End of the beginning

Lincoln D. Stein

Just over three years ago, it was announced that a first draft of the human genome sequence had been completed. Gaps and errors remained, but the job of fixing those problems is now largely done.

GETTY IMAGES

This issue of *Nature* features an article[1] entitled "Finishing the euchromatic sequence of the human genome". It has been authored by members of the International Human Genome Sequencing Consortium (IHGSC), and appears on page 931. The article marks the latest, but by no means the last, milestone in this historic project. But readers can be forgiven for being a bit confused by the announcement. Wasn't the human genome 'finished' several years ago?

The answer is 'yes' — and 'no'. Early in 2001, the duelling IHGSC (public) and Celera Corporation (private) groups published papers in *Nature*[2] and *Science*[3] describing the completion of so-called 'draft' sequences. These sequences have revolutionized molecular biology by largely eliminating the need to clone and sequence genes involved in human health and disease. Instead of going to the bench, biologists now go to the web to look up gene sequences in public online databases.

But despite their immediate usefulness, the draft sequences were far from perfect. Both drafts were missing some 10% of the so-called 'euchromatin' — the gene-rich portion of the genome — and some 30% of the genome as a whole (which includes the gene-poor regions of 'heterochromatin'). The drafts contained hundreds of thousands of gaps, and had misassembled regions where portions of the genome were flipped or misplaced. As a result, any large-scale analyses of the genome, such as studies of the mechanisms of gene evolution or the long-range structure of the genome, had to contend with numerous uncertainties and artefacts. For example, studies of 'pseudogenes', the dying remnants of genes that have accumulated mutations that render them non-functional, had to contend with the possibility that any apparent pseudogene was instead the result of a sequencing error.

Since the publication of the drafts, the IHGSC sequencing centres have quietly undertaken a laborious 'finishing' process, in which each gap in the draft was individually examined and subjected to a battery of steps involving cloning and resequencing stretches of DNA. The sequence announced today has just 341 gaps remaining, and consists of contiguous runs of sequence averaging 38 million base pairs. The authors estimate that the finished sequence covers 99% of the euchromatic portion of the genome and that the overall error rate is less than 1 error per 100,000 base pairs. This substantially exceeds the original goals for the project.

The finishing procedure roughly doubled the total time and cost of the project. Does it contribute anything new to our understanding of the genome? It does indeed, and to prove the point the authors of the current paper[1] describe several large-scale analyses of the genome that would have been difficult to perform on the draft sequence. One analysis studied the processes of gene birth and death. The authors find 1,183 human genes that show evidence of having been recently 'born' by a process of gene duplication and divergence. They also find 37 genes that seem to have recently 'died' by acquiring a mutation that rendered the gene non-functional. The resulting pseudogene then slowly degrades and disappears.

In a second analysis, the authors use the finished sequence to map out segmental duplications — large regions of the genome that have duplicated in recent evolution. They find that 5% of the genome is involved in segmental duplications, and that the distribution of these regions varies widely across the chromosomes. Knowing the nature and extent of such duplications is important for understanding the evolution of the human genome, and for studying the many medically relevant disorders that are involved in segmental duplications, such as DiGeorge syndrome and Charcot–Marie–Tooth syndrome.

Another paper in this issue, by She *et al.*[4] (page 927), directly compares the outcomes of this second analysis with results obtained on an unfinished version of the human genome (an improved version of the Celera draft). She *et al.* find that the draft version artefactually 'simplifies' the genome by eliminating many duplicated regions. Their results bear on one of the highly publicized differences between the public and private genome projects. The public project used an older strategy in which the genome was first cloned into bacterial artificial chromosomes (BACs); the clones were then mapped, and each clone was sequenced and their sequences assembled individually. Celera championed an untested technique, 'whole-genome shotgun' (WGS), in which the entire genome was shattered into bite-size pieces, sequenced, and then assembled by software in one conceptually simple step.

Celera proved that the WGS technique is both technically feasible and provides a dramatic cost-saving over the clone-by-clone approach. Although the Celera draft has languished because the availability of public data in free online databases undermined the company's business plan to sell genome-database subscriptions, the effort left a permanent mark on the public project. Almost all genome-sequencing projects since then have used some form of WGS. The cautionary results contained in the new

papers from the IHGSC[1] and She *et al.*[4] argue for a hybrid strategy in which WGS is supplemented by a modest amount of BAC cloning and mapping. This would protect draft WGS sequences from some of the 'simplification' reported by She *et al.* and provide the clones needed for finishing selected regions of special interest.

What is next for the human genome project? Even with a finished sequence in hand there is much still to do. Surprisingly, one task is to develop the definitive catalogue of protein-coding genes. In the current paper[1], the number is estimated to be between 20,000 and 25,000. This wide range reflects limitations to state-of-the-art gene-prediction software that leave doubts about the validity of many predicted genes. One promising approach is to use comparative genomics to align the human genome with the genomes of other animals. Because natural selection ensures that functional regions are more highly conserved than non-functional ones, this approach highlights candidate protein-coding regions. The same approach shows promise for finding other functional elements such as gene promoters, which control the timing and level of expression of genes, and micro-RNAs, which have been implicated as regulatory agents of many developmental processes.

Much farther in the future is the task of sequencing the remaining 20% of the genome that lies within heterochromatin, the gene-poor, highly repetitive sequence that is implicated in the processes of chromosome replication and maintenance. The repetitiveness of heterochromatin means that it cannot be tackled using current sequencing methods, and new technologies will have to be developed to attack it. So don't be shocked to see another paper announcing the 'finishing' of the human genome in 2010 — it will describe how the heterochromatin problem has been cracked.

In sequencing the human genome, researchers have already climbed mountains and travelled a long and winding road. But we are only at the end of the beginning: ahead lies another mountain range that we will need to map out and explore as we seek to understand how all the parts revealed by the genome sequence work together to make life. ■

Lincoln D. Stein is at Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.
e-mail: lstein@cshl.edu

1. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
2. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
3. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
4. She, X. *et al. Nature* **431,** 927–930 (2004).

Comparative genomics

Small genome, big insights

John Mulley and Peter Holland

The genome of a second pufferfish species has been sequenced. Why is this important? Because comparing this genome with that of other animals yields a wealth of information on genome evolution.

It is still early days for the field of comparative genomics. Only around a dozen species of animal have so far had their complete DNA sequence determined, even to draft coverage. These are predominantly the widely studied model species, such as mice, fruitflies and nematode worms, or species of particular interest to humans, such as the malaria-carrying mosquito.

It may come as a surprise, therefore, to find that the list now includes not one, but two species of Tetraodontiformes, a relatively obscure group of fish also known as puffers. Following on from the publication two years ago of the genome sequence of the Japanese pufferfish *Takifugu rubripes*[1], Jaillon and colleagues[2] report, on page 946 of this issue, the near-complete sequence of the spotted green pufferfish *Tetraodon nigroviridis*. *Takifugu* is a poisonous marine fish best known to connoisseurs of sushi restaurants, whereas *Tetraodon* is a small, brackish-water pufferfish commonly kept in aquaria. But, like all Tetraodontiformes, the two species share a feature of great convenience for genomics: their cells possess less DNA than those of any other group of backboned animals — about eight or nine times less than human cells.

Although the *Tetraodon* genome is small compared with that of other vertebrates, sequencing it was still a hugely formidable task. The research reported in this issue[2] was performed in a collaboration between Genoscope in France and the Broad Institute of the Massachusetts Institute of Technology and Harvard University in the United States. Together they have generated a genome sequence of impressive accuracy and coverage, with 64% of the DNA sequence mapped to specific chromosomes[3].

By comparing the *Tetraodon* genome sequence with that of humans, Jaillon *et al.* even allow us to peer into the genome of the last common ancestor of pufferfish and humans — a primitive bony fish that lived hundreds of millions of years ago. The descendants of this long-extinct ancestor split into two distinct evolutionary lineages: the actinopterygians (ray-finned fish), which include teleosts such as pufferfish and zebrafish, and the sarcopterygians (lobe-fins), which include lungfish, coelacanths and ourselves (Fig. 1). By matching up the genes on each pufferfish chromosome to the related genes on human chromosomes, Jaillon *et al.* deduce that the extinct ancestor had 12 pairs of chromosomes. Work on partially completed genome sequences had suggested this number[4,5], but the new analyses add fascinating detail to the picture. For example, it is now possible to say which genes were on which chromosomes, despite this unknown animal having been extinct for more than 400 million years.

One puzzling observation concerns the apparent stability of the genomes of ray-finned fish. It seems that the ancestral genome underwent as few as ten large inter-chromosomal rearrangements (exchanges, fissions or fusions) to give rise to the present-day *Tetraodon* genome. Indeed, 11 *Tetraodon* chromosomes have not experienced any such rearrangements. Only one human chromosome (14) can make the same claim, despite the timescale being identical.

Although the genomes of ray-finned fish may have been slowly evolving in terms of chromosome breakages and fusions, they have experienced a cataclysmic event in their history. Jaillon and colleagues' analyses of the complete *Tetraodon* genome sequence show clearly that a duplication of the whole genome occurred sometime within the ray-finned-fish lineage. This inference is not new, having been previously suggested from analyses of the Hox-gene clusters and other gene families in zebrafish, *Takifugu* and other teleosts[4–7], but the conclusion has remained controversial[8].

Two new analyses should now settle the issue, however. First, Jaillon and colleagues plotted the chromosome positions for about 750 pairs of 'ancient' duplicated genes within the *Tetraodon* genome, revealing related pairs of chromosomes or chromosomal regions. Every chromosome is involved, consistent with an ancient whole-genome duplication. In the second test, chromosome positions for more than 6,000 pufferfish genes were compared with the positions of related genes in the human genome. This revealed a striking pattern of 'double conserved synteny', meaning that one chromosomal region in humans matches two in pufferfish, across the entire genome. This is a clear echo of whole-genome duplication in the ray-finned-fish lineage. Every gene, on every chromosome, was duplicated, although there has since been a massive degree of gene loss and local gene shuffling.

When did this whole-genome duplication occur? Analysis of zebrafish genetic maps strongly suggests that this species also underwent such an event in its history[4].