

The Human Genome Project: A Player's Perspective

Maynard V. Olson

Department of Medicine, UW Genome Center, University of Washington, Box 352145, Seattle, WA 98195, USA

The Human Genome Project was a natural culmination of one of the great scientific triumphs of the 20th century—the elucidation of the means by which biological organisms store, replicate, and process information. Specifically, the discovery at mid-century that biological information is stored as a linear, digital code led directly to the concept of a genome sequence. Technical advances in DNA analysis during the second half of the 20th century made the determination of genome sequences a practical goal. The achievement of this goal, at the end of the century, occurred during an extraordinary confluence of rapid scientific progress, rampant technological optimism, and exuberant entrepreneurial capitalism. The resultant strains on basic scientific values are exemplified by the public–private competition that arose during the sequencing of the human genome. While this competition accelerated initial availability of the genome sequence, it did so at considerable cost to the health of the interface between science and society. Analysis of this episode may reveal important lessons since the human future will continue to be shaped by the same forces that were at play during the endgame of the Human Genome Project.

There are two stories of the Human Genome Project. One describes a century of scientific progress that began with the rediscovery of Mendel's laws in 1900 and ended in a frenzy of genome sequencing. The other is a story about contemporary societal values—particularly, those that framed the project's endgame and continue to shape public perceptions toward this defining event of our time. Both stories deserve attention. The first will help counter the illusion that the Human Genome Project was a sudden inspiration of the 1990s, while the second offers a sobering look at the forces that shape the current interface between science and society. We should pay attention to the ethic of this frontier since—like that of the American West—its influence will linger long after the present exuberance has faded.

I will start with the scientific story of the first triumphant century of genetics. In this narrative, the Human Genome Project will emerge as a natural step in the scientific quest to understand one of nature's deepest mysteries: How can a fertilized egg cell, an object too small to see with the unaided

eye, contain all the information required to guide the development of a unique human being?

Within a few years of the rediscovery of Mendel's laws, an initial synthesis emerged that linked Mendel's probabilistic rules of inheritance to the internal structure of cells. This synthesis placed the material basis of Mendelian traits on the chromosomes, tiny bodies within cells that are readily stained with colored dyes. Chromosomes are duplicated every time a cell divides and are distributed with extraordinary precision to the two progeny cells. Through the life cycle of an organism, as it builds itself through successive divisions of a fertilized egg and ultimately produces new egg or sperm cells, chromosomes obey precisely the same rules of transmission as those that govern the inheritance of Mendelian traits. Hence, early geneticists quickly inferred that chromosomes must mediate the transmission of Mendelian inheritance. But how? Within the scientific worldview of the early 20th century there was simply no plausible explanation. Chromosomes, as then known to cell biologists, were tiny, uniformly staining objects comprised of a featureless material referred to as "chromatin." This monotonous substance seemed an unpromising carrier of the rich universe of biological traits displayed by elephants, orchids, and human beings. William Bateson, an early geneticist, expressed this perspective in a 1916 review of the classic book *The Mechanism of Mendelian Heredity*:¹

... it is inconceivable that particles of chromatin or of any other substance, however complex, can possess those powers which must be assigned to our [Mendelian] factors ... The supposition that particles of chromatin, indistinguishable from each other and indeed almost homogeneous under any known test, can by their material nature confer all the properties of life surpasses the range of even the most convinced materialism.

We have developed such disdain for "vitalism," the idea that non-material forces act within living organisms, that we forget how recently biology has developed plausible materialist explanations for basic life processes. Bateson had no quarrel with the scientific evidence presented in *The*

Mechanism of Mendelian Heredity. He simply could not conceive of how such tiny, uniform objects as chromosomes could “confer all the properties of life” on a newborn baby. The critical phrase in Bateson’s commentary is “almost homogeneous under any known test.” How could the stunningly intricate structures of life be patterned by microscopic blobs of structureless matter?

This most basic of biological questions remained unanswered until mid-century, when developments in two seemingly unrelated fields—molecular genetics and information theory—laid the groundwork for the Human Genome Project. In biology, the seminal event was the discovery of the structure of DNA in 1953. Earlier data had pointed to DNA—a polymeric molecule comprised of a simple chain of four different monomers—as the chemical in chromatin that encoded genetic information. Watson and Crick solved the structure of DNA, which proved to be an austere beautiful molecule composed of two helically interwoven strands. This structure can accommodate any one of the four monomer units, conventionally symbolized as G, A, T, and C at any position in the polymer chain. In deference to their chemical properties, the G’s, A’s, T’s, and C’s in a DNA chain are referred to as “bases.” Complementary bases on the two strands of the double helix pair with one another, a structural arrangement that plays a critical role in the precise copying of the DNA chains at each cell division. Initial excitement over the double helix centered on this copying mechanism. Remarkably, it was only in a single sentence of a follow-up paper that elaborated on the initial description of the double helix that Watson and Crick pointed out—almost off handedly—that the problem of biological information storage appeared also to have been solved.²

It follows [from the properties of the double helix] that in a long molecule many different permutations are possible, and it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information.

An actual patch of the human genome sequence can be represented in the following notation:

... ACCCTCTTCAGAGCTGCACTAGACATTCA-CAGGGAATTGAGGTTCCCTAT ...

This patch of monomer sequences defines the structure on only one strand of DNA. An elegant feature of the double helical structure is that the sequence of monomers on one strand allows a simple prediction of the sequence on the other since an A is always opposite a T and a G opposite a C. These A–T and G–C pairings are referred to as “base pairs.” The informational redundancy is the key to the copying mechanism since either strand can guide the synthesis of its complement when copying occurs.

There is an obvious analogy between the notation for DNA sequence and the notation commonly used to describe information stored in computing devices. Computers rely on a simple base-two code of 0’s and 1’s. In contrast, the genetic code is a “base-four” code of G’s, A’s, T’s, and C’s, which could equally well be written in mathematical, rather than chemical, notation as a series of 0’s, 1’s, 2’s, and 3’s. For example, if we represent G as 0, A as 1, T as 2, and C as 3, the segment of human DNA sequence shown above would be written as the following base-four number:

... 13332322310103203132101312231310001122010022333212 ...

Obvious as the analogy between DNA sequence and the digital information of computers is to us, it was largely foreign to the biologists of the early 1950s. The mathematical and conceptual underpinnings of “information theory” were first clearly articulated by Claude Shannon in 1948, more than 30 years after Bateson peered through a microscope at featureless strands of chromatin and only a few years before the discovery of the double helix.

The Human Genome Project is the direct descendent of the wholly unexpected confluence of genetics and information theory. In a 1954 *Nature* paper, the cosmologist George Gamow pointed out, apparently for the first time, that “the hereditary properties of any given organism could be characterized by a long number written in a four-digital system.”³ The term “four-digital,” soon to be replaced by “base four,” sounds quaint to the modern ear. This archaism is a colorful reminder that both molecular biology and information theory were then young. The confluence of genetics and computer science must rank as one of the great coincidences in the history of science and technology. In the same historical instant, humans discovered that biological information is digital—a mechanism of information storage and processing that evolved within cells over billions of years—and, quite independently, invented new technological means of storing, processing, and transmitting information based on digital codes. Thus, the two technological forces that are most profoundly reshaping the future of human culture—genetics and computing—are linked at their historical and conceptual roots.

The goal of The Human Genome Project is simply to find out, for our own species, what Gamow’s “long number written in a four-digital system” actually is. During the 1950s and 1960s, molecular biologists worked out the basic mechanisms by which the digital information within cells is copied and read. In the 1970s, with the advent of recombinant-DNA techniques, there was an explosion of technical capabilities to analyze and manipulate DNA in the laboratory. By the late 1970s, practical methods to sequence the monomers on a DNA strand had been developed. Still, before the Human Genome Project was to

become feasible, enormous increases in the efficiency of DNA sequencing were required. Even in 1980, a ten-thousand-monomer sequencing project was a large undertaking, while the sequencing of the human genome's three billion monomers was off scale as a practical goal. Nonetheless, the efficiency of DNA analysis increased steadily in the early 1980s and several of biology's visionaries began to advocate an all-out assault on the human genome. Adopting a moderate position on the feasibility of such a project, a 1987 report of the U.S. National Research Council's Committee on Mapping and Sequencing the Human Genome reached the following conclusions (Ref. 4, p. 2):

- Acquiring a map, a sequence, and an increased understanding of the human genome merits a special effort that should be organized and funded specifically for this purpose. Such a special effort in the next two decades will greatly enhance progress in human biology and medicine.
- The technical problems associated with mapping and sequencing the human and other genomes are sufficiently great that a scientifically sound program [will] require a diversified, sustained effort to improve our ability to analyze complex DNA molecules. Although the needed capabilities do not yet exist, the broad outlines of how they could be developed are clear.

Over the following decade, the scientific and organizational foundations of human-genome sequencing were built. The scientific foundations largely involved the sorting out of candidate technologies, pilot projects on organisms with much smaller genomes than the human, and the development of detailed maps of the human chromosomes. These maps, which identified the positions of landmark sequences along the chromosomes, would ultimately guide the assembly of the full sequence. In the United States, the organizational foundations of the Human Genome Project included appropriation of major funding by the United States Congress and development within federal agencies of mechanisms to distribute peer-reviewed funding to geographically dispersed investigators. Parallel activities occurred in several other countries, most notably in the United Kingdom, where the Wellcome Trust made a major commitment to the project. A loosely knit international coalition evolved to provide overall coordination. By 1995, both the technology and the organizational framework looked sufficiently mature to go ahead. I articulated the case for action in a Policy Forum in *Science* entitled "A Time to Sequence."⁵ The allusion to *Ecclesiastes* in this title, with its implication that the sequencing of the human genome would fit comfortably into the natural rhythms of life, proved even more ironic than intended. As events developed, the culminating phase of the Human Genome Project was

to ride an expanding bubble of financial and technological optimism that was without precedent in human history—the great high-tech boom of the late 1990s.

The rapid expansion of this bubble, starting in 1997, provides an appropriate point for me to shift from the scientific to the societal stories of the Human Genome Project. Indeed, by this time, there was little need for new science or technology. What was left was a mad rush for the spoils.

I was at a meeting in April, 1998, of a group charged with drafting a plan for the next few years of the National Human Genome Research Institute's activities, when a report circulated that an announcement of a major private-sector initiative to sequence the human genome was imminent. While details were vague, I was not surprised by this development. I had thought for some time that conditions were ripe for a cream-skimming effort by a new-economy company seeking to harvest intellectual property from the human genome. The technology had stabilized to a point where much of the human-genome sequence could be determined for a few hundred million dollars, and the necessary capabilities could be deployed by any medium-sized biotechnology company. Less clear, was how such an initiative could be embedded in a viable business plan. But this was the late 1990's, and business plans did not need to be viable to attract billions of dollars of private investment.

On May 9, 1998, Perkin-Elmer, the parent company of Applied Biosystems—a major manufacturer of instrumentation and reagents for DNA sequencing—announced a plan to sequence the human genome with private funding. Craig Venter, a maverick molecular biologist who had long thrived by relying on investment capital to support genome-analysis projects, was to lead the effort. The new venture was ultimately incorporated under the name Celera, which is derived from the Latin word for speed. From the start, Celera enveloped its activities in a noisy public-relations campaign. A key ally proved to be the *New York Times* reporter Nicholas Wade. Wade scooped other journalists with the initial story and then, over the next three years, became such an uncritical component of Celera's public-relations campaign that there was often little distinction between front-page stories in the *Times* and unfiltered corporate press releases.

Celera's initial claims were too good to be true. The company's technology was advertised as so advanced and its management so bold that the human-genome sequence could be produced at a fraction of the cost and time projected by the public effort, while maintaining rigorous quality control and unencumbered public access to the data. Celera would determine the sequence largely to showcase its technological capabilities. A modest number of "gene patents" and sheer technological momentum would provide ample benefits for Celera's investors. There was simply no reason for the public project to continue. Venter suggested

that the public scientists should just move on to the mouse genome and leave the human to him. Wade's initial reporting indicated that the leadership of the public effort in the U.S. was ready to throw in the towel.^{6,7}

The events of the spring of 1998 are a sobering lesson of how a public-relations campaign with deep pockets that is reinforced by an uncritical press can hijack public perceptions of scientific developments that are of profound social interest. It was apparent from the first days of the controversy that the efforts of individual scientists to joust with Celera's public-relations blitz were counter-effective. Celera successfully cast the story as one of a maverick genius against an unimaginative, turf-conscious establishment. With Venter spewing extravagant sound bites at every turn, academic experts who attempted to explain the cumbersome realities of genome sequencing to the press simply added fuel to Celera's story. They were just playing the roles for which they had been cast.

Dismal as this situation was for supporters of a vigorous public role in the Human Genome Project, there were a few bright moments. While the scientific community and the press performed poorly, some of science's friends took up the slack. A key stalwart was the leadership of the Wellcome Trust in the U.K., which committed itself within days to increase funding for its critical component of the public effort. In the United States, the star role was played by science's supporters in the U.S. Congress.

While many scientists like to malign the ignorance, short-sightedness, and indifference to academic values of political leaders, U.S. politicians have been consistently ahead of the scientific community in grasping the importance of the Human Genome Project. This pattern was established in the mid-1980's during the project's infancy. With the exception of a few visionary leaders such as Robert Sinsheimer and James Watson, the scientific community simply failed to recognize the way in which genome sequences could create an enlarged world of biomedical research. Scientists were preoccupied with the risk that any re-slicing of the resource pie would disadvantage their favorite projects. Indeed, the intensity of scientific opposition to the Human Genome Project is remarkable given that the project was only slated, at peak operation, to absorb 2% of the budget of the National Institutes of Health. In contrast to scientists themselves, science's supporters in Congress consistently recognized that the Human Genome Project addressed scientific questions of compelling interest.⁸

In the spring of 1998, this reservoir of Congressional enthusiasm for the Human Genome Project played a critical role. This point is nicely documented by the transcript of a hearing of the Subcommittee on Energy and Environment of the Committee on Science of the U.S. House of Representatives, which was held on June 17, 1998. I

testified at this hearing, along with Craig Venter, Francis Collins, and others. Because of the critical timing of the hearing, little over a month after the initial Celera announcement, the hearing's transcripts are an important primary source in the archives of the Human Genome Project.⁹

The purpose of the hearings was to explore how the Celera announcement should affect the public program. Congressman Calvert, the Subcommittee Chairman, set the stage in his introductory remarks (Ref. 9, p. 1):

CALVERT—As the 15-year, \$3 billion federal program reached its halfway point this year, the scientific world was stunned on May 9 when one of the country's foremost genetic scientists, Dr. Craig Venter, and the Perkin-Elmer Corporation announced they would [launch] a new venture to, as they put it, "substantially complete the sequencing of the human genome" in 3 years at one-tenth the cost of the Federal program. Just how this should affect the government program is the focus of this hearing today. Press reports and some back and forth between critics and supporters of the federal program have raised as many questions as it has produced answers. For example are the goals of the initiative realistic or just an optimistic vision? Will this private sector initiative duplicate the Federal program and make it redundant or is it another approach that can complement the Federal program and make it stronger?

The testimony that followed, particularly during the question-and-answer session, captured a critical moment in the history of the Human Genome Project. On the issue of Celera's too-good-to-be-true plan to sequence the genome at its own expense and release the data to the public, Venter was unequivocal. In his prepared testimony, he stated that "an essential feature of the new company's business plan is to provide public availability of the sequence data (Ref. 9, p. 6)." He continued with the following statement:

VENTER—Because of the importance of this information to the entire biomedical research community, key elements of this database, including primary sequence data, will be made available. In this regard we will work closely with national DNA repositories like the National Center for Biotechnology Information.

The reference to the NCBI is of special importance since this agency is the U.S. curator of Genbank, the public repository of DNA sequence data, which provides all its data free to anyone who queries the NCBI's World Wide Web site, no questions asked. Beyond assuring the subcommittee that Celera's data would go to Genbank,

Venter committed Celera to frequent releases of data as the project developed: "It is our plan to release data into the public domain at least every 3 months including the complete human genome sequence at the end of the project (Ref. 9, p. 6)." There was much discussion of this claim during the hearing. At one point, Venter raised the stakes to an ethical level: "we feel morally compelled to release [the] genome sequence to the entire public (Ref. 9, p. 78)." At another point, he stated that "the raw sequence itself will be provided to the world for free (Ref. 9, p. 79)."

Francis Collins, the Director of the National Human Genome Research Institute of the National Institutes of Health, raised a prophetic point when he noted that, since nothing in Venter's testimony was in any way binding on Celera, the only way to assure free, unrestricted access of the public to the human-genome sequence, was to continue with the public project (Ref. 9, p. 80):

COLLINS—I believe having the public effort continue to be vigorously involved in [human-genome sequencing] as much or more so than they have been, is also the best insurance that the data is made publicly accessible. I do not question for a moment Dr. Venter's sincerity in his statement that his data will be made available on a quarterly basis in a database that anybody can look at. I know that that is what he is committed to doing. But, after all, the sequence of the human genome is of such profound importance, that I think a scenario where large quantities of it were only available within the database of a single private entity might be a rather unstable situation. If business demands were to change or personnel were to change or the stockholders were to decide its not such a good thing to be giving this all away anymore, one would not want to see a circumstance where the publicly-funded effort was suddenly found to have dropped the ball. We don't intend to drop the ball.

Three years after Collins's extemporaneous testimony, none of Celera's data has been released to Genbank, the company is attempting to subsist on database subscriptions, and even academic researchers who want single-query access to Celera's data must execute elaborate legal agreements with the company to acquire it. I will leave it to historians to judge Venter's "sincerity" in promising free, unrestricted public access to the sequence. Clearly, one possibility is that Celera's game plan, from the beginning, was a classic "bait-and-switch" scam. In this scenario, the company's strategy was to use the promise of free, unrestricted access to the data to undercut support for the public project and thereby set the stage for a lucrative monopoly in selling the sequence on a fee-for-service basis. Alternately, in Collins's words, changes in "business demands" or "person-

nel" at Celera may have led the company "to decide its not such a good thing to be giving this all away anymore." From a science-policy perspective, the distinction matters little. The real lesson is that if science is to remain an open, progressive force in society, in which scientists continually build on the insights of their predecessors and subject their own findings to the open review of their peers, governmental agencies and philanthropic organizations must pay the costs of basic research.

The need for public control over basic science goes beyond the question of the free, unrestricted access to data. There is also the question of who controls the scientific process through which the data are collected. For example, at the 1998 Congressional hearings, there was much discussion of the need to insure a high quality standard for the human-genome sequence. The prevailing quality-control standards of the public project were not in dispute. The live issue was whether or not Celera's technical strategy would be likely to meet these standards. Normally, discussions of an arcane technical question of this type would be limited to a small group of experts. However, the Celera public-relations apparatus had managed to make the relative merits of different sequencing strategies into front-page news. The company did so by portraying Venter as the unappreciated genius behind a radically new technology—the "whole-genome" approach—that was alleged to be a revolutionary improvement over previous methods. In contrast, the public project was portrayed as a bureaucratic quagmire of old ideas that was sticking to its "clone-by-clone" approach simply out of stubbornness and a general inability to change with the times. The technical issues associated with "whole-genome" *versus* "clone-by-clone" sequencing involve a level of detail that would normally be of little interest to the general reader. However, the supposed merits of whole-genome sequencing were so central to the Celera mystique that the controversy surrounding this method requires brief explication.

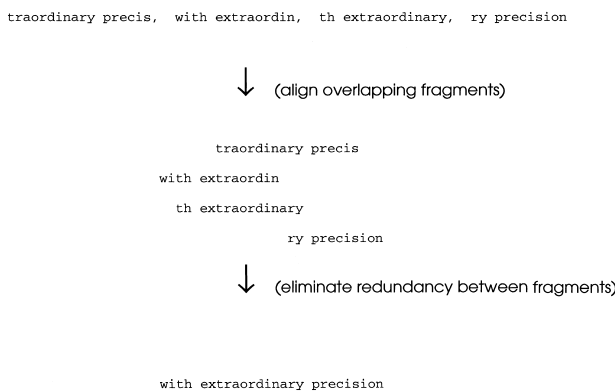
All contemporary sequencing methods depend on acquiring short tracts of DNA sequence in the form of raw sequencing "reads." A single read defines the sequence of approximately 500 of the 3,000,000,000 base pairs of DNA present in the human genome. The composite sequence of the genome must be built up by "assembling" tens of millions of raw sequencing reads. Collectively, these reads "oversample" the genome (i.e. a typical short sequence is sampled 5–10 times in different reads, each of which has a different starting and ending point). The whole-genome *versus* clone-by-clone controversy relates to how the raw reads are sampled from genomic DNA. In the whole-genome method, the reads are sampled directly from an organism's DNA. Hence, a particular 500-base-pair read is equally likely to come from any site in the genome. It would take 6,000,000 reads, laid end-to-end, to cover the genome. However, since the reads have randomly positioned end

points, it is necessary to sample 5–10 times that number in order to insure that few unsampled gaps are left. This oversampling also insures that most reads will overlap extensively with their nearest neighbors, thereby providing the information required to assemble the reads into a long, contiguous sequence.

These concepts are most easily grasped by analogy. Consider the following short fragments of text that have been sampled at random from this article:

endel's laws i, traordinary -
 precis, se in Bateson's c

These three samplings are little more than “teasers”—they provide glimpses into the article’s content but no clues as to how the sampled fragments relate to one another. Furthermore, most of the article’s content is simply missing. However, if a sufficient number of random samplings were taken, overlaps would begin to occur between different samplings and, ultimately, every bit of text would have been sampled many times. From a data set of this type, it would be possible to reassemble the whole chapter with minimal ambiguity. The example below illustrates how the assembly would work in a local region:



It would take approximately 30,000 random samplings of the length shown to allow a relatively good assembly of the whole article.

“Whole-genome” assembly is a method analogous to carrying out the initial sampling from the entire article. Since this article contains relatively few repetitions of character strings as long as the 10–20-character fragments shown above, a whole-article assembly from fragments of this length would work reasonably well. However, approximately half of the human genome is comprised of recognizably repetitive segments of DNA sequence, many of which are longer than typical sequencing reads. For this reason, the publicly funded Human Genome Project adopted a “clone-by-clone” approach to assembly. In clone-by-clone sequencing, the genome is initially broken into modules much larger than a sequencing read by recombinant-DNA methods. In the specific implementation of clone-by-clone sequencing

employed by the public project, the number of these modules, which are referred to as “clones,” is approximately 30,000. The order of these clones, is determined in advance of the sequencing by DNA-mapping methods that are far less expensive than complete sequencing. Once mapped, the clones are sequenced individually and the sequences are combined to produce the overall sequence of the genome. This approach is analogous to reassembly of the text of a large encyclopedia from a page-by-page sampling into fragments rather than from a single sampling of all pages in all volumes.

The actual sequencing is carried out in the same way in both methods and the total number of reads required is also approximately the same. Most of the cost of sequencing is in acquiring these reads, so whole-genome sequencing is only moderately quicker and cheaper than the clone-by-clone method. The enormous advantage of clone-by-clone sequencing is that modular sampling allows inevitable problems with the final assembly to be resolved locally since all reads associated with a particular clone are constrained to come from a small, contiguous segment of the genome. In contrast, in the whole-genome strategy, reads are easily misplaced entirely. Indeed, there is often so much ambiguity in their placement that no assignment is possible.

Venter’s claim to having “invented” whole-genome sequencing is based on his leadership of a project to sequence a tiny bacterial genome that was nearly devoid of repeats. Indeed, in several years of sequencing bacterial genomes, Venter’s group had never carried out a successful whole-genome assembly on a genome that was even 0.1% the size of the human genome, and his group had little track record in dealing with the ubiquitous repeats of mammalian genomes. Indeed, the leading proponent of whole-genome sequencing of the human had been a little-known, but highly innovative geneticist named Jim Weber. Weber had teamed up in 1997 with an assembly expert named Gene Myers, who was later to lead the assembly effort at Celera, and published a thoughtful article advocating whole-genome sequencing as an initial strategy in the Human Genome Project.¹⁰ A simultaneous statement of the case against the whole-genome approach was published by another expert in sequence assembly, my colleague Phil Green.¹¹ The Weber/Myers *versus* Green debate exemplified the traditions of open, peer-reviewed science. Both parties brought obvious expertise to the discussion and presented objective arguments for their positions. The strength of the Weber/Myers proposal was that it would produce more useful biological data faster. Its weakness was that it would leave a poorly assembled genome even after much of the total cost of the project had been expended, and the options for cleaning up the assembly were unattractive. All these points were clearly articulated in the relevant publications and extensively discussed among

participants in the Human Genome Project. The consensus conclusion was that clone-by-clone sequencing was the better strategy, a position that was vindicated by the technical difficulties that Celera ultimately encountered.

Given the history of this debate, the public-relations blitz in the spring of 1998 announcing Venter's development of whole-genome sampling as a dramatically improved approach to human sequencing was met with some surprise by experts in DNA sequencing. The issue was on full display at the June 17 Congressional hearings.⁹ In my testimony, I decided to wade briefly into the technical details. My goal was simply to illustrate, by example, the difference between scientific and political dialog (Ref. 9, p. 56):

OLSON—I, frankly, am a skeptic that the approaches as publicly described will lead to a product of sufficient quality to meet the long-term needs of the scientific community. I'm prepared to be proven wrong, as any scientist must be, but I am comfortable predicting that this approach, as the downside of its efficiency, will encounter reasonably catastrophic problems at the stage [at] which the tens of millions of independent sequencing tracks need to be melded together to produce a composite view of the human genome. To be specific, I'm comfortable predicting that there will be over 100,000 serious gaps in the final product and in this context, I define a serious gap as one in which there is uncertainty even as [to] how one should orient and align the islands of assembled sequence between the gaps. Furthermore, I'll predict that a substantial fraction, particularly [of] the smaller islands of sequence [produced], will be misassembled, that is they will not actually correspond to the organization of the human genome ...

I had no intention of educating members of Congress, or even their staffs, about the minutiae of DNA sequencing. I simply wanted to make clear that there were some experts who sharply challenged Venter's representation of the superiority of his methods. And, I wanted to go on record with a prediction that was sufficiently specific that it could be falsified by subsequent data if I was wrong.

I was exceptionally fortunate that one member of the subcommittee was Congressman Vernon Ehlers, a conservative Republican from Michigan. Ehlers was the only member of Congress with a Ph.D. in science (physics) and substantial research experience. Despite the distance of the discussion from his own scientific background, he immediately grasped the essence of the issue. Addressing Craig Venter, Ehlers pursued the following line of questioning (Ref. 9, p. 77):

EHLERS—Let me ask another question. I've

done some experiments which demand extreme precision, parts [in] ten to the ninth [*i.e., experiments demanding a precision of a few parts per billion*], and very, very careful work over some time. I've also done some which are called quick and dirty where you are just trying to outline the parameters of something to decide whether or not there is something worth investigating there. Is that, in a sense, the difference between the so-called human genome project and your work?

VENTER—Absolutely not. In fact, I appreciate you asking that question. Quick does not mean dirty. Quick means better technology, better approaches, new strategies. We're going to be sequencing the human genome 10 times [*a reference to the redundancy of sampling inherent in all current approaches to genome sequencing*]. The sequences that we've done in the past are some of the most accurate sequences ever put in the public domain by any scientist and we're going to have the same standard for the sequences that we do with the human genome ...

EHLERS—So your statement would be that your method is going to yield results with the same completeness and the same accuracy as the Human Genome Project?

VENTER—We actually feel that our approach is going to yield more completeness and at least the same level of accuracy as done by the best groups ...

EHLERS—So, basically, what I hear you saying is it's not the contrast between the precise, complete experiment and the quick-and-dirty experiment but rather the contrast between a bureaucratic risk-free approach and a more thoughtful modern approach.

VENTER—I think that would characterize my view quite well.

Ehlers knew exactly what he was doing in this exchange, and his effectiveness at making his point illustrates the importance of having scientists in critical positions of public responsibility. No other Congressman was going to challenge Ehlers on an issue of research methodology. He closed with the following statement (Ref. 9, p. 79):

EHLERS—Thank you. I find this very interesting and, as Dr. Collins observed, this is an experiment. I will be very interested in seeing the results of the experiment and it will be fun to get you back in about 3 or 4 years and read your prepared testimony and your answers back to you at that point.

When he said that, I knew that the Celera strategy to push the public sector out of the Human Genome Project would fail. Ehlers wanted to see how the Celera *versus* public-sector competition played out, and he was willing to appropriate the tax money of his constituents to insure that

the needed data were collected. I would suggest that scientists create an Ehlers award for politician–scientists who use their scientific experience and political power to support basic scientific values.

We have not yet had the final reckoning that Congressman Ehlers anticipated with such interest. However, some comments can now be made about this episode with considerable benefit of hindsight. After two-and-one half noisy years in which public interest in “who was ahead” in the race to sequence the genome never seriously lagged, rival publications appeared in *Nature* and *Science* in February, 2001.^{12,13}

These two publications were governed by wholly unequal rules of engagement. In contradiction to Venter’s sworn testimony in June, 1998, Celera had kept its data entirely secret. The public project had continued to release all data immediately. Hence, Celera was in the position to combine its data with the entire public data set, while the public project had to work with its data alone. In defense of this bizarre circumstance, Mark Adams, an assistant to Venter at Celera made the inspired comment, “We pay taxes, too.” This remark was a far cry from the confidence projected in the spring of 1998 that the public project was superfluous.

Science, a prestigious journal published by the American Association of Arts and Sciences, had caved in the face of Celera’s publicity campaign and widespread support for Celera within the scientific community. While *Science* would previously have refused to publish a paper of mine in which I limited access of other scientists to the underlying data, the journal agreed to publish the Celera paper despite the reality that no independent expert would have any real way to test its claims.

However, a full independent analysis proved unnecessary. In their paper, Celera scientists, after a defensive and technically problematical discussion of their efforts at whole-genome assembly, went on to base their entire biological analysis on an assembly achieved by superimposing the Celera data on the public sector’s clone-by-clone assembly. The documentation of how poorly the Celera assembly had worked was restricted to a few tables that only an expert could interpret and, of course, was little noted in the press hoopla that surrounded the *Science* and *Nature* publications. For the record, the number of “serious gaps” in the whole-genome assembly of the pooled Celera and public data was 118,968 (Ref. 12, Table 3, column 1—see “No. of scaffolds” in the “Whole-genome assembly” segment of the table). This result vindicated my prediction of “over 100,000 serious gaps” at the June, 1998, Congressional hearings.

We now come to the real question: What was the harm of this episode? I surmise from my anecdotal sampling of the views of my peers that even most scientists think that the competition between

Celera and the public sector was beneficial. The core argument is that “we got the sequence faster” because Celera prodded the public project to accelerate its pace. About this claim, there is no question. The Celera initiative undoubtedly accelerated the availability of an initial human-genome sequence by approximately two years. The public project committed itself, in order to compete effectively with Celera, to a rough-draft phase of the sequencing and mobilized more resources to achieve the job quickly than would have otherwise been available. The question of the value to science and society of this modest acceleration relative to the costs in distorted scientific priorities is too big an issue for this article. I will only say that I believe the enthusiasm of many scientists for accelerated availability of the human sequence was not based on an assessment of costs and benefits. Scientists do not think that way. Scientists are driven by short-term competitive pressures to get an edge on other scientists. Without doubt, this “red-queen” effect is the dominant drive behind the frenetic pace of modern science. Hence, enthusiasm for accelerated availability of the human sequence was concentrated among those scientists who considered themselves better positioned than their competitors to take advantage of it. Whether society would have been better or worse served by a more balanced allocation of resources—even at the expense of some delay in the availability of a rough draft of the human-genome sequence—is one of those “what-if” questions that we have no way to answer.

Certainly Celera, with a corporate identity based on speed, promulgated the idea that any delay in the availability of the human sequence would cause great human suffering. At some points, the irresponsibility with which this message was put forward was breathtaking. In the spring of 2000, I visited the Celera website and was startled to find myself staring into the eyes of two malnourished children, who were pleading by the expressions on their faces for my help. A banner appeared announcing that “Every minute, 10 children die from the effects of malnutrition.” The page was decorated with the mottos “Speed Matters” and “Discovery-Can’t-Wait.” Venter struck the same theme in prepared testimony to Congress on April 6, 2000: “Since the Congress began funding the human genome effort over 5 million Americans have died of cancer and over a million people have died because of adverse reactions to drugs.”¹⁴ The clear implication was that the dilly-dallying of myself and other participants in the Human Genome Project, was partly responsible.

I will limit my comments here to the bizarre reference to malnutrition. Science has known for well over 50 years—a time span that takes us back to before the discovery of the double helix—how to avoid malnutrition, safely, cheaply, and effectively. The reason for the horrific problem that children continue to die in large numbers for lack of adequate food is unrelated to genome science.

Indeed, it is even unrelated to nutritional science. To imply otherwise is morally wrong.

Returning to the question I posed earlier about the harm done by this episode, I suggest that it lies in three areas:

- (1) The politicization of scientific dialog.
- (2) The distortion of public expectations about the short-term practical payoff of basic research.
- (3) The undermining of support for a public commons of fundamental scientific knowledge.

In none of these areas, do we want to travel down the path that Celera blazed during the peak of the genome wars.

On the “politicization of scientific dialog,” I speak from personal experience. A full reading of the record of the June, 1998, Congressional hearings—and Venter’s further distortions of my views in the follow-up hearings on April 6, 2000—provides a lesson in why “attack ads” are so effective in political campaigns. The essential purpose of the attack ad is to do more damage to one’s opponent in 15 seconds than he can repair in a rejoinder of similar length. The matter of who has the stronger argument is irrelevant. In April, 2000, Venter referred back to my earlier testimony (Ref. 14, p. 18):

VENTER—One of the witnesses [in June of 1998] said, “show me the data!” He predicted we would fail—fail “catastrophically.” He was wrong—and I am happy to again show the Subcommittee and the world the data.

The “he” to whom Venter refers is clearly me. My “show me the data!” challenge had obviously hit a sensitive spot at the June, 1998, hearings. However, the claim that I predicted that Celera would fail “catastrophically” is an attack-ad-style misrepresentation of my testimony. My only use of the word “catastrophic” had been the one I quoted earlier in connection with my prediction that Celera’s human assembly would encounter “reasonably catastrophic problems” at the assembly stage. Indeed, I had explicitly addressed the question of potential failure of the Celera project in response to a question from Chairman Calvert (Ref. 9, p. 83):

CALVERT—I have just a quick question for Dr. Olson. Obviously you are a skeptic when it comes to the private sector initiative described here today. If this project is likely to fail, in your estimation, should we just ignore it and continue the federal program we have today unchanged?

OLSON—Well, I want to make clear that failure is a relative term. I have emphasized that I believe it will produce a huge amount of extremely useful data. I don’t believe that it will meet the quality standards that have

been outlined. And I think that the federal program would be well advised over the next 2 or 3 years to concentrate on defining the cost-benefit tradeoffs associated with the high-quality sequence product. No known approach is going to produce a perfect product. Indeed, perfect is not well-defined in the context of [an] intrinsically variable structure like the human genome, but I believe that...the unique niche for the federal program over the next few years is to refine the methods that are required to produce the best available product that can be achieved at reasonable cost, and I would define a reasonable cost as roughly current levels of funding.

Not only was Venter’s April, 2000, statement that I had predicted catastrophic failure for Celera’s human sequencing a misrepresentation of my testimony, but his claim that he had the data to prove me wrong was also specious. The April, 2000, hearings came as Celera published a paper on the fruit-fly genome, a genome that is a few percent the size of the human genome and has a much lower representation of repeated DNA. I had made no predictions about how well Celera’s methods would perform in this very different setting. Clearly, they would work much better. However, it is worth noting that, as of this writing (December, 2001), a coalition of publicly funded laboratories has been working in a clone-by-clone fashion for one and one-half years to straighten out Celera’s fruit-fly sequence and this effort remains well short of the goal. Until the clean-up job is complete, there will be no meaningful basis on which to assess how well Celera’s methods worked even on the fruit fly.

The above episode illustrates the dynamic of the attack ad. Venter said that I had predicted he would fail catastrophically and he now had the data to show that I was wrong. I had made no such prediction and he had no such data. However, his comments were highly quotable. A *Nature* reporter sent me an e-mail in which he provided the following summary of Venter’s testimony: “At yesterday’s science committee hearing Venter basically said your prediction about the result of his shotgun sequencing was wrong (Paul Smaglik, personal communication).” It was clear that any defense I might mount of my earlier testimony would not be nearly as quotable as Venter’s swashbuckling attack on it.

Healthy scientific dialog depends on several implicit rules. Scientists must confine their comments to subjects on which they are well informed. They must make a good-faith effort to communicate their ideas rather than simply to inflict damage on those who disagree with them. They must rely on objective evidence, when available, or, when such evidence is lacking, they must acknowledge that their position is speculative. Adherence to these rules should be the price of

admission. This code can only be enforced by scientists themselves. The broad failure of the biomedical research community to enforce it during the genome wars is cause for alarm.

While Celera broke new ground in politicizing scientific dialog, it had plenty of company when it came to distorting public expectations about the short-term practical payoff of basic research. Celera's effort to associate accelerated genome sequencing with the challenge of reducing childhood malnutrition is but one of many fantastic advertising claims made by biotech companies during the late 1990s. For example, Agilent, a biotechnology spinoff of Hewlett-Packard, ran advertisements in 1999 that featured an abstract representation of the double helix, drawn as a twisted ladder. The advertisement announced boldly that "At the top of this ladder is a world without disease."¹⁵ The key to rapid progress toward this utopia lay in the use of Agilent's ultrafast gene-analysis technology. The idea that genome analysis—or any other human activity—will lead to "a world without disease" is so ludicrous that it apparently works as an advertising hook. To be sure, proponents of the publicly supported Human Genome Project have not always been as temperate in their claims for the Project's practical benefits as I would have preferred. However, I know of no claims that the data will end childhood hunger or lead to a "world without disease." Indeed, it is a hallmark of public science that the processes by which resources are allocated and research is carried out benefit from the free play of open, pluralistic dialog. This process typically provides a powerful counterweight to the inevitable bouts of excessive exuberance by strong partisans of particular courses of action.

Finally, and perhaps most fundamentally, Celera did harm by undermining support for a public commons of fundamental scientific knowledge. The maintenance of a healthy public commons speaks to the very heart of the scientific enterprise. The two most central characteristics of science are its reliance on open, evidence-based dialog and its inherently progressive nature. Scientific dialog routinely leads to a robust consensus about what is reliably known, what remains uncertain, and what is unknowable. Because this process is rooted in the structure of the human brain, science has a unique ability to transcend cultural differences that are irremediably divisive in other areas of social activity.

The erosion of the public commons of science will compromise both science's philosophical and practical benefits for society. Science is progressive because it depends absolutely on incremental additions to previously acquired knowledge. If we create obstacles to the sharing of information—and construct toll gates at each step in the advancement of knowledge—the cost and sheer cumbersome of further progress will escalate uncontrollably. We can only be thankful that Perkin-Elmer's business plan took shape in the

1990's rather than the 1600's. If the first manufacturer of microscopes had operated in the same legal and business framework as Perkin-Elmer, scientists might now only be able to buy images of the microbial world rather than to look for themselves to see what is there.

Impassioned polemics will not lead to a reversal of the disturbing trends illustrated by the genome wars. We will need actual changes in public policy or sufficiently credible threats of such changes to alter the behavior of decision makers who fund and control private-sector science. Little can be done directly to tame the exuberant financial markets that provided most of Celera's capital. However, painful as the short-term effects of such exuberance may be, these markets are self-correcting. Indeed, Celera's stock has declined approximately 90% since the peak of the company's power in the winter of 2000 and new companies with similar game plans are unlikely to emerge in the present business climate. However, while bull markets come and go, a highly profitable pharmaceutical industry, which is the central driver of the trends I am discussing, remains.

Companies like Celera attract investors through the promise that they can either evolve into major pharmaceutical companies themselves or, more plausibly, tap into the profits of existing drug producers by licensing valuable technology, intellectual property, or proprietary data to them. Governments have two powerful tools with which to regulate this system in the public interest. First, governments control intellectual-property law. Intellectual-property law is under the malleable control of legislatures and administrative agencies. If the law is not serving the public interest, it can and should be changed. Present trends in the biotechnology industry suggest that the law has become overly protective of private claims to early-stage scientific knowledge.

Additional public leverage over the pharmaceutical industry's behavior arises because governments are a principal customer for their products. By controlling the prices they are willing to pay for drugs, governments can control the profitability of the industry. Presently, the pharmaceutical industry operates on profit margins that are far above those typically associated with high-volume manufacturing. This arrangement has been historically justified by the large research-and-development costs required to bring new drugs to market. Indeed, a strong case can be made that the high profitability of the pharmaceutical industry has been socially beneficial by enabling the development of an ever increasing variety of safe and effective drugs. The quid pro quo for this arrangement must be socially responsible behavior by the industry. There are many indications that pharmaceutical companies largely understand this tradeoff. For example, Merck funded the creation of a public database of the genomic sequences that code for human proteins simply to establish the

importance of keeping such early-stage knowledge in the public domain. Coalitions of pharmaceutical companies have similarly contributed to the creation of public-domain data on the mouse-genome sequence and on human genetic variation. However, other companies have declined to participate in these coalitions and, indeed, have provided major financial support for Celera's privatization schemes. Given the industry's dependence on publicly supported basic science, and its sensitivity to changes in public policy toward intellectual property and product pricing, there is reason to hope that the trend in the pharmaceutical industry will be toward increasingly responsible social behavior. If this trend fails to materialize—and drug companies attempt to use their high profitability to gain increasing control over basic scientific knowledge—the profitability of the industry can and should be reduced.

Although many of the forces that threaten basic scientific values originate outside the scientific community, there is also an enemy within. During the genome wars, the scientific community's support for the core values on which science's health depends has been disappointing. Perhaps science has assimilated the mores of the "new economy" a bit too readily. While the economic boom of the 1990's was not the first occasion on which science has felt the sudden, strong embrace of the larger society—World War II and much of the Cold War are other major examples—this new hug has been more encompassing than its predecessors, particularly for biomedical researchers. Knowledge has finally emerged as not just the most valuable, but actually the most valued, product of human civilization. The consequences for the scientific community, whose *raison d'être* is the creation of new knowledge, should be a sharpening of the distinct values that underlie the scientific enterprise, not the melding of these values with those of politics and commerce.

To achieve this sharpening of science's basic values, scientists will need to develop a greater willingness to subject the activities of other scientists to substantive public criticism. In one press interview, Craig Venter characterized criticism of his work as "one of the sadder parts of science."^{16,17} Then he went on to make the following odd comment:

There's two ways to get ahead in science, one is to do something that is significant, and the other way is to criticize someone who has done something significant. We've chosen the former; some of our critics have chosen the latter.

If Venter actually believes that criticizing those who "have done something significant" is a way to "get ahead" in science, he is displaying a tenuous grasp of contemporary scientific culture.

In reality, public criticism of other scientists—regardless of how substantive that criticism may be—has become a near taboo in the scientific community. Indeed, it is this taboo that has provided Celera with such a large space in which to maneuver. Scientists who wish to promote a healthy relationship between science and society should seek to recapture this space. To do so, they must learn to emulate our political tradition at its best, not its worst. In this tradition, the vigorous exercise of public criticism plays an essential role.

References

1. Bateson, W. (1916). Book review of *The Mechanism of Mendelian Heredity*. *Science*, **44**, 536–543.
2. Watson, J. D. & Crick, F. H. C. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.
3. Gamow, G. (1954). Possible relation between deoxyribonucleic acid and protein structures. *Nature*, **173**, 318.
4. National Research Council Committee on Mapping and Sequencing the Human Genome (1988). *Mapping and Sequencing the Human Genome*, National Academy Press, Washington, DC.
5. Olson, M. V. (1995). A time to sequence. *Science*, **270**, 394–396.
6. Wade, N. (May 10, 1998). Scientist's plan: map all DNA within 3 years. *The New York Times*, A1.
7. Wade, N. (May 12, 1998). Beyond sequencing of human DNA. *The New York Times*, C3.
8. Cook-Deegan, R. M. (1994). *The Gene Wars: Science, Politics, and the Human Genome*, W.W. Norton, New York.
9. Subcommittee on Energy and Environment of the Committee on Science, U.S. House of Representatives, 105th Congress, Second Session (1998). *The Human Genome Project: How Private Sector Developments Affect the Government Program*, June 17, 1998, U.S. Government Printing Office, Washington, DC.
10. Weber, J. L. & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409.
11. Green, P. (1997). Against a whole-genome shotgun. *Genome Res.* **7**, 410–417.
12. Venter, J. C. *et al.* (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.
13. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
14. Subcommittee on Energy and Environment of the Committee on Science, U.S. House of Representatives, 106th Congress, Second Session (2000). *The Human Genome Project*, April 6, 2000, U.S. Government Printing Office, Washington, DC.
15. Agilent Technologies (1999). *U.S. News & World Report*, September 6, 1999, pp. 2–3.
16. Broberg, B. (September, 2001). Scientists may be solving the mystery of the human genome, but the debate is getting hotter over profit motives and the rights to the human blueprint. In *Columns*, p. 31, University of Washington, Seattle, WA.
17. Manning, S. (2001). *From Genome Mapping to Drug Discovery*, May 28, Associated Press.

Maynard Olson is Professor of Medicine and Genome Sciences, as well as Director of the Genome Center at the University of Washington in Seattle, USA. He graduated from Caltech with a Bachelor's degree in chemistry and received his PhD in inorganic chemistry from Stanford University in 1970. After five years on the faculty at Dartmouth College, he changed his research emphasis to molecular genetics, working with Benjamin Hall in the Department of Genetics at the University of Washington. During that period, he participated in early applications of recombinant-DNA techniques to problems in yeast genetics. In 1979, he moved to the Department of Genetics at Washington University in St. Louis, where he became a Professor in 1986. At Washington University, he participated in the development of systematic approaches to the analysis of complex genomes before moving back to the University of Washington in 1992. Dr Olson has participated extensively in the formulation of policy for the Human Genome Project: in 1987, he served on the National Research Council Committee on Mapping and Sequencing the Human Genome, from 1989 to 1992, he was a member of the Program Advisory Committee on the Human Genome at the National Institutes of Health, and he presently serves on the National Human Genome Research Institute Council. Dr Olson received the Genetics Society of America Medal in 1992 and the City of Medicine Award in 2000; he was elected to the National Academy of Sciences of the USA in 1994.

(Received and Accepted 11 April 2002)