

The Human Genome Project

"Science is essentially a cultural activity. It generates pure knowledge about ourselves and about the universe we live in, knowledge that continually reshapes our thinking" [1] [John Sulston]

In 1953 James Watson and Francis Crick discovered the structure of DNA - the code of instructions for all life on earth...

...in 2003 - just 50 years later - humankind had developed and exploited the technology, the computing capability and the financial and social impetus to record one whole human DNA sequence: some 3 billion letters of genetic code.



[Genome Research Limited]

Introduction Sharing data The draft The gold standard Nearly there? References

The Human Genome Project: a new reality

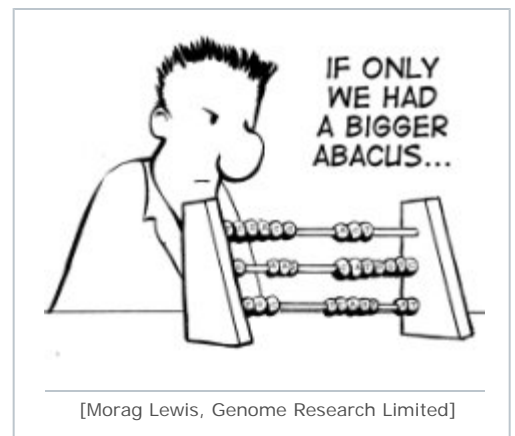
In June 1985, as dusk encroached on the second millennium, meetings aimed at outlining the practical task of sequencing the human genome began at the University of California, Santa Cruz. The scientific and technological conditions of the 1980s had become a catalyst for these discussions. DNA cloning and Fred Sanger's sequencing methods, developed in the mid- to late 1970s, were being exploited by scientists who felt that sequencing the human genome seemed possible at an experimental level. Crucially, researchers were, at the same time, beginning to apply computing solutions to genetics and DNA sequencing, developing methods that would make feasible the task of generating and handling genetic data globally.

This grand, new concept - a "Human Genome Project" - had strong supporters, who argued that deciphering the human genome would lead to new understanding and benefits for human health as well as and determined detractors, who feared such a project would provide a product that would bear little explanatory power for humans - perhaps merely a meaningless string of letters. Even before the Human Genome Project began in earnest, some commentators feared that this project had "engendered a controversy... that involves personalities and politics." [2]

The personalities, the politics and the controversy were only just emerging.

The Human Genome Project launched in 1990, through funding from the US National Institutes of Health (NIH) and Department of Energy, whose labs joined with international collaborators and resolved to sequence 95% of the DNA in human cells in just 15 years. Meanwhile in the UK, John Sulston and his colleagues at the MRC's Laboratory of Molecular Biology in Cambridge, had, for several years, been working at mapping the genome of the nematode worm and had resolved that sequencing the entire genome of the worm was finally feasible.

As the Human Genome Project was progressing in the US, in the UK the MRC approached the Wellcome Trust suggesting they form a new partnership to fund John's proposed worm sequencing, as a pilot for the Human Genome Project. From here things soon snowballed: the Wellcome Trust suggested that a much larger sequencing effort, to bolster the Human Genome Project should be embarked upon in the UK and appointed one of their senior administrators, Michael Morgan, to look into the viability of such a sequencing initiative. Eventually, in 1992, John Sulston submitted a grant application for an enormous £40-50 million to fund a new centre - the Sanger Centre - which was to form the British arm of the Human Genome Project's sequencing efforts.



In 1993 - with funding from the Wellcome Trust and MRC - the Sanger Centre was officially opened. One scientist recalls being struck by the scale of the task that lay ahead, on arriving at the Institute in 1993 Simon Gregory reflects: "it was just a huge lab, a huge empty lab, with boxes and boxes of equipment. It was all very exciting."

By the end of that year 87 scientists were working at the Sanger Centre, under the leadership of John Sulston, beginning to map and sequence the human genome.

The global sequencing effort

To sequence the human genome as accurately as possible, researchers developed the 'hierarchical shotgun' method. Researchers agreed that this was the best way to achieve the Human Genome Project's target of 95% coverage of the human genome by 2005.



The first challenge was to create a map of the human genome - a set of index marks on the genome code, used to position the sequences of letters of code that would come later.

Researchers essentially broke many copies of the genome into fragments, each around 150,000 letters of code (or base-pairs) long. They inserted the fragments into a bacterial artificial chromosome that could be grown in *E. coli* bacteria which divided, thereby replicating the DNA samples to create a stable resource - a 'library' of DNA clones. Where the cloned fragments came from or which overlapped was not known at this point.

Using special enzymes, researchers could cut the individual clones into diagnostic 'fingerprint' of fragments defined by each clone's sequence. They could then search among millions of fingerprints for shared

fragments that would reveal overlaps among the clones. Researchers then assembled the clones into longer contiguous regions and mapped these onto the human chromosomes. The result: a physical human genome map that would be crucial for the sequencing efforts.

To generate sequence of the individual bases that make up the genome, scientists needed to break the cloned fragments into smaller, more manageable, chunks, each around 1000 to 2000 base-pairs long. Researchers sequenced these fragments of human DNA using the shotgun method developed by Fred Sanger and his colleagues a dozen years before. Much as in mapping, researchers used overlaps, this time in the letters of genetic code itself, to reassemble the short stretches of determined sequence. Assembling the sequence from many short segments of sequence was a hugely intense compute task that depended on emerging technology and software to succeed.

Gradually labs around the world began producing DNA sequence. By 1994, the Sanger Institute had produced its first 100,000 bases of human DNA sequence. Remarkably, researchers at the Institute had already produced ten times that amount from the nematode worm genome. The worm project was a trailblazer - its methods, practices, collaborations and ethos would be integral to the development the social mores that would later lead to the successful completion of the Human Genome Project.

As the human sequence data was pouring out from centres across the globe, researchers were afforded glimpses of the kind of power that the human genome sequence might have for medical advance. In 1995, researchers from the Sanger Centre, with international collaborators, located the *BRCA2* gene, associated with increased risk of breast cancer. Elsewhere, as early as 1993, a US team had located the *MSH2* gene, which increases the risk of colon cancer for carriers. In Canada, researchers found five variants on the *FAD* gene, which together confer an almost 100 per cent risk of developing Alzheimer's disease.

Sharing the data

In 1996, representatives from sequencing centres around the world met in Bermuda to establish a set of principles for the release of data generated by the project. The meeting was initiated by Michael Morgan, the Wellcome Trust administrator who had been integral in the establishment of the Sanger Institute as the UK sequencing effort. The delegates agreed a set of Bermuda principles [3], which outlined that useful data should be made available prior to publication as part of a greater aim to distribute data and speed progress to maximize the public benefits of the Human Genome Project. In a move that stood biology on its head, the scientists agreed to share their results before they had had a chance to publish their findings in a scientific journal.

Sharing in the Wellcome Trust's determination that the principles of publically and freely available data would be crucial as the Human Genome Project continued, Michael Morgan worked beyond that



initial meeting in Bermuda to ensure that the commitments made were adhered to globally.

John Sulston was one of the champions of this ethos: he and Bob Waterston first began to develop a system of data sharing when they worked together on the nematode worm sequencing project. Described by Mike Dexter - the then Director of the Wellcome Trust - as "a torch bearer for the social conscience in science" [4], John maintained a stern resolve that researchers had a moral obligation to make their results freely available. The Sanger Institute and other sequencing centres agreed: every evening, sequencing centres either side of the Atlantic would make the results produced that day freely available on the internet.

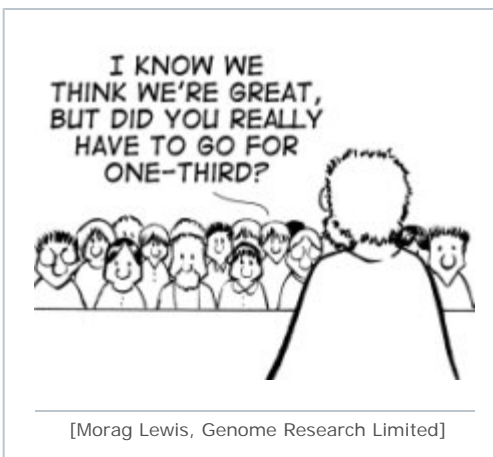
Data sharing, arguably not an instinctive practice for scientists, has - since the Human Genome Project - remained a resolute part of research practice at the Sanger Institute and other centres worldwide. The adoption of free release and data sharing has been among the major achievements of the Human Genome Project and will arguably prove as influential as the sequencing outputs themselves.

Around the time of the draft human genome publication in 2001, Mike Dexter, the then Director of the Wellcome Trust, reflected that "after our large-scale investment in the Human Genome Project, it is incumbent upon us to ensure that the public resources are fully utilised, in order to deliver the health benefits that will undoubtedly flow from the use of this information." [5]

Throughout the Human Genome Project, the Wellcome Trust had maintained these principles with an open-access publication policy to match its data release policy in ensuring that the research it funds can be accessed, read and built upon, thereby fostering a richer research culture.

A private effort launches

1998, a new, private venture was launched to sequence the human genome. The enterprise - named Celera Genomics - aimed to create its own database of human genomic data, which users would be able to subscribe to for a fee. They aimed to patent 300 clinically important genes, and it was reported that they held, at one point, 6500 applications on human genes [6,7]. The public consortium became still more resolute that the Human Genome Project goals would be met and that the shared human heritage - our genome sequence - would be freely and completely available, that access would not be fettered by subscription. The Sanger Centre received a funding boost from the Wellcome Trust and upped its projected contribution to one-third of the sequence.



History may well reflect that the two projects were engaged in a race toward completion. Those involved would likely reject such accounts. Jane Rogers, who project managed the Sanger Centre's Human Genome Project contribution, reflected afterwards that "it has not been a race, but a battle to ensure that the tools to speed biomedical research were available to all." [8] It was a quest to ensure the widest benefit worldwide for today and for the future: a successful quest that some have misinterpreted as the race to get to a finish line for glory. For the International Human Genome Sequencing Consortium, it was never a race to a finish line.

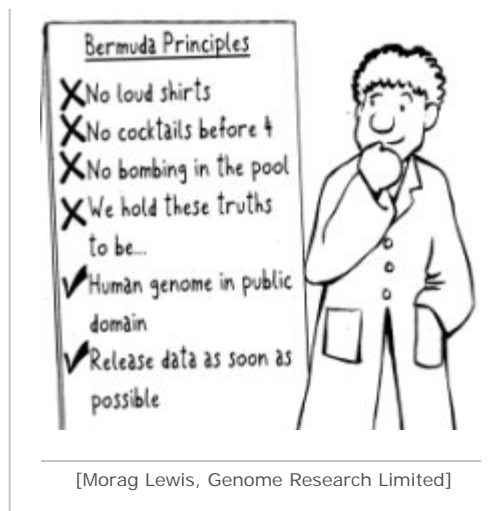
John Sulston warned that "If global capitalism gets complete control of the human genome, that is very bad news indeed." [9] Craig Venter, the head of Celera Genomics, attributed the unease among the participants of the Human Genome Project to scientific rivalry. The politics and personalities were in danger of detracting from the task at hand.

Talks aimed at consolidating the two projects, to bring more rapid public health benefits, were initiated in 1999, but broke down over concerns that Celera sought a ownership of the our shared code for human life. Martin Bobrow, a representative for the Human Genome Project and a then Governor of the Wellcome Trust announced: "Unfortunately, Celera's requirements seemed to amount to them establishing an effective monopoly over the human genome."

Each effort employed hundreds of sequencing machines - Celera reportedly had 300 of the latest machines [10] - to produce the raw results that would generate a human genome. Each effort worked day and night to complete a draft genome.

The draft human genome of 2000

On 26 June 2000 came two announcements - the public and private enterprises had completed their respective draft genome sequences. The UK's Prime Minister, Tony Blair in London, was linked live to President Bill Clinton as he described how "the effort to decipher the human genome...will be the scientific breakthrough of the century - perhaps of all time," maintaining that "we have a profound responsibility to ensure that the life-saving benefits of any cutting-



edge research are available to all human beings."

The researchers involved in the Human Genome Project had sequenced the order of bases in each chromosomal region four or five times. However, what remained was to improve the depth of coverage. The sequence had achieved broad whole genome coverage: but the sequence was by no means complete. As Sulston and others said, the "current human genome information marks just the "end of the beginning", not a completion of the task." [12]

In 2001, each effort published an account of its draft human genome sequence: Celera's effort appeared in *Science*, and the International Human Genome Sequencing Consortium's effort was published in *Nature*.

Celera's methodology was criticised: the company suggested that they had generated their sequence using Fred Sanger's "whole genome shotgun" method, skipping the mapping phase of the process. Critics argued that such an assembly would be impossible at the time without the Human Genome Project's data.

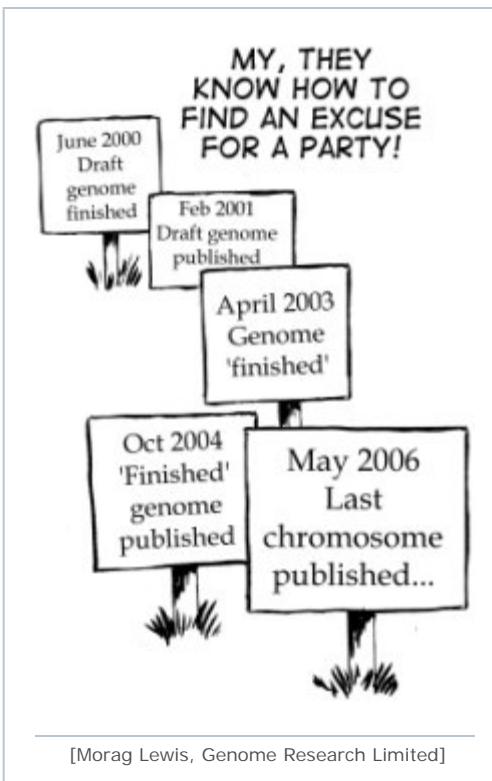
A year later, Venter stepped down as President of Celera, as the company cut its genomic research, "making room for additional senior level management experienced in pharmaceutical discovery and development" [13]. The genome sequence was freely and publically available.

Despite the fact that there was still much hard work to be done, John Sulston was convinced of the transforming power that the sequence would have. "Over the decades and centuries to come this sequence will inform all of medicine, all of biology, and will lead us to a total understanding of not only human beings but all of life," he said. "Life is a unity, and by understanding one part you understand another."

After the draft genome was announced, John Sulston stepped down as Director of the Sanger Institute, handing over to his successor, Allan Bradley. Much remained to be done to take the Human Genome Project to completion and John continued to work on the Sanger Centre's contribution until the announcement of the gold-standard sequence, as well as working on the last few sequences of his beloved worm. Allan Bradley took over with enthusiasm, looking to the future and developing new ways in which the Institute could contribute to genetic science in what he described as a new "postgenomic era".



Completion? The gold standard sequence of 2003



In 2003, two years ahead of schedule, with contributions from countless scientists from 20 institutions across the globe, the International Human Genome Sequencing Consortium announced that they had completed the gold-standard reference human genome, according to the guidelines of the original Human Genome Project, with 99.99% accuracy.

The sequence provided an invaluable resource for researchers. Launched in 2000, the Ensembl genome browser, co-run by the Sanger Institute and its neighbour, the European Bioinformatics Institute, was by this time receiving approximately 600,000 hits a week from researchers in over 120 countries looking to make use of the freely available genetic resources it provided.

This database system was secured and strengthened by the Human Genome Project's policies and continues to be an invaluable tool that not only allows open access to the fruits of the Human Genome Project and other sequencing projects, but also facilitates continual updating, annotation and comparison of sequences from numerous genomes, in a way that would be simply not have been possible in a privatised structure.

A year later, the publication of the gold standard sequence appeared in *Nature*, where the authors commented that the Human Genome Project "provides an essential foundation for the sequencing and analysis of additional large genomes." [14]

Scientists embarking on new sequencing projects would soon turn to the sequence as a reference for their work. It was a working example of scientific progress, its

very existence as a reference, allowed scientists to develop new, faster, cheaper and more effective

sequencing methods for their projects.

The finished sequence revealed a code over 3 billion letters long. Astonishingly the final sequence was found to contain only 25,000 genes - a quarter of the number originally suspected. The worldwide effort quickly turned to establishing what these genes and their code - the "book of life" - would mean for human health and disease.

Determined that the medical fruits of that "book of life" would be shared equally, irrespective of wealth or nationality, the Wellcome Trust extended its funding for the Sanger Institute and other biomedical projects around the world to work on developing genetic knowledge of some of the most devastating and human diseases: from cancer to malaria; from heart disease to typhoid.

Are we nearly there yet?

The "Human Genome Project" was officially complete in 2003 - researchers had sequenced the genome to 99.99% accuracy, two years ahead of schedule and under budget. But the "human genome project" without capitals - the project to understand the genetic nature of ourselves - is still a long way from completion. Some parts of the human genome have, to this day, refused to reveal themselves to researchers and the unknowns abound.



[Morag Lewis, Genome Research Limited]



[Morag Lewis, Genome Research Limited]

The controversy and fierce debate that suffused the Human Genome Project brought science and culture together. The product - a string of As, Ts, Cs and Gs - is far from a meaningless jumble of letters: it is the code that shapes our physical existence. That something so deeply personal is also what defines each and every human being on the globe has challenged some of mankind's most deep-rooted definitions of ownership and sharing.

The year 2003 marked 50 years of DNA and the completion of the Human Genome Project...

...but 2003 also marked the beginning of a new voyage - a quest to extract the shared, communal benefits that lay hidden among the billions of unique genome sequences carried in the cells of individuals across the globe.

References

1. Sulston J. (2001) Foreword: why we do science. In: *Frontiers 01: Science and Technology* ed. Tim Radford. 2001-02. p.11.
2. Smith L and Hood L. (1987) Mapping and sequencing the human genome: how to proceed. *Nature Biotechnology* 5: 933 - 939. doi: [10.1038/nbt0987-933](https://doi.org/10.1038/nbt0987-933)
3. [International Large-Scale Sequencing Meeting](#)
4. [Top international award for British genome champion](#)
5. [Wellcome Trust promotes open door policy of human genome treasure chest](#)
6. [6,000 human gene patents sought](#)
7. [Patent May Yield Myriad Tools](#)
8. Rogers J. (2003) Genome sequencing: Wellcome news? *Frontiers 03: New writing on cutting-edge science by leading scientists*, ed. Tim Radford, (Trowbridge: Atlantic Press, 2003), p.77.
9. [Mad scientist who wants to put a microbe in your tank](#)
10. [Celera Genomics completes the first assembly of the human genome](#)
11. Press conference, June 26, 2000
12. [The first draft of the Book of Humankind has been read](#)
13. [J. Craig Venter Steps Down as President of Celera Genomics](#)
14. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945. doi: [10.1038/nature03001](https://doi.org/10.1038/nature03001)

Our contribution

[Chromosome 1](#)

[Chromosome 6](#)

[Chromosome 9](#)

[Chromosome 10](#)

[Chromosome 11](#)

[Chromosome 13](#)

[Chromosome 20](#)

[Chromosome 22](#)

[Chromosome X](#)

Related links

[History](#)

[Achievements](#)

[Milestones](#)

[John Sulston](#)

[Frederick Sanger](#)

[Genome reference consortium](#)

[YourGenome](#)

[Help](#) | [Contact us](#) | [Legal](#) | [Cookies policy](#) | [Data sharing](#)

Connect with us:



Wellcome Trust Sanger Institute, Genome Research Limited (reg no. 2742969) is a charity registered in England with number 1021457

Last modified: Wed, 30 Jan 2013 10:54:59 GMT