



“Sequencing Human Genomes”

Adam Bostanci
ESRC Centre for Genomics in Society
University of Exeter
United Kingdom
a.w.s.bostanci@ex.ac.uk

(DRAFT – accepted for publication in *The Mapping Cultures of 20th Century Genetics*
ed. H.-J. Rheinberger and J.-P. Gaudillière, Routledge forthcoming 2003 – DRAFT)

The history of the Human Genome Project is a history of mapping projects. In its course, geneticists and molecular biologists surveyed human chromosomes with cytogenetic, genetic and physical markers. The maps featuring these landmarks were subsequently often collated or mapped onto one another, and eventually biologists began to 'sequence' human DNA, a process customarily explained as mapping the hereditary material at the highest possible resolution. But even this final phase of the Human Genome Project occurred not once, but twice. In February 2001 independent research groups described preliminary drafts of the human genome in separate publications: A consortium of laboratories commonly known as the Human Genome Project published its draft in *Nature* (International Human Genome Sequencing Consortium 2001). The other version of the human genetic code was the product of Celera Genomics, a company in Rockville, Maryland. This draft was described in the same week's issue of *Science* (Venter *et al.* 2001).

Visionary molecular biologists have always conceived of the human genome as a single natural object. Dubbing their fantastic plan of determining the sequence of its chemical building blocks as the 'holy grail of genetics', they predicted that the human genome sequence would eventually become "the central organising principle for human genetics in the next century" (Waterston and Sulston 1998: 53). Arguably, the publication of two human DNA sequences threatens to spoil these aspirations. Making a virtue out of necessity, one team of scientists concluded in February 2001: "We are in the enviable position of having two distinct drafts of the human genome sequence. Although gaps, errors, redundancy and incomplete annotation mean that individually each falls short of the ideal, many of these problems can be assessed by comparison" (Aach *et al.* 2001: 856).

Relating the discoveries of the Human Genome Project and Celera Genomics as imperfect, yet comparable representations of the same natural object is a common move, but by no means the only way of making sense of these discoveries. Alternatively, one might conceive of the two versions of the human genome as separate objects. This is not

to say that the drafts are irreconcilable in principle, but for the time being there are good reasons to speak of them as distinct objects in scientific practice. First, independent research groups working with different methods produced them: The Human Genome Project pieced together its version of the human genetic code in a "map-based sequencing" operation. Celera Genomics, in contrast, employed a "whole-genome shotgun". Second, the two drafts differ in ways that go beyond superficial sequence similarities and discrepancies, such as their topology. Third, it remains to be seen whether the two drafts of the human genome will be reconciled at all.

In the controversies surrounding efforts to sequence the human genome, the scientific strategies of the Human Genome Project and Celera Genomics were usually compared in terms of feasibility and cost. Each camp tried to persuade the public that its own approach was better suited for sequencing the human genome. My aim is not to address these disputes directly, but to extract from them information about the diverging scientific practices in the two projects. As one observer pointed out, from "an unbiased view, the two plans seem to be simply two methods to obtain genomic information" (Goodman 1998: 567). Hence I compare the research centres at which map-based sequencing and the whole-genome shotgun emerged and examine the social, epistemic and ontological roles genome maps came to play in these dissimilar sequencing regimes. This analysis shows that the disputes between the Human Genome Project and Celera Genomics emerged in the first instance from differences in scientific practice, which were later connected with gene patenting and the publication of human genome sequence data.

MAPPING AND SEQUENCING GENOMES

In scientific practice, mapping projects go hand in hand with the alignment of people, practices and instruments (Turnbull 2000, Ch. 3). The production of coherent maps requires the introduction not only of standardised measures but of standardised measuring practices, which, in turn, often impinge on moral and productive economies more

generally. The Human Genome Project is a case in point: Many biologists initially opposed the production of 'global' maps of the human genome because they feared such an endeavour would undermine their scientific autonomy (Balmer 1996; Cook-Deegan 1994). Even as the programme was slowly incorporated into the cottage industry of molecular biology of the early 1990s, alternative mapping and sequencing regimes were evaluated not only as experimental techniques but as quasi-political modes of organisation. As I will illustrate, genome maps themselves later came to play an important organisational role in the sequencing regime of the Human Genome Project.

Around 1990 several teams of scientists around the world were constructing 'physical maps' of the human genome. This type of genome map depicts the order and distance between genomic sites at which 'restriction enzymes' sever DNA. The enzyme *HindIII*, for instance, cuts DNA molecules in two wherever it encounters the sequence of chemical building blocks abbreviated by the letters AAGCTT. When DNA is exposed to *HindIII* and the resulting fragments are analysed, AAGCTT is said to become a 'landmark'. Physical mapping projects thus establish the order and spacing of such landmarks within a genome of interest. The resultant 'maps' comprise both an ordered collection of DNA fragments and a catalogue detailing the cleavage sites on them.

Fig.1 Popular Science: A diagram of common genome mapping techniques relates the progress of the Human Genome Project from low to high resolution (Cooper 1994: 205)

Yet, the sequence AAGCTT may occur hundreds of times in a large genome; knowledge of this landmark alone is not sufficient to describe a genomic location unambiguously. Consequently, molecular biologists initially only mapped particular chromosomes or chromosomal regions and developed additional landmarks, which often varied from laboratory to laboratory. As one scientist explained, one reason for dividing the early Human Genome Project by chromosome was to preserve the structure of scientific research: "No one wanted a large monolithic organisation dictating how the information would be gathered and disseminated" (R.K. Moyzis quoted in Cooper 1994: 111).

Others perceived this lack of standardisation as divisive. Maynard Olson, a geneticist at Washington University in St. Louis, recalled that he regarded large-scale physical mapping "as a kind of tower of Babel" because it gave rise to maps expressed in "completely incompatible languages" (M.V. Olson quoted in Cooper 1994: 123). The vision of the leaders of the Human Genome Project was to build one coherent map of the genome, which led Olson to the conclusion that, eventually all the maps "would have to be done again by whatever method proved most generic" (123). In response to this threat, he proposed to map the genome with sequence-tagged sites (STS), a new set of landmarks that eliminated the space for local variation.

Olson defined STS as sequence targets that are "operationally unique in the human genome," that is segments of DNA that can be detected in "presence of all other genomic information" (Olson *et al.* 1989: 245). For example, the site bounded by the pair of sequence tags AGTTCGGGAGTAAAATCTTG and GCTCTATAGGAGGCCCTGAG is a unique genomic site on chromosome 17 (example from Hilgartner 1995). Notably, STS are unique in the genome because they are defined by longer sequence tags than the enzymatic cleavage sites used to construct physical maps. Another advantage of STS was that these landmarks could be described fully as information in an electronic database and, unlike physical maps, required no exchange of delicate biological materials.

Olson presented STS as a "common language" that would "solve the problem of merging data from many sources" (Olson *et al.* 1989: 245). But as Hilgartner has argued, the new landmarks also embodied a particular management philosophy for the Human Genome Project. While the chromosomes initially emerged as the natural units of its organisation, STS later became the technological mode of management. STS were endorsed by scientific leaders because they represented those features of the human genome that permitted to tighten a network of dispersed laboratories by making their research commensurable and cumulative (Hilgartner 1995).

By 1996 the first global STS-map of the human genome had been assembled (Schuler *et al.*, 1996). This consensus or unified map of the genome, as it also came to be known,

contained more than 15,000 STS that had been mapped by an international consortium of laboratories. This map was 'global' in two senses: On the one hand, it made the genome accessible as one coherent domain, rather than a collection of 'local' physical maps. On the other hand, research centres anywhere in the world could henceforth work with landmarks that were conveniently available in an electronic database. STS made it possible to align existing physical maps with the consensus map. Later the Human Genome Project adopted the consensus map as an organisational framework for genome sequencing.

Sequencing

The term 'sequencing' refers to a routine form of analysis that establishes the succession of the four biochemical building blocks A, T, C and G in DNA molecules. In plain language, sequencing is commonly explained as mapping DNA at the highest possible resolution. Indeed, biologists often invoke the theme of increasing resolution in popular accounts of the Human Genome Project. The analogy makes it possible to relate the collection of genome maps that were produced in the course of the Human Genome Project as a series of ever more accurate representations, and to rationalise its history as a natural progression from mapping at low resolution to mapping at high resolution. In addition, the analogy permits to differentiate sequencing from previous mapping projects in the Human Genome Project: sequencing reveals the territory of the genome itself, not another partial map thereof. As one report predicted, the human DNA sequence would be "the ultimate map" containing "every piece of genetic information" (The Sanger Centre and The Washington University Genome Sequencing Center 1998: 1197). But alluring as this tale of progress may be, the practical relationship between genome mapping and DNA sequencing in the Human Genome Project has been more haphazard and more interesting.

Some of the first biochemical reagents and a strategy for large-scale DNA sequencing were developed during the 1970s by Cambridge biochemist Fred Sanger, who later shared one of his two Nobel Prizes for the invention of this method. During the 1980s, his

reagents were modified in such a manner that the results could be sequentially read by a computer, and commercial DNA sequencing machines became available in 1987. Since then, DNA sequencing machines have been the workhorses of all large-scale genome sequencing projects. But no matter by how much DNA sequencing technology has improved, one fundamental limitation of this form of biochemical analysis has remained: Only several hundred letters of the genetic code can be deciphered per sequencing experiment. The chemical composition of longer DNA molecules has to be re-constructed by aligning the partial sequence overlaps of DNA fragments. According to one of the first reports, this method, "in which the final sequence is built up as a composite of overlapping sub-fragment sequences, has been aptly termed "shotgun" DNA sequencing" (Anderson 1981: 3015). Notably, both the Human Genome Project and Celera Genomics eventually used versions of the shotgun strategy to sequence human DNA. The two genome sequencing strategies are schematically represented in Figure 2 and 3 (for a review see Green, E. D. 2001).

Fig. 2 Genome Sequencing: The strategy employed by the Human Genome Project simplified the problem of shotgun assembly by sequencing mapped fragments one at a time rather than the whole genome at once. (IHGSC 2001: 863)

Fig. 3 Map-based sequencing vs. Whole-genome shotgun: A scheme of the genome sequencing strategies of the Human Genome Project and Celera Genomics illustrates the different spatialisation steps involved (Waterston 2002: 3713)

In theory, there is no limit to the length of DNA molecules that can be sequenced by means of the shotgun method, the fundamental sequencing strategy. As long as a sufficiently large number of overlapping fragments is generated, the DNA sequence can be re-constructed by aligning overlaps. In practice, however, sequencing large genomes in this manner creates formidable challenges: Data from millions of sequencing experiments has to be stored, manipulated and assembled with special computer algorithms. Assembly is especially complicated if the genome contains repetitive sequences, which give rise to ambiguous overlaps. Ultimately, assembly may fail or

require data from additional experiments to resolve ambiguities. Despite these shortcomings of the sequencing technology available in the early 1990s, some biologists established large-scale genome sequencing projects at specialised research facilities. After all, one stated aim of the Human Genome Project was to sequence the entire human genome by 2005.

THE SANGER CENTRE

The Sanger Centre was opened in the small village Hinxton near Cambridge in 1993 (Fletcher and Porter 1997). Its founding director John Sulston had several years earlier constructed a complete physical map of the genome of the small nematode worm *C. elegans* and, more recently, embarked on a project to sequence its genome at the Laboratory of Molecular Biology in Cambridge. From the outset John Sulston organised this sequencing project as a transatlantic collaboration with Bob Waterston at Washington University in St. Louis. Generous funding from a biomedical charity, the Wellcome Trust, enabled Sulston to re-locate into some disused laboratory buildings on a country estate in Hinxton, and to expand the sequencing project. Waterston concurrently established a genome sequencing centre in St. Louis. Contemporary observers hailed their cooperation as the flagship of the Human Genome Project. In due course, it became a model for its organisation.

Spanning 100 million biochemical building blocks, the genome of *C. elegans* was two orders of magnitude larger than any genome that had been sequenced by the early 1990s. Sulston approached this formidable task by sequencing selected fragments of DNA from the map of the *C. elegans* genome, which he had constructed several years earlier together with Alan Coulson and Bob Waterston (Coulson *et al.* 1991). This sequencing strategy became known as 'map-based sequencing' or 'hierarchical shotgun sequencing'. The physical map served him as a convenient repository of DNA. In addition, sequencing in this step-by-step manner effectively eliminated technical problems associated with

whole-genome shotgun sequencing. As sequence information obtained from the fragments of the physical map was local, the problem of long-range misassembly was eliminated. Sequencing small portions of the genome also reduced the likelihood of short-range misassembly. As an early report from Sulston's laboratory justified, this strategy was feasible with existing technology (Sulston *et al.* 1992). Nonetheless, other scientists criticised the map-based approach for yielding much non-informative data from repetitive sequences at an early stage of the project.

While sequencing selected fragments from the physical DNA map simplified assembly, it also facilitated collaboration. The fragments introduced modularity, which provided a mechanism for the cooperation of research groups within the sequencing project. At the Sanger Centre, for instance, separate sequencing teams initially worked on different fragments of DNA. The collaboration with Washington University was organised similarly. As Alan Coulson, then a senior scientist at the Sanger Centre, explained in 1994: "We would not be able to divide the project in a sensible way between us and St. Louis, if we didn't have a map to do it in an ordered fashion" (A. Coulson at Sanger Centre Open Day 1994).

At the same time, map-based sequencing began to structure the daily work of the scientists and an increasing number of technicians at the Sanger Centre. Each sequencing team worked on a fragment of DNA from start to finish and routines associated with map-based sequencing began to be differentiated within the team. Later, especially when sequencing human DNA, specialists performed routine steps such as mapping, raw sequence production or checking for errors in the assembly. As the Sanger Centre moved into larger purpose-built facilities on the estate in 1996, the differentiated tasks were also separated spatially. Mapping and sequencing, for instance, were accommodated in separate parts of the new building (Sulston 2000). Increasingly, the Sanger Centre operated like a sequence assembly line, albeit the facade of the country estate was maintained by converting Hinxton Hall into a conference centre.

Genome maps not only facilitated cooperation with distant genome sequencing centres, but connected the Sanger Centre with other researchers studying *C. elegans*. Sequencing the genome in portions at a time meant that finished sequences could be published continuously in an electronic database. This provided other researchers with a continuous stream of intelligible information and helped convince the 'worm community' that sequencing at a large, centralised facility was worthwhile (Sulston 2000). The sequences published online were useful in conventional academic research because their genomic provenance had been mapped. Other researchers could draw on the sequences to locate and study the genes of *C. elegans*.

A diagram presented by Alan Coulson at its Open Day in 1994 (reproduced as Figure 4) further illustrates how the Sanger Centre articulated its place in the research community: As a central facility, it would maintain the physical genome map of *C. elegans* and supply the research community with DNA material. It would also curate a database of genomic information on the worm, continuously adding its own sequencing data and genetic mapping information generated by dispersed research groups. In short, the Sanger Centre presented and understood itself as a service provider. As John Sulston commented on this database at the summit of the Human Genome Organisation that year: "I feel it has not been so much a map as a means of genomic communication on *C. elegans*" (J.E. Sulston at the Human Genome Summit 1994). Such talk of cooperation appealed scientists who opposed 'big science' and justified the allocation of resources to genome sequencing facilities like the Sanger Centre. More importantly, it defined the role of the Sanger Centre in the research-industry-technology landscape and ultimately the Human Genome Project.

Fig. 4 Alan Coulson's Diagram: Presented at the Sanger Centre Open Day in 1994, the diagram articulates the place of the Sanger Centre within the traditional biological research community (Coulson 1994)

This historical sketch shows how the physical map of the *C. elegans* genome became the basis of a low-risk sequencing strategy at the Sanger Centre and came to underpin its

collaboration with the Washington University. In addition, this sequencing strategy produced a stream of potentially useful sequence data that was continuously released, convincing many biologists that this mode of organisation might be suitable for the Human Genome Project as a whole. Around 1994, laboratories around the world began to form consortia, planning to sequence parts of the human genome in arrangements resembling the collaboration between the Sanger Centre and Washington University. And Tom Caskey, the president of the Human Genome Organisation urged his colleagues that the *C. elegans* sequencing project should be regarded as a "model system of how we should be conducting ourselves" (C.T. Caskey at the Human Genome Summit 1994).

THE INSTITUTE OF GENOMIC RESEARCH

The Institute of Genomic Research was set up in 1992. Its founding director Craig Venter had previously overseen a sequence-based discovery programme of human genes at the National Institutes of Health (NIH). When he proposed to expand this programme, the extramural research community objected to his plan, which was subsequently dropped by NIH (Cook-Deegan 1994: 311-325). Venter accepted funding from a company instead and set up The Institute of Genomic Research (TIGR). TIGR – Venter pronounced the acronym as "tiger" – was initially embroiled in controversy because it kept sequence information on human genes obtained in contract research in private databases. Later TIGR – its detractors had begun to use the diminutive pronunciation "tigger" – was aggressively promoted as a new model for high-throughput DNA sequencing and analysis (Adams *et al.* 1995). However, the research of this genome sequencing centre (Figure 5) evolved into an operation that differed from the Sanger Centre with respect to implicit goals, specialist expertise and organisation. By 1995, TIGR had sequenced the first bacterial genome by means of what came to be known as the 'whole-genome shotgun'.

The gene discovery programme carried out at TIGR in the early 1990s differed from work at the Sanger Centre in two respects: First, by design sequencing experiments carried out at TIGR only yielded sequences corresponding to protein-encoding genes. Indeed, in the early 1990s some scientists argued that this approach should take precedence over map-based sequencing because the coding sequences constituted the major information content of the human genome (see e.g. Adams *et al.* 1991). Second, while the Sanger Centre produced sequences whose genomic provenance had been mapped prior to sequencing, TIGR generated sequence data whose origin in the genome was unknown.

On the one hand, these differences reflected endemic disagreements on the priorities of the Human Genome Project. Was the objective to discover all genes before revealing less promising sequences or to sequence genomes methodically from end to end? On the other hand, these differences in laboratory practice show how locally existing resources were re-invested in future research at both centres: The Sanger Centre drew on a recent accomplishment of its founders, the physical map of the *C. elegans* genome, as a DNA repository. TIGR redeployed computational methods from its early gene discovery programme in subsequent genome sequencing projects. As an early report from TIGR phrased it, computational methods developed to create assemblies from hundreds of thousands sequence shreds "led us to test the hypothesis that segments of DNA several megabases in size, including entire microbial genomes, could be sequenced rapidly, accurately, and cost-effectively by applying a shotgun sequencing strategy to whole genomes" (Fleischmann *et al.* 1995: 496). Features of the laboratory information management at TIGR system were "applicable or easily modified for a genomic sequencing project," too (497).

Fig. 5 Factory or Lab? The interior of the Institute of Genomic Research, a typical genome sequencing centre of the mid-1990s (Cover of ASM News, March 1996)

The DNA of *Haemophilus influenzae*, the pathogen that causes ear infections, was the first large bacterial genome to be sequenced in this manner at TIGR in 1995 (Fleischmann *et al.* 1995). The entire genome spanning 1.8 million chemical building blocks was fragmented with ultrasound. Fragments of an appropriate size were sequenced from both ends in 23,643 sequencing reactions, generating pairs of sequence data whose separation and orientation relative to one another was known. The data was then assembled with TIGR's specialised software in 30 hours of high performance computing and yielded 140 sequence assemblies. Gaps between them were then closed by means of targeted experiments. Finally, the correctness of the sequence was verified by comparison with an existing physical map of the genome. "Our finding that the restriction map of the *H. influenzae* Rd genome based on our sequence data is in complete agreement with that previously published further confirms the accuracy of the assembly," assured the report (508).

This account of a successful 'double-barrelled' whole-genome shotgun experiment shows that the physical map of the *H. influenzae* genome helped to validate the DNA sequence, but played no visible role in the organisation of the genome project itself. Besides, as raw data from shotgun sequencing was unmapped and unintelligible before assembly, TIGR deposited the finished genome sequence in a database, rather than publishing sequences incrementally online for the benefit of the scientific community. Comparison with the Sanger Centre suggests that although the term strategy conjures up the image of a purposeful plan determined by theoretical considerations, the genome sequencing strategies of both TIGR and the Sanger Centre, in practice, drew on locally existing resources and expertise. Both research centres strove to capture the impetus and align the objectives of the Human Genome Project with their own research agendas. Not surprisingly, Craig Venter asserted that the whole-genome shotgun was fit for carrying out the Human Genome Project: "this strategy has potential to facilitate the sequencing of the human genome," he concluded (511).

MAP-BASED SEQUENCING VERSUS WHOLE-GENOME SHOTGUN

In February 1996, representatives from all major genome sequencing centres convened in Bermuda to evaluate approaches for sequencing the human genome. The delegates – among them John Sulston and Craig Venter – considered a number of genome sequencing strategies, but in the end concluded that sequence data had to be generated by the various approaches before they could be evaluated properly. Nevertheless map-based sequencing became the status quo of the Human Genome Project. The collaborative mechanisms of the *C. elegans* project were extended to human DNA sequencing, albeit with some modifications. The main difference between the *C. elegans* project and the Human Genome Project was that a complete repository of mapped DNA fragments existed from the outset of the *C. elegans* sequencing cooperation, but only a digital set of ordered landmarks was available for the human genome at the time. Consequently, each genome sequencing centre had yet to align suitable pieces of human DNA with this unified map in preparation for sequencing. Although the production of 'sequence-ready maps' proved to be the bottleneck in the human DNA sequencing for some time, the leaders of the Human Genome Project felt that this drawback was outweighed by other advantages of map-based sequencing.

Between 1996 and 1998, the digital unified map of the human genome provided mechanisms and metaphors for collaboration and coordination in the Human Genome Project. For example, when the Wellcome Trust announced its intention to fund the sequencing of one sixth of the genome, the Sanger Centre could register corresponding 'sequencing claims' on an internet-based map known as the Human Genome Sequencing Index. The 'boundaries' of sequencing claims were defined by means of 'unique markers' from the genome map. In addition, the members of the human genome sequencing consortium endorsed a set of rules, known as 'sequencing etiquette', that specified which 'regions' each laboratory was entitled to claim based on past sequencing achievements (Bentley *et al.* 1998).

One year after the meeting in Bermuda, Jim Weber from the Marshfield Medical Research Foundation and Gene Myers, a computer scientist at the University of Arizona,

made an argument for sequencing the human genome by means of the whole-genome shotgun strategy in the journal *Genome Research* (Weber and Myers 1997). Phil Green, a former member of the *C. elegans* sequencing project, rebutted this proposal in an article published back-to-back in the same volume (Green, P. 1997). He criticised that the whole-genome shotgun offered no mechanisms for upholding the practical values that had held the *C. elegans* collaboration together and were being instituted in the Human Genome Project. In his words, it was "unclear how the project could be distributed among several laboratories" (416). Green also characterised the whole-genome shotgun regime as "inherently a monolithic approach" (411), which in terms of organisation might be read as a statement of Lewis Mumford's classic thesis that technologies are intrinsically authoritarian or democratic. After all, the whole-genome shotgun was at the time only carried out single-handedly at TIGR.

According to Green, the advantages of map-based sequencing went beyond coordination. The strategy also made room for accommodating differences between laboratories, as well as providing means for isolating problematic regions. In this manner the Human Genome Project could be partitioned between investigators "without forcing them to interact" (411), and laboratories could explore "alternative sequencing strategies and methods independently without redundancy of effort" (410). Any unintended replication of sequencing efforts was considered to amount to unwarranted and costly duplication. Besides, the map-based sequencing regime permitted "to isolate problematic regions" (411). Whole-genome shotgun sequencing, in contrast, offered no mechanism for separating contributions from individual laboratories and for isolating problematic regions: "Problems with data quality in one laboratory would affect all laboratories, because any region of the genome would have shotgun reads generated at all labs," lamented Green (416).

In making the case for whole-genome shotgun, Weber and Myers negated some of these alleged advantages and left others unanswered. Regarding collaboration, for instance, they suggested, contrary to Green's claim, that laboratories "throughout the world could participate in raw sequence generation" (401), effectively a proposal to decentralise

sequence production. Most importantly, they proposed a distinct role for the unified map of the genome. In their sequencing regime the consensus map would function as a 'scaffold' for the assembly of the human DNA sequence, not as a framework for the organisation of the Human Genome Project. According to Weber and Myers, the task of assembly was to build "scaffolds that span adjacent STSs" (403). Since it was available, information from the unified map of the human genome could assist assembly of the whole-genome shotgun, but would play no role in the organisation of the programme.

In short, the sequencing strategy of the Human Genome Project constituted a moral and productive economy, whose common ground, so to speak, was the map of the human genome. Reminiscent of the *C. elegans* project, the map underpinned cooperation and coordination, but also came to orchestrate access to resources, accountability for cost and allocation of scientific credit. Although the whole-genome shotgun was organised around the same automatic DNA sequencing machines, its moral, productive and epistemic standards were different to the extent that Green argued that it was "incompatible" with map-based sequencing (411). Having explored the mechanisms of collaboration and cooperation, the analysis will now focus on ontological differences emerging from the two genome sequencing strategies.

Ontological Assumptions and Consequences

Throughout the history of the Human Genome Project, the human genome – according to the vision of its leaders – had been envisaged to be a single and coherent domain – 24 unbroken sequences of genetic code corresponding to the human chromosomes. The dispute in the pages of *Genome Research* thrived on this assumption: The arguments of the adversaries were framed primarily in terms of feasibility and cost, albeit they also overtly disagreed on the objectives of the Human Genome Project. Between the lines they also acknowledged that their strategies were based on variant assumptions about the nature of the human genome, on the one hand, and would ultimately manifest different

versions of the genome, on the other. Their remarks are indicative of differences beyond superficial sequence similarities and discrepancies.

For Phil Green, the advocate of map-based sequencing, the ultimate aim of the Human Genome Project was to generate an accurate and complete human DNA sequence, which was to serve as a "reference against which human variation can be catalogued" (410). This aim, he felt, would be best attained by map-based sequencing, which permitted to isolate and carefully sequence problematic regions. For Weber and Myers, in contrast, the true objective of the Human Genome Project was to "sequence all human genes" or, more pragmatically, "to generate as much of the critical sequence information as rapidly as possible and leave clean-up of gaps and problematic regions for future years" (406). Weber and Myers argued that these objectives could be achieved with minimum effort with the whole-genome shotgun strategy, which had emerged from a sequence-based gene discovery programme at TIGR.

The opponents also admitted that their approaches would manifest different versions of the human genome. Phil Green, the advocate of map-based sequencing, anticipated that the whole-genome shotgun had "a high probability of failure" by which he implied that the final assembly would contain a very large number of gaps and fail to reconstruct problematic regions faithfully (411). Jim Weber and Gene Myers conceded that their strategy would by no means produce in a "single unbroken sequence for entire chromosomes," which was perfectly reconcilable with their priorities (404). They found fault with map-based sequencing because rearrangements in DNA fragments required for map-based sequencing could cause artefacts. In their eyes map-based sequencing was working towards an "arbitrary, mythical goal of 99.99% accuracy of a single, artifactual (in places) and nonrepresentative copy of the genome" (406). Green countered that the artefacts caused by rearrangements could be prevented easily in map-based sequencing.

Falling under the heading of assumptions about the nature of the human genome, Green criticised that, in computer simulation of the whole-genome shotgun, Weber and Myers had assumed that repeated sequences are distributed uniformly throughout the genome.

Green asserted this was an "incorrect assumption about the nature of the genome" (411). Similarly, one might suggest that Weber and Myers' failure to propose a means for managing problematic regions might have arisen from similar assumption about the nature of the genome. For Green, arguing from the vantage point of map-based sequencing, it was natural to assume that the human genome had 'regions', which were bounded by landmarks of the consensus map or embodied in physical DNA fragments. Crucially, by virtue of these regions problematic data could be isolated during the incremental analysis of the human DNA sequence. The spatial metaphors used to manage the Human Genome Project became properties of the human genome itself and a resource for its organisation. Weber and Myers, in contrast, articulated no mechanism for dealing with problematic data. They simply proposed that "only a few or possibly even one large informatics group would assay the primary task of sequence assembly" (401). In other words, they proposed not to regionalise the human genome, but to divide the community of genome researchers into those responsible for raw sequence production and those responsible for assembly.

Finally, Green's assertion that the whole-genome shotgun was "inherently monolithic" can be read as a hint at deeper differences. In a whole-genome shotgun regime the human genome sequence emerged as a whole only after a painstaking process of sequence assembly. In map-based sequencing, in contrast, the genome sequence emerged continuously and incrementally as a mosaic of contributions sequenced and assembled at dispersed genome sequencing centres. The differences between the genome sequences produced by means of the two strategies might therefore reside beyond the surface of sequence similarities and discrepancies. For example, if problems with sequence data or the assembly algorithms are discovered in the future, what will be the effect on the respective sequences? For a sequence derived by means of whole-genome shotgun one might expect global adjustments, whereas consequences for the product of the map-based regime might well be local. As a generalisation about these differences, one might say that the human genome, as envisaged and produced by the Human Genome Project, had the topology of a map. Its large-scale spatialisation was a result of preliminary physical mapping projects. The human genome produced by means of a whole-genome shotgun

might be better conceptualised as an algorithm. The spatialisation of sequence data emerged primarily during the assembly of sequence overlaps by means of specialised computer algorithms developed at TIGR.

CELERA GENOMICS

One year after the controversy in *Genome Research*, Craig Venter deserted the collaborative framework of the Human Genome Project to become the director of Celera Genomics, a company intending to sequence the human genome single-handedly by means of the whole-genome shotgun (Venter *et al.* 1998). The leaders of the Human Genome Project perceived the venture announced on 9 May as a threat. Venter's plan amounted to sequencing the human genome significantly faster and cheaper than the Human Genome Project, which had by 1998 merely deciphered five percent of the genome. The defence of the Human Genome Project had four prongs: First, in a series of scientific articles its leaders justified their sequencing strategy (The Sanger Centre and The Washington University Genome Sequencing Center 1998; Bentley *et al.* 1998; Waterston and Sulston 1998). Maynard Olson and Phil Green, for instance, argued that it was "essential to adopt a 'quality-first' credo" on "both scientific and managerial grounds" (Olson and Green 1998). Second, they explained the strategy to the scientific community alleging that it was the only approach that could be "efficiently coordinated to minimize overlap between collaborating groups" (quote from The Sanger Centre and The Washington University Genome Sequencing Center 1998: 1099; Dunham *et al.* 1999). Third, the strategy of the Human Genome Project was revised considerably several months later to allow for the production of a preliminary draft of the human genome. Finally, arguments for the continuation of the Human Genome Project went hand in hand with public criticism of Craig Venter and disapproval of his sequencing strategy. Craig Venter was cast in the role of an unscrupulous profiteer. In plain words, his method was often dismissed as 'quick and dirty'.

While I can judge neither the character nor the motivation of any of the actors, I suspect that attempting to 'shotgun' the human genome might have been a question of personal pride for Craig Venter. Having pushed the limits of shotgun sequencing at TIGR and later seen this method rejected by the scientific community, he may instead have teamed-up with a company willing to fund his endeavour. He may also have felt that inadequate use was being made of the facilities available at TIGR under a map-based sequencing regime. For, TIGR was designed for large-scale shotgun sequencing just as much as shotgun sequencing had originally been designed around the resources available at TIGR. At any rate, Craig Venter must have been well-aware that this endeavour would yield a product incongruent with the product of the Human Genome Project.

Only one popular article published on the website of the Lasker Foundation (1998) attempted to compare the sequencing strategies employed by Venter and the Human Genome Project in their own right, concluding that the order of mapping and sequencing was reversed in the two approaches: "In a sense, Venter's approach is to sequence first, map later. In the public project, genes are first mapped to a specific, relatively small region of the genome, then sequenced." The article highlighted that both sequencing strategies drew on existing STS-maps of the human genome: The Human Genome Project used STS-maps prior to sequencing as described above. Celera Genomics searched for STS in its primary sequence assemblies, thus establishing their genomic provenance in a secondary spatialisation step called 'anchoring'. Despite its methodological focus, this article pitted the two strategies against each other in terms of feasibility.

After the formation of Celera Genomics debate was polarised in public to suggest that Celera Genomics might keep its sequence data secret and establish a monopoly by patenting the majority of human genes. These motives were cited as the main reasons for the inability of the Human Genome Project to cooperate with Celera Genomics. As the controversy in the pages of *Genome Research* had shown, another important obstacle to cooperation resided in the fact that the whole-genome shotgun provided no proven mechanisms for upholding the practical values that had made the *C. elegans* project

work. In public, the lack of these mechanisms was articulated in terms of gene patenting and data release.

Patenting

In presenting his business plan, Craig Venter assured the readers of *Science* that "we do not plan to seek patents on primary human genome sequences" but to identify and patent no more than "100 to 300 novel gene systems from among the thousands of potential targets" for medical interventions (Venter *et al.* 1998: 1541). The leaders of the Human Genome Project, in contrast, portrayed the private venture as contrary to the public interest. "Underlying all this is concern about a land grab – that is we're concerned about massive patenting," commented one scientist (R. Gibbs quoted in *The Australian* 31 March 1999), and in a sense his remark was accurate: The venture of Celera Genomics amounted to a 'land grab' in so far as the human genome was conceptualised as a map by the Human Genome Project. In proposing the whole-genome shotgun, Venter ignored the sequencing etiquette and all sequencing claims that had already been registered on the Human Genome Sequencing Index. As Venter diagnosed the cause of irritation, "we are changing the rules and that upsets people" (J.C. Venter quoted in Larkin 1999: 2218).

Data Release

Another outcome of the talks in Bermuda had been an agreement to make the human DNA sequence publicly available as it was produced. According to the "Bermuda principles" human sequence data "should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society" (Smith and Carrano 1996). In practice, this meant that contributing laboratories had to publish assembled sequence data on the internet within 24 hours, or else faced exclusion from the human genome sequencing consortium. Another rudiment of the *C. elegans* sequencing project, this policy aimed to minimise competition among genome sequencing centres and to deter them from focusing their efforts on lucrative genes. Celera Genomics, in contrast, announced it would release data into the public domain at least every three months and assured that, being an information company, "complete

public availability of the sequence data" is "an essential feature of the business plan" (Venter *et al.* 1998: 1541).

Once again, the policy of the Human Genome Project had been justified on epistemic and managerial grounds in 1996. "The finished sequence should be released directly upon completion," argued a scientist of the Sanger Centre, insisting that this was required both to "optimise coordination" and to facilitate "independent checking" (Bentley 1996: 533). More generally, the laboratories of the Human Genome Project planned to implement a dispersed regime of data quality monitoring in 1996: Each laboratory sequenced regions of the genome but it was envisaged that the ultimate validation of the data would arise from independent checking by other laboratories. To this end the network of laboratories even proposed a data exchange exercise, in which the same raw data would be assembled at different laboratories to check for discrepancies. In contrast to the immediate data release policy of the Human Genome Project, the custom at TIGR was to sequence, check and annotate sequence data single-handedly before publication. Hence, on epistemic grounds, Craig Venter and his senior scientist Mark Adams objected that quality control would be compromised if immediate data release was enforced at all genome sequencing centres (Adams and Venter 1996).

Moreover, despite the managerial justification for immediate data release, compliance with the Bermuda principles was by no means a necessary condition for coordination, as illustrated by the data release practices of the yeast genome sequencing project. This genome sequencing project was carried out by a network of 79 laboratories and coordinated by means of a map of its genome. Delayed data release was an integral part of the collaboration. For example, when the completion of sequencing the yeast genome was announced in March 1996, twenty percent of it was still withheld from public databases until industrial sponsors and investigating scientists had satisfied their own interests. André Goffeau, the administrator of the yeast genome project, defended delayed data release "on the grounds that the scientists who did the work deserve to be the first to reap some benefits. 'We cannot just give this away,' he said – a view shared by some researchers – especially those from small labs" (Kahn 1996). If truth be told, several of

the large genome sequencing centres that participated in the Human Genome Project had earlier contributed in the yeast sequencing project. Although differences in data release practices caused tensions between the projects, both collaborations hung together as long as laboratories restricted themselves to sequencing regions of the genome within a map-based sequencing regime.

After the formation of Celera Genomics the leaders of the Human Genome Project suggested that, depending on future business expedients, the company might refuse to make its sequence data available to the public. Francis Collins, director of the National Human Genome Research Institute, argued as much on 17 June at a hearing in the U.S. House of Representatives. The Human Genome Project was "the best insurance that the data are publicly accessible," he said (F.S. Collins in U.S. House of Representatives 1998: 80). Admittedly, in March 2000, an attempt at cooperation between the two camps fell through because of disagreements on data release. However, on this occasion a failure in reconciling the legal and commercial obligations to which each of the data sets were bound caused the downfall of the negotiations, not a categorical refusal by Celera Genomics to publish or combine its data with that of the Human Genome Project (reported by Marshall 2000).

In February 2001, Celera Genomics and the Human Genome Project finally published their preliminary drafts of the human genome simultaneously but separately in *Nature* and *Science*. Celera Genomics made its sequence available to academic and non-profit researchers around the world. Albeit up-dates of the sequence will be available to paying customers only, the company has arguably honoured its assurance to make the sequence publicly available. Spokesmen of the Human Genome Project continue to criticise the terms that restrict the uses of the sequence data produced by the company. Debates over the conditions of access to the sequence data produced by publicly funded genome sequencing centres also continue to simmer (reported by Roberts 2002), as does the controversy about the propriety of patenting human genes. Celera Genomics has filed significantly more than 300 provisional patent applications, but maintains that, as anticipated, no more than 300 patents will be filed eventually.

Even after the publication of the two human genome sequences in February 2001, another controversy arose. To save time and money, Celera Genomics had supplemented its own shotgun sequence data with publicly available data from the Human Genome Project. The company had shredded this data, combined it with its own and assembled the combined set. Leaders of the Human Genome Project now charged that the data had not lost its positional information and that the combined assembly had succeeded only because of the inclusion of data produced by the Human Genome Project. In effect, they argued that Celera Genomics had not produced an independent human genome sequence but merely replicated the product of the Human Genome Project. This charge was repeated in a peer-reviewed journal in March 2002 (Waterston *et al.*) and rebutted in the same volume (Myers *et al.*). According to one newspaper report, Craig Venter countered that he had "shredded the data specifically to lose any positional information because he suspected the data had been misassembled in places" (Wade 2001).

CONCLUSION

Although the human genome has long been conceived of as single natural domain, Celera Genomics and the Human Genome Project have recently produced and will continue to produce different versions of the human genetic code. The Human Genome Project produced its draft as a mosaic of contributions from different laboratories and with the help of long-range maps. The genome produced at TIGR, in contrast, emerged from a whole-genome shotgun experiment. Here the primary spatialisation of data emerged during sequence assembly.

In February 2001, both drafts of the human genome had on the order of 100,000 gaps, were riddled by known and unknown inaccuracies and errors and differed in overall length. As both versions are gradually improved, it remains to be seen, whether the two products turn out to be as "incompatible" as the methods by which they were produced, "complementary", as suggested when leaders of the Human Genome Project attempted to make peace with Celera Genomics in 1999 (Lander 1999), or "hardly comparable"

because "different procedures and measures were used to process the data" (Bork and Copley 2001). As the most recent dispute between the Human Genome Project and Celera Genomics demonstrates (Waterston *et al.* 2002; Myers *et al.* 2002), the precise role of genome maps in the production of the two drafts published in February 2001 also continues to be debated.

Even if the two versions of the human genome were to be synthesised, practical mechanisms for doing so will have to be negotiated: One option would be to combine the underlying data for joint analysis, which is arguably what Celera Genomics has already done. However, as the latest in a long series of controversies between the two camps illustrates, such a cooperation would require negotiations about the status and value of positional information of map based data. Another conceivable approach would be to collate the two versions of the human DNA sequence and resolve discrepancies by further experiments. Finally, if all attempts to reconcile the two human genomes fail and the sequences remain separate, it remains to be seen whether they will ultimately be used in different ways, one as a reference sequence for basic research and public health, the other to satisfy the information needs of pharmaceutical companies.

NOTES

There are some aspects of the genomes controversy that I have deliberately not engaged with. In general, my argument has been a historical analysis of the two sequencing strategies as different experimental traditions, or epistemic cultures (Knorr-Cetina 1999), not a detailed analysis of the very latest papers published by the two camps. In the same vein, I have not recounted in detail the modifications of their respective strategies by Celera Genomics and Human Genome Project in the tumultuous years from 1999 until 2002. The details of the different assembly algorithms and quality assurance procedures developed by the Human Genome Project and Celera Genomics are yet to be examined. Finally, between May 1999 and March 2000 the genome of the fruit fly *Drosophila melanogaster* was successfully sequenced by means of a whole-genome shotgun in a formal collaboration of Celera Genomics with the Berkeley *Drosophila* Genome Project (Myers *et al.* 2000). Evidently, mechanisms that enabled the collaboration of the company with academic scientists were established in this project. And in December 2002 a consortium of publicly funded laboratories published a draft sequence of the mouse genome, which was produced by means of a whole genome shotgun. Evidently new mechanisms that made it possible to carry out the whole-genome shotgun in a network of dispersed laboratories were established in this project, too. (Mouse Genome Sequencing Consortium 2002).

ACKNOWLEDGEMENTS

I am indebted to John Sulston, who made himself available for an interview, and Alan Coulson, who provided a copy of his diagram from the Sanger Centre Open Day in 1994. Don Powell, press officer, has also been very helpful at the Sanger Centre, as has Barbara Skene of the Wellcome Trust. I thank the Studienstiftung des deutschen Volkes and Trinity College, Cambridge, for scholarships that enabled me undertake most of the research assembled in this chapter. Helen Verran and David Turnbull helped me in making sense of it. Katrina Dean read and criticised several of the preliminary drafts of this paper.

REFERENCES

- The Australian* (31 March 1999) "Clash over genome patents" p. 40.
- Aach, J., *et al.* (2001) "Computational comparison of two draft sequences of the human genome", *Nature*, 409: 856-859.
- Adams, M. D., *et al.* (1995) "A model for high-throughput automated DNA sequencing and analysis core facilities", *Nature*, 368: 474-475.
- Adams, M. D., *et al.* (1991) "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project", *Science*, 252: 1651-1656.
- Adams, M. D. and Venter, J. C. (1996) "Should non-peer-reviewed raw DNA sequence data release be forced on the scientific community?" *Science*, 274: 534-6.
- Anderson, S. (1981) "Shotgun DNA sequencing using cloned DNase I-generated fragments", *Nucleic Acids Research*, 9: 3015-3027.
- Balmer, B. (1996) "Managing Mapping in the Human Genome Project", *Social Studies of Science*, 26: 531-573.
- Bentley, D. R. (1996) "Genomic sequence information should be released immediately and freely in the public domain", *Science*, 274: 533-4.
- Bentley, D. R., *et al.* (1998) "Coordination of human genome sequencing via a consensus framework map", *TRENDS in Genetics*, 14: 381-384.
- Bork, P. and Copley, R. (2001) "Filling in the gaps", *Nature*, 409: 818-820.
- Cook-Deegan, R. (1994) *The Gene Wars: Science, Politics, and the Human Genome Project*, New York, London: W. W. Norton & Co.
- Cooper, N. G. (ed) (1994) *The Human Genome Project: Deciphering the Blueprint of Heredity*, Mill Valley, California: University Science Books.
- Coulson, A., *et al.* (1991) "YACs and the *C. elegans* genome", *Bioessays*, 13: 413-417.
- Dunham, I., *et al.* (1999) "The DNA sequence of human chromosome 22", *Nature*, 402: 489-495.
- Fleischmann, R. D., *et al.* (1995) "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd", *Science*, 269: 496-512.
- Fletcher, L. and Porter, R. (1997) *A Quest for the Code of Life: Genome Analysis at the Wellcome Trust Genome Campus*, London: The Wellcome Trust.
- Goodman, L. (1998) "Random Shotgun Fire", *Genome Research*, 8: 567-568.

- Green, E. D. (2001) "Strategies for the systematic sequencing of complex genomes", *Nature Review Genetics*, 2: 573-583.
- Green, P. (1997) "Against a Whole-Genome Shotgun", *Genome Research*, 7: 410-417.
- Hilgartner, S. (1995) "The Human Genome Project", pp. 302-315 in Jasanoff, S., Markle, G. E., Peterson, J. C. and Pinch, T. J. (eds) *Handbook of science and technology studies*, Thousand Oaks, California: Sage Publications.
- Human Genome Summit (January 1994) held at Rice University, Houston, Texas VHS recording held at Sanger Centre Library, Hinxton, U.K.
- International Human Genome Sequencing Consortium (2001) "Initial sequencing and analysis of the human genome", *Nature*, 409: 860-921.
- Kahn, P. (1996) "Sequencers split over data release", *Science*, 271: 1798.
- Knorr-Cetina, K. (1999) *Epistemic Cultures: How the Sciences Make Knowledge*, Cambridge, Massachusetts: Harvard University Press.
- Lander, E.S. (1999) "Shared Principles", letter to Celera Genomics (28 December 1999).
- Larkin, M. (1999) "J Craig Venter: sequencing genomes his way", *The Lancet*, 353: 2218.
- Lasker Foundation (1998) "Public, Private Projects at Odds Over Research Technique" published 30 September in *Comment*, a publication of Mary Woodard Lasker Charitable Trust, <http://www.laskerfoundation.org/comment/12/comm1.html> (accessed on 5 October 2001)
- Marshall, E. (2000) "Talks of public-private deal end in acrimony", *Science*, 287: 1723-5.
- Mouse Genome Sequencing Consortium (2002) "Initial sequencing and comparative analysis of the mouse genome", *Nature*, 420: 520-562.
- Myers, E. W. *et al.* (2000) "A whole-genome assembly of *Drosophila*", *Science*, 287: 2196-204.
- Myers, E. W., *et al.* (2002) "On the sequencing and assembly of the human genome", *Proc Natl Acad Sci U S A*, 99: 4145-4146.
- Olson, M. and Green, P. (1998) "A "Quality-First" Credo for the Human Genome Project", *Genome Research*, 8: 414-415.
- Olson, M. V., *et al.* (1989) "A Common Language for Physical Mapping of the Human Genome", *Science*, 245: 1434-1435.

- Roberts, L. (2002) "A Tussle Over the Rules for DNA Data Sharing", *Science*, 298: 1312-1313.
- Schuler, G. D., *et al.* (1996) "A Gene Map of the Human Genome", *Science*, 274: 540-546.
- Smith, D. and Carrano, A. (1996) "International Large-Scale Sequencing Meeting", *Human Genome News*, 7 (6): <http://www.ornl.gov/hgmis/publicat/hgn/hgn.html>
- Sulston, J. E. (19 March 2000), interviewed by the author, Sanger Centre, Hinxton, U.K.
- Sulston, J. E., *et al.* (1992) "The *C. elegans* genome sequencing project: a beginning", *Nature*, 356: 37-41.
- The Sanger Centre and The Washington University Genome Sequencing Center (1998) "Toward a Complete Human Genome Sequence", *Genome Research*, 8: 1197-1108.
- Sanger Centre Open Day (1994) VHS recording, Sanger Centre Library, Hinxton, U.K.
- Turnbull, D. (2000) *Masons, tricksters and cartographers : comparative studies in the sociology of scientific and indigenous knowledge*, Australia: Harwood Academic.
- U.S. House of Representatives (1998), Hearing of the Subcommittee on Energy and Environment of the Committee on Science, 17 June, "The Human Genome Project: How Private Sector Developments Affect the Government Program", Washington, DC: U.S. Government Printing Office.
- Venter, J. C., *et al.* (2001) "The sequence of the human genome", *Science*, 291: 1304-51.
- Venter, J. C., *et al.* (1998) "Shotgun sequencing of the human genome", *Science*, 280: 1540-2.
- Wade, N. (2 May 2001) "Genome Feud Heats Up as Academic Team Accuses Rival of Faulty Work", *The New York Times*, p. A15.
- Waterston, R. and Sulston, J.E. (1998) "The Human Genome Project: Reaching the Finish Line", *Science*, 282: 53-54.
- Waterston, R. H., *et al.* (2002) "On the sequencing of the human genome", *Proc Natl Acad Sci U S A*, 99: 3712-3716.
- Weber, J. L. and Myers, E. (1997) "Human Whole-Genome Shotgun Sequencing", *Genome Research*, 7: 401-409.