

**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**

---

### **ABSTRACT**

In this paper, we progressively de-center the Human Genome Project (HGP) in the history of genomics and human genomics. We show that the HGP, understood as an international effort to make the human reference genome sequence publicly available, constitutes a specific model of genomics: prominent and influential but nevertheless distinct from others that preceded, existed alongside, and succeeded it. Our analysis of a comprehensive corpus of publications describing human DNA sequences submitted to public databases from 1985 to 2005 reveals a plethora of authoring institutions, with only a few contributing to the HGP. Examining these publications in a co-authorship network enables us to propose two different sequencing approaches—horizontal and vertical sequencing—whose changing dynamics shaped the history of human genomics. We argue that investigating the extent to which different institutions combined these approaches or prioritized one of them captures the history of genomics better than using the categories of large-scale sequence production and sequence use, as much scholarly literature concerning the HGP has done. Sequence production and use became fully distinct only within the HGP model, and especially during the last stages of this endeavor. By exploring a collaboration between Celera Genomics, a large-scale sequencing institution, and two medical genetics laboratories, we show the potential of our co-authorship

\*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: AUTHSC, Archives of the University of Toronto Hospital for Sick Children, Toronto, Canada; CF, Cystic Fibrosis; DNA, deoxyribonucleic acid; EST, Expressed Sequence Tag; G5, Genomic 5; HD, Huntington's Disease; HGP, Human Genome Project; IHGSC, International Human Genome Sequencing Consortium; PMID, PubMed ID; SickKids, University of Toronto Hospital for Sick Children; US-HGP, the US national human genome project.

---

*Historical Studies in the Natural Sciences*, Vol. 52, Number 3, pps. 320–360. ISSN 1939-1811, electronic ISSN 1939-182X. © 2022 by the Regents of the University of California. All rights reserved. Please direct all requests for permission to photocopy or reproduce article content through the University of California Press's Reprints and Permissions web page, <https://www.ucpress.edu/journals/reprints-permissions>. DOI: <https://doi.org/10.1525/hsns.2022.52.3.320>.

network and its analysis for historical research. Our study connects the historiographies of medical genetics and human genomics and indicates that the so-called translational gap from sequence data to clinical outcomes may reflect the assumption that genomics was substantially different from prior and parallel genetics research. 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 Michael García-Sancho and James Lowe.

KEY WORDS: genomics, DNA sequencing, *Homo sapiens*, Human Genome Project, Celera Genomics, medical genetics, cystic fibrosis, Huntington's disease

## 1. THE HUMAN GENOME PROJECT AND THE HISTORIOGRAPHY OF GENOMICS

In this paper we draw attention to the different models of human genomics that preceded, paralleled, and succeeded the more prominent and dominant model represented by the Human Genome Project (HGP), an *ad-hoc* international initiative that led to the publication of the full reference DNA sequence of *Homo sapiens* (known as *the human genome*) in 2004. This will allow us to “thicken” the history of human genomics, in line with the general objective of this special issue.<sup>1</sup> To do this, we focus specifically on the genomic work that communities of medical geneticists—who had themselves pioneered some of the early methods and institutional frameworks of genomics from the 1970s onward—conducted and organized outside of the umbrella of the HGP. They became marginal to the HGP effort but found their own way toward the large-scale sequencing and analysis of the human genome through allying with the rival project of Celera Genomics, a private company that Craig Venter led from 1998. Celera pursued an aggressive strategy to sequence the human genome and establish a proprietary database of the sequence that could be commercialized. This created a heated acrimony with the HGP, whose scientists and funders—public and charitable organizations—had committed to release the sequence data in open-access databases.

In 2000, after lengthy negotiations, Celera and the organizations behind the HGP agreed to simultaneously make available a draft version of their sequences in the scientific literature. In February 2001, this carefully

1. 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?,” this issue.

orchestrated and staged draw materialized in the publication, the same week, of two papers reporting and analyzing the sequences of the human genome. Celera Genomics described their sequence in *Science*, and an International Human Genome Sequencing Consortium (IHGSC) representing the laboratories involved in the HGP published theirs in *Nature*.

In their publication, Venter and his co-authors described the HGP as a national, government-funded program that had started in 1990 in the United States with the aim “of completing the [human] genome sequence” and partially overlapped with their own effort.<sup>2</sup> This contrasted with how the IHGSC, an organization led by twenty genome sequencing centers and a number of bioinformatics laboratories and administrative entities, saw the HGP. These institutions regarded themselves as part of an “international collaboration” and referred to their joint effort to “produce and make freely available a draft sequence of the human genome” as “the Human Genome Project.”<sup>3</sup> Structurally, however, the HGP was an *ad-hoc* amalgamation of smaller programs funded primarily by national governments—especially in the United States, United Kingdom, France, Germany, China, and Japan—and charitable organizations, namely the Wellcome Trust.

The contrast between Venter and colleagues’ definition and the IHGSC’s view shows that the meaning of the HGP had varied across actors and time. As scholars have shown, the aim of mapping and sequencing the human genome started in the mid- to late 1980s as a mosaic of national, government-funded programs that operated with relative independence and began interacting—slowly and gradually—only during the following decade.<sup>4</sup> The US program,

2. J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al., “The Sequence of the Human Genome,” *Science* 291, no. 5507 (2001): 1304–51, 1305.

3. International Human Genome Sequencing Consortium, “Initial sequencing and analysis of the human genome,” *Nature* 409 (2001): 860–921, 860; see also 862–63 for the authors’ definition of the HGP.

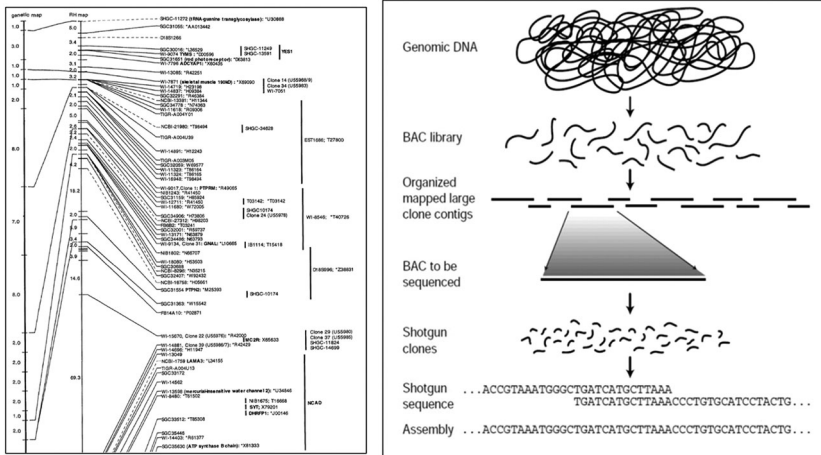
4. On the British Human Genome Mapping Project, see Brian Balmer, “Managing Mapping in the Human Genome Project,” *Social Studies of Science* 26, no. 3 (1996): 531–73; Miguel García-Sancho and James W. E. Lowe, *A History of Genomics Across Species, Communities and Projects* (Palgrave Macmillan, forthcoming), chap. 3. On the French effort, which received charitable as well as government funding, see Alain Kaufmann, “Mapping the Human Genome at Génethon Laboratory: The French Muscular Dystrophy Association and the Politics of the Gene,” in *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics*, eds. Jean-Paul Gaudillière and Hans-Jörg Rheinberger (Abingdon: Routledge, 2004), 129–57; Paul Rabinow, *French DNA. Trouble in Purgatory* (Chicago: University of Chicago Press, 1999); Vincent Ramillon, “Le Deux Génomiques. Mobiliser, Organiser, Produire: Du Séquençage À La Mesure De L’expression Des Gènes,” (PhD dissertation, École des Hautes Études en Sciences Sociales, Paris, France, 2007).

jointly sponsored by the National Institutes of Health (NIH) and the Department of Energy (DoE), was among the few that, since its inception in 1990, sought to comprehensively tackle the whole human genome.<sup>5</sup> Both the US program and the international effort of which it became part were called the Human Genome Project; the Celera-led article hinted at a continuity between the NIH and DoE initiative, and the more international IHGSC reference sequencing effort. Distinguishing between them is, however, historiographically important and in what follows we use the unqualified HGP acronym to designate only the international endeavor reported in the 2001 *Nature* paper and concluded in 2004 with the publication—also in *Nature*—of a closer-to-final version of the IHGSC reference genome. When referring to the US initiative, we label it the US-HGP.

Another implication of the contrast between Venter and the IHGSC's definition is the partiality of the HGP as a framework to capture the broader history of genomics. Strictly speaking, the HGP designated only a selective club of institutions—those that the 2001 and 2004 papers identified as members of the IHGSC—unifying their efforts toward the production of a full reference genome sequence. It excluded the sequencing project led by Celera and substantial parts of the national programs that had preceded the HGP. These national programs—including the US-HGP—had supported medical geneticists who mapped chromosomes, hunted genes and their variants (polymorphisms), and worked outward from these specific targets to map and sequence ever-larger stretches of the human genome. In spite of their apparent alignment with the objectives of the HGP, the IHGSC membership sidelined those medical genetics groups, which were often based in clinical settings and sought to connect their results to observed conditions of real-life patients.

Two key representatives of these clinically inclined geneticists were Victor McKusick and Frank Ruddle, based in the medical schools of Johns Hopkins University and Yale University, respectively. From the early 1970s onward, these two scientists had been active in the establishment of the chromosome workshops, an international forum devoted to the mapping of the human genome. Human and medical geneticists from all over the world attended these meetings, which occurred annually or biennially, and pooled the mapping results of the genes or chromosomal regions they were working on. By

5. Robert Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (New York: Norton, 1994), parts 2 and 3.



**FIGURE 1.** Left image: part of the genetic map and radiation hybrid map of chromosome 18, as reported in the Fourth International Workshop devoted to its mapping, held in Boston (United States) in 1996. The genetic map, also known as a linkage map (the term we use throughout this special issue) depicts the relative position of genes or markers on the chromosome, through horizontal lines drawn to the thick vertical line that represents the chromosome. The radiation hybrid map is a form of physical map that depicts the chromosomal location of markers or genes. Image obtained from Gary A. Silverman, Joan Overhauser, Steve Gerken, Rami Aburomia, Peter O'Connell, Ken S. Krauter, Sevilla D. Detera-Wadleigh, et al., "Report of the Fourth International Workshop on Human Chromosome 18 Mapping 1996," *Cytogenetics and Cell Genetics* 75 (1996): 111–31, 119. Reproduced with permission from Karger Publishers, Basel, Switzerland; Copyright © 1996. Right image: transition from physical mapping to sequencing as described by the International Human Genome Sequencing Consortium that published the draft reference sequence of the whole human genome in 2001. Image obtained from International Human Genome Sequencing Consortium, "Initial Sequencing and Analysis of the Human Genome," *Nature* 409 (2001): 860–921, 863. Reprinted with permission from Rightslink for Springer Nature, Copyright © 2001.

doing this, the workshop attendees achieved an increased resolution in both the knowledge of the human genome and the location of specific areas connected to genetic diseases (see figure 1).<sup>6</sup>

McKusick incorporated the workshop results, initially into his ongoing *Mendelian Inheritance in Man* catalog and later into an electronic repository called The Genome Database, housed at Johns Hopkins University with joint

6. Emma M. Jones and Elizabeth M. Tansey, eds., *Human Gene Mapping Workshops c.1973–c.1991: The Transcript of a Witness Seminar Held by the History of Modern Biomedicine Research Group, Queen Mary University of London, on 25 March 2014* (London: Queen Mary University of London, 2015). [www.histmodbiomed.org/sites/default/files/W54LoRes.pdf](http://www.histmodbiomed.org/sites/default/files/W54LoRes.pdf)

funding from the DoE and NIH.<sup>7</sup> With this systematic compilation, he sought to develop a growing and increasingly refined map connecting genetic diseases to different areas of the forty-six human chromosomes. In 1987, with Ruddle, he founded the journal *Genomics* as another step in this direction. In their first editorial, titled “A New Discipline, A New Name, A New Journal,” Ruddle and McKusick argued that “mapping all expressed genes” and “sequencing out from these” was “seen by many as the way to go.” They presented the human genome sequence as “the ultimate map” or a “Rosetta stone” from which “the complexities of gene expression in development” could be “translated and the genetic mechanisms of disease interpreted.”<sup>8</sup> Human and medical geneticists—and, to a lesser extent, evolutionary biologists—dominated the authorship of the first volume of *Genomics*.

Historians have observed and documented the strong presence of human and medical geneticists in the foundational years of genomics. Soraya de Chadarevian sees Ruddle and McKusick’s work as the culmination of the gene-mapping practices associated with human cytogenetics in the second half of the twentieth century, and argues that these practices offer an alternative route to genomics from histories that emphasize continuities with molecular biology.<sup>9</sup> In the same vein, Andrew Hogan traces the emergence of “the genomic gaze” of disease to the 1970s and challenges the idea that the “clinical gaze” of geneticists was replaced by a “molecular gaze.” Drawing on McKusick’s engagement in gene mapping, Susan Lindee proposes a “cataloguing imperative” that has shaped the history of genetics and connects with later initiatives such as the sequencing of the human genome.<sup>10</sup> Other monographs on the history of human and medical genetics suggest lineages between the practices of chromosome mapping and genome sequencing,

7. Peter Li, personal communication, November 2021.

8. Victor A. McKusick and Frank H. Ruddle, “EDITORIAL: A New Discipline, A New Name, A New Journal,” *Genomics* 1 (1987): 1–2, 1. See also Alexander Powell, Maureen A. O’Malley, Staffan Müller-Wille, Jane Calvert, and John Dupré, “Disciplinary Baptisms: A Comparison of the Naming Stories of Genetics, Molecular Biology, Genomics, and Systems Biology,” *History and Philosophy of the Life Sciences* 29, no. 1 (2007): 5–32.

9. Soraya de Chadarevian, *Heredity Under the Microscope: Chromosomes and the Study of the Human Genome* (Chicago: University of Chicago Press, 2020), esp. chap. 5. On genealogies with molecular biology, see Michel Morange, *The Black Box of Biology: A History of the Molecular Revolution* (Cambridge, MA: Harvard University Press, 2020), part 4 and chap. 27.

10. Andrew Hogan, *Life Histories of Genetic Disease: Patterns and Prevention in Postwar Medical Genetics* (Baltimore: Johns Hopkins University Press, 2016), 210; M. Susan Lindee, *Moments of Truth in Genetic Medicine* (Baltimore: Johns Hopkins University Press, 2005), 89.

without fully exploring the connections between the two endeavors beyond pointing to collaborations as well as rivalries.<sup>11</sup> This lack of specificity is partly due to the absence of intersecting actors: with the exception of McKusick, who appeared as a co-author in the Celera-led paper, none of the contributors to the first issue of *Genomics* was present in the two articles that in 2001—fourteen years later—reported the first draft sequences of the whole human genome.

The separation of the mapping and sequencing endeavors of medical geneticists from the project to sequence the whole human genome occurred gradually throughout the 1990s. Stephen Hilgartner has argued that in this period genomics established a “knowledge-control regime” that became distinct from other life science and biomedical research fields. This knowledge-control regime predicated the creation of large-scale centers that conducted various aspects of genomics, including DNA mapping and sequencing, in an increasingly concentrated and exclusive fashion. Their establishment was due to the successful implementation of a vision propounded by a set of early advocates of whole-genome sequencing, including senior figures in the DoE and James Watson, the renowned molecular biologist who co-elucidated the double helix structure of DNA and served as the initial leader of the NIH arm of the US-HGP. These centers were connected to reputed US universities and medical schools—sometimes due to their involvement in the mapping and sequencing of other organisms with smaller genomes—and sought to tackle the human genome in a rapid, efficient manner.<sup>12</sup>

As Hilgartner describes, during the early years of the US-HGP, researchers at large-scale centers combined comprehensive mapping and sequencing of the human genome with work on genes and chromosomal regions involved in diseases, often in collaboration with medical geneticists. In the mid- to late 1990s, the US-HGP and other national programs started coalescing into what became the IHGSC as the NIH, DoE, and other sponsors—mainly the Wellcome Trust in the UK—created a funding regime that enabled the large-scale centers to

11. Nathaniel Comfort, *The Science of Human Perfection: How Genes Became the Heart of American Medicine* (New Haven, CT: Yale University Press, 2014), chap. 7; Peter Harper, *A Short History of Medical Genetics* (Oxford: Oxford University Press, 2008), 207, 378. See also Jones and Tansy, “Human Gene” (n.6).

12. Stephen Hilgartner, *Reordering Life: Knowledge and Control in the Genomics Revolution* (Cambridge, MA: MIT Press, 2017), chaps. 3–4. On how the large-scale center vision has shaped historical and anthropological scholarship on genomics, see Leng et al., “The Sequences” (n.1). On the role of the US large-scale centers in the mapping and sequencing of the yeast genome and the contrast of their operation with the network organization of genomics in Europe, see Miguel García-Sancho, James Lowe, Gil Viry, Rhodri Leng, Mark Wong, and Niki Vermeulen, “Yeast Sequencing: ‘Network’ Genomics and Institutional Bridges,” this issue.

exclusively focus on the production of sequence data with advanced technology. They left the use of that data to other institutions, among them human and medical genetics laboratories that would access the reference sequence in their investigations of genetic diseases and evolutionary problems.<sup>13</sup>

While the NIH streamlined its funding to three large-scale sequencing centers—based in the Whitehead Institute, Washington University School of Medicine, and Baylor College of Medicine—the DoE amalgamated three preexisting genome centers into the Joint Genome Institute, and the Wellcome Trust substantially increased its support to the Sanger Institute, the UK-based genome center it had established in 1993 as the Sanger Centre.<sup>14</sup> These institutions became known as the Genomic 5 (G5) and featured as the largest sequence contributors to the 2001 *Nature* paper. The adoption of that model of organization represented a victory for Watson, the DoE officials, and those arguing for the concentration of efforts and a focus on technology development. Their vision had, at times, conflicted with established figures in medical genetics, who advocated continuing to focus on individual genes using the available mapping and sequencing instruments.

The decision to concentrate funding in the G5 followed the International Strategy Meeting for Human Genome Sequencing, convened chiefly at the initiative of the Sanger Institute and the Wellcome Trust and held in Bermuda in 1996. Out of this meeting arose new forms of international coordination, as well as the commitment of the IHGSC institutions and other attendees to rapidly release new sequence data to global, open-access databases. This commitment was aimed at countering the growing practice of patenting genes or parts thereof, something that meeting organizers feared would inhibit research on sequence data, and therefore stymie the purported medical benefits the HGP would deliver. The emergence of Celera in 1998 accentuated these concerns and pushed the IHGSC and its funders to prioritize whole-genome sequencing over mapping.<sup>15</sup>

The HGP thus triggered a dramatic rearrangement of funding priorities and organizational models and led to the demarcation of two distinct categories of

13. Hilgartner, *Reordering* (n.12), esp. chap. 7.

14. The name change materialized in 2001. For consistency across the special issue, we use the name Sanger Institute, although its involvement in the sequencing of the yeast and human genomes mainly occurred under the name Sanger Centre.

15. 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, no. 4 (2018): 693–805.



genomic actors: the large-scale centers that produced whole-genome sequences and the much more loosely defined *other laboratories* that were supposed to use the sequence data for medical and biological research purposes. Yet, as we will argue, this distinction captures only the final stage of the HGP (when it acquired its own and governance regime), and a relatively much later stage in the history of genomics (when comprehensive sequencing at large-scale centers consolidated as the dominant model). In other words, and in line with other organizational rearrangements that the historiography of science and technology has described, the alleged optimization of efficiency that the large-scale center model aimed to achieve promoted some actors—the G5 and other large-scale sequence producers—at the expense of others, such as the disease-oriented genome mappers.<sup>16</sup>

In this paper, we characterize human genomics, including human reference genomics, beyond the forms represented by a particular stage of the culmination of the HGP. This stage had undoubted significance in the history of genomics, and influence in shaping the norms, infrastructure, organizational architecture, and methods of genomics research more broadly. But however dominant, it was not the only approach to conducting genomics, and it was not the only approach that survived and thrived throughout the 1990s. In line with other contributions to this special issue, we will thicken the historiography of genomics by examining an unexplored trajectory of human genomics: that pursued by the medical geneticists who had little to do with the IHGSC but who came to intersect with the other large-scale effort led by Celera.

## 2. THE GOALS AND APPROACH OF THIS PAPER

Our historical approach starts with the chromosome mappers that, despite being so visible during the early years of genomics, were later relegated to the ill-defined category of users of the human reference sequence. We aim to

16. The organizational model that Watson and colleagues proposed for the HGP pursued higher efficiency through concentration of resources rather than just automation. Yet as we show later in this paper, the automation of sequencing practices at the large-scale centers connects their operation with other investigations of automation processes, their history, and social consequences. See David Noble, *Forces of Production: A Social History of Industrial Automation* (New York: Alfred A. Knopf, 1984); Harry Braverman, “Technology and Capitalist Control,” in *The Social Shaping of Technology*, 2nd edition, eds. Donald MacKenzie and Judy Wajcman (Buckingham: Open University Press, 1999), 158–60; Judy Wajcman, “Addressing Technological Change: The Challenge to Social Theory,” *Current Sociology* 50, no. 3 (2002): 347–63.

characterize these disease-oriented gene mappers as contributors to genomics rather than mere *users* of the resulting data. Our goal is historiographically relevant since it enables us to expand upon Hilgartner's scholarship and other literature through detailing the interactions—or the lack thereof—between medical geneticists and large-scale sequencing centers, from the foundation of the latter institutions to the years following the publication of the human reference sequence.<sup>17</sup> By doing this, we also bridge two separate bodies of historical scholarship—on medical genetics and genomics—showing that their divide is an artifact of reducing human genomics to the production of the human reference sequence. As we will show, the lack of a clear definition of *sequence user* is due mainly to the emphasis of policymakers on the production side following the implementation of the large-scale center model.

We specifically address the contribution of medical geneticists to the generation, analysis, and interpretation of new DNA sequences. To do this, we combine historical research with quantitative and visual analysis of a co-authorship network. The network, derived from the dataset we presented earlier in this special issue,<sup>18</sup> depicts co-authorship relationships among institutions publishing DNA sequences in the scientific literature for the first time. In the case of human DNA, the dataset comprises almost 25,000 publications that report sequences submitted to global, open-access repositories between 1985 and 2005—a period that precedes, encompasses, and follows the determination of the human reference genome. As we show below, most of the publications underlying the network map and further characterize genes implicated in diseases. Although large-scale sequencing centers co-authored some of these articles, the majority of the authoring institutions in the network are medical genetics laboratories. The network thus enables us to identify some of these medical genetics groups and their connections with large-scale sequencing centers, something that has remained rather obscure in the historiography of genomics.

17. On the divergence and hybridizations of the practices of collecting, comparing, and experimenting with data in biomedical research throughout the twentieth century, see Bruno Strasser, *Collecting Experiments: Making Big Data Biology* (Chicago: University of Chicago Press, 2019). On bioinformatics as a continuous “cycle of production and consumption” of sequence data, see Hallam Stevens, *Life Out of Sequence: A Data-Driven History of Bioinformatics* (Chicago: University of Chicago Press, 2013), 73.

18. Leng et al., “The Sequences” (n.1). See also 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>

In the next section of this paper (section 3), we discuss our approach to analyzing this large network. By examining the two strongest co-authorship ties, we identified two major medical genetics teams at teaching hospitals in Boston and Toronto that were both prolific and consistent publishers of DNA sequences throughout our 1985–2005 period. Analysis of their connections in the network revealed an unexpected co-authorship between these medical geneticists and Celera, which we further investigated through oral histories, archival research, and close reading of the article in which it materialized, alongside other publications underpinning the network. The co-authored paper turned out to be the tip of an iceberg of hidden lineages between the practice of mapping clinically relevant genes and Celera's large-scale sequencing effort. These lineages are substantially more absent when the IHGSC becomes the standpoint of the narratives of the determination of the human genome sequence.

In sections 4 and 5, we detail the Celera-medical genetics lineages by proposing the concepts of vertical and horizontal sequencing. Vertical sequencing captures a longstanding practice of geneticists based in medical schools and teaching hospitals, one that predated the determination of the reference human genome and continued during and after this large-scale sequencing exercise. It consists in compiling data across three dimensions: the linear string of DNA nucleotides of a chromosomal region, the sequence variation in that region across individuals (generations of a family or patients compared to controls), and the possible connection of that variability to the phenotype (the manifestation of disease).

Throughout the 1990s, and especially after the Bermuda conference of 1996, Venter and researchers involved in the IHGSC argued for an alternative approach that we call horizontal sequencing. In this approach, the emphasis shifted from addressing individual disease-linked genes toward mapping and sequencing whole chromosomes and later the whole human genome. This shift, however, involved focusing on just one dimension of the data—the string of nucleotides—and leaving the other two to the *users* of the sequence. As we will show, the rise of the horizontal sequencing approach led some medical geneticists to reconsider their sequencing strategies and address increasingly larger areas of the genome.

The concepts of horizontal and vertical sequencing offer an alternative to the dichotomy between sequence producers and users that emerges from the equation of the history of human genomics with the large-scale center model. As we show below, medical genetics institutions continued to produce

sequences vertically—across family lineages and disease patients—despite the HGP funders recasting them as users of the genome centers. These multidimensional, vertical sequences provided information about variability and clinical effects that was vital to link the (horizontal) reference genome data to medical problems. Our vertical/horizontal approach will enable us to argue that, from Venter's perspective, the separation between sequence producers and users was never fully dichotomous. From 1992 onward, when his work started to rely on corporate funding, he needed to adapt his sequence production to the necessities of customer-users. This dependency pre-dated Celera and materialized in contacts that led to the collaboration with medical geneticists that our co-authorship network captures.<sup>19</sup>

Building on this, we will conclude that rather than being a metonym standing for genomics as a whole, the HGP represents a highly contingent and historically situated episode, one substantially different from other forms of genomics that preceded, co-existed, and followed it. The separation between sequence producers and users that the HGP predicated was only fully accomplished between the late 1990s and early 2000s, and affected a limited number of contributors to the reference sequence: those grouped in the IHGSC and, especially, the G5.<sup>20</sup> The different historicity of Venter's endeavor suggests a genealogy between medical genetics and genomics that the HGP story conceals, and our vertical and horizontal approach renders visible. Within this pathway that emphasizes continuity over disruption and displacement, Venter co-constructed the sequences with medical genetics co-authors rather than casting this community as a mere user of the data. To start uncovering this

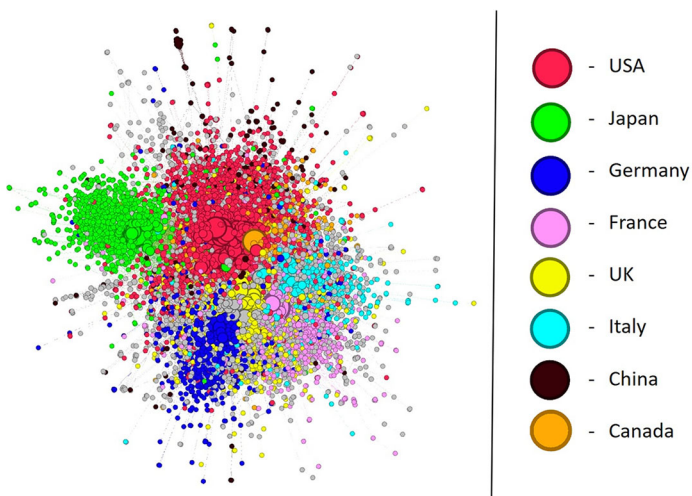
19. The collaboration between Celera and state-funded medical genetics institutions further challenges the rigid dichotomy between corporate and public actors that scholarship on genomics has consistently criticized, e.g., Michael Fortun, "Celera Genomics: The Race for the Human Genome Sequence," in *Living with the Genome: Ethical and Social Aspects of Human Genetics*, eds. Angus Clarke and Flo Ticehurst (Basingstoke: Palgrave, 2006), 27–32. More generally, it aligns with the historiography of commercialization of contemporary science, which has identified institutional entanglements that blur the boundaries between the private and public: Peter Weingart, "From 'Finalization' to 'Mode 2': Old Wine in New Bottles?," *Social Science Information* 36, no. 4 (1997): 591–613; Olle Edqvist, "Layered Science and Science Policies," *Minerva* 41, no. 3 (2003): 207–21; David Edgerton, "Time, Money, and History," *Isis* 103, no. 2 (2012): 316–27; Benoît Godin and Désirée Schauz, "The Changing Identity of Research: A Cultural and Conceptual History," *History of Science* 54, no. 3 (2016): 276–306.

20. On this contingency, see Bertrand Jordan's anthropological immersion into some of the IHGSC centers, conducted in 1993 when they started to debate how best to focus their mapping and sequencing operations: Bertrand Jordan, *Travelling around the Human Genome: An in situ Investigation* (Paris: INSERM, 1993).

eclipsed genealogy of genomics, we now introduce our networks of co-authored publications describing sequence data.

### 3. INTERROGATING A CO-AUTHORSHIP NETWORK

The scale of the activity that the human network captures is exponentially larger than in the other two we analyze in this special issue: 6,014 co-authoring institutions compared to 684 and 1,272 in the yeast and pig networks, respectively. The co-authored publications in the human network described new DNA sequences of *H. sapiens* submitted to the European Nucleotide Archive, GenBank, and DNA Data Bank of Japan—the three global, open-access sequence repositories—between 1985 and 2005. These sequence descriptions materialized in 24,726 publications that provided the data underlying the network. Overall, 39,565 co-authorship relationships (edges, represented as lines in the network) connect the authoring institutions (nodes, represented as circles). Figure 2 shows the network's main component, the largest interconnected group of nodes. This comprises 93% of the institutions and almost all of the co-authorship ties.



**FIGURE 2.** Main component of the co-authorship network of publications describing new human sequences. We sized the nodes representing institutions according to their number of cross-institutional publications and colored them by country, as indicated in the legend on the right side of the figure; we colored the rest of the nodes gray. Figure elaborated by the authors.

The network derives from over 2.5 million sequence submissions to the three repositories, which are a fraction of the almost 10.1 million human sequences submitted from 1985–2005, representing 21 billion DNA nucleotides. As the approximate size of the reference human genome is 3 billion nucleotides, this further demonstrates that our dataset contains both published and unpublished sequences whose determination occurred outside the HGP and Celera’s sequencing efforts.

As we discussed earlier in this special issue, the publishers and submitters in the dataset present a strong asymmetry. The institutions that published the highest number of articles describing new sequences do not necessarily coincide with those that submitted the largest volume of nucleotide data. Moreover, the sequence submissions are considerably more skewed and dominated by a few large institutions than the publication data.<sup>21</sup> In the case of the human dataset, this reflects different practices between the community of medical geneticists and the researchers based in the large-scale genome centers. The former promptly published any achievement resulting from their gene-mapping and sequencing experiments, along with a discussion of the clinical relevance of those results. The latter, though, prioritized rapid submission over publication, in order to complete the genome projects they were involved in on time. This meant that the G5 institutions—and also Celera—would publish their sequences only when they represented milestones toward the conclusion of the whole human genome, the 2001 first draft papers in *Nature* and *Science* being examples of this.<sup>22</sup> Their main mission was to produce sequence data rather than publications describing parts of it, and the different leaderboard of Tables 1 and 2 below—especially after 1996—reflects the different priorities of the genome centers compared to the medical genetics community.

Submitters and publishers of human sequences also markedly differ in their evolution over time. If we split our dataset into two batches, one limited to the period 1985–1995, and the other to 1996–2005, we observe striking differences across the submission and publication records.

21. Leng et al., “The Sequences” (n.1).

22. The journal *Cytogenetics and Cell Genetics* published the proceedings of the chromosome workshops where medical geneticists pooled their mapping and sequencing results. By contrast, throughout the 1990s biomedical journal editors increasingly demanded further analysis of the sequences they published. On the chromosome workshops, see Jones and Tansey, “Human Gene” (n.6). On the changing publication policies of journals, see Strasser, *Collecting* (n.17), 214; Stevens, *Life* (n.17), 58–60.

**TABLE 1.** Largest Submitters of Human DNA Sequences to Global Databases in the Periods 1985–1995 and 1996–2005\*

<b>Largest Submitters of Nucleotides 1985–1995</b>		<b>Largest Submitters of Nucleotides 1996–2005</b>	
<b>Rank</b>	<b>Institution</b>	<b>Rank</b>	<b>Institution</b>
1	Généthon	1	Celera Genomics
2	Wellcome Trust Sanger Institute	2	Whitehead Institute for Biomedical Research
3	University of Padua	3	Wellcome Trust Sanger Institute
4	Kazusa DNA Research Institute	4	Washington University St. Louis School of Medicine
5	National Institute of Genetics Japan	5	DOE Joint Genome Institute
6	Imperial Cancer Research Fund	6	Kazusa DNA Research Institute
7	Genzentrum München	7	Baylor College of Medicine
8	MRC Human Genome Mapping Project Resource Centre	8	Genoscope
9	Baylor College of Medicine	9	RIKEN Institute of Physical and Chemical Research
10	National Cancer Center Research Institute Japan	10	University of Washington
11	Harvard Medical School	11	University of Tokyo
12	Ludwig-Maximilians-Universität München	12	Hospital for Sick Children Canada
13	University of Oklahoma	13	National Institutes of Health Mammalian Gene Collection
14	European Molecular Biology Laboratory	14	Helix Research Institute
15	Kyoto Prefectural University of Medicine	15	Technische Universität München

\*We have highlighted institutions that were leading submitters in both periods.

Only three of the top fifteen submitters from the second batch occupied this position before 1996, the year of the Bermuda meeting. In contrast, the overlap among sequence publishers is much higher: ten out of the top fifteen institutions headed the leaderboard in both periods (tables 1 and 2). Furthermore, whereas the numbers of publications per institution tend to be equally distributed across the two batches, the volume of submitted nucleotides varies dramatically, with the overlapping institutions contributing virtually all their data after 1996 and some of the others halting their submissions in the second period. This suggests that the publication data reflect a stable practice of reporting experimental results, while the submissions radically shifted with the consolidation of the large-scale

**TABLE 2.** Largest Publishers of Human DNA Sequences Submitted to Global Databases in the Periods 1985–1995 and 1996–2005\*

<b>Largest Publishers of Papers Describing Newly Submitted DNA Sequences, 1985–1995</b>		<b>Largest Publishers of Papers Describing Newly Submitted DNA sequences, 1996–2005</b>	
<b>Rank</b>	<b>Institution</b>	<b>Rank</b>	<b>Institution</b>
1	National Cancer Institute Bethesda	1	Harvard University Medical School
2	Harvard University Medical School	2	University of Tokyo
3	University of California San Francisco	3	Inserm
4	Inserm	4	Baylor College of Medicine
5	Washington University in Saint Louis School of Medicine	5	National Institutes of Health Bethesda
6	Baylor College of Medicine	6	National Cancer Institute Bethesda
7	Imperial Cancer Research Fund and Cancer Research UK	7	University of Washington
8	University of Washington	8	University of California San Francisco
9	Massachusetts General Hospital	9	University of Toronto
10	Howard Hughes Medical Institute	10	Osaka University
11	Scripps Research Institute	11	Brigham and Women's Hospital
12	University of Texas Southwestern Medical School	12	Washington University in Saint Louis School of Medicine
13	University of Toronto	13	Johns Hopkins School of Medicine
14	Johns Hopkins School of Medicine	14	Massachusetts General Hospital
15	University of Chicago	15	Deutsches Krebsforschungszentrum (German Cancer Research Center)

\*We have highlighted institutions that were leading publishers in both periods.

centers as institutions whose funding and organization sought to make the sequences rapidly available in data repositories rather than the scientific literature.

The consistency of publication rankings across the two periods suggests that the reporting of sequencing results in the literature continued as a characteristic practice of the medical genetics community, regardless of the funding boost that the genome centers received from the NIH, DoE, and Wellcome Trust since 1996. Yet the direction that the HGP adopted during the mid- to late 1990s and the identification of the history of genomics with just reference



genomics has made it difficult to follow the footprints of disease-oriented gene mapping and sequencing, especially in the shadow of the enhanced submission capacities of large-scale centers. Our dataset and network thus represent an opportunity to uncover this history, tracing its parallels and connections with that of the genome centers.

The presence of institutions explicitly engaged with medical genetics research at the top of the publication leaderboard offers us a useful entry point into the crowded co-authorship network. Not unexpectedly, three of the institutions in the top fifteen are represented in the two strongest ties of the network, between (1) Harvard Medical School (HMS) and the Massachusetts General Hospital (MGH), and (2) the University of Toronto (UT) and the University of Toronto Hospital for Sick Children (SickKids). Together, these four institutions co-authored 113 articles from our dataset, 77 jointly signed by HMS and the MGH, and 76 by UT and SickKids (some articles include all four institutions). HMS and SickKids are also the first and fifth most degree-central institutions in the whole network: those with co-authorship ties with the greatest number of other institutions.<sup>23</sup> They thus constitute a suitable standpoint to start examining the thousands of ties and nodes in the network.

An analysis of the 113 co-authored publications reveals that, throughout 1985–2005, scientists at HMS, MGH, SickKids, and UT were mainly focused on determining the chromosomal position of genes or other DNA fragments connected to genetic diseases. Mapping often involves the sequencing of small portions of DNA to act as probes to find the desired location in the genome. The co-authorships between HMS, MGH, SickKids, and UT thus reflect the longstanding practice of mapping and sequencing genes or other DNA fragments of medical interest and reporting these results in the literature, as McKusick and Ruddle had asked scientists to do in the first editorial of the journal *Genomics*. Neither HMS, nor MGH, UT, or SickKids became part of the HGP or the Celera-led genome projects.

The co-authorships underpinning these articles often result from the double affiliation of one scientist—with both HMS and MGH, or SickKids and UT—rather than two or more colleagues publishing together at different

23. HMS and MGH, SickKids and UT, remain the two strongest ties if we limit the network to the first batch of our dataset (1985–1995). By then, HMS had already become the most degree-central institution of the whole network. More surprisingly, HMS was the eleventh largest submitter of DNA nucleotides between 1985 and 1995, something that further shows the dramatic effects of the concentration of funding and sequence production on the genome centers during the latter stages of the HGP (see table 1).

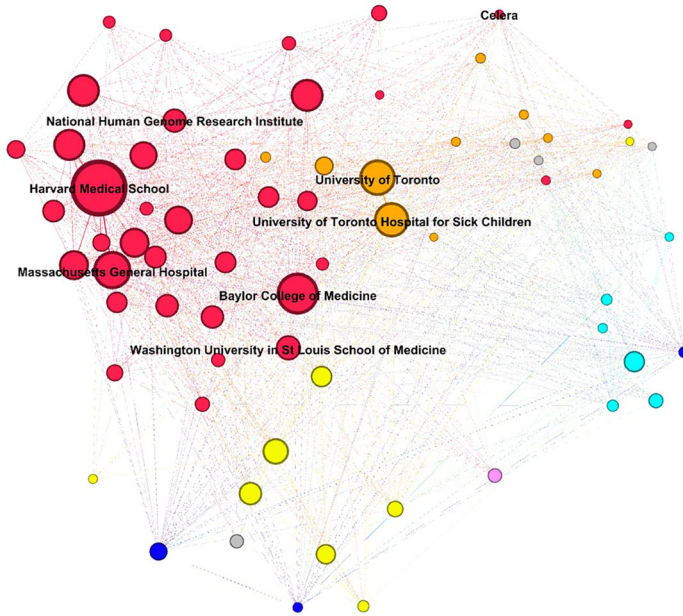
institutions. Although in such instances double affiliation does not reflect collaboration at the level of individual scientists, it may show a purposeful organizational arrangement that requires interaction between two different institutions. This was the case with Harvard and Toronto, where the double affiliations derived from a concerted institutional plan. In 1980, HMS founded its Department of Genetics, which evolved into a model that allowed faculty positions to be filled by researchers already associated with clinical divisions or laboratories of the MGH or other Harvard-affiliated teaching hospitals in the Boston area. This model fostered clinical innovation and allowed a career path that included academic tenure for researchers based in the hospitals.<sup>24</sup> A similar arrangement was in place at Toronto between the pre-clinical and clinical units of SickKids and the UT.<sup>25</sup> Being in the same city enabled the double-affiliated scientists to move from one of the spaces to the other and exploit resources—patients, techniques, knowledge, colleagues—that were present in only one of them.<sup>26</sup>

To further visualize the connections of these especially active medical genetics institutions, we displayed their intersecting ego-networks: networks centered on a focal node and its ties—an *ego* in the argot of social network analysis. To do this, we used the filters and subfilters of Gephi—our network analysis software—and showed only the nodes and edges representing HMS, MGH, UT, and SickKids, as well as institutions with at least one co-authorship tie with all of them. The resulting subnetwork comprises 68 institutions from ten countries connected by 1,102 co-authorship relationships (figure 3).

24. Joseph B. Martin, *Alfalfa to Ivy: Memoir of a Harvard Medical School Dean* (Edmonton, Canada: Gutteridge Books, 2011). James Gusella, interview with Miguel García-Sancho, Massachusetts General Hospital, Boston, November 2018; personal communication, November 2021.

25. Freda Zipursky, Aser Rothstein, and Manuel Buchwald, “A Decade of Research at the Hospital for Sick Children” (1984), 55–65. AUTHSC, file number 2007-288-001 6, consulted April 2018. Stephen Scherer, interview with Miguel García-Sancho, SickKids, Toronto, April 2018.

26. The other strongest ties of the human network share this trend: the ten largest number of co-authorships are between institutions based in the same city, and eight of them involve either a medical school or a teaching hospital. Historical and sociological scholarship has highlighted the importance of such local connections for what in the 1990s and early 2000s began to be known as clinical translation of biomedical research results; see Alison Kraft, “New Light through an Old Window? The ‘Translational Turn’ in Biomedical Research: A Historical Perspective,” in *Translational Medicine: The Future of Therapy?*, eds. James Mittra and Christopher-Paul Milne (Boca Raton, FL: CRC Press, 2013), 19–54; Peter Keating and Alberto Cambrosio, *Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine* (Cambridge, MA: MIT Press, 2003). The yeast and pig datasets offer a different co-authorship pattern.



**FIGURE 3.** Intersecting ego-networks of Harvard Medical School, the Massachusetts General Hospital, the University of Toronto, and the Toronto Hospital for Sick Children. The resulting visualization displays only nodes with at least one co-authorship tie with all four of these institutions and the edges between them. We sized the nodes according to their number of inter-institutional publications and colored them by country using the same legend as in figure 2. We discuss the labeled nodes in the main text. Figure elaborated by the authors.

The subnetwork shows international collaboration, especially with British and Italian institutions. Most of the co-authorship ties, however, are between US institutions, with an additional significant presence of Canadian nodes. What is most interesting about the position of HMS, MGH, UT, and Sick-Kids is that they are all connected with leading institutions in the other set of rankings we produced concerning numbers of nucleotides submitted: Celeris Genomics, Baylor College of Medicine, and Washington University in St. Louis School of Medicine, as well as the National Human Genome Research Institute (NHGRI), the arm of the NIH that funded and organized the US-HGP. This means that as well as instantiating the practice of publishing DNA

sequences, the ego-networks of HMS, MGH, UT, and SickKids open the opportunity to look at connections between this practice and the systematic submission of genome data. The intersecting co-authorship ties of these four institutions also provide a link between the two rival genome projects: the one led by Baylor College of Medicine, Washington University in St. Louis, and the other G5 institutions, and the one led by Celera.

An examination of the publications, though, reveals that none of the sub-network ties connect HMS, MGU, UT, and SickKids with co-authors at the genome centers of Washington University and the Baylor College of Medicine. Any joint articles between HMS, MGU, UT, and SickKids on the one hand, and Washington University and the Baylor College of Medicine on the other, are with authors from other departments of the latter two institutions. As we will see below, the activity of the genome centers is captured instead in the ties between Washington University and the NHGRI: HMS, MGU, UT and SickKids do not therefore manifest direct connections with the institutions involved in the HGP.

Conversely, the ties with Celera enable us to connect HMS, MGU, UT, and SickKids with the practice of large-scale sequencing. These ties correspond with just one publication, which appeared in *Science* in 2003 under the title “Human Chromosome 7: DNA Sequence and Biology,” signed by scientists affiliated with forty-three institutions (in our dataset: PMID 12690205). This publication shows that despite being mainly devoted to disease-oriented mapping throughout the 1990s, the Toronto and Boston institutions, along with the other co-authors, engaged in the full-length description of a chromosome with Celera, the largest sequence submitter in our dataset and a much less prolific publisher. The ego-networks, along with our analysis of the batched submission and publication data, capture this change of orientation of the Toronto and Boston institutions, and something distinctive about Celera’s strategy compared with the HGP. The moment in which the co-authorship occurred—two years after Celera and the IHGSC had simultaneously published their draft whole-genome sequences—is also historically relevant.

This network observation enabled us to develop a case study that seeks to interrogate the historical trajectories and intersection of the vertical and horizontal sequencing approaches. Why and how did HMS, MGH, UT, and SickKids end up co-authoring the full description of a chromosome sequence with Celera? Why did they choose Celera to work with instead of the IHGSC institutions? Why did this co-authorship follow the publication of the entire human genome sequence? What did the Toronto and Boston institutions

contribute to the chromosome 7 publication? Were these institutions involved in the production of the chromosome 7 sequence, even though they had been deemed as sequence users within the large-scale center model of genomics? In what follows, we augment our dataset and network analyses with qualitative historical work that addresses these questions and aims to explore the motivations behind the co-authorships and underlying collaborations among HMS, MGH, SickKids, UT, and later Celera.

#### 4. AUGMENTING SEQUENCES VERTICALLY

When they published with Celera, the Toronto- and Boston-based contributors to the chromosome 7 paper were researchers at various stages of their careers. Stephen Scherer, the lead author, was thirty-nine years old by the time of publication, in 2003, and had shortly beforehand become associate director of the Center for Applied Genomics at SickKids. Other more senior scientists who co-authored the paper, such as James Gusella and Lap-Chee Tsui, were established figures within the community of medical genetics for having mapped and sequenced the genes involved in Huntington's Disease (HD) and Cystic Fibrosis (CF) between the late 1980s and early 1990s.

The 2003 co-authorship led us to visit Gusella's laboratory in Boston, and Tsui and Scherer's groups in Toronto, with the objective of reconstructing the events linking the characterization of the disease genes to their subsequent participation in the sequencing of an entire chromosome. The objective of our visits was obtaining qualitative evidence from the archives of the universities and hospitals, as well as conducting oral histories that led us to personal, uncatalogued records. This evidence was what enabled us to transform the network ties and underlying datasets into a history of disease gene mapping that, in its later stages, penetrated the history of whole-genome sequencing.

Gusella is a Canadian-born geneticist who in the late 1970s started his career as a graduate student at the Massachusetts Institute of Technology (MIT) working on the globin gene system. In 1980, he moved to MGH following the award of an NIH grant that sought to create synergies between physicians and biomedical scientists investigating HD. The grant built on advances in the construction of a linkage map tracing the inheritance of the disease among affected families and enabled Gusella to obtain tenure in the HMS Genetics Department while working at the Neurology Department of MGH, thanks to double affiliation. Gusella's role was to trace the HD gene to a specific DNA

fragment and physical chromosomal position (see left part of figure 1, above). He found the first marker—a DNA fragment connected to the gene—in 1983 and isolated the gene in 1993 (in our dataset: PMID 8458085). In between, in 1987, Gusella’s team contributed to the first volume of *Genomics*, reporting on the identification of various DNA segments near the HD gene.<sup>27</sup>

Gusella considers that the main difference between MGH and his PhD graduate work at MIT was the focus on disease genes rather than technology development. This reflected a mindset ingrained earlier when he began MSc graduate studies at the Princess Margaret Hospital’s Ontario Cancer Institute, where he would regularly “see patients in the lunch room walking with their drips.” The pathologies that gene mutations may cause in patients and the possibility of making an impact on the diagnosis and treatment of the diseases at the hospital while also learning fundamental scientific knowledge was—and still is—what motivated Gusella’s endeavor. His MGH laboratory, like those at the MIT, was able to apply the latest DNA technologies, such as isolating small bits of human DNA by placing them in bacterial cells and then growing those cells to replicate and thereby produce biochemically analyzable quantities of each bit. A major route of analysis was to expose the DNA to bacterial enzymes known as restriction enzymes, each of which could make a cut if the requisite sequence of 4–6 base pairs was present. Gusella’s priority was repeating the same analysis on “many small bits” of DNA from all around the genome in families with members both affected and unaffected by HD to see if any of the differences in cutting tracked with the inheritance of the disease.<sup>28</sup>

From the mid-1980s onward, there was an intense debate in the United States and other countries with national human genome programs on whether the priority should be the development of more powerful mapping and sequencing methods or the characterization of genes using the available instruments.<sup>29</sup> Eric Lander, a researcher at the Whitehead Institute for Biomedical Research, was one of the strongest defenders of focusing on technology development. The Whitehead Institute—the second-largest sequence submitter in our dataset—was another Boston-based institution established in the early

27. See the memoirs of former interim general director of MGH and Dean of HMS Joseph Martin for a Boston-centered story of the HD gene: *Alfalfa* (n.24), chap. 6.

28. James Gusella, interview with Miguel García-Sancho, Massachusetts General Hospital, Boston, November 2018; personal communication, November 2021.

29. On this debate, see “Sequencing the Human Genome,” a special section of *Issues in Science and Technology* 3, no. 3 (1987): 25–56; and the “Debate” section of *The FASEB Journal* 5, no. 1 (1991): 75–78.

1980s to pursue basic research connected to, but independent from, the biological investigations at MIT. Lander worked on chromosomal structure using the advanced mapping and sequencing methods that MIT and other institutions were devising. Building on his experience and the growing capacity of those techniques, he argued that progress would be faster if scientists tackled larger genome areas than the markers of HD and other target diseases of medical genetics.<sup>30</sup> In other words, Lander was proposing to replace the focus on disease genes and variability characteristic of vertical sequencing with a horizontal approach that pursued chromosome and genome-wide analysis.

The Whitehead Institute, along with the MIT, founded a Center for Genome Research in 1990 and appointed Lander as its director. This center complemented other whole-genome mapping units that the DoE was establishing at its Lawrence Berkeley and Lawrence Livermore National Laboratories. A substantial part of the medical genetics community advocated for a different model in which the funding would be distributed among networks of laboratories interested in the same chromosome or chromosomal region—typically connected to vertical variability underlying a disease. The networks would still use the chromosome marker as the principle for grouping NIH grant awardees, contrary to Lander's approach. By creating complementary networks and making the results available to the community, the NIH could compile an ever-growing map and sequence, but still a partial one. This is due to disease-related regions comprising only a small part of the overall human genome. Despite this network model requiring a more modest initial investment, the NIH increasingly channeled its grants on the Whitehead Institute and other genome centers that it sponsored once the US-HGP started.

The advance of the genome center model was a sign of the NIH gradually shifting its grant award mechanisms and organizational preferences toward what Watson, the DoE, and other early advocates of sequencing the whole human genome proposed. NIH administrators and policymakers were increasingly persuaded that focusing on sequence production through the concentration of advanced technology and leaving the clinical interpretation of the data to other laboratories—including Gusella's—was more efficient than pursuing both sequencing and data interpretation under the same roof. Literature on the history and sociology of technology has shown that industrialization and other processes of rationalization of scientific and technical work always involve

30. See an oral history interview with Eric Lander at [www.cshl.edu/oral-history/eric-lander](http://www.cshl.edu/oral-history/eric-lander).

power battles hidden under the labels of optimization and efficiency.<sup>31</sup> In the case of human sequencing, Gusella and other medical geneticists saw their remit shrink to just using sequences to diagnose and treat disease. Yet the responsibility and funding to determine those sequences shifted to the large-scale genome centers at the Whitehead Institute and DoE laboratories.

This dichotomy between vertical, medically oriented, and horizontal, technology-driven approaches also featured in the mapping and sequencing of the CF gene. Tsui at Toronto aligned with medical geneticists and, contrarily to Lander, prioritized the isolation of the disease gene over the characterization of broader areas of the genome. A Chinese-born researcher, Tsui had joined the Department of Genetics at SickKids in 1981 and was also affiliated to the Department of Medical Genetics and Medical Biophysics of UT. A main driving force of his research, like Gusella's, was improving the life quality of hospital patients, as suggested by letters from parents and other relatives of children affected by CF that are available in Tsui's archive.<sup>32</sup> CF is a disease that affects people at early age and was, therefore, a priority at SickKids.<sup>33</sup>

The mapping of the CF gene involved collaboration between Tsui's group at Toronto and the University of Michigan. The team in Michigan included Francis Collins, a researcher who had devised the technique of chromosome jumping to make it easier to find markers and eventually isolate a gene by screening multiple DNA fragments across human chromosomes. The teams co-authored a series of *Nature* papers in which they reported a positive association between a DNA fragment and the CF gene in 1989 (in our dataset: PMID 2475911). However, the collaboration did not progress much beyond these joint publications, and it was Tsui and other Toronto-based researchers

31. On technology, automation, and industrialization, see note 16. On rationalization of genomic work, see Peter Keating, Camille Limoges, and Alberto Cambrosio, "The Automated Laboratory: The Generation and Replication of Work in Molecular Genetics," in *The Practices of Human Genetics*, eds. Michael Fortun and Everett Mendelsohn (Dordrecht: Springer, 1999), 125–42; Michael Fortun, "Projecting Speed Genomics," in *The Practices of Human Genetics*, eds. Michael Fortun and Everett Mendelsohn (Dordrecht: Springer, 1999), 25–48; Chris Mellingwood, "Amphibious Researchers: Working with Laboratory Automation in Synthetic Biology" (PhD dissertation, University of Edinburgh, Edinburgh, Scotland, 2018); Jenny Reardon, "The Genomic Open," *Limn* 6, 2016, <https://limn.it/articles/the-genomic-open>; Stevens, *Life* (n.17), 115–21; Andrew Bartlett, "Accomplishing Sequencing the Human Genome" (PhD dissertation, Cardiff University, Cardiff, Wales, 2008), sections C3 and C4.

32. Uncatalogued letters, Papers and Correspondence of L.C. Tsui, AUTHSC, consulted April 2018.

33. David Wright, *SickKids: The History of the Hospital for Sick Children* (Toronto: University of Toronto Press, 2016), chap. 14.



who sequenced the CF gene during the early 1990s (in our dataset: PMID 1710598).<sup>34</sup>

This distancing between Michigan and Toronto was partly due to Collins deciding to apply his jumping technique to other genes after 1989. Rather than looking at variants within the CF region, he focused on other genome locations and became a pioneer of positional cloning, a mapping technique in which it is not necessary to know the function of the gene prior to its isolation. This prioritization of horizontal over vertical approaches squared with the strategy that the NIH was beginning to favor and, in 1993, Collins succeeded Watson as director of the NHGRI.<sup>35</sup> Collins's move and his subsequent role as HGP champion has led some scientists and commentators to retrospectively cast the identification of the CF region as an antecedent of the whole-genome project, especially after the publication of the first draft reference sequence in 2001.<sup>36</sup> Although this narrative may correspond with Collins's experience, at the time of the events other co-workers reacted to the CF achievement in a significantly different way.

This was the case for Tsui, who after 1989 concentrated his efforts on the clinical possibilities that the CF finding opened. In 1995, he and other colleagues at SickKids filed a patent protecting the technique for the isolation of the gene. The patent application was restricted to the technique, without affecting the CF gene, its sequence, or the diagnosis of the condition. Medical geneticists, including Gusella, would often protect their techniques, so their home universities or hospitals could license the patents to companies conducting diagnostic tests.<sup>37</sup> On several occasions, the income resulting from this

34. For a first-person account, see Lap-Chee Tsui and Ruslan Dorfman, "The Cystic Fibrosis Gene: A Molecular Genetic Perspective," *Cold Spring Harbor Perspectives in Medicine* 3, no. 2 (2013): a009472.

35. Until 1997, the NHGRI was the National Center for Human Genome Research. For consistency throughout the special issue, we use the later name.

36. Nicole Kresge, Robert D. Simoni, and Robert L. Hill, "The Molecular Genetics of Cystic Fibrosis: The Work of Francis Collins," *Journal of Biological Chemistry* 286, no. 3 (2011): e8–e9.

37. Lap-Chee Tsui, John R. Riordan, Francis S. Collins, Johanna M. Rommens, Michael C. Iannuzzi, Bat-Sheva Kerem, Mitchell L. Drumm, and Manuel Buchwald, "Methods of Detecting Cystic Fibrosis Gene by Nucleic Acid Hybridization," US Patent number 5,776,677, filed 6 June 1995, and issued 7 July 1998; Subhashini Chandrasekharan, Christopher Heaney, Tamara James, Chris Conover, and Robert Cook-Deegan, "Impact of Gene Patents and Licensing Practices on Access to Genetic Testing for Cystic Fibrosis," *Genetics in Medicine* 12, no. 4 S (2010): S194. Gusella attempted to commercialize an HD test through the Massachusetts-based company Integrated Genetics, later acquired by Genzyme: MGH Archives and Special Collections, Gusella, James T., MD, Biographical Files, b. 9.

licensing became a source of funding for further research at the patent applicants' teams and their home institutions.

Tsui and Gusella also engaged in creating networks to facilitate the sharing of mapping data and the diagnosis of the diseases they were working on. A "Collaborative Research Group" signed the 1993 paper that reported the isolation of the HD gene and included eight institutions, among them the MGH, with Gusella as the corresponding author. Collins also featured as a co-author, still based in Michigan and expanding his interests from CF to other conditions before his move to the NIHGR (see figure 4). The authors presented the identification of the HD gene as a collective effort and the consequence of pooling the mapping results of each of the groups involved (in our dataset: PMID 8458085).

**\*The Huntington's Disease Collaborative Research Group comprises:**

**Group 1:**

Marcy E. MacDonald,<sup>1</sup> Christine M. Ambrose,<sup>1</sup> Mabel P. Duyao,<sup>1</sup> Richard H. Myers,<sup>2</sup> Carol Lin,<sup>1</sup> Lakshmi Srinidhi,<sup>1</sup> Glenn Barnes,<sup>1</sup> Sherryl A. Taylor,<sup>1</sup> Marianne James,<sup>1</sup> Nicolet Groot,<sup>1</sup> Heather MacFarlane,<sup>1</sup> Barbara Jenkins,<sup>1</sup> Mary Anne Anderson,<sup>1</sup> Nancy S. Wexler,<sup>3</sup> and James F. Gusella<sup>1\*</sup>

<sup>1</sup>Molecular Neurogenetics Unit  
Massachusetts General Hospital  
and Department of Genetics  
Harvard Medical School  
Boston, Massachusetts 02114

<sup>2</sup>Department of Neurology  
Boston University Medical School  
Boston, Massachusetts 02118

<sup>3</sup>Hereditary Disease Foundation  
1427 7th Street, Suite 2  
Santa Monica, California 90401

**Group 2:**

Gillian P. Bates, Sarah Baxendale, Holger Hummerich, Susan Kirby, Mike North, Sandra Youngman, Richard Mott, Gunther Zehetner, Zdenek Sedlacek, Annemarie Poustka, Anna-Maria Frischauf, and Hans Lehrach

Genome Analysis Laboratory  
Imperial Cancer Research Fund  
Lincoln's Inn Fields  
London, WC2A 3PX, England

**Group 3:**

Alan J. Buckler,<sup>1</sup> Deanna Church,<sup>1</sup> Lynn Doucette-Stamm,<sup>1</sup> Michael C. O'Donovan,<sup>1</sup>

Laura Riba-Ramirez,<sup>1</sup> Manish Shah,<sup>1</sup> Vincent P. Stanton,<sup>1</sup> Scott A. Strobel,<sup>2</sup> Karen M. Draths,<sup>2</sup> Jennifer L. Wales,<sup>2</sup> Peter Dervan,<sup>2</sup> and David E. Housman<sup>1</sup>

<sup>1</sup>Center for Cancer Research  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

<sup>2</sup>Division of Chemistry and Chemical Engineering  
California Institute of Technology  
Pasadena, California 91125

**Group 4:**

Michael Altherr, Rita Shiang, Leslie Thompson, Thomas Felder, and John J. Wasmuth  
Department of Biological Chemistry  
University of California  
Irvine, California 92717

**Group 5:**

Danilo Tagle, John Valdes, Lawrence Elmer, Marc Allard, Lucio Castilla, Manju Swaroop, Kris Blanchard, and Francis S. Collins  
Department of Internal Medicine and Human Genetics  
and The Howard Hughes Medical Institute  
University of Michigan  
Ann Arbor, Michigan 48109

**Group 6:**

Russell Snell, Tracey Holloway, Kathleen Gillespie, Nicole Datson, Duncan Shaw, and Peter S. Harper  
Institute of Medical Genetics  
University of Wales College of Medicine  
Cardiff, CF4 4XN, Wales

<sup>1</sup>Correspondence should be addressed to James F. Gusella.

**FIGURE 4.** The "Huntington's Disease Collaborative Research Group," as reported in Marcy E. MacDonald, Christine M. Ambrose, Mabel P. Duyao, Richard H. Myers, Carol Lin, Lakshmi Srinidhi, Glenn Barnes, et al., "A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes," *Cell* 72, no. 6 (1993): 971–83, 971 (in our dataset: PMID 8458085). Republished with permission of Elsevier Science & Technology Journals, copyright © 1993; permission conveyed through Copyright Clearance Center Inc.

#52

Genetic Anal-

**CYSTIC FIBROSIS MUTATION DATA** (December 16, 1994)  
(Privileged communication prepared for members of the CF Genetic Analysis Consortium)

Summary of CF mutations:

Name	Nucleotide change	Exon	Consequence	CFTR domain	Reference
-816C→T	C→T at -816	5'upstream	promoter mutation?		Bienvenu et al. (NL#60)
-741T→G	T→G at -741	5'upstream	promoter mutation?		Bienvenu et al. (NL#59)
-471delAGG	deletion of AGG from -471	5'upstream	promoter mutation?		Grade et al. 1994
MIV	A→G at 133	1	No initiation codon		Cheadle et al. 1993c
MIK	T→A at 134	1	No initiation codon		Claustres et al. 1993
MII	G→A at 135	1	No initiation codon		Axton & Brock (NL#61)
Q2X (together with R3W?)	C→T at 136	1	Gln→Stop at codon 2		Savov et al. 1994a
R3W (together with Q2X?)	A→T at 139	1	Arg→Trp at codon 3		Savov et al. 1994a
S4X	C→A at 143	1	Ser→Stop at 4		Glavac et al. 1993
P5L	C→T at 146	1	Pro→Leu at 5		Chillón et al. (NL#59)
K14X	A→T at 172	1	Lys→Stop at 14		Ferec et al. (NL#56)
175delC	deletion of C at 175	1	frameshift		Ferec et al. (NL#56)

**FIGURE 5.** List of mutations identified by the Cystic Fibrosis Genetic Analysis Consortium and compiled by Lap-Chee Tsui to aid the diagnosis of the disease at the University of Toronto Hospital for Sick Children and other clinical settings (uncatalogued files, AUTHSC, consulted April 2018). Reprinted with permission from Hospital Archives, The Hospital for Sick Children, Toronto. We thank David Wencer for help in managing the permission request.

Similarly, after the 1989 *Nature* paper, Tsui's prestige among the medical genetics community led him to become co-coordinator of the mapping workshops devoted to chromosome 7, where the CF gene is. He also took the lead in the creation of a Cystic Fibrosis Genetic Analysis Consortium that compiled a database of mutations associated with the disease. His archive reveals an intense correspondence among the consortium members and extensive lists of mutations, which suggest that this activity occupied a considerable amount of Tsui's time (see figure 5). The information in the database helped, and still helps, SickKids and other hospitals to diagnose the condition.<sup>38</sup>

These genetic disease consortia operated as spaces to pool results and create knowledge that transcended individual laboratories, thus amplifying the mapping and sequencing approach embodied by Tsui and Gusella. By accumulating knowledge about associated markers and mutations of a given gene, the consortia augmented the reach of the maps and sequences and eased the diagnosis of CF, HD, or any other genetic condition. Within the

38. "Cystic fibrosis mutation data (16 December 1994) (Privileged communication prepared for members of the CF Genetic Analysis Consortium)," uncatalogued file, AUTHSC, consulted April 2018. The database was still active at the time of writing. [www.genet.sickkids.on.ca](http://www.genet.sickkids.on.ca)

consortia, this augmentation affected only the vertical rather than the horizontal reach of the sequence: it related to the specificity and variability of a given gene rather than attempting to extend to other areas of the genome. There were, however, broader forums that also accomplished the horizontal extension, such as the chromosome workshops and the journal *Genomics*.

Tsui and Gusella's approach contrasted with that of Lander, Collins, the NHGRI and the Genome Center of the Whitehead Institute. Instead of being constrained by particular chromosomes, genes, markers, or disease-causing mutations, these latter scientists and institutions focused on producing a reference sequence of the whole human genome. The projected reach of this sequence overcame the boundaries of any medical genetic forum, but at the expense of abstracting—at least initially—the information about vertical variability. Toward the mid- to late 1990s, the sequencing approach that prioritized horizontal over vertical reach gained momentum and forced the disease consortia to redefine their alliances.

## 5. AUGMENTING SEQUENCES HORIZONTALLY

Throughout the 1990s, a variety of both publicly funded and commercial institutions arose with the objective of delivering increased amounts of genome data building on the emergent horizontal sequencing approach. In 1993, the NIH sponsored their second Genome Sequencing Center at Washington University in St. Louis (WU), reinforcing the shift toward concentrating the production of a reference genome sequence in a reduced number of technologically intensive facilities.<sup>39</sup> These centers, however, co-existed with other institutions that received genome mapping grants from the NIH and combined horizontal reference genome work with parallel vertical projects to locate genes and explore potential pathological variants. Researchers at these institutions had not separated the production of genome data from their use in medical genetics research. Yet the new NIH-funded genome centers were able to concentrate on producing horizontal sequences without the necessity of pursuing other lines of research.<sup>40</sup>

39. On the foundation of the genome center in St. Louis, see Christopher Donohue's interview with David Schlessinger, Oral History Collection of the National Human Genome Research Institute. [www.youtube.com/watch?v=N\\_oSUvzMTQo&t=2s](http://www.youtube.com/watch?v=N_oSUvzMTQo&t=2s)

40. During the early years of these genome centers, there was a limited number of gene-hunting projects that researchers conducted in parallel to the large-scale mapping and sequencing work and, crucially, at separate laboratories or start-up companies: see Hilgartner, *Reordering* (n.12), 140.

The horizontal sequencing approach also flourished in the commercial side when, in 1992, Venter left the NIH to found The Institute for Genomic Research (TIGR), a nonprofit organization. From this new institution, he devised a sequencing strategy that made use of what he called Expressed Sequence Tags (ESTs). This strategy selectively yielded regions across the whole genome potentially involved in diseases.<sup>41</sup> Venter's sequencing was thus still focused on locating genes, but did not limit its scope to specific genome regions or disease markers, in line with what Lander at the Whitehead Institute had recommended.

The rise of these new players affected practices at MGH and SickKids. Medical genetics laboratories feared that, in their horizontal sequencing endeavors, Venter or the genome centers would tackle the genes they were pursuing and either file patents or otherwise prevent them from asserting their priority. In Boston, Gusella started collaborating with Cynthia Morton, a researcher at Brigham and Women's Hospital, another local teaching hospital associated with HMS. Morton directed the cytogenetics laboratory in the Department of Pathology and was interested in exploring the breakpoints of balanced chromosome rearrangements as locations of genes involved in clinical phenotypes. Her interests were thus not connected to any particular chromosome or disease, and had enabled her to compile a list of chromosomal alterations. Morton and Gusella's collaboration focused on a series of structural rearrangements that chromosomes experience during embryonic development that are connected to cleft palate, mental disabilities, limb defects, and other anomalies in newborns. Morton pioneered chromosomal in situ hybridization (FISH) to localize chromosomal breakpoints, and Gusella, along with another researcher at the Boston Children's Hospital (Gail Bruns), used their cloning expertise to discover the gene(s) in these chromosomal regions. This culminated in a project that used the breakage points of these chromosomes as signposts to systematically identify new genes.<sup>42</sup>

In Toronto, Scherer also transitioned to broader genome areas and devised a strategy to map chromosome 7 beyond the CF gene. Building on the growing

41. Myles W. Jackson, *The Genealogy of a Gene: Patents, HIV/AIDS, and Race* (Cambridge MA: MIT Press, 2015), esp. chap. 1.

42. Cynthia Morton, interview with Miguel García-Sancho, Brigham and Women's Hospital, Boston, November 2018; personal communication, November 2021. See also "DGAP: Developmental Genome Anatomy Project," application submitted to the NIH in 1999 and funded that same year, in Morton's personal archive (consulted November 2018). The co-authorship relationships of HMS with the Children's Hospital in Boston and Brigham and Women's Hospital underpin the fourth and fifth strongest ties in the whole human network, respectively.

collaborative network around Tsui, who had supervised his PhD dissertation, Scherer offered a mapping service to both CF geneticists wanting to expand from the regions that the 1989 *Nature* paper had identified and groups attending the Chromosome 7 Workshops. The Genome Sequencing Center of WU also required Scherer's input, since it was determining the whole sequence of chromosome 7 as part of the US-HGP, along with NHGRI.<sup>43</sup> In 1998, together with Tsui as director, Scherer became associate director of the Center for Applied Genomics, a new unit that UT and SickKids founded with support from the Canadian government.

Also in 1998, Celera started its activity as the corporate competitor of the genome centers. Venter became Celera's CEO and continued to focus on functionally relevant regions of the genome. These regions were the main commercial target of the new firm, created to provide sequence data to pharmaceutical industry laboratories. Yet, just before the launch of Celera, a new automatic DNA sequencer with enhanced capacity and speed (the Applied Biosystems Prism 3700) had entered the market and persuaded Venter that tackling the whole genome would be more efficient than screening for disease associations through the EST strategy. Further, the tightening of criteria for patenting DNA sequences in the light of growing public controversy made developing a proprietary database of the full genome a more secure business opportunity. For this database to be commercially viable, Celera needed to be the first in determining the sequence and this led Venter to pursue his whole-genome approach without developing a prior physical map.<sup>44</sup>

Celera's competition precipitated the strategic shift that prioritized horizontal over vertical sequencing. By 1999, the NIH, the DoE, and the Wellcome Trust had concentrated their support in the G5 and tasked these institutions with completing and publicly releasing the reference sequence before Celera could restrict access. They also decided that the G5 and other large-scale sequencers would conduct their own *ad-hoc* mapping rather than relying on

43. Stephen Scherer, interview with Miguel García-Sancho, SickKids, Toronto, April 2018.

44. On Celera's whole genome strategy and its differences with the IHGSC, see 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: Routledge, 2004), 158–79. On the Prism 3700 sequencer, see Miguel García-Sancho, *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA, 1945–2000* (Basingstoke: Palgrave Macmillan, 2012), chap. 6. On the patenting controversy, see Robert Cook-Deegan and Christopher Heaney, "Patents in Genomics and Human Genetics," *Annual Review of Genomics and Human Genetics* 11 (2010): 383–425; Maxson Jones et al., "The Bermuda" (n.15), 763–71.

a prior, more broadly and collectively produced physical map.<sup>45</sup> This new regime consolidated the institutional differentiation of the genome centers compared to other life sciences laboratories,<sup>46</sup> and completed the power shift that had started with the introduction of the large-scale center model.

The division between (horizontal) sequence production and use channeled the funding of the genome programs—public, commercial, and charitably sponsored—toward the large-scale centers and separated their whole-genome sequencing practices from medical research goals. Yet the kind of separation differed between the G5 and Celera. Whereas for the former, the fulfillment of medical goals was supposed to occur after the determination of the reference genome, following what began to be called the *translation* of the data,<sup>47</sup> Celera still needed to persuade its potential customers of the suitability of their sequence for the development of diagnostic and therapeutic interventions.

The connections between Celera and medical genetics laboratories, as documented in our co-authorship network, were traces of this persuasion exercise.<sup>48</sup> In what follows, we qualitatively examine those with Gusella, Tsui, and Scherer's groups, and argue that they derived from overlaps in the sequencing strategies of Venter and the medical genetics community throughout the 1990s. These overlaps created synergies that materialized in the co-authored chromosome 7 article that brought together not only Celera and the Boston and Toronto teams but also other medical genetics groups scattered around the network space of our visualization.

The first overlap between Venter and medical geneticists was that they both patented some of their research results: the sequences in the case of TIGR, and the mapping techniques in the case of Tsui and Gusella. Unlike the G5 centers,

45. On this mapping strategy and the role of the Sanger Institute in shaping it within the IHGSC, see García-Sancho and Lowe, *A History* (n.4), chap. 4.

46. Hilgartner, *Reordering* (n.12).

47. Kraft, "New Light" (n.26).

48. In line with historiography that has challenged the public–private dichotomy in the production and commercialization of scientific knowledge (see note 19), the connections between Celera and noncommercial medical genetics laboratories qualify the idea of a race between *corporate* and *public* human genome projects. Scholars have already questioned a monolithic separation between public and commercial actors, and documented alliances between the G5 institutions and large pharmaceutical companies, such as Merck: Maxson Jones et al., "The Bermuda" (n.15); Steve Sturdy, "Public versus Private Interests in the Creation of the Genomic Commons," unpublished draft. On a more recent collaboration between a private sequencing company (Illumina) and a publicly funded clinical annotation initiative (the ClinVar repository), see Emmanuel Didier, "Open-Access Genomic Databases: A Profit-Making Tool?," *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 659–72.

whose finances were provided by the DoE, NIH, or Wellcome Trust, the work of isolating and sequencing disease-related genes was expected to be supported, at least partially, by corporate funders. There was some intersection between Venter's funders and those of medical geneticists, since some of the pharmaceutical companies with which Celera worked were interested in using genomic data for diagnostic purposes. When Celera's proprietary database came into the picture, the commercial protection of this company's results was thus not as anathema for Gusella and Tsui, as it was for the G5 and its commitment to the unrestricted release of data.

Second, Venter's goal was and continued to be the protein-coding regions of the genome, as it was for the medical genetics community. One of Venter's first recruits at Celera was Peter Li, a software engineer at Johns Hopkins University working on both the online version of *Mendelian Inheritance in Man* and The Genome Database. The rationale behind Li's appointment was to develop bioinformatic approaches modeled on those databases in order to annotate Celera's sequence with information about the position of genes and other clinical data of relevance to the interests of the pharmaceutical industry customers. Celera also used information from the *Online Mendelian Inheritance in Man* to annotate the sequence it published in 2001 in *Science*, explaining the presence of McKusick as a co-author of this article.<sup>49</sup>

Third, Venter was always engaged with the medical potential or any other possible exploitation of the sequences he determined. In TIGR, this materialized in a preferential agreement with the company Human Genome Sciences to clinically commercialize the EST data. In Celera, Venter never lost sight of the necessities of his target customers, despite becoming a large-scale sequence producer and not having direct contact with them.<sup>50</sup> Celera's user engagement manifested in the redeployment of approaches from the medical genetics community to curate their draft genome sequence—via Li's recruitment and McKusick's co-authorship of the *Science* paper. Another way in which this

49. Peter Li, Skype interview with Miguel García-Sancho and James Lowe, September 2020. Johns Hopkins University had lost its NIH human genome mapping grant following the decision to concentrate funding and operations on the G5 institutions. On the practice of annotating DNA sequences, see García-Sancho and Lowe, *A History* (n.4), chap. 6.

50. We further develop this different degree of definition of the sequence users elsewhere in this special issue, through the concepts of directed and undirected, as well as proximate and distal sequencing. As with the cases of TIGR, Celera and the G5, knowing or not knowing in advance who the beneficiaries of their sequences would be shaped both strategy and institutional positioning of different laboratories during the sequencing of the yeast genome: García-Sancho et al., "Yeast Sequencing" (n.12).



producer–user interface operated was through the so-called *jamboree meetings*, whose rationale and format resembled the human chromosome workshops.

The jamborees gathered Celera’s personnel and a group of scientists representing the sequence users in the company’s premises in Gaithersburg, close to Washington, DC. The first was in 1999, when over two weeks Venter’s team and a group of expert geneticists discussed the first whole-genome sequence that the company had co-determined. This was the reference genome of the fruit fly, *Drosophila melanogaster*. As a large-scale producer not directly in touch with *Drosophila* genetics, Celera’s initial annotation of the sequence—indicating the position of the genes, the proteins they coded for, and other relevant biological characteristics—had been exclusively based on computational methods. According to Li, the company needed the know-how and biological expertise from the users to improve that layer of metadata that always accompanied fully sequenced genomes. More generally, “getting the *Drosophila* geneticists on board” was essential for Celera’s sequence being trusted, valued, and used in the long term by both the jamboree attendees and the community to which they belonged.<sup>51</sup> The firm had partially achieved this through an agreement with a rival, NIH-funded *Drosophila* sequencing project at the University of California, Berkeley. The jamboree, scheduled six months after this agreement, was aimed at life scientists with more general expertise on the fly, as well as a European consortium that had competed with the US sequencing projects (see figure 6).

In April 2001, seventeen months after the *Drosophila* gathering and two months after the publication of the human reference sequence, Venter convened a second jamboree. Attended by users of the human sequence who were based in hospitals and medical schools, this second jamboree presented some differences with the *Drosophila* antecedent. First, while the drosophilists had been prominently involved in the determination of the fly sequence, medical geneticists were more absent in the last stages of the completion of the human reference genome, due to the implementation of the large-scale center model. They were therefore outside of the structures of the IHGSC and free to collaborate rather than compete with Celera in annotating the human sequence. Second, and related to the IHGSC competition, Celera had pursued

51. Peter Li, Skype interview with Miguel García-Sancho and James Lowe, September 2020. Invitations, preparation documents and materials produced during the *Drosophila* jamboree are available at the Papers and Correspondence of Michael Ashburner, Wellcome Library, London, collection reference PP/MIA/A/5/5 (consulted October 2020; includes nine separate files).

**Drosophila Annotation Jamboree**  
November 7 - 20, 1999

**Vignette Report**  
*Tell us about your day.*

Please e-mail this form to [REDACTED] for posting.

**Date:** 11/17/99  
**Last Name:** Cravchik  
**First Name:** Anibal

**What did you seek?**  
G-protein coupled receptors (non-olfactory)


**How did you look for it?**  
Blastp and tblastn, protein hydrophobicity predictions, HMMs.

**What did you find?**  
As of Thursday night we have about 80 non-olfactory GPCRs. Of these, 70% are new receptors.

**What did you conclude?**  
So far, it appears that the number of these receptors in *Drosophila* is much lower than in *C. elegans*.

**Did you change your course or pursue a tangent? How?**

**Do you wish you had changed your course or pursued a tangent? How?**



**FIGURE 6.** A feedback form from an attendee at the Celera *Drosophila* jamboree. Retrieved from Papers and Correspondence of Michael Ashburner, Wellcome Library, London, file number PP/MIA/A/5/5/8. We have edited the image to anonymize the name of the intended recipient of the form. Reproduced with permission from Anibal Cravchik.

a commercial goal with the human sequence in the form of a proprietary database. Given that after the 2001 publication the option of restricting access was no longer viable—Celera had agreed to release its draft sequence data—Venter cultivated a proximate relationship with the prospective users to differentiate his sequence from the G $\zeta$ 's.<sup>52</sup> Clinical annotations, which the published draft had not made available, could broker lucrative collaborations between Celera and the pharmaceutical industry, potentially mediated by the

52. Indeed, Venter criticized the Bermuda Protocol, suggesting it would lead to a “dangerous bifurcation” between data production and analysis; Maxson Jones et al., “The Bermuda” (n.15), 740.

geneticists attending the jamboree. In other words, Celera sought to reverse-engineer from the horizontal sequence to the vertical and clinically relevant sequence annotations and mapping data that medical geneticists had long been producing and compiling in their chromosome workshops and disease-oriented consortia.

Scherer was among the attendees to the human jamboree. During the proceedings, he proposed to Richard Mural—head of the Annotation Team at Celera—to use the company’s draft sequence to construct a complete annotated version of chromosome 7. This would enable Celera to close the gaps of its unmapped draft and, more importantly, gather an extra layer of information from Scherer’s physical mapping network, comprised of scientists working on genetic diseases affecting chromosome 7. The proposal was particularly timely and appealing for Celera, which after the first draft publication was reorienting its business model. Among its plans was the establishment of a diagnosis division and resequencing substantial areas of the genome with the aim of entering the drug development market. The polished sequence of chromosome 7, along with the physical mapping details from the medical geneticists, would be invaluable assets to develop and patent potential therapeutic targets.<sup>53</sup>

Scherer and Mural’s negotiation resulted in a collaborative Chromosome 7 Annotation Project. Its website and database, still active online at the time of writing, describe the initiative as “a weighing station for testing community ideas” and producing “highly curated data.” A particularly crucial aspect that the project envisaged was to make the sequence “available in a user-friendly manner having every biological and medically relevant feature annotated along its length” to “facilitate biological discovery, disease gene research and medical genetic applications.” In other words, Scherer and Mural aimed to mobilize the knowledge of biologists and geneticists working on chromosome 7 and incorporate data about its physiological and clinical implications to the horizontal sequence. This way, the sequence would also represent vertical “chromosome alterations, variants, and polymorphisms.” The project’s organizers would deem it a success “when an

53. Paul Rabinow and Talia Dan-Cohen, *A Machine to Make a Future: Biotech Chronicles* (Princeton, NJ: Princeton University Press, 2004). According to our dataset, Celera submitted all its eight billion DNA nucleotides between 2000 and 2005, coinciding with the agreement to publish the first draft sequence and the resequencing exercise. Venter left the company in 2003, the same year in which *Science* published the chromosome 7 paper.

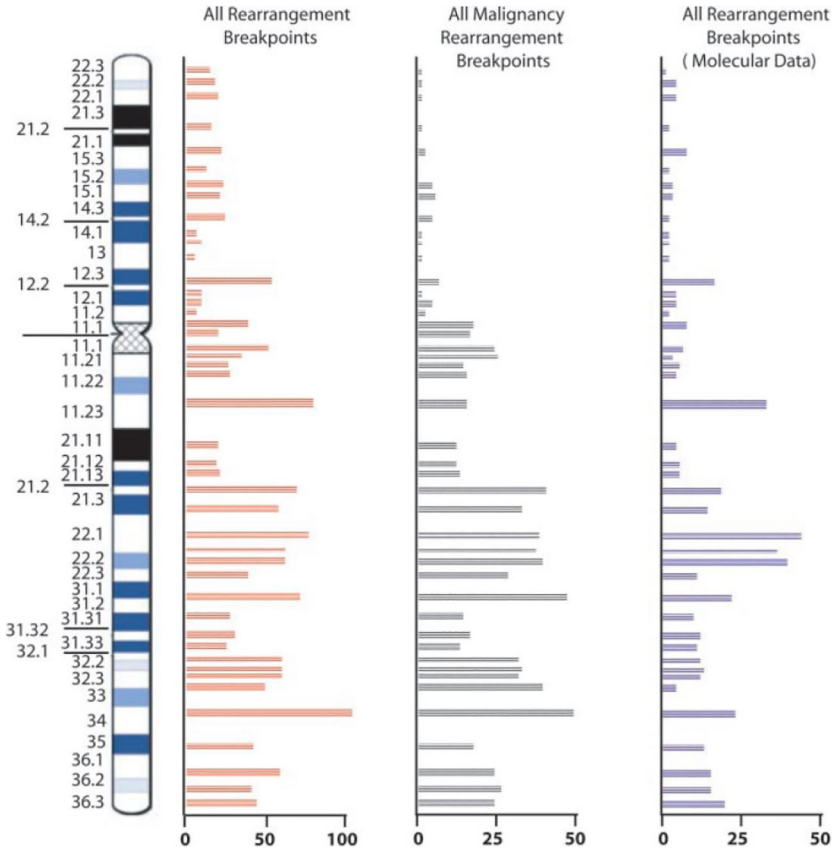
equal number of molecular biologists, medical geneticists, and physicians utilize the information.”<sup>54</sup>

The project culminated with the article describing the “sequence” and “biology” of chromosome 7, which appeared in *Science* in May 2003 (in our dataset: PMID 12690205). Its list of authors included forty-three different affiliations, among them Celera and SickKids as main institutional contributors. Only two of all the co-authoring institutions had been involved in the draft whole-genome publications two years previously: Baylor College of Medicine in the IHGSC *Nature* paper and Yale University School of Medicine in the Celera-led *Science* paper. The chromosome 7 article had a different scope and did not limit itself to the description of the sequence. Apart from the 2001 data, Celera contributed a number of unpublished scaffolds—sembled DNA sequences—that their scientists mapped on the chromosome. The rest of the signing institutions were human and medical genetics groups who either belonged to Scherer’s network or had attended the jamboree. They provided information that enriched Celera’s horizontal sequence with vertical variants associated with CF or other chromosome 7 conditions, such as Williams-Beuren Syndrome or Shwachman-Diamond Syndrome. To identify these regions, the rearrangement breakpoints of chromosome 7 were crucial: Morton and Gusella, as mappers of these points in all human chromosomes, contributed to the paper as co-authors from their HMS-MGH, and HMS-Brigham and Women’s Hospital groups (see figure 7).<sup>55</sup> Overall, the paper gathered the efforts of teams and consortia working on over thirty diseases in nine North American, European, and Asian countries.

In July 2003, *Nature* published another full sequence of chromosome 7 as a co-authored article led by the genome center at WU (in our dataset: PMID 12853948). The rationale of this article substantially differed from that of Scherer and Mural, although its publication occurred only two months afterward. Apart from WU and their sequencing partners at NHGRI, most of the other co-authors of the *Nature* publication were from institutions that had

54. All quotes from [www.chr7.org/project.php](http://www.chr7.org/project.php).

55. Stephen Scherer, interview with Miguel García-Sancho, SickKids, Toronto, April 2018. See also Stephen W. Scherer, Joseph Cheung, Jeffrey R. MacDonald, Lucy R. Osborne, Kazuhiko Nakabayashi, Jo-Anne Herbrick, Andrew R. Carson, et al., “Human Chromosome 7: DNA Sequence and Biology,” *Science* 300, no. 5620 (2003): 767–72.



**FIGURE 7.** A diagram of the location of the breakpoints in which chromosome 7 experiences structural rearrangements, as reported in the article describing its full sequence and biology. Cynthia Morton, James Gusella, Gail Bruns, and other colleagues at teaching hospitals affiliated with Harvard Medical School identified the breakpoints. Stephen Scherer and his network of collaborators within and outside the University of Toronto Hospital for Sick Children associated some of these breakpoints with malignancy (genetic diseases). Richard Mural and other scientists at Celera Genomics connected the breakpoints with their sequence and other molecular data. Researchers could explore the results in a genome browser that displayed together and interlinked the mapping information, DNA sequence, and clinical annotations: Stephen W. Scherer, Joseph Cheung, Jeffrey R. MacDonald, Lucy R. Osborne, Kazuhiko Nakabayashi, Jo-Anne Herbrick, Andrew R. Carson, et al., "Human Chromosome 7: DNA Sequence and Biology," *Science* 300, no. 5620 (2003), 767–72, 771. Republished with permission of the American Association for the Advancement of Science, copyright © 2003; permission conveyed through Copyright Clearance Center Inc.

contributed to the IHGSC first draft sequence; Case Western Reserve University in Cleveland additionally contributed to Celera's draft genome.<sup>56</sup> The *Nature* publication was part of an agreement with the journal, according to which a mapped and fully polished sequence of each human chromosome would follow from the 2001 first draft. It pursued only the refinement of the horizontal reference sequence of chromosome 7, with no significant vertical exploration. This narrower focus resulted in a correspondingly thinner co-authorship list and a smaller bridging role in our network: the WU-led article displayed eight institutional affiliations, seven from the United States and one from Germany.

The timing of the two chromosome 7 articles, and the fact that one was co-authored by Celera and the other by institutions from the IHGSC, seemed to emulate the race between these two counterparts in the sequencing of the human genome. Yet if one looks at these publications from outside the boundaries of the HGP—as our co-authorship network has enabled us to do—the one led by Scherer and Celera had a deeper historical significance. By looking at the collaborative annotation project behind this article, we could see the trajectories of vertical and horizontal sequencing, two different but intersecting approaches that the large-scale center model of the HGP had attempted to separate. This separation, however, was fully operational only within the confines of the G5 institutions: Celera kept an umbilical cord with the vertical sequencers that resulted in the chromosome 7 co-authorship following the publication of the first draft of the human genome. This suggests that the large-scale center model, especially as applied during the conclusion of the HGP, represents only a partial picture of human genomics. Beyond this organizational and historiographical framework, we can critically interrogate the perceived discontinuities between the production of the human genome sequence and the use of its data in medical genetics research.

56. According to an NIH co-author, the interest of some of the contributors who were not based in genome centers was exploring the evolutionary implications of comparing sequences across species: Matthew Portnoy, interview with Miguel García-Sancho, National Human Genome Research Institute, Bethesda, Maryland, November 2018. Elsewhere in this special issue, we identify this practice of interspecies comparison—which became prevalent in pig and other animal genomics as well as human genomics—with an extensive as opposed to intensive way of conducting DNA sequencing: James Lowe, Rhodri Leng, Gil Viry, Mark Wong, Niki Vermeulen, and Miguel García-Sancho, “The Bricolage of Pig Genomics,” this issue.

## 6. CONCLUSIONS

In a telling ethnographic vignette, historian and ethnographer Hallam Stevens discusses two separate buildings with wholly different architectural schemes at the Broad Institute, founded in 2004 as a development of the Center for Genome Research at the Whitehead Institute. The “back region” is a closed building functioning as a genome center and structured like a manufacturing facility that devotes itself to the production of DNA sequences. The “front region” is the main public-facing and open Broad Institute building, and consists of offices and laboratories that make use of the data.<sup>57</sup> However much these realms are designed to work in tandem at the Broad Institute, their physical separation represents the conceptual and practical distances between the production and use of sequence data that emerged and consolidated during the HGP. Rather than be confined by this division, we have sought to open up historical inquiry into human genomics that moves beyond a purely contingent differentiation.

Throughout this paper, we have approached the HGP as generating and instantiating a particular—dominant and influential—model of conducting genomics rather than being constitutive of genomics itself, or the primary model and object for the history of genomics. We have demonstrated this by exploring a dataset and a co-authorship network and using new categories to interpret these analyses. Through examining publications that drew our attention in the analysis of the network, we have concluded that the separation between sequence producers and users that the HGP predicated operated only within a reduced number of institutions and limited time range: twenty large-scale genome centers and, especially, five institutions—the so-called G5—that led the completion of the publicly and charitably funded draft reference human sequence between 1996 and 2001.

The majority of our co-authoring institutions explored the medical implications of the sequences they had produced in the publications. Other historians had documented this entanglement between sequence production and use but without detailing how it operated within genomics research. Our historical interpretation of the co-authorship network has enabled us to identify and detail a collaboration between medical genetics laboratories and Celera Genomics, a large-scale sequence producer.

57. Stevens, *Life* (n.17), esp. chap. 3.

The study of this collaboration has led us to propose the categories of vertical and horizontal sequencing as alternatives to sequence production and use. The medical genetics laboratories were and continued to be producers of vertical sequences, despite appearing to be just users of the reference human genome during and after the last stages of the HGP. Apart from compiling sequences, this vertical approach gathered information about variability in the data and its potential clinical implications. Celera Genomics, though it was a large-scale producer that competed with the G5 and other participants in the HGP, never lost sight of more proximal potential medical uses of its sequences. They generated horizontal sequences that encompassed much broader genome areas than those the vertical medical geneticists produced. Yet the expertise these vertical sequencers embodied and the insights they offered about clinical variability were essential for connecting Celera's sequence with medical problems.

Within our vertical and horizontal categories, the two chromosome 7 publications that our network helped us to identify emerge as tips of an iceberg. The one that Celera co-authored reflects historical synergies between medical genetics and genomics that are embodied in the ways this company interacted with the other contributing institutions. The one led by WU shows the exceptionality of the funding and organizational arrangements of the last stages of the HGP. Rather than representing the history of human genomics or genomics more generally, we have portrayed the success story of the HGP and the later publication of detailed chromosomal sequences as the skillful execution of a well-funded (and time-limited) model of organizing genomics involving only a handful of institutions.

The distinctness of the HGP model is even more visible in nonhuman genomics, as the other papers of this special issue show.<sup>58</sup> Within the history of human genomics, however, it is worth noting that the (apparent) increasing absence of medical geneticists may be an artifact of taking the contingent HGP story as a whole. This apparent absence may have also created the impression of a translational gap in which medical geneticists and other external *users* would need to find applications for the reference sequence once the *producers* had made this data available in the public domain.

58. García-Sancho et al., "Yeast Sequencing" (n.12); Lowe et al., "The Bricolage" (n.56)—these publications direct more attention toward European genomic endeavors than this mainly North American-focused paper.



Medical geneticists were never fully absent if we change our historiographical lens and approach the production of the human genome sequence from the perspective of Celera rather than the G5 or any other participant in the HGP. From this standpoint, Celera is no longer the private counterpart of the HGP story but an institution that pursued the potential clinical uses of the sequence with a more proximal relationship to the community of medical geneticists. These potential uses became actionable through collaborations in which the horizontal sequence of Celera became entangled with mapping information and other clinically pertinent data that the vertical sequencers had produced. Uncovering this reciprocal process and the multiple dimensionalities it confers to the sequences—rather than assigning absolute priority to their horizontal determination—enables a better understanding of the translation of genomic data.

### **Acknowledgments**

See a full list of people and institutions whose support has been essential at the end of the introductory article of this special issue “The Sequences and the Sequencers: What Can a Mixed-Methods Approach Reveal about the History of Genomics?” The research and writing of this paper were sponsored by 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. For more details on the project, see <https://transgene.sps.ed.ac.uk/>.