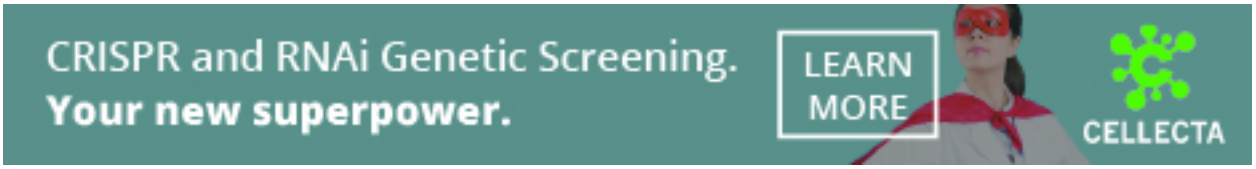
Contemplating the End of the Beginning

Francis S. Collins

Genome Res. 2001 11: 641-643
Access the most recent version at doi:10.1101/gr.1898



Commentary

Contemplating the End of the Beginning

Francis S. Collins

National Human Genome Research Institute, Bethesda, Maryland 20892-2152, USA

On February 12, 2001, an unprecedented collection of papers describing the initial sequencing and analysis of the human genome was published in *Nature* and *Science* (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). Although much celebration and press attention had surrounded the earlier announcement in June 2000 of the coverage of the vast majority of the human genome sequence in working draft form, the publications in February 2001 carried with them the kind of satisfying scientific significance that laborers in the genome fields had longed for—the full description of the methods used to determine the letters of over 90% of the human instruction book, and a host of surprising revelations from the analysis of its contents. This brief essay represents a personal reflection on how we got here, and where we are going.

The achievement of these landmarks, coming years ahead of the original schedule, was only possible because of the advances begun early in the preceding decade, reflecting the polyphonic set of interconnected goals that the planners of the Human Genome Project (HGP) wisely included as part of the original master plan. Science traditionally operates by the process of researchers standing on the shoulders of those who came before, and that has certainly been true for the HGP. Building detailed genetic and physical maps, developing better, cheaper, and faster technologies for handling DNA, and mapping and sequencing the more modest-sized genomes of model organisms were all critical stepping stones on the path to initiating the large-scale sequencing of the human genome.

Pilot efforts to sequence the human genome began in the mid-1990s. When the International Human Genome Sequencing Consortium met for the first time in Bermuda in 1996, there was a sense of excitement, but the magnitude of the task at hand was sobering–throughput was too low, costs were too high, technology was still immature. Despite that anxiety, the assembled scientific leaders from several countries at that meeting endorsed the importance of high quality sequence, and made one of the most crucial decisions of the genome era–immediate data release. Led by John Sulston and Bob Waterston, who had adopted this same policy for the sequence of *C. elegans*, the assembled sequencing center directors unanimously adopted a statement that all assembled contigs greater than 1 or 2 kb would be placed in public databases within 24 hours. The argument was simple: The sequence would only benefit the public fully if it could be understood, and that required making it immediately available so that all the creative minds of the planet could work on it. The establishment of this principle was one of the defining moments of the HGP.

Over the next three years the rate-limiting steps of large-scale sequencing began to yield to creative innovations. The genome centers implemented major improvements in library production, template preparation, and laboratory information management, so that less and less human intervention was required in the main production pipelines. The advent of capillary sequencing machines from Amersham and ABD provided a much-needed boost in efficiency. Much has also been made of the appearance of a commercial entity on the scene in May 1998 (Celera Genomics) as an additional nudge to the HGP. Whereas it is fair to say that the resulting sense of competition provided an additional incentive to the genome centers, it would be misguided to say that the HGP was previously operating in a relaxed fashion, or that the significant advances in throughput would otherwise not have happened. After all, most of those advances were born of previous accomplishments of the HGP itself.

From my perspective, a major turning point occurred in Houston in February 1999. The largest NIH-funded centers (at the Whitehead Institute, Washington University, and Baylor College of Medicine) had just undergone rigorous peer review of their proposals to scale up sequencing throughput and were about to receive a significant increase in funding. The Sanger Centre in Hinxton (UK) and the Joint Genome Institute (Walnut Creek, CA) of the Department of Energy were also scaling up production rapidly. An experiment carried out the preceding summer had documented the high degree of utility of a "working draft" human genome sequence; the half-dozen labs that compared draft and finished sequences found that the draft could answer most of the scientific queries they posed (although it was harder to work with), and suggested that the majority of the HGP's efforts might well be devoted to obtaining working-draft coverage of the genome as quickly as possible, as long as the commitment to finishing was not diminished. Accordingly, the sequencing plans for the NIH and DOE that were

E-MAIL: fc23a@nih.gov; FAX (301) 402-0837.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1898.

published in October 1998 included the new goal of obtaining such coverage by the end of 2001.

But in February of 1999 that goal seemed awfully far off. Less than 15% of the genome had been sequenced. In preparing for the Houston meeting, I estimated that the five genome centers mentioned above (by then already referred to as the "G5", and expected to do about 80–85% of the work) might be able to collectively produce some 30—36 million reads/yr after the ramp-up had occurred. For a working draft, 4×–5× coverage of the genome is desirable. At 500 bp/read and a 75% success rate, one could estimate that 36M reads would represent ~4.5× coverage of a 3-Gb genome. So, I came to the Houston meeting prepared to argue that we should set a dramatic new goal: To cover 90% of the genome in at least working-draft form sometime in the spring of 2000, a good 18 months ahead of the previous goal.

I was not at all confident about how this proposal would be received. There was an intense debate. All agreed that if such a goal could be met, it would be of profound benefit to the scientific community. All agreed with the importance of getting the human sequence into the public domain as quickly as possible. But this dramatic speeding up of the enterprise would place acute pressures on the genome centers, some of whom would have to scale up sequence production by an order of magnitude in a matter of a few months. Everything we had learned in the preceding years indicated that ramps of over two- or threefold in a year were almost never successful. Perhaps even more profoundly, this plan would mean moving away from a carefully-orchestrated regional program of selection of bacterial artificial clone (BAC) templates, and toward a centralized, highly coordinated program for clone selection, tracking, and distribution, with very little time for regional curation. This would place a major stress on the mapping centers; particularly on Bob Waterston, John McPherson, and their able team at Washington University, who were in the process of building a whole-genome BAC fingerprint map but had not expected to have to complete it under such tight timelines.

The debate continued through most of the day. At the end—to the credit of the adventurous spirit of the center directors (Branscomb, Gibbs, Lander, Sulston, and Waterston)—there was a consensus to make the working draft the highest priority and to aim for coverage by spring of 2000. Thus began a new and intensely demanding collaborative phase of the effort, where the G5, bonded together by this rather outrageous promise, began weekly conference calls (some of us will never again note that it's 11 AM on Friday without feeling a mental tug), established multiple working groups on specific tasks, and subjected themselves to monthly (and sometimes bruising) detailed reviews of production levels and quality assessment.

If the G5 served as the "Security Council" of the sequencing effort, the complete set of sequencing centers (often referred to as the G16) served as the "General Assembly". This was a wonderfully talented group, and the diversity and international nature of the sequencing effort was one of its most important attributes. The G16 lost no time in enthusiastically endorsing the accelerated plan. At the G16 meetings in May and September 1999, a detailed assignment of regional responsibilities was made, so that no part of the genome was left untouched and that duplication was minimized. Although some modest adjustments were, by necessity, made along the way, it is greatly to the credit of these centers (located in six countries around the world) that nearly all of the individual center goals were met. I consider it one of the great privileges of my career to have played the role of project manager for this remarkable group, along with my dedicated staff and my colleagues Ari Patrinos (DOE) and Michael Morgan (Wellcome Trust).

Of course there were a host of problems. The capillary sequencing machines, on which the ramp depended, were fresh off the instrument production lines and performed poorly during their first few months. Feeding the sequencing pipeline with BACs was a major challenge, especially as this very accelerated schedule gave little or no time for sequential walking steps. By September of 1999 the collective effort had fallen well behind the projections that had been set in May. But later that year the hard work of planning began to pay off, and by January 2000 the G16 were collectively sequencing 1000 bp a second, 24 h/d, 7 d/wk. After celebrating 1 Gb in public databases in November 1999, it was possible to celebrate 2 Gb in March 2000. By early June 2000 it was clear that the 90% coverage level was close at hand and a plan was made to announce this milestone jointly with colleagues at Celera Genomics, who had carried out their own intense sequencing production during that time.

The announcement at the White House on June 26, 2000, was a heady experience, but a cynical observer might also have described this as a bit of an "odometer moment". After all, the point of sequencing the genome was to understand its meaning, not just tick off the letters. Over the summer and fall of 2000, a remarkable group of sequence analysts, many from institutions who had not been part of the sequencing effort, coalesced around the shared goal of trying to understand the overall landscape, repeats, genes, and proteins of the human genome at first reading. These four dozen experts, guided by the able leadership of Eric Lander and expressing very little concern about individual credit, dug deeply into the work—once again held together by extended conference calls,

thousands of e-mails, and occasional intense face-to-face meetings decorated by high-end computing capabilities. By October 2000 the analysis was taking shape, and with it emerged a series of major surprises about the human genome. Among those were the surprisingly small number of genes, the more complex architecture of human proteins compared to their homologs in worms and flies, the profoundly important lessons that could be learned from the human repeat sequences, and the discovery of apparent horizontal transfer from bacterial species. These revelations were particularly gratifying; perhaps all of us had worried a bit, without saying it aloud, that when the full sequence of the human genome emerged it might be rather unsurprising, or (worse yet) even a bit dull. We need not have worried!

The drafting of the major sequencing manuscript proceeded apace, in close concert with a spate of other companion papers describing important related features of the genome. This time, the public announcements on February 12, 2001 were intensely satisfying. The assembly of the sequence proved to be extremely robust, and the determination of the International Consortium to stick to the map-based strategy was strongly validated. Furthermore, the depth of sequence analysis was reported to produce "chills down the spine" and "goosebumps" in many readers. This was no odometer turning over; this was a genome of elegant and sweeping proportion.

What, then, of the future? First, let us collectively agree to cease using the phrase "the post-genomic era", at least for a decade or two. This is the beginning of genomics, not the end. Critical understanding of gene expression, the connection between sequence variations and phenotype, large-scale protein—protein interactions, and a host of other global analyses of human biology can now get seriously underway. In fact, we should probably also agree to deep-six the use of the term "the post-sequencing era", at least for a while. After all, finishing the sequence of the human genome to the same standard already achieved for chromosomes 21 and 22 will take major efforts by the genome centers over the next two years. Among vertebrates, large- scale sequencing of the mouse, rat, zebrafish, and two pufferfish species are already underway or planned, and the sequence of many more genomes will be of intense biological interest. In fact, a major challenge for the next phase of the HGP will be to define the process of prioritizing such sequencing projects and to continue investments in technology to drive the costs downward.

So, to borrow from Winston Churchill, this is not the end of genomics. This is not even the beginning of the end. But it may be the end of the beginning. For me, as a physician, the true payoff from the HGP will be the ability to better diagnose, treat, and prevent disease, and most of those benefits to humanity still lie ahead. With these immense data sets of sequence and variation now in hand, we are now empowered to pursue those goals in ways undreamed of a few years ago. If research support continues at vigorous levels, it is hard to imagine that genomic science will not soon reveal the mysteries of hereditary factors in heart disease, cancer, diabetes, mental illness, and a host of other conditions; in fact, the public accessibility of the HGP data has already jump-started many of these endeavors. Working in effective partnerships with private industry, we can anticipate the application of these basic science efforts to the development of dramatic new therapies, although the pace of medical breakthroughs will always be slower than we wish. However, for this vision to come true, we must place as much emphasis on solving some of the ethical, legal, and social issues that accompany this rapid pace of genomic discovery as we do on the hard science—or we will face the likelihood that the public will be fearful of taking advantage of otherwise powerful and highly beneficial information (for further discussion, see Clayton 2001). We must also work tirelessly to ensure that these medical advances provide benefits to all, and not just the privileged few.

At the celebration accompanying the publications on February 12, 2001, my musical colleagues in "The Directors' Band" unveiled a few new songs about the genome. Most were rather tongue-in-cheek. But the chorus and final verse of the last song, sung to the tune of Woody Guthrie's most famous composition, sums up why we did all this and what some of our hopes are:

> "This draft is your draft, this draft is my draft,
> And it's a free draft, no charge to see draft.
> It's our instruction book, so come on have a look,
> This draft was made for you and me.
>
> We only do this once, it's our inheritance,
> Joined by this common thread - black, yellow, white, or red,
> It is our family bond, and now its day has dawned.
> This draft was made for you and me."

REFERENCES

Clayton, E. 2001. *Genome Research* **11:** 659–664.

International Human Genome Sequencing Consortium. 2001. *Nature* **409:** 860–921.

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. *Science* **291:** 1304–1351.