Multiple personal genomes await

Genomic data will soon become a commodity; the next challenge — linking human genetic variation with physiology and disease — will be as great as the one genomicists faced a decade ago, says **J. Craig Venter**.

Nearly ten years after Francis Collins and I stood at the White House with President Bill Clinton to announce the first two drafts of the human genome, the technology for DNA sequencing has progressed more dramatically than any of us could have predicted. The Human Genome Project took a worldwide effort and billions of dollars to reach what some had thought was an impossible goal. Today, thanks to innovation inspired in part by the race for the first draft between my company Celera Genomics, then in Rockville, Maryland, and the public effort led by Collins, it is possible to sequence a human genome in a day on a single machine for just a few thousand dollars.

Yet there is still some way to go before this capability can have a significant effect on medicine and health. As sequencing costs continue to plummet, data quality needs to improve. The generation of genomic data will have little value without corresponding phenotypic information about individuals' observable characteristics, and computational tools for linking the two. The challenges facing researchers today are at least as daunting as those my colleagues and I faced a decade ago.

In hindsight

The Human Genome Project was controversial from the start for several reasons, in particular the likelihood that it would divert funds from other biological research projects. Some of the early decisions had long-term effects on research strategies. In 1989, with sequencing costs projected to fall to US$1 a base pair within a few years, a group of those involved decided to ask the US Congress for $3 billion to cover the costs of sequencing a haploid genome consisting of 3 billion base pairs, rather than $6 billion to cover a diploid genome of 6 billion base pairs representing both sets of chromosomes, which was considered too expensive.

It was also agreed that, because of the hugely ambitious nature of the project, armies of scientists would be needed to sequence fragmented pieces of the genome. Once these decisions were made, there was little room for substantive innovation. I believe that history has proved they were a mistake. Although for example Leroy Hood argued that his first sequencing machines were equivalent to the Model A Ford and that a major effort was needed on technology development; the project moved forwards regardless.

In 1994, frustrated with the slow progress and inefficient use of labour, my team at the Institute for Genome Research in Rockville, Maryland, developed the 'whole-genome shotgun sequencing' approach, which we used to sequence an entire bacterial genome in three months[1]. Five years later at Celera we applied this approach to the *Drosophila* and human genomes[2,3]. Once highly controversial, whole-genome shotgun sequencing has been used for almost every genome sequenced since 2001.

> "The genome revolution is only just beginning."

Work at Celera was done in a single large facility with 300 automated DNA sequencers and a powerful computer. The public project used around 600 DNA sequencers distributed among several laboratories around the world. Together, the two projects gave a remarkable early insight into the human species, the most significant findings being the small number of human genes — 26,000 compared with earlier estimates of up to 300,000 — and the small amount of variation (0.1%) between individual humans[3,4].

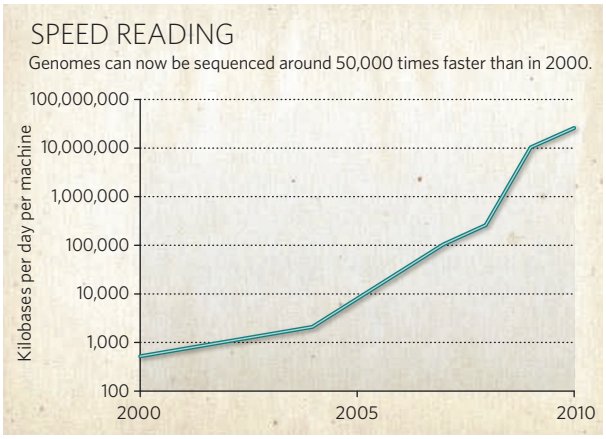
With the publication of the human genome drafts in 2001, many analysts predicted the end of the market for DNA sequencing technology. As we now know, a very different story unfolded. Sequencing centres turned to zoology, and the number of sequenced genomes of non-human species grew to today's tally of more than 3,800 (see 'Completed genomes'). At the same time, data from labs around the world continued to add to the draft human genome, resulting in an improved version in 2004 (ref. 5). My team concentrated on completing the sequencing of my personal genome, resulting in the publication of the first diploid human genome from an individual in 2007 (known as the HuRef genome)[6].

Bigger differences

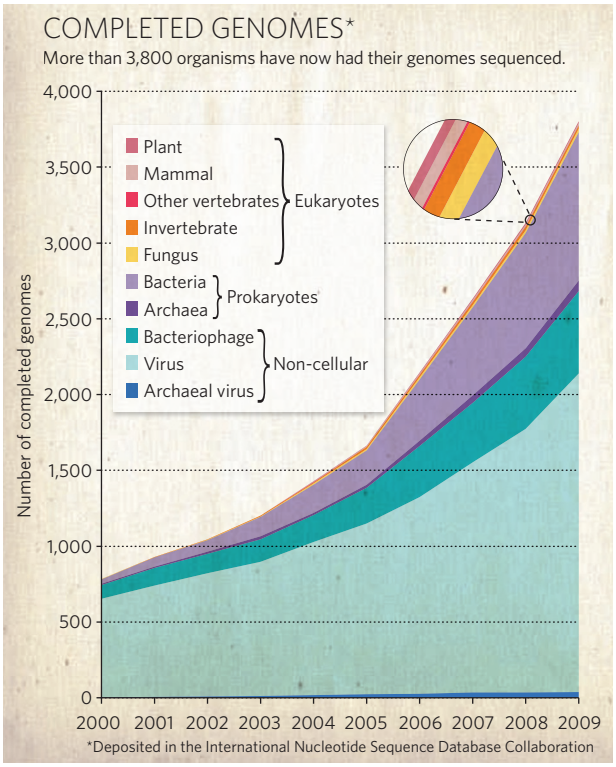
This first diploid human genome ushered in a new picture of human diversity. The sequence showed that my two parental genomes differed from each other by 0.5% when insertions and deletions of nucleotides in the DNA sequence were included along with single nucleotide polymorphisms (SNPs) — another common form of genetic variation. This was a dramatic increase over the 0.1% estimated in 2001 from looking at SNPs alone. It was subsequently discovered that the genomes of different individuals differ by between 1% and 3%.

Why did the data from the first two draft human genomes fail to show that individual genomes differ so significantly? The public effort was by design a haploid genome project. Construction of the haploid sequence involved sequencing cloned segments; there was therefore no way to directly detect polymorphisms, insertions and deletions. In contrast, the problem with the Celera programme was that there was too much genetic variation. The DNA came from two males and three females of various ethnicities including African American, Chinese, Hispanic and Caucasian[3]. Early versions of the Celera genome assembly software used a 'majority rule' approach to generate a single consensus sequence. This left out a substantial number of insertions and deletions from each genome. Had we used only one person's DNA, we would have had a much more complete understanding back then of the extent to which individuals vary genetically.

Despite the limitations of both projects, the race to sequence the



SOURCE: M. STRATTON/WELLCOME TRUST SANGER INST.



human genome inspired many in basic research labs and companies to develop the sequencing technologies that have emerged over the past few years. In 2003, to spur commercial and government investment, my institute set up a $500,000 prize to be awarded to the team that made the most substantial progress towards the sequencing of a human genome for $1,000 or less. This has since evolved into the $10-million Archon X-Prize, to go to the first team to build a device that can sequence 100 human genomes to a high degree of accuracy within 10 days for minimal costs[7].

These and other incentives rapidly accelerated the pace of innovation (see 'Speed reading'). Early this year, two companies, Illumina in San Diego, California, and Life Technologies in Carlsbad, California, announced sales of new sequencing instruments that can generate 25 billion base pairs per day and 100 billion base pairs per day, respectively. Life Technologies also announced a future version that will produce 300 billion base pairs per day. Both companies claim that their current instruments can sequence a human genome in a day for less than $6,000. Another company, Complete Genomics in Mountain View, California, maintains that it can sequence a human genome for $5,000–$8,000, but it is not producing instruments for sale. This incredible progress matches or exceeds anything that has happened in high-performance computing over the same period. Consider that the first Applied Biosystems sequencer in my National Institutes of Health lab in 1987 processed 4,800 base pairs a day. Twenty-three years on, the latest Life Technology instrument has improved on that by about eight orders of magnitude.

Yet these impressive increases have come with a big penalty. Most of the high-speed instruments sequence DNA in very short segments (or 'reads') of less than 100 base pairs at a time. This is significantly shorter than the reads produced by the first generation Sanger instruments, which manage 800–900 base pairs per read, or the second generation Roche 454 technology, whose reads approached 500 base pairs. Short reads greatly hamper one's ability to assemble sequences into long stretches representing the chromosomes. Sequencing groups have tried to overcome these limitations by layering their results on one of the already published human genome sequences instead of trying to assemble the whole sequence from scratch. This gives a distorted view of any single genome and, despite all the advances in processing, the resulting data quality is still well below diagnostic standards.

Improving data quality is crucial, because if a human genome cannot be independently assembled then the sequence data cannot be sorted into the two sets of parental chromosomes, or haplotypes. This process — haplotype phasing — will become one of the most useful tools in genomic medicine. Establishing the complete set of genetic information that we received from each parent is crucial to understanding the links between heritability, gene function, regulatory sequences and our predisposition to disease. Fortunately there are some exciting developments on the way that could help, such as new methods from Pacific Biosciences in Menlo Park, California, and Life Technologies that can produce sequence information from a single DNA strand. This approach promises sequence reads, in the range of thousands of base pairs, that will result in substantially higher-quality genome sequence data.

The next challenge

At the current rate of technological progress, DNA sequencing is soon likely to become a commodity, and the generation of cheap, high-quality sequence data will cease to be an issue. Phenotypes — the next hurdle — present a much greater challenge than genotypes because of the complexity of human biological and clinical information. The experiments that will change medicine, revealing the relationship between human genetic variation and biological outcomes such as physiology and disease, will require the complete genomes of tens of thousands of humans together with comprehensive digitized phenotype data. A simplistic query could be easily scored, for example, "do you have diabetes: yes or no". A more comprehensive view would include such things as age of onset and a scoring of the range of clinical manifestations associated with the disease, including extent of nerve damage, vascular issues, doses of medications used and family history. The scoring system would subcategorize characteristics such as disease type, progression and severity.

Even if we had all this information today, we wouldn't be able to make use of it because we don't have the computational infrastructure to compare even thousands of genotypes and phenotypes with each other. The need for such an analysis could be the best justification for building a proposed 'exascale' supercomputer, which would run 1,000 times faster than today's fastest computers. Scientists need to work together on a global basis to set the criteria for phenotype data; the incentive could come from academia, governments or industry.

Where will genomics be ten years from now? As sequencing capacity increases globally and the data quality improves, we will move beyond the current goal of one genome per person to sequencing multiple genomes per person from sources including sperm and egg cells, blastocysts, stem cells, pre-tumour cells and cancer cells. This will enable us to select healthy cells for reproduction and tissue transplants, or to better understand ageing and tumour development. Equally important for medical progress is the sequencing of the genomes of the millions of microbacteria that dwell within all of us[8]. The genome revolution is only just beginning. ■

J. Craig Venter is at the J. Craig Venter Institute, La Jolla, California 92121, USA.
e-mail: jcventer@jcvi.org

1. Fleischmann, R. D. *et al. Science* **269,** 496–512 (1995).
2. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
3. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
4. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
5. International Human Genome Sequencing Consortium *Nature* **431,** 931 (2004).
6. Levy, S. *et al. PLoS Biol.* **5,** e254 (2007).
7. http://genomics.xprize.org/archon-x-prize-for-genomics/prize-overview
8. NIH HMP Working Group *et al. Genome Res.* **19,** 2317–2323 (2009).

See Editorial, page 649, and human genome special at www.nature.com/humangenome.