The Human Genome Revealed

James D. Watson

Genome Res. 2001 11: 1803-1804
Access the most recent version at doi:10.1101/gr.211601

Commentary

The Human Genome Revealed

James D. Watson

President, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Seeing the International Sequencing Consortium's draft of the human genome is highly satisfying. The way in which its 3 billion bases have been determined closely followed the course outlined more than a decade ago by the National Academy of Sciences (NAS) Committee on "Mapping and Sequencing the Human Genome." Bruce Alberts, now the President of the NAS, was its chairman and I one of its 14 other members. The predictions in our 1988 report, that the human genome could be sequenced over a 15-year period for a cost of three billion dollars, were more accurate than we dared guess. Two more years of work, to fill in gaps and correct mistakes, will result in an almost errorless genetic script for human existence.

That the human script would become available within our lifetimes never passed through my mind or that of Francis Crick when we found the double helix in 1953. At that time, just learning how cells read the genetic instructions within DNA seemed a tall order. Happily, progress was faster than expected, and by 1966 we knew how the genetic code utilizes groups of three DNA bases to specify the amino acid constituents of proteins—the main "actors" in the plays of life. Things speeded up even more after the recombinant DNA procedures of Stanley Cohen and Herb Boyer burst upon the scene in 1973. Gene cloning and manipulation metamorphosed from being dreams to becoming facts of life. Simultaneously, Fred Sanger and Walter Gilbert each developed a powerful way to determine the order of bases along DNA molecules. This meant that humans, like cells, could read the messages of genes. The way was open to ascertain the complete genetic instructions, i.e. to sequence the genome, of any organism (subject to the usual constraints of money, personnel, and technology).

The first genomes tackled were those of viruses, with the first sequenced viral genomes containing only several thousand bases. By the early 1980s, viral genomes containing more than 100,000 bases had been completed, and bacterial genomes containing more than a million bases became realistic objectives. Completion of such genomes would at last tell us the number of different proteins necessary for bacterial existence. Back then I thought that the human genome, at several billion bases long, was much, much too large to take on. Soon, however, I became a strong proponent of an internationally-based Human Genome Project (HGP), believing that the large-scale mapping and sequencing resources that it would command would greatly hasten our discovery of the genetic underpinnings of many important human diseases.

Our NAS committee wasted little time on whether we needed a HGP; instead we focused on how it should be organized and financed. It seemed best to begin modestly and end with a sequencing crescendo, hopefully fueled by much lower sequencing costs. We agreed unanimously that the first big sequencing efforts should not focus on human DNA but on DNA from a model organism of genetics, such as baker's yeast and the fruit fly, *Drosophila*. We knew that many human genes were likely to be homologous to those of model organisms, and these provided good systems for studying gene function.

That we proposed a 15-year effort reflected our belief that those starting the project should also be part of the finishing team. Richard Gibbs, Eric Lander, Maynard Olson, John Suiston, Bob Waterston, and Jean Weissenbach all have stayed the course, running increasingly larger megabase sequencing labs. Only one of our original NAS committee is no longer in science. Sadly, Dan Nathans died of leukemia three years ago, at the age of 70. During our committee deliberations, no one proposed a shorter time frame—technology had to improve too much. Later, I learned that Congress likes big projects to be finished within 10 years so that key initial backers are still in Washington when the achievement is celebrated. Luckily, Tom Harkin recently became that Congress rarity: a three-term Democratic senator from Iowa. So, like New Mexico's Republican Pete Domenici, he will see the HGP from its beginnings to its finish as a senator.

The improvements in technology that the HGP would need for its success materialized almost on schedule. They largely involved modifications in pre-existing methods, as opposed to great leaps forward that generate Nobel Prize-like rewards. The current DNA sequencing machines, the workhorses of our big sequencing labs, are 1000-fold-improved descendents of the original sequencing machine put together by Mike Hunkepillar and Lloyd Smith in Lee Hood's Caltech lab. The computers and software that now compare new raw DNA sequences to preexisting ones also do their tasks 1000 times faster than was possible when the HGP began.

A major obstacle to the correct assembly of the human genome was the vast amount of the repetitive DNA (~50%). So the HGP labs decided early on to sequence DNA coming from known chromosomal locations. Their map-based strategy, however, was suddenly challenged in May 1998 by the new private company Celera Genomics, led by Craig Venter. Celera proposed an alternative strategy whereby the genome was randomly shredded into pieces that were sequenced and then reassembled in a single process without the construction of a map—a strategy known as "whole-genome shotgun sequencing." The key to their approach was to be the 200 new, high-capacity capillary DNA sequencers that were about to be launched in the market, as well as new proprietary shotgun assembly software for use on high-powered computers. So armed, Cetera promised a first draft of the human genome in only two years.

I first heard of Celera in a telephone call from my former associate, Richards Roberts, who organized the first (1988) Cold Spring Harbor meeting on Genome Mapping and Sequencing. Rich told me that Celera would blow the international consortium out of the water and asked me to consider joining him on its scientific advisory board. Expecting to learn more about Celera's game plan at our soon-to-be-held spring 1988 Genome Meeting, I quickly phoned the National Institutes of Health (NIH) Genome Office and the Wellcome Trust to report that Celera had marked them out for obsolescence. Later that week, Craig Venter visited the NIH to tell Harold Varmus and Francis Collins that the HGP's future effort might best be devoted to sequencing the mouse.

From the moment of Rich Roberts's call, I found it unthinkable that a private company should effectively control much of the human genome through key patents. This was a gene power-play that, at all costs, must be contained. To my relief, the Wellcome

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.211601.

Trust's immediate response was to double the budget for human genome sequencing at the Sanger Centre. Although the merits of each approach were yet to be tested, Celera's "super shotgun" method quickly caught the fancy of the serious press, who reported that the HGP was off-course. In fact, two years earlier at its spring 1996 Bermuda meeting, HGP leaders had seriously discussed Jim Weber's proposal for a low-resolution, whole-genome shotgun effort to complement the high-resolution map-based thrust. There, Phil Green's off-the-cuff calculations, later redone and published (Green 1997), indicated that human DNA is too repetitive for a pure shotgun approach to assemble the genome correctly.

In September 1998, I returned to Washington to tell key congressional leaders that expanded federal support of the publicly-funded sequencing effort was necessary to prevent a monopoly on human genetic information. Much of "big pharma" rooted for the public HGP, believing that Celera's future databases could only be validated through checking with publicly obtained sequences. To my relief, Congress increased public sequencing monies significantly. Thus encouraged, the HGP announced that it, like Celera, would complete a rough draft of the human genome in the spring of 2000. But unlike Celera, it would further pursue a highly accurate final product.

The February 2001 publications of the human genome by the HGP (International Human Genome Sequencing Consortium 2001) and Celera (Venter et al. 2001) represent a milestone in human history, revealing the basic features of the human genetic script. They will allow us to identify most of the genes that underlie human existence. Using the genetic code to translate their message into protein products, we now have the first comprehensive overview of the molecules that make up our bodies. And it is immediately obvious that these are very similar to the molecular building blocks of other forms of life. Darwinian evolution can be increasingly described through incremental changes in underlying DNA scripts.

It is, however, unclear whether either draft is accurate enough for confident protein structure predictions. In fact, proteome predictions from the two human drafts may be seriously misleading; only a virtually errorless "gold standard" human DNA script will move us confidently into proteome waters. That so much more sequencing needs to be done, however, should in no way lessen our admiration for what both groups have accomplished.

Until we saw the first DNA scripts underlying multicellular existence, it seemed natural that increasing organismal complexity would involve corresponding increases in gene numbers. So, I and virtually all of my scientific peers were surprised last year when the number of genes of the fruit fly, *Drosophila melanogaster*, was found to be much lower than that of a less complex animal, the roundworm *Caenorhabditis elegans* (13,500 vs. 18,500). More shocking still was the recent finding that the small mustard plant, *Arabadopsis thaliana*, contains many thousand more genes (~28,000) than does *C. elegans*. Now we are jolted again by the conclusion that the number of human genes may not be much more than 30,000. Until a year ago, I anticipated that human existence would require 70,000–100,000 genes.

Why organismal complexity fails to correlate with gene numbers is not fully clear. It may be partly due to RNA splicing events, which generate multiple protein products from single genes: Vertebrate genes give rise to more splicing products than do invertebrate genes. But equally relevant may be the quality of respective nervous systems. The roundworm, being dumber than the fruit fly, may need more specific proteins (and therefore genes) to respond to enemies or changes in its environment; the fruit fly's more advanced nervous system lets it respond to potential enemies and stresses by flying away. Plants, being totally dumb, must continually evolve new genes to respond to new enemies and climatic changes.

Many more vertebrate genomes need to be sequenced before we have a sense of how often the generation of new genes has underlain evolutionary change. We also need to know why vertebrate genomes contain so many more repetitive sequences than do invertebrate genomes. Most human repetitive sequences appear to have risen as the result of the generation and movement of transposable genetic elements. Conceivably, many of the mutations that underlie vertebrate evolution arise from transposon movement into regulatory regions, thus changing gene expression patterns. The very high levels of repetitive DNA in amphibians and lungfish may reflect their past needs to evolve fast for survival in their ever-changing ecological niches.

It should be possible to test the idea that changes in regulatory segments, as opposed to changes in amino acid coding segments, have dominated vertebrate evolution. For example, sequencing information from morphologically different breeds of dog may be informative, and hopefully funds will be made available to produce draft genomes of several breeds. How soon we shall be able to meaningfully compare the chimpanzee genome with that of our own, remains unclear. Obviously we would like to know the genetic changes that make possible the larger and more powerful human brain.

Of the many new facts emerging from the human genome draft, I am most excited by the finding that repetitive sequences are almost absent from the four clusters of homeobox genes. Unlike most functionally-related human genes, the chromosomal order of homeobox genes reflects their temporal expression patterns during embryonic development. In this respect, they resemble the genes of bacterial operons that are transcribed from single messenger RNA molecules: Genes located at the start of bacterial operons are transcribed first by RNA polymerase molecules moving along their respective region of DNA. Conceivably, much of early developmental timing in humans may be a reflection of the time needed for RNA polymerase molecules to transcribe the lengthy introns of homeobox genes. If so, insertions of sizable transposable sequences into them would lethally mis-set key timing events in embryonic development.

Many, many more unanticipated observations and hypotheses will emerge as the reading of the human script extends beyond those individuals who produced it, to the much larger world of interested biologists. Even the heartiest, however, will find themselves stretched if they take on too much. The most triumphs of the near future will likely come from focusing on human homologs of genes functionally understood in one or more model organisms. Eventually, even more important dividends will come from focusing on ourselves as human beings and making sense of the often-seemingly intractable relationsbetween nature and nurture. There is much more to human life than interactions between its DNA script and the RNA and protein "actors" that carry out its instructions. The culturally-derived facts and traditions that our brains pass onward from one generation to the next equally affect our lives.

Our genomes, thus, can never accurately predict our futures. But we would be more than silly if we did not use their information to the fullest. The human genetic script that we are now finalizing will be regarded as the most important book ever to be read.

NOTE

Published in modified form from *A Passion for DNA: Genes, Genomes, and Society* by James D. Watson, (2001). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

REFERENCES

Green, P. 1997. *Genome Res.* **7:** 410–417.

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.