Quantitative Biology 2021, 9(1): 4–7
https://doi.org/10.15302/J-QB-021-0243

FEATURE

The Human Genome Project: the Beginning of the Beginning

Michael S. Waterman[1,2,*]

[1] Biocomplexity Institute, University of Virginia, Chancellorsville, VA 22904, USA
[2] Department of Biological Sciences, Quantitative and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: msw@usc.edu

Received February 19, 2021

In May 1985 there was at University of California Santa Cruz an influential meeting that was the first serious discussion of sequencing the entire human genome. The author was one of the participants and described the meeting and related issues.

The Human Genome Project (HGP) is a historical and landmark scientific project. In spite of initial controversy it has become a bedrock foundation for much progress in biological science and human health. After the Human Genome Project was completed in the early 2000s, next generation sequencing technologies were developed and that has revolutionized genomics. Here is a brief account of the May 1985 meeting at University of California Santa Cruz. Historical accounts often begin with a the Department of Energy (DOE) meeting in Santa Fe in March 1986 and neglect including the Santa Cruz meeting [1], although sometimes it is discussed [2].

It was May 1985. I drove up narrow Highway 17 out of San Jose, over the mountains, down to the sea to turn right on the famous Route 1, and eventually turned right again to the campus of the University of California at Santa Cruz (UCSC). I drove across the open fields of the bench land and then into the redwood forested fingers of the university. Harry Noller, a RNA biologist, has his labs somewhere in the maze of redwoods, small canyons and occasional buildings. Harry told me that when he interviewed in 1968 he similarly became lost among the magnificent trees, and that is what decided him to join the newly formed university. Harry with his beard and abiding love of jazz is identical to Santa Cruz for me. UC Santa Cruz was begun in 1965 to promote progressive and interdisciplinary undergraduate education, and these 20 years later it is building a serious scientific reputation.

And in 1985 Robert Sinsheimer was chancellor of the university (1977–1987) and he is famous for his work in isolating, purifying, and replicating synthetically the DNA of the virus ϕX 174. Sinsheimer had the vision and courage to be the first seriously to propose sequencing the human genome and that is why I and others are visiting UCSC.

With Noller, Edgar, Moldave, and Ludwig from UCSC organizing, on May 24 and 25 1985 there was a meeting of a dozen experts that assembled at Santa Cruz (Fig.1). Those attending were Bart Barrell, David Botstein, George Church, Ronald Davis, Helen Donis-Keller, Walter Gilbert, Lee Hood, Hans Lerach, Leonard Lerman, David Schwartz, John Sulston, and Michael Waterman. My inclusion as the only computational and mathematical person was surely due to Noller, and the meeting was transformational for me as well as several others attending.

I had met Noller, Gilbert and Hood before, but no one else. One of my dominant impressions the first day was negative. David Botstein seemed to be constantly offering strong and outrageous opinions. I wondered how he had survived being so needlessly confrontational. By the morning of the second day I grudgingly realized that every time he opened his mouth he said something smart. In the end we became friends and I think highly of him. But if you do not want to know what he thinks about something, do not ask him. You probably have no choice

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

Figure 1. Meeting organizers: Sinsheimer, Edgar, Lugwig, Noller. Permission to use photograph: David Haussler, UC Santa Cruz.

if he's in the same room. At the conclusion of the meeting he asked me about some issues with genetic mapping, and as David was at Massachusetts Institute of Technology (MIT) and I was on the opposite coast, I strongly suggested he contact a person named Eric Lander. The rest of that story became, as they say, history.

Much of the meeting consisted of people wondering if the project was technically feasible. Schwartz who had not completed his PhD yet told us about his revolutionary pulsed-field electrophoresis techniques for separating and mapping large DNA molecules. Gilbert had a Nobel Prize for sequencing methods, and Lee Hood and especially George Church were planning new and improved methods. Ron Davis knew about making clones with large inserts and he talked about how to maintain accuracy. Church, Davis and Schwartz are among the true geniuses of biotechnology. I wondered if computational methods were able to store and process that much data, and concluded it might be just possible. I was naive about the repeats in the human genome, although the repetitive nature of the human genome had been established by Britten and Davidson beginning in 1969. And reading the non repeated DNA would be triumph enough. My head was spinning at a project of this size. It was not until 1995, ten years later, that the first complete genome of a free-living organism was sequenced, and that was a bacterium of only 1.8 million base pairs while the human genome is 3 billion base pairs. Even with excellent data assembling a genome was not going to be easy!

While drinking brandy after dinner, Wally Gilbert wondered how to find the labor to do the boring repetitive sequencing. Sinsheimer wanted there to be an institute in Santa Cruz to do the job. Gilbert proposed that we use prisoners to do the work. Give one group the Crick strand and another group the Watson strand, that'd create competition and quality control, he said. In graduate school at Michigan State University I had a part-time job at the Michigan State Highway Department where I worked with prisoners at Michigan's Jackson Prison. From that experience I knew that Gilbert had no idea what he was talking about. Still this was a big issue that everyone could see: how to manage science when it becomes more than single-investigator projects. Today we are in an era of "big science" where large multi-disciplinary projects in biology are common.

At the end of the meeting most everyone agreed that sequencing the human genome might be possible. Now what would it cost? To read one base cost around $15 then and the human genome was 3 billion letters long. And a redundancy of coverage of at least 5, and probably 8, was needed. This was getting into inconceivable numbers. But someone declared that we had to be optimistic about scientific and technological progress. Let's assume that the genome can be sequenced at $1 per base, that's only three-billion dollars. Only! This seemed outlandish and unwise. Then someone quoted what a battleship cost and I had a realization that the three-billion dollar price tag was cheap. With the US military budget in mind, a mere $3,000,000,000 was a small price to learn our genetic identity, what our grandparents has given us. Later much was made out of the promoters of the genome project saying "the human genome" as if it were unique. Of

course it escaped no one's attention that there is a diverse population of humans, but you have to start somewhere and the sequencing projects used DNA from multiple people.

After the main meeting the "big shots," who must have included Hood and Botstein, met with Sinsheimer and gave a negative evaluation of the project, at least the project to be done at Santa Cruz. Ignorant of this, I was then and remained steady in my belief that this was barely possible and truly important. If there were no medical benefits, and for sure there would be, just deciphering the code that our ancestors passed down to us as our genetic heritage was priceless. We had little knowledge of the complex details of human genetics and this scientific project would be one of the greatest endeavors and accomplishments in the history of science. Plus we would get a glimpse of how biology worked, actually a harder problem than understanding atomic physics.

Sinsheimer did not get his institute and for some time it looked as if the US National Institutes of Health (NIH) would pass on the controversial project. But my friend Charles Delisi had left Los Alamos to become director of biology at the Department of Energy and with David Smith he organized a meeting the next March in Santa Fe that was consequential. And Delisi managed to set up a DOE human genome project. NIH could not let such an important project in human biology slip out of their hands and the stage was set for the Human Genome Program. *California Magazine* later had an article about the Santa Cruz meeting. I treasure the fact that I was referred to as a computer specialist from University of Southern California. They didn't bother to use my name, I assume because such a person was obviously peripheral and completely unimportant.

By the mid 1980s and through the 1990s it was clear that many more analytical people were going to be required to study genomic DNA sequences and then the substantially more challenging problem of learning real biology from the sequence. I served on many NSF and NIH committees to evaluate proposals and set policy. Often I was working from a disadvantage of not having enough deep knowledge of biological science but just as often there was no one available who was better qualified. It was usually possible to get a sense of whether the computational analysis was feasible and made good sense, and then if the proposal had people associated who could handle it. I was also asked to give many lectures. To mathematics departments I emphasized statistics and computer science along with the biological motivations and realities. To statistics departments I emphasized computer science and biology, and to computer science departments I emphasized statistics and biology. To biology departments I tried to show why mathematical analysis is important. Once at the Pasteur Institute in Paris I gave a talk debunking a so-called discovery and showed a biology-free simulation of random sequences. "Just look at that alignment, it looks good enough to publish in *Nature*," I remarked. A friend in the audience told me later, "When you showed that simulation you terrified almost everyone in the room." My point was that just because the results of a computation look good they are not necessarily meaningful.

By June 1986 the idea of a human genome project was fiercely debated. Dedicating the funding required for such a project was a big change on how biological science was done and clearly some people felt their current funding was threatened. Plus it was said that only the genes were interesting and important; why spend that kind of money sequencing what was called junk DNA? Gilbert weighed in with "The total human sequence is the grail of human genetics." In 1987 the National Academy of Science published a report of the upcoming project, "seeing progress" [3]. I was asked to review the draft report whose authors included only one person involved in computation, the biochemist Russ Doolittle who had a shallow but extremely confident understanding of the issues. My review touched several points which were completely ignored and turned out to be critical. One matter, not initially appreciated, was the issue of whether genome sequences could be patented. I found the idea of human genes being commercial property offensive, and when I would attend sessions with lawyers talking about the issue, I would become upset. Finally that has been properly sorted out but it took a depressingly long while. See [4] for a recent discussion of these issues.

The HGP officially started in 1990 jointly by NIH and DOE and became a truly international project with many participating countries. James Watson headed the NIH HGP and David Galas the DOE HGP project. In 1993 they were replaced by Francis Collins (NIH) and Aristides Patrino (DOE). See [1] for a nice account of this process. The first draft of the genome sequence was announced in 2000. There was a parallel effort from Craig Venter and his private company Celera. Venter recruited Hamilton Smith who built high quality libraries of clones to sequence, and Eugene Myers who developed critical algorithms to assemble sequence from the whole genome (as opposed to sequencing one chromosome at a time as the NIH project did). Although my paper with Eric Lander [5] clearly implied that whole genome sequencing was as efficient as chromosome-by-chromosome sequencing, this point was not widely understood and the idea of whole genome sequencing was controversial. Today whole-genome sequencing is common. I spent some time at Celera and there was a magical exciting atmosphere in Myers' group. He brought together very

capable people who were inspired by the project.

The pressure from Celera caused a speedup of the public effort. Celera completed their genome sequence using some sequence from the public project; this complicated the issue of "who's on first" and created one of the arguments in many unpleasant and unfortunate exchanges between the groups that appeared in the press and elsewhere. At this point there was great concern within the public HGP that Celera would be first to publish the human genome sequence. Eric Lander called David Haussler at the University of California at Santa Cruz asking for help in finding the genes in yet to be assembled chromosomes, and Haussler called on Jim Kent who did not have his PhD yet heroically stepped up to write his genome assembly program when other efforts for assembly did not succeed. Then Haussler's student David Kulp and others could work to find the genes.

And the rough draft of the human genome sequence was announced in 2000 jointly between the public and private projects [6, 7]. The DNA was not from a single person and it was incomplete and full of errors. The cost was around three billion dollars. Repetitive sequence at centromeres and telomeres (centers and ends of chromosomes) made it very difficult to close gaps in sequence. It was not until 2020 that the first chromosome (the X chromosome) was completely end-to-end sequenced [8].

REFERENCES

1. Galas, D. J., Patrinos, A. and Delisi, C. (2017) Notes from a revolution: lessons from the Human Genome Project. Issues Sci. Technol., 57–62
2. Robert Cook-Deegan. (1994) The Gene Wars: Science, Politics, and the Human Genome. New York: W. W. Norton and Company, Inc.
3. National Research Council. (1988) Mapping and Sequencing the Human Genome. Washington, DC: The National Academies
4. Green, E. D., Watson, J. D. and Collins, F. S. (2015) Human Genome Project: Twenty-five years of big biology. Nature, 526, 29–31
5. Lander, E. S. and Waterman, M. S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2, 231–239
6. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860–921
7. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) The sequence of the human genome. Science, 291, 1304–1351
8. Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., *et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585, 79–84