edna Suárez-Díaz and Victor H. Anaya-Muñoz, "History, Objectivity, and the Construction of Molecular Phylogenies," *Studies in History and Philosophy of Biological and Biomedical Sciences* 39, no. 4 (2008): 451–68; Bruno J. Strasser, "Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965," *Journal of the History of Biology* 43 (2010): 623–60.

network structure formed by these relationships. Qualitatively, we pursued oral histories, archival searches, and close reading of the publications to investigate the aims and practices of the research groups involved, and whether their co-authorship ties reflected sustained collaboration. By shuttling between quantitative and qualitative analysis, we identified and interpreted clusters of institutions bound together by joint publications that contributed to genomics research despite not being necessarily focused on large-scale whole-genome sequencing.

A key principle underlying this approach is that our analysis avoids operating from an *a priori* definition of what genomics is, or from a set of representative case studies. We rather investigate how genomics—or various forms of genomics—materialized through different, coordinated, and historically evolving sequencing practices aimed at the yeast, human, and pig genomes. These practices included the HGP and other large-scale sequencing projects conducted at genome centers, but these were not the only or most visible projects present in the co-authorship networks. Work in smaller laboratories that used the sequence data they produced for their own research goals and pooled their results with other institutions to achieve a broader understanding of their target organisms was more prominently captured by our co-authorship data. This considerably expands the boundaries of genomics to incorporate sequencing practices that did not *directly* contribute to the completion of reference genomes.[13] In other words, our history of genomics is a history of connections among institutions, species, and communities, with temporal pathways being just one dimension among many. Addressing these synchronic and dynamic connections that shaped genomics and made it a diverse and continually evolving object is as important as attempting to reconstruct the continuities and discontinuities between a particular incarnation of genomics and its antecedents, or exploring the consequences of the release of fully sequenced reference genomes.[14]

13. This broadening of the historiography of genomics chimes with Gabrielle Hecht's portrayal of nuclear science through the lens of "nuclearity": a framework that includes the labor of uranium producers in the Global South and not just the prominent actors and institutions behind the *final* atomic bombs; see Gabrielle Hecht, *Being Nuclear: Africans and the Global Uranium Trade* (Cambridge, MA: MIT Press, 2012).

14. In a similar vein, but building on pre-selected case studies rather than a systematic, quantitative review, Bruno Strasser portrays genomics and other data-intensive biomedical fields today as the result of particular connections—among many adopted throughout the twentieth century—of comparative and experimental practices; see Strasser, *Collecting* (n.1), chaps. 5 and 6.

In the rest of this introductory essay, we present our datasets and methodology alongside initial analyses of the ENA sequence submissions and associated publication data. The quantitative analysis demonstrates that the large-scale sequencing centers dominated the sequence submissions, and increasingly so over time. However, there remained many institutions that, while unable to rival these centers in terms of output, consistently contributed data to genomic repositories. Some of these smaller institutions contributed quantities of sequence data to databases that were significant at the time of submission, given the productive capacities of the time; others were important publishers and analyzers of the biological significance of sequence data, as evidenced in the number of publications produced and/or co-authored.

The network visualizations we created using the co-authorship data show a complex ecosystem of institutions held together by joint publications describing DNA sequences. These networks were structured by geography and also by shared research aims and resources among the co-authoring institutions. We offer an initial comparison between the structure of the human, yeast, and pig networks, and reflect on the advantages and limitations of our approach.

The papers of this special issue mobilize our analysis of the co-authorship networks to examine in more depth the practices and configurations involved in the production and use of genomic data, beyond the well-studied HGP. In so doing we not only challenge a rigid dichotomy between sequence production and use but also propose categories that operationalize this producer–user entanglement. This leads us to propose a historiographical framework in which genomics, as well as constituting a *thin* field characterized by a new breed of genome centers (in charge of producing whole-genome sequences), was also a *thick* platform that enabled existing institutions, communities, and disciplines to converge around the shared goals of producing and using genome data.[15] In its *thick* version, genomics was co-constituted by—and contributed to new developments in—other fields such as genetics, biotechnology, immunology, biochemistry, and cell biology. Recognizing this enables us to capture the interactions of genomics with medical, agricultural, and industrial practice

15. Our proposal of a "thicker" historiography of genomics mirrors and expands the concepts of *thin* and *thick* sequencing that one of us formulated elsewhere: James W. E. Lowe, "Sequencing through Thick and Thin: Historiographical and Philosophical Implications," *Studies in History and Philosophy of Biological and Biomedical Sciences* 72 (2018): 10–27. In both thick genomics and thick sequencing, the products acquire full utility only through material and computational processes preceding and succeeding the delineation of the order of sequence bases, executed by a wide community of practitioners.

more fluidly, as we analyze the submission and publication of sequence data rather than assuming that the translation of those sequences—the use of them in scientific and practical contexts—occurred only after the determination of the reference genome.

3. DATA COLLECTION

We compiled qualitative evidence by delving into several archives, many of which were still uncatalogued or only recently open for historical research (see the complete list in the appendix at the end of this paper). These archives were either the personal records of prominent scientists—such as Alan Archibald or Lap-Chee Tsui, considered to be pioneers of pig genomics and human gene mapping, respectively—or the administrative files of concerted genomic initiatives, such as the European Commission's Yeast Genome Sequencing Project (YGSP). The evidence these archives offer is thus constrained by the administrative boundaries of each project or the selective memory and compiling strategies of the scientists and support staff who worked with them.[16]

Scholars investigating the historiography of recent science have documented the limitations of working with personal or institutional project-based archives.[17] A strategy they suggest to overcome them is using the archival materials as springboards to locate less well-known actors who may hold other records or be the sources of oral histories. Our archives pointed to a range of other genomic scientists and enabled us to gather their oral recollections and personal archives, when available. However, the best springboards for our search of historical evidence were the quantitative data that we simultaneously compiled, encompassing submissions and publications of DNA sequences.

16. The limitations of these archives resonate with the nature of genomics as a big science enterprise: unlike physics, genomics is not concentrated around centralized instrumentation but distributed in a geographically dispersed network; Niki Vermeulen, "Big Biology," *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin* 24 (2016): 195–223. Staffan Müller-Wille and Hans-Jörg Rheinberger observe that this means that new technologies or insights may therefore "emerge anywhere in the network, prove themselves locally and spread in capillary fashion." Hans-Jörg Rheinberger and Staffan Müller-Wille, *A Cultural History of Heredity* (Chicago: University of Chicago Press, 2012), 201.

17. Ronald E. Doel and Thomas Söderqvist, eds., *The Historiography of Contemporary Science, Technology, and Medicine: Writing Recent Science* (New York: Routledge, 2006); Miguel García-Sancho, "The Proactive Historian: Methodological Opportunities Presented by the New Archives Documenting Genomics," *Studies in History and Philosophy of Biological and Biomedical Sciences* 55 (2016): 70–82.

We retrieved these data from the ENA, a database founded in the 1980s as the first centralized bank of DNA sequences and housed today in the European Bioinformatics Institute, a UK-based station of the European Molecular Biology Laboratory.[18] The ENA is part of the International Nucleotide Sequence Database Collaboration, which facilitates information sharing with GenBank—a sequence submission database provided by the US National Center for Biotechnology Information—and the DNA Data Bank of Japan. Because the entries of the three databases are mirrored, users access the same sequence information regardless of which one they choose to query. Researchers, especially those who are publicly funded, have increasingly submitted the DNA sequences they determine to one of these repositories, either voluntarily or compelled by their funders or publishers.[19]

We downloaded all ENA sequence submissions for *H. sapiens*, *S. cerevisiae*, and *S. scrofa*. We chose a different time range for each species to capture sequence submissions before, during, and after the concerted projects to sequence the whole genome of each species. Given that the European YGSP started in 1989 and the reference genome of this species was completed in 1996, the *S. cerevisiae* data comprise submissions between 1980 and 2000. The *H. sapiens* dataset encompasses submissions between 1985 and 2005: the official start date of the HGP was 1990 and its conclusion was in 2003. Finally, *S. scrofa* data comprise submissions between 1990 and 2015: the first mapping projects date back to the 1990s; the first full submission of a reference genome of this species became available online in 2009, and was described in the scientific literature in 2012.

All ENA submissions display a unique identifier—an accession number—and contain information on the date of submission and DNA nucleotide length. If specified by the submitter or subsequently added by the database curators, the entries incorporate additional fields on the identity and institutional affiliation of submitting individual(s), and associated publications.[20] Our search involved over 30 million interactions with the ENA's API

18. Website: www.ebi.ac.uk/ena/browser/home. On the history of the ENA, see Miguel García-Sancho, "From Metaphor to Practices: The Introduction of 'Information Engineers' into the First DNA Sequence Database," *History and Philosophy of the Life Sciences* 74 (2011): 71–104.

19. Hallam Stevens, "Globalizing Genomics: The Origins of the International Nucleotide Sequence Database Collaboration," *Journal of the History of Biology* 51 (2018): 657–91. The tendency to submit to repositories was not without contestation or occasional reversals: see Strasser, *Collecting Experiments* (n.1), chaps. 5 and 6.

20. Weizhong Li, Andrew Cowley, Mahmut Uludag, Tamer Gur, Hamish McWilliam, Silvano Squizzato, Young Mi Park, Nicola Buso, and Rodrigo Lopez, "The EMBL-EBI Bioinformatics Web and Programmatic Tools Framework," *Nucleic Acids Research* 43 (2015): W580–84.

TABLE 1. Total ENA Sequence Submissions and Retrieved Publication Records by Species

	Accession numbers representing total sequence submissions	*Total number of sequenced and submitted nucleotides*	*Accession numbers that contain institutional submitter records (% of overall submitted nucleotides)*	*Publications retrieved describing for the first time submissions in the scientific literature (number of sequence submissions they represent; % of total)*
Yeast (1980–2000)	18,521	32,726,254	5,421 (70.18%)	2,887 (3,343; 18.05%)
Human (1985–2005)	10,091,109	21,034,707,659	2,654,378 (80.54%)	24,726 (2,582,496; 25.6%)
Pig (1990–2015)	3,322,337	18,890,916,045	2,593,209 (96.69%)	1,947 (1,435,419; 43.21%)
Total	13,431,967	39,958,349,958	5,253,008 (88.17%)	29,560 (4,021,258; 30%)

(Application Programming Interface, which allows the user to query and search the archive on a large scale) and returned a total of 13,431,967 records (see table 1). We used the R programming language and statistical environment to scrape and structure the data,[21] which comprise the list of accession numbers, dates of submission, number of nucleotides sequenced, and—if available—individual submitters, their institutional affiliation, and the first publication describing the sequences in the scientific literature.[22]

We also systematically gathered information about publications associated with sequence submissions, initially as a proxy to both compensate for the

21. R Core Team, *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing, 2016).

22. For a detailed description of our data collection strategy, including the R scripts, see Mark Wong and Rhodri Leng, "On the Design of Linked Datasets Mapping Networks of Collaboration in the Genomic Sequencing of *Saccharomyces cerevisiae*, *Homo sapiens*, and *Sus scrofa*," *F1000 Research* 8 (2019): 1200. https://doi.org/10.12688/f1000research.18656.2

patchiness of submitter details (see table 1) and to better capture the context of production of the sequence data. We used a bibliometric database, Europe PubMed Central (Europe PMC), which indexes accession numbers as metadata for publications that focus on describing and analyzing the sequence rather than merely mentioning it in the text.[23] Using the list of human, yeast, and pig accession numbers, we generated queries to the Europe PMC's API and yielded a corpus of publications in which the submitted sequences—both with and without an identifiable submitter—appeared as indexed metadata.

Among all the publications associated with a given accession number, we decided to select only the chronologically earliest record and compiled information about all the co-authors and their affiliated institutions. This is because we wanted to identify the scientists and institutions who had either collaborated in the determination of the sequence or in the first discussion about its potential use in their lines of research. Our strategy excluded publications that reported updates of a particular sequence or discussed later investigations. Using two bibliometric databases proved necessary, as while Europe PMC allows searches for publications with indexed accession numbers, it only holds information of the corresponding author for articles published before 2014.[24] The SCOPUS citation database holds bibliometric records of all authors and their institutions, particularly for biomedical and natural science literature.[25] However, unlike Europe PMC, SCOPUS does not index literature by sequence accession number. Therefore, we used the publications' PubMed IDs (PMIDs) that we retrieved from Europe PMC to extract data from SCOPUS on all authors, their institutional affiliations, the city and country of institution, and the date of publication for our corpus of first sequence-reporting

23. Rodrigo Lopez, Andrew Cowley, Weizhong Li, and Hamish McWilliam, "Using EMBL-EBI Services via Web Interface and Programmatically via Web Services," *Current Protocols in Bioinformatics* 48, no. 1 (2014): 3.12.1–50. The parallel with the use of the term "description" in natural history in connection with the reporting of new species is deliberate, and is an actor's category. The ENA distinguishes between papers describing the sequence and those citing it. While Europe PMC takes this distinction into account when indexing accession numbers contained in an article, other bibliometric databases such as SCOPUS do not make this differentiation (see below).

24. The Europe PMC Consortium, "Europe PMC: A Full-Text Literature Database for the Life Sciences and Platform for Innovation," *Nucleic Acids Research* 43 (2015): D1042–48.

25. Daniele Rotolo and Loet Leydesdorff, "Matching Medline/PubMed data with Web of Science: A Routine in R Language," *Journal of the Association for Information Science and Technology* 66, no. 10 (2015): 2155–59.

articles.[26] When we refer to a publication in this dataset throughout the special issue, we signal it in the following way: "in our dataset: PMID [number]."[27]

4. ANALYSIS OF SUBMISSION AND PUBLICATION DATA: DIVERGING STORIES AND GENEALOGIES

When comparing the submission and publication records in the quantitative datasets, one notices various kinds of asymmetries. First, the sequence submissions are substantially larger than the publications. This is explained, in part, by the fact that only 30% of the accession numbers—the overall DNA sequence submissions—were described in the scientific literature when considering our three species. This finding is not surprising in itself as it relates to changing policies in scientific journals.[28] The submission of several thousand nucleotides in the 1980s involved considerably more cost and time than it would by the 2000s. At first, researchers in charge of the sequence determination could publish their sequences as peer-reviewed journal articles without much additional embellishment in journals such as *Nucleic Acids Research* and *Genomics*. As the ability to sequence improved and concerted projects developed, the value of producing a given amount of sequence declined over time. Increasingly, publishers demanded the description of larger volumes of sequence—whole chromosomes and whole genomes—or the addition of

26. We have detected some differences between the total numbers of publications represented in our datasets and those in the individual repositories from which we extracted the submission and publication records. This is likely the result of the cleaning and triangulation strategies through which we constructed our datasets by combining records from three repositories: the ENA, Europe PMC, and SCOPUS. See Wong and Leng, "On the Design" (n.22) for a full description of the datasets and an outline of their construction. On the discrepancies between and absences within SCOPUS and PubMed, see Cynthia M. Schmidt, Roxanne Cox, Alissa V. Fial, Teresa L. Hartman, and Martha L. Magee, "Gaps in Affiliation Indexing in Scopus and PubMed," *Journal of the Medical Library Association* 104, no. 2 (2016): 138–42.

27. The datasets are available without restrictions at https://datashare.is.ed.ac.uk/handle/10283/3517. They provide the names, institutional affiliations, and countries for all co-authors associated with a PMID, as well as the date on which the paper was published according to SCOPUS and the date on which the sequence it describes was submitted to the ENA. PubMed and other bibliometric databases retrieve full publication details from PMIDs. We also provide details of all institutions submitting to the ENA, number of nucleotides sequenced, and year of submission to the database in aggregate form in the human and pig datasets, and sequence per sequence in the yeast dataset.

28. Strasser, *Collecting Experiments*, (n.1), 214, 232–35. See also Hilgartner, *Reordering Life* (n.2), chap. 6; Stevens, *Life Out of Sequence* (n.1), 58–60.

a more extensive analytical and interpretive framework in which authors discussed the actual or potential use of the sequence in the article.

Our data, however, indicate that an increasing proportion of sequence submissions were reported in the scientific literature over time. For yeast (covering 1980–2000), only about 18% of accession numbers are linked to publications, for human (1985–2005) about 26%, and for pig (1990–2015) about 43%. How can this be the case? One answer is that accession numbers were increasingly reported in a smaller number of publications. The yeast publications in our dataset described an average of 1.2 accession numbers per paper, while human publications had an average of 104.5 accession numbers, and pig publications reported an average of 737 accession numbers, something consistent with the journals' requirement of publishing beyond individual sequence descriptions as determining those sequences became more standard practice. More generally, the asymmetries between our submission and publication data questioned our initial assumption that publications would be one-to-one proxies of submissions and we could just use the publications to compensate for the sparsity of information in the ENA about individual or institutional submitters. In what follows, we zoom into the detail of those asymmetries and show the potential of the publication data to illuminate aspects of the history of genomics that the submissions render invisible.

4.1. Submission Analysis

The ENA dataset provides information on trends in the production of DNA sequences, as well as allowing us to gauge the scale of contribution by specific sequencers. In figure 1, we see a dramatic difference in the number of nucleotides sequenced for each species during the time periods we studied. From 1980–2000, sequence submissions to the ENA totaled c32.7 million nucleotides for *S. cerevisiae*. The amount submitted for *H. sapiens* was c21 billion nucleotides from 1985–2005, and c18.9 billion nucleotides were submitted for *S. scrofa* from 1990–2015. The evolution of submitted sequences presents a similar pattern in each species, with a steady increase in the volume during the concerted projects to determine the yeast, human, and pig genomes—especially in the years before the publication of the reference sequences—and a plateau afterward.

For each of these species, we found a heavily skewed distribution of sequence production, as gauged by the number of sequenced nucleotides contributed by the top ten submitters. For human submissions, ten institutions were

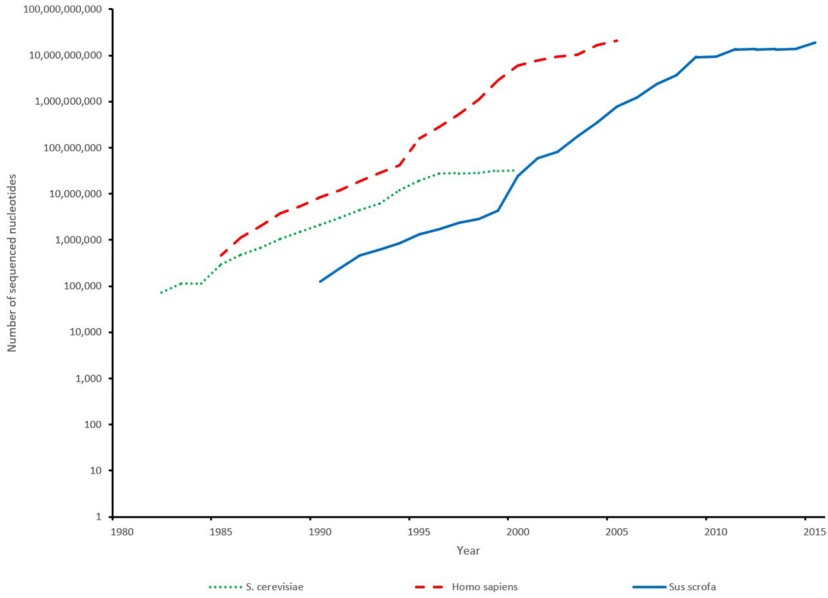

FIGURE 1. Cumulative growth of sequenced nucleotides by year of submission to the European Nucleotide Archive (ENA) for *S. cerevisiae*, *H. sapiens*, and *S. scrofa*. Figure elaborated by the authors.

responsible for over 91% of all nucleotides with submitter details recorded in the ENA during the whole period. In the yeast dataset, the top ten institutions submitted over 88% of all nucleotides with submitter data, while in pig this percentage rose to almost 96%. As reflected in tables 2 to 4, a substantial proportion of these top submitters were either genome centers or institutions that incorporated genome centers to their departmental structure in the 1990s. This chronology and the evolution of sequence production shown in figure 1 suggest that the contributions of those top submitters were concentrated on the second part of the periods we explore in the human, yeast, and pig datasets.

The contributions of the major submitters clearly dwarfed all others. Considering this, it is easy to see why historians and other scholars have focused primarily on these major producers of sequence data. Institutions like the Sanger Institute, the Beijing Genomics Institute, and Celera Genomics represented a new form of organization of scientific work that merited scholarly attention. Indeed, we are *not* arguing that these institutions are historically uninteresting, or that the vast majority of sequencing work occurred outside such centers, if sequencing work is simply equated with the number of nucleotides sequenced and submitted to databases.

TABLE 2. Top Ten Submitters to the ENA of Nucleotides of Human DNA, 1985–2005

Rank	*Institution*	*% of Total Submitted Nucleotides*
1	Celera Genomics, US	50.46
2	Whitehead Institute for Biomedical Research, US	10.27
3	Wellcome Trust Sanger Centre, UK	7.98
4	Washington University School of Medicine St. Louis, US	5.39
5	DOE Joint Genome Institute, US	4.51
6	Kazusa DNA Research Institute, Japan	4.34
7	Baylor College of Medicine, US	2.76
8	Genoscope—Centre National de Séquençage, France	2.18
9	RIKEN—The Institute of Physical and Chemical Research, Japan	2.07
10	University of Washington, US	1.31
Top 10 % of total submitted nucleotides		***91.26***
Total nucleotides in all sequence data with submitter data		***16942665389***

TABLE 3. Top Ten Submitters to the ENA of Nucleotides of Yeast DNA, 1980–2000

Rank	*Institution*	*% of Total Submitted Nucleotides*
1	MIPS at Max-Planck-Institut für Biochemie, Germany	38.67
2	Sanger Centre, UK	16.19
3	Stanford University, US	11.88
4	Washington University in St. Louis, US	11.07
5	Université Catholique de Louvain, Belgium	3.92
6	McGill University, US	2.13
7	European Molecular Biology Laboratory, Germany	1.46
8	RIKEN—The Institute of Physical and Chemical Research, Japan	1.26
9	Ludwig-Maximilians-Universität München, Germany	1.15
10	Institut Curie Research Centre in Orsay, France	0.95
Top 10 % of total submitted nucleotides		88.68
Total nucleotides in all sequence data with submitter data		***22967465***

TABLE 4. Top Ten Submitters to the ENA of Nucleotides of Pig DNA, 1990–2015

Rank	*Institution*	*% of Total Submitted Nucleotides*
1	Genome Analysis Centre, Norwich Research Park, UK	36.24
2	Wellcome Trust Sanger Institute, UK	31.38
3	Beijing Genomics Institute, China	12.93
4	Novogene Bioinformatics Institute, China	12.85
5	National Institute of Agrobiological Sciences, Japan	1.42
6	NIH Intramural Sequencing Center, US	0.47
7	Genoscope—Centre National de Séquençage, France.	0.21
8	Roslin Institute, UK	0.16
9	Institut National de la Recherche Agronomique Castanet-Tolosan, France	0.09
10	National Institute of Animal Science, South Korea	0.08
Top 10 % of total submitted nucleotides		95.85
Total nucleotides in all sequence data with submitter data		***18265626724***

There is, however, another way of approaching sequencing work historically. Beyond the prolific sequence submitters, we observed a large number of institutions in all three species that consistently contributed sequence data to archives, despite contributing only a fraction of the total over the whole time period. Some of these institutions made considerable contributions of nucleotides in the early years of the periods we have used, which were masked by the more prodigious quantities pumped out later; examples include Harvard Medical School in the human dataset, the European Molecular Biology Laboratory in yeast, and the Meat Animal Research Center of the US Department of Agriculture in pig. Furthermore, not only did these institutions submit sequence data, and continue to do so, but many published scientific papers in journals describing both their sequences and the utility (actual or potential) of the data for biochemical, medical, agricultural, or other forms of research. These publications thus reflect an entanglement of sequence production and use (actual or intended) that occurred much more sporadically and decreasingly within the genome center model.

4.2. Publication Analysis

The publication data are less skewed than the submission data. For human, the top ten publishing institutions co-authored only 12.9% (3,191) of the 24,726

journal articles describing for the first time a particular sequence in the literature. For yeast, this percentage was 20.3% (585) of the 2,887 publications, and for pig 23.0% (448) of the 1,947 publications—in yeast, we considered the top fourteen publishing institutions due to ties in our top ten table.[29]

Overall, there was little overlap among institutions that were the most prolific sequence publishers and those that were the most prolific submitters for each species: three institutions in the human leaderboard tables, three in the yeast tables, and two in the pig tables—eight institutions of a total of thirty-four, considering the four extra publishers of yeast sequences. As tables 5 to 7 show, some of the overlapping institutions—Baylor College of Medicine, Washington University, and Stanford University—are large universities or research institutes that incorporated genome centers, so a substantial number of the co-authors of the publications were based in different departments than the top submitters.

TABLE 5. Top Ten Publishers of Papers Describing Human DNA Sequences for the First Time in the Literature, 1985–2005*

Rank	*Institution*	*Number of papers authored (% of total in the dataset)*
1	Harvard University Medical School, US	453 (1.83)
2	National Cancer Institute Bethesda, US	406 (1.64)
3	INSERM, France	362 (1.46)
4	Baylor College of Medicine, US	350 (1.42)
5	University of California San Francisco, US	325 (1.31)
6	Washington University in St. Louis School of Medicine, US	299 (1.21)
7	University of Washington, US	298 (1.21)
8	University of Tokyo, Japan	288 (1.16)
9	National Institutes of Health Bethesda, US	264 (1.07)
10	Massachusetts General Hospital, US	261 (1.06)

*We have highlighted the institutions that are also leading submitters in table 2.

29. These figures are all lower than the sum of the numbers and percentages of papers authored in the respective tables 5, 6, and 7, as some publications were authored by more than one institution represented in each table.

TABLE 6. Top Fourteen Publishers of Papers Describing Yeast DNA Sequences for the First Time in the Literature, 1980–2000*

Rank	*Institution*	*Number of papers authored (% of total in the dataset)*
1	Massachusetts Institute of Technology, US	77 (2.67)
2	University of California Berkeley, US	69 (2.39)
3	University of California San Francisco, US	57 (1.97)
4	University of Tokyo, Japan	48 (1.66)
5	Ludwig-Maximilians-Universität München, Germany	42 (1.45)
6	Harvard University Medical School, US	39 (1.35)
7	University of Washington, US	38 (1.32)
8	Columbia University in the City of New York, US	36 (1.25)
=	Yale University, US	36 (1.25)
10	Centre National de la Recherche Scientifique, France	33 (1,14)
=	Cornell University, US	33 (1.14)
=	Cold Spring Harbor Laboratory, US	33 (1.14)
=	Stanford University, US	33 (1.14)
=	European Molecular Biology Laboratory, Germany	33 (1.14)

*We have highlighted the institutions that are also leading submitters in table 3.

TABLE 7. Top Ten Publishers of Papers Describing Pig DNA Sequences for the First Time in the Literature, 1990–2015*

Rank	*Institutions*	*Number of papers authored (% of total in the dataset)*
1	Huazhong Agricultural University, China	132 (6.78)
2	USDA ARS Meat Animal Research Center, US	53 (2.72)
3	Institut National de la Recherche Agronomique Castanet-Tolosan, France	52 (2.67)
4	Iowa State University, US	47 (2.41)

(continued)

TABLE 7. *(continued)*

Rank	Institutions	Number of papers authored (% of total in the dataset)
5	China Agricultural University, China	41 (2.11)
6	Kobenhavns Universitet, Denmark	38 (1.95)
7	Chinese Academy of Agricultural Sciences, China	37 (1.90)
8	National Institute of Agrobiological Sciences, Japan	37 (1.90)
9	Universität Göttingen, Germany	37 (1.90)
10	Institute of Animal Physiology and Genetics of the Academy of Sciences of the Czech Republic, Czech Republic	34 (1.75)

*We have highlighted the institutions that are also leading submitters in table 4.

This disjunction between the leading submitters and publishers prompts the question of what the frequently publishing but less-prolific submitters were doing, what they were trying to achieve, and how this was different from the aims and outputs of the large-scale genome sequencing centers and projects. Publishing institutions engaged in a form of sequencing that was not large scale, but rather focused on describing each sequence submission in the literature—as opposed to the genome centers that would typically only publish whole chromosomes or whole genomes, formed by multiple submissions to databases. We had first sought publication data in order to identify likely contributors to sequencing work that were not listed formally in the ENA submissions—the majority of ENA entries did not include submitter details (see table 1). Yet many of the publications linked to particular accession numbers included multiple co-authors. Many of these co-authors were not involved in the determination of the sequence but performed other key tasks concerning the resulting data. These authors were often from different institutions to the submitters.

This suggested that the publications, rather than being a proxy of the submissions, told a different story; a story of the entanglement of the *production* and *use* of sequence data.[30] That is, the co-authorship relationships

30. This accords with scientometric research that demonstrates different collaborative dynamics between sequence submitters and publishers using metadata extracted from GenBank:

underpinning the articles reflected a connection between sequence production and the mobilization of the data in peer-reviewed scientific journal publications—and this story did not fit the familiar narrative of genomics, a narrative dominated by large-scale sequence producers. This was instead a more complex story, one with a diversity of co-authoring institutions pursuing different research goals. The details of this story are the subject of the following three articles of the special issue. Here, though, we will show the general trends by comparing the human, yeast, and pig co-authorship networks into which we transformed our publication data.

5. GENOMIC COLLABORATION AMONG INSTITUTIONS EXPLORED THROUGH CO-AUTHORSHIP NETWORK ANALYSIS

To better understand the collaborative relationships among institutions that published analyses of DNA sequences, we transformed our publication datasets into institution-to-institution co-authorship networks. The rationale behind this was that, while large-scale sequencing centers had ample resources and coordinated their activity through official genome projects, smaller institutions needed to gather these resources and collaborate outside of these projects. One way of doing this *might* have been to pool resources in small groups of collaborating institutions that sat outside of the major sequencing initiatives and sought to publish their analyses, as well as submitting their sequences to the ENA. Alternatively, institutions outside of the formal genome projects may have participated by working alongside institutions that were themselves involved—the genome centers—and crediting their contributions in co-authored publications. We used co-authorship network analysis to visualize and identify which institutions collaborated in the writing of papers that first described specific sequences in the literature. We then used these networks to identify links among institutions in order to guide qualitative investigations into *why* those institutions were involved, and in what capacity, in the description of the sequences.

Mark R. Costa, Jian Qin, and Sarah Bratt, "Emergence of Collaboration Networks around Large Scale Data Repositories: A Study of the Genomics Community Using GenBank," *Scientometrics* 108 (2016): 21–40; on comparing these with metadata on patents as well, see Jeff Hemsley, Jian Qin, and Sarah E. Bratt, "Data to Knowledge in Action: A Longitudinal Analysis of GenBank Metadata," *Proceedings of the Association for Information Science and Technology* 57, no. 1 (2020): e253.

5.1. The Potential and Limitations of Our Approach

Scholars have extensively used co-authorship network analysis to study scientific collaboration between individual researchers or research institutions in different disciplines.[31] The core idea of this network approach is that scientific collaboration is analyzed in terms of a proxy indicator: patterns of co-authorship relationships, such as the network size, structure, or composition. For example, analyzing co-authorship in sociological publications from the 1960s to the 1990s, James Moody showed that authors who wrote in historical, qualitative, and interpretive subfields published more often alone and formed a more fragmented scientific network than those writing in experimental and quantitative subdisciplines.[32]

Our network approach takes *direct* and *indirect* social connections as the fundamental unit of analysis, and aims to examine the structure of the interconnections between actors (e.g., centralized vs. decentralized, more or less highly clustered into communities), based on the assumption that this structure matters. There are limitations to this approach, particularly regarding the use of institutional co-authorship as a proxy of collaboration. As quantitative social science literature has shown, authors may decide to co-publish a paper or to collaborate for many reasons: collaboration need not result in co-authorship, and co-authorship may not result from prior collaboration.[33] Both co-authorship and collaboration are complex and multifaceted processes and do not stand in simple relation to each other.

Distinguishing between co-authorship and collaboration was challenging in our dataset and networks. Each publication just represented a number of

31. Sameer Kumar, "Co-Authorship Networks: A Review of the Literature," *Aslib Journal of Information Management* 67, no. 1 (2015): 55–73; Mark E. J. Newman, "Coauthorship Networks and Patterns of Scientific Collaboration," *Proceedings of the National Academy of Sciences* 101, Supplement 1 (2004): 5200–5.

32. James Moody, "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999," *American Sociological Review* 69, no. 2 (2004): 213–28.

33. J. Sylvan Katz and Ben R. Martin, "What Is Research Collaboration?," *Research Policy* 26 (1997): 1–18; see also Wolfgang Glänzel and András Schubert, "Analysing Scientific Networks Through Co-Authorship," in *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, eds. Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch (Dordrecht, Netherlands: Springer: 2004): 257–76; Terttu Luukkonen, Robert J. W. Tijssen, Olle Persson, Gunnar Sivertsen, "The Measurement of International Scientific Collaboration," *Scientometrics* 28 (1993): 15–36; Diane H. Sonnenwald, "Scientific Collaboration," *Annual Review of Information Science and Technology* 41 (2007): 643–81.

institutions having co-authored an article that described a DNA sequence for the first time. These co-authorships ties may have occurred occasionally or reflect intense collaborative links that pursued some use of the sequence data, and may or may not have translated into additional publications. This meant that supplementary—and often qualitative—information from beyond the networks and data on the publications was essential to appreciate the nature and significance of co-authorship relationships. Like other scholars using mixed-methods, we found qualitative evidence and expertise crucial to making our datasets and visualizations sociologically and historically meaningful. Yet these quantitative and network data provided evidence that could not have been obtained through standard historical research tools.[34]

There were two additional limitations with our data. The first was that in order to achieve a readable network size, we needed to remove departmental affiliations from the co-authoring institutions and label them according to the higher level of organization as universities, hospitals, research institutes, or companies.[35] Because of this, publications written by one or several scientists all based in the same institution did not display as co-authorship relationships in our networks. These publications represented about half of the human and pig datasets (52% of the human articles; 47% of the pig articles), and about three-quarters of the yeast dataset (76% of articles). We did, however, consider single-author and single-institutional publications when computing the overall number of sequence-reporting articles per institution in tables 5, 6, and 7, and retrieve them from our dataset in qualitative analyses.

Second, in line with other scientometric studies,[36] different fields and countries in our networks showed their own specific levels of co-authorship

34. Deryc T. Painter, Bryan C. Daniels, and Jürgen Jost, "Network Analysis for the Digital Humanities: Principles, Problems, Extensions," *Isis* 110 (2019): 538–54; Manfred D. Laubichler, Jane Maienschein, and Jürgen Renn, "Computational History of Knowledge: Challenges and Opportunities," *Isis* 110 (2019): 502–12.

35. The exception to this is medical schools, which we labeled separately from any universities they are part of, given that their scientific and administrative independence tends to be stronger than in any other department or faculty, especially in the United States. The levels of organization that we chose—medical schools, universities, companies, hospitals, and research institutes—represent more stable entities than departments, which tend to exhibit considerable change in name and structure over decades.

36. Moody, "The Structure" (n.32); Newman, "Coauthorship Networks" (n.31); David Pontille, "Authorship Practices and Institutional Contexts in Sociology: Elements for a Comparison of the United States and France," *Science, Technology, & Human Values* 28, no. 2 (2003): 217–43.

and structure of scientific collaboration. More generally, co-authorship patterns shifted over time and shared a general trend toward an increasing number of authors of scientific papers, a continued decline in the proportion of papers published by single authors, and a general increase in international co-authorships.[37] This meant that it was difficult to use our publication data to directly compare human, yeast, and pig genomics, as the articles covered different time periods, exhibited different disciplinary make-ups, and involved different communities with distinct publication and collaboration cultures, and moral economies.[38] There were also intimations of different geographical patterns in the co-authorship data across the three species.

We have therefore been careful in drawing direct comparisons from the network data alone. Instead, we have sought to interpret the differences and similarities across the three species from a deep engagement with the particularities of genomics research concerning each species, and relating the analytical distinctions and concepts we generated through this to each other, rather than directly comparing the networks and network data alone. In doing so, however, we have had to be attentive to avoid separating our characterization into species-centered silos based on differential data collection between them. In the following section, we offer an initial analysis of the networks within and across our three species, each of which are explored more deeply in the rest of the special issue.

5.2. Network Visualization and Initial Explorations

The networks are composed of a set of institutions represented by nodes (circles) and a set of co-authorship relationships represented by edges (lines between the circles). If an institution is associated with more than one publication with another institution, we reflect this in the edge weight (or edge value): an edge weight of seven between two institutions means that institution *A* has co-authored seven publications with institution *B*. Our method counts unique institutions associated with a unique publication only once. Thus, if a publication has four authors, with three from a single institution and another

37. Wolfgang Glänzel, "Coauthorship Patterns and Trends in the Sciences (1980–1998): A Bibliometric Study with Implications for Database Indexing and Search Strategies," *Library Trends* 50 (2002): 461–73; Terttu Luukkonen, Olle Persson, and Gunnar Sivertsen, "Understanding Patterns of International Scientific Collaboration," *Science, Technology, & Human Values* 17 (1992): 101–26.

38. On moral economies in science, see Lorraine Daston, "The Moral Economy of Science," *Osiris* 10 (1995): 2–24; Robert E. Kohler, *Lords of the Fly: Drosophila Genetics and the Experimental Life* (Chicago: University of Chicago Press, 1994).

from a second institution, in the network this will result in an edge weight of one between these two institutions.

Using Gephi as network visualization and analysis software (version 0.9.2), figures 2 to 4 show the main component of the co-authorship networks for human, yeast, and pig sequence publications. The main component is the largest connected subnetwork—that is, the biggest subset of nodes connected through co-authorship paths consisting of one or more edges. Co-authorship network analysis often focuses on the main component and disregards smaller components composed of more marginal institutions, including isolated institutions that did not publish with any other institutions.

In what follows, we quantitatively and visually analyze each network and interpret them according to the existing literature and knowledge on the history of human, yeast, and pig genomics. We then calculate some network indices and synoptically compare them in order to push our interpretation and pave the way to the more detailed study we offer in the next three articles of the special issue, specifically devoted to the analysis of the human, yeast, and pig networks. The node size in the figures represents the number of publications that the institutions co-authored with other institutions (their weighted degree); the node color denotes the home country of an institution. We used ForceAtlas2, a force-based network layout algorithm in Gephi, to produce an appropriate and comprehensible visualization by adjusting the different layout settings. In particular, the attraction-repulsion strength settings of the algorithm determine how strongly it pulls closer together connected nodes directly or indirectly (the latter through edges with common neighbors) and forces apart distantly connected nodes.[39]

Figure 2 depicts the main component of the human co-authorship network. It comprises 5,573 institutions (93% of the total when considering the whole network) and 39,448 co-authorship relationships (99.7% of the total). The average number of publications per connected pair is 1.24; in about 87% of co-authorship relationships there is only one underlying publication. Although the proportion of US institutions is lower than in yeast, and European institutions constitute the largest continental group (see table 8), the most connected institutions in the human network forming a densely connected cluster at its

39. Repeating the process from the same starting point will not result in the same visualization, as there is a random element in Gephi and other network display algorithms and software. For commentary on the caution required when interpreting network visualizations, see these two blogposts of Mathieu Jacomy, one of the main architects of Gephi's algorithms: "The Problem with Network Maps," https://reticular.hypotheses.org/1724 and "Is Gephi a Black Box?," https://reticular.hypotheses.org/976.

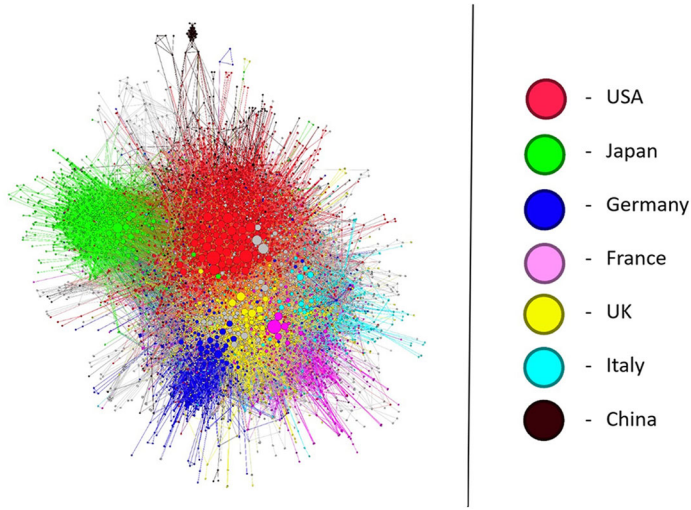

FIGURE 2. Main component of the human co-authorship network. The node size corresponds to the number of publications that the represented institution co-authored with other institutions (weighted degree). We colored the nodes according to the home country of the represented institution, as indicated in the legend on the right side of the figure; we colored the rest of the nodes in gray. Figure elaborated by the authors.

TABLE 8. Distribution of Institutions by Continents and Leading Countries: Human

Continent	*% of institutions*	*Country*	*% of institutions*
Europe	42.5	US	27.5
North America	30.5	Japan	12.8
Asia	21.0	Germany	8.2
Oceania	2.7	France	7.7
Africa	0.8	UK	6.8
Latin America	1.8	Italy	5.2
		China	3.0

core are from the United States. Germany, in spite of national qualms over human genetics research due to its eugenic past, is also a player of perhaps unexpected significance in the network.[40] Indeed, grouping the nodes into

40. This is also the case for the yeast network and suggests that historiographical emphasis on the consequences of Nazism on German science may have masked the continued impetus and importance of genetics research in this country; see García-Sancho et al., "Yeast Sequencing" (n.10).

more or less national clusters is far more evident in the human network than in the yeast or pig networks.

A possible explanation of this strong national clustering may be the initial forming and continued strength of human genome programs in many countries, especially European, Asian, and North American. As the literature has shown, the HGP was never a fully international endeavor from an administrative viewpoint, emerging from the convergence of government and charitably funded schemes that originated nationally in the early to mid-1980s. Most of these schemes, in their beginnings, focused on specific genetic diseases and involved the formation of consortia of institutions, either for research and regulatory convenience or the local importance of a particular condition.[41] Through presenting this pattern of national clustering, the network encourages us to direct attention to this rather overlooked initial stage of human genomics.[42]

The diagram in figure 3 shows the main component of the yeast co-authorship network. It comprises 590 institutions (71% of the total) and 1,959 co-authorship relationships (98% of the total). The average number of publications per connected pair is 1.07, so most institutions have co-authored only one publication with partner institutions. The percentages of nodes by country and continent (table 9) reflect that European institutions are the most numerous. However, as shown in our prior analysis of the publication dataset (see table 6, above), eleven of the top fourteen publishers of yeast sequences are institutions from Japan and the United States. These publishers have more often co-authored within their own institution compared with the European publishers, which have more frequently co-authored with other institutions and thus produced visible edges in the network.[43]

41. Michael Fortun, "Mapping and Making Genes and Histories: The Genomics Project in the United States, 1980–1990" (PhD dissertation, Harvard University, Cambridge, Massachusetts, 1993); Peter Glasner and Harry Rothman, "Does Familiarity Breed Concern? Bench Scientists and the Human Genome Mapping Project," *Science and Public Policy* 26, no. 4 (1999), 233–40; Vololona Rabeharisoa and Michel Callon, "Patients and Scientists in French Muscular Dystrophy Research," in *States of Knowledge: The Co-Production of Science and Social Order*, ed. Sheila Jasanoff (New York: Routledge, 2004), 142–60.

42. We detail the role of medical geneticists during the early years of genomics and the interactions between whole-genome initiatives and sequencing work targeted to specific genes or conditions in García-Sancho et al., "The Human Genome Project" (n.8).

43. The Japanese and US co-authorship pattern was not unusual in comparison to other genomic or life science research in the 1980s and 1990s. Levels of inter-institutional and international co-authorship had increased prior to and throughout the period covered by the yeast

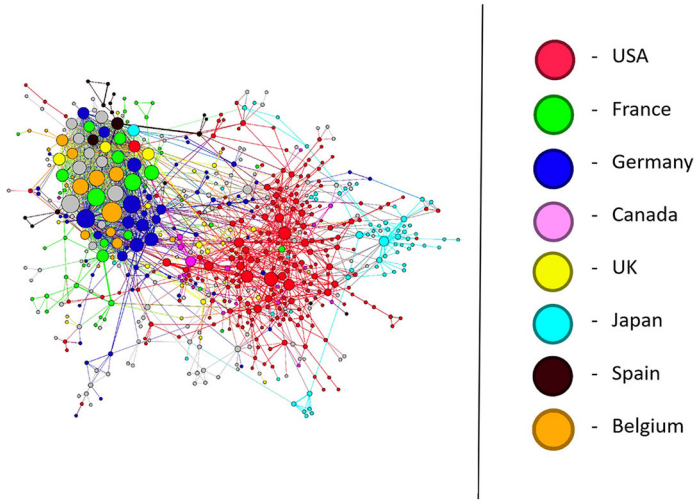

FIGURE 3. Main component of the yeast co-authorship network. The node size corresponds to the number of publications that the represented institution co-authored with other institutions (weighted degree). We colored the nodes according to the home country of the represented institution, as indicated in the legend on the right side of the figure; we colored the rest of the nodes in gray. Figure elaborated by the authors.

TABLE 9. Distribution of Institutions by Continents and Leading Countries: Yeast

Continent	*% of institutions*	*Country*	*% of institutions*
Europe	42.5	US	36.9
North America	40.3	Japan	10.3
Asia	12.8	Germany	10.3
Oceania	2.2	France	6.8
Latin America	1.0	UK	4.7
Middle East	0.7	Canada	3.4
Africa	0.3	Belgium and Spain	2.9 each

The European institutions embody the mode of collaboration spurred by the YGSP. This project was funded by the European Commission between 1989 and 1996, and involved the formation of a consortium that included

network data, but this had not yet risen to the levels visible later, for instance in the human and pig networks. At the time, the co-authorship pattern represented by the European yeast sequencers was the outlier. See Glänzel, "Coauthorship Patterns" (n.37).

institutions from nearly every country in the European Economic Community (subsequently, the European Union). The aim was to systematically sequence the whole yeast genome.[44] The organization of the YGSP provides a strong continental clustering to this network: connections between Europe, North America, and Asia are markedly more tenuous than in the human and pig networks. This makes the study of the bridging institutions that co-authored across continents particularly appealing.[45]

Figure 4 presents the main component of the pig co-authorship network. It comprises 1,021 institutions (80% of the total) and 3,196 co-authorship relationships (97% of the total). The average number of publications per connected pair is 1.18; in about 90% of co-authorship relationships there is only one underlying publication. French and US institutions occupy the most central positions, but unlike in the yeast and especially the human network, the North American nodes tend to be more scattered and less structured in bounded clusters. There is also a stronger presence of Asian institutions, with three clusters of densely connected Chinese, Japanese, and South Korean nodes in the periphery of the network. Overall, Asian institutions rank second in the pig network, above North American and below European nodes (see table 10). This is due to relative Japanese strength in the 1990s persisting post-2000, and the capturing of the rapidly increasing Chinese sequencing and publishing output in the 2000s, which the other datasets either mostly or fully miss.

In spite of their strong presence, Asian institutions occupy peripheral positions in the pig network, as in the human and yeast ones. This is because of high levels of intra-country collaborations, which we explain by observing the lack of a transnational body like the European Commission or transnational funding policies like those promoted by the United States Department of Agriculture (USDA, a major sponsor of pig genomics). Funds from USDA grants could move to institutions in other countries, enabling international collaboration and partly accounting for the lack of nationally bounded US clusters in the network. Furthermore, the significance of research directed

44. Giuditta Parolini, *Building Human and Industrial Capacity in European Biotechnology: The Yeast Genome Sequencing Project (1989–1996)* (Berlin: Technische Universität Berlin, 2018). https://depositonce.tu-berlin.de/bitstream/11303/7470/4/parolini_guiditta.pdf

45. For our full analysis of these institutions, their qualitative significance, and their historiographical import, see García-Sancho et al., "Yeast Sequencing" (n.10).

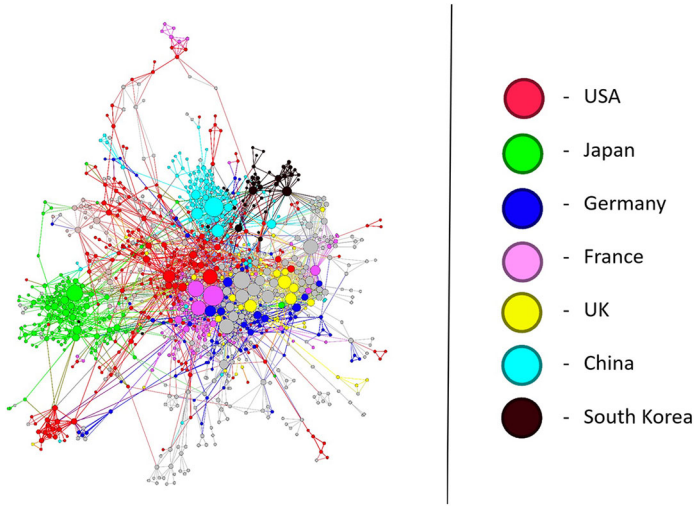

FIGURE 4. Main component of the pig co-authorship network. The node size corresponds to the number of publications that the represented institution co-authored with other institutions (weighted degree). We colored the nodes according to the home country of the represented institution, as indicated in the legend on the right side of the figure; we colored the rest of the nodes in gray. Figure elaborated by the authors.

TABLE 10. Distribution of Institutions by Continents and Leading Countries: Pig

Continent	*% of institutions*	*Country*	*% of institutions*
Europe	38.3	US	21.9
Asia	31.3	Japan	12.7
North America	25.6	China	10.6
Oceania	2.4	Germany	7.7
Africa	1.2	France	6.8
Latin America	1.3	UK	4.3
		South Korea	3.9

toward the problems of animal breeding in pig genomics meant that the genetically distinctive breeds that were domestically important in particular Asian countries were not so important for European and North American markets, and therefore did not spark an interest among breeding companies and publicly funded institutions working on pig genetics there. There were

exceptions to this, but not enough to overcome the isolation of the country-based clusters for China, Japan, and South Korea in our network.[46]

To compare the structure of the yeast, human, and pig main components, we calculated some network indices (table 11) and synoptically analyzed them.

Despite having, by far, the largest number of nodes and connections—as reflected by the average weighted and unweighted degrees in table 11—the human network has the smallest diameter and average path length. Diameter is the length of the longest possible path of a network, with a path measured as the number of edges separating any two nodes. The diameter of the human network is nine, while the yeast and pig networks have more chain-like structures beyond their core clusters that translate into diameters of eleven and twelve, respectively (see figures 3 and 4 above).

TABLE 11. Network Parameters of the Main Components of the Yeast, Human, and Pig Co-authorship Networks

Network parameters	***Yeast***	***Human***	***Pig***
Average unweighted degree *Average number of cross-institutional edges per institution*	6.6	14.2	6.3
Average weighted degree *Average number of cross-institutional publications per institution*	7.1	17.7	7.4
Network degree centralization (unweighted) (%) *Variations in the number of cross-institutional edges (degrees) across institutions*	8.4	7.8	5.6
Network density (unweighted) (‰) *Proportion of cross-institutional edges compared to all possible edges*	11	2.6	6.1
Clustering coefficient (%) *Probability that two randomly chosen institutions will co-author if they have both co-authored with another third institution*	58	14	35
Diameter *Longest possible distance between two institutions*	11	9	12
Average path length *Average distance between the institutions*	4.2	3.3	4.3

46. On the importance of the materiality of pig breeds and pig circulation as a structuring factor in the network—not only among Asian populations but also concerning the Spanish Iberian pig—see our detailed analysis in Lowe et al., "The Bricolage" (n.10).

We interpret this with reference to the significance of institutions that act as highly collaborative hubs in the human network, as they channel co-authorship ties from a large number of other institutions. The collaborative hubs present institutional variety—including universities, medical schools, and hospitals—and feature some of the largest sequence publishers, among them the Harvard Medical School and the Massachusetts General Hospital. Their ability to link disparate parts of the network in a relatively small number of steps accounts for how a network that is huge in its appearance can nonetheless have a shorter diameter and average path length than the much smaller yeast and pig networks. This structure embodies collaboration between preclinical and clinical institutions in the co-authored description of disease-associated genes in the scientific literature.[47] As we will show below, the yeast and pig networks feature publications motivated by more varied aims and practices, implying the presence of many distinct collaborative communities with weaker co-authorships between them and a more fractured layout.

The yeast network has the highest clustering coefficient, the chance that two randomly chosen institutions co-author if they both co-author with another third institution: 58% in yeast versus 35% in pig and only 14% in the human network. These results show that the co-authorship patterns in the yeast network are particularly cohesive, largely due to the strongly connected cluster of mainly European institutions on the left-hand side of the main component (see figure 3). Most of these institutions belonged to the YGSP consortium. Members of this consortium collaborated in the sequencing of yeast chromosomes and regularly published their results as co-authored articles in *Yeast* and other specialist journals.

Outside of the European cluster, the yeast network features two large-scale sequencing centers in the United States, at Stanford University and Washington University in St. Louis.[48] The sequencing practices of these two genome centers differed from the European consortium: their work was more

47. For a case study reflecting this form of collaboration, see our analysis in García-Sancho et al., "The Human Genome Project" (n.8). The uneven distribution of co-authorships between this relatively small number of collaborative hubs and the rest of the institutions in the human network—along with its larger size—also explains its smaller average path length, and exponentially smaller clustering coefficient and density compared to the yeast and pig networks (see table 11 for figures and definitions).

48. Erika Szymanski, Niki Vermeulen, and Mark Wong, "Yeast: One Cell, One Reference Sequence, Many Genomes?," *New Genetics and Society* 38 (2019): 430–50.

concentrated and they could individually determine whole yeast chromosomes without the necessity of collaborating with others. Another difference with the European institutions was that the US genome centers were not directly interested in the biology of yeast and undertook the sequencing effort as a pilot exercise in aid of the HGP. The network position and properties of these institutions are thus shaped by the different organizations of the American and European arms of the whole yeast genome sequencing effort, with the YGSP—referred to as the *cottage industry* approach by US actors—representing an alternative to the large-scale center model.[49]

The main peculiarity of the pig network is its low degree centralization. Degree centralization measures the extent to which co-authorship ties are equally or unequally distributed across institutions. Low degree centralization signals a rather dispersed network, in which co-authorship tends to be evenly distributed across the institutions, with a rather modest proportion of core nodes that dominate the ties. The pig co-authorship network is the least centralized (with a score of 5.6%), followed by human (7.8%) and yeast (8.4%). This, together with the relatively high diameter of the pig network, accounts for its visual appearance, less compact than the human network and more formed of separated clusters of moderate density.

In the Swine Genome Sequencing Project, previous historical literature has identified the key role of a wider group of institutions beyond the Sanger Institute, the second largest pig sequence submitter and the main institution responsible for the determination of the order of nucleotides of the reference genome of this species—the *thin sequencing*. The *thick sequencing* involved groups of institutions that operated in autonomous, but coordinated, ways and produced the libraries containing the DNA that was sequenced, along with conducting sequence assembly and annotation.[50] Some of these assemblages of *thick* sequencers correspond to our network clusters. By following the co-authorship ties between these institutions, it is thus possible to explore the circulation of tools, resources, and other material entities of genomics—including the pigs from which they extracted the DNA—in a process that

49. These different organizations are also reflected in the high degree centralization score of the yeast network compared with the human and pig ones (see table 11 for figures and definitions). The high degree centralization of the yeast network suggests that a core of European institutions collaborated extensively within the YGSP and the rest occupied a much more peripheral position.

50. Lowe, "Sequencing" (n.15).

blurs the boundaries between sequence production and use, and further thickens the historiography of genomics.[51]

Overall, the co-authorship networks reflect modes of conducting genomics—including human genomics—beyond the HGP. They highlight institutions, particularly in Europe, that organized sequencing differently from the large-scale center model that US funders increasingly endorsed throughout the 1990s. In spite of this, some of the most central US institutions in the networks shared alternative modes of organization in which the practices of producing and using sequence data were blurred. We explore this entanglement between the production and use of DNA sequences further in the remaining articles of this special issue. In what follows here, we offer a roadmap to these remaining articles, along with some initial conclusions.

6. CONCLUSIONS AND SPECIAL ISSUE ROADMAP

In this special issue, we approach the history of genomics through two of its constitutive practices: the submission of DNA sequences to global, open-access databases and the much less-explored publication of those sequences in the scientific literature. Rather than being mirror images, these two practices trace different genealogies within the history of genomics for the three species we address: human, yeast, and pig. One of these genealogies is based on the total numbers of submitted nucleotides; in this, the history of genomics is the history of a distinctive and emergent field of research epitomized by a new breed of large-scale sequencing centers whose activities and goals are distinct from those of the end-users of the sequence data. The other genealogy is based on the published description of DNA sequences and suggests a broader historiography: one of mutual and inseparable entanglement between the production of genomic data and its use in biological, agricultural, and medical inquiry, meaning that the history of genomics and the history of research on evolution, immunology, genetics, biochemistry, and cell biology are co-constitutive of each other.

The observed quantitative inequalities in submitted nucleotides across institutions represent considerable differences in magnitude—and, certainly, organization—of work. However, it was not the case that the leading sequence

51. For a demonstration, see our interpretation of the pig network in Lowe et al., "The Bricolage" (n.10).

publishers were uninterested in compiling data. It was just that their sequencing activity either occurred at a notable level before the considerable ramp-up in mass and speed for their species (which masked their contemporary prominence), or that their sequencing output, despite never being at a high level relative to other institutions, was appropriate for the purposes to which they wanted to put the sequences.

We analyze specific publications and sets of publications in the rest of the special issue, allowing us to relate sequencing and genomics to existing programs of research, such as cell biochemistry and molecular biology, livestock genetics, medical genetics, and systematics. Although we seek to qualify the narrative based on numbers of submitted nucleotides, which dominates the literature on genomics, we do not seek to displace it by promoting our account as the one that truly grasps the nature and history of genomics. Instead, we have sought to demonstrate that the stories of submitting and publishing DNA sequences—and the forms of work and actors they represent—can be complementary and build on accounts that detail and query the advent of a bifurcation between two different domains of genomics research: sequence production and use.

To do this, considering a longer time frame is crucial. But, as we show throughout the special issue, it is not enough to simply establish new genealogies or to place our findings in a long-term historical trajectory. We must also thicken beyond the perspective of time and consider other dimensions that interact in scientific practice such as disciplines, research communities, target species, and home institutions. We return to this in the concluding essay of the special issue.[52] For now, we stress that the data collection strategy and mixed-methods approach we have formulated throughout this paper is crucial to capture the synchronic connections among communities, disciplines, and species, as well as their diachronic transformation over time. The large and wide body of evidence we compiled, along with its analysis, is what operationalized our historical framework and transformed the portrayal of genomics: from a *thin* field of research in search of practical application of its resulting data into a *thick* set of practices and tools in permanent entanglement with biological research, and linked medical and agricultural domains. Even outside of the history of genomics, the practices of collecting, connecting, and interpreting large datasets—such as our corpuses of sequence submissions and

52. James Lowe, Miguel García-Sancho, Rhodri Leng, Mark Wong, Niki Vermeulen, and Gil Viry, "Across and within Networks: Thickening the History of Genomics," this issue.

publications—may help scholars to grasp phenomena that can escape case study approaches, among them interdisciplinary and interspecies work, or the moving boundaries between research and its application.[53]

Throughout the special issue, we identify different modes and organizational models of genomics research and elucidate historically relevant connections and entanglements between production and use of human, yeast, and pig sequence data. This work builds on the initial network analysis detailed above to highlight particular features of the networks of each species. We use our quantitative and visual analysis of the networks, along with qualitative historical work, to develop case studies that help to articulate analytical distinctions that characterize distinct and heterogeneous modes of organizing sequencing work and conducting genomics research. The distinctions are:

- **Horizontal and vertical sequencing**, referring to the object of sequencing, adding more along the single dimension of the string of nucleotides (horizontal) or adding dimensionality through incorporating genomic variation (vertical). This distinction enables us to connect the history of genomics with the practices of locating genes associated to diseases, pooling data about these findings, and publishing the results in the scientific literature.
- **Directed/undirected and proximate/distal sequencing**, being the extent to which a given sequence is intended to satisfy a specific research goal (directed/undirected) or produced with a given subject-user in mind (proximate/distal). These distinctions enable us to differentiate a range of actors in the history of genomics of which the genome centers were just one example. Most of these actors engaged in collaborative, co-authored publications, and combined the practices of sequence production and use.

53. On the social and historical dimensions of interdisciplinary work, see Mitchell G. Ash, "Interdisciplinarity in Historical Perspective," *Perspectives on Science* 27, no. 4 (2019): 619–42; Samantha Muka, "Historiography of Marine Biology" in *Handbook of the Historiography of Biology: Historiographies of Science*, eds. Michael Dietrich, Mark Borrello, and Oren Harman (Cham, Switzerland: Springer, 2021). On work across species, see Rachel Mason Dentinger and Abigail Woods, eds., "Working across Species: Comparative Practices in Modern Medical, Biological and Behavioral Sciences," special collection of *History and Philosophy of the Life Sciences* 40, no. 30 (2018): articles 18, 20, 21, 22, 24, 27, 30. On the genealogies of applied science as a research and policy category, see Robert Bud, ed., "Applied Science," Focus Section of *Isis* 103, no. 3 (2012): 515–63.

- **Intensive/extensive sequencing**, pertaining to different modes by which sequence data becomes a scaffold, either internal to the individual, population, or species that the sequence data is supposed to represent (intensive), or to help produce genomic representations of other individuals, populations, or species (extensive). This distinction enables us to better appreciate the temporality of the entanglements between sequence production and use, the continuous reconfiguration of these entanglements, and how the advent of a reference genome is an inflection point that changes co-authorship patterns and practices in the history of genomics rather than representing its culmination.

We address the first distinction in the next article of this special issue, "The Human Genome Project as a Singular Episode in the History of Genomics." There we distinguish two genomic strategies that we use to navigate the considerably populated human co-authorship network: producing sequences horizontally by incorporating additional nucleotides to the single strings of human chromosomes and producing sequences vertically by identifying variants, for instance through the study of particular mutant genes. The large-scale center model that characterized the HGP from the mid-1990s onward was an example of pursuing horizontal sequencing to its fullest extent. We focus on another genealogy of genomics by analyzing the vertical sequencing pursued by Harvard Medical School, Massachusetts General Hospital, and other medical genetics institutions in Toronto. These were linked by co-authorships to the private company Celera Genomics, which contributed to the acceleration of the HGP's horizontal sequencing approach while also pursuing it themselves. However, Celera's commercial strategy led them to collaborate with the vertical sequencers in Boston and Toronto, leading to a project to incorporate clinical annotations to the sequence of human chromosome 7. This study, which materialized in a co-authored publication, bridges the history of genomics with the history of human and medical genetics. It also contributes to better understanding the rationalization of scientific work and the formation of private–public partnerships in genomics and late twentieth-century life science.

Then, in the article titled "Yeast Sequencing: 'Network' Genomics and Institutional Bridges," we use the distinctions between proximate and distal, as well as directed and undirected, sequencing. We contrast the nascent large-scale center model of genomics to the distributed, networked strategy of

genome sequencing pioneered by the European Commission for the YGSP. The comprehensive sequencing work at the genome centers, we argue, can be characterized as both undirected and distal, since the data was produced without reference to a specific user with which the sequencer was connected. The European model, by contrast, exhibited a variety of combinations of proximate and distal with directed and undirected sequencing. We detail these combinations by exploring three institutions that connect different clusters in our coauthorship network: two small German biotechnology companies specializing in DNA sequencing services (GATC and Genotype) and Biozentrum, a large research institute at Universität Basel founded with funds contributed by the local government and the pharmaceutical industry. Our focus on the European side of the yeast genome sequencing effort enables us to connect the history of yeast genomics to that of yeast biochemistry and molecular biology, and through the German companies to the history of biotechnology.

In the next article, titled "The Bricolage of Pig Genomics," we address the distinction between intensive and extensive sequencing. We track continuities of collaboration around particular tools and resources developed primarily by a set of agriculturally inclined institutions pursuing research aimed at improving the effectiveness of livestock breeding. Through filtering the network, we find that many of the institutions at the core pursued the characterization of genes and their variants, and were also key participants in the Swine Genome Sequencing Project. This leads us to show how the generation of particular tools for specific purposes can underpin collaboration, as well as subsequent reuse and adaptation for new purposes in what we call *bricolage*. To explore these observations further, we examine two French institutions: CEA-INRA Jouy-en-Josas (who produced a Bacterial Artificial Chromosome library and then set up the means to distribute it to the community) and the Institut National de la Recherche Agronomique station at Castanet-Tolosan near Toulouse (which produced a radiation hybrid panel for genome mapping in conjunction with the University of Minnesota). The continuities also operate across the shift from one mode of conducting sequencing to another: from an intensive mode of sequencing to an extensive mode in which genomic representations of the species are compared to, and help scaffold, new representations of particular populations or species. The availability of the pig reference genome sequence in the mid-2000s marked an inflection point from one form of sequencing predominating in the literature to the other doing so. Examining intensive and extensive sequencing allows us to connect the history of genomics with that of

agricultural genetics, immunology, and systematics research, and highlight the historiographical value of examining networks of co-authorship to identify the circulation and repurposing of tools and resources, and uncover their salience in underpinning collaboration.

Overall, our distinctions help to characterize the first decades of genomics research without reinforcing categories and concepts forged in a particular stage of the history of this field, notably the later years of the HGP. In this way, they enable us to rebalance the history of genomics away from a focus on the production of sequence data and the separation of production and use. Our distinctions both transcend this separation and operationalize the entanglement of sequence production and use, especially when applied to case studies. Exploring the institutions and collaborations at the heart of our case studies provoked us into conducting new qualitative research, which in turn led us to new analyses and interpretations of our quantitative data, network visualizations, and metrics derived from the networks. Our mixed-methods approach, as explained in this paper, was therefore not composed of separate parts but rather involved a constant dialogue and iteration between and across them.

Acknowledgments

We would like to thank colleagues within and outside the University of Edinburgh for their invaluable feedback, some of it at conferences and other events at which we presented work underpinning this special issue. Stephen Hilgartner, Steve Sturdy, Robert Cook-Deegan, and other members of the Advisory Board of our project devoted considerable time to provide us with guidance and commented on earlier drafts of the papers. Staff at the European Bioinformatics Institute worked with Mark Wong in designing a strategy to collect and parse our datasets. Rodrigo Liscovsky Barrera helped in cleaning the data, and Ann Bruce offered useful input for interpreting the co-authorship networks, especially those devoted to pig genomics. Jarmo de Vries provided helpful advice and carefully copy edited the manuscripts in his capacity of research assistant during the last stages of the project.

We are grateful to all those who have provided us with access to archives or agreed to be interviewed, including those not cited in the publications but who nevertheless have informed our analyses. Catriona Byers digitized and remotely delivered archival records during the travel restrictions of the Covid-19 pandemic, and some interviewees shared uncatalogued materials, as well as further

recollections and qualifications by email ahead of publication. Erika Milam's guidance and support was crucial in her capacity of editor of *Historical Studies in the Natural Sciences*, and one of the referees offered helpful and constructive comments.

The research and writing of this special issue were conducted through the "TRANSGENE: Medical Translation in the History of Modern Genomics" Starting Grant, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No. 678757. Without this support it would not have been possible to produce this work. We also thank the Research Support Office of the School of Social and Political Science of the University of Edinburgh for their help in managing the grant.

APPENDIX

Presented in table A1 are the archives and oral histories used in the special issue. Those listed in *italics* are not directly quoted or referred to in the special issue but were used to obtain background information (due to the nature of our work, it is impossible to offer a comprehensive list of the sources we have used).

TABLE A1. Archives and Oral Histories Used in the Special Issue

Species	***Archives***	***Oral histories***
Human	Archives of the University of Toronto Hospital for Sick Children (records of the Department of Genetics and personal collection of Lap-Chee Tsui) Archives of the Massachusetts General Hospital (personal collection of James Gusella and *records of the Department of Molecular Biology*) Personal archive of Cynthia Morton (Department of Obstetrics and Gynecology, Harvard Medical School) Papers and Correspondence of Michael Ashburner (Wellcome Library) *Papers and Correspondence of John Sulston (Wellcome Library)*	Stephen Scherer (Toronto Hospital for Sick Children) *Johanna Rommens (Toronto Hospital for Sick Children)* James Gusella (Massachusetts General Hospital) Cynthia Morton (Harvard Medical School) *Holly Zheng (formerly Celera Genomics)* Peter Li (formerly Celera Genomics) Matthew Portnoy (National Human Genome Research Institute of the US National Institutes of Health)
Yeast	University of Basel Biozentrum: Biennial Report—Zweijahresbericht, 1991–1993, 1993–1995, 1996–1997, 1998–1999, and 2000–2001 (Universität Basel, provided by Peter	Mark Johnston (formerly Washington University Genome Sequencing Center) Peter Philippsen (formerly Biozentrum)

(continued)

TABLE A1. *(continued)*

Species	Archives	Oral histories
	Philippsen) Max-Planck Gesellschaft Jahrbuch [Yearbook of the Max Planck Institutes], 1985, 1986, and 1992 (Library of the Max Planck Institute of Biochemistry, Martinsried, Munich) *Hoechst GmbH Firmenarchiv (Frankfurt, Germany)—records on the establishment of the Genzentrum at Ludwig-Maximilians-Universität München* Personal archives of Karl Kleine (in particular, Programme of the Final European Conference of the Yeast Genome Sequencing Network, Trieste, September 25–28, 1996) *Personal archive of Thomas Pohl (formerly GATC)* *Records of the Yeast Genome Sequencing Project, Historical Archives of the European Union*	Kostas Tokatlidis (formerly Biozentrum) Thomas Pohl (formerly GATC) Michael Rieger (formerly Genotype) Karl Kleine (formerly Martinsried Institute for Protein Sequences) H. Werner Mewes (formerly Martinsried Institute for Protein Sequences)
Pig	Records of Pig Gene Mapping Project and Swine Genome Sequencing Project (personal archive of Alan Archibald) and *documents pertaining to Pig Gene Mapping Project and pig genomics research networks (University of Edinburgh Centre for Research Collections)* Records of the distribution of Bacterial Artificial Clones of pig DNA from the INRA BAC-YAC Resource Center Newsletters and minutes of the United States Department of Agriculture's Animal Genome Program and the Swine Genome Sequencing Consortium. Available online: www.animalgenome.org/pig/community/NRSP8/index.html www.animalgenome.org/pig/newsletter/index.html	Patrick Chardon, Christine Renard, and Marcel Vaiman; joint-interview (formerly CEA-INRA Jouy-en-Josas) Miguel Pérez Enciso (Universitat Autònoma de Barcelona) Claire Rogel-Gaillard (INRA Jouy-en-Josas) *Max Rothschild (Iowa State University)* *Lawrence Schook (University of Illinois at Urbana-Champaign; formerly University of Minnesota)*

(continued)

TABLE A1. *(continued)*

Species	Archives	Oral histories
	www.animalgenome.org/pig/genome Personal archive of Lawrence Schook (University of Illinois at Urbana-Champaign; formerly University of Minnesota) *Personal archive of Louis Ollivier (formerly INRA Jouy-en-Josas)*	