Origins of the Human Genome Project: Why Sequence the Human Genome When 96% of It Is Junk?

I was not much involved in the discussion and debate about initiating a program to determine the base-pair sequence of the human genome, until the idea surfaced publicly. As I recall the genesis of the Human Genome Project, the idea for sequencing the human genome was initiated independently and nearly simultaneously by Robert Sinsheimer, then Chancellor of the University of California–Santa Cruz (UCSC), and Charles DeLisi of the United States Department of Energy. Each had his own purpose in promoting such an audacious undertaking, but the goals of their ambitious plans are best left for them to tell. The proposal was initially aired at a meeting of a small group of scientists convened by Sinsheimer at UCSC in May 1985 and received the backing of those who attended. I became aware of the project through an editorial or op-ed–style piece by Renato Dulbecco in *Science,* March 1986. Dulbecco's enthusiasm for the project was based on his conviction that only by having the complete human genome sequence could we hope to identify the many oncogenes, tumor suppressors, and their modifiers. Although that particular goal seemed problematic, I was enthusiastic about the likelihood that the sequence would reveal important organizational, structural, and functional features of mammalian genes.

That conviction stemmed from having seen, firsthand, the tremendous advantages of knowing the sequence of SV40 (in 1978) and adenovirus genomic DNAs (in 1979–1980), particularly for deciphering their biological properties. In each of these instances, as well as for the longer and more complex genomic DNAs of the herpes virus and cytomegalovirus, knowing the sequences was critical for accurately mapping their mRNAs, identifying the introns, and making pretty good guesses about the transcriptional regulatory elements. Even more significant was the ability to engineer precisely targeted modifications to their genomes (e.g., base changes, deletions and additions, sequence rearrangements, and substitutions of defined segments with nonviral DNA). One could easily imagine that knowing the human DNA sequence would enable us to manipulate the sequences of specific genes for a variety of hitherto-undoable experiments.

Aware of the upcoming 1986 Cold Spring Harbor (CSH) Symposium on the "Molecular Biology of *Homo sapiens,*" I suggested to Jim Watson that it might be interesting to convene a small group of interested people to discuss the proposal's feasibility. I thought that such a rump session might attract people who would be engaged by the proposal, and Watson agreed to set aside some time during the first free afternoon. As the attendees assembled, it was clear that the project was on the minds of many, and almost everyone who attended the symposium showed up for the session at the newly dedicated Grace Auditorium. Wally Gilbert and I were assigned the task of guiding the discussion. Needless to say, what followed was highly contentious; the reactions ranged from outrage to moderate enthusiasm—the former outnumbering the latter by about five to one.

Gilbert began the discussion by outlining his favored approach: fragment the entire genome's DNA into a collection of overlapping fragments, clone the individual fragments, sequence the cloned segments with the then-existing sequencing technology, and assemble their original order with appropriate computer software. In his most self-assured manner, Gilbert estimated that such a project could be completed in ~10–20 years at a net cost of ~$1 per base, or ~$3 billion. Even before he finished, one could hear the rumblings of discontent and the audience's gathering outrage. It was not just his matter-of-fact manner and self-assurance about his projections that got the discussion off on the wrong foot, for there was also the rumor (which may well have been planted by Gilbert) that a company he was contemplating starting would undertake the project on its own, copyright the sequence, and market its content to interested parties.

One could sense the fury of many in the audience, and there was a rush to speak out in protest. Among the more vociferous comments, three points stood out:

1. The cost of doing this project would diminish federal funding for individual investigator–initiated science and thereby would shift the culture of basic biological research from "Little Science" to "Big Science." Some feared that biology would experience the same consequence that physics did when massive projects like the Stanford Linear Accelerator Center were undertaken in that field.
2. Many thought that Gilbert's approach was boring and thus would not attract well-experienced people, which, most likely, would make the product suspect. Moreover, the benefits of the sequence project might not materialize until the very late stages.
3. A surprisingly vocal group argued that, because <5% of the DNA sequence was informational (i.e., repre-

Address for correspondence and reprints: Dr. Paul Berg, Beckman Center B-062, Stanford University Medical Center, Stanford, CA 94305. E-mail: pberg@stanford.edu
Am. J. Hum. Genet. 2006;79:603–605. 0002-9297/2006/7904-0002$15.00

sented by genes encoding proteins and RNAs), there was no point in sequencing what was unaffectionately labeled "junk"—junk was defined as all the stuff between genes and within introns. Why, many asked, should we spend a lot of money and effort to sequence what was clearly irrelevant?

The fury of the reactions of some of our most respected molecular geneticists startled me. Several of the speakers argued that certain areas of research, usually their own specialty, were far more valuable than the sequence of the human genome. I was particularly irked by the claims that there was no need to sequence the entire 3 billion base pairs and that knowing the sequences of only the genes would suffice. Frankly, I was shocked by what seemed to me to be a display of what I termed an "arrogance of ignorance." Why, I asked, should we foreclose on the likelihood that noncoding regions within and surrounding genes contain signals that we have not yet recognized or learned to assay? Furthermore, wasn't it conceivable that there are DNA sequences for functions other than encoding proteins and RNAs? For example, the DNA sequence might serve for other organismal functions (e.g., chromosomal replication, packaging of the DNA into highly condensed chromatin, or control of development). It seemed surprising and disconcerting to hear that many were prepared to discard, a priori, a potential source of such information, and it was even more surprising that this myopic view persisted both throughout the meeting and for some time afterward.

During the session, I tried to steer the discussion away from the cost issue and the fuzzy arguments about Little Science versus Big Science. Perhaps it was better, I thought, to tempt the creative minds in the audience. After all, this was a scientific meeting with some of the most creative minds sitting in the audience. What if, I said, some philanthropic source descended into our midst and offered $3 billion to produce the sequence of the human genome at the end of 10 years? And, I suggested, assume that we were assured that there would be no impact on existing sources of funding. Would the project be worth doing? If so, how should we proceed with it? Gilbert had offered his approach, but, I asked, are there better ways?

To get that discussion started, I proposed that we might consider sequencing only cloned cDNAs from a variety of libraries made from different tissues and conditions. Knowledge of the expressed sequences would enable us to bootstrap our way to cloning the genomic versions of the cDNAs and, thereby, enable us to identify the introns and the likely promoters. Such an approach, I argued, would very likely yield valuable and interesting cloned material for many investigators to work on long before we knew the entire sequence. The premise was that the effort would identify the chromosomal versions of the expressed sequences and, with some cleverness, their flanking sequences.

However, try as I might, I could not engage the audience in that exercise. Their concerns were about the price that would be paid by traditional ways of doing science and that many more-interesting and important problems would be abandoned or neglected. The meeting ended with most people unconvinced of the value of proceeding with a project to sequence the human genome.

At the end of the meeting, I flew to Basel, Switzerland, where I was part of an advisory group to the Basel Institute of Immunology. At the hotel, I found a group of American and European colleagues perched on the veranda overlooking the fast-flowing Rhine River. They were clearly aware of the discussion at CSH and my participation in it. I again had to defend my support for the sequencing project against arguments that were a repetition of those expressed at CSH.

Soon thereafter, the National Academy of Sciences convened a blue-ribbon committee, many members of which had been among the critical voices at CSH. Their report recast the scope and direction of the project in a more constructive way; the principal change was the proposal to proceed in phases: determine the genetic map by use of principally polymorphic markers, create a physical map consisting of linked cloned cosmids, and focus on developing more cost- and time-efficient means of sequencing DNA. The most important recommendation, in my view, was to include in the project the sequencing of the then-favorite model organisms: *Escherichia coli, Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans,* and the mouse. It was clear that the new formulation did not threaten research support for those who worked on prokaryotes and lower eukaryotes. More likely, the additional funding would energize research on these organisms. It also provided a livelihood for those interested in mapping their favorite organism and for those committed to cloning and mapping large segments of DNA. In the end, people were mollified by the realization that they would not be left out of the project's funding. Also, the proposal had a logic for how to proceed and the acceptance that useful information would be generated long before the project was completed.

Sometime after the project was under way, Watson became the director of the project and set the agenda for how the project would proceed. He was committed to a razor-like focus on the development of genetic and physical maps, discouraging and even dismissing proposals that focused on making the work relevant to the biology. Indeed, that strategy was enforced by the study sections that reviewed genome-project grant proposals; proposals involving methods that would further the two mapping projects received preference, whereas those that hinted at deviation from that goal went unfunded. There is little question that Watson's forceful and committed leadership ensured the project's success.

It is interesting, in retrospect, that the course Gilbert had proposed for obtaining the human genome sequence—shotgun cloning, sequencing, and assembly of completed bits into the whole—was what carried the day.

Also, people who had dismissed the necessity of knowing the sequence of the junk now readily admit that the junk may very well be the crown jewels, the stuff that orchestrates the coding sequences in biologically meaningful activities. The past few years have revealed unexpected findings regarding noncoding genomic sequences, giving assurance that there is much more to discover in the genome sequences. Moreover, understanding the function of the noncoding genome sequences is very likely to accelerate, as the tools for mining the sequence and the application of robust and large-scale methods for detecting transcription become more refined.

PAUL BERG
Stanford University Medical Center
Stanford